



# Apprendre à représenter et à générer du texte en utilisant des mesures d'information

Pierre Colombo

## ► To cite this version:

Pierre Colombo. Apprendre à représenter et à générer du texte en utilisant des mesures d'information. Document and Text Processing. Institut Polytechnique de Paris, 2021. English. NNT : 2021IP-PAT033 . tel-03471220

**HAL Id: tel-03471220**

**<https://theses.hal.science/tel-03471220>**

Submitted on 8 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Learning to Represent and Generate Text using Information Measures

Thèse de doctorat de l'Institut Polytechnique de Paris  
préparée à Telecom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)  
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 17 November 2021, par

**PIERRE COLOMBO**

Composition du Jury :

Pierre Zweigenbaum CNRS/LISN, Orsay, France	President
Claire Gardent CNRS/LORIA, Nancy, France	Rapporteur
Alexandre Allauzen ESPCI, Paris, France	Rapporteur
Chloe Clavel Telecom Paris, Palaiseau, France	Directrice de thèse
Giovanna Varni Telecom Paris, Palaiseau, France	Co-encadrante de thèse
Loic Barrault University of Sheffield, Sheffield, United Kingdom	Examineur
Jackie Chi Kit Cheung Mila/McGill, Montreal Canada	Examineur
Pablo Piantanida CNRS/L2S, Gif-sur-Yvette, France	Examineur
Emmanuel Vignon Sicara	Invité
Joffrey Martinez IBM GBS	Invité

---

*Cette thèse est dédiée à ma famille qui est toujours là pour moi et particulièrement à ma mère qui est un exemple pour nous tous.*





# Résumé

Le traitement du langage naturel (NLP) permet la compréhension et la génération automatiques du langage naturel. Le traitement du langage naturel a récemment fait l'objet d'un intérêt croissant de la part de l'industrie et des chercheurs, car l'apprentissage profond (AD) a exploité la quantité stupéfiante de textes disponibles (*e.g* web, youtube, médias sociaux) et a atteint des performances similaires à celles de l'homme dans plusieurs tâches (*e.g* traduction, classification de textes). Par ailleurs, la théorie de l'information (TI) et la DL ont développé un partenariat de longue date. En effet, l'informatique a favorisé l'adoption des réseaux neuronaux profonds grâce à des principes célèbres tels que la longueur minimale de description (LMD), le goulot d'étranglement de l'information (GIO) ou le célèbre principe InfoMax. Dans tous ces principes, différentes mesures de l'information (*e.g* entropie, MI, divergences) sont l'un des concepts fondamentaux.

Dans cette thèse, nous abordons l'interaction entre le NLP et les mesures d'information. Nos contributions se concentrent sur deux types de problèmes PNL : la compréhension du langage naturel (NLU) et la génération du langage naturel (NLG). La NLU vise à comprendre et à extraire automatiquement des informations sémantiques d'un texte d'entrée, tandis que la NLG vise à produire un langage naturel à la fois bien formé (*i.e* grammaticalement correct, cohérent) et informatif.

La construction d'agents conversationnels parlés est un défi et le traitement des données conversationnelles parlées reste un problème difficile et négligé. Ainsi, nos premières contributions sont tournées vers l'UAL et nous nous concentrons sur l'apprentissage de représentations de transcriptions. Notre contribution se concentre sur l'apprentissage de meilleures représentations de transcriptions qui incluent deux caractéristiques importantes des conversations humaines parlées : la dimension conversationnelle et la dimension multimodale. Pour ce faire, nous nous appuyons sur diverses mesures d'information et nous tirons parti du principe de maximisation de l'information mutuelle. Le deuxième groupe de contributions aborde les problèmes liés au NLG. Cette thèse se concentre spécifiquement sur deux problèmes centraux. Premièrement, nous proposons une nouvelle limite supérieure de l'information mutuelle pour aborder le problème de la génération contrôlée via l'apprentissage de la représentation démêlée (transfert de style *i.e* et génération de phrases conditionnelles). Deuxièmement, nous abordons le problème de l'évaluation automatique des textes générés en développant une nouvelle famille de métriques utilisant diverses mesures d'information.

---

# Abstract

Natural language processing (NLP) allows for the automatic understanding and generation of natural language. NLP has recently received growing interest from both industry and researchers as deep learning (DL) has leveraged the staggering amount of available text (*e.g* web, youtube, social media) and reached human-like performance in several tasks (*e.g* translation, text classification). Besides, Information theory (IT) and DL have developed a long-lasting partnership. Indeed, IT has fueled the adoption of deep neural networks with famous principles such as Minimum Description Length (MDL), Information Bottleneck (IB) or the celebrated InfoMax principle. In all these principles, different measures of information (*e.g* entropy, MI, divergences) are one of the core concepts.

In this thesis, we address the interplay between NLP and measures of information. Our contributions focus on two types of NLP problems: natural language understanding (NLU) and natural language generation (NLG). NLU aims at automatically understand and extract semantic information from an input text where NLG aims at producing natural language that is both well-formed (*i.e* grammatically correct, coherent) and informative.

Building spoken conversational agents is a challenging issue and dealing with spoken conversational data remains a difficult and overlooked problem. Thus, our first contributions, are turned towards NLU and we focus on learning transcript representations. Our contribution focuses on learning better transcript representations that include two important characteristics of spoken human conversations: namely the conversational and the multi-modal dimension. To do so, we rely on various measures of information and leverage the mutual information maximization principle. The second group of contributions addresses problems related to NLG. This thesis specifically focuses on two core problems. First, we propose a new upper bound on mutual information to tackle the problem of controlled generation via the learning of disentangled representation (*i.e* style transfer and conditional sentence generation). Secondly, we address the problem of automatic evaluation of generated texts by developing a new family of metrics using various measures of information.

---

# Acknowledgments

I am grateful to all the persons that I have met during this PhD journey. I would like to express my gratitude to all of them for contributing one way or another to this PhD.

I would like to thank Pr. Clavel and Pr. Varni for being my thesis supervisors and I would like to address a special thanks to Emmanuel Vignon for the mentoring and the numerous efforts he invests in finding the necessary funding from IBM to create this thesis. I am also grateful to Geoffrey Martinez for taking the succession of Emmanuel on the IBM side and for the freedom he granted me.

I thank Pr. Gardent and Pr. Allauzen for accepting to review this thesis. I also thank Pr. Zweigenbaum, Pr. Barrault, Pr. Cheung, Pr. Piantanida, Emmanuel Vignon and Joffrey Martinez for having accepted to be part of my committee. I am very grateful to you, and hope that the reading of this manuscript will be pleasant for you.

This thesis has benefited from the expert and sharp eyes from Cyril and Gaël. I am grateful for their relevant suggestions.

I owe a lot to my numerous co-authors I learned a lot from Tanvi Dinkar, Florence D'Alche-Buc, Anne Sabourin, Marine Picot, Pablo Piantanida, Matthieu Labeau, Matteo Manica, Eric Gaussier, Slim Essid, Chouchang Jack Yang, Emile Chapuis, Alexandre Garcia, Hamid Jalalzai, Guillaume Staerman, Nathan Noiry, Gael Gibbon, Lucien Maman, Malik Boudiaf, Günther Koliander, Georg Pichler. I am thankful to all of them for the hard work and the fun moments we had, and we will have.

A special thanks Emile, Tanvi, Lucien, Charles T, Quentin, Junjie, Robin, Alexandre, Dimitri, Amaury, Ebenge, Louis, Myrto, Mathieu, Kimia, Nathan, Kamélia, Marc, Tamim, JR, Mastane, Vincent, Emilia, Nidham, Anas, Arturo, Cyril, Luc, Yannick and Rémi as well as other members of the Telecom Paris, D.R and IBM Research for the amazing moments and passionate discussions we shared over the years.

Of course, I owe a lot to all the people who shared my daily routine and I would like to give a special thanks to my friends Emile, Lucien, Amel, Alexandre, Alex, Nicolas (with many s), Mark Henry, Charles, Christophe, Enrique, Francesco and many other from Les Etoiles du 8eme who with I spent countless hours on the track over the years.

I would like to express my gratitude to people that had a decisive impact (most likely without knowing) on my decision to start a PhD: my mentor Matteo Manica, Hadrien, Louis, Marc, Mathilde and Prof. Geist, Prof Barret, Prof. Gastpar.

Last but not least, I am forever in debt with my parents Florence and Alain; my brothers François, Michel and Daniel; my grandparents Françoise (especially her punchlines on "superficial intelligence"), Edith, Jean and Adrien for my upbringing. I also have a special thought for all the rest of my family: my godmother Béatrice, aunts Valérie, Danielle, Michèle and uncles Marcels, Philippe, Rémi and all my cousins Marie, Stephs, Mathilde, Adrien, Olivier, Paul, Jean for supporting me over the years, they have contributed more than they will ever imagine to this tough journey.



# Contents

<b>Contents</b>	<b>9</b>
<b>List of Figures</b>	<b>11</b>
<b>List of Tables</b>	<b>15</b>
<b>1 Introduction and Overview</b>	<b>17</b>
1.1 Introduction . . . . .	17
1.2 Research Questions . . . . .	18
1.3 Organization of the Thesis . . . . .	20
1.4 List of Publications . . . . .	22
1.5 References . . . . .	23
<b>I Background</b>	<b>29</b>
<b>2 Measures of Information</b>	<b>31</b>
2.1 Shannon's Information Measures . . . . .	31
2.2 Computational Aspects of MI . . . . .	34
2.3 Contrastive Learning and NLP . . . . .	35
2.4 Beyond KL Divergence as a Measure of Similarity . . . . .	35
2.5 Multivariate Extensions . . . . .	38
2.6 References . . . . .	39
<b>3 Representing Textual Transcripts</b>	<b>45</b>
3.1 Importance of Conversational and Multimodal Dimensions to Learn Transcripts Representations . . . . .	45
3.2 Pretrained Representations and MI Maximization . . . . .	47
3.3 Supervised Fine-tuning For Multimodal Data . . . . .	50
3.4 References . . . . .	52
<b>4 Controlled sentence generation and automatic evaluation of NLG</b>	<b>59</b>
4.1 Controlled Sentence Generation . . . . .	59
4.2 Evaluation of NLG . . . . .	62
4.3 References . . . . .	65
<b>II Integrating Conversational and Multimodal Dimensions in Transcripts Representations via MI Maximization</b>	<b>71</b>
<b>5 Integrating Conversational Dimension in Pretrained Representation</b>	<b>75</b>
5.1 Introduction . . . . .	75

5.2	Method . . . . .	77
5.3	Information Theoretic Justification of Pretraining Losses . . . . .	83
5.4	Evaluation of Sequence Labelling . . . . .	83
5.5	Results on SILICONE . . . . .	86
5.6	Model Analysis . . . . .	89
5.7	References . . . . .	93
<b>6</b>	<b>Including multimodal dimension in representation of spoken transcripts</b>	<b>101</b>
6.1	Introduction . . . . .	101
6.2	Problem Formulation & Related Work . . . . .	103
6.3	Model and Training Objective . . . . .	105
6.4	Experimental Setting . . . . .	108
6.5	Numerical Results . . . . .	110
6.6	References . . . . .	116
<b>III</b>	<b>Text Generation using the Measures of Information</b>	<b>123</b>
<b>7</b>	<b>Learning to Disentangle Textual Representations and Attributes via MI</b>	<b>127</b>
7.1	Context . . . . .	127
7.2	Main Definitions and Related Works . . . . .	129
7.3	Model and Training Objective . . . . .	130
7.4	Experimental Setting . . . . .	134
7.5	Numerical Results . . . . .	135
7.6	References . . . . .	149
<b>8</b>	<b>Automatic Text Generation Evaluation</b>	<b>155</b>
8.1	Context . . . . .	155
8.2	InfoLM . . . . .	157
8.3	Experimental Frameworks . . . . .	158
8.4	Numerical Results on Summarization . . . . .	160
8.5	Numerical Results on Data2Text . . . . .	164
8.6	References . . . . .	166
<b>9</b>	<b>Conclusions, Limitations and Future Work</b>	<b>173</b>
9.1	Conclusions . . . . .	173
9.2	Limitation and Future Work . . . . .	174
9.3	References . . . . .	176
<b>A</b>	<b>Annexes</b>	<b>179</b>
A.1	Proofs of Chapter 7 . . . . .	179



# List of Figures

2.1	Venn Diagrams connecting entropy, conditional entropy, joint entropy and MI for two r.v X and Y. . . . .	33
5.1	General structure of our proposed hierarchical dialog encoder, with a decoder: $f_{\theta}^u$ , $f_{\theta}^d$ and the sequence label decoder ( $g_{\theta}^{dec}$ ) are colored respectively in green, blue and red. . . . .	78
5.2	This figure shows an example of corrupted context. Here $p_C$ is randomly set to 2 meaning that two utterances will be corrupted. $u_1$ and $u_4$ are randomly picked in 5.2b, 5.2d and then masked in 5.2c, 5.2e. . .	79
5.3	Schema of the different models evaluated on SILICONE. In this figure $f_{\theta}^u$ , $f_{\theta}^d$ and the sequence label decoder ( $g_{\theta}^{dec}$ ) are respectively colored in green, blue and red for the hierarchical encoder (see Figure 5.3a and Figure 5.3d). For BERT there is no hierarchy and embedding is performed through $f_{\theta}^u$ colored in grey (see Figure 5.3c, Figure 5.3d) . .	82
5.4	Histograms showing the utterance length for each dataset of SILICONE.	86
5.5	A comparison of pre-trained encoders being fine-tuned on different percentage the training set of SEM. Validation and test set are fixed over all experiments, reported scores are averaged over 10 different random splits. . . . .	89
5.6	Illustration of improvement of accuracy during pre-training stage on SEM for both a TINY and SMALL models. . . . .	91
5.7	A comparison of different parameters initialisation on MELD <sub>s</sub> . Training is performed using a different percentage of complete training set. Validation and test set are fixed over all experimentation. Each score is the averaged accuracy over 10 random runs. . . . .	92
6.1	Estimation of different dependency measures for multivariate Gaussian random variables for different degree of correlation. . . . .	105
6.2	Illustration of the method describes in Algorithm 2 for the different estimators derived from Theorem 1. $\mathcal{B}$ and $\tilde{\mathcal{B}}$ stands for the batch of data sample from the joint probability distribution and the product of the marginal distribution respectively. $Z_{lav}$ denotes the fusion representation of linguistic, acoustic and visual (resp. $l$ , $a$ and $v$ ) modalities provided by a multimodal architecture $f_{\theta_e}$ for the batch $\mathcal{B}$ . $\bar{Z}_{lav}$ denotes the same quantity as described before for the batch $\tilde{\mathcal{B}}$ . $A_{\theta_p}$ denotes the linear projection before classification or regression. . . . .	108

6.3	Study of the robustness of the representations against drop of the linguistic modality. Studied model is MAGBERT on CMU-MOSI. The ratio between the accuracy achieved with a corrupted linguistic modality $Acc_2^{corrupt}$ and the accuracy $Acc_2$ without any corruption is reported on $y$ -axis. The preserved modalities during inference are reported on $x$ -axis. A, V respectively stands for acoustic and visual modality. . . . .	114
6.4	Study of the robustness of the representations against a drop of the linguistic modality. Studied model is MAGXLNET on CMU-MOSI. The ratio between the accuracy achieved with a corrupted linguistic modality $Acc_2^{corrupt}$ and the accuracy $Acc_2$ without any corruption is reported on $y$ -axis. The preserved modalities during inference are reported on $x$ -axis. A, V respectively stands for the acoustic and visual modality. . .	115
7.1	Proposed methods. As described in Theorem 2. . . . .	132
7.2	Baselines methods, theses models use an adversarial loss for disentanglement. $f_{\theta_e}$ represents the input sentence encoder; $f_{\theta_e}^s$ denotes the style encoder (only used for sentence generation tasks); $C_{\theta_c}$ represents the adversarial classifier; $f_{\theta_d}$ represents the decoder that can be either a classifier (Figure 7.2a) or a sequence decoder (Figure 7.2b). . . . .	132
7.3	Numerical results on fair classification. Trade-offs between target task and attacker accuracy are reported in Figure 7.3a, Figure 7.3b for mention task, and Figure 7.3c, Figure 7.3d for sentiment task. For low values of $\lambda$ some points coincide. As $\lambda$ increases the level of disentanglement increases and the proposed methods using both KL (KL) and Reny divergences ( $\mathcal{D}_\alpha$ ) clearly offer better control than existing methods. . .	136
7.4	Disentanglement of representation learnt by $f_{\theta_e}$ in the binary (left) and multi-class ( <i>i.e.</i> , $ \mathcal{Y}  = 5$ ) (right) sentence generation scenario. In the multi-class scenario the <i>Adv</i> degenerates for $\lambda \geq 0.01$ and offer no fined-grained control over the degree of disentanglement. . . . .	138
7.5	Numerical experiments on binary style transfer. Quality of generated sentences are evaluated using BLEU (Figure 7.5a); style transfer accuracy (Figure 7.5a); sentence fluency (Figure 7.5c). We report existing trade-offs between disentanglement and sentence generation quality. Human evaluation is reported in Table 7.1. . . . .	138
7.6	Numerical experiments on conditional sentence generation. Results include BLEU (Figure 7.6a), style transfer accuracy (Figure 7.6b) and sentence fluency (Figure 7.6c). . . . .	138
7.7	Numerical experiments on multiclass style transfer using categorical labels. Results include: BLEU (Figure 7.7a); style transfer accuracy (Figure 7.7b); sentence fluency (Figure 7.7c); cosine similarity (Figure 7.7d)	139
7.8	Disentanglement of the learnt embedding when training an off-line adversarial classifier for the sentence generation with gender data. . .	141
7.9	Numerical experiments on binary style transfer using gender labels. Results include: BLEU (Figure 7.9a); cosine similarity (Figure 7.9d); style transfer accuracy (Figure 7.9b); sentence fluency (Figure 7.9c). . .	142
7.10	Numerical experiments on conditional sentence generation using gender labels. Results includes: BLEU (Figure 7.10a); cosine similarity (Figure 7.10d); style transfer accuracy (Figure 7.10b); sentence fluency (Figure 7.10c). . . . .	142

7.11 Numerical experiments on the multi-class conditionnal sentence generation. Results include: BLEU (Figure 7.11a); cosine similarity (Figure 7.11d); style transfer accuracy (Figure 7.11b); sentence fluency (Figure 7.11c). . . . .	143
7.12 Content preservation measured by the cosine similarity. . . . .	148
8.1 Results of the correlation between metrics and human judgement on the CNN dataset. First column reports the report correlations as measured by the Person ( $r$ ), second column reports Spearman ( $\rho$ ) and third column reports Kendall ( $\tau$ ) coefficient. . . . .	161
8.2 Pearson correlation at the system level between metrics when considering abstractive system outputs. . . . .	162
8.3 Impact of Calibration on system-level correlation (with Pearson ( $r$ ), Spearman ( $\rho$ ) or Kendall ( $\tau$ )) for CNN. The chosen measure is Rao as it is parameter-free. Calibration is changed using temperature scaling. . . . .	162
8.4 Impact of change in $\alpha$ and $\beta$ for $\mathcal{D}_{AB}$ . System level correlation, as measured by Pearson ( $r$ ) or Spearman ( $\rho$ ), is presented on abstractive (first column) and extractive system (second column). . . . .	164
8.5 Results of William's Significance Test: the tested hypothesis is: "is the increase of correlation significant". For clarity and due to space constraints the p-values are truncated and multiply par 100. Only p-values that are lower than 5.00 are displayed. . . . .	165



# List of Tables

2.1	Expression of the divergences (upper group) and distance between two positives measures $\mathbf{p} = (p_1, \dots, p_N)$ and $\mathbf{q} = (q_1, \dots, q_N)$ as well as the definition domain. For sake of clarity we omit the index in the summations. . . . .	36
3.1	Example of dialog taken from the Switchboard Dialog Act Corpus. . . .	47
3.2	Choices of the different parameters used in Equation 3.2 for the different pretraining objectives (borrowed from KONG and collab. [2019]) . .	49
5.1	Examples of dialogs labelled with DA taken from SwDA. The labels qw, sd, b, bk respectively correspond to wh-question, statement-non-opinion, backchannel and response acknowledgement. . . . .	77
5.2	Statistics of datasets composing SILICONE. E stands for emotion label and S for sentiment label; * stands for datasets with available official split. Sizes of Train, Val and Test are given in number of conversations. . . . .	85
5.3	Architecture hyperparameters used for the hierarchical pre-training. . .	87
5.4	Performances of different encoders when decoding using a MLP on SILICONE. The datasets are grouped by label type (DA vs E/S) and ordered by decreasing size. MT stands for Map Task, IEM for IEMOCAP and Sem for Semaine. . . . .	87
5.5	Experiments comparing decoder performances. Results are given on SILICONE for two types of baseline encoders (pre-trained BERT models and hierarchical recurrent encoders $\mathcal{H}\mathcal{R}$ ). . . . .	88
5.6	Results of ablation studies on SILICONE . . . . .	89
5.7	Comparison of GAP and MLM with a comparable number of parameters. For all models a MLP decoder is used on top of a TINY pre-trained encoder. . . . .	91
5.8	Number of parameters for the encoders. Sizes are given in million of parameters. . . . .	92
6.1	Statistics network description. $d_{in}$ denotes the dimension of $Z_{avl}$ . . . .	110
6.2	Results on sentiment analysis on both CMU-MOSI and CMU-MOSEI for a EF-LSTM. $Acc_7$ denotes accuracy on 7 classes and $Acc_2$ the binary accuracy. MAE denotes the Mean Absolute Error and $Corr$ is the Pearson correlation. $^h$ means higher is better and $^l$ means lower is better. The choice of the evaluation metrics follows standard practices [RAHMAN and collab., 2020]. Underline results demonstrate significant improvement (p-value belows 0.05) against the baseline when performing the Wilcoxon Mann Whitney test [WILCOXON, 1992] on 10 runs using different seeds. . . . .	112

6.3	Results on sentiment and emotion prediction on both CMU-MOSI and CMU-MOSEI dataset for the different neural architectures presented in section 6.4 relying on various fusion mechanisms. . . . .	113
6.4	Examples from the CMU-MOSI dataset using MAGBERT. The last column is computed using the statistical network $T_\theta$ . L stands for low values and H stands for high values. Green, grey, red highlight positive, neutral and negative expression/behaviours respectively . . . . .	113
6.5	Examples from the CMU-MOSI dataset using MAGXLNET trained with $\mathcal{L}_W$ . The last column is computed using the statistic network $T_\theta$ . L stands for low values and H stands for high values. Green, grey, red highlight positive, neutral and negative expression/behaviours respectively. . . .	115
7.1	Human annotations of generated samples. For the comparison we rely on the sentences provided in <a href="https://github.com/rpryzant/delete_retrieve_generate">https://github.com/rpryzant/delete_retrieve_generate</a> . Human annotations are also provided by LI and collab. [2018]. We have reprocessed the provided sentence using a tokenizer based on SentencePiece [KUDO, 2018; SENNRICH and collab., 2016]. Since there is a trade-off between automatic evaluation metrics ( <i>i.e</i> BLEU, Perplexity and Accuracy of Style Transfer), we set minimum thresholds on BLEU and on style transfert accuracy. The best model that met the threshold on validation is selected. We will release—along with our code—new generated sentences for comparison.	140
7.2	Sequences generated by the different models on the binary sentiment transfer task. . . . .	145
7.3	Sequences generated by the different models on the binary sentiment conditional sentence generation task. . . . .	147
8.1	Correlation at the system level with human judgement along five different axis: correctness, data coverage, fluency, relevance and text structure for the WebNLG task. Best results by group are underlined, overall best results are bolted. . . . .	165

# Chapter 1

## Introduction and Overview

### 1.1 Introduction

Natural Language Processing (NLP) allows computers to automatically read, interpret and generate natural language. Today's algorithms can automatically analyze, classify and generate texts in a consistent way. Considering the staggering amount of textual data that is generated every day on various online platforms (e.g social network, online marketplace, transcripts of conversations) improving automated processing to analyze textual data in an efficient manner will be critical. This task is made complex by the variability of both spoken and written human language [EVANS and LEVINSON, 2009]. On top of that, algorithms have to be flexible and robust to multiple undesirable variations such as misspelling [HU and collab., 2021], abbreviations [MOON and collab., 2012], lack of punctuation [EK and collab., 2020], typos [DUTREY and collab., 2012] when dealing with written texts, or stutters [LU and collab., 2018], disfluencies [DUTREY and collab., 2014], presence of fillers (e.g. “um” or “uh”) [DINKAR and collab., 2018, 2020], transcript errors [PYE and collab., 1988] when processing spoken transcripts. Because of the almost infinite variability of language, the most thriving approaches in NLP are nowadays fully data-driven. Those systems keep improving while increasing the amount of data as they are exposed to more diverse linguistic variations [LIU and collab., 2019]. In NLP, data driven systems relying on neural networks have been widely adopted since they have reached state-of-the art results and human like performances on many NLP tasks (e.g translation, sentiment analysis) [SEJNOWSKI, 2020]. Thus nowadays, the use of data driven methods such as neural networks is one of the dominant paradigms.

The breakthrough and the adoption of deep neural networks have been fueled by Information Theory (IT) [SHANNON, 2001]. Perhaps the most famous loss to train neural networks is the cross-entropy loss which can be linked to a specific measure of information named Entropy introduced by Shannon [COVER, 1999]. Entropy which lies at the root of IT measures the information and the redundancy contained in a message. As an example, Shannon found 11.82 bits per word when computing the entropy of English over 8000 words [SHANNON, 1951]. Over the years multiple concepts of IT has encountered much success when applied to neural networks: most recent embeddings [CLARK and collab., 2020] take inspiration in the noisy channel model [VINCENT and collab., 2010] and maximize a lower bound of mutual information (MI) [KONG and collab., 2019], tokenizers use concepts from universal source coding [GAGE, 1994], neural networks have been explained using the information bottleneck principle [SAXE and collab., 2019], deep architectures [VELICKOVIC and collab.,

2019] are trained using InfoMax principle [CARDOSO, 1997], neural networks are compressed using lossless compression [WIEDEMANN and collab., 2020]. In all these applications, different measures of information (*e.g* entropy, MI, divergences) are one of the core concepts.

This thesis addresses the interplay between two different classes of NLP problems and measures of information. The first class gathers problems related to representation learning for natural language understanding (NLU). For this class of problem, our work is inscribed within the MI maximization framework which aims at learning representations by maximizing the MI between the inputs and a latent representation of the encoder. The second class is related to natural language generation (NLG) and our work mainly focuses on two problems namely controlled generation (*i.e* style transfer) and automatic evaluation of text generation.

In the next section we present our two sets of research questions RQ1 and RQ2, where RQ1 is turned towards NLU and RQ2 is dedicated to NLG.

## 1.2 Research Questions

### 1.2.1 RQ1: How to adapt the MI maximization principle to learn transcripts representations with conversational and multimodal dimensions?

One of the big payoffs of deep learning is to allow the learning of a higher level of abstraction which allows both better generalization and better transfer [BENGIO and collab., 2007]. These abstract representations are deeply linked to the invariant in the data distribution [BENGIO and collab., 2013]. The MI maximization principle, which leverages the invariance property of MI, has been successfully applied to learning representation of diverse types of data, including text [DEVLIN and collab., 2018; MIKOLOV and collab., 2013a]. However, these studies are mainly focusing on written text [MEHRI and collab., 2019]. There is a pressing need for spoken conversational agents [CHEN and collab., 2017] as the field of business have shown a growing interest in using them to improve both service quality and market competitiveness [GAO and collab., 2018a]. Thus, it is interesting to adapt the MI maximization principle to conversational data. Conversations are well-structured interaction [ARORA and collab., 2013], they are a sequence of turns (or utterances) which contains a variable number of words. Additionally, augmenting text with additional modalities (*e.g* audio, video) is of high importance in the context of designing conversational agents as interactions are intrinsically multimodal and thus multimodal signal carries more information than the commonly used textual representation [MORENCY and BALTRUŠAITIS, 2017]. As an example, prosodic cues (*e.g* change of pitch, laughs, pauses) and corporal expressions (*e.g* gaze, gestures) are carried by the audio and video respectively.

The two aforementioned characteristics of human conversations (*i.e* conversational nature and multi-modalities) are among the challenges that researchers need to address while building generic representations for conversational agent [GAO and collab., 2018b]. Thus, RQ1 is linked to the generic questions: *how to adapt the MI maximization principle for learning better transcript representations and including conversational information? How to enhance representation with additional information carried by multimodal signal (e.g audio, video)?*



**Learning generic text transcript representations with conversational dimension.** Available generic representations such as Word2Vect [MIKOLOV and collab., 2013b], Glove [PENNINGTON and collab., 2014] or BERT [DEVLIN and collab., 2018] have been shown to be an effective way to achieve state-of-the-art results on written benchmarks. However, they are not suited to the hierarchical structure of conversations which can not be considered flat contrarily to non-conversational text. This setting raises the following sub-questions:

- How to adapt the MI maximization principle to the hierarchy of conversations and to build generic representation for transcripts that takes into account the specifics of dialogue?
- What are the consequences of introducing hierarchy? How can this inductive bias be further leveraged to improve the learning phase?

**Enriching transcripts representation with additional modalities.** In multi-modal learning, one of the difficulties is to join information from the different modalities. The information coming from various sources has different nature and thus a fusion mechanism is needed. A good fusion mechanism should retain as much information as possible from different modalities [GAO and collab., 2020]. This new setting raises the following sub-questions:

- Does it make sense to apply the MI maximization principle to learn representations of multi-modal conversations?
- If so, how can we adapt it to multi-modal data?
- What new properties are learnt by the representations when using the MI maximization principle?

### 1.2.2 RQ2: How to use the geometrical properties of the measures of information to generate and evaluate generated text?

In NLG, the goal is to produce natural language that is both well formed (*i.e* grammatically correct, coherent) and informative [GATT and KRAHMER, 2018]. Both the nature of input data and the output highly depends on the application. Popular types of business applications include producing personalized text [COLOMBO and collab., 2019], translating texts [SON and collab., 2012], summarizing documents [ALLAHYARI and collab., 2017]. Since, the text is highly variable and there are multiple ways to express the same idea, automatic evaluation of text generation systems is also a challenging problem: text quality needs to be assessed along multiple axes (*e.g* informativeness, relevance coherence) where each axis is task-specific. One of the flexibilities of using different types of information measure (*e.g* MI, Fisher-Rao distance [ATKINSON and MITCHELL, 1981], F-divergences [SASON and VERDÚ, 2016]) is the ability to measure different geometrical properties of the distributions [AMARI and CICHOCKI, 2010]. Thus the use of these measures seems to be particularly suited for NLG where variability and diversity plays a major role. The set of RQ2 is related to the following research questions: *can we use the geometrical properties of the measures of information to control different aspects of text generation? Since the different measures of information measure different properties of the distributions, what are the best measures of information to automatically assess the quality of generated texts?*

**Generating Textual Data.** For this problem, we focus on controlled text generation. A popular application is style transfer which aims at controlling several factors of the generated text. Examples of factors include style formality [RAO and TETREAULT, 2018], polarity [HU and collab., 2017], sarcasm [MISHRA and collab., 2019], gender [PRABHUMOYE and collab., 2018] or product type [LAMPLE and collab., 2018]. One of the existing dominant approaches in the context of text data has been to learn to embed the input sentence into a style-independent vector. This vector, along with the desired attribute is feed to a decoder that generate a new sentence. As previously mentioned, measures of information such as MI are related to the invariant. Our research subquestions boil down to:

- What conditions can we introduce to learn disentangled representations to remove attribute information from the latent space?
- How do these conditions affect the learned representations? What is the trade-off that exists between the disentanglement and the quality of the representations?

**Evaluation of Text Generation.** The goal of NLG is to generate coherent, readable and informative texts from some input data (*e.g.*, texts, images and tables). However, the exact definition of each of these three criteria remains task-dependent and thus, making it hard to provide a unique metric for all tasks. As an example, NMT focuses on fluency, fidelity and adequate [HOVY, 1999; WHITE and collab., 1994] in contrast to summarization where annotators have to focus on coherence, content, readability, syntactic coherence and conciseness [MANI, 2001]. Thus this setting raises the following sub-questions:

- Can we use the measure of information to propose a new metric that automatically evaluates text generation?
- Are the measures of information flexible enough to correlate well with different task-specific criteria?

## 1.3 Organization of the Thesis

This thesis presentation focuses on the use of using the measure of information with application to NLP using deep neural networks. After introducing the measure of information and presenting relevant previous work on representation learning and generation of textual data (see [Part I](#)), we focus in [Part II](#) on the problem of using MI (a specific measure of information) to learn better representation. In [Part III](#), we tackle two NLG problems using the measure of information.

[[Part I](#)]. In this first part, we introduce the background and mathematical tools related to IT and NLP useful to understand the contributions in [Part II](#) and [Part III](#).

[[Chapter 2](#)]: A plethora of measures of information has been used in various applications. We start by defining the entropy and its derivatives (conditional entropy, differential entropy) and then move to MI which is a central concept of this thesis. In many cases, the exact computation of the MI is intractable so we review the most used alternatives which rely on variational bounds. Then, we focus on the connections

between MI and KL divergence to introduce other types of information measures in the discrete case. These information measure will be useful in [Chapter 8](#).

[[Chapter 3](#)]: In this chapter, we explore previous work linked to the learning representation of textual data in NLU. State of the art techniques rely on pretrained representations that are learnt through self supervised objectives. We draw connections between these objectives and MI information maximization and discuss the limitations of current pretraining objectives. The second part of the chapter will be useful in [Chapter 3](#) and is dedicated to learning multimodal representations.

[[Chapter 4](#)]: In this chapter, we present the two different NLG problems we will address, namely style transfer and automatic evaluation of text generation. The first problem involves learning disentangled representations. The learning of such disentangled representations can be set as a multitask learning problem where the first term is a task specific term and the second term involves computing the MI. In the second problem, we recall the framework of automatic evaluation and review existing metrics.

[[Part II](#)]. The second part gathers the contributions related to the RQ1. MI maximization is applied to learn transcripts representations on two different dimension namely interactional and multimodal.

[[Chapter 5](#)]: The first setting involves learning representations of transcripts that integrate the conversational dimension. In this chapter, we propose two sets of new losses that can be connected to the MI maximization framework introduced in [Chapter 2](#). These new pretraining losses are tailored for spoken dialog and allow the model to learn the hierarchical nature of the data. The novel form of these losses has a direct influence on the choice of the deep neural network architecture. Additionally, the new pretraining objective allows to reduce the number of parameters and train the representation at a reduced cost. The experiments are conducted using various corpora composed of spoken dialogues.

[[Chapter 6](#)]: In the second setting, we focus on learning representations that integrate the multimodal dimension. In this chapter, we show how to leverage the MI maximization principle as an alternative to complex fusion mechanisms. Our method involves using the total correlation or multivariate MI introduced in [Chapter 2](#). Not only our solution can be used as an alternative to complex fusion mechanism but it also improves the fusion of state of the art models (presented in [Chapter 3](#)). We show that the resulting representations are more robust but can also be better explained. Experiments are conducted using multimodal corpora composed of monologues (a particular type of interaction where only one speaker is involved).

[[Part III](#)]. In this third part, we gather the contributions related to the RQ2 and apply the measures of information to generate and evaluate generated texts.

[[Chapter 7](#)]: This chapter gathers our contributions to learning disentangled representations for style transfer. In this chapter, we develop a new trainable upper bound on MI. We start by experimenting with this bound on fair classification where we find that our bound does not suffer from existing problems of existing MI estimators (e.g saturation, degeneracy). Then we experiment on textual style and show that our new method achieves a better trade-off while allowing to reach better disentangled representations. As a matter of fact, there is no free-lunch [[WOLPERT](#)










and MACREADY, 1997] for sentence generation tasks: although transferring style is easier with disentangled representations, it also removes important information about the content.


[Chapter 8]: In the last chapter, we study the use of discrete measures of information (see Chapter 3) to build automatic metrics. Currently, two main categories of untrained metrics can be distinguished: word or character based-metrics that compute a score based on string representation of the texts and embedding-based metrics that rely on a continuous representation of the text. In this chapter, we propose a new metric, that belongs to both classes. This metric called InfoLM leverages different measures of information and a pretrained language model it outperforms available untrained metrics on both summarization and data2text generation.

The following references have been published during the thesis, underlined references are discussed in this thesis.


## 1.4 List of Publications

### 1.4.1 Conferences



1.  **P. Colombo**, C. Clavel and P. Piantanida. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. **AAAI 2022**
2.  **P. Colombo\***, E. Chapuis\*, M. Labeau, and C. Clavel. Code-switched inspired losses for generic spoken dialog representations. **EMNLP 2021**
3.  **P. Colombo**, G. Staerman, C. Clavel and P. Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. **EMNLP 2021**
4.  **P. Colombo**, E. Chapuis, M. Labeau, and C. Clavel. Improving multimodal fusion via mutual dependency maximisation. **EMNLP 2021**
5.  **P. Colombo**, C. Clavel and P. Piantanida. A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations (oral) **ACL 2021**
6.  E. Chapuis\*, **P. Colombo\***, M. Manica, M. Labeau, and C. Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. **Finding of EMNLP 2020**
7.  T. Dinkar\*, **P. Colombo\***, M. Labeau, and C. Clavel. The importance of fillers for text representations of speech transcripts. **EMNLP 2020**
8.  H. Jalalzai\*, **P. Colombo\***, C. Clavel, E. Gaussier, G. Varni, E. Vignon, and A. Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. **NeurIPS 2020**
9.  **P. Colombo\***, E. Chapuis\*, M. Manica, E. Vignon, G. Varni, and C. Clavel. Guiding attention in sequence-to-sequence models for dialogue act prediction. **(oral) AAAI 2020**

10.  A. Garcia\*, **P. Colombo\***, S. Essid, F. d'Alché-Buc, and C. Clavel. From the token to the review: A hierarchical multimodal approach to opinion mining. **EMNLP 2019**
11.  **P. Colombo\***, W. Witon\*, A. Modi, J. Kennedy, and M. Kapadia. Affect-driven dialog generation. **NAACL 2019**

#### 1.4.2 Workshop

1.  W. Witon\*, **P. Colombo\***, A. Modi, and M. Kapadia. Disney at IEST 2018: Predicting emotions using an ensemble. In Proceedings of the 9th Workshop WASSA@EMNLP 2018

#### 1.4.3 Preprints

1.  Georg Pichler\*, **P. Colombo\***, Malik Boudiaf\*, Günther Koliander, Pablo Piantanida. KNIFE: Kernelized-Neural Differential Entropy Estimation. Submitted at **NeurIPS 2021**
2.  **P. Colombo**, C Yang, G. Varni, and C. Clavel. Beam search with bidirectional strategies. 2020

#### 1.4.4 Patent

1. Affect Driven Dialog Generation, A. Modi, M. Kapadia, Douglas A. Fidaleo, J. Kennedy, W. Witon and **P. Colombo**, US Patent 16,226,166, A framework for Affective Conversational System

### 1.5 References

- ALLAHYARI, M., S. POURIYEH, M. ASSEFI, S. SAFAEI, E. D. TRIPPE, J. B. GUTIERREZ and K. KOCHUT. 2017, «Text summarization techniques: a brief survey», *arXiv preprint arXiv:1707.02268*. [19](#)
- AMARI, S.-I. and A. CICHOCKI. 2010, «Information geometry of divergence functions», *Bulletin of the polish academy of sciences. Technical sciences*, vol. 58, n° 1, p. 183–195. [19](#)
- ARORA, S., K. BATRA and S. SINGH. 2013, «Dialogue system: A brief review», *arXiv preprint arXiv:1306.4134*. [18](#)
- ATKINSON, C. and A. F. MITCHELL. 1981, «Rao's distance measure», *Sankhyā: The Indian Journal of Statistics, Series A*, p. 345–365. [19](#)
- BENGIO, Y., A. COURVILLE and P. VINCENT. 2013, «Representation learning: A review and new perspectives», *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, n° 8, p. 1798–1828. [18](#)
- BENGIO, Y., Y. LECUN and collab.. 2007, «Scaling learning algorithms towards ai», *Large-scale kernel machines*, vol. 34, n° 5, p. 1–41. [18](#)

- CARDOSO, J.-F. 1997, «Infomax and maximum likelihood for blind source separation», *IEEE Signal processing letters*, vol. 4, n° 4, p. 112–114. [18](#)
- CHEN, H., X. LIU, D. YIN and J. TANG. 2017, «A survey on dialogue systems: Recent advances and new frontiers», *Acm Sigkdd Explorations Newsletter*, vol. 19, n° 2, p. 25–35. [18](#)
- CLARK, K., M.-T. LUONG, Q. V. LE and C. D. MANNING. 2020, «Electra: Pre-training text encoders as discriminators rather than generators», *arXiv preprint arXiv:2003.10555*. [17](#)
- COLOMBO, P., W. WITON, A. MODI, J. KENNEDY and M. KAPADIA. 2019, «Affect-driven dialog generation», *arXiv preprint arXiv:1904.02793*. [19](#)
- COVER, T. M. 1999, *Elements of information theory*, John Wiley & Sons. [17](#)
- DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805*. [18](#), [19](#)
- DINKAR, T., I. VASILESCU, C. PELACHAUD and C. CLAVEL. 2018, «Disfluencies and teaching strategies in social interactions between a pedagogical agent and a student: Background and challenges», in *SEMDIAL 2018 (AixDial), The 22nd workshop on the Semantics and Pragmatics of Dialogue*, Laurent Prévot, Magalie Ochs and Benoît Favre, p. 188–191. [17](#)
- DINKAR, T., I. VASILESCU, C. PELACHAUD and C. CLAVEL. 2020, «How confident are you? exploring the role of fillers in the automatic prediction of a speaker’s confidence», in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 8104–8108. [17](#)
- DUTREY, C., C. CLAVEL, S. ROSSET, I. VASILESCU and M. ADDA-DECKER. 2014, «A crf-based approach to automatic disfluency detection in a french call-centre corpus», in *Fifteenth Annual Conference of the International Speech Communication Association*. [17](#)
- DUTREY, C., A. PERADOTTO and C. CLAVEL. 2012, «Analyse de forums de discussion pour la relation clients: du text mining au web content mining», *Actes JADT*. [17](#)
- EK, A., J.-P. BERNARDY and S. CHATZIKYRIAKIDIS. 2020, «How does punctuation affect neural models in natural language inference», in *Proceedings of the Probability and Meaning Conference (PaM 2020)*, p. 109–116. [17](#)
- EVANS, N. and S. C. LEVINSON. 2009, «The myth of language universals: Language diversity and its importance for cognitive science», *Behavioral and brain sciences*, vol. 32, n° 5, p. 429–448. [17](#)
- GAGE, P. 1994, «A new algorithm for data compression», *C Users Journal*, vol. 12, n° 2, p. 23–38. [17](#)
- GAO, J., M. GALLEY and L. LI. 2018a, «Neural approaches to conversational ai», in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 1371–1374. [18](#)



- GAO, J., M. GALLEY and L. LI. 2018b, «Neural approaches to conversational ai», in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 1371–1374. [18](#)
- GAO, J., P. LI, Z. CHEN and J. ZHANG. 2020, «A survey on deep learning for multimodal data fusion», *Neural Computation*, vol. 32, n° 5, p. 829–864. [19](#)
- GATT, A. and E. KRAHMER. 2018, «Survey of the state of the art in natural language generation: Core tasks, applications and evaluation», *Journal of Artificial Intelligence Research*, vol. 61, p. 65–170. [19](#)
- HOVY, E. H. 1999, «Toward finely differentiated evaluation metrics for machine translation», in *Proceedings of the EAGLES Workshop on Standards and Evaluation Pisa, Italy, 1999*. [20](#)
- HU, Y., X. JING, Y. KO and J. T. RAYZ. 2021, «Misspelling correction with pre-trained contextual language model», *arXiv preprint arXiv:2101.03204*. [17](#)
- HU, Z., Z. YANG, X. LIANG, R. SALAKHUTDINOV and E. P. XING. 2017, «Toward controlled generation of text», in *International Conference on Machine Learning*, PMLR, p. 1587–1596. [20](#)
- KONG, L., C. D. M. D’AUTUME, W. LING, L. YU, Z. DAI and D. YOGATAMA. 2019, «A mutual information maximization perspective of language representation learning», *arXiv preprint arXiv:1910.08350*. [17](#)
- LAMPLE, G., S. SUBRAMANIAN, E. SMITH, L. DENOYER, M. RANZATO and Y.-L. BOUREAU. 2018, «Multiple-attribute text rewriting», in *International Conference on Learning Representations*. [20](#)
- LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTMAYER and V. STOYANOV. 2019, «Roberta: A robustly optimized bert pretraining approach», *arXiv preprint arXiv:1907.11692*. [17](#)
- LU, S., Y. ZHU, W. ZHANG, J. WANG and Y. YU. 2018, «Neural text generation: Past, present and beyond», *arXiv preprint arXiv:1803.07133*. [17](#)
- MANI, I. 2001, *Automatic summarization*, vol. 3, John Benjamins Publishing. [20](#)
- MEHRI, S., E. RAZUMOVSKAIA, T. ZHAO and M. ESKENAZI. 2019, «Pretraining methods for dialog context representation learning», *arXiv preprint arXiv:1906.00414*. [18](#)
- MIKOLOV, T., K. CHEN, G. CORRADO and J. DEAN. 2013a, «Efficient estimation of word representations in vector space», *arXiv preprint arXiv:1301.3781*. [18](#)
- MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. S. CORRADO and J. DEAN. 2013b, «Distributed representations of words and phrases and their compositionality», in *Advances in neural information processing systems*, p. 3111–3119. [19](#)
- MISHRA, A., T. TATER and K. SANKARANARAYANAN. 2019, «A modular architecture for unsupervised sarcasm generation», in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 6146–6155. [20](#)

- MOON, S., S. PAKHOMOV and G. B. MELTON. 2012, «Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations», in *AMIA annual symposium proceedings*, vol. 2012, American Medical Informatics Association, p. 1310. 17
- MORENCY, L.-P. and T. BALTRUŠAITIS. 2017, «Multimodal machine learning: integrating language, vision and speech», in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, p. 3–5. 18
- PENNINGTON, J., R. SOCHER and C. D. MANNING. 2014, «Glove: Global vectors for word representation», in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 19
- PRABHUMOYE, S., Y. TSVETKOV, R. SALAKHUTDINOV and A. W. BLACK. 2018, «Style transfer through back-translation», *arXiv preprint arXiv:1804.09000*. 20
- PYE, C., K. A. WILCOX and K. A. SIREN. 1988, «Refining transcriptions: the significance of transcriber ‘errors’», *Journal of Child Language*, vol. 15, n° 1, p. 17–37. 17
- RAO, S. and J. TETREAULT. 2018, «Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer», in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, p. 129–140, doi: 10.18653/v1/N18-1012. URL <https://www.aclweb.org/anthology/N18-1012>. 20
- SASON, I. and S. VERDÚ. 2016, « $f$ -divergence inequalities», *IEEE Transactions on Information Theory*, vol. 62, n° 11, p. 5973–6006. 19
- SAXE, A. M., Y. BANSAL, J. DAPELLO, M. ADVANI, A. KOLCHINSKY, B. D. TRACEY and D. D. COX. 2019, «On the information bottleneck theory of deep learning», *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, n° 12, p. 124 020. 17
- SEJNOWSKI, T. J. 2020, «The unreasonable effectiveness of deep learning in artificial intelligence», *Proceedings of the National Academy of Sciences*, vol. 117, n° 48, p. 30 033–30 038. 17
- SHANNON, C. E. 1951, «Prediction and entropy of printed english», *Bell system technical journal*, vol. 30, n° 1, p. 50–64. 17
- SHANNON, C. E. 2001, «A mathematical theory of communication», *ACM SIGMOBILE mobile computing and communications review*, vol. 5, n° 1, p. 3–55. 17
- SON, L. H., A. ALLAUZEN and F. YVON. 2012, «Continuous space translation models with neural networks», in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 19
- VELICKOVIC, P., W. FEDUS, W. L. HAMILTON, P. LIÒ, Y. BENGIO and R. D. HJELM. 2019, «Deep graph infomax.», in *ICLR (Poster)*. 17



- VINCENT, P., H. LAROCHELLE, I. LAJOIE, Y. BENGIO, P.-A. MANZAGOL and L. BOTTOU. 2010, «Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.», *Journal of machine learning research*, vol. 11, n° 12. [17](#)
- WHITE, J. S., T. A. O'CONNELL and F. E. O'MARA. 1994, «The arpa mt evaluation methodologies: evolution, lessons, and future approaches», in *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. [20](#)
- WIEDEMANN, S., H. KIRCHHOFFER, S. MATLAGE, P. HAASE, A. MARBAN, T. MARINČ, D. NEUMANN, T. NGUYEN, H. SCHWARZ, T. WIEGAND and collab.. 2020, «Deep-cabac: A universal compression algorithm for deep neural networks», *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, n° 4, p. 700–714. [18](#)
- WOLPERT, D. H. and W. G. MACREADY. 1997, «No free lunch theorems for optimization», *IEEE transactions on evolutionary computation*, vol. 1, n° 1, p. 67–82. [21](#)



# **Part I**

## **Background**



# Chapter 2

## Measures of Information

### Chapter 2 abstract

This thesis explores the use of measures of information for two NLP problems: representation learning and generation of textual data. In this chapter, we present formal definitions of the various measures of information that we will use in the rest of this thesis. We start by reviewing the well-known Shannon's information measures namely the entropy, the conditional entropy and the mutual information. MI is one of the most important measures of information in IT, when used to train deep neural networks they are often intractable, thus we often rely on surrogates. In the second part, we review current techniques to estimate them with a particular focus on MI. Last, we introduce additional measures of information that have been introduced over the years and will be useful in [Chapter 8](#).

## 2.1 Shannon's Information Measures

We begin this section dedicated to information measures by introducing Shannon's information measures namely the entropy, the conditional entropy as well as the MI. For each measure, we start by the definition with discrete random variables and then extend it to the continuous case.

### 2.1.1 Entropy

In this section, we define the concept of entropy. Linear combination of entropy will be further used to define other Shannon's information measures. The entropy measures the level of information available considering a random variable's possible outcomes. If an event is unlikely to occur, the observation of such an event brings more information than observing an event that is likely to append.

**Definition 2.1.1** (Entropy (Discrete Case)). Let  $X$  be a random variable (r.v) taking values in a discrete space  $\mathcal{X}$ , the associated pdf is denoted  $p_X$  with  $p_X(x)$  which denotes the probability of  $X$  to take the value  $x \in \mathcal{X}$ . The entropy  $H(X)$  of a r.v  $X$  is defined as:

$$H(X) = - \sum_{x \in \mathcal{X}} p_X(x) \log p_X(x). \quad (2.1)$$

We suppose in [Equation 2.1](#) and all the definitions that  $\forall x \in \mathcal{X} p_X(x) > 0$ . From [Equation 2.1](#) it can be deduced that the maximum entropy distribution for the discrete case is obtained for the uniform distribution. In that case  $H(X) = \log|\mathcal{X}|$ .

We can extend the definition of entropy by defining the joint entropy between two discrete r.v. The joint entropy measures the level of information of the set of variables  $X, Y$ . It is the entropy of the couple  $(X, Y)$

**Definition 2.1.2** (Joint Entropy (discrete Case)). Let  $X$  and  $Y$  two r.v taking value in a discrete space  $\mathcal{X}$  and  $\mathcal{Y}$  respectively with joint pdf  $p_{X,Y}$ . The joint entropy  $H(X, Y)$  between r.v  $X$  and  $Y$  is defined as:

$$H(X, Y) = H(Y, X) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log p_{X,Y}(x, y). \quad (2.2)$$

For two r.v we can also define the conditional entropy which characterizes the quantity of information needed to know the behavior of  $Y$  when  $X$  is known.

**Definition 2.1.3** (Conditional Entropy (discrete Case)). The conditional entropy  $H(X|Y)$  between  $X$  and  $Y$  is defined as:

$$H(Y|X) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log p_{Y|X}(y|x). \quad (2.3)$$

In [Chapter 7](#), we will use the following equality:  $H(Y|X) = \sum_{x \in \mathcal{X}} p_X(x) H(Y|X = x)$  which is a direct consequence of [Equation 2.3](#) by denoting

$$H(Y|X = x) = - \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log p_{Y|X}(y|x)$$

Previous definitions can be extended to continuous random variables. Let  $X$  and  $Y$  be two random variables taking values in a continuous space  $\mathcal{X}$  and  $\mathcal{Y}$  respectively. As previously, pdf are denoted  $p_X$  and  $p_Y$ .

**Definition 2.1.4** (Differential Entropy). For continuous r.v the entropy becomes the differential entropy. Formally, the differential entropy  $h(X)$  of a r.v  $X$  is defined as:

$$h(X) = - \int_{x \in \mathcal{X}} p_X(x) \log p_X(x) dx. \quad (2.4)$$

In [Equation 2.6](#) and what follows, we assume that when  $h(X)$  is written it exists (*i.e* that the integral is defined and the pdf exists). We then generalize the differential entropy to a set of continuous r.v as well as the conditional case.

**Definition 2.1.5** (Differential Entropy of a set). The differential entropy of a set of continuous r.v  $X_1, \dots, X_n$  with pdf  $p_{X_1, \dots, X_n}$  is defined as:

$$h(X_1, \dots, X_n) = - \int_{(x_1, \dots, x_n) \in (\mathcal{X}^1 \times \dots \times \mathcal{X}^n)} p_{X_1, \dots, X_n}(x_1, \dots, x_n) \log p_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1, \dots, dx_n. \quad (2.5)$$

**Definition 2.1.6** (Conditional Differential Entropy). The conditional differential entropy of two of continuous r.v  $X, Y$  with joint pdf  $p_{X,Y}$  is defined as:

$$h(X|Y) = - \int_{(x,y) \in (\mathcal{X} \times \mathcal{Y})} p_{X,Y}(x, y) \log p_{X|Y}(x|y) dx dy. \quad (2.6)$$

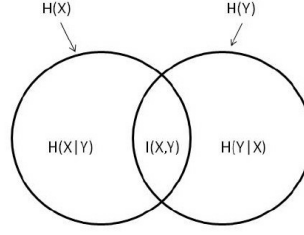


Figure 2.1 – Venn Diagrams connecting entropy, conditional entropy, joint entropy and MI for two r.v  $X$  and  $Y$ .

## 2.1.2 Mutual Information

The mutual information (MI) can be defined from the entropy. MI can be seen as a quantity that characterizes the amount of information between two r.v.. Given the observation of the first r.v, it measures how much information can be deducted on the second. Formally: how will the knowledge of  $X$  be affected if a specific event of  $Y$  happens.

**Definition 2.1.7** (MI for Discrete r.v). Given two discrete r.v  $X$  and  $Y$  the MI is defined as:

$$I(X; Y) = I(Y; X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)}. \quad (2.7)$$

Similarly to what is done with entropy the MI can be extended to continuous r.v.

**Definition 2.1.8** (MI for Continuous r.v). Given two continuous r.v  $X$  and  $Y$  the MI  $I(X; Y)$  is defined as:

$$I(X; Y) = I(Y; X) = \int_{x \in \mathcal{X}} \int_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} dx dy. \quad (2.8)$$

**Connection between MI and entropy.** We then recall the connection between MI and entropy. In what follows, we will work in the case of discrete variables but the formalism hold with a continuous random variable by changing replacing the entropy with the conditional entropy.

The MI can be link to entropy by the following formula:

$$I(X; Y) = H(X) - H(X|Y). \quad (2.9)$$

Further link can be drawn between joint entropy and MI. They can be summarized using the Venne Diagram given in [Figure 2.1](#).

**Positivity of the MI.** The MI is always non negative. It reflects the intuitive fact that knowing the outcome of the first random variable can only decrease the uncertainty on the second one. If the two variables are independent, no information is gained on the second knowing the first one.

**Data Processing Inequality.** The data processing inequality states that any transformation or a r.v can only decrease the quantity of information available in the r.v. Formally, given a function  $f$  applied to  $X$ :

$$I(X; Y) > I(f(X); Y). \quad (2.10)$$

### 2.1.3 Connections Between MI and Kullback-Leibler (KL) Divergence.

**Kullback-Leibler (KL) divergence.** The KL divergence measures the difference between a distribution  $p_X$  and a reference distribution  $q_X$  with same support  $\mathcal{X}$  (we also suppose that  $p_X$  is absolutely continuous with respect to  $q_X$ ).

**Definition 2.1.9** (KL divergence). In the discrete case, the KL divergence is defined as:

$$\text{KL}(p_X; q_X) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)}. \quad (2.11)$$

In the continuous case, the KL divergence is defined as:

$$\text{KL}(p_X || q_X) = \int_{x \in \mathcal{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)} dx. \quad (2.12)$$

The KL divergence can be linked to the MI:

$$I(X; Y) = \text{KL}[p_{Y|X}(y|x) || p_Y(y)]. \quad (2.13)$$

## 2.2 Computational Aspects of MI

Estimating MI has been a long-standing challenge as the exact computation is often intractable, in particular when dealing with high-dimensional data [PANINSKI, 2003; PICHLER and collab., 2020]. Thus in practice most of the methods rely on variational bounds (see POOLE and collab. [2019] for a comprehensive study). Although a plethora of estimators are available [AGAKOV, 2004; ALEMI and collab., 2016; BLEI and collab., 2017; MCALLESTER and STRATOS, 2020], in this section we focus on the fourth we will use in Part II and Part III namely InfoNCE [OORD and collab., 2018a], MINE [BELGHAZI and collab., 2018], NWJ and CLUB [CHENG and collab., 2020] as they are the one that are commonly used for textual data.

In practice, to compute a bound of the MI between two r.v, we have pairs  $\{(x_i, y_i)\}_{i=1}^N$  that are sampled from an unknown distribution  $p_{X,Y}$ .

In this setting, the empirical MINE estimator is given by:

$$\hat{I}_{\text{MINE}}(X; Y) = \frac{1}{N} \sum_{i=1}^N f_{\theta}(x_i, y_i) - \log \frac{1}{N} \sum_{i=1}^N \exp(f_{\theta}(x_i, y_{k_i})), \quad (2.14)$$

where  $f_{\theta}(\cdot, \cdot)$  is a neural network that approximate a score function. This estimator is a direct consequence of the Donsker-Varadhan representation of the KL divergence [DONSKEK and VARADHAN, 1985].

A similar estimator can be derived for the MI  $f$ -divergence representation and is due to Nguyen, Wainwright, and Jordan (NWJ) [NGUYEN and collab., 2017, 2010]:

$$\hat{I}_{\text{NWJ}}(X; Y) = \frac{1}{N} \sum_{i=1}^N f_{\theta}(x_i, y_i) - \frac{1}{N} \sum_{i=1}^N \exp(f_{\theta}(x_i, y_{k_i}) - 1). \quad (2.15)$$

Using the the Noise Contrastive estimation principle [GUTMANN and HYVÄRINEN, 2010], a lower bound on MI can be derived called InfoNCE:

$$\hat{I}_{\text{InfoNCE}}(X; Y) = \frac{1}{N} \sum_{i=1}^N \log \frac{\exp(f_{\theta}(x_j, y_j))}{\frac{1}{N} \sum_{j=1}^N \exp(f_{\theta}(x_j, y_j))}, \quad (2.16)$$



where  $f_\theta(.,.)$  is a function with parameter  $\theta$  (classical choice includes dot products between encoded representations).

One of the most recent MI estimator called CLUB (Log-ratio Upper Bound) which takes inspiration from [FEUTRY and collab. \[2018\]](#) is defined by:

$$\hat{I}_{\text{CLUB}}(X; Y) = \frac{1}{N} \sum_{i=1}^N \log q_\theta(y_i | x_i) - \log q_\theta(y_{k_i} | x_i), \quad (2.17)$$

where  $q_\theta$  is a variational distribution that approximate  $p_{Y|X}$  and with  $k_i$  uniformly selected from  $[1, N]$ .

InfoNCE will be mainly used in [Chapter 6](#) for learning representation of textual data, MINE and NWJ will be used for learning multimodal representation in [Chapter 7](#) and CLUB will be compared to our estimator in [Chapter 8](#).

## 2.3 Contrastive Learning and NLP

InfoNCE which can be linked to contrastive learning [CHOPRA and collab. \[2005\]](#) offer satisfactory approximation of MI with theoretical guarantees (we refer the reader to [OORD and collab. \[2018b\]](#) for further details). Contrastive learning has first been introduced in [GUNEL and collab. \[2020\]](#); [KHOSLA and collab. \[2020\]](#) and is connected to triplet loss [SCHROFF and collab. \[2015\]](#). It has since been used to tackle the different problems including self-supervised or unsupervised representation learning (e.g. audio [QIAN and collab. \[2021\]](#), image [YAMAGUCHI and collab. \[2019\]](#), text [GIORGI and collab. \[2020\]](#); [LOGESWARAN and LEE \[2018\]](#); [REIMERS and GUREVYCH \[2019\]](#)). It consists in bringing closer pairs of similar inputs, called *positive pairs* and further dissimilar ones, called *negative pairs*. The positive pairs can be obtained by data augmentation techniques [CHEN and collab. \[2020\]](#) or using various heuristic (e.g. similar sentences belong to the same document [GIORGI and collab. \[2020\]](#), backtranslation [FANG and collab. \[2020\]](#) or more complex techniques [GILLICK and collab. \[2019\]](#); [QU and collab. \[2020\]](#); [SHEN and collab. \[2020\]](#)). For a deeper dive in mining techniques used in NLP, we refer the reader to [RETHMEIER and AUGENSTEIN \[2021\]](#). In contrast, recent *supervised contrastive learning* methods take advantage of the label to create positive and negative pairs. In both cases, the sampling strategy adopted to obtain positive and negative examples is instrumental for the performance [CHEN and collab. \[2020\]](#); [ZHANG and STRATOS \[2021\]](#); [? \[2021\]](#). Additional important factors to tune to ensure good performance of contrastive learning include to choose the temperature parameter [WANG and LIU \[2021\]](#); [WANG and ISOLA \[2020\]](#) and working with large batch size [BACHMAN and collab. \[2019\]](#); [HENAFF \[2020\]](#); [MITROVIC and collab. \[2020\]](#); [OORD and collab. \[2018b\]](#). In practice, hardware limits the maximum number of sample that can be stored in memory. Although several works [GAO and collab. \[2021\]](#); [HE and collab. \[2020\]](#), have been conducted to go beyond the memory usage limitation, every experiment we conducted was performed on a single GPU.

## 2.4 Beyond KL Divergence as a Measure of Similarity

MI can be linked to the KL divergence. However, the KL divergence is not the only measure that can be adopted to measure the similarity between two distributions. In machine learning, a plethora of distances have been utilized among which we can mention the euclidian distance, f-divergences, Bregman divergences [[BREGMAN,](#)

Name	Notation	Domain	Expression
$\alpha$ -divergence [CSISZÁR, 1967]	$\mathcal{D}_\alpha$	$\alpha \notin \{0, 1\}$	$\frac{1}{\alpha(\alpha-1)}(1 - \sum q_i^{1-\alpha} p_i^\alpha)$
$\gamma$ divergence [FUJISAWA and EGUCHI, 2008]	$\mathcal{D}_\gamma^\beta$	$\beta \notin \{0, -1\} \beta \in \mathbb{R}$	$\frac{1}{\beta(\beta+1)} \log \sum p_i^{\beta+1} + \frac{1}{\beta+1} \log \sum q_i^{\beta+1} - \frac{1}{\beta} \log \sum p_i q_i^\beta$
AB Divergence [CICHOCKI and collab., 2011]	$\mathcal{D}_{sAB}^{\alpha,\beta}$	$(\alpha, \beta) \in (\mathbb{R}^*)^2$ $\beta + \alpha \neq 0$	$\frac{1}{\beta(\beta+\alpha)} \log \sum p_i^{\beta+\alpha} + \frac{1}{\beta+\alpha} \log \sum q_i^{\beta+\alpha} - \frac{1}{\beta} \log \sum p_i^\alpha q_i^\beta$
$\mathcal{L}_1$ distance	$\mathcal{L}_1$		$\sum  p_i - q_i $
$\mathcal{L}_2$ distance	$\mathcal{L}_2$		$\sqrt{\sum (p_i - q_i)^2}$
$\mathcal{L}_\infty$ distance	$\mathcal{L}_\infty$		$\max_i  p_i - q_i $
Fisher-Rao distance	R		$\frac{2}{\pi} \arccos \sum \sqrt{p_i \times q_i}$

Table 2.1 – Expression of the divergences (upper group) and distance between two positives measures  $\mathbf{p} = (p_1, \dots, p_N)$  and  $\mathbf{q} = (q_1, \dots, q_N)$  as well as the definition domain. For sake of clarity we omit the index in the summations.

1967], Rao [RAO, 1987] or Wasserstein distances [PEYRÉ and collab., 2019]. In this section, we introduce the one used in Chapter 8 and thus we focus on discrete distribution. We call information measure any function of one or more probability distributions (see [BASSEVILLE, 2013; CROOKS, 2017] for an exhaustive study). Here we focus on comparing a pair of discrete probability distributions. We call distance, a function that is symmetric, positive, respect the triangle inequality and is equal to zero if (and only if) the two considered distributions are strictly identical. The divergence is a measure of dissimilarity that is always positive or equal to zero if (and only if) the two considered distributions are strictly identical. Here, we focus, on information measures that belong to either Csiszar  $f$ -divergences [CSISZÁR, 1967] or that are distances.

## 2.4.1 Divergence Measures

Various divergence measures have been proposed for a large variety of applications [BASSEVILLE, 2013; CROOKS, 2017]. The full expression of the studied divergences can be found in Table 2.1. We focus here on three families of divergences  $\alpha$  Divergences,  $\gamma$  Divergences and AB Divergences. Note that there exist other families of divergences such as Bregman divergence [BREGMAN, 1967],  $\beta$  divergences [BASU and collab., 1998], Chernoff divergence [CHERNOFF and collab., 1952; KAKIZAWA and collab., 1998] or  $\alpha$ -Rényi Divergences [RÉNYI and collab., 1961; VAN ERVEN and HARREMOS, 2014] to cite a few of them.

### $\alpha$ -Divergences.

This divergence was introduced by RÉNYI and collab. [1961] and are a special case of the  $f$ -divergences [ALI and SILVEY, 1966; CSISZÁR, 1967]. They are widely used in variational inference [LI and TURNER, 2016] and closely related to Rényi divergences but are not a special case [PÓCZOS and SCHNEIDER, 2011]. From Table 2.1 we note special cases of  $\alpha$ -Divergences:

- Kullback-Leiber (KL) is recovered by letting  $\alpha \rightarrow 1$ ;
- Hellinger distance [HELLINGER, 1909] follows by choosing  $\alpha = 0.5$ .

For this family,  $\alpha$  can be seen as weighting the influence of  $\frac{p}{q}$ .

### $\gamma$ -Divergences.

This divergence has been introduced by [EGUCHI and KATO \[2010\]](#); [FUJISAWA and EGUCHI \[2008\]](#) as a scale-invariant modification of the robust  $\beta$ -divergences.<sup>1</sup> For the  $\gamma$  divergences the parameter  $\beta$  is used to control the importance of the element of small probabilities (e.g., outliers in some scenario, words with low probability in our case). If  $\beta > 1$ , the importance of large  $q_i$  is reduced which gives more weights to the outliers. Special cases include the  $L_2$  distance (i.e.,  $\beta = 2$ ) and KL divergence (i.e.,  $\beta \rightarrow 1$ ).

### AB-Divergences.

The family of AB-divergences is flexible and allows to respectively control the mass coverage or the robustness. [CICHOCKI and collab. \[2011\]](#); [REGLI and SILVA \[2018\]](#) propose to use AB divergences. As can be seen in [Table 2.1](#) these divergences have two parameters  $\alpha, \beta$  allow to tune the mass coverage and the robustness independently. Special cases of this divergence include:

- The KL divergence which is recovered by choosing  $\alpha = 1, \beta = 1$ ;
- The  $\beta$ -divergence is obtained by choosing  $\alpha = 1, \beta \in \mathbb{R}$ .

## 2.4.2 Distances

### $\mathcal{L}_p$ Distances

The  $\mathcal{L}_p$  distances  $p \in \mathbb{R}_{>0}$  can be used to measure the similarity between two distributions. We restrict ourselves to the special case where  $p \in \{1, 2, \dots, +\infty\}$ .

### Fisher-Rao Distance

The Fisher-Rao distance represents the Geodesic Distance [[MENÉNDEZ and collab., 1997](#); [RAO, 1987](#)] between two distributions. Interestingly, this distance remains overlooked in the ML community but has been recently used to achieve robustness against adversarial attacks [[PICOT and collab., 2021](#)].

## 2.4.3 From Information Divergences to Discrimination

For our application in [Chapter 8](#), we would like to produce a metric between two texts regardless of the source (system or human). Thus we are interested in symmetric divergence: such divergences are called discrimination. To obtain discrimination two tricks are commonly applied either the Jeffrey's symmetrization, which is obtained by averaging  $KL(p \parallel q)$  and  $KL(q \parallel p)$ , or the Jensen's symmetrization, which is obtained by averaging  $KL(p \parallel \frac{p+q}{2})$  and  $KL(q \parallel \frac{p+q}{2})$ . We choose to use Jeffreys symmetrization as it does not require computing  $\frac{p+q}{2}$ .

<sup>1</sup>In our setting we work with normalised distributions, thus scale invariance is not a mandatory property. However, it is worth mentioning as it could cause practical issues when optimising our metric.

## 2.5 Multivariate Extensions

In the previous section, we focused on computing the MI between two r.v, however, in some case it can be interesting to measure the statistical dependency of multiple r.v.. In this part, we present the extension of the MI to multiple r.v..

### 2.5.1 Extension of MI to Different Metrics

The KL divergence seems to be limited when used for estimating MI [MCALLESTER and STRATOS, 2020]. A natural step is to replace the KL divergence in Equation 6.2 with different divergences such as the f-divergences or distances such as the Wasserstein distance. Hence, we introduce new mutual dependency measures (MDM): the f-Mutual Information [BELGHAZI and collab., 2018], denoted  $I_f$  and the Wasserstein Measures [OZAIR and collab., 2019], denoted  $I_{\mathcal{W}}$ . As previously,  $p_{XY}$  denotes the joint pdf, and  $p_X$ ,  $p_Y$  denote the marginal pdfs. The new measures are defined as follows:

$$I_f \triangleq \mathcal{D}_f(p_{XY}(x, y); p_X(x) p_Y(y)), \quad (2.18)$$

where  $\mathcal{D}_f$  denotes any  $f$ -divergences and

$$I_{\mathcal{W}} \triangleq \mathcal{W}(p_{XY}(x, y); p_X(x) p_Y(y)), \quad (2.19)$$

where  $\mathcal{W}$  denotes the Wasserstein distance [PEYRÉ and collab., 2019].

### 2.5.2 From Bivariate to Multivariate

In our Chapter 7, we will maximize cross-view interactions involving three modalities (*i.e* text, audio, video), thus we need to generalize bivariate dependency measures to multivariate dependency measures.

**Definition 2.5.1** (Multivariate Dependencies Measures). *Let  $X_a, X_v, X_l$  be a set of random variables with joint pdf  $p_{X_a X_v X_l}$  and respective marginal pdf  $p_{X_j}$  with  $j \in \{a, v, l\}$ . Then we defined the multivariate MI  $I_{kl}$ , also refered as total correlation [WATANABE, 1960] or multi-information [STUDENÝ and VEJNAROVÁ, 1998]:*

$$I_{kl} \triangleq \text{KL}(p_{X_a X_v X_l}(x_a, x_v, x_l) || \prod_{j \in \{a, v, l\}} p_{X_j}(x_j)).$$

Simarly for any  $f$ -divergence we define the multivariate  $f$ -MI  $I_f$  as:

$$I_f \triangleq \mathcal{D}_f(p_{X_a X_v X_l}(x_a, x_v, x_l); \prod_{j \in \{a, v, l\}} p_{X_j}(x_j)).$$

Finally, we also extend Equation 6.3 to obtain the multivariate Wasserstein dependency measure  $I_{\mathcal{W}}$ :

$$I_{\mathcal{W}} \triangleq \mathcal{W}(p_{X_a X_v X_l}(x_a, x_v, x_l); \prod_{j \in \{a, v, l\}} p_{X_j}(x_j)).$$

where  $\mathcal{W}$  denotes the Wasserstein distance.

**Remark.** Aforementioned multivariate measures suffer from the same problem as MI: the exact value is often intractable, thus in Chapter 6 we will work with variational bounds.

**Chapter 2 conclusion**

In this chapter, we provided an informal introduction of the main mathematical tools belonging to the field of information theory required to understand the contributions of the thesis. We introduced the main measures of information we will use in this thesis. In the next chapter, we will see the connection between learning textual representations and the mutual information introduced at the beginning of this chapter.

**2.6 References**

- AGAKOV, D. B. F. 2004, «The im algorithm: a variational approach to information maximization», *Advances in neural information processing systems*, vol. 16, p. 201. [34](#)
- ALEMI, A. A., I. FISCHER, J. V. DILLON and K. MURPHY. 2016, «Deep variational information bottleneck», *arXiv preprint arXiv:1612.00410*. [34](#)
- ALI, S. M. and S. D. SILVEY. 1966, «A general class of coefficients of divergence of one distribution from another», *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 28, n° 1, p. 131–142. [36](#)
- BACHMAN, P., R. D. HJELM and W. BUCHWALTER. 2019, «Learning representations by maximizing mutual information across views», *arXiv preprint arXiv:1906.00910*. [35](#)
- BASSEVILLE, M. 2013, «Divergence measures for statistical data processing—an annotated bibliography», *Signal Processing*, vol. 93, n° 4, p. 621–633. [36](#)
- BASU, A., I. R. HARRIS, N. L. HJORT and M. JONES. 1998, «Robust and efficient estimation by minimising a density power divergence», *Biometrika*, vol. 85, n° 3, p. 549–559. [36](#)
- BELGHAZI, M. I., A. BARATIN, S. RAJESWAR, S. OZAI, Y. BENGIO, A. COURVILLE and R. D. HJELM. 2018, «Mine: mutual information neural estimation», *arXiv preprint arXiv:1801.04062*. [34](#), [38](#)
- BLEI, D. M., A. KUCUKELBIR and J. D. MCAULIFFE. 2017, «Variational inference: A review for statisticians», *Journal of the American statistical Association*, vol. 112, n° 518, p. 859–877. [34](#)
- BREGMAN, L. M. 1967, «The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming», *USSR computational mathematics and mathematical physics*, vol. 7, n° 3, p. 200–217. [35](#), [36](#)
- CHEN, T., S. KORNBLITH, M. NOROUZI and G. HINTON. 2020, «A simple framework for contrastive learning of visual representations», in *International conference on machine learning*, p. 1597–1607. [35](#)

- CHENG, P., W. HAO, S. DAI, J. LIU, Z. GAN and L. CARIN. 2020, «Club: A contrastive log-ratio upper bound of mutual information», in *International Conference on Machine Learning*, p. 1779–1788. [34](#)
- CHERNOFF, H. and collab.. 1952, «A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations», *The Annals of Mathematical Statistics*, vol. 23, n° 4, p. 493–507. [36](#)
- CHOPRA, S., R. HADSELL and Y. LECUN. 2005, «Learning a similarity metric discriminatively, with application to face verification», in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, p. 539–546. [35](#)
- CICHOCKI, A., S. CRUCES and S.-I. AMARI. 2011, «Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization», *Entropy*, vol. 13, n° 1, p. 134–170. [36](#), [37](#)
- CROOKS, G. E. 2017, «On measures of entropy and information», *Tech. Note*, vol. 9, p. v4. [36](#)
- CSISZÁR, I. 1967, «Information-type measures of difference of probability distributions and indirect observation», *studia scientiarum Mathematicarum Hungarica*, vol. 2, p. 229–318. [36](#)
- DONSKER, M. and S. VARADHAN. 1985, «Large deviations for stationary gaussian processes», *Communications in Mathematical Physics*, vol. 97, n° 1-2, p. 187–210. [34](#)
- EGUCHI, S. and S. KATO. 2010, «Entropy and divergence associated with power function and the statistical application», *Entropy*, vol. 12, n° 2, p. 262–274. [37](#)
- FANG, H., S. WANG, M. ZHOU, J. DING and P. XIE. 2020, «Cert: Contrastive self-supervised learning for language understanding», *arXiv preprint arXiv:2005.12766*. [35](#)
- FEUTRY, C., P. PIANTANIDA, Y. BENGIO and P. DUHAMEL. 2018, «Learning anonymized representations with adversarial neural networks», . [35](#)
- FUJISAWA, H. and S. EGUCHI. 2008, «Robust parameter estimation with a small bias against heavy contamination», *Journal of Multivariate Analysis*, vol. 99, n° 9, p. 2053–2081. [36](#), [37](#)
- GAO, L., Y. ZHANG, J. HAN and J. CALLAN. 2021, «Scaling deep contrastive learning batch size under memory limited setup», in *Proceedings of the 6th Workshop on Representation Learning for NLP (ReL4NLP-2021)*, p. 316–321. [35](#)
- GILLICK, D., S. KULKARNI, L. LANSING, A. PRESTA, J. BALDRIDGE, E. IE and D. GARCIA-OLANO. 2019, «Learning dense representations for entity retrieval», *arXiv preprint arXiv:1909.10506*. [35](#)
- GIORGI, J. M., O. NITSKI, G. D. BADER and B. WANG. 2020, «Declutr: Deep contrastive learning for unsupervised textual representations», *arXiv preprint arXiv:2006.03659*. [35](#)



- GUNEL, B., J. DU, A. CONNEAU and V. STOYANOV. 2020, «Supervised contrastive learning for pre-trained language model fine-tuning», *arXiv preprint arXiv:2011.01403*. 35
- GUTMANN, M. and A. HYVÄRINEN. 2010, «Noise-contrastive estimation: A new estimation principle for unnormalized statistical models», in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, p. 297–304. 34
- HE, K., H. FAN, Y. WU, S. XIE and R. GIRSHICK. 2020, «Momentum contrast for unsupervised visual representation learning», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 9729–9738. 35
- HELLINGER, E. 1909, «Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.», *Journal für die reine und angewandte Mathematik*, vol. 1909, n° 136, p. 210–271. 36
- HENAFF, O. 2020, «Data-efficient image recognition with contrastive predictive coding», in *International Conference on Machine Learning*, PMLR, p. 4182–4192. 35
- KAKIZAWA, Y., R. H. SHUMWAY and M. TANIGUCHI. 1998, «Discrimination and clustering for multivariate time series», *Journal of the American Statistical Association*, vol. 93, n° 441, p. 328–340. 36
- KHOSLA, P., P. TETERWAK, C. WANG, A. SARNA, Y. TIAN, P. ISOLA, A. MASCHINOT, C. LIU and D. KRISHNAN. 2020, «Supervised contrastive learning», *arXiv preprint arXiv:2004.11362*. 35
- LI, Y. and R. E. TURNER. 2016, «R\'enyi divergence variational inference», *arXiv preprint arXiv:1602.02311*. 36
- LOGESWARAN, L. and H. LEE. 2018, «An efficient framework for learning sentence representations», *arXiv preprint arXiv:1803.02893*. 35
- MCALLESTER, D. and K. STRATOS. 2020, «Formal limitations on the measurement of mutual information», in *International Conference on Artificial Intelligence and Statistics*, p. 875–884. 34, 38
- MENÉNDEZ, M. L., D. MORALES, L. PARDO and M. SALICRÚ. 1997, « $(h, \phi)$ -entropy differential metric», *Applications of Mathematics*, vol. 42, n° 2, p. 81–98. 37
- MITROVIC, J., B. MCWILLIAMS and M. REY. 2020, «Less can be more in contrastive learning», . 35
- NGUYEN, T., T. LE, H. VU and D. PHUNG. 2017, «Dual discriminator generative adversarial nets», in *Advances in Neural Information Processing Systems*, p. 2670–2680. 34
- NGUYEN, X., M. J. WAINWRIGHT and M. I. JORDAN. 2010, «Estimating divergence functionals and the likelihood ratio by convex risk minimization», *IEEE Transactions on Information Theory*, vol. 56, n° 11, p. 5847–5861. 34
- OORD, A. V. D., Y. LI and O. VINYALS. 2018a, «Representation learning with contrastive predictive coding», *arXiv preprint arXiv:1807.03748*. 34

- OORD, A. V. D., Y. LI and O. VINYALS. 2018b, «Representation learning with contrastive predictive coding», *arXiv preprint arXiv:1807.03748*. 35
- OZAI, S., C. LYNCH, Y. BENGIO, A. VAN DEN OORD, S. LEVINE and P. SERMANET. 2019, «Wasserstein dependency measure for representation learning», in *Advances in Neural Information Processing Systems*, p. 15 604–15 614. 38
- PANINSKI, L. 2003, «Estimation of entropy and mutual information», *Neural computation*, vol. 15, n° 6, p. 1191–1253. 34
- PEYRÉ, G., M. CUTURI and collab.. 2019, «Computational optimal transport: With applications to data science», *Foundations and Trends® in Machine Learning*, vol. 11, n° 5-6, p. 355–607. 36, 38
- PICHLER, G., P. PIANTANIDA and G. KOLIANDER. 2020, «On the estimation of information measures of continuous distributions», . 34
- PICOT, M., F. MESSINA, M. BOUDIAF, F. LABEAU, I. BEN AYED and P. PIANTANIDA. 2021, «Adversarial robustness via fisher-rao regularization», *arXiv preprint arXiv:*. 37
- PÓCZOS, B. and J. SCHNEIDER. 2011, «On the estimation of alpha-divergences», in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, p. 609–617. 36
- POOLE, B., S. OZAI, A. VAN DEN OORD, A. ALEMI and G. TUCKER. 2019, «On variational bounds of mutual information», in *Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol. 97, édité par K. Chaudhuri and R. Salakhutdinov, PMLR, p. 5171–5180. URL <http://proceedings.mlr.press/v97/poole19a.html>. 34
- QIAN, R., T. MENG, B. GONG, M.-H. YANG, H. WANG, S. BELONGIE and Y. CUI. 2021, «Spatiotemporal contrastive video representation learning», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 6964–6974. 35
- QU, Y., D. SHEN, Y. SHEN, S. SAJEEV, J. HAN and W. CHEN. 2020, «Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding», *arXiv preprint arXiv:2010.08670*. 35
- RAO, C. R. 1987, «Differential metrics in probability spaces», *Differential geometry in statistical inference*, vol. 10, p. 217–240. 36, 37
- REGLI, J.-B. and R. SILVA. 2018, «Alpha-beta divergence for variational inference», *arXiv preprint arXiv:1805.01045*. 37
- REIMERS, N. and I. GUREVYCH. 2019, «Sentence-bert: Sentence embeddings using siamese bert-networks», *arXiv preprint arXiv:1908.10084*. 35
- RÉNYI, A. and collab.. 1961, «On measures of entropy and information», in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California. 36



- RETHMEIER, N. and I. AUGENSTEIN. 2021, «A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives», *arXiv preprint arXiv:2102.12982*. [35](#)
- SCHROFF, F., D. KALENICHENKO and J. PHILBIN. 2015, «Facenet: A unified embedding for face recognition and clustering», in *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 815–823. [35](#)
- SHEN, D., M. ZHENG, Y. SHEN, Y. QU and W. CHEN. 2020, «A simple but tough-to-beat data augmentation approach for natural language understanding and generation», *arXiv preprint arXiv:2009.13818*. [35](#)
- STUDENÝ, M. and J. VEJNAROVÁ. 1998, «The multiinformation function as a tool for measuring stochastic dependence», in *Learning in graphical models*, Springer, p. 261–297. [38](#)
- VAN ERVEN, T. and P. HARREMOS. 2014, «Rényi divergence and kullback-leibler divergence», *IEEE Transactions on Information Theory*, vol. 60, n° 7, p. 3797–3820. [36](#)
- WANG, F. and H. LIU. 2021, «Understanding the behaviour of contrastive loss», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 2495–2504. [35](#)
- WANG, T. and P. ISOLA. 2020, «Understanding contrastive representation learning through alignment and uniformity on the hypersphere», in *International Conference on Machine Learning*, p. 9929–9939. [35](#)
- WATANABE, S. 1960, «Information theoretical analysis of multivariate correlation», *IBM Journal of research and development*, vol. 4, n° 1, p. 66–82. [38](#)
- YAMAGUCHI, S., S. KANAI, T. SHIODA and S. TAKEDA. 2019, «Multiple pretext-task for self-supervised learning via mixing multiple image transformations», *arXiv preprint arXiv:1912.11603*. [35](#)
- ZHANG, W. and K. STRATOS. 2021, «Understanding hard negatives in noise contrastive estimation», *arXiv preprint arXiv:2104.06245*. [35](#)



# Chapter 3

## Representing Textual Transcripts

### Chapter 3 abstract

Representing the meaning of natural language in a mathematically grounded way is a scientific challenge that has received increasing attention with the explosion of digital content and text data in the last decade. Relying on the richness of contents, several embeddings have been proposed [DEVLIN and collab., 2019; PETERS and collab., 2018a; RADFORD and collab., 2018] with demonstrated efficiency for the considered tasks when learned on massive written datasets. This chapter is dedicated to the related work specific to the problem of representation learning for transcript data, more precisely we focus on integrating the conversational and multimodal dimensions. First, we review the specifics of written transcript, then we review the connection between the InfoNCE and classical word embeddings (*e.g* Skipgram objective [MIKOLOV and collab., 2013b]) as well as contextualized embeddings based on the denoising autoencoder framework (*e.g* Masked Language Model (MLM) [DEVLIN and collab., 2018] or the Generalized Autoregressive Pretraining objective (GAP) [YANG and collab., 2019]). Second, different features not present in the text modality such as prosodic features (*e.g* pitch, word duration and intensity) and corporal expressions (*e.g* gaze, gestures) can be of interest when learning representations of transcripts. Thus, we will enrich our representation with different multi-modal signals of Chapter 6 and we finish this chapter by recalling the principal challenges of multimodal learning that aims at providing better textual embedding when additional modalities (*e.g* audio, video) are available. We review existing architectures with a particular emphasis on the most recent state-of-the-art models that rely on pretrained textual representations.

### 3.1 Importance of Conversational and Multimodal Dimensions to Learn Transcripts Representations

Conversational AI or dialogue systems is a booming field that attracts researchers from various communities [CHEN and collab., 2017] (*e.g* Natural Language Processing (NLP), Linguistics, Psychology, Information Retrieval (IR), Machine Learning (ML)). There is a pressing need for conversational agents as the field of business have shown a growing interest in using conversational agents to improve both service quality

and market competitiveness. Thus learning representations of transcripts of spoken conversations is a challenging research topic. A conversation is a well-structured interaction [ARORA and collab., 2013]. As can be seen in Table 3.1, a dialogue is a sequence of turns (or utterances) which contains a variable number of words. Each utterance can be classified by a kind of “action” performed by the speaker named *dialog act or speech act*. Along with speech acts each speaker ground each other’s utterances; meaning that each listener implicitly (or explicitly) acknowledges that he has understood the speaker. These characteristics of human conversations are among the challenges that researchers need to address while building conversational agent [GAO and collab., 2018].

**Specifics of Spoken Language.** The example in Table 3.1 also illustrates two specific phenomena that appear when working with spoken language (as opposed to written text):

- *Disfluencies.* In many applications, the input of the conversational agent comes from spoken language. Spoken language is rarely fluent. Disfluencies that are interruptions in the regular flow of speech, such as pausing silently, repeating words, or interrupting oneself to correct something previously said. They commonly occur in spoken language, as spoken language is rarely fluent. Fillers are a type of disfluency that can be a sound (“um” or “uh”) filling a pause in an utterance or conversation [STOLCKE and SHRIBERG, 1996].
- *Segmentation Issues.* In spoken dialog segmentation is not given as the text usually comes from an automatic speech recognition system (ASR), thus utterances are not given. Finding the beginning and the end of an utterance is a tough problem [ANG and collab., 2005; ZIMMERMANN and collab., 2005]. For instance, one can assume that an utterance ends when the user ceases to speak (by detecting a certain amount of silence, or non-speech). As previously mentioned spontaneous speech usually contains silent pauses inside utterances, for instance when a hesitation occurs. Efficient and effective segmentation of spoken language remains an open problem in spoken dialog.

Further specifics of spoken text could be listed (*e.g* surface formality change, the role of prosody, lexical diversity, and grammatical complexity and accuracy), for an exhaustive comparison one could refer to [CHAFE and TANNEN, 1987; REDEKER, 1984]. Different features not present in the text only modality such as prosodic (*i.e* pitch, sound lengths) or visual (*e.g* gaze, glance, face expressions) features can be of interest when learning transcript representations which motivate the multimodal setting of Chapter 6.

Caller	Utterance
A	um, did you do through a public school system or private?
B	Yeah,
B	well, I went through private an until ninth grade.
A	Uh-huh,
A	did you notice a big difference?
B	Oh, yeah,
B	a big difference.
A	Like in what sense?
B	Well, um, in, uh, public schools I guess there wer-, there are a lot of, of, you know,
B	people can take lower level courses an get away with learning nothing.
A	Uh-huh.
B	But, um, private school you couldn't do that,
A	Uh-huh.
B	you had to learn.
A	Yeah,
A	I work in a temporary agency

Table 3.1 – Example of dialog taken from the Switchboard Dialog Act Corpus.

## 3.2 Pretrained Representations and MI Maximization

In this section, we gather the related work useful to understand our contributions described in [Chapter 5](#). Generic representations are an effective way to adapt models across different sets of labels [[DEVLIN and collab., 2018](#); [LIU and collab., 2019](#); [MIKOLOV and collab., 2013b](#); [PENNINGTON and collab., 2014](#); [PETERS and collab., 2018b](#); [YANG and collab., 2019](#)]. Those representations are usually trained on large written corpora such as OSCAR [[SUÁREZ and collab., 2019](#)], Book Corpus [[ZHU and collab., 2015](#)] or Wikipedia [[DENOYER and GALLINARI, 2006](#)]. In this section, we review the most famous existing pretraining objectives for learning generic representations (*e.g* Skipgram [[MIKOLOV and collab., 2013b](#)], MLM [[DEVLIN and collab., 2018](#)] and GAP [[YANG and collab., 2019](#)]) and show how they relate to MI. This section is heavily borrowed from the work of [KONG and collab. \[2019\]](#).

### 3.2.1 Relationship between InfoNCE, MI and Cross Entropy

In the MI Maximization framework, which is inspired from the Infomax principle [[LINSKER, 1988](#)], the training of the encoder is done by maximizing the MI. However, the direct maximization of the MI is often intractable when the encoder is a deep neural network. Thus, the maximization often involves a variational lower bound. In this section, we particularly focus on InfoNCE which was previously introduced (see [Chapter 2](#)). As a recall, InfoNCE is defined as:

$$\hat{I}_{\text{InfoNCE}}(X; Y) = \frac{1}{N} \sum_{i=1}^N \log \frac{\exp(f_{\theta}(x_i, y_i))}{\frac{1}{N} \sum_{j=1}^N \exp(f_{\theta}(x_i, y_j))}, \quad (3.1)$$

where the pairs  $\{(x_i, y_i)\}_{i=1}^N$  are sampled from an unknown distribution  $p_{X,Y}$  and  $f_{\theta}(\cdot, \cdot)$  is a function with parameter  $\theta$  (classical choice includes dot products between encoded representations).

Following the steps described in [KONG and collab. \[2019\]](#); [OORD and collab. \[2018a\]](#), InfoNCE can be rewritten and further connected to MI.

$$I_{\text{InfoNCE}}(X; Y) = f_{\theta}(x_i, y_i) - \frac{1}{N} \sum_{i=1}^N \log \frac{1}{N} - \log \sum_{j=1}^N \exp(f_{\theta}(x_i, y_j)) \quad (3.2)$$

$$= f_{\theta}(x_i, y_i) - \mathbb{E}_{q(\tilde{\mathcal{B}})} \log \sum_{y_j \in \tilde{\mathcal{B}}} \exp(f_{\theta}(x_i, y_j) + \log |\tilde{\mathcal{B}}|) \quad (3.3)$$

$$\leq I(X; Y). \quad (3.4)$$

In practice, in order to use Equation 3.2, we have to:

- select  $X, Y$ ,
- choose the form of the function  $f_{\theta}$ ,
- fix the set  $\tilde{\mathcal{B}}$  of negative samples,
- choose the proposal distribution  $q$  from which we sample the negative examples.

Last, the sampling strategy from the first expectation can be an additional parameter.

**InfoNCE and Cross Entropy.** Following the work from KONG and collab. [2019], we can see that InfoNCE is related to cross entropy when  $\tilde{\mathcal{B}}$  includes all possible values of  $Y$  and  $q$  is the uniform distribution. Thus, maximizing the InfoNCE is similar to maximizing the cross-entropy defined by:

$$f_{\theta}(x_j, y_j) - \log \sum_{y_j \in \tilde{\mathcal{B}}} \exp(f_{\theta}(x_j, y_j)). \quad (3.5)$$

**On the relationship of InfoNCE and contrastive learning.** InfoNCE can be linked to contrastive learning surrogates CHOPRA and collab. [2005] which offer satisfactory approximations of MI with theoretical guarantees (we refer the reader to OORD and collab. [2018b] for further details). Contrastive learning has first been introduced in GUNEL and collab. [2020]; KHOSLA and collab. [2020] and is connected to triplet loss SCHROFF and collab. [2015]. It has since been used to tackle the different problems including self-supervised or unsupervised representation learning (e.g. audio QIAN and collab. [2021], image YAMAGUCHI and collab. [2019], text GIORGI and collab. [2020]; LOGESWARAN and LEE [2018]; REIMERS and GUREVYCH [2019]). It consists in bringing closer pairs of similar inputs, called *positive pairs* and further dissimilar ones, called *negative pairs*. The positive pairs can be obtained by data augmentation techniques CHEN and collab. [2020] or using various heuristic (e.g. similar sentences belong to the same document GIORGI and collab. [2020], back-translation FANG and collab. [2020] or more complex techniques GILLICK and collab. [2019]; QU and collab. [2020]; SHEN and collab. [2020]). For a deeper dive in mining techniques used in NLP, we refer the reader to RETHMEIER and AUGENSTEIN [2021]. In contrast, recent *supervised contrastive learning* methods take advantage of the label to create positive and negative pairs. We then discuss the important parameters when working with contrastive losses. First, the sampling strategy adopted to obtain positive and negative examples is instrumental for the performance CHEN and collab. [2020]; KARPUKHIN and collab. [2020]; WU and collab. [2021]; ZHANG and STRATOS [2021]; ?. Second, as discussed in WANG and LIU [2021]; WANG and ISOLA [2020] a

	X	Y	$p(x, y)$	$g_v$	$g_\psi$
Skip-gram	word	word.	words + context	lookup	lookup
MLM	context	masked word	masked token probability	transformer	lookup
XLNet	context	masked word	factorization permutation	TXL	lookup

Table 3.2 – Choices of the different parameters used in Equation 3.2 for the different pretraining objectives (borrowed from KONG and collab. [2019])

good choice of temperature parameter is also crucial for contrastive learning. Lastly, works on contrastive losses BACHMAN and collab. [2019]; HENAFF [2020]; MITROVIC and collab. [2020]; OORD and collab. [2018b] also argue for using large batch sizes to achieve good performances. In practice, hardware limits the maximum number of sample that can be stored in memory. Although several works GAO and collab. [2021]; HE and collab. [2020], have been conducted to go beyond the memory usage limitation.

### 3.2.2 MI and Pretraining Objectives

In their work, KONG and collab. [2019] show the connection between Equation 3.2 and different pretraining objectives. They focus on a particular form of  $f_\theta$  by defining  $f_\theta(x, y) = g_\psi(x)g_v(y)$  where  $\{\psi, v\} = \theta$ . Popular choices for  $g$  include:

- $g_\psi : \mathcal{V} \rightarrow \mathbb{R}^d$ , where  $\mathcal{V}$  stands for the vocabulary, a lookup function that maps a token index to a vector.
- a transformer encoder that processes a sentence and returns the final hidden state.
- a combination of Transformer XL [DAI and collab., 2019] with a two-stream of attention [YANG and collab., 2019].

Table 3.2 relates classical pretraining objectives to Equation 3.2. We refer the curious reader to KONG and collab. [2019] for exhaustive details.

### 3.2.3 Limitations of Current Pretrained Representations

For our application we will focus on learning representation for spoken conversation where hierarchy plays an import role (see Chapter 1). However, research on learning generic representation focuses either on word level [MIKOLOV and collab., 2013a] or sentence level objectives [LAMPLE and CONNEAU, 2019]; while modeling the conversational aspect of data requires to capture discourse-level features [THORNBURY and SLADE, 2006] (*i.e* information presents at different levels of the dialogue hierarchy: discourse-level information within the dialogue context and utterance-level information). Moreover, the aforementioned research focuses on written data which makes them not suited for spoken dialogue as there is a discrepancy between spoken and written language (*e.g* disfluencies [DINKAR and collab., 2020; STOLCKE and SHRIBERG, 1996], a change in grammar and lexical accuracy [CHAFE and TANNEN, 1987; REDEKER, 1984] and modifications of surface formality form [HEYLIGHEN and DEWAELE, 1999]).

In Chapter 5, we will propose new pretraining objectives tailored for conversational data and we will show how they relate to the previously introduced framework of Mutual Information Maximization.



### 3.3 Supervised Fine-tuning For Multimodal Data

This section describes related work useful to understand our contributions in [Chapter 4](#) on integrating the multimodal dimension on transcript representations.

#### 3.3.1 Introduction

##### Importance of Multimodality

Humans employ several different modalities to communicate in a coordinated manner for example, the language modality with the use of words and sentences, the vision modality with gestures, poses and facial expressions and, the acoustic modality through change in vocal tones. Multimodal representation learning has shown great progress in a large variety of tasks including emotion recognition, sentiment analysis [[SOLEYMANI and collab., 2017](#)], speaker trait analysis [[PARK and collab., 2014](#)] and fine-grained opinion mining [[GARCIA and collab., 2019](#)]. Learning from different modalities is an efficient way to improve performance on the target tasks [[XU and collab., 2013](#)]. For example, multimodal data has been shown to provide a mean to disambiguate some hard to understand opinion expressions such as irony and sarcasm [[ATTARDO and collab., 2003](#)] and contain crucial information indicating the level of engagement and the persuasiveness of the speaker [[BEN YOUSSEF and collab., 2019](#); [CLAVEL and CALLEJAS, 2016](#); [NOJAVANASGHARI and collab., 2016](#)]. Nevertheless, heterogeneities across modalities increase the difficulty of learning multimodal representations and raise specific challenges.

##### Challenges for Multimodal Learning

[BALTRUŠAITIS and collab. \[2018\]](#) identifies fusion as one of the five core challenges in multimodal representation learning, the four other being: representation, modality alignment, translation and co-learning. Fusion aims at integrating the different unimodal representations into one common synthetic representation. Effective fusion is still an open problem: the best multimodal models in sentiment analysis [[RAHMAN and collab., 2020](#)] improve over their unimodal counterparts, relying on text modality only (through BERT), by less than 1.5% on accuracy. Additionally, the fusion should not only improve accuracy but also make representations more robust to missing modalities.

#### 3.3.2 A Formalization of Learning Multimodal Representations

Plethora of neural architectures have been proposed to learn multimodal representations for sentiment classification. Models often rely on a fusion mechanism (*e.g* multi-layer perceptron [[KHAN and collab., 2012](#)], tensor factorisation [[LIU and collab., 2018](#); [ZADEH and collab., 2019](#)] or complex attention mechanisms [[ZADEH and collab., 2018a](#)]) that is fed with modality-specific representations. These complex fusion mechanism involves new trainable parameters and lack of effectiveness. Formally, the fusion problem boils down to learning a model  $\mathcal{M}_f : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathcal{R}^d$ .  $\mathcal{M}_f$  is fed with uni-modal representations of the inputs  $X_{a,v,l} = (X_a, X_v, X_l)$  obtained through three embedding networks  $f_a$ ,  $f_v$  and  $f_l$ .  $\mathcal{M}_f$  has to retain both modality-specific interactions (*i.e* interactions that involve only one modality) and cross-view interactions (*i.e* more complex, they span across both views).



### 3.3.3 Existing Deep Encoders

A large body of work has focused on the design of the encoding functions that will solve the aforementioned challenges. Multimodal fusion can be divided into early and late fusion techniques: early fusion takes place at the feature level [YE and collab., 2017], while late fusion takes place at the decision or scoring level [KHAN and collab., 2012]. Current research in multimodal sentiment analysis mainly focuses on developing new fusion mechanisms relying on deep architectures (e.g TFN [ZADEH and collab., 2017], LFN [LIU and collab., 2018], MARN [ZADEH and collab., 2018c], MISA [HAZARIKA and collab., 2020], MCTN [PHAM and collab., 2019], HFNN [MAI and collab., 2019], ICCN [SUN and collab., 2020]). In this section, we quickly review the architectures we will use in Chapter 6. These architectures can be broadly divided in two categories: randomly initialized and pretrained models.

#### Randomly Initialized Multimodal Encoders

**Early Fusion LSTM (EF-LSTM)** EF-LSTM is the most basic architecture used in the current multimodal analysis where each sequence view is encoded separately with Long Short Term Memory Units (LSTM) channels. LSTM cells [HUANG and collab., 2015] have been shown to provide good results on tasks implying the encoding or decoding of a sentence in or from a fixed size representation. Such a problem is encountered in automatic machine translation [LUONG and collab., 2015], automatic summarization [NALLAPATI and collab., 2017] or image captioning and visual question answering [ANDERSON and collab., 2018]. Sequential models build their inner state based on observations from the past. One can thus naturally use the hidden state computed at the last observation of a sequence to represent the entire sequence. In the case of the EF-LSTM, the input objects are the concatenation of the 3 feature representations coming from text, audio and video.

**Memory Fusion Networks (MFN):** MFN enriches the previous EF-LSTM architecture with an attention module that computes a cross-view representation at each time step. It belongs to the family of multi-view sequential models built upon a set of LSTM per modality feeding a joint delta memory. This architecture has been designed to carry some information in the memory even with very long sequences due to the choice of a complex retain/forget mechanism. The Memory Fusion Network [ZADEH and collab., 2018b] is made of 3 blocks:

- Each modality based sequence of feature is represented by the hidden state of a LSTM model. These hidden state are fed in the next part of the model:
- A delta attention memory takes the concatenation of two consecutive input vectors (taken from the sequence of hidden representations of the LSTM) which are fed to a feedforward model to compute an attention score for each component of these inputs. The name delta memory is only indicating the fact that the inputs are taken by pairs of inputs.
- The output of the attention layer is then sent to a Multi-view Gated Memory generalizing the GRU layer to multiview data by taking into account a modality specific and a cross modality hidden representations.

**Low Rank Fusion Network (TFN).** TFN [ZADEH and collab., 2017] computes a representation of each view, and then applies a fusion operator. Acoustic and visual views

are first mean-pooled then encoded through a 2-layers perceptron. Linguistic features are computed with a LSTM channel. Here, the fusion function is a cross-modal product capturing unimodal, bimodal and trimodal interactions across modalities.

### Pretrained Multimodal Encoders

MAG-BERT and MAG-XLNET Transformers have been recently introduced to learn multi-modal representations [TSAI and collab., 2019]. The current state of the art involves pretrained transformers and more precisely by MAG-BERT and MAG-XLNET [RAHMAN and collab., 2020]. They are based on pre-trained transformer architectures (such as BERT [DEVLIN and collab., 2018] or XLNET [YANG and collab., 2019]) allowing inputs on each of the transformer units to be multimodal, thanks to a special gate inspired by WANG and collab. [2018]. The representations are pulled thanks to the representation of the [CLS] token provided by the last transformer head.

### 3.3.4 Limitation of Current Architectures

The aforementioned architectures are usually trained by minimising either a  $L_1$  loss or a Cross-Entropy loss between the predictions and the ground-truth labels. The learning of the fusion involves both the minimisation of the downstream task loss (to retain task specific information) and the maximization of the information between the different modalities. To the best of our knowledge, few efforts have been dedicated to deriving new losses to take into account both phenomenon at once. Additionally, the use of measure of information remains overlooked to address the fusion problem. That is our contribution of Chapter 6.

#### Chapter 3 conclusion

In this chapter, we presented the related work needed to understand the contribution of Part II. We recalled the connection between MI and the pretraining objectives and showed how the problem of multi-modal representation learning can be linked to the measures of information. In both case we presented the limitation of existing approaches. In the next chapter, we will present the related work useful for Part III.

## 3.4 References

- ANDERSON, P., X. HE, C. BUEHLER, D. TENEY, M. JOHNSON, S. GOULD and L. ZHANG. 2018, «Bottom-up and top-down attention for image captioning and visual question answering», in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 6077–6086. 51
- ANG, J., Y. LIU and E. SHRIBERG. 2005, «Automatic dialog act segmentation and classification in multiparty meetings», in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, IEEE, p. 1–1061. 46
- ARORA, S., K. BATRA and S. SINGH. 2013, «Dialogue system: A brief review», *arXiv preprint arXiv:1306.4134*. 46

- ATTARDO, S., J. EISTERHOLD, J. HAY and I. POGGI. 2003, «Multimodal markers of irony and sarcasm», *Humor*, vol. 16, n° 2, p. 243–260. 50
- BACHMAN, P., R. D. HJELM and W. BUCHWALTER. 2019, «Learning representations by maximizing mutual information across views», *arXiv preprint arXiv:1906.00910*. 49
- BALTRUŠAITIS, T., C. AHUJA and L.-P. MORENCY. 2018, «Multimodal machine learning: A survey and taxonomy», *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, n° 2, p. 423–443. 50
- BEN YOUSSEF, A., C. CLAVEL and S. ESSID. 2019, «Early detection of user engagement breakdown in spontaneous human-humanoid interaction», *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2019.2898399, p. 1–1, ISSN 1949-3045. 50
- CHAFE, W. and D. TANNEN. 1987, «The relation between written and spoken language», *Annual Review of Anthropology*, vol. 16, n° 1, p. 383–407. 46, 49
- CHEN, H., X. LIU, D. YIN and J. TANG. 2017, «A survey on dialogue systems: Recent advances and new frontiers», *Acm Sigkdd Explorations Newsletter*, vol. 19, n° 2, p. 25–35. 45
- CHEN, T., S. KORNBLITH, M. NOROUZI and G. HINTON. 2020, «A simple framework for contrastive learning of visual representations», in *International conference on machine learning*, p. 1597–1607. 48
- CHOPRA, S., R. HADSELL and Y. LECUN. 2005, «Learning a similarity metric discriminatively, with application to face verification», in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, p. 539–546. 48
- CLAVEL, C. and Z. CALLEJAS. 2016, «Sentiment analysis: from opinion mining to human-agent interaction», *IEEE Transactions on affective computing*, vol. 7, n° 1, p. 74–93. 50
- DAI, Z., Z. YANG, Y. YANG, J. CARBONELL, Q. V. LE and R. SALAKHUTDINOV. 2019, «Transformer-xl: Attentive language models beyond a fixed-length context», *arXiv preprint arXiv:1901.02860*. 49
- DENOYER, L. and P. GALLINARI. 2006, «The wikipedia xml corpus», in *International Workshop of the Initiative for the Evaluation of XML Retrieval*, Springer, p. 12–19. 47
- DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805*. 45, 47, 52
- DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2019, «BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding», in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171–4186, doi: 10.18653/v1/N19-1423. URL <https://www.aclweb.org/anthology/N19-1423>. 45

- DINKAR, T., P. COLOMBO, M. LABEAU and C. CLAVEL. 2020, «The importance of fillers for text representations of speech transcripts», *arXiv preprint arXiv:2009.11340*. 49
- FANG, H., S. WANG, M. ZHOU, J. DING and P. XIE. 2020, «Cert: Contrastive self-supervised learning for language understanding», *arXiv preprint arXiv:2005.12766*. 48
- GAO, J., M. GALLEY and L. LI. 2018, «Neural approaches to conversational ai», in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 1371–1374. 46
- GAO, L., Y. ZHANG, J. HAN and J. CALLAN. 2021, «Scaling deep contrastive learning batch size under memory limited setup», in *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, p. 316–321. 49
- GARCIA, A., P. COLOMBO, S. ESSID, F. D’ALCHÉ BUC and C. CLAVEL. 2019, «From the token to the review: A hierarchical multimodal approach to opinion mining», *arXiv preprint arXiv:1908.11216*. 50
- GILLICK, D., S. KULKARNI, L. LANSING, A. PRESTA, J. BALDRIDGE, E. IE and D. GARCIA-OLANO. 2019, «Learning dense representations for entity retrieval», *arXiv preprint arXiv:1909.10506*. 48
- GIORGI, J. M., O. NITSKI, G. D. BADER and B. WANG. 2020, «Declutr: Deep contrastive learning for unsupervised textual representations», *arXiv preprint arXiv:2006.03659*. 48
- GUNEL, B., J. DU, A. CONNEAU and V. STOYANOV. 2020, «Supervised contrastive learning for pre-trained language model fine-tuning», *arXiv preprint arXiv:2011.01403*. 48
- HAZARIKA, D., R. ZIMMERMANN and S. PORIA. 2020, «Misa: Modality-invariant and-specific representations for multimodal sentiment analysis», *arXiv preprint arXiv:2005.03545*. 51
- HE, K., H. FAN, Y. WU, S. XIE and R. GIRSHICK. 2020, «Momentum contrast for unsupervised visual representation learning», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 9729–9738. 49
- HENAFF, O. 2020, «Data-efficient image recognition with contrastive predictive coding», in *International Conference on Machine Learning*, PMLR, p. 4182–4192. 49
- HEYLIGHEN, F. and J.-M. DEWAELE. 1999, «Formality of language: definition, measurement and behavioral determinants», *Interner Bericht, Center “Leo Apostel”, Vrije Universiteit Brussel*, vol. 4. 49
- HUANG, Z., W. XU and K. YU. 2015, «Bidirectional lstm-crf models for sequence tagging», *arXiv preprint arXiv:1508.01991*. 51
- KARPUKHIN, V., B. OĞUZ, S. MIN, P. LEWIS, L. WU, S. EDUNOV, D. CHEN and W.-T. YIH. 2020, «Dense passage retrieval for open-domain question answering», *arXiv preprint arXiv:2004.04906*. 48

- KHAN, F. S., R. M. ANWER, J. VAN DE WEIJER, A. D. BAGDANOV, M. VANRELL and A. M. LOPEZ. 2012, «Color attributes for object detection», in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, p. 3306–3313. [50](#), [51](#)
- KHOSLA, P., P. TETERWAK, C. WANG, A. SARNA, Y. TIAN, P. ISOLA, A. MASCHINOT, C. LIU and D. KRISHNAN. 2020, «Supervised contrastive learning», *arXiv preprint arXiv:2004.11362*. [48](#)
- KONG, L., C. D. M. D’AUTUME, W. LING, L. YU, Z. DAI and D. YOGATAMA. 2019, «A mutual information maximization perspective of language representation learning», *arXiv preprint arXiv:1910.08350*. [15](#), [47](#), [48](#), [49](#)
- LAMPLE, G. and A. CONNEAU. 2019, «Cross-lingual language model pretraining», *arXiv preprint arXiv:1901.07291*. [49](#)
- LINSKER, R. 1988, «Self-organization in a perceptual network», *Computer*, vol. 21, n° 3, p. 105–117. [47](#)
- LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTMAYER and V. STOYANOV. 2019, «Roberta: A robustly optimized bert pretraining approach», *arXiv preprint arXiv:1907.11692*. [47](#)
- LIU, Z., Y. SHEN, V. B. LAKSHMINARASIMHAN, P. P. LIANG, A. ZADEH and L.-P. MORENCY. 2018, «Efficient low-rank multimodal fusion with modality-specific factors», *arXiv preprint arXiv:1806.00064*. [50](#), [51](#)
- LOGESWARAN, L. and H. LEE. 2018, «An efficient framework for learning sentence representations», *arXiv preprint arXiv:1803.02893*. [48](#)
- LUONG, M.-T., H. PHAM and C. D. MANNING. 2015, «Effective approaches to attention-based neural machine translation», *arXiv preprint arXiv:1508.04025*. [51](#)
- MAI, S., H. HU and S. XING. 2019, «Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing», in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 481–492. [51](#)
- MIKOLOV, T., Q. V. LE and I. SUTSKEVER. 2013a, «Exploiting similarities among languages for machine translation», *arXiv preprint arXiv:1309.4168*. [49](#)
- MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. S. CORRADO and J. DEAN. 2013b, «Distributed representations of words and phrases and their compositionality», in *Advances in neural information processing systems*, p. 3111–3119. [45](#), [47](#)
- MITROVIC, J., B. MCWILLIAMS and M. REY. 2020, «Less can be more in contrastive learning», . [49](#)
- NALLAPATI, R., F. ZHAI and B. ZHOU. 2017, «Summarunner: A recurrent neural network based sequence model for extractive summarization of documents», in *Thirty-First AAAI Conference on Artificial Intelligence*. [51](#)



- NOJAVANASGHARI, B., D. GOPINATH, J. KOUSHIK, T. BALTRUŠAITIS and L.-P. MORENCY. 2016, «Deep multimodal fusion for persuasiveness prediction», in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ACM, p. 284–288. 50
- OORD, A. V. D., Y. LI and O. VINYALS. 2018a, «Representation learning with contrastive predictive coding», *arXiv preprint arXiv:1807.03748*. 47
- OORD, A. V. D., Y. LI and O. VINYALS. 2018b, «Representation learning with contrastive predictive coding», *arXiv preprint arXiv:1807.03748*. 48, 49
- PARK, S., H. S. SHIM, M. CHATTERJEE, K. SAGAE and L.-P. MORENCY. 2014, «Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach», in *Proceedings of the 16th International Conference on Multimodal Interaction*, p. 50–57. 50
- PENNINGTON, J., R. SOCHER and C. D. MANNING. 2014, «Glove: Global vectors for word representation», in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 47
- PETERS, M. E., M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE and L. ZETTLEMOYER. 2018a, «Deep contextualized word representations», in *Proc. of NAACL*. 45
- PETERS, M. E., M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE and L. ZETTLEMOYER. 2018b, «Deep contextualized word representations», *arXiv preprint arXiv:1802.05365*. 47
- PHAM, H., P. P. LIANG, T. MANZINI, L.-P. MORENCY and B. PÓCZOS. 2019, «Found in translation: Learning robust joint representations by cyclic translations between modalities», in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 6892–6899. 51
- QIAN, R., T. MENG, B. GONG, M.-H. YANG, H. WANG, S. BELONGIE and Y. CUI. 2021, «Spatiotemporal contrastive video representation learning», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 6964–6974. 48
- QU, Y., D. SHEN, Y. SHEN, S. SAJEEV, J. HAN and W. CHEN. 2020, «Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding», *arXiv preprint arXiv:2010.08670*. 48
- RADFORD, A., K. NARASIMHAN, T. SALIMANS and I. SUTSKEVER. 2018, «Improving language understanding by generative pre-training», URL <https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language-understanding-paper.pdf>. 45
- RAHMAN, W., M. K. HASAN, S. LEE, A. B. ZADEH, C. MAO, L.-P. MORENCY and E. HOQUE. 2020, «Integrating multimodal information in large pretrained transformers», in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2359–2369. 50, 52
- REDEKER, G. 1984, «On differences between spoken and written language», *Discourse processes*, vol. 7, n° 1, p. 43–55. 46, 49

- REIMERS, N. and I. GUREVYCH. 2019, «Sentence-bert: Sentence embeddings using siamese bert-networks», *arXiv preprint arXiv:1908.10084*. 48
- RETHMEIER, N. and I. AUGENSTEIN. 2021, «A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives», *arXiv preprint arXiv:2102.12982*. 48
- SCHROFF, F., D. KALENICHENKO and J. PHILBIN. 2015, «Facenet: A unified embedding for face recognition and clustering», in *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 815–823. 48
- SHEN, D., M. ZHENG, Y. SHEN, Y. QU and W. CHEN. 2020, «A simple but tough-to-beat data augmentation approach for natural language understanding and generation», *arXiv preprint arXiv:2009.13818*. 48
- SOLEYMANI, M., D. GARCIA, B. JOU, B. SCHULLER, S.-F. CHANG and M. PANTIC. 2017, «A survey of multimodal sentiment analysis», *Image and Vision Computing*, vol. 65, p. 3–14. 50
- STOLCKE, A. and E. SHRIBERG. 1996, «Statistical language modeling for speech disfluencies», in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, IEEE, p. 405–408. 46, 49
- SUÁREZ, P. J. O., B. SAGOT and L. ROMARY. 2019, «Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures», *Challenges in the Management of Large Corpora (CMLC-7) 2019*, p. 9. 47
- SUN, Z., P. SARMA, W. SETHARES and Y. LIANG. 2020, «Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis», in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, p. 8992–8999. 51
- THORNBURY, S. and D. SLADE. 2006, *Conversation: From description to pedagogy*, Cambridge University Press. 49
- TSAI, Y.-H. H., S. BAI, P. P. LIANG, J. Z. KOLTER, L.-P. MORENCY and R. SALAKHUTDINOV. 2019, «Multimodal transformer for unaligned multimodal language sequences», in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, NIH Public Access, p. 6558. 52
- WANG, F. and H. LIU. 2021, «Understanding the behaviour of contrastive loss», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 2495–2504. 48
- WANG, T. and P. ISOLA. 2020, «Understanding contrastive representation learning through alignment and uniformity on the hypersphere», in *International Conference on Machine Learning*, p. 9929–9939. 48
- WANG, Y., Y. SHEN, Z. LIU, P. P. LIANG, A. ZADEH and L.-P. MORENCY. 2018, «Words can shift: Dynamically adjusting word representations using nonverbal behaviors», . 52
- WU, C., F. WU and Y. HUANG. 2021, «Rethinking infonce: How many negative samples do you need?», *arXiv preprint arXiv:2105.13003*. 48

- XU, C., D. TAO and C. XU. 2013, «A survey on multi-view learning», *arXiv preprint arXiv:1304.5634*. 50
- YAMAGUCHI, S., S. KANAI, T. SHIODA and S. TAKEDA. 2019, «Multiple pretext-task for self-supervised learning via mixing multiple image transformations», *arXiv preprint arXiv:1912.11603*. 48
- YANG, Z., Z. DAI, Y. YANG, J. CARBONELL, R. R. SALAKHUTDINOV and Q. V. LE. 2019, «Xlnet: Generalized autoregressive pretraining for language understanding», in *Advances in neural information processing systems*, p. 5754–5764. 45, 47, 49, 52
- YE, J., H. HU, G.-J. QI and K. A. HUA. 2017, «A temporal order modeling approach to human action recognition from multimodal sensor data», *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, n° 2, p. 1–22. 51
- ZADEH, A., M. CHEN, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2017, «Tensor fusion network for multimodal sentiment analysis», *arXiv preprint arXiv:1707.07250*. 51
- ZADEH, A., P. P. LIANG, N. MAZUMDER, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2018a, «Memory fusion network for multi-view sequential learning», *arXiv preprint arXiv:1802.00927*. 50
- ZADEH, A., P. P. LIANG, N. MAZUMDER, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2018b, «Memory fusion network for multi-view sequential learning», in *Thirty-Second AAAI Conference on Artificial Intelligence*. 51
- ZADEH, A., P. P. LIANG, S. PORIA, P. VIJ, E. CAMBRIA and L.-P. MORENCY. 2018c, «Multi-attention recurrent network for human communication comprehension», in *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2018, NIH Public Access, p. 5642. 51
- ZADEH, A., C. MAO, K. SHI, Y. ZHANG, P. P. LIANG, S. PORIA and L.-P. MORENCY. 2019, «Factorized multimodal transformer for multimodal sequential learning», *arXiv preprint arXiv:1911.09826*. 50
- ZHANG, W. and K. STRATOS. 2021, «Understanding hard negatives in noise contrastive estimation», *arXiv preprint arXiv:2104.06245*. 48
- ZHU, Y., R. KIROS, R. ZEMEL, R. SALAKHUTDINOV, R. URTASUN, A. TORRALBA and S. FIDLER. 2015, «Aligning books and movies: Towards story-like visual explanations by watching movies and reading books», in *Proceedings of the IEEE international conference on computer vision*, p. 19–27. 47
- ZIMMERMANN, M., Y. LIU, E. SHRIBERG and A. STOLCKE. 2005, «Toward joint segmentation and classification of dialog acts in multiparty meetings», in *International Workshop on Machine Learning for Multimodal Interaction*, Springer, p. 187–193. 46



## Chapter 4

# Controlled sentence generation and automatic evaluation of NLG

### Chapter 4 abstract

Recently sequence-to-sequence (seq2seq) neural networks have been widely used in various language-based applications as they have flexible capabilities. Although seq2seq generally output grammatical, coherent sentences, controlling discrete attributes of the generated text (*e.g* polarity, tense) remains an open problem. In this chapter, we first introduce the problem of textual style transfer which aims at controlling the style of a generated sentence. Then, we present the problem of Automatic Evaluation (AE) of NLG. AE is a key problem towards better NLG systems [SPECIA and collab., 2010] as it allows to assess the quality of generated text without relying on human evaluation campaigns that are expensive and time consuming [BELZ and REITER, 2006]. Thus designing automatic and effective metrics has two simultaneous goals: (i) to be able to compare, to control and to debug systems without relying on human annotators [PEYRARD, 2018]; and (ii) to improve the learning phase of models by deriving losses that are a better surrogate of human judgment than the widely used cross-entropy loss [CLARK and collab., 2019]. In this chapter, we present the two aforementioned problems that will be tackled in Part III.

## 4.1 Controlled Sentence Generation

### 4.1.1 Context and Problem Statement

Due to recent breakthroughs in Artificial Intelligence, the use of chatbots has become more prevalent. Existing systems are mainly focused on functional aspects of chatbots: keywords extraction, natural language understanding, and pertinence of generated responses. Although these aspects are indeed key features for building a commercial chatbot, most of the existing solutions lack social intelligence. From a functional point of view, social intelligence could help by (1) avoiding interaction problems (anger, user indifference) that arise when the bot does not understand the user request [MASLOWSKI and collab.], and (2) building a relationship with the user. One step to make chatbots more social is to output sequences expressing emotion in a controlled manner, without sacrificing either grammatical correctness, coherence,

or relevance. In this chapter, we propose to explore the problem of conditional text generation. Formally, given an input text  $X$  and a target label  $Y$ , the goal is to produce a grammatically correct sentence that contains the label  $Y$  while preserving most of the content of  $X$ . Two popular tasks corresponding to this framework are conditional sentence generation and style transfer.

### 4.1.2 Related Work

The task of conditional sentence generation consists of taking as input a text containing specific stylistic properties to generate then a realistic (synthetic) text containing potentially different stylistic properties. It requests to learn a model  $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$  that maps a pair of inputs  $(x, y^g)$  to a sentence  $x^g$ , where the outcome sentence should retain as much as possible of the original content from the input sentence while having (potentially a new) attribute  $y^g$ . Proposed approaches to tackle textual style transfer [XU and collab., 2019; ZHANG and collab., 2020] can be divided into two main categories. The first category [LAMPLE and collab., 2018; PRABHUMOYE and collab., 2018] uses cycle losses based on back translation [WIETING and collab., 2017] to ensure that the content is preserved during the transformation. Whereas, the second category looks to explicitly separate attributes from the content. This constraint is enforced using either adversarial training [FU and collab., 2017; HU and collab., 2017; ZHANG and collab., 2018] or MI minimisation using CLUB [CHENG and collab., 2020]. Traditional adversarial training is based on an encoder that aims to fool the adversary discriminator by removing attribute information from the content embedding [ELAZAR and GOLDBERG, 2018].

### 4.1.3 Problem Formulation

For this task, aforementioned previous works rely on an encoder  $f_{\theta_e}$  taking as input a random sentence  $X$  and maps it to a random representation  $Z$  using  $f_{\theta_e}$ . Then, classification and sentence generation are performed using either a classifier or an auto-regressive decoder denoted by  $f_{\theta_d}$ . We aim at minimizing MI between the latent code represented by the r.v.  $Z = f_{\theta_e}(X)$  and the desired attribute represented by the r.v.  $Y$ . The objective of interest  $\mathcal{L}(f_{\theta_e})$  is defined as:

$$\mathcal{L}(f_{\theta_e}) \equiv \underbrace{\mathcal{L}_{down.}(f_{\theta_e})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_{\theta_e}(X); Y)}_{\text{disentangled}}, \quad (4.1)$$

where  $\mathcal{L}_{down.}$  represents a downstream specific (target task) loss and  $\lambda$  is a meta-parameter that controls the sensitive trade-off between disentanglement (*i.e.*, minimizing MI) and success in the downstream task (*i.e.*, minimizing the target loss). In [section 7.5](#), we illustrate theses different trade-offs.

### 4.1.4 Evaluation Approach

Automatic evaluation of generative models for text is still an open research problem. Sentences generated by the model are expected to be fluent, to preserve the input content and to contain the desired style. Concurrent works rely both on perceptive evaluation and automatic measures to evaluate the model quality through four criteria:

- C1 measures fluency: are the generated sequences grammatically correct and fluent?
- C2 evaluates label transfer: does the label present in the generated sequence match the target label?
- C3 measures the content preservation: does the generated sentence have the same content as the input sentence?
- C4 measures the disentanglement of the latent space: are we learning disentangled representations?

In the following we present the automatic metrics used for our evaluation.

**C1: Fluency evaluation.** Motivated by previous work, we evaluate the fluency of the language with the perplexity given by a GPT-2 [RADFORD and collab., 2018] pretrained model performing fine-tuning on the training corpus. We choose to report the log-perplexity since we believe it can better reflect the uncertainty of the language model (a small variation in the model loss would induce a large change in the perplexity due to the exponential term).

**C2: Style transfer.** For style transfer, the desired style is different from the input style while for conditional sentence generation, both input and output styles should be similar. We measure the style of the output sentence by using a fastText classifier [JOULIN and collab., 2016].

**C3: Content preservation.** For content preservation, we follow JOHN and collab. [2018] and compute both: (i) the cosine measure between source and generated sentence embeddings, which are the concatenation of min, max, and mean of word embeddings (sentiment words removed), and (ii) the BLEU score between generated text and the input using SACREBLEU from POST [2018].

**C4: Efficiency measure of the disentanglement methods.** BARRETT and collab. [2019] reports that offline classifiers (post training) outperform clearly adversarial discriminators. We will re-train a classifier on the latent representation learned by the model and we will report its accuracy.

#### 4.1.5 Limitations of Previous Approaches

Learning disentangled representations of textual data is essential for many natural language tasks such as style transfer and sentence generation, among others. The existent dominant approaches in the context of text data have been based on training an adversary (discriminator or teacher) that aims at making attribute values difficult to be inferred from the latent code. Although these approaches are remarkably simple and even though the adversary seems to be performing perfectly during the training phase, after training is completed a fair amount of sensitive information to infer the attribute still remains. In Chapter 7 we propose a novel objective to train disentangled representations from attributes. It overcomes some known limitations of adversarial losses to learn disentangled representations as we propose to minimize our novel bound on Mutual Information (MI) between the latent code and the attribute.

## 4.2 Evaluation of NLG

### 4.2.1 Introduction

A plethora of applications of natural language processing (NLP) perform text-to-text transformations [BELZ and REITER, 2006; MELLISH and DALE, 1998; SPECIA and collab., 2018] that is, given a text, these systems are required to produce a text that is coherent, readable and informative. Due to both high annotation costs and time requirements, researchers tend to rely on automatic evaluation to compare the output of such systems. Reference-based automatic evaluation relies on comparing a candidate text produced by the NLG system and one or multiple reference texts ('gold standard') created by a human annotator. Generic automatic evaluation of NLG is a huge challenge as it requires building a metric that evaluates semantic equivalence between a candidate and one or several gold-standard reference texts. However, the definition of semantic equivalence is task-specific: as an example, evaluation of text summarization focuses on content, coherence, grammatically correctness, conciseness, and readability [MANI, 2001], whereas machine translation focuses on fidelity, fluency and adequacy of the translation [HOVY, 1999; WHITE and collab., 1994] and data2text generation [DUŠEK and collab., 2020; GARDENT and collab., 2017; TIAN and collab., 2019] considers criteria such as data coverage, correctness and text structure.

Automatic text evaluation is an active area of research and a plethora of metrics have been previously proposed. They fall into two categories: metrics that are trained to maximize their correlation using human annotation (*e.g.*, RUSE [SHIMANAKA and collab., 2018], BLANC [LITA and collab., 2005], BEER [STANOJEVIĆ and SIMA'AN, 2014], BLEND [MA and collab., 2017], Q-Metrics [NEMA and KHAPRA, 2018], SIMILE [WIETING and collab., 2019]) and untrained metrics (*e.g.*, BLEU [PAPINENI and collab., 2002], ROUGE [LIN, 2004], BERTSCORE [ZHANG and collab., 2019], Word Mover Distance [KUSNER and collab., 2015]). In Chapter 8, we focus on untrained metrics as they do not require costly training<sup>1</sup>. Two categories of untrained metrics can be distinguished: word or character based-metrics that compute a score based on string representation of the texts and embedding-based metrics that rely on a continuous representation of the text. String-based metrics (*e.g.*, BLEU, METEOR) often fail to robustly match paraphrases [REITER and BELZ, 2009] as they mainly focus on the surface form (*e.g.*, string representation of the metric) as opposed to embedding-based metrics that leverage continuous representations.

In this section, we start by introducing notations and formulate the problem of both evaluating text generation and metrics. Then, we identify and present the most relevant related work and the existing approaches for the studied tasks.

### 4.2.2 Problem statement

**NLG evaluation.** Given a dataset  $\mathcal{D} = \{\mathbf{x}_i, \{\tilde{\mathbf{x}}_i^s, h_{\mathbf{x}_i}(\tilde{\mathbf{x}}_i^s)\}_{s=1}^S\}_{i=1}^N$  where  $\mathbf{x}_i$  is the  $i$ -th reference text;  $\tilde{\mathbf{x}}_i^s$  is the  $i$ -th candidate text generated by the  $s$ -th NLG system;  $N$  is the number of text in the dataset and  $S$  the number of systems available. The vector  $\mathbf{x}_i = (\omega_1, \dots, \omega_M)$  is composed of  $M$  tokens (*e.g.*, words or subwords) and

---

<sup>1</sup>Existing labelled corpora are of small size thus trained metrics may not generalize well to new data.

$\tilde{\mathbf{x}}_i^s = (\tilde{\omega}_1, \dots, \tilde{\omega}_L)$  is composed of  $L$  tokens<sup>2</sup>.  $h_{\mathbf{x}_i}(\tilde{\mathbf{x}}_i^s)$  is the score associated by a human annotator to the candidate text  $\tilde{\mathbf{x}}_i^s$  when comparing it with the reference text  $\mathbf{x}_i$ . We aim at building an evaluation metric  $f$  such that  $f(\mathbf{x}_i, \tilde{\mathbf{x}}_i) \in \mathbb{R}_{>0}$ .

**Evaluating evaluation metrics.** To assess the relevance of an evaluation metric  $f$ , correlation with human judgment is considered to be one of the most important criterion [BANERJEE and LAVIE, 2005; CHATZIKOUMI, 2020; KOEHN, 2009; SPECIA and collab., 2010]. Debate on the relative merits of different correlation for the evaluation of automatic metrics is ongoing, but classical correlation measures are Pearson [LEUSCH and collab., 2003], Spearman [MELAMED and collab., 2003] or Kendall test [KENDALL, 1938]. Two meta evaluation strategies are commonly used: (1) text-level correlation or (2) system-level correlation. Formally, the sentence-level correlation  $C_{t,f}$  is computed as follows:

$$C_{t,f} \triangleq \frac{1}{N} \sum_{i=1}^N K(\mathbf{F}_i^t, \mathbf{H}_i^t), \quad (4.2)$$

where  $\mathbf{F}_i = [f(\mathbf{x}_i, \tilde{\mathbf{x}}_i^1), \dots, f(\mathbf{x}_i, \tilde{\mathbf{x}}_i^S)]$  and  $\mathbf{H}_i = [h_{\mathbf{x}_i}(\tilde{\mathbf{x}}_i^1), \dots, h_{\mathbf{x}_i}(\tilde{\mathbf{x}}_i^S)]$  are the vectors composed of scores assigned by the automatic metric  $f$  and the human respectively. and  $K: \mathbb{R}^N \times \mathbb{R}^N \rightarrow [-1, 1]$  is the chosen correlation measure (e.g., Pearson, Kendall or Spearman). Similarly, the system level correlation  $C_{sy,f}$  is obtained by

$$C_{sy,f} \triangleq K(\mathbf{F}^{sy}, \mathbf{H}^{sy}), \quad (4.3)$$

$$\mathbf{F}^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \tilde{\mathbf{x}}_i^1), \dots, \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \tilde{\mathbf{x}}_i^S) \right]$$

$$\mathbf{H}^{sy} = \left[ \frac{1}{N} \sum_{i=1}^N h_{\mathbf{x}_i}(\tilde{\mathbf{x}}_i^1), \dots, \frac{1}{N} \sum_{i=1}^N h_{\mathbf{x}_i}(\tilde{\mathbf{x}}_i^S) \right],$$

where the latter are the vectors composed of the averaged scores assigned by  $f$  and the human, respectively. For the significance analysis, we follow GRAHAM [2015]; GRAHAM and BALDWIN [2014]; GRAHAM and collab. [2015]<sup>3</sup>. They use a William test to validate a significant improvement for dependent correlations [STEIGER, 1980].

### 4.2.3 Existing metrics

In this section, we start by reviewing existing untrained metrics which can be grouped into two categories (e.g., string-based and embedding-based metrics) and then provide a short overview of training-based metrics.

#### String-Based Metrics

There are two types of string-based metrics: *N-Grams matching* metrics and *Edit distance-based* metrics.

**N-Grams matching metrics** count the number of  $N$ -grams in common between the candidate text and the reference text. Thus they are usually lightweight. The three most-used metrics are BLEU [PAPINENI and collab., 2002], ROUGE [LIN, 2004] and

<sup>2</sup>The reference and candidate text can be composed of several sentences as it is the case in summarization.

<sup>3</sup>Code of the authors is available at <https://github.com/ygraham/nlp-williams>

METEOR [BANERJEE and LAVIE, 2005]. If no n-gram is in common between the input text candidate and the reference, these metrics fail to produce meaningful scores. Several revised version of BLEU [DODDINGTON, 2002; GALLEY and collab., 2015; POPOVIĆ, 2015, 2017] and METEOR [DENKOWSKI and LAVIE, 2014; GUO and HU, 2019] have been proposed in the recent years.

**Edit distance based metrics.** The second category of metrics measures the number of basic operations such as edit/delete/insert to measure semantic equivalence (*i.e.*, using Levenshtein distance [LEVENSHTAIN, 1966]). Variants include TER [SNOVER and collab., 2006], CDER [LEUSCH and collab., 2006], EED [STANCHEV and collab., 2019], CHARACTER [WANG and collab., 2016]. Similarly to previous approaches these metrics do not handle synonyms and focus on surface form.

### Embedding-Based Metrics

Another class of metrics relies on word embeddings. These metrics either use static word embeddings such as Glove [PENNINGTON and collab., 2014], word2vec [MIKOLOV and collab., 2013] or contextualized embeddings such as ELMO [PETERS and collab., 2018], BERT [DEVLIN and collab., 2018] and its variants [LIU and collab., 2019; SANH and collab., 2019]. Among the most popular metrics we can mention MoverScore [ZHAO and collab., 2019], BERTScore [ZHANG and collab., 2019], SentenceMover [CLARK and collab., 2019], WMD [KUSNER and collab., 2015], WMDO [CHOW and collab., 2019], MEANT [LO, 2017; LO and WU, 2011] and YISI [LO and collab., 2018]. Contextualized embeddings achieve better results, but it remains an open question how to find the combination of layers that leads to the best results<sup>4</sup>. InfoLM addresses this issue by relying on the language model only.

### Learning-Based Metrics

Various trained metrics have been proposed such as BEER [STANOJEVIĆ and SIMA'AN, 2014], BLEND [MA and collab., 2017], RUSE [SHIMANAKA and collab., 2018], CIDER [VEDANTAM and collab., 2015]. Because of the learning phase, these methods require a training validation and testing set composed of human evaluations. Different from these approaches InfoLM relies on a pretrained LM.

**Use of pretrained LM as a metric.** In text generation (*e.g.*, style transfer, news generation), LM is (optionnaly) fine-tuned to measure perplexity and assess the fluency of the generated sentences.

#### 4.2.4 Weakness of Pretrained Embedding-Based Metrics

Current metrics based on pretrained embeddings such as BERT (or other contextualized embeddings) take advantage of the contextualized pretrained representations. However, they rely on a first arbitrary operation to aggregate layer information (average or single layer selection) followed by a second arbitrary operation (optimal transport, cosine similarity) to transform the previously obtained vector in a real number. In Chapter 8, we first get rid of the first operation by working with the output distribution of the LM. Then, we leverage the discrete nature of the output distributions to use discrete measures of information. The geometrical interpretations of the discrete measures of information allow us to better interpret the proposed metric.

---

<sup>4</sup>BertScore uses a different layer for each model while MoverScore uses the 5 last layers.



### Chapter 4 conclusion

In this chapter, we introduced two problems related to NLG, namely controlled conditional text generation and text evaluation. The latter is closely related to the former as it could be used to derive new losses. In both cases, we presented the limitation of existing approaches that we will address in [Part III](#). In the next chapter, we will present our contributions related to RQ1.

## 4.3 References

- BANERJEE, S. and A. LAVIE. 2005, «Meteor: An automatic metric for mt evaluation with improved correlation with human judgments», in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, p. 65–72. [63](#), [64](#)
- BARRETT, M., Y. KEMENTCHEDJHIEVA, Y. ELAZAR, D. ELLIOTT and A. SØGAARD. 2019, «Adversarial removal of demographic attributes revisited», in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 6331–6336. [61](#)
- BELZ, A. and E. REITER. 2006, «Comparing automatic and human evaluation of nlg systems», in *11th Conference of the European Chapter of the Association for Computational Linguistics*. [59](#), [62](#)
- CHATZIKOUMI, E. 2020, «How to evaluate machine translation: A review of automated and human metrics», *Natural Language Engineering*, vol. 26, n° 2, p. 137–161. [63](#)
- CHENG, P., M. R. MIN, D. SHEN, C. MALON, Y. ZHANG, Y. LI and L. CARIN. 2020, «Improving disentangled text representation learning with information-theoretic guidance», *arXiv preprint arXiv:2006.00693*. [60](#)
- CHOW, J., L. SPECIA and P. MADHYASTHA. 2019, «WMDO: Fluency-based word mover’s distance for machine translation evaluation», in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Association for Computational Linguistics, Florence, Italy, p. 494–500, doi: 10.18653/v1/W19-5356. URL <https://www.aclweb.org/anthology/W19-5356>. [64](#)
- CLARK, E., A. CELIKYILMAZ and N. A. SMITH. 2019, «Sentence mover’s similarity: Automatic evaluation for multi-sentence texts», in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 2748–2760, doi: 10.18653/v1/P19-1264. URL <https://www.aclweb.org/anthology/P19-1264>. [59](#), [64](#)
- DENKOWSKI, M. and A. LAVIE. 2014, «Meteor universal: Language specific translation evaluation for any target language», in *Proceedings of the ninth workshop on statistical machine translation*, p. 376–380. [64](#)

- DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805*. 64
- DODDINGTON, G. 2002, «Automatic evaluation of machine translation quality using n-gram co-occurrence statistics», in *Proceedings of the second international conference on Human Language Technology Research*, p. 138–145. 64
- DUŠEK, O., J. NOVIKOVA and V. RIESER. 2020, «Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge», *Computer Speech & Language*, vol. 59, p. 123–156. 62
- ELAZAR, Y. and Y. GOLDBERG. 2018, «Adversarial removal of demographic attributes from text data», *arXiv preprint arXiv:1808.06640*. 60
- FU, Z., X. TAN, N. PENG, D. ZHAO and R. YAN. 2017, «Style transfer in text: Exploration and evaluation», *arXiv preprint arXiv:1711.06861*. 60
- GALLEY, M., C. BROCKETT, A. SORDONI, Y. JI, M. AULI, C. QUIRK, M. MITCHELL, J. GAO and B. DOLAN. 2015, «deltaleu: A discriminative metric for generation tasks with intrinsically diverse targets», *arXiv preprint arXiv:1506.06863*. 64
- GARDENT, C., A. SHIMORINA, S. NARAYAN and L. PEREZ-BELTRACHINI. 2017, «Creating training corpora for nlg micro-planning», in *55th annual meeting of the Association for Computational Linguistics (ACL)*. 62
- GRAHAM, Y. 2015, «Improving evaluation of machine translation quality estimation», in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 1804–1813. 63
- GRAHAM, Y. and T. BALDWIN. 2014, «Testing for significance of increased correlation with human judgment», in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 172–176. 63
- GRAHAM, Y., T. BALDWIN and N. MATHUR. 2015, «Accurate evaluation of segment-level machine translation metrics», in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 1183–1191. 63
- GUO, Y. and J. HU. 2019, «Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation», in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Association for Computational Linguistics, Florence, Italy, p. 501–506, doi: 10.18653/v1/W19-5357. URL <https://www.aclweb.org/anthology/W19-5357>. 64
- HOVY, E. H. 1999, «Toward finely differentiated evaluation metrics for machine translation», in *Proceedings of the EAGLES Workshop on Standards and Evaluation Pisa, Italy, 1999*. 62
- HU, Z., Z. YANG, X. LIANG, R. SALAKHUTDINOV and E. P. XING. 2017, «Toward controlled generation of text», *arXiv preprint arXiv:1703.00955*. 60



- JOHN, V., L. MOU, H. BAHULEYAN and O. VECHTOMOVA. 2018, «Disentangled representation learning for non-parallel text style transfer», *arXiv preprint arXiv:1808.04339*. 61
- JOULIN, A., E. GRAVE, P. BOJANOWSKI and T. MIKOLOV. 2016, «Bag of tricks for efficient text classification», *arXiv preprint arXiv:1607.01759*. 61
- KENDALL, M. G. 1938, «A new measure of rank correlation», *Biometrika*, vol. 30, n° 1/2, p. 81–93. 63
- KOEHN, P. 2009, *Statistical machine translation*, Cambridge University Press. 63
- KUSNER, M., Y. SUN, N. KOLKIN and K. WEINBERGER. 2015, «From word embeddings to document distances», in *International conference on machine learning*, PMLR, p. 957–966. 62, 64
- LAMPLE, G., S. SUBRAMANIAN, E. SMITH, L. DENOYER, M. RANZATO and Y.-L. BOUREAU. 2018, «Multiple-attribute text rewriting», in *International Conference on Learning Representations*. 60
- LEUSCH, G., N. UEFFING and H. NEY. 2006, «CDER: Efficient MT evaluation using block movements», in *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Trento, Italy. URL <https://www.aclweb.org/anthology/E06-1031>. 64
- LEUSCH, G., N. UEFFING, H. NEY and collab.. 2003, «A novel string-to-string distance measure with applications to machine translation evaluation», in *Proceedings of Mt Summit IX*, p. 240–247. 63
- LEVENSHTEIN, V. I. 1966, «Binary codes capable of correcting deletions, insertions, and reversals», in *Soviet physics doklady*, vol. 10, Soviet Union, p. 707–710. 64
- LIN, C.-Y. 2004, «ROUGE: A package for automatic evaluation of summaries», in *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, p. 74–81. URL <https://www.aclweb.org/anthology/W04-1013>. 62, 63
- LITA, L. V., M. ROGATI and A. LAVIE. 2005, «Blanc: Learning evaluation metrics for mt», in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 740–747. 62
- LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTMAYER and V. STOYANOV. 2019, «Roberta: A robustly optimized bert pretraining approach», *arXiv preprint arXiv:1907.11692*. 64
- LO, C.-K. 2017, «Meant 2.0: Accurate semantic mt evaluation for any output language», in *Proceedings of the second conference on machine translation*, p. 589–597. 64
- LO, C.-K., M. SIMARD, D. STEWART, S. LARKIN, C. GOUTTE and P. LITTELL. 2018, «Accurate semantic textual similarity for cleaning noisy parallel corpora using

- semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task», in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Association for Computational Linguistics, Belgium, Brussels, p. 908–916, doi: 10.18653/v1/W18-6481. URL <https://www.aclweb.org/anthology/W18-6481>. 64
- LO, C.-K. and D. WU. 2011, «Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles», in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, p. 220–229. 64
- MA, Q., Y. GRAHAM, S. WANG and Q. LIU. 2017, «Blend: a novel combined mt metric based on direct assessment—casict-dcu submission to wmt17 metrics task», in *Proceedings of the second conference on machine translation*, p. 598–603. 62, 64
- MANI, I. 2001, *Automatic summarization*, vol. 3, John Benjamins Publishing. 62
- MASLOWSKI, I., D. LAGARDE and C. CLAVEL. «In-the-wild chatbot corpus: from opinion analysis to interaction problem detection», . 59
- MELAMED, I. D., R. GREEN and J. TURIAN. 2003, «Precision and recall of machine translation», in *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, p. 61–63. 63
- MELLISH, C. and R. DALE. 1998, «Evaluation in the context of natural language generation», *Computer Speech & Language*, vol. 12, n° 4, p. 349–373. 62
- MIKOLOV, T., K. CHEN, G. CORRADO and J. DEAN. 2013, «Efficient Estimation of Word Representations in Vector Space», *arXiv preprint arXiv:1301.3781*. URL <https://arxiv.org/abs/1301.3781>. 64
- NEMA, P. and M. M. KHAPRA. 2018, «Towards a better metric for evaluating question generation systems», *arXiv preprint arXiv:1808.10192*. 62
- PAPINENI, K., S. ROUKOS, T. WARD and W.-J. ZHU. 2002, «Bleu: a method for automatic evaluation of machine translation», in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 311–318, doi: 10.3115/1073083.1073135. URL <https://www.aclweb.org/anthology/P02-1040>. 62, 63
- PENNINGTON, J., R. SOCHER and C. D. MANNING. 2014, «Glove: Global vectors for word representation», in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 64
- PETERS, M. E., M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE and L. ZETTMELMOYER. 2018, «Deep contextualized word representations», in *Proc. of NAACL*. 64
- PEYRARD, M. 2018, «A simple theoretical model of importance for summarization», *arXiv preprint arXiv:1801.08991*. 59
- POPOVIĆ, M. 2015, «chrF: character n-gram f-score for automatic mt evaluation», in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392–395. 64

- POPOVIĆ, M. 2017, «chrf++: words helping character n-grams», in *Proceedings of the second conference on machine translation*, p. 612–618. 64
- POST, M. 2018, «A call for clarity in reporting bleu scores», *arXiv preprint arXiv:1804.08771*. 61
- PRABHUMOYE, S., Y. TSVETKOV, R. SALAKHUTDINOV and A. W. BLACK. 2018, «Style transfer through back-translation», *arXiv preprint arXiv:1804.09000*. 60
- RADFORD, A., K. NARASIMHAN, T. SALIMANS and I. SUTSKEVER. 2018, «Improving language understanding by generative pre-training», URL <https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language-understanding-paper.pdf>. 61
- REITER, E. and A. BELZ. 2009, «An investigation into the validity of some metrics for automatically evaluating natural language generation systems», *Computational Linguistics*, vol. 35, n° 4, p. 529–558. 62
- SANH, V., L. DEBUT, J. CHAUMOND and T. WOLF. 2019, «Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter», *arXiv preprint arXiv:1910.01108*. 64
- SHIMANAKA, H., T. KAJIWARA and M. KOMACHI. 2018, «Ruse: Regressor using sentence embeddings for automatic machine translation evaluation», in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, p. 751–758. 62, 64
- SNOVER, M., B. DORR, R. SCHWARTZ, L. MICCIULLA and J. MAKHOUL. 2006, «A study of translation edit rate with targeted human annotation», in *Proceedings of association for machine translation in the Americas*, vol. 200, Cambridge, MA. 64
- SPECIA, L., D. RAJ and M. TURCHI. 2010, «Machine translation evaluation versus quality estimation», *Machine translation*, vol. 24, n° 1, p. 39–50. 59, 63
- SPECIA, L., C. SCARTON and G. H. PAETZOLD. 2018, «Quality estimation for machine translation», *Synthesis Lectures on Human Language Technologies*, vol. 11, n° 1, p. 1–162. 62
- STANCHEV, P., W. WANG and H. NEY. 2019, «Eed: Extended edit distance measure for machine translation», in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, p. 514–520. 64
- STANOJEVIĆ, M. and K. SIMA'AN. 2014, «BEER: BEtter evaluation as ranking», in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Baltimore, Maryland, USA, p. 414–419, doi: 10.3115/v1/W14-3354. URL <https://www.aclweb.org/anthology/W14-3354>. 62, 64
- STEIGER, J. H. 1980, «Tests for comparing elements of a correlation matrix.», *Psychological bulletin*, vol. 87, n° 2, p. 245. 63
- TIAN, R., S. NARAYAN, T. SELLAM and A. P. PARIKH. 2019, «Sticking to the facts: Confident decoding for faithful data-to-text generation», *arXiv preprint arXiv:1910.08684*. 62

- VEDANTAM, R., C. LAWRENCE ZITNICK and D. PARIKH. 2015, «Cider: Consensus-based image description evaluation», in *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4566–4575. [64](#)
- WANG, W., J.-T. PETER, H. ROSENDAHL and H. NEY. 2016, «Character: Translation edit rate on character level», in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, p. 505–510. [64](#)
- WHITE, J. S., T. A. O’CONNELL and F. E. O’MARA. 1994, «The arpa mt evaluation methodologies: evolution, lessons, and future approaches», in *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. [62](#)
- WIETING, J., T. BERG-KIRKPATRICK, K. GIMPEL and G. NEUBIG. 2019, «Beyond bleu: Training neural machine translation with semantic similarity», *arXiv preprint arXiv:1909.06694*. [62](#)
- WIETING, J., J. MALLINSON and K. GIMPEL. 2017, «Learning paraphrastic sentence embeddings from back-translated bitext», *arXiv preprint arXiv:1706.01847*. [60](#)
- XU, R., T. GE and F. WEI. 2019, «Formality style transfer with hybrid textual annotations», *arXiv preprint arXiv:1903.06353*. [60](#)
- ZHANG, T., V. KISHORE, F. WU, K. Q. WEINBERGER and Y. ARTZI. 2019, «Bertscore: Evaluating text generation with bert», *arXiv preprint arXiv:1904.09675*. [62](#), [64](#)
- ZHANG, Y., N. DING and R. SORICUT. 2018, «Shaped: Shared-private encoder-decoder for text style adaptation», *arXiv preprint arXiv:1804.04093*. [60](#)
- ZHANG, Y., T. GE and X. SUN. 2020, «Parallel data augmentation for formality style transfer», *arXiv preprint arXiv:2005.07522*. [60](#)
- ZHAO, W., M. PEYRARD, F. LIU, Y. GAO, C. M. MEYER and S. EGER. 2019, «Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance», *arXiv preprint arXiv:1909.02622*. [64](#)

## **Part II**

# **Integrating Conversational and Multimodal Dimensions in Transcripts Representations via MI Maximization**



---

## Part II Introduction

In this part, we aim at providing answers to the set of research questions RQ1 and other related subquestions presented in [Chapter 1](#). We thus present our contribution to the problem of learning transcript representations using MI. This second part is split in two chapters that covers two different aspects:

- In [Chapter 5](#), we describe a new method to learn generic text spoken transcript representations by integrating the conversational dimension. We obtain our representations with a hierarchical encoder based on transformer architectures, for which we propose two new pre-training objectives that are tightly linked to MI. Pre-training is performed on OpenSubtitles: a large corpus of movie subtitles containing over 2.3 billion of tokens. These representations are evaluated on a new benchmark we call Sequence labelling evaluation benchmark for spoken language benchmark (SILICONE). Our hierarchical pre-training method achieves competitive results with consistently fewer parameters compared to state-of-the-art models and we show their importance for both pre-training and fine-tuning.
- In [Chapter 6](#), we enrich transcripts representation with multi-modal dimensions. So far, a consequent effort has been made on developing complex architectures for multimodal representation learning allowing the fusion of these modalities. We investigate unexplored penalties and propose a set of new objectives that measure the dependency between modalities. Our new fusion method, which can be combined with both randomly initialized encoder and pretrained representations, not only achieves a new SOTA on sentiment analysis but also produces representations that are more robust to modality drops. Finally, a by-product of our methods includes a statistical network which can be used to interpret the high dimensional representations learnt by the model.

---



# Chapter 5

## Integrating Conversational Dimension in Pretrained Representation

### Chapter 5 abstract

This chapter is dedicated to our first contribution where we propose a new approach to learn generic representations adapted to spoken transcripts by including the conversational dimension. We evaluate the learnt representations on a new benchmark we call Sequence labelling evaluation benchmark for spoken language benchmark (SILICONE). SILICONE<sup>a</sup> is model-agnostic and contains 10 different datasets of various sizes. We obtain our representations with a hierarchical encoder based on transformer architectures, for which we extend two well-known pre-training objectives (*i.e* Masked Language Model and Generalized Autoregressive Pretraining). We study the connection of the new pretraining objectives to MI [section 5.3](#). We process and gather a large pre-training corpora extracted from OpenSubtitles: a large corpus of spoken dialog containing over 2.3 billion of tokens. We demonstrate how hierarchical encoders achieve competitive results with consistently fewer parameters compared to state-of-the-art models and we show their importance for both pre-training and fine-tuning.

<sup>a</sup>Benchmark can be found in the dataset library from HuggingFace [[WOLF and collab., 2020](#)] at <https://huggingface.co/datasets/silicone>

### 5.1 Introduction

Generic representations have been shown to be an effective way to adapt models across different sets of labels [[DEVLIN and collab., 2018](#); [LIU and collab., 2019](#); [MIKOLOV and collab., 2013](#); [PENNINGTON and collab., 2014](#); [PETERS and collab., 2018](#); [YANG and collab., 2019](#)]. Those representations are usually trained on large written corpora such as OSCAR [[SUÁREZ and collab., 2019](#)], Book Corpus [[ZHU and collab., 2015](#)] or Wikipedia [[DENoyer and GALLINARI, 2006](#)]. Although achieving state-of-the-art (SOTA) results on written benchmarks [[WANG and collab., 2018](#)], they are neither tailored for transcript nor for spoken dialog (SD) representation. Indeed, [TRAN and collab. \[2019\]](#) have suggested that training a parser on conversational speech data can improve results, due to the discrepancy between spoken and written

language (e.g. disfluencies [STOLCKE and SHRIBERG, 1996], fillers [DINKAR and collab., 2020; SHRIBERG, 1999]). Furthermore, capturing discourse-level features, which distinguish transcripts from other types of text [THORNBURY and SLADE, 2006], e.g., capturing multi-utterance dependencies, is a key to embed dialog that is not explicitly present in pre-training objectives [DEVLIN and collab., 2018; LIU and collab., 2019; YANG and collab., 2019], as they often treat sentences as a simple stream of tokens. *The goal of this work is to train on SD data a generic dialog encoder capturing discourse-level features that produce representations which integrate the conversational nature of the transcripts.*

**Evaluation.** To evaluate these pretrained representations we focus on Dialog Acts (DA) and Emotion/Sentiment (E/S). The automated identification of both DA and E/S in spoken language is an important step toward improving model performances on spontaneous dialogue tasks. Especially, it is essential to avoid the generic response problem, *i.e.*, having an automatic dialog system generate an unspecific response — that can be an answer to a very large number of user utterances COLOMBO and collab. [2019]; YI and collab. [2019]. DA and emotion identification JALALZAI and collab. [2020a]; WITON and collab. [2018a] are done through sequence labelling systems that are usually trained on large corpora (with over 100k labelled utterances) such as Switchboard [GODFREY and collab., 1992], MRDA [SHRIBERG and collab., 2004] or Daily Dialog Act [LI and collab., 2017]. Even though large corpora enable learning complex models from scratch (e.g., seq2seq [COLOMBO and collab., 2020]), those models are very specific to the labelling scheme employed. Adapting them to different sets of emotions or dialog acts would require more annotated data. We evaluate these representations on both DA and E/S labelling through a new benchmark SILICONE (Sequence labelling evaluation benchmark for spoken language) composed of datasets of varying sizes using different sets of labels.

We follow the general trend of using smaller models to obtain lightweight representations JIAO and collab. [2019]; LAN and collab. [2019] that can be trained without a costly computation infrastructure while achieving good performance on several downstream tasks [HENDERSON and collab., 2020]. Concretely, since hierarchy is an inherent characteristic of dialog [THORNBURY and SLADE, 2006], we propose the first hierarchical generic multi-utterance encoder based on a hierarchy of transformers. This allows us to factorise the model parameters, getting rid of long term dependencies and enabling training on a reduced number of GPUs.

Based on this hierarchical structure, we generalise two existing pre-training objectives. As embeddings highly depend on data quality [LE and collab., 2019] and volume [LIU and collab., 2019], we preprocess OpenSubtitles [LISON and collab., 2019]: a large corpus of spoken dialog from movies. This corpora is an order of magnitude bigger than corpora [BUDZIANOWSKI and collab., 2018b; DANESCU-NICULESCU-MIZIL and LEE, 2011; LOWE and collab., 2015] used in previous works [HAZARIKA and collab., 2019; MEHRI and collab., 2019]. Lastly, we evaluate our encoder along with other baselines on SILICONE, which lets us draw finer conclusions of the generalisation capability of our models<sup>1</sup>.

---

<sup>1</sup>Upon publication, we will release the code, models and especially the preprocessing scripts to replicate our results.

## 5.2 Method

We start by formally defining the Sequence Labelling Problem. At the highest level, we have a set  $D$  of conversations composed of utterances, *i.e.*,  $D = (C_1, C_2, \dots, C_{|D|})$  with  $Y = (Y_1, Y_2, \dots, Y_{|D|})$  being the corresponding set of labels (*e.g.*, DA, E/S). At a lower level each conversation  $C_i$  is composed of utterances  $u$ , *i.e.*  $C_i = (u_1, u_2, \dots, u_{|C_i|})$  with  $Y_i = (y_1, y_2, \dots, y_{|C_i|})$  being the corresponding sequence of labels: each  $u_i$  is associated with a unique label  $y_i$ . At the lowest level, each utterance  $u_i$  can be seen as a sequence of words, *i.e.*  $u_i = (\omega_1^i, \omega_2^i, \dots, \omega_{|u_i|}^i)$ . Concrete examples with dialog act can be found in Table 5.1.

Utterances	DA
How long does that take you to get to work?	qw
Uh, about forty-five, fifty minutes.	sd
How does that work, work out with, uh, storing your bike and showering and all that?	qw
Yeah ,	b
It can be a pain .	sd
It's, it's nice riding to school because it's all along a canal path, uh,	sd
Because it's just,	sd
it's along the Erie Canal up here.	sd
So, what school is it?	qw
Uh, University of Rochester.	sd
Oh, okay.	bk

Table 5.1 – Examples of dialogs labelled with DA taken from SwDA. The labels qw, sd, b, bk respectively correspond to wh-question, statement-non-opinion, backchannel and response acknowledgement.

### 5.2.1 Pre-training Objectives

Our work builds upon existing objectives designed to pre-train encoders: the Masked Language Model (MLM) from [DEVLIN and collab. \[2018\]](#); [LAN and collab. \[2019\]](#); [LIU and collab. \[2019\]](#); [ZHANG and collab. \[2019a\]](#) and the Generalized Autoregressive Pre-training (GAP) from [YANG and collab. \[2019\]](#).

**MLM Loss:** The MLM loss corrupts sequences (or in our case, utterances) by masking a proportion  $p_\omega$  of tokens. The model learns bidirectional representations by predicting the original identities of the masked-out tokens. Formally, for an utterance  $u_i$ , a random set of indexed positions  $m^{u_i}$  is selected and the associated tokens are replaced by a masked token [MASK] to obtain a corrupted utterance  $u_i^{\text{masked}}$ . The set of parameters  $\theta$  is learnt by maximizing :

$$\mathcal{L}_{\text{MLM}}^u(\theta, u_i) = \mathbb{E} \left[ \sum_{t \in m^{u_i}} \log(p_\theta(\omega_t^i | \tilde{u}_i)) \right] \quad (5.1)$$

where  $\tilde{u}_i$  is the corrupted utterance,  $m_j^{u_i} \sim \text{unif}[1, |u_i|] \forall j \in [1, p_\omega]$  and  $p_\omega$  is the proportion of masked tokens.

**GAP Loss:** the GAP loss consists in computing a classic language modelling loss across different factorisation orders of the tokens. In this way, the model will learn to gather

information across all possible positions from both directions. The set of parameters  $\theta$  is learnt by maximising:

$$\mathcal{L}_{\text{GAP}}^u(\theta, u_i) = \mathbb{E} \left[ \mathbb{E}_{\mathbf{z} \sim \mathbb{Z}_{|u_i|}} \left[ \sum_t \log p_{\theta}(\omega_{z_t}^i | u_i^{\mathbf{z}^{<t}}) \right] \right] \quad (5.2)$$

where  $\mathbb{Z}_{|u_i|}$  is the set of permutations of length  $|u_i|$  and  $u_i^{\mathbf{z}^{<t}}$  represent the first  $t$  tokens of  $u_i$  when permuting the sequence according to  $\mathbf{z} \in \mathbb{Z}_{|u_i|}$ .

### 5.2.2 Hierarchical Encoding

Capturing dependencies at different granularity levels is key for dialog embedding. Thus, we choose a hierarchical encoder [CHEN and collab., 2018b; LI and collab., 2018a]. It is composed of two functions  $f^u$  and  $f^c$ , satisfying:

$$\mathcal{E}_{u_i} = f_{\theta}^u(\omega_1, \dots, \omega_{|u_i|}) \quad (5.3)$$

$$\mathcal{E}_{C_j} = f_{\theta}^d(\mathcal{E}_{u_1}, \dots, \mathcal{E}_{C_j}) \quad (5.4)$$

where  $\mathcal{E}_{u_i} \in \mathbb{R}^{d_u}$  is the embedding of  $u_i$  and  $\mathcal{E}_{C_j} \in \mathbb{R}^{d_d}$  the embedding of  $C_j$ . The structure of the hierarchical encoder is depicted in Figure 5.1.

### 5.2.3 Hierarchical Pre-training

#### General Motivation

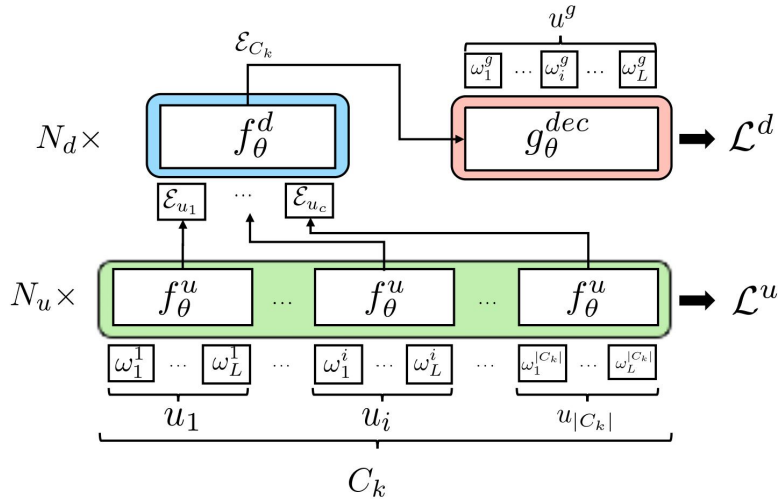


Figure 5.1 – General structure of our proposed hierarchical dialog encoder, with a decoder:  $f_{\theta}^u$ ,  $f_{\theta}^d$  and the sequence label decoder ( $g_{\theta}^{\text{dec}}$ ) are colored respectively in green, blue and red.

Current self-supervised pre-training objectives such as MLM and GAP are trained at the sequence level, which for us translates to only learning  $f_{\theta}^u$ . In this section, we extend both the MLM and GAP losses at the dialog level in order to pre-train  $f_{\theta}^d$ . Following previous work on both multi-task learning [ARGYRIOU and collab., 2007; RUDER, 2017] and hierarchical supervision [GARCIA and collab., 2019; SANH and collab., 2019], we argue that optimising simultaneously at both levels rather than separately improves the quality of the resulting embeddings. Thus, we write our global hierarchical loss as:

$$\mathcal{L}(\theta) = \lambda_u * \mathcal{L}^u(\theta) + \lambda_d * \mathcal{L}^d(\theta) \quad (5.5)$$



(a) Initial context composed by 5 utterances.



(b)  $u_1$  is chosen to be masked.



(c) Corrupted context with utterance  $u_1$  masked.



(d)  $u_4$  is chosen to be masked.



(e) Corrupted context with utterance  $u_4$  masked.

Figure 5.2 – This figure shows an example of corrupted context. Here  $p_C$  is randomly set to 2 meaning that two utterances will be corrupted.  $u_1$  and  $u_4$  are randomly picked in 5.2b, 5.2d and then masked in 5.2c, 5.2e.

where  $\mathcal{L}^u(\theta)$  is either the MLM or GAP loss at the utterance level and  $\mathcal{L}^d(\theta)$  is its generalisation at the dialog level.

### MLM Loss

The MLM loss at the utterance level is defined in Equation 5.1. Our generalisation at the dialog level masks a proportion  $p_{\mathcal{C}}$  of utterances and generates the sequences of masked tokens. Thus, at the dialog level the MLM loss is defined as:

$$\mathcal{L}_{\text{MLM}}^d(\theta, C_k) = \mathbb{E} \left[ \sum_{j \in m^{C_k}} \sum_{i=1}^{|u_j|} \log(p_{\theta}(w_i^j | \tilde{C}_k)) \right] \quad (5.6)$$

where  $m_j^{C_k} \sim \text{unif}\{1, |C_k|\} \forall j \in [1, p_{\mathcal{C}}]$  is the set of positions of masked utterances in the context  $C_k$ ,  $\tilde{C}_k$  is the corrupted context, and  $p_{\mathcal{C}}$  is the proportion of masked utterances. We propose a visual illustration of the MLM Loss with corrupted context in Figure 5.2.

### GAP Loss

The GAP loss at the utterance level is defined in Equation 5.2. A possible generalisation of the GAP at the dialog level is to compute the loss of the generated utterance across

all factorization orders of the context utterances. Formally, the GAP loss is defined at the dialog level as:

$$\mathcal{L}_{\text{GAP}}^d(\theta, C_k) = \mathbb{E} \left[ \mathbb{E}_{\mathbf{z} \sim \mathbb{Z}_T} \left[ \sum_{t=1}^{|C_k|} \sum_{i=1}^{|u_{z_t}|} \log p_{\theta}(\omega_i^{z_t} | C_k^{\mathbf{z}^{<t}}) \right] \right] \quad (5.7)$$

where  $\omega_i^{z_t}$  denotes the first  $i$ -th tokens of the permuted  $t$ -th utterance when permuting the context according to  $\mathbf{z} \in \mathbb{Z}_T$  and  $C_k^{\mathbf{z}^{<t}}$  the first  $t$  utterances of  $C_k$  when permuting the context according to  $\mathbf{z}$ .

### 5.2.4 Architecture

Commonly, The functions  $f_{\theta}^u$  and  $f_{\theta}^d$  are either modelled with recurrent cells [SERBAN and collab., 2015] or Transformer blocks [VASWANI and collab., 2017]. Transformer blocks are more parallelizable, offer shorter paths for the forward and backward signals and require significantly less time to train compared to recurrent layers. To the best of our knowledge this is the first attempt to pre-train a hierarchical encoder based only on transformers<sup>2</sup>.

The structure of the model can be found in Figure 5.1. In order to optimize dialog level losses as described in Equation 5.5, we generate (through  $g_{\theta}^{dec}$ ) the sequence with a Transformer Decoder ( $\mathcal{T}_{dec}$ ). For downstream tasks, the context embedding  $\mathcal{E}_{C_k}$  is fed to a simple MLP (simple classification), or to a CRF/GRU/LSTM (sequential prediction). In the remainder, we will name our hierarchical transformer-based encoder  $\mathcal{HT}$  and the hierarchical RNN-based encoder  $\mathcal{HR}$ . We use  $\theta_y^x$  to refer to the set of model parameters learnt using the pre-training objective  $y$  (either MLM or GAP) at the level  $x$ <sup>3</sup>.

Thus our proposed losses can be seen as minimizing a weighted sum of two losses which are lower bounds on the MI. They are complementary as they considered different views of the sentence.

### 5.2.5 Pre-training Datasets

Pretraining datasets used to pre-train dialog encoders [HAZARIKA and collab., 2019; MEHRI and collab., 2019] are often medium-sized (e.g. Cornell Movie Corpus [DANESCU-NICULESCU-MIZIL and LEE, 2011], Ubuntu [LOWE and collab., 2015], MultiWOz [BUDZIANOWSKI and collab., 2018a]). In our work, we focus on OpenSubtitles [LISON and TIEDEMANN, 2016]<sup>4</sup> because: (1) it contains spoken language, contrarily to the Ubuntu corpus [LOWE and collab., 2015] based on logs; (2) as Wizard of Oz [BUDZIANOWSKI and collab., 2018a] and Cornell Movie Dialog Corpus [DANESCU-NICULESCU-MIZIL and LEE, 2011], it is a multi-party dataset; and (3) OpenSubtitles is an order of magnitude larger than any other spoken language dataset used in previous work. We segment OpenSubtitles by considering the duration of the silence between two consecutive utterances. Two consecutive utterances belong to the same conversation

<sup>2</sup>Although it is possible to relax the fixed size imposed by transformers [DAI and collab., 2019] in this paper we follow COLOMBO and collab. [2020] and fix the context size to 5 and the max utterance length to 50 — these choices are made to work with OpenSubtitles, since the number of available dialogs drops when considering a number of utterances greater than 5.

<sup>3</sup>if  $x = u$  solely utterance level training is used, if  $x = d$  solely dialog level is used and if  $x = u, d$  multi level supervision is used ( $\lambda_u, \lambda_d \in \{0, 1\}$ <sup>2</sup> according to the case.)

<sup>4</sup><http://opus.nlpl.eu/OpenSubtitles-alt-v2018.php>

if the silence is shorter than  $\delta_T$ <sup>5</sup>. Conversations shorter than the context size  $T$  are dropped<sup>6</sup>. After preprocessing, Opensubtitles contains subtitles from 446520 movies or series which represent 54642424 conversations and over 2.3 billion of words.

### 5.2.6 Baseline Encoder

We compare the different methods we presented with two different types of baseline encoders: pre-trained encoders, and hierarchical encoders based on recurrent cells. The latter, achieve current SOTA performance in many sequence labelling tasks [COLOMBO and collab., 2020; LI and collab., 2018a; LIN and collab., 2017].

**Pre-trained Encoder Models.** We use BERT [DEVLIN and collab., 2018] through the Pytorch implementation provided by the Hugging Face transformers library [WOLF and collab., 2019]. The pre-trained model is fed with a concatenation of the utterances. Formally given an input context  $C_k = (u_1, \dots, u_T)$  the concatenation  $[u_1, \dots, u_T]$  is fed to BERT.

**Hierarchical Recurrent Encoders.** In this work we rely on our own implementation of the model based on  $\mathcal{H}\mathcal{R}$ .

A representation for all the baselines can be found in Figure 5.3. For all models, both hidden dimension and embedding dimension is set to 768 to ensure fair comparison with the proposed model. The MLP used for decoding contains 3 layers of sizes (768, 348, 192). We use ReLU AGARAP [2018] to introduce non linearity inside our architecture.

---

<sup>5</sup>We choose  $\delta_T = 6s$

<sup>6</sup>Using pre-training method based on the next utterance proposed by MEHRI and collab. [2019] requires dropping conversation shorter than  $T + 1$  leading to a non-negligible loss in the preprocessing stage.



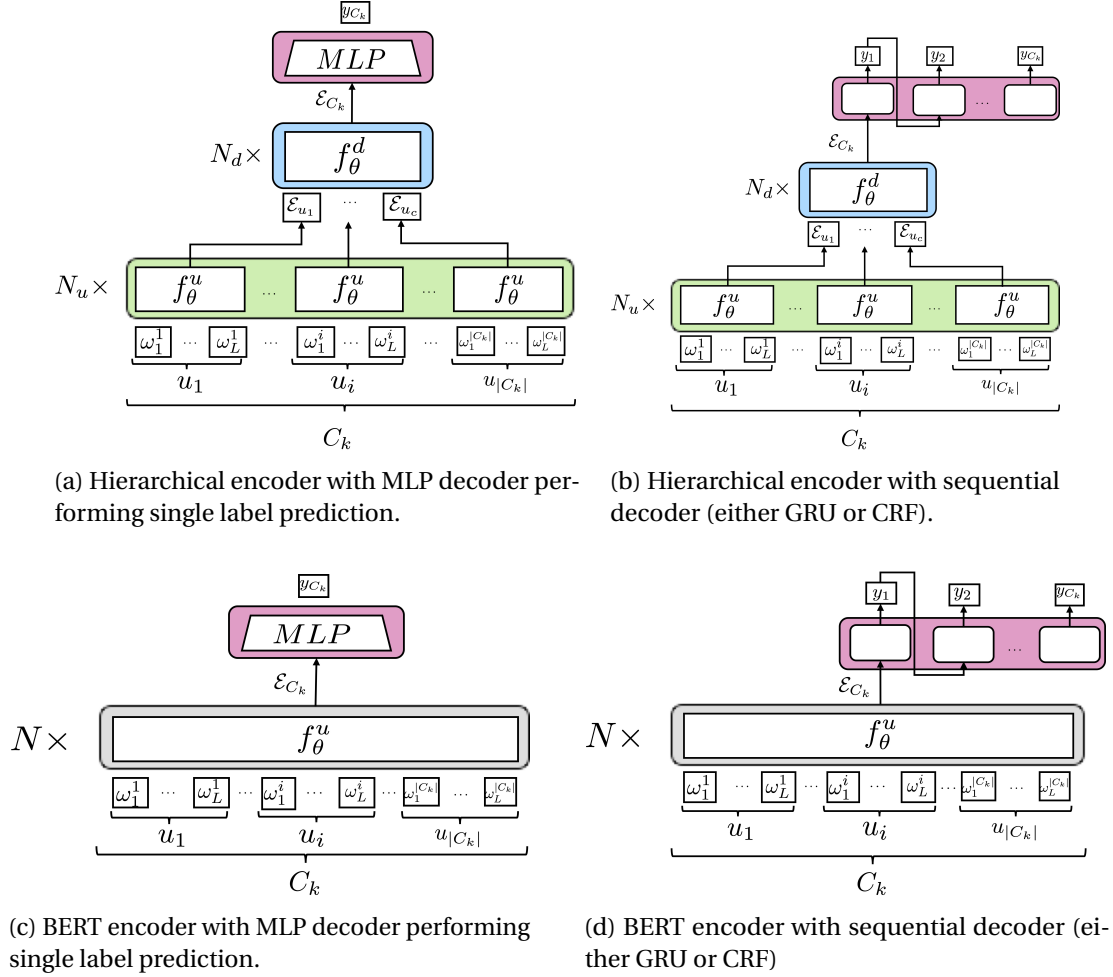


Figure 5.3 – Schema of the different models evaluated on SILICONE. In this figure  $f_\theta^u$ ,  $f_\theta^d$  and the sequence label decoder ( $g_\theta^{dec}$ ) are respectively colored in green, blue and red for the hierarchical encoder (see Figure 5.3a and Figure 5.3d). For BERT there is no hierarchy and embedding is performed through  $f_\theta^u$  colored in grey (see Figure 5.3c, Figure 5.3d)



## 5.3 Information Theoretic Justification of Pretraining Losses

In this section we draw connections between our losses and the MI information framework from KONG and collab. [2019] described in Chapter 3.

**MLM and GAP:** For these losses the parameters used in Equation 3.2 are similar to the one reported in Table 3.2 where the model used for  $g_v$  refers to the first level of hierarchy of our encoders.

**MLM loss at the utterance level:** Consider an input context  $C_k = \{u_1, \dots, u_{|C_k|}\}$  and the corrupted context  $\tilde{C}_k = \{u_1, \dots, \hat{u}_j, \dots, u_{|C_k|}\}$ . Following the notations of Chapter 3, we consider  $f_\theta(x, y)$  where  $x$  is the  $k$ -th masked token of the masked utterance  $\hat{u}_j$  and  $y$  be the masked context  $\tilde{C}_k$ . Let us consider  $g_\psi: \mathcal{V} \rightarrow \mathbb{R}^d$ , where  $\mathcal{V}$  stands for the vocabulary, a lookup function that maps a token index to a vector and  $g_\omega(\tilde{C}_k)$  that returns the final state corresponding to the  $k$ -th masked token of  $\hat{u}_j$ . In this case  $g_\omega$  includes both the hierarchical encoder as well as the transformer decoder. The masking strategy impacts the choice of the joint distribution  $p$ .

**GAP loss at the utterance level:** Following similar steps as done in KONG and collab. [2019], the GAP loss can be casted in the framework described in Chapter 2. The main differences with the MLM loss at the utterance level lie in the choice of both  $g_\omega$  and  $p$ . In this case,  $g_\omega$  is composed of the two level of TXL++ with the transformer decoder. Similar to XLNET,  $p$  is composed of factorization permutations at the sentence level (thus is composed of  $T!$  elements).

## 5.4 Evaluation of Sequence Labelling

### 5.4.1 Related Work

Sequence labelling tasks for spoken dialog mainly involve two different types of labels: DA and E/S. Early work has tackled the sequence labelling problem as an independent classification of each utterance. Deep neural network models that currently achieve the best results [KEIZER and collab., 2002; STOLCKE and collab., 2000; SURENDRAN and LEVOW, 2006] model both contextual dependencies between utterances [COLOMBO and collab., 2020; LI and collab., 2018b] and labels [CHEN and collab., 2018b; KUMAR and collab., 2018; LI and collab., 2018c].

The aforementioned methods require large corpora to train models from scratch, such as: Switchboard Dialog Act (SwDA) [GODFREY and collab., 1992], Meeting Recorder Dialog Act (MRDA) [SHRIBERG and collab., 2004], Daily Dialog Act [LI and collab., 2017], HCRC Map Task Corpus (MT) [THOMPSON and collab., 1993]. This makes harder their adoption to smaller datasets, such as: Loqui human-human dialogue corpus (Loqui) [PASSONNEAU and SACHAR., 2014], BT Oasis Corpus (Oasis) [LEECH and WEISSER, 2003], Multimodal Multi-Party Dataset (MELD) [PORIA and collab., 2018a], Interactive emotional dyadic motion capture database (IEMO), SEMAINE database (SEM) [MCKEOWN and collab., 2013].

### 5.4.2 Presentation of SILICONE

Despite the similarity between methods usually employed to tackle DA and E/S sequential classification, studies usually rely on a single type of label. Moreover, despite

the variety of small or medium-sized labelled datasets, evaluation is usually done on the largest available corpora (*e.g.*, SwDA, MRDA). We introduce SILICONE, a collection of sequence labelling tasks, gathering both DA and E/S annotated datasets. SILICONE is built upon preexisting datasets which have been considered by the community as challenging and interesting. Any model that is able to process multiple sequences as inputs and predict the corresponding labels can be evaluated on SILICONE. We especially include small-sized datasets, as we believe it will ensure that well-performing models are able to both distil substantial knowledge and adapt to different sets of labels without relying on a large number of examples. The description of the datasets composing the benchmark can be found in the following sections, while corpora statistics are gathered in Table 5.2.

### DA Datasets

**Switchboard Dialog Act Corpus** (SwDA) is a telephone speech corpus consisting of two-sided telephone conversations with provided topics. This dataset includes additional features such as speaker id and topic information. The SOTA model, based on a seq2seq architecture with guided attention, reports an accuracy of 85.5% [COLOMBO and collab., 2020] on the official split.

**ICSI MRDA Corpus** (MRDA) has been introduced by SHRIBERG and collab. [2004]. It contains transcripts of multi-party meetings hand-annotated with DA. It is the second biggest dataset with around 110k utterances. The SOTA model reaches an accuracy of 92.2% [LI and collab., 2018a] and uses Bi-LSTMs with attention as encoder as well as additional features, such as the topic of the transcript.

**DailyDialog Act Corpus** (DyDA<sub>a</sub>) has been produced by LI and collab. [2017]. It contains multi-turn dialogues, supposed to reflect daily communication by covering topics about daily life. The dataset is manually labelled with dialog act and emotions. It is the third biggest corpus of SILICONE with 102k utterances. The SOTA model reports an accuracy of 88.1% [LI and collab., 2018a], using Bi-LSTMs with attention as well as additional features. We follow the official split introduced by the authors.

**HCRC MapTask Corpus** (MT) has been introduced by THOMPSON and collab. [1993]. To build this corpus, participants were asked to collaborate verbally by describing a route from a first participant’s map by using the map of another participant. This corpus is small (27k utterances). As there is no standard train/dev/test split<sup>7</sup> performances depends on the split. TRAN and collab. [2017] make use of a Hierarchical LSTM encoder with a GRU decoder layer and achieves an accuracy of 65.9%.

**Bt Oasis Corpus** (Oasis) contains the transcripts of live calls made to the BT and operator services. This corpus has been introduced by LEECH and WEISSER [2003] and is rather small (15k utterances). There is no standard train/dev/test split<sup>8</sup> and few studies use this dataset.

### S/E Datasets

In S/E recognition for spoken language, there is no consensus on the choice the evaluation metric (*e.g.*, GHOSAL and collab. [2019]; PORIA and collab. [2018b] use a weighted F-score while ZHANG and collab. [2019b] report accuracy). For SILICONE, we choose to stay consistent with the DA research and thus follow ZHANG and collab.

<sup>7</sup>We split according to the code in <https://github.com/NathanDuran/Maptask-Corpus>.

<sup>8</sup>We use a random split from <https://github.com/NathanDuran/BT-Oasis-Corpus>.

Corpus	Train	Val	Test	Utt.	Labels	Task	Utt./ Labels
SwDA*	1k	100	11	200k	42	DA	4.8k
MRDA*	56	6	12	110k	5	DA	2.6k
DyDA <sub>a</sub>	11k	1k	1k	102k	4	DA	25.5k
MT*	121	22	25	36k	12	DA	3k
Oasis*	508	64	64	15k	42	DA	357
DyDA <sub>e</sub>	11k	1k	1k	102k	7	E	2.2k
MELD <sub>s</sub> *	934	104	280	13k	3	S	4.3k
MELD <sub>e</sub> *	934	104	280	13k	7	S	1.8k
IEMO	108	12	31	10k	6	E	1.7k
SEM	62	7	10	5,6k	3	S	1.9k

Table 5.2 – Statistics of datasets composing SILICONE. E stands for emotion label and S for sentiment label; \* stands for datasets with available official split. Sizes of Train, Val and Test are given in number of conversations.

[2019b] by reporting the accuracy. Additionally, emotion/sentiment labels are neither merged nor pre-processed<sup>9</sup>.

**DailyDialog Emotion Corpus** (DyDA<sub>e</sub>) has been previously introduced and contains eleven emotional labels. The SOTA model [DE BRUYNE and collab., 2019] is based on BERT with additional Valence Arousal and Dominance features and reaches an accuracy of 85% on the official split.

**Multimodal EmotionLines Dataset** (MELD) has been created by enhancing and extending EmotionLines dataset [CHEN and collab., 2018a] where multiple speakers participated in the dialogues. There are two types of annotations MELD<sub>s</sub> and MELD<sub>e</sub>: three sentiments (positive, negative and neutral) and seven emotions (anger, disgust, fear, joy, neutral, sadness and surprise). The SOTA model with text only is proposed by ZHANG and collab. [2019b] and is inspired by quantum physics. On the official split, it is compared with a hierarchical bi-LSTM, which it beats with an accuracy of 61.9% (MELD<sub>s</sub>) and 67.9% (MELD<sub>e</sub>) against 60.8% and 65.2.

**IEMOCAP database** (IEMO) is a multimodal database of ten speakers. It consists of dyadic sessions where actors perform improvisations or scripted scenarios. Emotion categories are: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other. There is no official split on this dataset. Previous SOTA model is built with bi-LSTMs and achieves 35.1%, with text only [ZHANG and collab., 2019b].

**SEMAINE database** (SEM) comes from the Sustained Emotionally coloured Machine human Interaction using Nonverbal Expression project [MCKEOWN and collab., 2013]. This dataset has been annotated on three sentiments labels: positive, negative and neutral by BARRIERE and collab. [2018]. It is built on a Multimodal Wizard of Oz experiment where participants held conversations with an operator who adopted various roles designed to evoke emotional reactions. There is no official split on this dataset.

### Diversity of SILICONE

We illustrate the diversity of the dataset composing SILICONE. In Figure 5.4, we plot two histograms representing the different utterance lengths for DA and E/S. As expected, for spoken dialog, lengths are shorter than for written benchmarks (e.g., GLUE).

<sup>9</sup>Comparison with concurrent work is more difficult as system performance heavily depends on the number of classes and label processing varies across studies CLAVEL and CALLEJAS [2015].

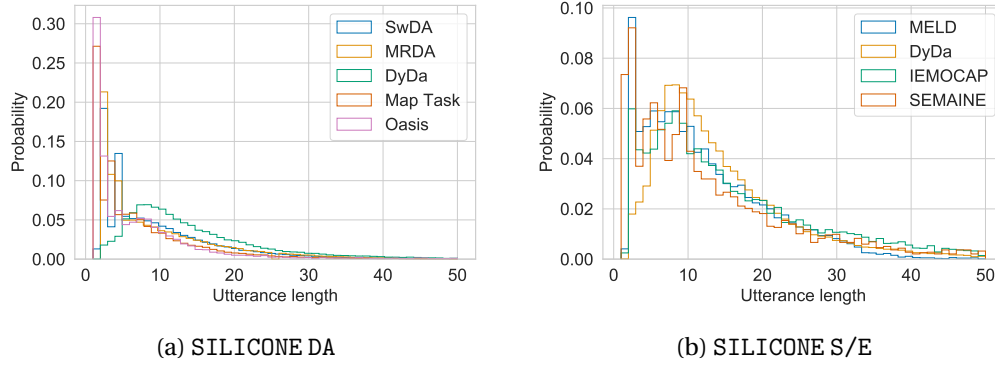


Figure 5.4 – Histograms showing the utterance length for each dataset of SILICONE.

## 5.5 Results on SILICONE

This section describes experiments performed on the SILICONE benchmark. We first provide the hyperparameters and then analyze an appropriate choice for the decoder. The decoder is selected over a set of experiments on our baseline encoders: a pre-trained BERT model and a hierarchical RNN-based encoder ( $\mathcal{HR}$ ). Since we focus on small-sized pre-trained representations, we limit the sizes of our pre-trained models to TINY and SMALL (see Table 5.8). We then study the results of the baselines and our hierarchical transformer encoders ( $\mathcal{HT}$ ) on SILICONE along three axes: the accuracy of the models, the difference in performance between the E/S and the DA corpora, and the importance of pre-training. As we aim to obtain robust representations, we do not perform an exhaustive grid search on the downstream tasks.

### 5.5.1 Parameter choices

For all models, we use a batch size of 64 and automatically select the best model on the validation set according to its loss. We do not perform exhaustive grid search either on the learning rate (that is set to  $10^{-4}$ ), nor on other hyper-parameters to perform a fair comparison between all the models. We use ADAMW [KINGMA and BA, 2014; LOSHCHILOV and HUTTER, 2017] with a linear scheduler on the learning rate and the number of warm-up steps is set to 100. For all models we use a tokenizer based on WordPiece [WU and collab., 2016]. We used GELU [HENDRYCKS and GIMPEL, 2016] activations and the dropout rate [SRIVASTAVA and collab., 2014] is set to 0.1.

### 5.5.2 Decoder Choice

Current research efforts focus on single label prediction, as it seems to be a natural choice for sequence labelling problems (subsection 5.2.1). Sequence labelling is usually performed with CRFs [CHEN and collab., 2018b; KUMAR and collab., 2018] and GRU decoding [COLOMBO and collab., 2020], however, it is not clear to what extent inter-label dependencies are already captured by the contextualised encoders, and whether a plain MLP decoder could achieve competitive results. As can be seen in Table 5.5, we found that in the case of E/S prediction there is no clear difference between CRFs and MLPs, while GRU decoders exhibit poor performance, probably due to a lack of training data. It is also important to notice, that training a sequential

	TINY	SMALL
Nbs of heads	1	6
$N_d$	2	4
$N_u$	2	4
T	50	50
C	5	5
$\mathcal{T}_d$ nbs of heads	6	6
Inner dimension	768	768
Model Dimension	768	768
Vocab length	32000	32000
$\mathcal{T}_d$ : Emb. size	768	768
$d_k$ :	64	64
$d_v$ :	64	64

Table 5.3 – Architecture hyperparameters used for the hierarchical pre-training.

	Avg	SwDA	MRDA	DyDA <sub>DA</sub>	MT	Oasis	DyDA <sub>E</sub>	MELD <sub>S</sub>	MELD <sub>E</sub>	IEMO	SEM
BERT-4layers	70.4	77.8	90.7	79.0	88.4	66.8	90.3	55.3	53.4	43.0	58.8
BERT	72.8	79.2	90.7	<b>82.6</b>	88.2	66.9	91.9	59.3	<b>61.4</b>	<b>45.0</b>	62.7
$\mathcal{H}\mathcal{R}$	69.8	77.5	90.9	80.1	82.8	64.3	91.5	59.3	59.9	40.3	51.1
$\mathcal{H}\mathcal{T}(\theta_{\text{MLM}}^{u,d})_{\text{(TINY)}}$	73.3	<b>79.3</b>	92.0	80.1	90.0	68.3	92.5	62.6	59.9	42.0	66.6
$\mathcal{H}\mathcal{T}(\theta_{\text{GAP}}^d)_{\text{(TINY)}}$	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{H}\mathcal{T}(\theta_{\text{MLM}}^{u,d})_{\text{(SMALL)}}$	<b>74.3</b>	79.2	<b>92.4</b>	81.5	<b>90.6</b>	<b>69.4</b>	<b>92.7</b>	<b>64.1</b>	60.1	<b>45.0</b>	<b>68.2</b>

Table 5.4 – Performances of different encoders when decoding using a MLP on SILICONE. The datasets are grouped by label type (DA vs E/S) and ordered by decreasing size. MT stands for Map Task, IEM for IEMOCAP and Sem for Semaine.

	Avg	Avg DA	Avg E/S
BERT (+MLP)	72.8	81.5	64.0
BERT (+GRU)	69.9	80.4	59.3
BERT (+CRF)	72.8	81.5	64.1
$\mathcal{HR}$ (+MLP)	69.8	79.1	60.4
$\mathcal{HR}$ (+GRU)	67.6	79.4	55.7
$\mathcal{HR}$ (+CRF)	70.5	80.3	60.7

Table 5.5 – Experiments comparing decoder performances. Results are given on SILICONE for two types of baseline encoders (pre-trained BERT models and hierarchical recurrent encoders  $\mathcal{HR}$ ).

decoder usually requires thorough hyper-parameter fine-tuning. As our goal is to learn and evaluate general representations that are decoder agnostic, in the following, we will use a plain MLP decoder for all the models compared.

### 5.5.3 General Performance Analysis

Table 5.4 provides an exhaustive comparison of the different encoders over the SILICONE benchmark. As previously discussed, we adopt a plain MLP as a decoder to compare the different encoders. We show that SILICONE covers a set of challenging tasks as the best performing model achieves an average accuracy of 74.3. Moreover, we observe that despite having half the parameters of a BERT model, our proposed model achieves an average result that is 2% higher on the benchmark. SILICONE covers two different sequence labelling tasks: DA and E/S. In Table 5.4 and Table 5.5, we can see that all models exhibit a consistently higher average accuracy (up to 14%) on DA tagging compared to E/S prediction. This performance drop could be explained by the different sizes of the corpora (see Table 5.2). Despite having a larger number of utterances per label ( $u/l$ ), E/S tasks seem generally harder to tackle for the models. For example, on Oasis, where the  $u/l$  is inferior than those of most E/S datasets (MELD<sub>s</sub>, MELD<sub>e</sub>, IEMO and SEM), models consistently achieve better results.

### 5.5.4 Importance of Pre-training for SILICONE

Results reported in Table 5.4 and Table 5.5 show that pre-trained transformer-based encoders achieve consistently higher accuracy on SILICONE, even when they are not explicitly considering the hierarchical structure. This difference can be observed both in small-sized datasets (e.g. MELD and SEM) and in medium/large size datasets (e.g. SwDA and MRDA). To validate the importance of pre-training in a regime of low data, we train different  $\mathcal{HT}$  (with random initialisation) on different portions of SEM and MELD<sub>s</sub>. Results shown in Figure 5.5 illustrate the importance of pre-trained representations.

#### Negative Results on GAP

We briefly describe few ideas we tried to make GAP works at both the utterance and dialog level. We hypothesise that:

- giving the same weight to the utterance level and the dialog level (see Equation 5.3) was responsible of the observed plateau. Different combinations lead to fairly poor improvements.

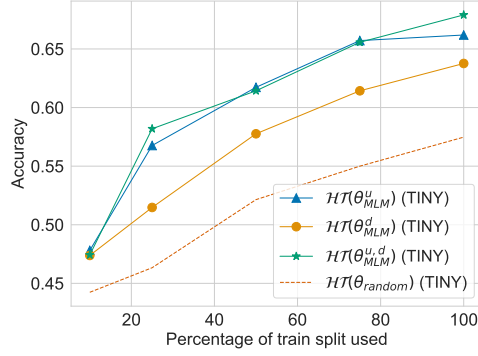


Figure 5.5 – A comparison of pre-trained encoders being fine-tuned on different percentage the training set of SEM. Validation and test set are fixed over all experiments, reported scores are averaged over 10 different random splits.

- the limited model capacity was part of the issue as reported for multilingual pretraining [LIU and collab., 2020].

## 5.6 Model Analysis

In this section, we dissect our hierarchical pre-trained models in order to better understand the relative importance of each component. We show how a hierarchical encoder allows us to obtain a light and efficient model.

### 5.6.1 Pre-training on Spoken vs Written Data

First, we explore the differences in training representations on spoken and written corpora. Experimentally, we compare the predictions on SILICONE made by  $\mathcal{HT}(\theta_{MLM}^u)$  and the one made by  $\mathcal{HT}(\theta_{BERT-2layers})$ . The latter is a hierarchical encoder where utterance embeddings are obtained with the hidden vector representing the first token [CLS] (see DEVLIN and collab. [2018]) of the second layer of BERT. In both cases, predictions are performed using an MLP<sup>10</sup>. Results in Table 5.6 show higher accuracy when the pre-training is performed on spoken data. Since SILICONE is a spoken language benchmark, this result might be due to the specific features of colloquial speech (e.g. disfluencies, sentence length, vocabulary, word frequencies).

<sup>10</sup>We consider the two first layer for a fair comparison based on the number of model parameters. We reproduce the setting of COLOMBO and collab. [2021a, 2020, 2021b,c]; DINKAR and collab. [2020]; GARCIA and collab. [2019]; JALALZAI and collab. [2020b]; STAERMAN and collab. [2021a, 2020, 2021b, 2019, 2021c]; WITON and collab. [2018b]; ?; ?; ?; ?.

	Avg DA	Avg E/S
BERT (4 layers)	80.5	60.2
$\mathcal{HT}(\theta_{BERT-2layers})$	80.5	61.1
$\mathcal{HT}(\theta_{MLM}^u)$	<b>80.8</b>	<b>64.0</b>

Table 5.6 – Results of ablation studies on SILICONE



### 5.6.2 Hierarchy and Multi-Level Supervision

We study the relative importance of three aspects of our hierarchical pre-training with multi-level supervision. We first show that accounting for the hierarchy increases the performance of fine-tuned encoders, even without our specific pre-training procedure. We then compare our two proposed hierarchical pre-training procedures based on the GAP or MLM loss. Lastly, we look at the contribution of the possible levels of supervision on reduced training data from SEM.

#### Importance of hierarchical fine-tuning

We compare the performance of BERT-4layers with the  $\mathcal{H}\mathcal{T}(\theta_{\text{BERT-2layer}})$  previously described. Results reported in Table 5.6 demonstrate that fine-tuning on downstream tasks with a hierarchical encoder yields to higher accuracy, with fewer parameters, even when using already pre-trained representations.

#### MLM vs GAP

In this experiment, we compare the different pre-training objectives at utterance and dialog level. As a reminder  $\mathcal{H}\mathcal{T}(\theta_{\text{MLM}}^u)$  and  $\mathcal{H}\mathcal{T}(\theta_{\text{GAP}}^u)$  are respectively trained using the standard MLM loss [DEVLIN and collab., 2018] and the standard GAP loss [YANG and collab., 2019]. In Table 5.7 we report the different pre-training objective results. We observe that pre-training at the dialog level achieves comparable results to the utterance level pre-training for MLM and slightly worse for GAP. Interestingly, we observe that  $\mathcal{H}\mathcal{T}(\theta_{\text{GAP}}^u)$  compared to  $\mathcal{H}\mathcal{T}(\theta_{\text{MLM}}^u)$  achieves worse results, which is not consistent with the performance observed on other benchmarks, such as GLUE [WANG and collab., 2018]. The lower accuracy of the models trained using a GAP-based loss could be due to several factors (e.g., model size, pre-training using the GAP loss could require a finer choice of hyper-parameters). Finally, we see that supervising at both dialog and utterance level helps for MLM<sup>11</sup>.

#### Multi level Supervision for pre-training

In this section, we illustrate the advantages of learning using several levels of supervision on small datasets. We fine-tune different model on SEM using different size of the training set. Results are shown in Figure 5.5. Overall we see that introducing sequence level supervision induces a consistent improvement on SEM. Results on MELD<sub>s</sub> are provided in Figure 5.7.

### 5.6.3 Improvement over pre-training

In this experiment, we illustrate how pre-training improves performance on SEM (see Figure 5.6). As expected accuracy improves when pre-training.

### 5.6.4 Multi level Supervision for pre-training MELD

In this part, we report results of the experiment mentioned in section 5.6.2. In this experiment we see that the training process seems to be noisier for fractions lower

<sup>11</sup>We investigate a similar setting for GAP which lead to poor results, the loss hit a plateau suggesting that objectives are competing against each other. More advanced optimisations techniques [SENER and KOLTUN, 2018] are left for future work.



	Avg DA	Avg E/S
$\mathcal{HT}(\theta_{\text{MLM}}^u)$	80.8	64.0
$\mathcal{HT}(\theta_{\text{MLM}}^d)$	80.8	64.0
$\mathcal{HT}(\theta_{\text{GAP}}^u)$	80.7	62.0
$\mathcal{HT}(\theta_{\text{GAP}}^d)$	80.4	62.8
$\mathcal{HT}(\theta_{\text{MLM}}^{u,d})$	<b>81.9</b>	<b>64.7</b>

Table 5.7 – Comparison of GAP and MLM with a comparable number of parameters. For all models a MLP decoder is used on top of a TINY pre-trained encoder.

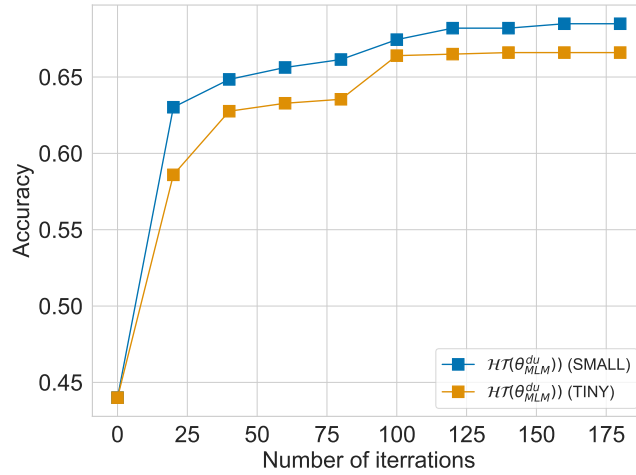


Figure 5.6 – Illustration of improvement of accuracy during pre-training stage on SEM for both a TINY and SMALL models.

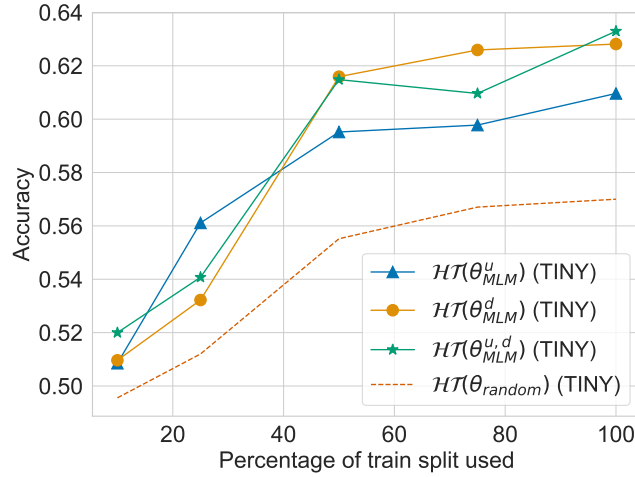


Figure 5.7 – A comparison of different parameters initialisation on MELD<sub>s</sub>. Training is performed using a different percentage of complete training set. Validation and test set are fixed over all experimentation. Each score is the averaged accuracy over 10 random runs.

	Emb.	Word	Seq	Total
BERT			87	110
BERT (4-layer)			43	66
HMLP	23	8.6	7.8	40
(TINY)		2.9	2.8	28.7
(SMALL)		10.6	10.6	45

Table 5.8 – Number of parameters for the encoders. Sizes are given in million of parameters.

than 40%. For larger percentages, we observe that including higher supervision (at the dialog level) during pre-training leads to a consistent improvement.

### 5.6.5 Other advantages of hierarchy

Introducing a hierarchical design in the encoder allows us to break a dialog into utterances and to consider inputs of size  $T$  instead of size 512. First, it allows parameters sharing, reducing the number of model parameters. The different model sizes are reported in Table 5.8. Our TINY model contains half the parameters of BERT (4-layers). Furthermore, modelling long-range dependencies hierarchically makes learning faster and allows to get rid of learning tricks (e.g., partial order prediction [YANG and collab., 2019], two-stage pre-training based on sequence length [DEVLIN and collab., 2018]) required for non-hierarchical encoders. Lastly, original BERT and XLNET are pre-trained using respectively 16 and 512 TPUs. Pre-training lasts several days with over 500K iterations. Our TINY hierarchical models are pre-trained during 180K iterations (1.5 days) on 4 NVIDIA V100.

### Chapter 5 Conclusion

In this chapter, we propose a hierarchical transformer-based encoder that integrate the conversational nature of transcript. We extend two well-known pre-training objectives to adapt them to a hierarchical setting and use OpenSubtitles, the largest spoken language dataset available, for encoder pre-training. Our hierarchical pretraining objectives can be connected to MI and thus are methods can be studies through the lens of the InfoMax principle. Further improvements include extension to multimodal and multi-lingual settings. Additionally, we provide an evaluation benchmark dedicated to comparing sequence labelling systems for the NLP community, SILICONE, on which we compare our models and pre-training procedures with previous approaches. By conducting ablation studies, we demonstrate the importance of using a hierarchical structure for the encoder, both for pre-training and fine-tuning. Finally, we find that our approach is a powerful method to learn generic representations of transcripts, with less parameters than state-of-the-art transformer models.

We hope that the SILICONE benchmark, will encourage further research to build stronger sequence labelling systems for NLP.

## 5.7 References

- AGARAP, A. F. 2018, «Deep learning using rectified linear units (relu)», *arXiv preprint arXiv:1803.08375*. [81](#)
- ARGYRIOU, A., T. EVGENIOU and M. PONTIL. 2007, «Multi-task feature learning», in *Advances in neural information processing systems*, p. 41–48. [78](#)
- BARRIERE, V., C. CLAVEL and S. ESSID. 2018, «Attitude classification in adjacency pairs of a human-agent interaction with hidden conditional random fields», in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 4949–4953. [85](#)
- BUDZIANOWSKI, P., T.-H. WEN, B.-H. TSENG, I. CASANUEVA, U. STEFAN, R. OSMAN and M. GAŠIĆ. 2018a, «Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling», in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [80](#)
- BUDZIANOWSKI, P., T.-H. WEN, B.-H. TSENG, I. CASANUEVA, S. ULTES, O. RAMADAN and M. GAŠIĆ. 2018b, «Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling», *arXiv preprint arXiv:1810.00278*. [76](#)
- CHEN, S.-Y., C.-C. HSU, C.-C. KUO, L.-W. KU and collab.. 2018a, «Emotionlines: An emotion corpus of multi-party conversations», *arXiv preprint arXiv:1802.08379*. [85](#)
- CHEN, Z., R. YANG, Z. ZHAO, D. CAI and X. HE. 2018b, «Dialogue act recognition via crf-attentive structured network», in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 225–234. [78](#), [83](#), [86](#)

- CLAVEL, C. and Z. CALLEJAS. 2015, «Sentiment analysis: from opinion mining to human-agent interaction», *IEEE Transactions on affective computing*, vol. 7, n° 1, p. 74–93. [85](#)
- COLOMBO, P., E. CHAPUIS, M. LABEAU and C. CLAVEL. 2021a, «Improving multimodal fusion via mutual dependency maximisation», *arXiv preprint arXiv:2109.00922*. [89](#)
- COLOMBO, P., E. CHAPUIS, M. MANICA, E. VIGNON, G. VARNI and C. CLAVEL. 2020, «Guiding attention in sequence-to-sequence models for dialogue act prediction», *arXiv preprint arXiv:2002.08801*. [76](#), [80](#), [81](#), [83](#), [84](#), [86](#), [89](#)
- COLOMBO, P., G. STAERMAN, C. CLAVEL and P. PIANTANIDA. 2021b, «Automatic text evaluation through the lens of wasserstein barycenters», *arXiv preprint arXiv:2108.12463*. [89](#)
- COLOMBO, P., W. WITON, A. MODI, J. KENNEDY and M. KAPADIA. 2019, «Affect-driven dialog generation», *arXiv preprint arXiv:1904.02793*. [76](#)
- COLOMBO, P., C. YANG, G. VARNI and C. CLAVEL. 2021c, «Beam search with bidirectional strategies for neural response generation», *arXiv preprint arXiv:2110.03389*. [89](#)
- DAI, Z., Z. YANG, Y. YANG, J. CARBONELL, Q. V. LE and R. SALAKHUTDINOV. 2019, «Transformer-xl: Attentive language models beyond a fixed-length context», *arXiv preprint arXiv:1901.02860*. [80](#)
- DANESCU-NICULESCU-MIZIL, C. and L. LEE. 2011, «Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.», in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*. [76](#), [80](#)
- DE BRUYNE, L., P. ATANASOVA and I. AUGENSTEIN. 2019, «Joint emotion label space modelling for affect lexica», *arXiv preprint arXiv:1911.08782*. [85](#)
- DENOYER, L. and P. GALLINARI. 2006, «The wikipedia xml corpus», in *International Workshop of the Initiative for the Evaluation of XML Retrieval*, Springer, p. 12–19. [75](#)
- DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805*. [75](#), [76](#), [77](#), [81](#), [89](#), [90](#), [92](#)
- DINKAR, T., P. COLOMBO, M. LABEAU and C. CLAVEL. 2020, «The importance of fillers for text representations of speech transcripts», *arXiv preprint arXiv:2009.11340*. [76](#), [89](#)
- GARCIA, A., P. COLOMBO, S. ESSID, F. D’ALCHÉ BUC and C. CLAVEL. 2019, «From the token to the review: A hierarchical multimodal approach to opinion mining», *arXiv preprint arXiv:1908.11216*. [78](#), [89](#)
- GHOSAL, D., N. MAJUMDER, S. PORIA, N. CHHAYA and A. GELBUKH. 2019, «Dialoguecn: A graph convolutional neural network for emotion recognition in conversation», *arXiv preprint arXiv:1908.11540*. [84](#)

- GODFREY, J. J., E. C. HOLLIMAN and J. MCDANIEL. 1992, «Switchboard: Telephone speech corpus for research and development», in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1, ICASSP'92*, IEEE Computer Society, USA, ISBN 0780305329, p. 517–520. [76](#), [83](#)
- HAZARIKA, D., S. PORIA, R. ZIMMERMANN and R. MIHALCEA. 2019, «Emotion recognition in conversations with transfer learning from generative conversation modeling», *arXiv preprint arXiv:1910.04980*. [76](#), [80](#)
- HENDERSON, P., J. HU, J. ROMOFF, E. BRUNSKILL, D. JURAFSKY and J. PINEAU. 2020, «Towards the systematic reporting of the energy and carbon footprints of machine learning», *arXiv preprint arXiv:2002.05651*. [76](#)
- HENDRYCKS, D. and K. GIMPEL. 2016, «Gaussian error linear units (gelus)», *arXiv preprint arXiv:1606.08415*. [86](#)
- JALALZAI, H., P. COLOMBO, C. CLAVEL, E. GAUSSIER, G. VARNI, E. VIGNON and A. SABOURIN. 2020a, «Heavy-tailed representations, text polarity classification & data augmentation», *arXiv preprint arXiv:2003.11593*. [76](#)
- JALALZAI, H., P. COLOMBO, C. CLAVEL, É. GAUSSIER, G. VARNI, E. VIGNON and A. SABOURIN. 2020b, «Heavy-tailed representations, text polarity classification & data augmentation», *arXiv preprint arXiv:2003.11593*. [89](#)
- JIAO, X., Y. YIN, L. SHANG, X. JIANG, X. CHEN, L. LI, F. WANG and Q. LIU. 2019, «Tinybert: Distilling bert for natural language understanding», *arXiv preprint arXiv:1909.10351*. [76](#)
- KEIZER, S., R. OP DEN AKKER and A. NIJHOLT. 2002, «Dialogue act recognition with bayesian networks for dutch dialogues», in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*. [83](#)
- KINGMA, D. P. and J. BA. 2014, «Adam: A method for stochastic optimization», *arXiv preprint arXiv:1412.6980*. [86](#)
- KONG, L., C. D. M. D'AUTUME, W. LING, L. YU, Z. DAI and D. YOGATAMA. 2019, «A mutual information maximization perspective of language representation learning», *arXiv preprint arXiv:1910.08350*. [83](#)
- KUMAR, H., A. AGARWAL, R. DASGUPTA and S. JOSHI. 2018, «Dialogue act sequence labeling using hierarchical encoder with crf», in *Thirty-Second AAAI Conference on Artificial Intelligence*. [83](#), [86](#)
- LAN, Z., M. CHEN, S. GOODMAN, K. GIMPEL, P. SHARMA and R. SORICUT. 2019, «Albert: A lite bert for self-supervised learning of language representations», *arXiv preprint arXiv:1909.11942*. [76](#), [77](#)
- LE, H., L. VIAL, J. FREJ, V. SEGONNE, M. COAVOUX, B. LECOUTEUX, A. ALLAUZEN, B. CRABBÉ, L. BESACIER and D. SCHWAB. 2019, «Flaubert: Unsupervised language model pre-training for french», *arXiv preprint arXiv:1912.05372*. [76](#)
- LEECH, G. and M. WEISSER. 2003, «Generic speech act annotation for task-oriented dialogues.», . [83](#), [84](#)

- LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018a, «A dual-attention hierarchical recurrent neural network for dialogue act classification», *CoRR*, vol. abs/1810.09154. URL <http://arxiv.org/abs/1810.09154>. 78, 81, 84
- LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018b, «A dual-attention hierarchical recurrent neural network for dialogue act classification», *CoRR*. 83
- LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018c, «A dual-attention hierarchical recurrent neural network for dialogue act classification», *arXiv preprint arXiv:1810.09154*. 83
- LI, Y., H. SU, X. SHEN, W. LI, Z. CAO and S. NIU. 2017, «Dailydialog: A manually labelled multi-turn dialogue dataset», . 76, 83, 84
- LIN, Z., M. FENG, C. N. D. SANTOS, M. YU, B. XIANG, B. ZHOU and Y. BENGIO. 2017, «A structured self-attentive sentence embedding», *arXiv preprint arXiv:1703.03130*. 81
- LISON, P. and J. TIEDEMANN. 2016, «Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles», . 80
- LISON, P., J. TIEDEMANN, M. KOUYLEKOV and collab.. 2019, «Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora», in *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA). 76
- LIU, Y., J. GU, N. GOYAL, X. LI, S. EDUNOV, M. GHAZVININEJAD, M. LEWIS and L. ZETTLEMOYER. 2020, «Multilingual denoising pre-training for neural machine translation», *Transactions of the Association for Computational Linguistics*, vol. 8, p. 726–742. 89
- LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER and V. STOYANOV. 2019, «Roberta: A robustly optimized bert pretraining approach», *arXiv preprint arXiv:1907.11692*. 75, 76, 77
- LOSHCHILOV, I. and F. HUTTER. 2017, «Decoupled weight decay regularization», *arXiv preprint arXiv:1711.05101*. 86
- LOWE, R., N. POW, I. SERBAN and J. PINEAU. 2015, «The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems», *CoRR*, vol. abs/1506.08909. URL <http://arxiv.org/abs/1506.08909>. 76, 80
- MCKEOWN, G., M. VALSTAR, R. COWIE, M. PANTIC and M. SCHRODER. 2013, «The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent», *Affective Computing, IEEE Transactions on*, vol. 3, doi: 10.1109/T-AFFC.2011.20, p. 5–17. 83, 85
- MEHRI, S., E. RAZUMOVSAKAIA, T. ZHAO and M. ESKENAZI. 2019, «Pretraining methods for dialog context representation learning», *arXiv preprint arXiv:1906.00414*. 76, 80, 81
- MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. S. CORRADO and J. DEAN. 2013, «Distributed representations of words and phrases and their compositionality», in *Advances in neural information processing systems*, p. 3111–3119. 75



- PASSONNEAU, R. and E. SACHAR. 2014, «Loqui human-human dialogue corpus (transcriptions and annotations)», . 83
- PENNINGTON, J., R. SOCHER and C. D. MANNING. 2014, «Glove: Global vectors for word representation», in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 75
- PETERS, M. E., M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE and L. ZETTMLOYER. 2018, «Deep contextualized word representations», *arXiv preprint arXiv:1802.05365*. 75
- PORIA, S., D. HAZARIKA, N. MAJUMDER, G. NAIK, E. CAMBRIA and R. MIHALCEA. 2018a, «Meld: A multimodal multi-party dataset for emotion recognition in conversations», . 83
- PORIA, S., D. HAZARIKA, N. MAJUMDER, G. NAIK, E. CAMBRIA and R. MIHALCEA. 2018b, «Meld: A multimodal multi-party dataset for emotion recognition in conversations», *arXiv preprint arXiv:1810.02508*. 84
- RUDER, S. 2017, «An overview of multi-task learning in deep neural networks», *arXiv preprint arXiv:1706.05098*. 78
- SANH, V., T. WOLF and S. RUDER. 2019, «A hierarchical multi-task approach for learning embeddings from semantic tasks», in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 6949–6956. 78
- SENER, O. and V. KOLTUN. 2018, «Multi-task learning as multi-objective optimization», in *Advances in Neural Information Processing Systems*, p. 527–538. 90
- SERBAN, I. V., A. SORDONI, Y. BENGIO, A. C. COURVILLE and J. PINEAU. 2015, «Hierarchical neural network generative models for movie dialogues», *CoRR*, vol. abs/1507.04808. URL <http://arxiv.org/abs/1507.04808>. 80
- SHRIBERG, E., R. DHILLON, S. BHAGAT, J. ANG and H. CARVEY. 2004, «The ICSI meeting recorder dialog act (MRDA) corpus», in *Proceedings of the 5th SIG-dial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, Association for Computational Linguistics, Cambridge, Massachusetts, USA, p. 97–100. URL <https://www.aclweb.org/anthology/W04-2319>. 76, 83, 84
- SHRIBERG, E. E. 1999, «Phonetic consequences of speech disfluency», cahier de recherche, SRI INTERNATIONAL MENLO PARK CA. 76
- SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER and R. SALAKHUTDINOV. 2014, «Dropout: a simple way to prevent neural networks from overfitting», *The journal of machine learning research*, vol. 15, n° 1, p. 1929–1958. 86
- STAERMAN, G., P. LAFORGUE, P. MOZHAROVSKIY and F. D’ALCHÉ BUC. 2021a, «When ot meets mom: Robust estimation of wasserstein distance», in *International Conference on Artificial Intelligence and Statistics*, PMLR, p. 136–144. 89
- STAERMAN, G., P. MOZHAROVSKIY, S. CLÉMEN and collab.. 2020, «The area of the convex hull of sampled curves: a robust functional statistical depth measure», in *International Conference on Artificial Intelligence and Statistics*, PMLR, p. 570–579. 89

- STAERMAN, G., P. MOZHAROVSKIY and S. CLÉMENÇON. 2021b, «Affine-invariant integrated rank-weighted depth: Definition, properties and finite sample analysis», *arXiv preprint arXiv:2106.11068*. 89
- STAERMAN, G., P. MOZHAROVSKIY, S. CLÉMENÇON and F. D'ALCHÉ BUC. 2019, «Functional isolation forest», in *Asian Conference on Machine Learning*, PMLR, p. 332–347. 89
- STAERMAN, G., P. MOZHAROVSKIY, P. COLOMBO, S. CLÉMENÇON and F. D'ALCHÉ BUC. 2021c, «A pseudo-metric between probability distributions based on depth-trimmed regions», *arXiv preprint arXiv:2103.12711*. 89
- STOLCKE, A., K. RIES, N. COCCARO, E. SHRIBERG, R. BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. V. ESS-DYKEMA and M. METEER. 2000, «Dialogue act modeling for automatic tagging and recognition of conversational speech», *Computational linguistics*, vol. 26, n° 3, p. 339–373. 83
- STOLCKE, A. and E. SHRIBERG. 1996, «Statistical language modeling for speech disfluencies», in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, IEEE, p. 405–408. 76
- SUÁREZ, P. J. O., B. SAGOT and L. ROMARY. 2019, «Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures», *Challenges in the Management of Large Corpora (CMLC-7) 2019*, p. 9. 75
- SURENDRAN, D. and G.-A. LEVOW. 2006, «Dialog act tagging with support vector machines and hidden markov models», in *Ninth International Conference on Spoken Language Processing*. 83
- THOMPSON, H., A. ANDERSON, E. BARD, G. DOHERTY-SNEDDON, A. NEWLANDS and C. SOTILLO. 1993, «The hcrc map task corpus: natural dialogue for speech recognition», doi: 10.3115/1075671.1075677. 83, 84
- THORNBURY, S. and D. SLADE. 2006, *Conversation: From description to pedagogy*, Cambridge University Press. 76
- TRAN, Q. H., G. HAFFARI and I. ZUKERMAN. 2017, «A generative attentional neural network model for dialogue act classification», in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 524–529. 84
- TRAN, T., J. YUAN, Y. LIU and M. OSTENDORF. 2019, «On the role of style in parsing speech with neural models», *Proc. Interspeech 2019*, p. 4190–4194. 75
- VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER and I. POLOSUKHIN. 2017, «Attention is all you need», in *Advances in neural information processing systems*, p. 5998–6008. 80
- WANG, A., A. SINGH, J. MICHAEL, F. HILL, O. LEVY and S. R. BOWMAN. 2018, «Glue: A multi-task benchmark and analysis platform for natural language understanding», *arXiv preprint arXiv:1804.07461*. 75, 90



- WITON, W., P. COLOMBO, A. MODI and M. KAPADIA. 2018a, «Disney at iest 2018: Predicting emotions using an ensemble», in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, p. 248–253. [76](#)
- WITON, W., P. COLOMBO, A. MODI and M. KAPADIA. 2018b, «Disney at iest 2018: Predicting emotions using an ensemble», in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, p. 248–253. [89](#)
- WOLF, T., L. DEBUT, V. SANH, J. CHAUMOND, C. DELANGUE, A. MOI, P. CISTAC, T. RAULT, R. LOUF, M. FUNTOWICZ and J. BREW. 2019, «Huggingface’s transformers: State-of-the-art natural language processing», *ArXiv*, vol. abs/1910.03771. [81](#)
- WOLF, T., Q. LHOEST, P. VON PLATEN, Y. JERNITE, M. DRAME, J. PLU, J. CHAUMOND, C. DELANGUE, C. MA, A. THAKUR, S. PATIL, J. DAVISON, T. L. SCAO, V. SANH, C. XU, N. PATRY, A. MCMILLAN-MAJOR, S. BRANDEIS, S. GUGGER, F. LAGUNAS, L. DEBUT, M. FUNTOWICZ, A. MOI, S. RUSH, P. SCHMIDD, P. CISTAC, V. MUŠTAR, J. BOUDIER and A. TORDJMAN. 2020, «Datasets», *GitHub. Note: https://github.com/huggingface/datasets*, vol. 1. [75](#)
- WU, Y., M. SCHUSTER, Z. CHEN, Q. V. LE, M. NOROUZI, W. MACHEREY, M. KRIKUN, Y. CAO, Q. GAO, K. MACHEREY and collab.. 2016, «Google’s neural machine translation system: Bridging the gap between human and machine translation», *arXiv preprint arXiv:1609.08144*. [86](#)
- YANG, Z., Z. DAI, Y. YANG, J. CARBONELL, R. R. SALAKHUTDINOV and Q. V. LE. 2019, «Xlnet: Generalized autoregressive pretraining for language understanding», in *Advances in neural information processing systems*, p. 5754–5764. [75](#), [76](#), [77](#), [90](#), [92](#)
- YI, S., R. GOEL, C. KHATRI, A. CERVONE, T. CHUNG, B. HEDAYATNIA, A. VENKATESH, R. GABRIEL and D. HAKKANI-TUR. 2019, «Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators», *arXiv preprint arXiv:1904.13015*. [76](#)
- ZHANG, X., F. WEI and M. ZHOU. 2019a, «Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization», *arXiv preprint arXiv:1905.06566*. [77](#)
- ZHANG, Y., Q. LI, D. SONG, P. ZHANG and P. WANG. 2019b, «Quantum-inspired interactive networks for conversational sentiment analysis.», . [84](#), [85](#)
- ZHU, Y., R. KIROS, R. ZEMEL, R. SALAKHUTDINOV, R. URTASUN, A. TORRALBA and S. FIDLER. 2015, «Aligning books and movies: Towards story-like visual explanations by watching movies and reading books», in *Proceedings of the IEEE international conference on computer vision*, p. 19–27. [75](#)



# Chapter 6

## Including multimodal dimension in representation of spoken transcripts

### Chapter 6 Abstract

Inspired by the success of the mutual information maximization principle, we investigate its application to integrate the multimodal dimension in transcript representations. Multimodal representation learning is a trending area of research, and multimodal fusion is one of its most active topics. Acknowledging humans communicate through a variety of channels (*e.g* visual, acoustic, linguistic), one of the challenges in multimodal systems is to integrating different unimodal representations into a synthetic one. In this chapter, we investigate the use of the measure of information as described in [Chapter 2](#) and propose a set of new objectives that measure the dependency between modalities. We show that these losses are an alternative to complex architectures allowing the fusion of these modalities. We demonstrate that our new penalties are an efficient mean to integrate multi-modality to both randomly initialized and pretrained representations. Our method leads to a consistent improvement (up to 4.3 on accuracy) across a large variety of state-of-the-art models on two well-known sentiment analysis datasets: CMU-MOSI and CMU-MOSEI. Our penalties not only achieves a new SOTA on both datasets but also produce representations that are more robust to modality drops. Finally, a by-product of our methods includes a statistical network that can be used to interpret the high dimensional representations learned by the model.

### 6.1 Introduction

Humans employ three different modalities to communicate in a coordinated manner: the language modality with the use of words and sentences, the vision modality with gestures, poses and facial expressions and the acoustic modality through change in vocal tones. Multimodal representation learning has shown great progress in a large variety of tasks including emotion recognition, sentiment analysis [[SOLEYMANI and collab., 2017](#)], speaker trait analysis [[PARK and collab., 2014](#)] and fine-grained opinion mining [[GARCIA and collab., 2019a](#)]. Keeping in mind our objective of integrating the multimodal dimension in transcript representation we want to learn

from different modalities to obtain better representations and obtain better performance on the target tasks [XU and collab., 2013]. Nevertheless, heterogeneities across modalities increase the difficulty of learning multimodal representations and raise specific challenges. BALTRUŠAITIS and collab. [2018] identifies fusion as one of the five core challenges in multimodal representation learning, the four other being: representation, modality alignment, translation and co-learning. Fusion aims at integrating the different unimodal representations into one common synthetic representation. Effective fusion is still an open problem: the best multimodal models in sentiment analysis [RAHMAN and collab., 2020] improve over their unimodal counterparts, relying on text modality only, by less than 1.5% on accuracy. Additionally, the fusion should not only improve accuracy but also make representations more robust to missing modalities.

Multimodal fusion can be divided into early and late fusion techniques: early fusion takes place at the feature level [YE and collab., 2017], while late fusion takes place at the decision or scoring level [KHAN and collab., 2012]. Current research in multimodal sentiment analysis mainly focuses on developing new fusion mechanisms relying on deep architectures (*e.g.* TFN [ZADEH and collab., 2017], LFN [LIU and collab., 2018], MARN [ZADEH and collab., 2018b], MISA [HAZARIKA and collab., 2020], MCTN [PHAM and collab., 2019], HFNN [MAI and collab., 2019], ICCN [SUN and collab., 2020]). These models are evaluated on several multimodal sentiment analysis benchmark such as IEMOCAP [BUSSO and collab., 2008], MOSI [WÖLLMER and collab., 2013], MOSEI [ZADEH and collab., 2018c] and POM [GARCIA and collab., 2019b; PARK and collab., 2014]. Current state-of-the-art on these datasets uses architectures based on pre-trained transformers [SIRIWARDHANA and collab., 2020; TSAI and collab., 2019] such as MultiModal Bert (MAGBERT) or MultiModal XLNET (MAGXLNET) [RAHMAN and collab., 2020].

The aforementioned architectures are trained by minimising either a  $L_1$  loss or a Cross-Entropy loss between the predictions and the ground-truth labels. To the best of our knowledge, few efforts have been dedicated to exploring alternative losses. In this work, we propose a set of new objectives to perform and improve over existing fusion mechanisms. These improvements are inspired by the InfoMax principle [LINSKER, 1988], *i.e.* choosing the representation maximising the mutual information (MI) between two possibly overlapping views of the input. The MI quantifies the dependence of two random variables; contrarily to correlation, MI also captures non-linear dependencies between the considered variables. Different from previous work, which mainly focuses on comparing two modalities, our learning problem involves multiple modalities (*e.g.* text, audio, video). Our proposed method, which induces no architectural changes, relies on jointly optimising the target loss with an additional penalty term measuring the mutual dependency between different modalities.

### 6.1.1 Our Contributions

We study new objectives to learn multimodal representation of transcripts and obtain more performant and robust multimodal representations through an enhanced fusion mechanism. We evaluate these representations on multimodal sentiment analysis. Our method also allows us to explain the learnt high dimensional multimodal embeddings. The paper contributions can be summarised as follows:

**A set of novel objectives using multivariate dependency measures.** We introduce

three new trainable surrogates to maximise the mutual dependencies between the three modalities (*i.e* audio, language and video). We provide a general algorithm inspired by MINE [BELGHAZI and collab., 2018], which was developed in a bi-variate setting for estimating the MI. Our new method enriches MINE by extending the procedure to a multivariate setting that allows us to maximise different Mutual Dependency Measures: the Total Correlation [WATANABE, 1960], the f-Total Correlation and the Multivariate Wasserstein Dependency Measure [OZAI and collab., 2019].

**Applications and numerical results.** We apply our new set of objectives to five different architectures relying on LSTM cells [HUANG and collab., 2015] (*e.g* EF-LSTM, LFN, MFN) or transformer layers (*e.g* MAGBERT, MAG-XLNET). Our proposed method (1) brings a substantial improvement on two different multimodal sentiment analysis datasets (*i.e* MOSI and MOSEI, subsection 6.5.1), (2) makes the encoder more robust to missing modalities (*i.e* when predicting without language, audio or video the observed performance drop is smaller, subsection 6.5.3), (3) provides an explanation of the decision taken by the neural architecture (subsection 6.5.4).

## 6.2 Problem Formulation & Related Work

In this section, we formulate the problem of learning multi-modal representation (subsection 6.2.1) and we review both existing measures of mutual dependency (see subsection 6.2.2) and estimation methods (subsection 6.2.3). We will focus on learning from three modalities (*i.e* language, audio and video), however our approach can be generalised to any arbitrary number of modalities.

### 6.2.1 Learning multimodal representations

Plethora of neural architectures have been proposed to learn multimodal representations for sentiment classification. Models often rely on a fusion mechanism (*e.g* multi-layer perceptron [KHAN and collab., 2012], tensor factorisation [LIU and collab., 2018; ZADEH and collab., 2019] or complex attention mechanisms [ZADEH and collab., 2018a]) that is fed with modality-specific representations. The fusion problem boils down to learning a model  $\mathcal{M}_f: \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathcal{R}^d$ .  $\mathcal{M}_f$  is fed with uni-modal representations of the inputs  $X_{a,v,l} = (X_a, X_v, X_l)$  obtained through three embedding networks  $f_a$ ,  $f_v$  and  $f_l$ .  $\mathcal{M}_f$  has to retain both modality-specific interactions (*i.e* interactions that involve only one modality) and cross-view interactions (*i.e* more complex, they span across both views). Overall, the learning of  $\mathcal{M}_f$  involves both the minimisation of the downstream task loss and the maximisation of the mutual dependency between the different modalities.

### 6.2.2 Mutual dependency maximisation

**Mutual information as mutual dependency measure:** the core ideas we rely on to better learn cross-view interactions are not new. They consist of mutual information maximisation [LINSKER, 1988], and deep representation learning. Thus, one of the most natural choices is to use the MI that measures the dependence between two random variables, including high-order statistical dependencies [KINNEY and ATWAL, 2014]. Given two random variables  $X$  and  $Y$ , the MI is defined by

$$I(X; Y) \triangleq \mathbb{E}_{XY} \left[ \log \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)} \right], \quad (6.1)$$

where  $p_{XY}$  is the joint probability density function (pdf) of the random variables  $(X, Y)$ , and  $p_X, p_Y$  represent the marginal pdfs. MI can also be defined with a the KL divergence:

$$I(X; Y) \triangleq \text{KL} [p_{XY}(x, y) || p_X(x) p_Y(y)]. \quad (6.2)$$

**Extension of mutual dependency to different metrics:** the KL divergence seems to be limited when used for estimating MI [MCALLESTER and STRATOS, 2020]. A natural step is to replace the KL divergence in Equation 6.2 with different divergences such as the f-divergences or distances such as the Wasserstein distance. Hence, we introduce new mutual dependency measures (MDM): the f-Mutual Information [BELGHAZI and collab., 2018], denoted  $I_f$  and the Wasserstein Measures [OZAI and collab., 2019], denoted  $I_W$ . As previously,  $p_{XY}$  denotes the joint pdf, and  $p_X, p_Y$  denote the marginal pdfs. The new measures are defined as follows:

$$I_f \triangleq \mathcal{D}_f(p_{XY}(x, y); p_X(x) p_Y(y)), \quad (6.3)$$

where  $\mathcal{D}_f$  denotes any  $f$ -divergences and

$$I_W \triangleq \mathcal{W}(p_{XY}(x, y); p_X(x) p_Y(y)), \quad (6.4)$$

where  $\mathcal{W}$  denotes the Wasserstein distance [PEYRÉ and collab., 2019].

### 6.2.3 Estimating mutual dependency measures

The computation of MI and other mutual dependency measures can be difficult without knowing the marginal and joint probability distributions, thus it is popular to maximise lower bounds to obtain better representations of different modalities including image [HJELM and collab., 2018; TIAN and collab., 2019], audio [DILPAZIR and collab., 2016] and text [KONG and collab., 2019] data. Several estimators have been proposed: MINE [BELGHAZI and collab., 2018] uses the Donsker-Varadhan representation [DONSKE and VARADHAN, 1985] to derive a parametric lower bound holds, NGUYEN and collab. [2017, 2010] uses variational characterisation of f-divergence and a multi-sample version of the density ratio (also known as noise contrastive estimation [OORD and collab., 2018; OZAI and collab., 2019]). These methods have mostly been developed and studied in a bi-variate setting.

**Illustration of neural dependency measures on a bivariate case.** In Figure 6.1 we can see the aforementioned dependency measures (*i.e* see Equation 6.2, Equation 6.4, Equation 6.3) when estimated with MINE [BELGHAZI and collab., 2018] for multivariate Gaussian random variables,  $X_a$  and  $X_b$ . The component wise correlation for the considered multivariate Gaussian is defined as follow:  $\text{corr}(X_i, X_k) = \delta_{i,k} \rho$ , where  $\rho \in (-1, 1)$  and  $\delta_{i,k}$  is Kronecker's delta. We observe that the dependency measure based on Wasserstein distance is different from the one based on the divergences and thus will lead to different gradients. Although theoretical studies have been done on the use of different metrics for dependency estimations, it remains an open question to know which one is the best suited. In this work, we will provide an experimental response in a specific case.

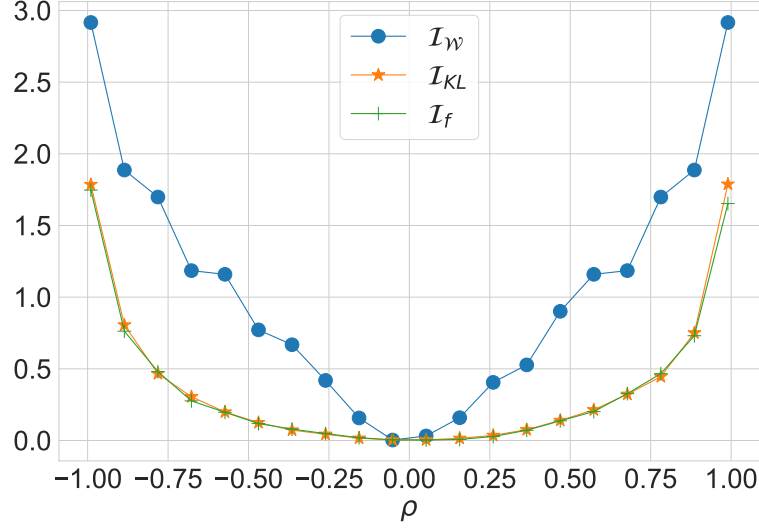


Figure 6.1 – Estimation of different dependency measures for multivariate Gaussian random variables for different degree of correlation.

### 6.3 Model and Training Objective

In this section, we introduce our new set of losses to improve fusion. In [subsection 6.3.1](#), we first extend widely used bi-variate dependency measures to multivariate dependencies measures (MDM) [JAMES and CRUTCHFIELD, 2017]. We then introduce variational bounds on the MDM, and in [subsection 6.3.2](#), we describe our method to minimise the proposed variational bounds.

**Notations** We consider  $X_a, X_v, X_l$  as the multimodal data from the audio, video and language modality with joint probability distribution  $p_{X_a X_v X_l}$ . We denote  $p_{X_j}$  the marginal distribution of  $X_j$  with  $j \in \{a, v, l\}$  corresponding to the  $j$ th modality.

**General loss** As previously mentioned, we rely on the InfoMax principle [LINSKER, 1988] and aim at jointly maximising the MDM between the different modalities and minimising the task loss; hence, we are in a multi-task setting [ARGYRIOU and collab., 2007; RUDER, 2017] and the objective of interest can be defined as:

$$\mathcal{L} \triangleq \underbrace{\mathcal{L}_{down.}}_{\text{downstream task}} - \underbrace{\lambda \cdot \mathcal{L}_{MDM}}_{\text{mutual dependency term}}. \quad (6.5)$$

$\mathcal{L}_{down.}$  represents a downstream specific (target task) loss *i.e* a binary cross-entropy or a  $L_1$  loss,  $\lambda$  is a meta-parameter and  $\mathcal{L}_{MDM}$  is the multivariate dependencies measures (see [subsection 6.3.2](#)). Minimisation of our newly defined objectives requires to derive lower bounds on the  $\mathcal{L}_{MDM}$  terms, and then to obtain trainable surrogates.

#### 6.3.1 From bivariate to multivariate dependencies

In our setting, we aim at maximising cross-view interactions involving three modalities, thus we need to generalise bivariate dependency measures to multivariate dependency measures.

**Definition 6.3.1** (Multivariate Dependencies Measures). *Let  $X_a, X_v, X_l$  be a set of random variables with joint pdf  $p_{X_a X_v X_l}$  and respective marginal pdf  $p_{X_j}$  with  $j \in \{a, v, l\}$ . Then we defined the multivariate mutual information  $I_{kl}$ , also referred as*



total correlation [WATANABE, 1960] or multi-information [STUDENÝ and VEJNAROVÁ, 1998]:

$$\mathbf{I}_{kl} \triangleq \text{KL}(p_{X_a X_v X_l}(x_a, x_v, x_l) \parallel \prod_{j \in \{a, v, l\}} p_{X_j}(x_j)).$$

Similarly for any  $f$ -divergence we define the multivariate  $f$ -mutual information  $\mathbf{I}_f$  as:

$$\mathbf{I}_f \triangleq \mathcal{D}_f(p_{X_a X_v X_l}(x_a, x_v, x_l); \prod_{j \in \{a, v, l\}} p_{X_j}(x_j)).$$

Finally, we also extend Equation 6.3 to obtain the multivariate Wasserstein dependency measure  $\mathbf{I}_{\mathcal{W}}$ :

$$\mathbf{I}_{\mathcal{W}} \triangleq \mathcal{W}(p_{X_a X_v X_l}(x_a, x_v, x_l); \prod_{j \in \{a, v, l\}} p_{X_j}(x_j)).$$

where  $\mathcal{W}$  denotes the Wasserstein distance.

### 6.3.2 From theoretical bounds to trainable surrogates

To train our neural architecture we need to estimate the previously defined multivariate dependency measures. We rely on neural estimators that are given in Theorem 1.

**Theorem 1. Multivariate Neural Dependency Measures** Let the family of functions  $T(\theta) : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}$  parametrized by a deep neural network with learnable parameters  $\theta \in \Theta$ . The multivariate mutual information measure  $\mathbf{I}_{kl}$  is defined as:

$$\mathbf{I}_{kl} \triangleq \sup_{\theta} \mathbb{E}_{p_{X_a X_v X_l}}[T_{\theta}] - \log[\mathbb{E}_{\prod_{j \in \{a, v, l\}} p_{X_j}}[e^{T_{\theta}}]]. \quad (6.6)$$

The neural multivariate  $f$ -mutual information measure  $\mathbf{I}_f$  is defined as follows:

$$\mathbf{I}_f \triangleq \sup_{\theta} \mathbb{E}_{p_{X_a X_v X_l}}[T_{\theta}] - \mathbb{E}_{\prod_{j \in \{a, v, l\}} p_{X_j}}[e^{T_{\theta}^{-1}}]. \quad (6.7)$$

The neural multivariate Wasserstein dependency measure  $\mathbf{I}_{\mathcal{W}}$  is defined as follows:

$$\mathbf{I}_{\mathcal{W}} \triangleq \sup_{\theta: T_{\theta} \in \mathbb{L}} \mathbb{E}_{p_{X_a X_v X_l}}[T_{\theta}] - \log[\mathbb{E}_{\prod_{j \in \{a, v, l\}} p_{X_j}}[T_{\theta}]]. \quad (6.8)$$

Where  $\mathbb{L}$  is the set of all 1-Lipschitz functions from  $\mathcal{R}^d \rightarrow \mathcal{R}$

**Sketch of proofs:** Equation 6.6 is a direct application of the Donsker-Varadhan representation of the KL divergence (we assume that the integrability constraints are satisfied). Equation 6.7 comes from the work of NGUYEN and collab. [2017]. Equation 6.8 comes from the Kantorovich-Rubenstein: we refer the reader to PEYRÉ and collab. [2019]; VILLANI [2008] for a rigorous and exhaustive treatment.

**Practical estimate of the variational bounds.** The empirical estimator that we derive from Theorem 1 can be used in practical way: the expectations in Equation 6.6, Equation 6.7 and Equation 6.8 are estimated using empirical samples from the joint distribution  $p_{X_a X_v X_l}$ . The empirical samples from  $\prod_{j \in \{a, v, l\}} p_{X_j}$  are obtained

by shuffling the samples from the joint distribution in a batch. We integrate this into minimising a multi-task objective (6.5). We refer to the losses obtained with the penalty based on the estimators described in Equation 6.6, Equation 6.7 and Equation 6.8 as  $\mathcal{L}_{kl}$ ,  $\mathcal{L}_f$  and  $\mathcal{L}_{\mathcal{W}}$  respectively. Details on the practical minimisation of our variational bounds are provided in Algorithm 2.



**Algorithm 1** Two-stage procedure to minimise multivariate dependency measures.

---

**INPUT:**  $\mathcal{D}_n = \{(x_a^j, x_v^j, x_l^j), \forall j \in [1, n]\}$  multimodal training dataset,  $m$  batch size,  $\sigma_a, \sigma_v, \sigma_l : [1, m] \rightarrow [1, m]$  three permutations,  $\theta_c$  weights of the deep classifier,  $\theta$  weights of the statistical network  $T_\theta$ .

**Initialization:** parameters  $\theta$  and  $\theta_c$

**Build Negative Dataset:**

$$\tilde{\mathcal{D}}_n = \{(x_a^{\sigma_a(j)}, x_v^{\sigma_v(j)}, x_l^{\sigma_l(j)}), \forall j \in [1, n]\}$$

**Optimization:**

**while**  $(\theta, \theta_c)$  not converged **do**

**for**  $i \in [1, \text{Unroll}]$  **do**

        Sample from  $\mathcal{D}_n$ ,  $\mathcal{B} \sim p_{X_a X_v X_l}$

        Sample from  $\tilde{\mathcal{D}}_n$ ,  $\tilde{\mathcal{B}} \sim \prod_{j \in \{a, v, l\}} p_{X_j}$

        Update  $\theta$  based on the empirical version of Equation 6.6 or Equation 6.7 or Equation 6.8.

**end for**

    Sample a batch  $\mathcal{B}$  from  $\mathcal{D}$

    Update  $\theta_c$  with  $\mathcal{B}$  using Equation 6.5.

**end while**

**OUTPUT:** Classifiers weights  $\theta_c$

---

**Remark.** In this work we choose to generalise MINE to compute multivariate dependencies. Comparing our algorithm to other alternatives mentioned in section 6.2 is left for future work. This choice is driven by two main reasons: (1) our framework allows the use of different types of contrast measures (e.g Wasserstein distance,  $f$ -divergences); (2) the critic network  $T_\theta$  can be used for interpretability purposes as shown in subsection 6.5.4.

As can be seen in Figure 6.2 and in Algorithm 2, to compute the mutual dependency measure the statistic network  $T_\theta$  takes the two embeddings of the different batch  $\mathcal{B}$  and  $\tilde{\mathcal{B}}$ .

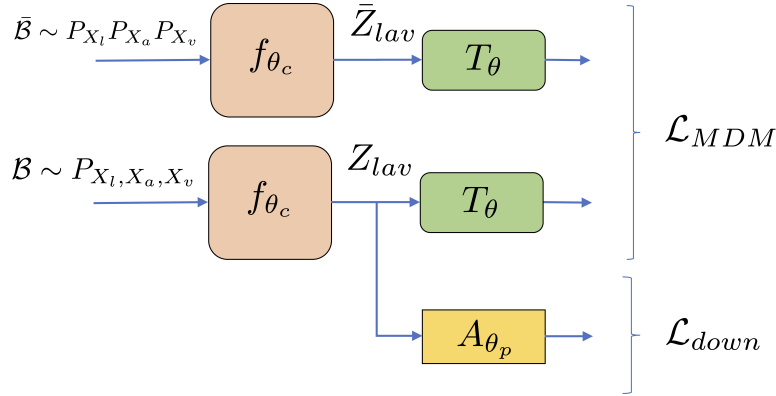


Figure 6.2 – Illustration of the method describes in Algorithm 2 for the different estimators derived from Theorem 1.  $\mathcal{B}$  and  $\bar{\mathcal{B}}$  stands for the batch of data sample from the joint probability distribution and the product of the marginal distribution respectively.  $Z_{lav}$  denotes the fusion representation of linguistic, acoustic and visual (resp.  $l$ ,  $a$  and  $v$ ) modalities provided by a multimodal architecture  $f_{\theta_c}$  for the batch  $\mathcal{B}$ .  $\bar{Z}_{lav}$  denotes the same quantity as described before for the batch  $\bar{\mathcal{B}}$ .  $A_{\theta_p}$  denotes the linear projection before classification or regression.

## 6.4 Experimental Setting

In this section, we present our experimental settings including the neural architectures we compare, the datasets, the metrics and present our methodology.

### 6.4.1 Datasets

We empirically evaluate our methods on two english datasets: CMU-MOSI and CMU-MOSEI. Both datasets have been frequently used to assess model performance in human multimodal sentiment and emotion recognition.

**CMU-MOSI:** Multimodal Opinion Sentiment Intensity [WÖLLMER and collab., 2013] is a sentiment annotated dataset gathering 2,199 short monologue video clips.

**CMU-MOSEI:** CMU-Multimodal Opinion Sentiment and Emotion Intensity [ZADEH and collab., 2018c] is an emotion and sentiment annotated corpus consisting of 23,454 movie review videos taken from YouTube. Both CMU-MOSI and CMU-MOSEI are labelled by humans with a sentiment score in  $[-3, 3]$ .

For each dataset, three modalities are available; we follow prior work [RAHMAN and collab., 2020; ZADEH and collab., 2017, 2018b] and the features that have been obtained as follows<sup>1</sup>:

**Language:** Video transcripts are converted to word embeddings using either Glove [PENNINGTON and collab., 2014], BERT or XLNET contextualised embeddings. For Glove, the embeddings are of dimension 300, where for BERT and XLNET this dimension is 768.

**Vision:** Vision features are extracted using Facet which results into facial action units corresponding to facial muscle movement. For CMU-MOSEI, the video vectors are composed of 47 units, and for CMU-MOSI they are composed of 35.

**Audio :** Audio features are extracted using COVAREP [DEGOTTEX and collab., 2014]. This results into a vector of dimension 74 which includes 12 Mel-frequency cepstral

<sup>1</sup>Data from CMU-MOSI and CMU-MOSEI can be obtained from [https://github.com/WasifurRahman/BERT\\_multimodal\\_transformer](https://github.com/WasifurRahman/BERT_multimodal_transformer)

coefficients (MFCCs), as well as pitch tracking and voiced/unvoiced segmenting features, peak slope parameters, maxima dispersion quotients and glottal source parameters.

Video and audio are aligned on text-based following the convention introduced in CHEN and collab. [2017] and the forced alignment described in YUAN and LIBERMAN [2008].

### 6.4.2 Evaluation metrics

Multimodal Opinion Sentiment Intensity prediction is treated as a regression problem. Thus, we report both the Mean Absolute Error (MAE) and the correlation of model predictions with true labels. In the literature, the regression task is also turned into a binary classification task for polarity prediction. We follow standard practices [RAHMAN and collab., 2020] and report the Accuracy<sup>2</sup> ( $Acc_7$  denotes accuracy on 7 classes and  $Acc_2$  the binary accuracy) of our best performing models.

### 6.4.3 Neural architectures

In our experiments, we choose to modify the loss function of the different models that have been introduced for multi-modal sentiment analysis on both CMU-MOSI and CMU-MOSEI: Memory Fusion Network (MFN [ZADEH and collab., 2018a]), Low-rank Multimodal Fusion (LFN [LIU and collab., 2018]) and two state-of-the-art transformers based models [RAHMAN and collab., 2020] for fusion rely on BERT [DEVLIN and collab., 2018] (MAG-BERT) and XLNET [YANG and collab., 2019] (MAG-XLNT). To assess the validity of the proposed losses, we also apply our method to a simple early fusion LSTM (EF-LSTM) as a baseline model.

**Model overview:** Aforementioned models can be seen as a multi-modal encoder  $f_{\theta_e}$  providing a representation  $Z_{avl}$  containing information and dependencies between modalities  $X_l, X_a, X_v$  namely:

$$f_{\theta_e}(X_a, X_v, X_l) = Z_{avl}.$$

As a final step, a linear transformation  $A_{\theta_p}$  is applied to  $Z_{avl}$  to perform the regression. EF-LSTM: is the most basic architecture used in the current multimodal analysis where each sequence view is encoded separately with LSTM channels. Then, a fusion function is applied to all representations.

TFN: computes a representation of each view, and then applies a fusion operator. Acoustic and visual views are first mean-pooled then encoded through a 2-layers perceptron. Linguistic features are computed with a LSTM channel. Here, the fusion function is a cross-modal product capturing unimodal, bimodal and trimodal interactions across modalities.

MFN enriches the previous EF-LSTM architecture with an attention module that computes a cross-view representation at each time step. They are then gathered and a final representation is computed by a gated multi-view memory [ZADEH and collab., 2018a].

MAG-BERT and MAG-XLNT are based on pre-trained transformer architectures [DEVLIN and collab., 2018; YANG and collab., 2019] allowing inputs on each of the transformer units to be multimodal, thanks to a special gate inspired by WANG and collab. [2018].

---

<sup>2</sup>The regression outputs are turned into categorical values to obtain either 2 or 7 categories (see LIU and collab. [2018]; RAHMAN and collab. [2020]; ZADEH and collab. [2018a])

The  $Z_{avl}$  is the [CLS] representation provided by the last transformer head. For each architecture, we use the optimal architecture hyperparameters provided by the associated papers.

### Hyperparameters selection

We use dropout [SRIVASTAVA and collab., 2014] and optimise the global loss Equation 6.5 by gradient descent using AdamW [KINGMA and BA, 2014; LOSHCHILOV and HUTTER, 2017] optimiser. The best learning rate is found in the grid {0.002, 0.001, 0.0005, 0.0001}. The best model is selected using the lowest MAE on the validation set. We Unroll to 10.

### Architectures of $T_\theta$

Across the different experiment we use a statistic network with an architecture as described in Table 6.1. We follow BELGHAZI and collab. [2018] and use LeakyReLU [AGARAP, 2018; XU and collab., 2015] as activation function.

Statistic Network		
Layer	Number of outputs	Activation function
$[Z_{lav}, \bar{Z}_{lav}]$	$d_{in}, d_{in}$	-
Dense layer	$d_{in}/2$	LeakyReLU
Dropout	0.4	-
Dense layer	$d_{in}$	LeakyReLU
Dropout	0.4	-
Dense layer	$d_{in}$	LeakyReLU
Dropout	0.4	-
Dense layer	$d_{in}/4$	LeakyReLU
Dropout	0.4	-
Dense layer	$d_{in}/4$	LeakyReLU
Dropout	0.4	-
Dense layer	1	Sigmoid

Table 6.1 – Statistics network description.  $d_{in}$  denotes the dimension of  $Z_{avl}$ .

## 6.5 Numerical Results

We present and discuss here the results obtained using the experimental setting described in section 6.4. To better understand the impact of our new methods, we propose to investigate the following points:

**Efficiency of the  $\mathcal{L}_{MDM}$ :** to gain understanding of the usefulness of our new objectives, we study the impact of adding the mutual dependency term on the basic multimodal neural model EF-LSTM.

**Improving model performance and comparing multivariate dependency measures:** the choice of the most suitable dependency measure for a given task is still an open problem (see section 6.3). Thus, we compare the performance – on both multimodal sentiment and emotion prediction tasks – of the different dependency measures. The compared measures are combined with different models using various fusion

mechanisms.

**Improving the robustness to modality drop:** a desirable quality of multimodal representations is the robustness to a missing modality. We study how the maximisation of mutual dependency measures during training affects the robustness of the representation when a modality becomes missing.

**Towards explainable representations:** the statistical network  $T_\theta$  allows us to compute a dependency measure between the three considered modalities. We carry out a qualitative analysis in order to investigate if a high dependency can be explained by complementarity across modalities.

### 6.5.1 Efficiency of the MDM penalty

For a simple EF-LSTM, we study the improvement induced by addition of our MDM penalty. The results are presented in Table 6.2, where a EF-LSTM trained with no mutual dependency term is denoted with  $\mathcal{L}_\emptyset$ . On both studied datasets, we observe that the addition of a MDM penalty leads to stronger performances on all metrics. For both datasets, we observe that the best performing models are obtained by training with an additional mutual dependency measure term. Keeping in mind the example shown in Figure 6.1, we can draw a first comparison between the different dependency measures. Although in a simple case  $\mathcal{L}_f$  and  $\mathcal{L}_{kl}$  estimate a similar quantity (see Figure 6.1), in more complex practical applications they do not achieve the same performance. Even though, the Donsker-Varadhan bound used for  $\mathcal{L}_{kl}$  is stronger<sup>3</sup> than the one used to estimate  $\mathcal{L}_f$ ; for a simple model the stronger bound does not lead to better results. It is possible that most of the differences in performance observed come from the optimisation process during training<sup>4</sup>.

**Takeaways:** *On the simple case of EF-LSTM adding MDM penalty improves the performance on the downstream tasks.*

### 6.5.2 Improving models and comparing multivariate dependency measures

In this experiment, we apply the different penalties to more advanced architectures, using various fusion mechanisms.

**General analysis.** Table 6.3 shows the performance of various neural architectures trained with and without MDM penalty. Results are coherent with the previous experiment: we observe that jointly maximising a mutual dependency measure leads to better results on the downstream task: for example, a MFN on CMU-MOSI trained with  $\mathcal{L}_W$  outperforms by 4.6 points on  $Acc_7^h$  the model trained without the mutual dependency term. On CMU-MOSEI we also obtain subsequent improvements while training with MMD. On CMU-MOSI the TFN also strongly benefits from the mutual dependency term with an absolute improvement of 3.7% (on  $Acc_7^h$ ) with  $\mathcal{L}_W$  compared to  $\mathcal{L}_\emptyset$ . Table 6.3 shows that our methods not only perform well on recurrent architectures but also on pretrained Transformer-based models, that achieve higher results due to a superior capacity to model contextual dependencies (see RAHMAN and collab. [2020]).

**Improving state-of-the-art models.** MAGBERT and MAGXLNET are state-of-the art

<sup>3</sup>For a fixed  $T_\theta$  the right term in Equation 6.6 is greater than Equation 6.7

<sup>4</sup>Similar conclusion have been drawn in the field of metric learning problem when comparing different estimates of the mutual information [BOUDIAF and collab., 2020].

	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$
CMU-MOSI				
$\mathcal{L}_\phi$	31.1	76.1	1.00	0.65
$\mathcal{L}_{kl}$	<u>31.7</u>	<b>76.4</b>	1.00	<b>0.66</b>
$\mathcal{L}_f$	<b>33.7</b>	76.2	1.02	<b>0.66</b>
$\mathcal{L}_W$	<u>33.5</u>	<b>76.4</b>	<b>0.98</b>	<b>0.66</b>
CMU-MOSEI				
$\mathcal{L}_\phi$	44.2	75.0	0.72	0.52
$\mathcal{L}_{kl}$	44.5	<u>75.6</u>	<u>0.70</u>	<u>0.53</u>
$\mathcal{L}_f$	<b>45.5</b>	75.2	<u>0.70</u>	0.52
$\mathcal{L}_W$	<u>45.3</u>	<b>75.9</b>	<b>0.68</b>	<b>0.54</b>

Table 6.2 – Results on sentiment analysis on both CMU-MOSI and CMU-MOSEI for a EF-LSTM.  $Acc_7$  denotes accuracy on 7 classes and  $Acc_2$  the binary accuracy. MAE denotes the Mean Absolute Error and  $Corr$  is the Pearson correlation.  $^h$  means higher is better and  $^l$  means lower is better. The choice of the evaluation metrics follows standard practices [RAHMAN and collab., 2020]. Underline results demonstrate significant improvement (p-value belows 0.05) against the baseline when performing the Wilcoxon Mann Whitney test [WILCOXON, 1992] on 10 runs using different seeds.

models on both CMU-MOSI and CMU-MOSEI. From Table 6.3, we observe that our methods can improve the performance of both models. It is worth noting that, in both cases,  $\mathcal{L}_W$  combined with pre-trained transformers achieves good results. This performance gain suggests that our method is able to capture dependencies that are not learnt during either pretraining of the language model (*i.e* BERT or XLNET) or by the Multimodal Adaptation Gate used to perform the fusion.

**Comparing dependency measures.** Table 6.3 shows that there is no dependency measure that achieves the best results in all cases. This result tends to confirm that the optimisation process during training plays an important role (see hypothesis in subsection 6.5.1). However, we can observe that optimising the multivariate Wasserstein dependency measure is usually a good choice, since it achieves state of the art results in many configurations. It is worth noting that several pieces of research point the limitations of mutual information estimators [MCALLESTER and STRATOS, 2020; SONG and ERMON, 2019].

**Takeaways:** *The addition of MMD not only benefits simple models (e.g EF-LSTM) but also improves performance when combined with both complex fusion mechanisms and pretrained models. For practical applications, the Wasserstein distance is a good choice of contrast function.*

### 6.5.3 Improved robustness to modality drop

Although fusion with visual and acoustic modalities provided a performance improvement [WANG and collab., 2018], the performance of Multimodal systems on sentiment prediction tasks is mainly carried by the linguistic modality [ZADEH and collab., 2017, 2018a]. Thus, it is interesting to study how a multimodal system behaves when the text modality is missing because it gives insights on the robustness of the representation.

**Experiment description.** In this experiment, we focus on the MAGBERT and MAGXL-NET since they are the best performing models. As before, the considered models are trained using the losses described in section 6.3 and all modalities are kept during

	CMU-MOSI				CMU-MOSEI			
	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$	$Acc_7^h$	$Acc_2^h$	$MAE^l$	$Corr^h$
MFN								
$\mathcal{L}_\emptyset$	31.3	76.6	1.01	0.62	44.4	74.7	0.72	0.53
$\mathcal{L}_{kl}$	<u>32.5</u>	76.7	<b>0.96</b>	<b>0.65</b>	44.2	74.7	0.72	<b>0.57</b>
$\mathcal{L}_f$	<u>35.7</u>	<u>77.4</u>	<b>0.96</b>	<b>0.65</b>	<u>46.1</u>	<b>75.4</b>	<b>0.69</b>	<u>0.56</u>
$\mathcal{L}_W$	<b>35.9</b>	<b>77.6</b>	<b>0.96</b>	<b>0.65</b>	<b>46.2</b>	75.1	<b>0.69</b>	<u>0.56</u>
LFN								
$\mathcal{L}_\emptyset$	31.9	76.9	1.00	0.63	45.2	74.2	0.70	0.54
$\mathcal{L}_{kl}$	<u>32.6</u>	<b>77.7</b>	0.97	0.63	<u>46.1</u>	75.3	0.68	<b>0.57</b>
$\mathcal{L}_f$	<b>35.6</b>	77.1	0.97	0.63	45.8	<b>75.4</b>	0.69	<b>0.57</b>
$\mathcal{L}_W$	<b>35.6</b>	<b>77.7</b>	<b>0.96</b>	<b>0.67</b>	<b>46.2</b>	<b>75.4</b>	<b>0.67</b>	<b>0.57</b>
MAGBERT								
$\mathcal{L}_\emptyset$	40.2	84.7	0.79	0.80	46.8	84.9	<b>0.59</b>	0.77
$\mathcal{L}_{kl}$	<b>42.0</b>	<b>85.6</b>	<b>0.76</b>	<b>0.82</b>	47.1	85.4	<b>0.59</b>	<b>0.79</b>
$\mathcal{L}_f$	<u>41.7</u>	<b>85.6</b>	0.78	<b>0.82</b>	46.9	<b>85.6</b>	<b>0.59</b>	<b>0.79</b>
$\mathcal{L}_W$	<u>41.8</u>	85.3	<b>0.76</b>	<b>0.82</b>	<b>47.8</b>	85.5	<b>0.59</b>	<b>0.79</b>
MAGXLNET								
$\mathcal{L}_\emptyset$	43.0	86.2	0.76	<b>0.82</b>	46.7	84.4	<b>0.59</b>	0.79
$\mathcal{L}_{kl}$	<b>44.5</b>	86.1	<b>0.74</b>	<b>0.82</b>	<u>47.5</u>	<u>85.4</u>	<b>0.59</b>	0.81
$\mathcal{L}_f$	<u>43.9</u>	86.6	<b>0.74</b>	<b>0.82</b>	47.4	85.0	<b>0.59</b>	0.81
$\mathcal{L}_W$	<u>44.4</u>	<b>86.9</b>	<b>0.74</b>	<b>0.82</b>	<b>47.9</b>	<b>85.8</b>	<b>0.59</b>	<b>0.82</b>

Table 6.3 – Results on sentiment and emotion prediction on both CMU-MOSI and CMU-MOSEI dataset for the different neural architectures presented in section 6.4 relying on various fusion mechanisms.

Spoken Transcripts	Acoustic and visual behaviour	$T_\theta$
um the story was all right	low energy monotonous voice + headshake	L
i mean its a Nicholas Sparks book it must be good	disappointed tone + neutral facial expression	L
the action is fucking awesome	head nod + excited voice	H
it was cute you know the actors did a great job bringing the smurfs to life such as joe george lopez neil patrick harris katy perry and a fourth	multiple smiles	H

Table 6.4 – Examples from the CMU-MOSI dataset using MAGBERT. The last column is computed using the statistical network  $T_\theta$ . L stands for low values and H stands for high values. Green, grey, red highlight positive, neutral and negative expression/behaviours respectively



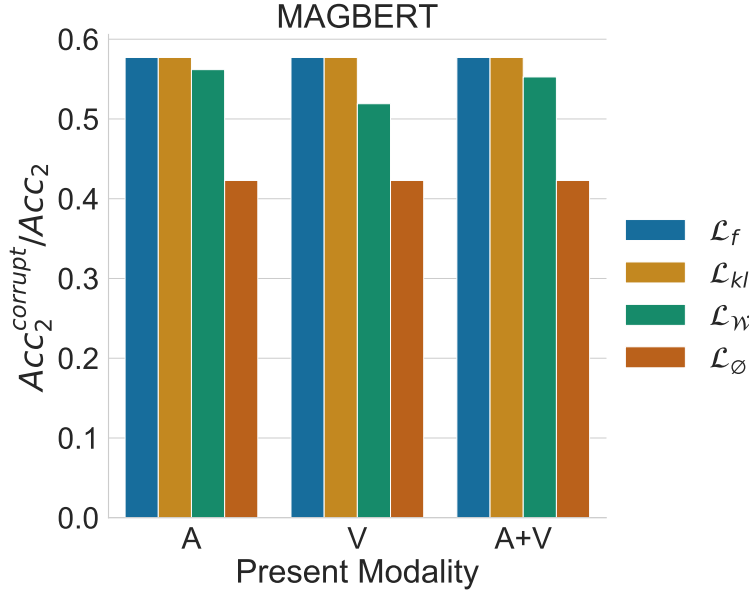


Figure 6.3 – Study of the robustness of the representations against drop of the linguistic modality. Studied model is MAGBERT on CMU-MOSI. The ratio between the accuracy achieved with a corrupted linguistic modality  $Acc_2^{corrupt}$  and the accuracy  $Acc_2$  without any corruption is reported on y-axis. The preserved modalities during inference are reported on x-axis. A, V respectively stands for acoustic and visual modality.

training time. During inference, we either keep only one modality (Audio or Video) or both. Text modality is always dropped.

**Results.** Results with MAG-BERT the experiments conducted on CMU-MOSI are shown in Figure 6.3, giving values for the ratio  $Acc_2^{corrupt} / Acc_2$  where  $Acc_2^{corrupt}$  is the binary accuracy in the corrupted configuration and  $Acc_2$  the accuracy obtained when all modalities are considered. We observe that models trained with an MDM penalty (either  $\mathcal{L}_{kl}$ ,  $\mathcal{L}_f$  or  $\mathcal{L}_w$ ) resist better to missing modalities than those trained with  $\mathcal{L}_\emptyset$ . For example, when trained with  $\mathcal{L}_{kl}$  or  $\mathcal{L}_f$ , the drop in performance is limited to  $\approx 25\%$  in any setting. Interestingly, for MAGBERT  $\mathcal{L}_w$  and  $\mathcal{L}_{KL}$  achieve comparable results;  $\mathcal{L}_{KL}$  is more resistant to dropping the language modality, and thus, could be preferred in practical applications. Figure 6.4 shows the results of the robustness text on MAGXLNET. Similarly to Figure 6.3 we observe more robust representation to modality drop when jointly maximising the  $\mathcal{L}_w$  and  $\mathcal{L}_{kl}$  with the target loss. Figure 6.4 shows no improvement when training with  $\mathcal{L}_f$ . This can also be linked to Table 6.3 which similarly shows no improvement in this very specific configuration.

**Takeaway:** Maximising MMD allows an information transfer between modalities.

#### 6.5.4 Towards explainable representations

In this section, we propose a qualitative experiment allowing us to interpret the predictions made by the deep neural classifier. During training,  $T_\theta$  estimates the mutual dependency measure, using the surrogates introduced in Theorem 1. However, the inference process only involves the classifier, and  $T_\theta$  is unused. Equation 6.6, Equation 6.7, Equation 6.8 show that  $T_\theta$  is trained to discriminate between valid representations (coming from the joint distribution) and corrupted representations (coming from the product of the marginals). Thus,  $T_\theta$  can be used, at inference time, to measure the mutual dependency of the representations used by the neural



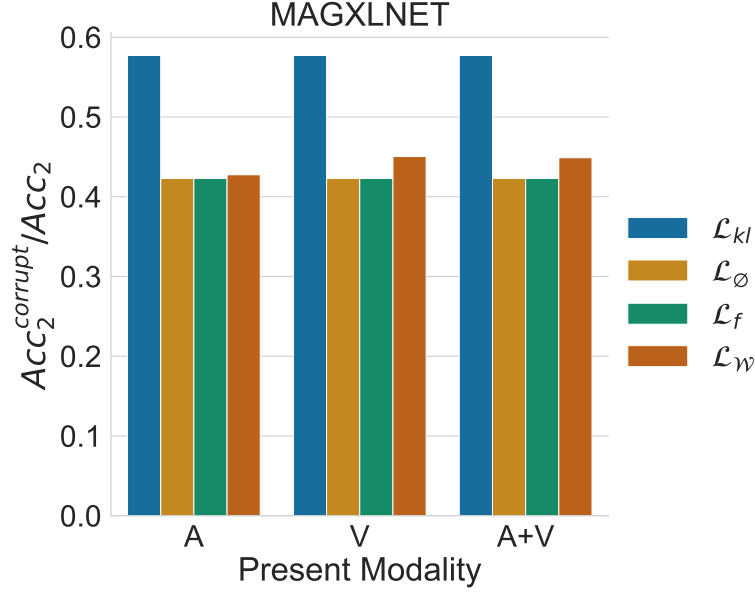


Figure 6.4 – Study of the robustness of the representations against a drop of the linguistic modality. Studied model is MAGXLNET on CMU-MOSI. The ratio between the accuracy achieved with a corrupted linguistic modality  $Acc_2^{corrupt}$  and the accuracy  $Acc_2$  without any corruption is reported on y-axis. The preserved modalities during inference are reported on x-axis. A, V respectively stands for the acoustic and visual modality.

Spoken Transcripts	Acoustic and visual behaviour	$T_{\theta}$
but the m the script is corny	high energy voice + headshake + (many) smiles	L
as for gi joe was it was just like laughing	high energy voice + laughs + smiles	L
its the the plot the the acting is terrible	headshake + long sigh	L
but i think this one did beat scream 2 now	static head + low energy monotonous voice	L
the xxx sequence is really well done	smiles + high energy voice + high pitch	H
you know of course i was waithing for the princess and the frog	smiles + high energy voice	H
dennis quaid i think had a lot of fun	low energy voice + frown eyebrows	H
it was very very very boring	angry voice + angry facial expression	H
i do not wanna see any more of this		H

Table 6.5 – Examples from the CMU-MOSI dataset using MAGXLNET trained with  $\mathcal{L}_W$ . The last column is computed using the statistic network  $T_{\theta}$ . L stands for low values and H stands for high values. Green, grey, red highlight positive, neutral and negative expression/behaviours respectively.

model. In Table 6.4 we report examples of low and high discrepancy measures for MAGBERT on CMU-MOSI. We can observe that high values correspond to video clips where audio, text and video are complementary (e.g use of head node [McCLAVE, 2000]) and low values correspond to the case where there exists contradictions across several modalities.

Table 6.5 reports the results for MAG-XLNET. Similarly to Table 6.1 we observe that high values correspond to complementarity across modalities and low values are related to contradictoriness across modalities.

**Takeaways:**  $T_{\theta}$  used to estimate the MDM provides a mean to interpret representations learnt by the encoder.

### Chapter 6 Conclusion

In this paper, we introduced three new losses based on MDM. We have shown that our new losses are an efficient way to integrate the multi-modality into unimodal transcript representations. Through an extensive set of experiments on CMU-MOSI and CMU-MOSEI, we have shown that SOTA architectures can benefit from our new losses with little modifications. A by-product of our method involves a statistical network that is a useful tool to explain the learned high dimensional multi-modal representations. This work paves the way for using and developing new methods to estimate mutual dependency in a multivariate setting. We believe that the generality of our method allows our method to be combined with the unimodal pretrained representations presented in the previous chapter with minor modifications. One other possible extension of this work would be to adapt the estimator developed in [Chapter 7](#) to improve the MI estimation and improve the fusion loss.

## 6.6 References

- AGARAP, A. F. 2018, «Deep learning using rectified linear units (relu)», *arXiv preprint arXiv:1803.08375*. [110](#)
- ARGYRIOU, A., T. EVGENIOU and M. PONTIL. 2007, «Multi-task feature learning», in *Advances in neural information processing systems*, p. 41–48. [105](#)
- BALTRUŠAITIS, T., C. AHUJA and L.-P. MORENCY. 2018, «Multimodal machine learning: A survey and taxonomy», *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, n° 2, p. 423–443. [102](#)
- BELGHAZI, M. I., A. BARATIN, S. RAJESWAR, S. OZAI, Y. BENGIO, A. COURVILLE and R. D. HJELM. 2018, «Mine: mutual information neural estimation», *arXiv preprint arXiv:1801.04062*. [103](#), [104](#), [110](#)
- BOUDIAF, M., J. RONY, I. M. ZIKO, E. GRANGER, M. PEDERSOLI, P. PIANTANIDA and I. B. AYED. 2020, «A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses», in *European Conference on Computer Vision*, Springer, p. 548–564. [111](#)
- BUSO, C., M. BULUT, C.-C. LEE, A. KAZEMZADEH, E. MOWER, S. KIM, J. N. CHANG, S. LEE and S. S. NARAYANAN. 2008, «Iemocap: Interactive emotional dyadic motion capture database», *Language resources and evaluation*, vol. 42, n° 4, p. 335. [102](#)
- CHEN, M., S. WANG, P. P. LIANG, T. BALTRUŠAITIS, A. ZADEH and L.-P. MORENCY. 2017, «Multimodal sentiment analysis with word-level fusion and reinforcement learning», in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, p. 163–171. [109](#)
- DEGOTTEX, G., J. KANE, T. DRUGMAN, T. RAITIO and S. SCHERER. 2014, «Covarep—a collaborative voice analysis repository for speech technologies», in *2014 IEEE international conference on acoustics, speech and signal processing (icassp)*, IEEE, p. 960–964. [108](#)

- DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805*. [109](#)
- DILPAZIR, H., Z. MUHAMMAD, Q. MINHAS, F. AHMED, H. MALIK and H. MAHMOOD. 2016, «Multivariate mutual information for audio video fusion», *Signal, Image and Video Processing*, vol. 10, n° 7, p. 1265–1272. [104](#)
- DONSKER, M. and S. VARADHAN. 1985, «Large deviations for stationary gaussian processes», *Communications in Mathematical Physics*, vol. 97, n° 1-2, p. 187–210. [104](#)
- GARCIA, A., P. COLOMBO, S. ESSID, F. D’ALCHÉ BUC and C. CLAVEL. 2019a, «From the token to the review: A hierarchical multimodal approach to opinion mining», *arXiv preprint arXiv:1908.11216*. [101](#)
- GARCIA, A., S. ESSID, F. D’ALCHÉ BUC and C. CLAVEL. 2019b, «A multimodal movie review corpus for fine-grained opinion mining», *arXiv preprint arXiv:1902.10102*. [102](#)
- HAZARIKA, D., R. ZIMMERMANN and S. PORIA. 2020, «Misa: Modality-invariant and-specific representations for multimodal sentiment analysis», *arXiv preprint arXiv:2005.03545*. [102](#)
- HJELM, R. D., A. FEDOROV, S. LAVOIE-MARCHILDON, K. GREWAL, P. BACHMAN, A. TRISCHLER and Y. BENGIO. 2018, «Learning deep representations by mutual information estimation and maximization», *arXiv preprint arXiv:1808.06670*. [104](#)
- HUANG, Z., W. XU and K. YU. 2015, «Bidirectional lstm-crf models for sequence tagging», *arXiv preprint arXiv:1508.01991*. [103](#)
- JAMES, R. G. and J. P. CRUTCHFIELD. 2017, «Multivariate dependence beyond shannon information», *Entropy*, vol. 19, n° 10, p. 531. [105](#)
- KHAN, F. S., R. M. ANWER, J. VAN DE WEIJER, A. D. BAGDANOV, M. VANRELL and A. M. LOPEZ. 2012, «Color attributes for object detection», in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, p. 3306–3313. [102](#), [103](#)
- KINGMA, D. P. and J. BA. 2014, «Adam: A method for stochastic optimization», *arXiv preprint arXiv:1412.6980*. [110](#)
- KINNEY, J. B. and G. S. ATWAL. 2014, «Equitability, mutual information, and the maximal information coefficient», *Proceedings of the National Academy of Sciences*, vol. 111, n° 9, p. 3354–3359. [103](#)
- KONG, L., C. D. M. D’AUTUME, W. LING, L. YU, Z. DAI and D. YOGATAMA. 2019, «A mutual information maximization perspective of language representation learning», *arXiv preprint arXiv:1910.08350*. [104](#)
- LINSKER, R. 1988, «Self-organization in a perceptual network», *Computer*, vol. 21, n° 3, p. 105–117. [102](#), [103](#), [105](#)
- LIU, Z., Y. SHEN, V. B. LAKSHMINARASIMHAN, P. P. LIANG, A. ZADEH and L.-P. MORENCY. 2018, «Efficient low-rank multimodal fusion with modality-specific factors», *arXiv preprint arXiv:1806.00064*. [102](#), [103](#), [109](#)

- LOSHCHILOV, I. and F. HUTTER. 2017, «Decoupled weight decay regularization», *arXiv preprint arXiv:1711.05101*. [110](#)
- MAI, S., H. HU and S. XING. 2019, «Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing», in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 481–492. [102](#)
- MCALLESTER, D. and K. STRATOS. 2020, «Formal limitations on the measurement of mutual information», in *International Conference on Artificial Intelligence and Statistics*, p. 875–884. [104](#), [112](#)
- MCCLAIVE, E. Z. 2000, «Linguistic functions of head movements in the context of speech», *Journal of pragmatics*, vol. 32, n° 7, p. 855–878. [115](#)
- NGUYEN, T., T. LE, H. VU and D. PHUNG. 2017, «Dual discriminator generative adversarial nets», in *Advances in Neural Information Processing Systems*, p. 2670–2680. [104](#), [106](#)
- NGUYEN, X., M. J. WAINWRIGHT and M. I. JORDAN. 2010, «Estimating divergence functionals and the likelihood ratio by convex risk minimization», *IEEE Transactions on Information Theory*, vol. 56, n° 11, p. 5847–5861. [104](#)
- OORD, A. V. D., Y. LI and O. VINYALS. 2018, «Representation learning with contrastive predictive coding», *arXiv preprint arXiv:1807.03748*. [104](#)
- OZAI, S., C. LYNCH, Y. BENGIO, A. VAN DEN OORD, S. LEVINE and P. SERMANET. 2019, «Wasserstein dependency measure for representation learning», in *Advances in Neural Information Processing Systems*, p. 15 604–15 614. [103](#), [104](#)
- PARK, S., H. S. SHIM, M. CHATTERJEE, K. SAGAE and L.-P. MORENCY. 2014, «Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach», in *Proceedings of the 16th International Conference on Multimodal Interaction*, p. 50–57. [101](#), [102](#)
- PENNINGTON, J., R. SOCHER and C. D. MANNING. 2014, «Glove: Global vectors for word representation.», in *EMNLP*, vol. 14, p. 1532–1543. [108](#)
- PEYRÉ, G., M. CUTURI and collab.. 2019, «Computational optimal transport: With applications to data science», *Foundations and Trends® in Machine Learning*, vol. 11, n° 5-6, p. 355–607. [104](#), [106](#)
- PHAM, H., P. P. LIANG, T. MANZINI, L.-P. MORENCY and B. PÓCZOS. 2019, «Found in translation: Learning robust joint representations by cyclic translations between modalities», in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 6892–6899. [102](#)
- RAHMAN, W., M. K. HASAN, S. LEE, A. B. ZADEH, C. MAO, L.-P. MORENCY and E. HOQUE. 2020, «Integrating multimodal information in large pretrained transformers», in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2359–2369. [15](#), [102](#), [108](#), [109](#), [111](#), [112](#)
- RUDER, S. 2017, «An overview of multi-task learning in deep neural networks», *arXiv preprint arXiv:1706.05098*. [105](#)

- SIRIWARDHANA, S., A. REIS, R. WEERASEKERA and S. NANAYAKKARA. 2020, «Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition», *arXiv preprint arXiv:2008.06682*. 102
- SOLEYMANI, M., D. GARCIA, B. JOU, B. SCHULLER, S.-F. CHANG and M. PANTIC. 2017, «A survey of multimodal sentiment analysis», *Image and Vision Computing*, vol. 65, p. 3–14. 101
- SONG, J. and S. ERMON. 2019, «Understanding the limitations of variational mutual information estimators», *arXiv preprint arXiv:1910.06222*. 112
- SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER and R. SALAKHUTDINOV. 2014, «Dropout: a simple way to prevent neural networks from overfitting», *The journal of machine learning research*, vol. 15, n° 1, p. 1929–1958. 110
- STUDENÝ, M. and J. VEJNAROVÁ. 1998, «The multiinformation function as a tool for measuring stochastic dependence», in *Learning in graphical models*, Springer, p. 261–297. 106
- SUN, Z., P. SARMA, W. SETHARES and Y. LIANG. 2020, «Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis», in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, p. 8992–8999. 102
- TIAN, Y., D. KRISHNAN and P. ISOLA. 2019, «Contrastive multiview coding», *arXiv preprint arXiv:1906.05849*. 104
- TSAI, Y.-H. H., S. BAI, P. P. LIANG, J. Z. KOLTER, L.-P. MORENCY and R. SALAKHUTDINOV. 2019, «Multimodal transformer for unaligned multimodal language sequences», in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, NIH Public Access, p. 6558. 102
- VILLANI, C. 2008, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media. 106
- WANG, Y., Y. SHEN, Z. LIU, P. P. LIANG, A. ZADEH and L.-P. MORENCY. 2018, «Words can shift: Dynamically adjusting word representations using nonverbal behaviors», . 109, 112
- WATANABE, S. 1960, «Information theoretical analysis of multivariate correlation», *IBM Journal of research and development*, vol. 4, n° 1, p. 66–82. 103, 106
- WILCOXON, F. 1992, «Individual comparisons by ranking methods», in *Breakthroughs in statistics*, Springer, p. 196–202. 15, 112
- WÖLLMER, M., F. WENINGER, T. KNAUP, B. SCHULLER, C. SUN, K. SAGAE and L.-P. MORENCY. 2013, «Youtube movie reviews: Sentiment analysis in an audio-visual context», *IEEE Intelligent Systems*, vol. 28, n° 3, p. 46–53. 102, 108
- XU, B., N. WANG, T. CHEN and M. LI. 2015, «Empirical evaluation of rectified activations in convolutional network», . 110
- XU, C., D. TAO and C. XU. 2013, «A survey on multi-view learning», *arXiv preprint arXiv:1304.5634*. 102

- YANG, Z., Z. DAI, Y. YANG, J. CARBONELL, R. R. SALAKHUTDINOV and Q. V. LE. 2019, «Xlnet: Generalized autoregressive pretraining for language understanding», in *Advances in neural information processing systems*, p. 5753–5763. [109](#)
- YE, J., H. HU, G.-J. QI and K. A. HUA. 2017, «A temporal order modeling approach to human action recognition from multimodal sensor data», *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, n° 2, p. 1–22. [102](#)
- YUAN, J. and M. LIBERMAN. 2008, «Speaker identification on the scotus corpus», *Journal of the Acoustical Society of America*, vol. 123, n° 5, p. 3878. [109](#)
- ZADEH, A., M. CHEN, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2017, «Tensor fusion network for multimodal sentiment analysis», *arXiv preprint arXiv:1707.07250*. [102](#), [108](#), [112](#)
- ZADEH, A., P. P. LIANG, N. MAZUMDER, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2018a, «Memory fusion network for multi-view sequential learning», *arXiv preprint arXiv:1802.00927*. [103](#), [109](#), [112](#)
- ZADEH, A., P. P. LIANG, S. PORIA, P. VIJ, E. CAMBRIA and L.-P. MORENCY. 2018b, «Multi-attention recurrent network for human communication comprehension», in *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2018, NIH Public Access, p. 5642. [102](#), [108](#)
- ZADEH, A., C. MAO, K. SHI, Y. ZHANG, P. P. LIANG, S. PORIA and L.-P. MORENCY. 2019, «Factorized multimodal transformer for multimodal sequential learning», *arXiv preprint arXiv:1911.09826*. [103](#)
- ZADEH, A. B., P. P. LIANG, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2018c, «Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph», in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2236–2246. [102](#), [108](#)

## **Part II Conclusions**

In this part, we showed how to adapt the mutual information principle to learn representations of transcripts by successively integrating the conversational and multimodal dimension of the interaction. In the case of text only ([Chapter 5](#)) we showcase that incorporating the conversational dimension through new hierarchical pretraining objective (that can be connected to MI) not only obtains better results but also to have smaller and cheaper to train models. In [Chapter 6](#), we devise a new method that can be use to integrate the multi-modal dimension both while fine tuning pretrained representations or when training new representations from scratch. Our new losses that takes inspiration in IT through the concept of total correlation allow having more robust representations. The statistic network used for MI estimation can be leveraged to explain the learned representations. In the next part, we move to NLG problems and we address RQ2.





## **Part III**

# **Text Generation using the Measures of Information**



---

### Part III Introduction

This part addresses RQ2 and is dedicated to the application of measures of information to natural text generation. In this part we tackle two tightly linked aspects of NLG:

- In [Chapter 7](#), we use MI to perform style transfer and we introduce a new variational estimator of MI. We show that this estimator leads to both better-disentangled representations and, in particular, allows for a precise control of the desired degree of disentanglement than state-of-the-art methods proposed for textual data. Furthermore, it does not suffer from the degeneracy of other losses in multi-class scenarios. We show the superiority of this method on both fair classification and textual style transfer tasks.
- In [Chapter 8](#), we leverage the flexibility of information measures to tackle the problem of AE. We present InfoLM, a family of untrained metrics which can be adapted to different evaluation criteria. Using direct assessment, we demonstrate that InfoLM achieves statistically significant improvement in many configurations than existing metrics on both summarization and data2text generation.

---

# Chapter 7

## Learning to Disentangle Textual Representations and Attributes via MI

### Chapter 7 Abstract

Learning disentangled representations of textual data is essential for many natural language tasks such as fair classification, style transfer and sentence generation, among others. The dominant approaches in the context of text data either rely on training an adversary (discriminator) that aims at making attribute values difficult to be inferred from the latent code or rely on minimising variational bounds of the mutual information between latent code and the value attribute. However, the available methods suffer of the impossibility to provide fine-grained control of the degree (or force) of disentanglement. In contrast to adversarial methods, which are remarkably simple, although the adversary seems to be performing perfectly well during the training phase after it is completed, a fair amount of information about the undesired attribute still remains. This chapter introduces a novel variational upper bound to the mutual information between an attribute and the latent code of an encoder. Our bound aims at controlling the approximation error via Renyi's divergence, leading to both better-disentangled representations and, in particular, precise control of the desired degree of disentanglement than state-of-the-art methods proposed for textual data. Furthermore, it does not suffer from the degeneracy of other losses in multi-class scenarios. We show the superiority of this method on fair classification and on textual style transfer tasks. Additionally, we provide new insights illustrating various trade-off in style transfer when attempting to learn disentangled representations and the quality of the generated sentence.

### 7.1 Context

Learning disentangled representations hold a central place to build rich embeddings of high-dimensional data. For a representation to be disentangled implies that it factorizes some latent cause or causes of variation as formulated by **BENGIO and collab. [2013]**. For example, if there are two causes for the transformations in the data that do not generally happen together and are statistically distinguishable (e.g.,

factors occur independently), a maximally disentangled representation is expected to present a sparse structure that separates those causes. Disentangled representations have been shown to be useful for a large variety of data, such as video [HSIEH and collab., 2018], image [SANCHEZ and collab., 2019], text [JOHN and collab., 2018], audio [HUNG and collab., 2018], among others, and applied to many different tasks, *e.g.*, robust and fair classification [ELAZAR and GOLDBERG, 2018], visual reasoning [VAN STEENKISTE and collab., 2019], style transfer [FU and collab., 2017], conditional generation [BURGESS and collab., 2018; DENTON and collab., 2017], few shot learning [KUMAR VERMA and collab., 2018], among others.

In this work, we focus our attention on learning disentangled representations for text, as it remains overlooked by JOHN and collab. [2018]. Perhaps, one of the most popular applications of disentanglement in textual data is fair classification [BARRETT and collab., 2019; ELAZAR and GOLDBERG, 2018] and sentence generation tasks such as style transfer [JOHN and collab., 2018] or conditional sentence generation [CHENG and collab., 2020b]. For fair classification, perfectly disentangled latent representations can be used to ensure fairness as the decisions are taken based on representations which are statistically independent from—or at least carrying limited information about—the protected attributes. However, there exists a trade-offs between full disentangled representations and performances on the target task, as shown by FEUTRY and collab. [2018], among others. For sequence generation and in particular, for style transfer, learning disentangled representations aim at allowing an easier transfer of the desired style. To the best of our knowledge, an in-depth study of the relationship between disentangled representations based either on adversarial losses solely or on CLUB and quality of the generated sentences remains overlooked. Most of the previous studies have been focusing on either trade-offs between metrics computed on the generated sentences [TIKHONOV and collab., 2019] or performance evaluation of the disentanglement as part of (or convoluted with) more complex modules. This emphasizes the need to provide a fair evaluation of disentanglement methods by isolating their individual contributions [CHENG and collab., 2020b; YAMSHCHIKOV and collab., 2019].

Methods to enforce disentangled representations can be grouped into two different categories. The first category relies on an adversarial term in the training objective that aims at ensuring that sensitive attribute values (*e.g.* race, sex, style) as statistically independent as possible from the encoded latent representation. Interestingly enough, ELAZAR and GOLDBERG [2018] have recently shown that even though the adversary teacher seems to be performing remarkably well during training, after the training phase, a fair amount of information about the sensitive attributes still remains, and can be extracted from the encoded representation. The second category aim at minimising Mutual Information (MI) between encoded latent representation and the sensitive attribute values, *i.e.*, without resorting to an adversarial discriminator. MI acts as an universal measure of dependence since it captures non-linear and statistical dependencies of high orders between the involved quantities [KINNEY and ATWAL, 2014]. However, estimating MI has been a long-standing challenge, in particular when dealing with high-dimensional data [PANINSKI, 2003; PICHLER and collab., 2020]. Recent methods rely on variational upper bounds. For instance, CHENG and collab. [2020b] study CLUB [CHENG and collab., 2020a] for sentence generation tasks. Although this approach improves on previous state-of-the-art methods, it does not allow to fine-tuning of the desired degree of disentanglement, *i.e.*, it enforces light or strong levels of disentanglement where only few features relevant to the input

sentence remain (see FEUTRY and collab. [2018] for further discussion).

### 7.1.1 Our Contributions

We develop new tools for building disentangled textual representations and evaluate them on fair classification and two sentence generation tasks, namely, style transfer and conditional sentence generation. Our main contributions are summarized below:

- *A novel objective to train disentangled representations from attributes.* To overcome some of the limitations of both adversarial losses and CLUB we derive a novel upper bound to the MI which aims at correcting the approximation error via either the Kullback-Leibler [ALI and SILVEY, 1966] or Renyi [RÉNYI and collab., 1961] divergences. This correction term appears to be a key feature for fine-tuning the degree of disentanglement compared to CLUB.
- *Applications and numerical results.* First, we demonstrate that the aforementioned surrogate is better suited than the widely used adversarial losses as well as CLUB as it can provide better disentangled textual representations while allowing *fine-tuning of the desired degree of disentanglement*. In particular, we show that our method offers a better accuracy versus disentanglement trade-offs for fair classification tasks. We additionally demonstrate that our surrogate outperforms both methods when learning disentangled representations for style transfer and conditional sentence generation while not suffering (or degenerating) when the number of classes is greater than two, which is an apparent limitation of adversarial training. By isolating the disentanglement module, we identify and report existing trade-offs between different degree of disentanglement and quality of generated sentences. The later includes content preservation between input and generated sentences and accuracy on the generated style.

## 7.2 Main Definitions and Related Works

We introduce notations, tasks, and closely related work. Consider a training set  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$  of  $n$  sentences  $x_i \in \mathcal{X}$  paired with attribute values  $y_i \in \mathcal{Y} \equiv \{1, \dots, |\mathcal{Y}|\}$  which indicates a discrete attribute to be disentangled from the resulting representations. We study the following scenarios:

**Disentangled representations.** Learning disentangled representations consists in learning a model  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^d$  that maps feature inputs  $X$  to a vector of dimension  $d$  that retains as much as possible information of the original content from the input sentence but as little as possible about the undesired attribute  $Y$ . In this framework, content is defined as any relevant information present in  $X$  that does not depend on  $Y$ .

**Applications to binary fair classification.** The task of fair classification through disentangled representations aims at building representations that are independent of selective discrete (sensitive) attributes (e.g., gender or race). This task consists in learning a model  $\mathcal{M} : \mathcal{X} \rightarrow \{0, 1\}$  that maps any input  $x$  to a label  $l \in \{0, 1\}$ . The goal of the learner is to build a predictor that assigns each  $x$  to either 0 or 1 “oblivious” of the protected attribute  $y$ . Recently, much progress has been made on devising

appropriate means of fairness, *e.g.*, [MOHRI and collab., 2019; ZAFAR and collab., 2017; ZEMEL and collab., 2013]. In particular, BARRETT and collab. [2019]; ELAZAR and GOLDBERG [2018]; XIE and collab. [2017] approach the problem based on adversarial losses. More precisely, these approaches consist in learning an encoder that maps  $x$  into a representation vector  $h_x$ , a critic  $C_{\theta_c}$  which attempts to predict  $y$ , and an output classifier  $f_{\theta_d}$  used to predict  $l$  based on the observed  $h_x$ . The classifier is said to be fair if there is no statistical information about  $y$  that is present in  $h_x$  [ELAZAR and GOLDBERG, 2018; XIE and collab., 2017].

**Applications to conditional sentence generation.** The task of conditional sentence generation consists in taking an input text containing specific stylistic properties to then generate a realistic (synthetic) text containing potentially different stylistic properties. It requests to learn a model  $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$  that maps a pair of inputs  $(x, y^l)$  to a sentence  $x^g$ , where the outcome sentence should retain as much as possible of the original content from the input sentence while having (potentially a new) attribute  $y^g$ . Proposed approaches to tackle textual style transfer [XU and collab., 2019; ZHANG and collab., 2020] can be divided into two main categories. The first category [LAMPLE and collab., 2018; PRABHUMOYE and collab., 2018] uses cycle losses based on back translation [WIETING and collab., 2017] to ensure that the content is preserved during the transformation. Whereas, the second category look to explicitly separate attributes from the content. This constraint is enforced using either adversarial training [FU and collab., 2017; HU and collab., 2017; YAMSHCHIKOV and collab., 2019; ZHANG and collab., 2018] or MI minimisation using vCLUB-S [CHENG and collab., 2020b]. Traditional adversarial training is based on an encoder that aims to fool the adversary discriminator by removing attribute information from the content embedding [ELAZAR and GOLDBERG, 2018]. As we will observe, the more the representations are disentangled the easier is to transfer the style but at the same time the less the content is preserved. In order to approach the sequence generation tasks, we build on the Style-embedding Model by JOHN and collab. [2018] (StyleEmb) which uses adversarial losses introduced in prior work for these dedicated tasks. During the training phase, the input sentence is fed to a sentence encoder, namely  $f_{\theta_e}$ , while the input style is fed to a separated style encoder, namely  $f_{\theta_e^s}$ . During the inference phase, the desired style—potentially different from the input style—is provided as input along with the input sentence.

## 7.3 Model and Training Objective

This section describes the proposed approach to learn disentangled representations. We first present the model overview and then, we derive the variational bound we will use, and discuss connections with adversarial losses.

### 7.3.1 Model Overview

Our models for fair classification and sequence generation share a similar structure. These rely on an encoder that takes as input a random sentence  $X$  and maps it to a random representation  $Z$  using a deep encoder denoted by  $f_{\theta_e}$ . Then, classification and sentence generation are performed using either a classifier or an auto-regressive decoder denoted by  $f_{\theta_d}$ . We aim at minimizing MI between the latent code represented by the Random Variable (RV)  $Z = f_{\theta_e}(X)$  and the desired attribute represented



by the RV  $Y$ . The objective of interest  $\mathcal{L}(f_{\theta_e})$  is defined as:

$$\mathcal{L}(f_{\theta_e}) \equiv \underbrace{\mathcal{L}_{down.}(f_{\theta_e})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_{\theta_e}(X); Y)}_{\text{disentangled}}, \quad (7.1)$$

where  $\mathcal{L}_{down.}$  represents a downstream specific (target task) loss and  $\lambda$  is a meta-parameter that controls the sensitive trade-off between disentanglement (*i.e.*, minimizing MI) and success in the downstream task (*i.e.*, minimizing the target loss). In [section 7.5](#), we illustrate these different trade-offs.

**Applications to fair classification and sentence generation.** For fair classification, we follow standard practices and optimize the cross-entropy between prediction and ground-truth labels. In the sentence generation task  $\mathcal{L}_{down.}$  represents the negative log-likelihood between individual tokens.

### 7.3.2 A Novel Upper Bound on MI

Estimating the MI is a long-standing challenge as the exact computation [[PANINSKI, 2003](#)] is only tractable for discrete variables, or for a limited family of problems where the underlying data-distribution satisfies smoothing properties, see recent work by [PICHLER and collab., \[2020\]](#). Different from previous approaches leading to variational lower bounds [[BELGHAZI and collab., 2018](#); [HJELM and collab., 2018](#); [OORD and collab., 2018](#)], in this paper we derive an estimator based on a variational upper bound to the MI which control the approximation error based on the Kullback-Leibler and the Renyi divergences [[DAUDEL and collab., 2020](#)].

**Theorem 2.** (*Variational upper bound on MI*) Let  $(Z, Y)$  be an arbitrary pair of RVs with  $(Z, Y) \sim p_{ZY}$  according to some underlying pdf, and let  $q_{\hat{Y}|Z}$  be a conditional variational distribution on the attributes satisfying  $p_{ZY} \ll p_Z \cdot q_{\hat{Y}|Z}$ , *i.e.*, absolutely continuous. Then, we have that

$$\begin{aligned} I(Z; Y) &\leq \mathbb{E}_Y \left[ -\log \int q_{\hat{Y}|Z}(Y|z) p_Z(z) dz \right] + \\ &\mathbb{E}_{YZ} \left[ \log q_{\hat{Y}|Z}(Y|Z) \right] + KL(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}), \end{aligned} \quad (7.2)$$

where  $KL(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z})$  denotes the KL divergence. Similarly, we have that for any  $\alpha > 1$ ,

$$\begin{aligned} I(Z; Y) &\leq \mathbb{E}_Y \left[ -\log \int q_{\hat{Y}|Z}(Y|z) p_Z(z) dz \right] + \\ &\mathbb{E}_{YZ} \left[ \log q_{\hat{Y}|Z}(Y|Z) \right] + D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}), \end{aligned} \quad (7.3)$$

where  $(\alpha - 1)D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}) = \log \mathbb{E}_{ZY} [R^{\alpha-1}(Z, Y)]$  denotes the Renyi divergence and  $R(z, y) = \frac{p_{Y|Z}(y|z)}{q_{\hat{Y}|Z}(y|z)}$ , for  $(z, y) \in \text{Supp}(p_{ZY})$ .

*Proof:* The upper bound on  $H(Y)$  is a direct application of the [DONSKEP and VARADHAN \[1985\]](#) representation of KL divergence while the lower bound on  $H(Y|Z)$  follows from the monotonicity property of the function:  $\alpha \mapsto D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z})$ . All proofs can be found in [section A.1](#).

**Remark:** It is worth to emphasise that the KL divergence in (7.2) and Renyi divergence in (7.3) control the approximation error between the exact entropy and its corresponding bound.

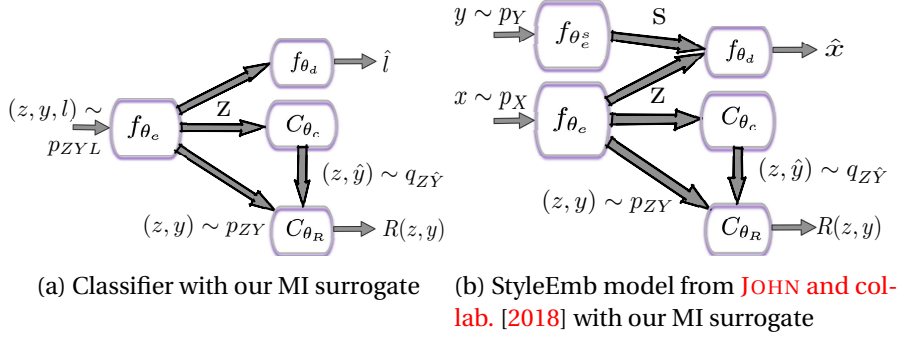


Figure 7.1 – Proposed methods. As described in Theorem 2.

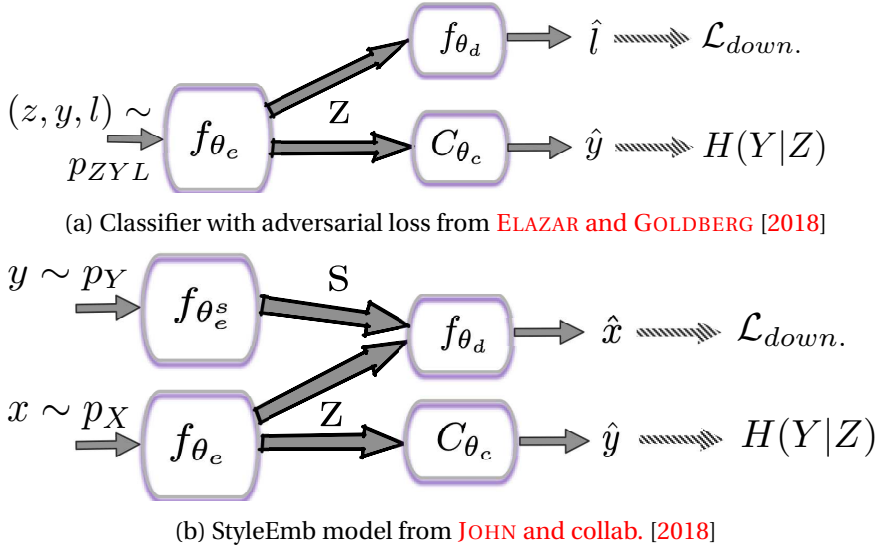


Figure 7.2 – Baselines methods, these models use an adversarial loss for disentanglement.  $f_{\theta_e}$  represents the input sentence encoder;  $f_{\theta_e^s}$  denotes the style encoder (only used for sentence generation tasks);  $C_{\theta_c}$  represents the adversarial classifier;  $f_{\theta_d}$  represents the decoder that can be either a classifier (Figure 7.2a) or a sequence decoder (Figure 7.2b).

**From theoretical bounds to trainable surrogates to minimize MI:** It is easy to check that the inequalities in (Equation 7.2) and (Equation 7.3) are tight provided that  $p_{ZY} \equiv p_Z \cdot q_{\hat{Y}|Z}$  almost surely for some adequate choice of the variational distribution. However, the evaluation of these bounds requires to obtain an estimate of the density-ratio  $R(z, y)$ . Density-ratio estimation has been widely studied in the literature (see SUGIYAMA and collab. [2012] and references therein) and confidence bounds has been reported by KPOTUFE [2017] under some smoothing assumption on underlying data-distribution  $p_{ZY}$ . In this work, we will estimate this ratio by using a critic  $C_{\theta_R}$  which is trained to differentiate between a balanced dataset of positive i.i.d samples coming from  $p_{ZY}$  and negative i.i.d samples coming from  $q_{\hat{Y}|Z} \cdot p_Z$ . Then, for any pair  $(z, y)$ , the density-ratio can be estimated by  $R(z, y) \approx \frac{\sigma(C_{\theta_R}(z, y))}{1 - \sigma(C_{\theta_R}(z, y))}$ , where  $\sigma(\cdot)$  indicates the sigmoid function and  $C_{\theta_R}(z, y)$  is the unnormalized output of the critic. It is worth to mention that after estimating this ratio, the previous upper bounds may not be strict bounds so we will refer them as surrogates.

We report in Figure 7.1 the schema of the proposed approach and the baselines are describes in Figure 7.2.

---

**Algorithm 2** Our method for the fair classification task

---

**INPUT:** training dataset for the encoder  $\mathcal{D}_n = \{(x_1, y_1, l_1), \dots, (x_n, y_n, l_n)\}$ , batch size  $m$ , training dataset for the classifiers and decoder  $\mathcal{D}'_n = \{(x'_1, y'_1, l'_1), \dots, (x'_n, y'_n, l'_n)\}$ .

**Initialization:** parameters  $(\theta_e, \theta_R, \theta_c, \theta_d)$  of the encoder  $f_{\theta_e}$ , classifiers  $C_{\theta_R}, C_{\theta_c}, f_{\theta_d}$

**Optimization:**

```

1: while  $(\theta_e, \theta_R, \theta_c, \theta_d)$  not converged do
2:   for  $i \in [1, Unroll]$  do                                      $\triangleright$  Train  $C_{\theta_c}, C_{\theta_R}, f_{\theta_d}$ 
3:     Sample a batch  $\mathcal{B}'$  from  $\mathcal{D}'$ 
4:     Update  $\theta_R$  based  $\mathcal{B}'$  and using  $C_{\theta_c}$ 
5:     Update  $\theta_c$  with  $\mathcal{B}'$ 
6:     Update  $\theta_d$  with  $\mathcal{B}'$ 
7:   end for
8:   Sample a batch  $\mathcal{B}$  from  $\mathcal{D}$                                       $\triangleright$  Train  $f_{\theta_e}$ 
9:   Update  $\theta_e$  with  $\mathcal{B}$  using Equation 7.1 with  $\theta_d$ .
10: end while

```

**OUTPUT:**  $f_{\theta_e}, f_{\theta_d}$

---

### 7.3.3 Comparison to existing methods

**Adversarial approaches:** In order to enhance our understanding of why the proposed approach based on the minimization of the MI using our variational upper bound in Theorem 2 may lead to a better training objective than previous adversarial losses, we discuss below the explicit relationship between MI and cross-entropy loss. Let  $Y \in \mathcal{Y}$  denote a random attribute and let  $Z$  be a possibly high-dimensional representation that needs to be disentangled from  $Y$ . Then,

$$\begin{aligned} I(Z; Y) &\geq H(Y) - \mathbb{E}_{YZ} \left[ \log q_{\hat{Y}|Z}(Y|Z) \right] \\ &= \text{Const} - \text{CE}(\hat{Y}|Z), \end{aligned} \quad (7.4)$$

where  $\text{CE}(\hat{Y}|Z)$  denotes the cross-entropy corresponding to the adversarial discriminator  $q_{\hat{Y}|Z}$ , noting that  $Y$  comes from an unknown distribution on which we have no influence  $H(Y)$  is an unknown constant, and using that the approximation error:  $\text{KL}(q_{ZY} \| q_{\hat{Y}|Z} \cdot p_Z) = \text{CE}(\hat{Y}|Z) - H(Y|Z)$ . Equation 7.4 shows that the cross-entropy loss leads to a lower bound (up to a constant) on the MI. Although the cross-entropy can lead to good estimates of the conditional entropy, the adversarial approaches for classification and sequence generation by BARRETT and collab. [2019]; JOHN and collab. [2018] which consists in maximizing the cross-entropy, induces a degeneracy (unbounded loss) as  $\lambda$  increases in the underlying optimization problem. As we will observe in next section, our variational upper bound in Theorem 2 can overcome this issue, in particular for  $|\mathcal{Y}| > 2$ .

CLUB: Different from our method, CHENG and collab. [2020a] introduce  $I_{\text{CLUB}}$  which is an upper bound on MI defined by

$$\begin{aligned} I_{\text{CLUB}}(Y; Z) &= \mathbb{E}_{YZ} [\log p_{Y|Z}(Y|Z)] \\ &\quad - \mathbb{E}_Y \mathbb{E}_Z [\log p_{Y|Z}(Y|Z)]. \end{aligned} \quad (7.5)$$

It would be worth to mention that this bound follows a similar approach to the previously introduced bound in FEUTRY and collab. [2018].

## 7.4 Experimental Setting

### 7.4.1 Datasets

**Fair classification task.** We follow the experimental protocol of ELAZAR and GOLDBERG [2018]. The main task consists in predicting a binary label representing either the sentiment (positive/negative) or the mention. The mention task aims at predicting if a tweet is conversational. The considered protected attribute is the race. This dataset has been automatically constructed from DIAL corpus [BLODGETT and collab., 2016] which contained race annotations over 50 Million of tweets. Sentiment tweets are extracted using a list of predefined emojis and mentions are identified using @mentions tokens. The final dataset contains 160k tweets for the training and two splits of 10K tweets for validation and testing. Splits are balanced such that the random estimator is likely to achieve 50% accuracy.

**Style Transfer** For our sentence generation task, we conduct experiments on three different datasets extracted from restaurant reviews in Yelp. The first dataset, referred to as SYelp, contains 444101, 63483, and 126670 labelled short reviews (at most 20 words) for train, validation, and test, respectively. For each review a binary label is assigned depending on its polarity. Following LAMPLE and collab. [2018], we use a second version of Yelp, referred to as FYelp, with longer reviews (at most 70 words). It contains five coarse-grained restaurant category labels (e.g., Asian, American, Mexican, Bars and Dessert). The multi-category FYelp is used to access the generalization capabilities of our methods to a multi-class scenario.

### 7.4.2 Metrics for Performance Evaluation

**Measure of the disentanglement methods.** BARRETT and collab. [2019] report that offline classifiers (post training) outperform clearly adversarial discriminators. We will re-training a classifier on the latent representation learnt by the model and we will report its accuracy. We follow previous work [LAMPLE and collab., 2018] that implements a two layers perceptron [ROSENBLATT, 1958]. We use LeakyRelu [XU and collab., 2015] as activation functions and set the dropout [SRIVASTAVA and collab., 2014] rate to 0.1.

**Measures of performance within the fair classification task.** In the fair classification task we aim at maximizing accuracy on the target task and so we will report the corresponding accuracy.

**Measure of performance within sentence generation tasks.** Sentences generated by the model are expected to be fluent, to preserve the input content and to contain the desired style. For style transfer, the desired style is different from the input style while for conditional sentence generation, both input and output styles should be similar. Nevertheless, automatic evaluation of generative models for text is still an open problem. We measure the style of the output sentence by using a fastText classifier [JOULIN and collab., 2016b]<sup>1</sup>. For content preservation, we follow JOHN and collab. [2018] and compute both: (i) the cosine measure between source and generated sentence embeddings, which are the concatenation of min, max, and

---

<sup>1</sup>This procedure also follows COLOMBO and collab. [2019] (e.g., polarity, gender or category) which uses a fasttext [BOJANOWSKI and collab., 2017; JOULIN and collab., 2016a,b] classifier <https://fasttext.cc/docs/en/supervised-tutorial.html>. The validation corpus is used to select the best model. Preliminary comparisons with deep classifiers (based on either convolutional layers or recurrent layers) show that fasttext obtains similar result while being lighter and faster.

mean of word embedding (sentiment words removed)<sup>2</sup>, and (ii) the BLEU score between generated text and the input<sup>3</sup>. Motivated by previous work, we evaluate the fluency of the language with the perplexity given by a GPT-2 [RADFORD and collab., 2019] pretrained model performing fine-tuning on the training corpus<sup>4</sup>. We choose to report the log-perplexity since we believe it can better reflects the uncertainty of the language model (a small variation in the model loss would induce a large change in the perplexity due to the exponential term). Besides the automatic evaluation, we further test our disentangled representation effectiveness by human evaluation results are presented in Table 7.1.

**Conventions and abbreviations.** *Adv* refers to a model trained using the adversarial loss; vCLUB-S, KL refers to a model trained using the vCLUB-S and KL surrogate (see Equation A.7) respectively; and  $D_\alpha$  refers to a model trained based on the  $\alpha$ -Renyi surrogate (Equation A.8), for  $\alpha \in \{1.3, 1.5, 1.8\}$ .

### 7.4.3 Architecture Hyperparameters

We use an encoder parameterized by a 2-layer bidirectional GRU [CHUNG and collab., 2014] and a 2-layer decoder GRU. Both GRU and our word embedding lookup tables, trained from scratch, and have a dimension of 128 (as already reported by GARCIA and collab. [2019], building experiments on higher dimensions produces marginal improvement). The style embedding is set to a dimension of 8. The attribute classifier are MLP and are composed of 3 layer MLP with 128 hidden units and LeakyReLU [XU and collab., 2015] activations, the dropout [SRIVASTAVA and collab., 2014] rate is set to 0.1. All models are optimised with AdamW [KINGMA and BA, 2014; LOSHCHILOV and HUTTER, 2017] with a learning rate of  $10^{-3}$  and the norm is clipped to 1.0. Our model's hyperparameters have been set by a preliminary training on each downstream task: a simple classifier for the fair classification and a vanilla seq2seq [COLOMBO and collab., 2020; SUTSKEVER and collab., 2014] for the conditional generation task. The models requested for the classification task are trained during 100k steps while 300k steps are used for the generation task.

## 7.5 Numerical Results

In this section, we present our results on the fair classification and binary sequence generation tasks, see subsection 7.5.1 and subsection 7.5.2, respectively. We additionally show that our variational surrogates to the MI—contrarily to adversarial losses—do not suffer in multi-class scenarios (see section 7.5.4). We choose to present results on fair classification first as the evaluation of fair classification is easier than the evaluation of text generation (see Chapter 8) as it only relies on accuracy. Thus,

<sup>2</sup>For computing the embedding we rely on the bag of word model and take the mean pooling of word embedding. We choose to use the pre-trained word vectors provided in <https://fasttext.cc/docs/en/pretrained-vectors.html>. They are trained on Wikipedia using fastText. These vectors in dimension 300 were obtained using the skip-gram model described in BOJANOWSKI and collab. [2017]; JOULIN and collab. [2016b] with default parameters.

<sup>3</sup>For computing the BLEU score we choose to use the corpus level method provided in python sacrebleu [POST, 2018] library <https://github.com/mjpost/sacrebleu.git>. It produces the official WMT scores while working with plain text.

<sup>4</sup>This procedure follows JALALZAI and collab. [2020]. GPT-2 is pre-trained on the BookCorpus dataset ? (around 800M words). The model has been taken from the HuggingFace Library [WOLF and collab., 2019]. Default hyperparameters have been used for the finetuning.

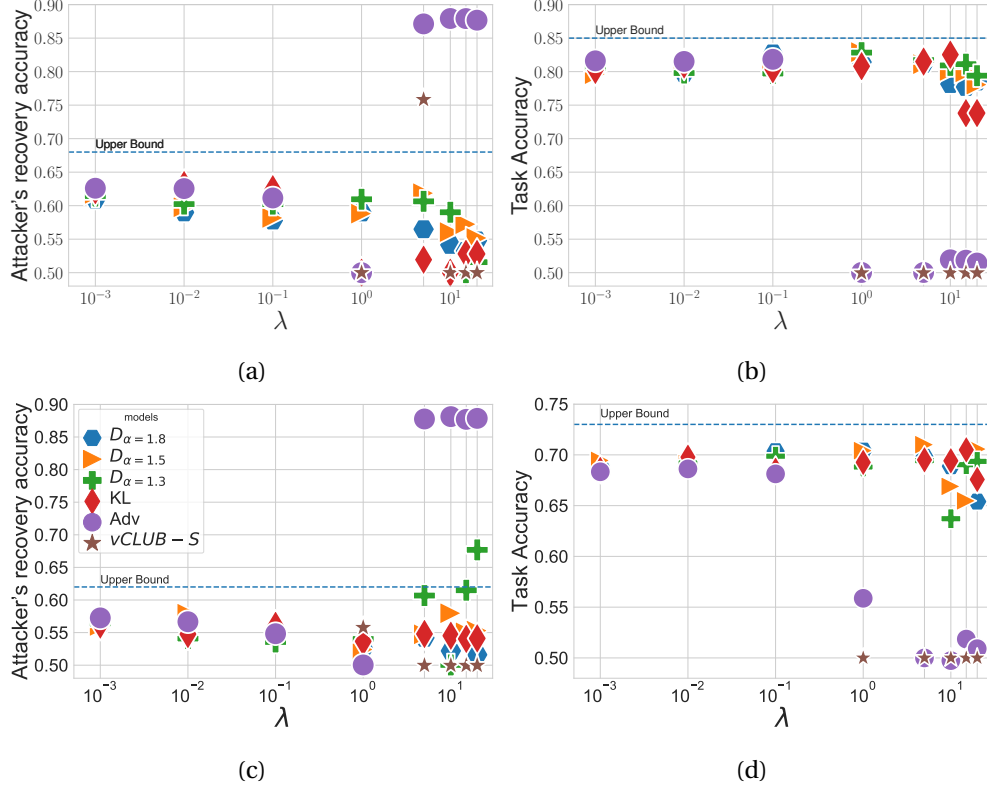


Figure 7.3 – Numerical results on fair classification. Trade-offs between target task and attacker accuracy are reported in Figure 7.3a, Figure 7.3b for mention task, and Figure 7.3c, Figure 7.3d for sentiment task. For low values of  $\lambda$  some points coincide. As  $\lambda$  increases the level of disentanglement increases and the proposed methods using both KL (KL) and Reny divergences ( $\mathcal{D}_\alpha$ ) clearly offer better control than existing methods.

conclusions and existing trade-off in fair classification will guide our analysis on the text generation tasks.

### 7.5.1 Applications to Fairness

**Upper bound on performances.** We first examine how much of the protected attribute we can be recovered from an unfair classifier (*i.e.*, trained without adversarial loss) and how well does such classifier perform. Results are reported in Figure 7.3. We observe that we achieve similar scores than the ones reported in previous studies [BARRETT and collab., 2019; ELAZAR and GOLDBERG, 2018]. This experiment shows that, when training to solve the main task, the classifier learns information about the protected attribute, *i.e.*, the attacker's accuracy is better than random guessing. In the following, we compare the different proposed methods to disentangle representations and obtain a fairer classifier.

**Methods comparisons.** Figure 7.3 shows the results of the different models and illustrates the trade-offs between disentangled representations and the target task accuracy. Results are reported on the test set for both sentiment and mention tasks when race is the protected. We observe that the classifier trained with an adversarial loss degenerates for  $\lambda > 5$  since the adversarial term in Equation 7.1 is influencing much the global gradient than the downstream term (*i.e.*, cross-entropy loss between predicted and golden distribution). Remarkably, both models trained to minimize either the KL or the Renyi surrogate do not suffer much from the afore-



mentioned multi-class problem. For both tasks, we observe that the KL and the Renyi surrogates can offer better disentangled representations than those induced by adversarial approaches. In this task, both the KL and Renyi achieve perfect disentangled representations (*i.e.*, random guessing accuracy on protected attributes) with a 5% drop in the accuracy of the target task, when perfectly masking the protected attributes. As a matter of fact, we observe that vCLUB-S provides only two regimes: either a “light” protection (attacker accuracy around 60%), with almost no loss in task accuracy ( $\lambda < 1$ ), or a strong protection (attacker accuracy around 50%), where a few features relevant to the target task remain.<sup>5</sup> On the sentiment task, we can draw similar conclusions. However, the Renyi’s surrogate achieves slightly better-disentangled representations. Overall, we can observe that our proposed surrogate enables good control of the degree of disentangling. Additionally, we do not observe a degenerated behaviour—as it is the case with adversarial losses—when  $\lambda$  increases. Furthermore, our surrogate allows simultaneously better disentangled representations while preserving the accuracy of the target task.

### 7.5.2 Binary Sentence Generation: Application to Binary Sentiment Labels

In the previous section, we have shown that the proposed surrogates do not suffer from limitations of adversarial losses and allow to achieve better disentangled representations than existing methods relying on CLUB. Disentanglement modules are a core block for a large number of both style transfer and conditional sentence generation algorithms [FU and collab., 2017; TIKHONOV and collab., 2019; YAMSHCHIKOV and collab., 2019] that place explicit constraints to force disentangled representations. First, we assess the disentanglement quality and the control over desired level of disentanglement while changing the downstream term, which for the sentence generation task is the cross-entropy loss on individual token. Then, we exhibit the existing trade-offs between quality of generated sentences, measured by the metric introduced in subsection 7.4.2, and the resulting degree of disentanglement. The results are presented for SYelp

#### Evaluating disentanglement

Figure 7.4a shows the adversary accuracy of the different methods as a function of  $\lambda$ . Similarly to the fair classification task, a fair amount of information can be recovered from the embedding learnt with adversarial loss. In addition, we observe a clear degradation of its performance for values  $\lambda > 1$ . In this setting, the Renyi surrogates achieves consistently better results in terms of disentanglement than the one minimizing the KL surrogate. The curve for Renyi’s surrogates shows that exploring different values of  $\lambda$  allows good control of the disentanglement degree. Renyi surrogate generalizes well for sentence generation. Similarly to the fairness task CLUB only offers two regimes: “light” disentanglement with very little polarity transfer and “strong” disentanglement.

---

<sup>5</sup>This phenomenon is also reported in FEUTRY and collab. [2018] on a picture anonymization task.

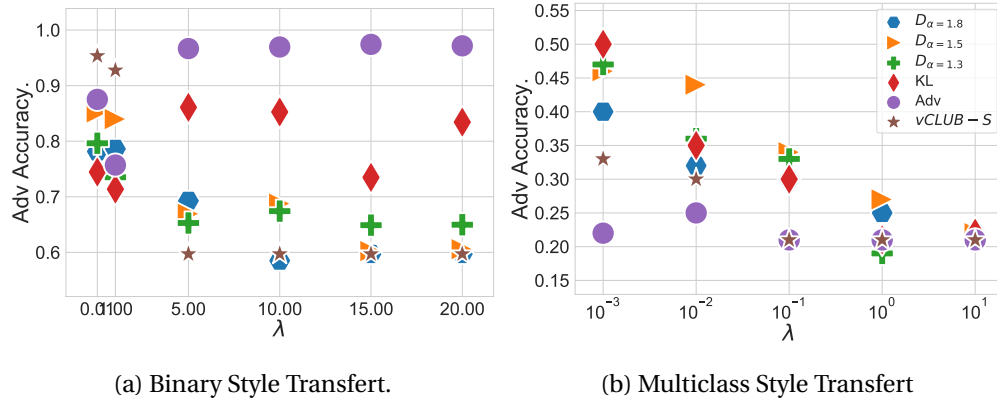


Figure 7.4 – Disentanglement of representation learnt by  $f_{\theta_e}$  in the binary (left) and multi-class (*i.e.*,  $|\mathcal{Y}| = 5$ ) (right) sentence generation scenario. In the multi-class scenario the *Adv* degenerates for  $\lambda \geq 0.01$  and offer no fined-grained control over the degree of disentanglement.

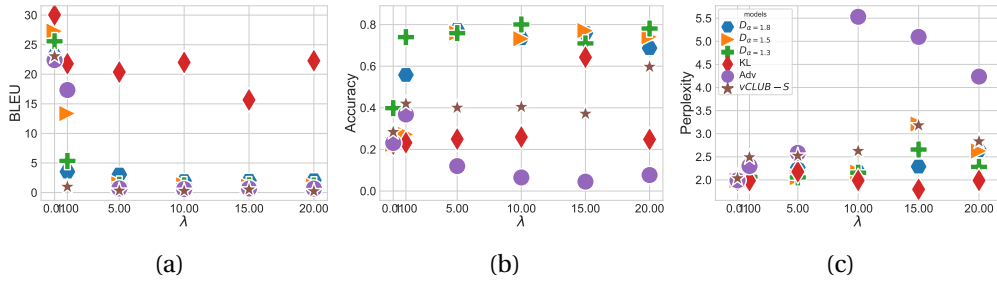


Figure 7.5 – Numerical experiments on binary style transfer. Quality of generated sentences are evaluated using BLEU (Figure 7.5a); style transfer accuracy (Figure 7.5a); sentence fluency (Figure 7.5c). We report existing trade-offs between disentanglement and sentence generation quality. Human evaluation is reported in Table 7.1.

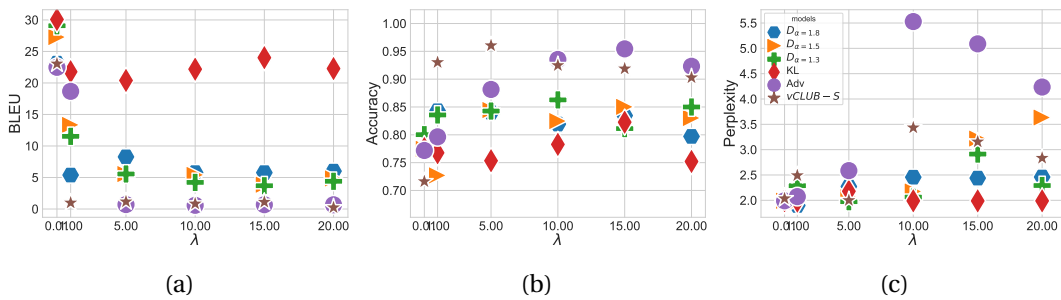


Figure 7.6 – Numerical experiments on conditional sentence generation. Results include BLEU (Figure 7.6a), style transfer accuracy (Figure 7.6b) and sentence fluency (Figure 7.6c).



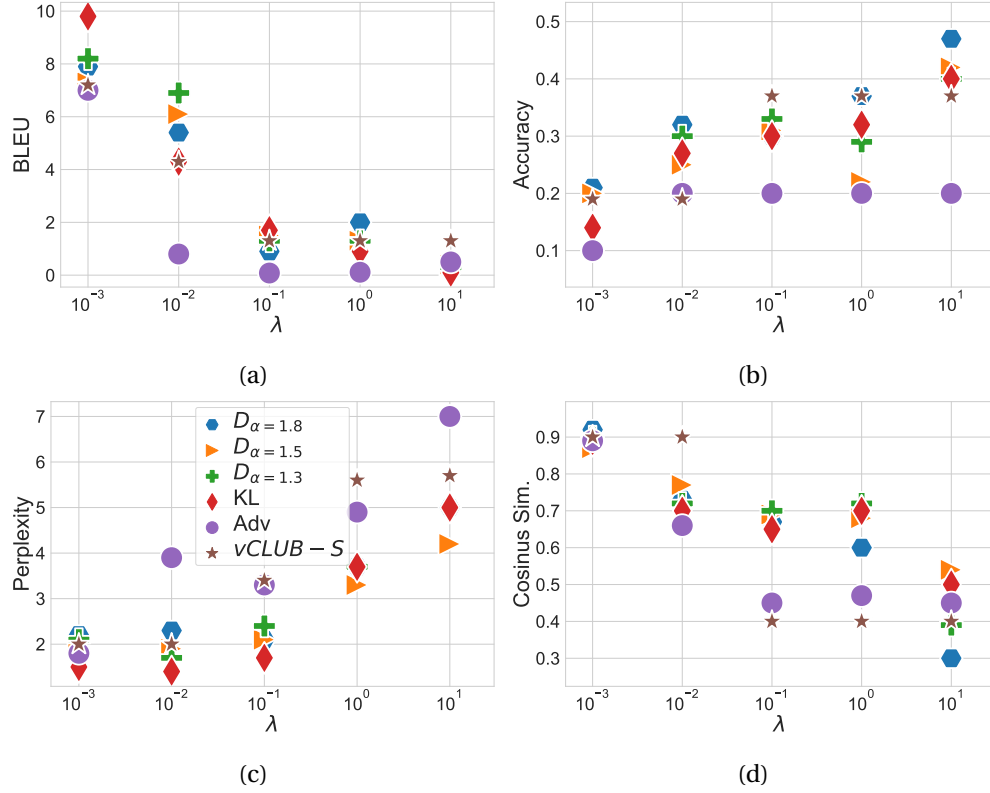


Figure 7.7 – Numerical experiments on multiclass style transfer using categorical labels. Results include: BLEU (Figure 7.7a); style transfer accuracy (Figure 7.7b); sentence fluency (Figure 7.7c); cosine similarity (Figure 7.7d)

### Disentanglement with Polarity Labels

The quality of generated sentences are evaluated using the fluency (see Figure 7.5c), the content preservation (see Figure 7.5a), additional results using a cosine similarity are given in Figure 7.12, and polarity accuracy (see Figure 7.5b). For style transfer, and for all models, we observe trade-offs between disentanglement and content preservation (measured by BLEU) and between fluency and disentanglement. Learning disentangled representations leads to poorer content preservation. As a matter of fact, similar conclusions can be drawn while measuring content with the cosine similarity (see ??). For polarity accuracy, in non-degenerated cases (see below), we observe that the model is able to better transfer the sentiment in presence of disentangled representations. *Transferring style is easier with disentangled representations, however there is no free lunch here since disentangling also removes important information about the content.* It is worth noting that even in the "strong" disentanglement regime vCLUB-S struggles to transfer the polarity (accuracy of 40% for  $\lambda \in \{1, 2, 10, 15\}$ ) where other models reach 80%. It is worth noting that similar conclusions hold for two different sentence generation tasks: style transfer and conditional generation, which tends to validate the current line of work that formulates text generation as generic text-to-text [RAFFEL and collab., 2019].

**Quality of generated sentences.** Table 7.1 gathers results of human evaluation and show that our surrogates can better disentangle style while preserving more content than available methods. In Table 7.1, we report the performances of systems when evaluated by humans on the polarity transfer task. 100 sentences are generated by each system and 3 english native speakers are asked to annotate each sentence

Model	Fluency	Content	Sentiment
Human	0.80	3.4	0.78
<i>Adv</i>	0.60	2.4	0.63
$\nu$ CLUB – S	0.62	2.6	0.65
KL	0.68	2.6	0.63
$D_{\alpha=1.3}$	0.70	2.4	0.65
$D_{\alpha=1.5}$	0.68	2.9	0.70
$D_{\alpha=1.8}$	0.76	3.0	0.58

Table 7.1 – Human annotations of generated samples. For the comparison we rely on the sentences provided in [https://github.com/rpryzant/delete\\_retrieve\\_generate](https://github.com/rpryzant/delete_retrieve_generate). Human annotations are also provided by [LI and collab. \[2018\]](#). We have reprocessed the provided sentence using a tokenizer based on SentencePiece [[KUDO, 2018](#); [SENNRICH and collab., 2016](#)]. Since there is a trade-off between automatic evaluation metrics (*i.e.* BLEU, Perplexity and Accuracy of Style Transfer), we set minimum thresholds on BLEU and on style transfer accuracy. The best model that met the threshold on validation is selected. We will release—along with our code—new generated sentences for comparison.

along 3 dimensions (*i.e.* fluency, sentiment and content preservation). Turkers assign binary labels to fluency and sentiment (following the protocol introduced in [JALALZAI and collab. \[2020\]](#)) while content is evaluated on a likert scale from 1-5. For content preservation, both the input sentence and the generated sentence are provided to the turker. The annotator agreement is measure by the Krippendorff Alpha<sup>6</sup> [[KRIPPENDORFF, 2018](#)]. The Krippendorff Alpha is:  $\alpha = 0.54$  on the sentiment classification,  $\alpha = 0.20$  for fluency and  $\alpha = 0.18$  for content preservation.

### Example of generated sentences

[Table 7.2](#) gathers sentences generated by the different sentences for different values of  $\lambda$ . They provide qualitative examples that illustrate the previously observed trade-offs. The adversarial loss degenerates for values  $\lambda \geq 5$  and a stuttering phenomenon appears [[HOLTZMAN and collab., 2019](#)].

**Style transfert.** From [Table 7.2](#), we can observe that the impact of disentanglement on a qualitative point of view. For small values of  $\lambda$  the models struggle to do the style transfer (see example 2 for instance). As  $\lambda$  increases disentanglement becomes easier, however, the content becomes more generic which is a known problem (see [LI and collab. \[2015\]](#) for instance).

**Conditional sentence generation.** From qualitative example displayed in [Table 7.3](#), we can draw similar conclusions than those for quantitative metrics previously displayed: as the disentanglement increases, the common content which is shared between input and generated sentences decreases.

**Example of “degeneracy” for large values of  $\lambda$ .** For sentences generated with the baseline model a repetition phenomenon appears for greater values of  $\lambda$ . For certain sentences, models ignore the style token (*i.e.*, the sentence generated with a positive sentiment is the same as the one generated with the negative sentiment). We attribute this degeneracy to the fact that the model is only trained with  $(x_i, y_i)$  sharing the same sentiment which appears to be an intrinsic limitation of the model introduced by [JOHN and collab. \[2018\]](#).

<sup>6</sup>Krippendorff Alpha measures of inter-rater reliability in  $[0, 1]$ : 0 is perfect disagreement and 1 is perfect agreement.

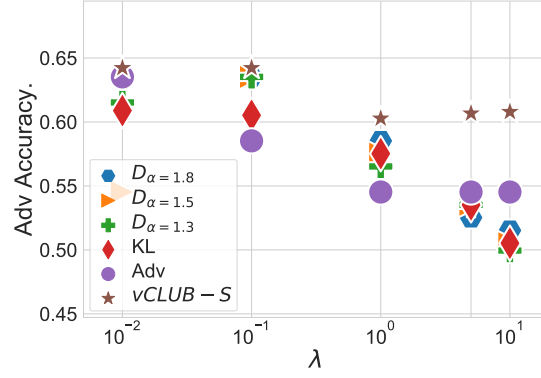


Figure 7.8 – Disentanglement of the learnt embedding when training an off-line adversarial classifier for the sentence generation with gender data.

**Analysis of performances of vCLUB-S** Similarly to what can be observed with automatic evaluation Table 7.2 shows that the system based on vCLUB-S has only two regimes: “light” disentanglement and strong disentanglement. With light disentanglement the decoder fail at transferring the polarity and for strong disentanglement few content features remain and the system tends to output generic sentences.

### 7.5.3 Binary Sentence Generation: Application to Gender Data

#### Quality of the Disentanglement

In Figure 7.8, we report the adversary accuracy of the different methods for the values of  $\lambda$ . It is worth noting that gender labels are noisier than sentiment labels [LAMPLE and collab., 2018]. We observe that the adversarial loss saturates at 55% where a model trained on MI bounds can achieve a better disentanglement. Additionally, the models trained with MI bounds allow better control of the desired degree of disentanglement.

#### Quality of Generated Sentences

Results on the sentence generation tasks are reported in Figure 7.9 and in Figure 7.10. We observe that for  $\lambda > 1$  the adversarial loss degenerates as observed in the sentiment experiments. Compared to sentiment score we observe a lower score of BLEU which can be explained by the length of the review in the FYelp dataset. On the other hand, we observe a similar trade-off between style transfer accuracy and content preservation in the non degenerated case: as style transfer accuracy increases, content preservation decreases. Overall, we remark a behaviour similar to the one we observe in sentiment experiments.

### 7.5.4 Results on Multi class Sentence Generation

Results on the multi-class style transfer and on conditional sentence generation are reported in Figure 7.7b and Figure 7.6b.

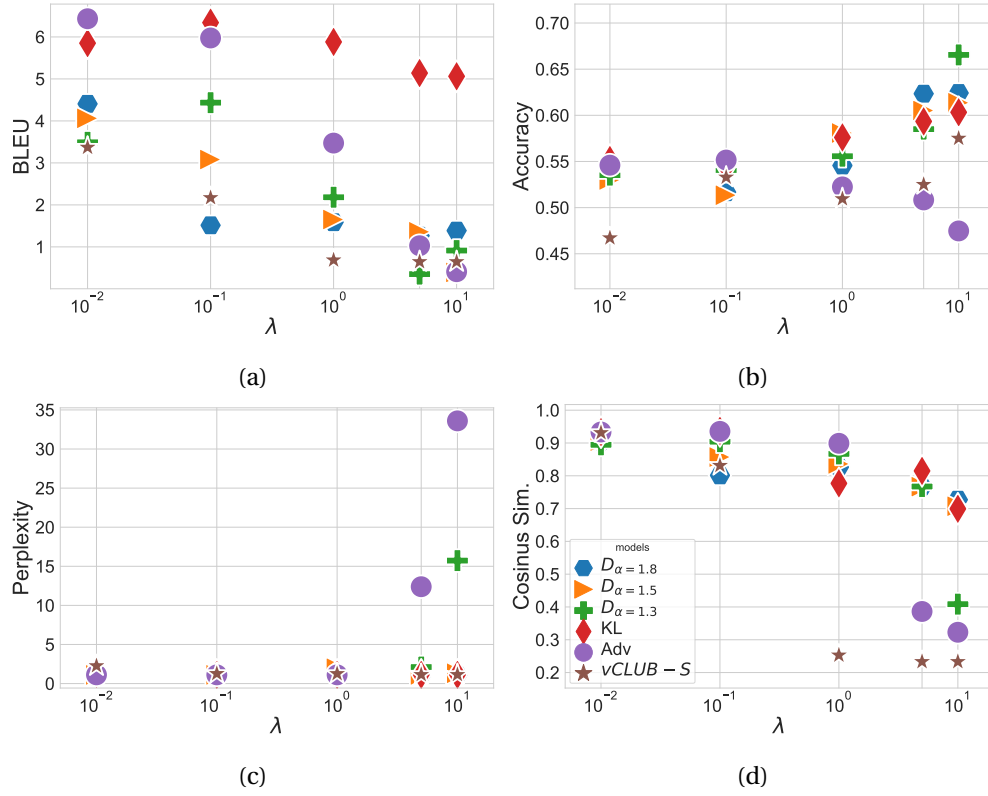


Figure 7.9 – Numerical experiments on binary style transfer using gender labels. Results include: BLEU (Figure 7.9a); cosine similarity (Figure 7.9d); style transfer accuracy (Figure 7.9b); sentence fluency (Figure 7.9c).

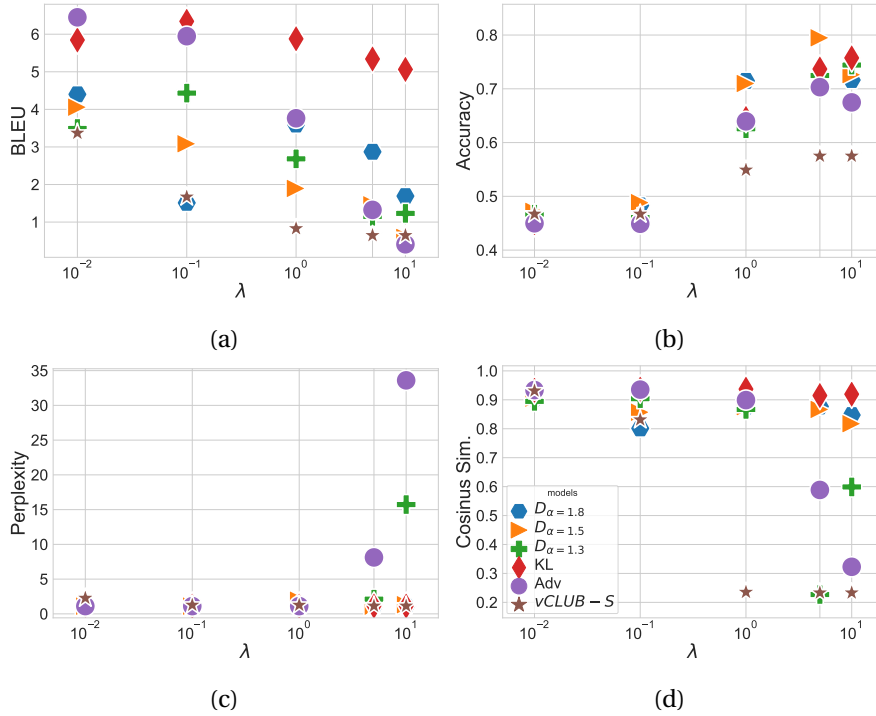


Figure 7.10 – Numerical experiments on conditional sentence generation using gender labels. Results includes: BLEU (Figure 7.10a); cosine similarity (Figure 7.10d); style transfer accuracy (Figure 7.10b); sentence fluency (Figure 7.10c).

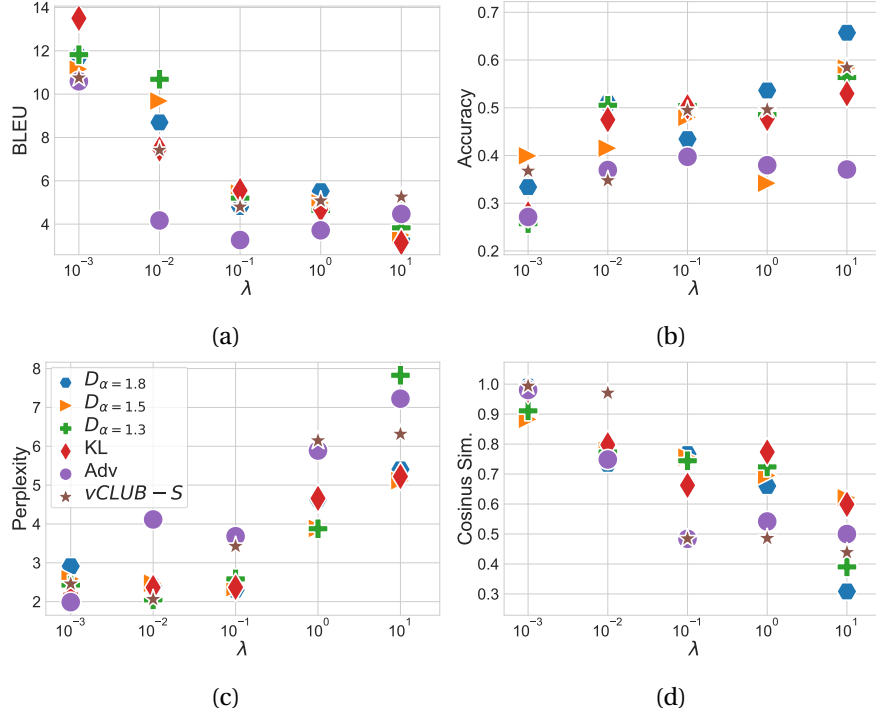


Figure 7.11 – Numerical experiments on the multi-class conditional sentence generation. Results include: BLEU (Figure 7.11a); cosine similarity (Figure 7.11d); style transfer accuracy (Figure 7.11b); sentence fluency (Figure 7.11c).

### Adversarial Loss Fails to Disentangle when $|\mathcal{Y}| \geq 3$

In Figure 7.4b we report the adversary accuracy of our different methods for the values of  $\lambda$  using FYelp dataset with category label. In the binary setting for  $\lambda \leq 1$ , models using adversarial loss can learn disentangled representations while in the multi-class setting, the adversarial loss degenerates for small values of  $\lambda$  (*i.e* sentences are no longer fluent as shown by the increase in perplexity in Figure 7.7c). Minimizing MI based on our surrogates seems to mitigate the problem and offer a better control of the disentanglement degree for various values of  $\lambda$  than vCLUB-S.

### Automatic Evaluation of generated sentences

Results on the multi-class style transfer and on conditional sentence generation are reported in Figure 7.7b and Figure 7.6b. Similarly than in the binary case, there exists a trade-off between content preservation and style transfer accuracy. We observe that the BLEU score in this task is in a similar range than the one in the gender task, which is expected because data come from the same dataset where only the labels changed.

Similarly than in the binary case, there exists a trade-off between content preservation and style transfer accuracy. We observe that the BLEU score in this task is in a similar range than the one in the gender task, which is expected because data come from the same dataset where only the labels changed.

$\lambda$	Model	Sentence
	<b>Input</b>	<b>the food was the best food i've ever experienced.</b>
	Adv	the food was the best i've ever had in.
	vCLUB-S	the food was the best i've ever had.

	KL	the food was the best food i've ever experienced.
	$D_{\alpha=1.3}$	the food was the best food i've experienced.
	$D_{\alpha=1.5}$	the food was so good and the best i ever had.
	$D_{\alpha=1.8}$	the food is so good i will be going back.
	Input	the food was the best food i've ever experienced.
	Adv	the food was the best i've ever eaten here.
	vCLUB-S	the food was the best.
1	KL	the food was the best food i've ever experienced.
	$D_{\alpha=1.3}$	the food was the best i've ever eaten at.
	$D_{\alpha=1.5}$	the food was amazing as well as i am extremely satisfied.
	$D_{\alpha=1.8}$	the food was very good and the service good.
	Input	the food was the best food i've ever experienced.
	Adv	i love this place.
	vCLUB-S	i love it.
5	KL	the food was the best i've ever eaten here.
	$D_{\alpha=1.3}$	the food is ok, but the service is terrible.
	$D_{\alpha=1.5}$	the food is always good but the service is always bad.
	$D_{\alpha=1.8}$	the food was ok and very good.
	Input	the food was the best food i've ever experienced.
	Adv	i love this place.
	vCLUB-S	i love it.
10	KL	the food was excellent, but i love this food.
	$D_{\alpha=1.3}$	the food was worst at best.
	$D_{\alpha=1.5}$	the food was not well cooked with the sauce.
	$D_{\alpha=1.8}$	the food wasn't bad but it was not good.
	Input	<b>It's freshly made, very soft and flavorful.</b>
	Adv	it's crispy and too nice and very flavor.
	vCLUB-S	It's freshly made, and great.
0.1	KL	it's a huge, crispy and flavorful.
	$D_{\alpha=1.3}$	it's hard, and the flavor was flavorless.
	$D_{\alpha=1.5}$	it's very dry and not very flavorful either.
	$D_{\alpha=1.8}$	it's a good place for lunch or dinner.
	Input	it's freshly made, very soft and flavorful.
	Adv	it's not crispy and not very flavorful flavor.
	vCLUB-S	It's bad.
1	KL	it's very fresh, and very flavorful and flavor.
	$D_{\alpha=1.3}$	it's not good, but the prices are good.
	$D_{\alpha=1.5}$	it's not very good, and the service was terrible.
	$D_{\alpha=1.8}$	it was a very disappointing experience and the food was awful.
	Input	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
5	KL	it's very fresh, flavorful and flavorful.
	$D_{\alpha=1.3}$	it's not worth the money, but it was wrong.
	$D_{\alpha=1.5}$	it's not worth the price, but not worth it.
	$D_{\alpha=1.8}$	it's hard to find, and this place is horrible.
	Input	it's freshly made, very soft and flavorful.
	Adv	i hate this place.
	vCLUB-S	i hate it.
10	KL	it's a little warm and very flavorful flavor.
	$D_{\alpha=1.3}$	it was a little overpriced and not very good.
	$D_{\alpha=1.5}$	it's a shame, and the service is horrible.

	$D_{\alpha=1.8}$	it's not worth the \$ NUM.
	<b>Input</b>	<b>Only then did our waitress show up with another styrofoam cup full of water.</b>
	Adv	then she didn't get a glass of coffee she was full full full full water.
0.1	vCLUB-S	the waitress broke the cup of water
	KL	only NUM hours of us in the water and no gratuity of a water.
	$D_{\alpha=1.3}$	waited NUM minutes at the front with us and offered to an ice glass water.
	$D_{\alpha=1.5}$	after NUM minutes of a table with a table and two entrees arrived.
	$D_{\alpha=1.8}$	after NUM minutes of a table with a table and NUM entrees arrived.
	<b>Input</b>	<b>Only then did our waitress show up with another styrofoam cup full of water.</b>
	Adv	only NUM minutes of our waiter was able to get a refilled ice cream.
	vCLUB-S	the waitress is bad i hate this place
1	KL	even the refund of them were brought out to refill the plate of our order.
	$D_{\alpha=1.3}$	NUM stars for the short NUM minute wait and recommend the perfect patio.
	$D_{\alpha=1.5}$	NUM minutes later, my food came out NUM minutes after our order.
	$D_{\alpha=1.8}$	i've been many years at the same time and great service.
	<b>Input</b>	<b>Only then did our waitress show up with another styrofoam cup full of water.</b>
	Adv	great price.
	vCLUB-S	i love it
5	KL	she was able to get us in for a table.
	$D_{\alpha=1.3}$	they are very friendly and have a great selection of beers and drinks.
	$D_{\alpha=1.5}$	i have been here several times and it's always a good experience.
	$D_{\alpha=1.8}$	he's a great guy and a very nice person with a smile.
	<b>Input</b>	<b>Only then did our waitress show up with another styrofoam cup full of water.</b>
	Adv	our server was very friendly and attentive.
	vCLUB-S	i love it
10	KL	great food, great prices, and great prices for a good price.
	$D_{\alpha=1.3}$	and i've been to this place since NUM years and love it.
	$D_{\alpha=1.5}$	only did the refill on us for about NUM mins with water tables.
	$D_{\alpha=1.8}$	i love the place.

Table 7.2 – Sequences generated by the different models on the binary sentiment transfer task.

$\lambda$	Model	Sentence
	<b>Input</b>	<b>Definitely every flavor for every person.</b>
	Adv	every thing have every other time.
	vCLUB-S	definitely very flavorful.
0.1	KL	definitely a good time to visit.
	$D_{\alpha=1.3}$	definitely worth every way every way.
	$D_{\alpha=1.5}$	definitely worth a try for all.
	$D_{\alpha=1.8}$	definitely worth a try to eat.
	<b>Input</b>	<b>Definitely every flavor for every person.</b>
	Adv	definitely my wife and i love.
	vCLUB-S	i like it. good place to eat.
1	KL	definitely worth every penny every time.
	$D_{\alpha=1.3}$	definitely worth the drive to earth.
	$D_{\alpha=1.5}$	definitely a recommend the whole family.
	$D_{\alpha=1.8}$	thank you for your help.
	<b>Input</b>	<b>Definitely every flavor for every person.</b>
	Adv	definitely a good place to eat.
	vCLUB-S	i love it !
5	KL	always a great experience.

	$D_{\alpha=1.3}$	a great place to eat.
	$D_{\alpha=1.5}$	definitely my go - to spot.
	$D_{\alpha=1.8}$	great service and great food.
	Input	Definitely every flavor for every person.
	Adv	i love this place!
	vCLUB-S	i love the flavor !
10	KL	definitely get my good time there.
	$D_{\alpha=1.3}$	very good and fast service.
	$D_{\alpha=1.5}$	i would recommend this place to anyone.
	$D_{\alpha=1.8}$	definitely worth the drive.
	Input	<b>needless to say, i will be paying them a visit and contacting corporate.</b>
	Adv	needless to say i will never be back with this vet... unacceptable.
	vCLUB-S	needless to say i will be back and don't recommend it.
0.1	$D_{\alpha=1.3}$	needless to say, i will never be back to a new office and walked away.
	$D_{\alpha=1.5}$	needless to say, i will never be back to this location with my flight.
	$D_{\alpha=1.8}$	needless to say, i'm not sure what i wanted to get it.
	Input	needless to say, i will be paying them a visit and contacting corporate.
	Adv	needless to say, i will never be back, and i am a member.
	vCLUB-S	needless to say i hate it.
1	KL	needless to say i will be back for a year and i am completely satisfied.
	$D_{\alpha=1.3}$	i wouldn't recommend this place to anyone who needs a good job.
	$D_{\alpha=1.5}$	needless to say, i will not be going back to this particular location again.
	$D_{\alpha=1.8}$	i'm not sure what i've had at this place....
	Input	needless to say, i will be paying them a visit and contacting corporate.
	Adv	i'm not sure what i'm going to this place.
	vCLUB-S	i won't be back again.
5	KL	needless to say, i will never go back, and i am completely unhappy.
	$D_{\alpha=1.3}$	they aren't even that busy, but the food isn't good.
	$D_{\alpha=1.5}$	if you're looking for a good deal, you'll find better.
	$D_{\alpha=1.8}$	needless to say, i didn't have a bad experience.
	Input	needless to say, i will be paying them a visit and contacting corporate.
	Adv	i'm not sure what i've been to.
	vCLUB-S	i hate it.
10	KL	needless to say, i will be back again, and a complete complete joke.
	$D_{\alpha=1.3}$	i'm not sure what the other reviews are to the worst.
	$D_{\alpha=1.5}$	needless to say, i will not be going back to this location.
	$D_{\alpha=1.8}$	i've been to this location NUM times and it's not good.
	Input	<b>We had to wait for a table maybe NUM min.</b>
	Adv	we had to wait for a table NUM mins.
	vCLUB-S	we had to wait for a table NUM mins.
0.1	KL	we had to wait for a wait for NUM min.
	$D_{\alpha=1.3}$	we had to wait a table for NUM min.
	$D_{\alpha=1.5}$	we had a NUM minute wait for over two minutes.
	$D_{\alpha=1.8}$	we had a bad experience with agroupon for NUM.
	Input	we had to wait for a table maybe NUM min.
	Adv	we went to wait for NUM minutes for no one.
	vCLUB-S	i dislike it.
1	KL	we had a wait time for us to order NUM.
	$D_{\alpha=1.3}$	we waited for NUM minutes for a refill order.
	$D_{\alpha=1.5}$	we had a bad experience.
	$D_{\alpha=1.8}$	we had a NUM minute wait for a table.
	Input	we had to wait for a table maybe NUM min.



Adv	i'm not sure what i paid for.
vCLUB-S	i don't like it.
KL	we ordered a table for NUM minutes of our table.
$D_{\alpha=1.3}$	we were seated immediately and we weren't even acknowledged.
$D_{\alpha=1.5}$	we ordered a chicken parm chicken and it was very bland.
$D_{\alpha=1.8}$	we had a bad experience with my boyfriend's birthday.
Input	we had to wait for a table maybe NUM min.
Adv	i'm not sure what happened.
vCLUB-S	i don't like it.
10 KL	we had a table to get a table for NUM.
$D_{\alpha=1.3}$	we ordered NUM for a lunch special and was very disappointed.
$D_{\alpha=1.5}$	we were seated immediately and we waited.
$D_{\alpha=1.8}$	we ordered NUM wings, NUM of NUM tacos and we waited.

Table 7.3 – Sequences generated by the different models on the binary sentiment conditional sentence generation task.

### Chapter 7 Conclusion

In this chapter, we devised a new alternative method to adversarial losses capable of learning disentangled textual representation based on MI. Our new variationnal bound on MI does not require adversarial training and hence, it does not suffer in presence of multi-class setups. A key feature of this new estimator is to account for the approximation error incurred when bounding the mutual information. Experiments show better trade-offs than both adversarial training and CLUB on two fair classification tasks and demonstrate the efficiency to learn disentangled representations for sequence generation. As a matter of fact, there is no free-lunch for sentence generation tasks: *although transferring style is easier with disentangled representations, it also removes important information about the content*. In this chapter, we believe that our conclusion have to be tempered by the weakness automatic metrics. Indeed, to assess content preservation we only rely on two simple heuristics (*i.e* token overlap and cosine similarity). Although, these metric are commonly used to assess style transfer we believe they might not be fully representatives of the quality of generated text. In the next chapter, we address AE in the specific case of text summarization and data2text generation by using different information measures.

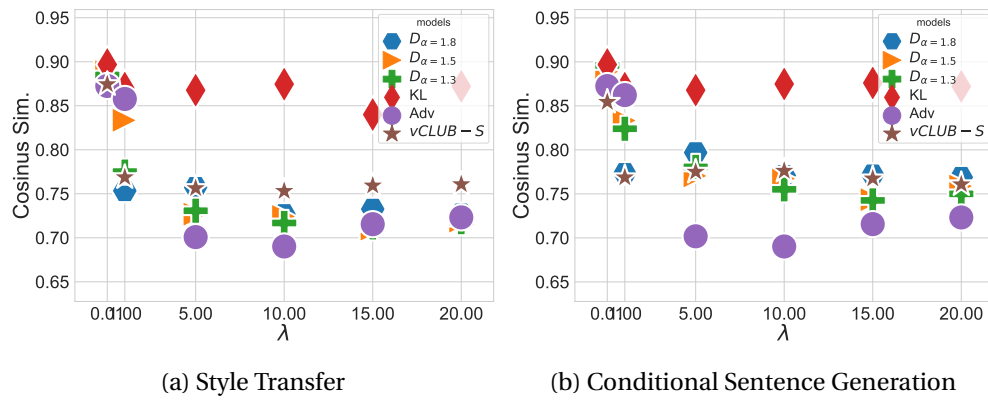


Figure 7.12 – Content preservation measured by the cosine similarity.

## 7.6 References

- ALI, S. M. and S. D. SILVEY. 1966, «A general class of coefficients of divergence of one distribution from another», *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 28, n° 1, p. 131–142. [129](#)
- BARRETT, M., Y. KEMENTCHEDJIEVA, Y. ELAZAR, D. ELLIOTT and A. SØGAARD. 2019, «Adversarial removal of demographic attributes revisited», in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 6331–6336. [128](#), [130](#), [133](#), [134](#), [136](#)
- BELGHAZI, M. I., A. BARATIN, S. RAJESWAR, S. OZAI, Y. BENGIO, A. COURVILLE and R. D. HJELM. 2018, «Mine: mutual information neural estimation», *arXiv preprint arXiv:1801.04062*. [131](#)
- BENGIO, Y., A. COURVILLE and P. VINCENT. 2013, «Representation learning: A review and new perspectives», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, n° 8, p. 1798–1828. [127](#)
- BLODGETT, S. L., L. GREEN and B. O’CONNOR. 2016, «Demographic dialectal variation in social media: A case study of African-American English», in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, p. 1119–1130, doi: 10.18653/v1/D16-1120. URL <https://www.aclweb.org/anthology/D16-1120>. [134](#)
- BOJANOWSKI, P., E. GRAVE, A. JOULIN and T. MIKOLOV. 2017, «Enriching word vectors with subword information», *Transactions of the Association for Computational Linguistics*, vol. 5, p. 135–146. [134](#), [135](#)
- BURGESS, C. P., I. HIGGINS, A. PAL, L. MATTHEY, N. WATTERS, G. DESJARDINS and A. LERCHNER. 2018, «Understanding disentangling in  $\beta$ -vae», *arXiv preprint arXiv:1804.03599*. [128](#)
- CHENG, P., W. HAO, S. DAI, J. LIU, Z. GAN and L. CARIN. 2020a, «Club: A contrastive log-ratio upper bound of mutual information», in *International Conference on Machine Learning*, PMLR, p. 1779–1788. [128](#), [133](#)
- CHENG, P., M. R. MIN, D. SHEN, C. MALON, Y. ZHANG, Y. LI and L. CARIN. 2020b, «Improving disentangled text representation learning with information-theoretic guidance», *arXiv preprint arXiv:2006.00693*. [128](#), [130](#)
- CHUNG, J., C. GULCEHRE, K. CHO and Y. BENGIO. 2014, «Empirical evaluation of gated recurrent neural networks on sequence modeling», *arXiv preprint arXiv:1412.3555*. [135](#)
- COLOMBO, P., E. CHAPUIS, M. MANICA, E. VIGNON, G. VARNI and C. CLAVEL. 2020, «Guiding attention in sequence-to-sequence models for dialogue act prediction.», in *AAAI*, p. 7594–7601. [135](#)
- COLOMBO, P., W. WITON, A. MODI, J. KENNEDY and M. KAPADIA. 2019, «Affect-driven dialog generation», *arXiv preprint arXiv:1904.02793*. [134](#)

- DAUDEL, K., R. DOUC and F. PORTIER. 2020, «Infinite-dimensional gradient-based descent for alpha-divergence minimisation», URL <https://hal.telecom-paris.fr/hal-02614605>, working paper or preprint. 131
- DENTON, E. L. and collab.. 2017, «Unsupervised learning of disentangled representations from video», in *Advances in neural information processing systems*, p. 4414–4423. 128
- DONSKER, M. and S. VARADHAN. 1985, «Large deviations for stationary gaussian processes», *Communications in Mathematical Physics*, vol. 97, n° 1-2, p. 187–210. 131
- ELAZAR, Y. and Y. GOLDBERG. 2018, «Adversarial removal of demographic attributes from text data», *arXiv preprint arXiv:1808.06640*. 128, 130, 132, 134, 136
- FEUTRY, C., P. PANTANIDA, Y. BENGIO and P. DUHAMEL. 2018, «Learning anonymized representations with adversarial neural networks», . 128, 129, 133, 137
- FU, Z., X. TAN, N. PENG, D. ZHAO and R. YAN. 2017, «Style transfer in text: Exploration and evaluation», *arXiv preprint arXiv:1711.06861*. 128, 130, 137
- GARCIA, A., P. COLOMBO, S. ESSID, F. D'ALCHÉ BUC and C. CLAVEL. 2019, «From the token to the review: A hierarchical multimodal approach to opinion mining», *arXiv preprint arXiv:1908.11216*. 135
- HJELM, R. D., A. FEDOROV, S. LAVOIE-MARCHILDON, K. GREWAL, P. BACHMAN, A. TRISCHLER and Y. BENGIO. 2018, «Learning deep representations by mutual information estimation and maximization», *arXiv preprint arXiv:1808.06670*. 131
- HOLTZMAN, A., J. BUYS, L. DU, M. FORBES and Y. CHOI. 2019, «The curious case of neural text degeneration», *arXiv preprint arXiv:1904.09751*. 140
- HSIEH, J.-T., B. LIU, D.-A. HUANG, L. F. FEI-FEI and J. C. NIEBLES. 2018, «Learning to decompose and disentangle representations for video prediction», in *Advances in Neural Information Processing Systems*, p. 517–526. 128
- HU, Z., Z. YANG, X. LIANG, R. SALAKHUTDINOV and E. P. XING. 2017, «Toward controlled generation of text», *arXiv preprint arXiv:1703.00955*. 130
- HUNG, Y.-N., Y.-A. CHEN and Y.-H. YANG. 2018, «Learning disentangled representations for timbre and pitch in music audio», *arXiv preprint arXiv:1811.03271*. 128
- JALALZAI, H., P. COLOMBO, C. CLAVEL, E. GAUSSIER, G. VARNI, E. VIGNON and A. SABOURIN. 2020, «Heavy-tailed representations, text polarity classification & data augmentation», *arXiv preprint arXiv:2003.11593*. 135, 140
- JOHN, V., L. MOU, H. BAHULEYAN and O. VECHTOMOVA. 2018, «Disentangled representation learning for non-parallel text style transfer», *arXiv preprint arXiv:1808.04339*. 128, 130, 132, 133, 134, 140
- JOULIN, A., E. GRAVE, P. BOJANOWSKI, M. DOUZE, H. JÉGOU and T. MIKOLOV. 2016a, «Fasttext.zip: Compressing text classification models», *arXiv preprint arXiv:1612.03651*. 134

- JOULIN, A., E. GRAVE, P. BOJANOWSKI and T. MIKOLOV. 2016b, «Bag of tricks for efficient text classification», *arXiv preprint arXiv:1607.01759*. 134, 135
- KINGMA, D. P. and J. BA. 2014, «Adam: A method for stochastic optimization», *arXiv preprint arXiv:1412.6980*. 135
- KINNEY, J. B. and G. S. ATWAL. 2014, «Equitability, mutual information, and the maximal information coefficient», *Proceedings of the National Academy of Sciences*, vol. 111, n° 9, p. 3354–3359. 128
- KPOTUFE, S. 2017, «Lipschitz Density-Ratios, Structured Data, and Data-driven Tuning», PMLR, Fort Lauderdale, FL, USA, p. 1320–1328. URL <http://proceedings.mlr.press/v54/kpotufe17a.html>. 132
- KRIPPENDORFF, K. 2018, *Content analysis: An introduction to its methodology*, Sage publications. 140
- KUDO, T. 2018, «Subword regularization: Improving neural network translation models with multiple subword candidates», *arXiv preprint arXiv:1804.10959*. 16, 140
- KUMAR VERMA, V., G. ARORA, A. MISHRA and P. RAI. 2018, «Generalized zero-shot learning via synthesized examples», in *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4281–4289. 128
- LAMPLE, G., S. SUBRAMANIAN, E. SMITH, L. DENOYER, M. RANZATO and Y.-L. BOUREAU. 2018, «Multiple-attribute text rewriting», in *International Conference on Learning Representations*. 130, 134, 141
- LI, J., M. GALLEY, C. BROCKETT, J. GAO and B. DOLAN. 2015, «A diversity-promoting objective function for neural conversation models», *arXiv preprint arXiv:1510.03055*. 140
- LI, J., R. JIA, H. HE and P. LIANG. 2018, «Delete, retrieve, generate: A simple approach to sentiment and style transfer», *arXiv preprint arXiv:1804.06437*. 16, 140
- LOSHCHILOV, I. and F. HUTTER. 2017, «Decoupled weight decay regularization», *arXiv preprint arXiv:1711.05101*. 135
- MOHRI, M., G. SIVEK and A. T. SURESH. 2019, «Agnostic federated learning», *arXiv preprint arXiv:1902.00146*. 130
- OORD, A. V. D., Y. LI and O. VINYALS. 2018, «Representation learning with contrastive predictive coding», *arXiv preprint arXiv:1807.03748*. 131
- PANINSKI, L. 2003, «Estimation of entropy and mutual information», *Neural computation*, vol. 15, n° 6, p. 1191–1253. 128, 131
- PICHLER, G., P. PIANTANIDA and G. KOLIANDER. 2020, «On the estimation of information measures of continuous distributions», . 128, 131
- POST, M. 2018, «A call for clarity in reporting bleu scores», *arXiv preprint arXiv:1804.08771*. 135

- PRABHUMOYE, S., Y. TSVETKOV, R. SALAKHUTDINOV and A. W. BLACK. 2018, «Style transfer through back-translation», *arXiv preprint arXiv:1804.09000*. 130
- RADFORD, A., J. WU, R. CHILD, D. LUAN, D. AMODEI and I. SUTSKEVER. 2019, «Language models are unsupervised multitask learners», *OpenAI Blog*, vol. 1, n° 8, p. 9. 135
- RAFFEL, C., N. SHAZEER, A. ROBERTS, K. LEE, S. NARANG, M. MATENA, Y. ZHOU, W. LI and P. J. LIU. 2019, «Exploring the limits of transfer learning with a unified text-to-text transformer», *arXiv preprint arXiv:1910.10683*. 139
- RÉNYI, A. and collab.. 1961, «On measures of entropy and information», in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California. 129
- ROSENBLATT, F. 1958, «The perceptron: a probabilistic model for information storage and organization in the brain.», *Psychological review*, vol. 65, n° 6, p. 386. 134
- SANCHEZ, E. H., M. SERRURIER and M. ORTNER. 2019, «Learning disentangled representations via mutual information estimation», *arXiv preprint arXiv:1912.03915*. 128
- SENNRICH, R., B. HADDOW and A. BIRCH. 2016, «Neural machine translation of rare words with subword units», in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, p. 1715–1725, doi: 10.18653/v1/P16-1162. URL <https://www.aclweb.org/anthology/P16-1162>. 16, 140
- SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER and R. SALAKHUTDINOV. 2014, «Dropout: a simple way to prevent neural networks from overfitting», *The journal of machine learning research*, vol. 15, n° 1, p. 1929–1958. 134, 135
- VAN STEENKISTE, S., F. LOCATELLO, J. SCHMIDHUBER and O. BACHEM. 2019, «Are disentangled representations helpful for abstract visual reasoning?», in *Advances in Neural Information Processing Systems*, p. 14 245–14 258. 128
- SUGIYAMA, M., T. SUZUKI and T. KANAMORI. 2012, *Density Ratio Estimation in Machine Learning*, 1<sup>re</sup> éd., Cambridge University Press, USA, ISBN 0521190177. 132
- SUTSKEVER, I., O. VINYALS and Q. V. LE. 2014, «Sequence to sequence learning with neural networks», in *Advances in neural information processing systems*, p. 3104–3112. 135
- TIKHONOV, A., V. SHIBAEV, A. NAGAEV, A. NUGMANOVA and I. P. YAMSHCHIKOV. 2019, «Style transfer for texts: Retrain, report errors, compare with rewrites», in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3927–3936. 128, 137
- WIETING, J., J. MALLINSON and K. GIMPEL. 2017, «Learning paraphrastic sentence embeddings from back-translated bitext», *arXiv preprint arXiv:1706.01847*. 130

- WOLF, T., L. DEBUT, V. SANH, J. CHAUMOND, C. DELANGUE, A. MOI, P. CISTAC, T. RAULT, R. LOUF, M. FUNTOWICZ, J. DAVISON, S. SHLEIFER, P. VON PLATEN, C. MA, Y. JERNITE, J. PLU, C. XU, T. L. SCAO, S. GUGGER, M. DRAME, Q. LHOEST and A. M. RUSH. 2019, «Huggingface’s transformers: State-of-the-art natural language processing», *ArXiv*, vol. abs/1910.03771. 135
- XIE, Q., Z. DAI, Y. DU, E. HOVY and G. NEUBIG. 2017, «Controllable invariance through adversarial feature learning», in *Advances in Neural Information Processing Systems*, p. 585–596. 130
- XU, B., N. WANG, T. CHEN and M. LI. 2015, «Empirical evaluation of rectified activations in convolutional network», *arXiv preprint arXiv:1505.00853*. 134, 135
- XU, R., T. GE and F. WEI. 2019, «Formality style transfer with hybrid textual annotations», *arXiv preprint arXiv:1903.06353*. 130
- YAMSHCHIKOV, I. P., V. SHIBAEV, A. NAGAEV, J. JOST and A. TIKHONOV. 2019, «Decomposing textual information for style transfer», *arXiv preprint arXiv:1909.12928*. 128, 130, 137
- ZAFAR, M. B., I. VALERA, M. GOMEZ RODRIGUEZ and K. P. GUMMADI. 2017, «Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment», in *Proceedings of the 26th international conference on world wide web*, p. 1171–1180. 130
- ZEMEL, R., Y. WU, K. SWERSKY, T. PITASSI and C. DWORK. 2013, «Learning fair representations», PMLR, Atlanta, Georgia, USA, p. 325–333. URL <http://proceedings.mlr.press/v28/zemel13.html>. 130
- ZHANG, Y., N. DING and R. SORICUT. 2018, «Shaped: Shared-private encoder-decoder for text style adaptation», *arXiv preprint arXiv:1804.04093*. 130
- ZHANG, Y., T. GE and X. SUN. 2020, «Parallel data augmentation for formality style transfer», *arXiv preprint arXiv:2005.07522*. 130



### **Part II Conclusions**

In this part, we studied the application of measures of information to two NLG problems: textual style transfer and NLG evaluation. In [Chapter 7](#), we provided a new upper bound on mutual information and use it to disentangled representations. Our experiments include both binary and multi-class style transfer. Our experiments show that our novel upper bound allow fine control over the degree of disentanglement. In [Chapter 8](#), we showed how to leverage different information measures to evaluate text generations. Our new metric InfoLM is flexible and correlates well with human judgments on summarization and data2text generation.



# Chapter 8

## Automatic Text Generation Evaluation

### Chapter 8 Abstract

Assessing the quality of natural language generation (NLG) systems through human annotation is very expensive. In this chapter, we show how to use the measures of information to measure the similarity between two sentences and assess their semantic equivalence. In practice, researchers rely on automatic metrics as a proxy of quality. In the last decade, many string-based metrics (*e.g.*, BLEU or ROUGE) have been introduced. More precisely, we introduce InfoLM a family of untrained metrics that can be viewed as a string-based metric and used different family of measures of information as well as a pre-trained masked language model. The use of information measures allowing the possibility to adapt InfoLM to different evaluation criteria. We apply these different information measures and demonstrate that InfoLM achieves statistically significant improvement in many configurations than existing metrics on both summarization and data2text generation.

### 8.1 Context

A plethora of applications of natural language processing (NLP) perform text-to-text transformations [BELZ and REITER, 2006; MELLISH and DALE, 1998; SPECIA and collab., 2018] that is, given a text, these systems are required to produce a text that is coherent, readable and informative. Due to both high annotation costs and time requirements, researchers tend to rely on automatic evaluation to compare the output of such systems. Reference-based automatic evaluation relies on comparing a candidate text produced by the NLG system and one or multiple reference texts ('gold standard') created by a human annotator. Generic automatic evaluation of NLG is a huge challenge as it requires building a metric that evaluates semantic equivalence between a candidate and one or several gold-standard reference texts. However, the definition of semantic equivalence is task-specific: as an example, evaluation of text summarization focuses on content, coherence, grammatically, conciseness, and readability [MANI, 2001], whereas machine translation focuses on fidelity, fluency and adequacy of the translation [HOVY, 1999; WHITE and collab., 1994] and data2text generation [DUŠEK and collab., 2020; GARDENT and collab., 2017; TIAN and collab., 2019] considers criteria such as data coverage, correctness and text structure.

Automatic text evaluation is an active area of research and a plethora of metrics have been previously proposed. They fall into two categories: metrics that are trained to maximise their correlation using human annotation (*e.g.*, RUSE [SHIMANAKA and collab., 2018], BLANC [LITA and collab., 2005], BEER [STANOJEVIĆ and SIMA'AN, 2014], BLEND [MA and collab., 2017], Q-Metrics [NEMA and KHAPRA, 2018], SIMILE [WIETING and collab., 2019]) and untrained metrics (*e.g.*, BLEU [PAPINENI and collab., 2002], ROUGE [LIN, 2004], BERTSCORE [ZHANG and collab., 2019a], Word Mover Distance [KUSNER and collab., 2015]). In this work, we focus on untrained metrics as they do not require costly training<sup>1</sup>. Two categories of untrained metrics can be distinguished: word or character based-metrics that compute a score based on string representation of the texts and embedding-based metrics that rely on a continuous representation of the text. String-based metrics (*e.g.*, BLEU, METEOR) often fail to robustly match paraphrases [REITER and BELZ, 2009] as they mainly focus on the surface form (*e.g.*, string representation of the metric) as opposed to embedding-based metrics that leverage continuous representations.

In this paper, we introduce InfoLM a family of new untrained metrics to evaluate text summarization and data2text generation. At the highest level InfoLM key components include: (1) a pre-trained language model that is used to compute two probability distributions  $p_r$  and  $p_c$ . They represent the probability of each token in the vocabulary to appear in each place of the reference and candidate text respectively; (2) a contrast function  $\mathcal{J}$  that computes the similarity between  $p_r$  and  $p_c$ . InfoLM relies on statistics on uni-gram, thus can be seen as belonging to the category of string-based metrics. However, contrarily to the existing string-based metric, the pre-trained language model allows InfoLM to assign a high score to paraphrases, capture distant dependencies and do not penalise semantically critical order changes.

### 8.1.1 Our contributions

Our main contributions are summarised below:

- *A set of novel metrics to automatically evaluate summarization and data2text generation.* To overcome the common pitfall of string matching metrics we introduce InfoLM. InfoLM combine a pre-trained model and a contrast function denoted by  $\mathcal{J}$  between two probability distributions. We explore the use of different choices of contrast functions such as  $f$ -divergences (one of the many generalizations of the Kullback Leibler divergence),  $\mathcal{L}_p$  distances or Fisher-Rao distances.
- *Tasks.* First, we demonstrate on both summarization and data2text that InfoLM is better suited than a wide set of concurrent metrics. A rigorous comparison is conducted, using multiple correlation measures with human judgement both at the text and system level. Second, we dissect our best performing metric to better understand the relative importance of each component. Last, the various performance of different  $\mathcal{C}_f$  allows us to gain valuable linguistic insights on how to build better metrics.

<sup>1</sup>Existing labelled corpora are of small size thus trained metrics may not generalize well to new data.

## 8.2 InfoLM

In this section, we first introduce a novel family of metrics called InfoLM and then detail the different components of these novel metrics, namely the pre-trained language model and the different information measures.

### 8.2.1 General overview

The task of evaluating the similarity between a candidate text  $\tilde{\mathbf{x}}_i^s$  and a reference text  $\mathbf{x}_i$  can be seen as measuring the similarity between the probability of observing each words of the vocabulary given the observation of the candidate text, denoted by  $p_{\Omega|\mathbf{X}}(\cdot|\tilde{\mathbf{x}}_i^s)$ , and the probability of observing each words given the observation of the reference text, denoted by  $p_{\Omega|\mathbf{X}}(\cdot|\mathbf{x}_i)$ . Formally, InfoLM is defined as follows:

$$\text{InfoLM}(\mathbf{x}_i, \tilde{\mathbf{x}}_i^s) \triangleq \mathcal{J}[p_{\Omega|\mathbf{X}}(\cdot|\tilde{\mathbf{x}}_i^s), p_{\Omega|\mathbf{X}}(\cdot|\mathbf{x}_i)], \quad (8.1)$$

where  $\mathcal{J} : [0, 1]^{| \Omega |} \times [0, 1]^{| \Omega |} \rightarrow \mathbb{R}_{\geq 0}$  is an information measure (see ??) and quantifies the similarity between  $p_{\Omega|\mathbf{X}}(\cdot|\tilde{\mathbf{x}}_i^s)$  and  $p_{\Omega|\mathbf{X}}(\cdot|\mathbf{x}_i)$ .

### 8.2.2 Computing the conditional distributions

To compute the conditional probability  $p_{\Omega|\mathbf{X}}(\omega|\tilde{\mathbf{x}})$  and  $p_{\Omega|\mathbf{X}}(\omega|\mathbf{x})$  we adopt an approach based on bags of distributions and rely on a pre-trained language model.

**Masked language modelling.** Language models based on masked language pre-training objectives aim at reconstructing a corrupt version  $\tilde{\mathbf{x}}$  of an input text  $\mathbf{x}$ . Formally, during training, the LM minimizes the following loss:

$$\mathcal{L} \triangleq \mathbb{E}_{\mathbf{M}} \left[ \sum_{k \in \mathbf{M}} p_{\Omega|\mathbf{X}}(\omega_k|\tilde{\mathbf{x}}; \theta) \right], \quad (8.2)$$

where  $\mathbf{M} \in \{1, \dots, M\}$  denotes a random vector indicating the selected set of masked positions and  $p_{\Omega|\mathbf{X}}(\cdot|\tilde{\mathbf{x}}; \theta)$  represents the output of the softmax layer of the considered LM.

The current state of the art masked language models is BERT and its improvements [DEVLIN and collab. \[2018\]](#); [LAN and collab. \[2019\]](#); [LIU and collab. \[2019\]](#); [ZHANG and collab. \[2019b\]](#). Other alternatives exist such as auto-regressive models [[BROWN and collab., 2020](#); [RADFORD and collab., 2018, 2019](#)] but they do not use bidirectional context. To provide a fair comparison with previous work [[ZHANG and collab., 2019a](#); [ZHAO and collab., 2019](#)] we choose not to work with XLNet [[YANG and collab., 2019](#)].

**Computing  $p_{\Omega|\mathbf{X}}(\cdot|\tilde{\mathbf{x}})$  and  $p_{\Omega|\mathbf{X}}(\cdot|\mathbf{x})$ .** Given a pre-trained masked language model  $p_\theta$  where  $\theta \in \Theta$  is fixed,  $p_{\Omega|\mathbf{X}}(\omega|\tilde{\mathbf{x}})$  and  $p_{\Omega|\mathbf{X}}(\omega|\mathbf{x})$  are respectively defined as follows:

$$p_{\Omega|\mathbf{X}}(\omega|\mathbf{x}; \theta) \triangleq \frac{1}{M} \sum_{k=1}^M p_\theta(\omega|\omega_1, \dots, \star_k, \dots, \omega_M),$$

$$p_{\Omega|\mathbf{X}}(\omega|\tilde{\mathbf{x}}; \theta) \triangleq \frac{1}{L} \sum_{k=1}^L p_\theta(\omega|\tilde{\omega}_1, \dots, \star_k, \dots, \tilde{\omega}_L),$$

where  $\star_k$  denotes a mask at the  $k$ -th position.

**Importance weighting.** It has been shown in the score of a similarity measure that

rare words can be more indicative of text similarity than common words [BANERJEE and LAVIE, 2005; VEDANTAM and collab., 2015]. The previous formula is flexible enough to allow us to weigh each distribution. Thus,  $p_{\Omega|X}(\cdot|\tilde{\mathbf{x}};\theta)$  and  $p_{\Omega|X}(\cdot|\mathbf{x};\theta)$  write as:

$$p_{\Omega|X}(\omega|\mathbf{x};\theta) = \frac{1}{M} \sum_{k=1}^M \gamma_k f_{\theta}(\omega|\omega_1, \dots, \star_k, \dots, \omega_M),$$

$$p_{\Omega|X}(\omega|\tilde{\mathbf{x}};\theta) = \frac{1}{L} \sum_{k=1}^N \tilde{\gamma}_k f_{\theta}(\omega|\tilde{\omega}_1, \dots, \star_k, \dots, \tilde{\omega}_L),$$

where  $\tilde{\gamma}_k$  and  $\gamma_k$  are normalized measures of the importance of the  $k$ -th word in the candidate and reference text, respectively, i.e., satisfying  $\sum_{j=1}^M \gamma_j = \sum_{j=1}^M \tilde{\gamma}_j = 1$ . These are computed using the idf scores determined at the corpus level [KUSNER and collab., 2015; ZHANG and collab., 2019a; ZHAO and collab., 2019] such as done in the current literature.

**LM Calibration.** Modern deep neural networks are known to be overconfident [GUO and collab., 2017] especially when they are deep. Several techniques have been proposed to alleviate the calibration problem relying on entropy rates [BRAVERMAN and collab., 2020], temperature [PLATT and collab., 1999], joint energy based training [HE and collab., 2021] or contextual calibration procedures scaling [ZHAO and collab., 2021]. In this work, we choose to study how calibration affects Inf oLM by relying on temperature scaling motivated by simplicity and speed.

## 8.3 Experimental Frameworks

In this section, we describe our experimental setting. We present the tasks and the baselines metrics use for each task.

### 8.3.1 Text summarization

The goal of text summarization is to compress long texts or documents into fluent, short sentences that preserve the most salient information.

**Datasets.** To compare the different metrics previous work [BHANDARI and collab., 2020; FABBRI and collab., 2020] either relies on the TAC datasets [DANG and OW CZARZAK, 2008; MCNAMEE and DANG, 2009] or on new summarization datasets extracted from CNN/DailyMail [HERMANN and collab., 2015; NALLAPATI and collab., 2016]. As pointed out by BHANDARI and collab. [2020]; PEYRARD [2019]; RANKEL and collab. [2013] TAC datasets are old and contain flaws (*e.g* systems used to generate summaries were of poor quality, human judgement when looking at the best summary only), we choose to work with the newly assemble dataset from CNN/Daily News proposed in BHANDARI and collab. [2020]. This dataset gathers 11,490 summaries coming from 14 abstractive systems [DONG and collab., 2019; KEDZIE and collab., 2018; LIU and LAPATA, 2019; NARAYAN and collab., 2018; WANG and collab., 2020; ZHONG and collab., 2020, 2019; ZHOU and collab., 2018,?] and 11 extractive systems [CHEN and BANSAL, 2018; DONG and collab., 2019; GEHRMANN and collab., 2018; LEWIS and collab., 2019; LIU and LAPATA, 2019; RAFFEL and collab., 2019; SEE and collab., 2017; YOON and collab., 2020].

**Annotations.** This dataset is annotated using the pyramid method [NENKOVA and collab., 2007; NENKOVA and PASSONNEAU, 2004; SHAPIRA and collab., 2019].

**Metrics.** For text summarization different metrics have been proposed. The most known metrics are string-based metrics and metrics based on ROUGE [LIN, 2004] and its extensions [GANESAN, 2018; NG and ABRECHT, 2015; SHAFIEI BAVANI and collab., 2018], or METEOR [BANERJEE and LAVIE, 2005] and its variants [DENKOWSKI and LAVIE, 2014; GUO and HU, 2019]. Recently, a new set of metrics (e.g BertScore [ZHANG and collab., 2019a], MoverScore [ZHAO and collab., 2019]) have been introduced for text summarization.

### 8.3.2 Data2Text generation

Prior work mainly rely on two task-oriented dialogue datasets (*i.e.*, BAGEL [MAIRESSE and collab., 2010], SFHOTEL [WEN and collab., 2015]). As sentence generated in these data-sets are unlikely to be representative of the progress of recent NLG systems we instead rely on a different dataset coming from the WebNLG2020 [GARDENT and collab., 2017; PEREZ-BELTRACHINI and collab., 2016] challenge<sup>2</sup> which consists in mapping data to text. The task consist in producing text from a set of triples in RDF language<sup>3</sup> extracted from DBpedia [AUER and collab., 2007].

Given the following example of triple (see bellow): (John\_Blaha birthDate 1942-08-26) (John\_Blaha birthPlace San\_Antonio) (John\_E\_Blaha job Pilot) the goal is to generate John Blaha, born in San Antonio on 1942-08-26, worked as a pilot. The goal of the WebNLG challenge is to develop efficient Knowledge Base Verbalizers [GARDENT and collab., 2017] (*i.e* generation algorithms that can verbalise knowledge base fragments) and thus handle complex interaction that can occur during the micro-planning phase when generating sentences [FERREIRA and collab., 2018]. Details on the WebNLG2020 task are given in FERREIRA and collab. [2020]<sup>4</sup>. The dataset is composed of generated sentences coming from 15 different systems using various approaches such as symbolic approaches or neural-based systems.

**Annotations.** The WebNLG task is evaluated by human annotators along four different axes:

- Data Coverage: Are all the description presented in the data included in the text?
- Relevance: Does the text contains only predicates found in the data?
- Correctness: Are predicates found in the data correctly mentioned and adequately introduced?
- Text structure: Is the produced text well-structured, grammatically correct and written in acceptable English?
- Fluency: Does the text progress naturally? Is it easy to understand? Is it a coherent whole?

<sup>2</sup>All data and system performance can be found in <https://webnlg-challenge.loria.fr/>

<sup>3</sup>RDF format is a widely used format for many dataset such as LinkedGeoData <http://linkedgeodata.org/About>, FOAF <http://www.foaf-project.org/> or MusicBrainz <https://musicbrainz.org/>.

<sup>4</sup>All data are freely available on github [https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release\\_v3.0](https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release_v3.0)

**Metrics.** For this task, organisers rely on untrained metrics such as BLEU, METEOR, BERTScore, TER, CHRF++ to compare the performance of the different systems. Thus, we will focus on system-level correlation.

## 8.4 Numerical Results on Summarization

In this section, we study the performance of InfoLM on text summarization. We first start by comparing the correlation between different metrics and human judgement (as measure following the pyramid method), we then study the correlation matrix between the different metrics to better understand the difference of score between InfoLM and other standard metrics (*e.g.*, BertScore, MoverScore, BLEU, Rouge). Finally, we study how changing the language model calibration affects the metric performance and how robust is  $\mathcal{D}_{AB}$  to the choice of  $\alpha$  and  $\beta$ .

### 8.4.1 Correlation analysis with human judgement

**General Analysis.** Figure 8.1 gathers the results of the correlation study between score output by different metrics and human judgement. We are able to reproduce results from BHANDARI and collab. [2020]. Similarly to other metric correlation between human judgement and InfoLM depends on the type of system to be evaluated (*e.g.*, abstractive or extractive) and the considered correlation level (*e.g.* text or system level). However, we can observe that InfoLM is in any case among the top-scoring metrics. By comparing  $\mathcal{D}_{AB}$  with other BERT-based metrics such as MoverScore or BertScore perform poorly at the text or system-level when considering output from extractive systems.  $\mathcal{D}_{AB}$  largely outperforms n-gram matching metrics (*e.g.*, ROUGE based metrics) on all datasets when measuring correlation with the Kendall  $\tau$  and in almost all configuration (except when considering abstractive outputs at the systems level) when using the Person  $r$ .

**Choice of information geometric measure for InfoLM.** In Figure 8.1, we can observe three different types of groups depending on the global behaviour. First we can notice that using  $\mathcal{L}_p$ ,  $p \in \{1, 2, \dots, +\infty\}$  lead to poor performances in many configurations<sup>5</sup>. The second group gathers JS,  $\mathcal{D}_\alpha$ ,  $\mathcal{D}_\beta$  and  $\mathcal{D}_{AB}$  and achieves the best performance overall.  $\mathcal{D}_\alpha$  and  $\mathcal{D}_{AB}$  achieve similar performances suggesting that the flexibility (*e.g.*, robustness to outliers) introduced by the  $\beta$  parameter in  $\mathcal{D}_{AB}$  is not useful in our task. This phenomenon is strengthened by the lower performance of  $\mathcal{D}_\beta$ . JS can be seen as a special case of  $\mathcal{D}_\alpha$  (corresponding to  $\alpha$ ). The difference of results between the two measures is due to the flexibility introduced by  $\alpha$  (*i.e.*,  $\alpha$  controls the relative importance of the ration  $\frac{p_i}{q_i}$ ) which can be interpreted in our case as the ability to control the importance attributed to less likely words according to the language model.

**Takeaways.** The best performing metric is obtained with  $\mathcal{D}_{AB}$ . The Fisher-Rao distance, denoted by  $\mathcal{R}$ , achieves good performance in many scenarios and has the advantage to be parameter-free. In this experiment, the temperature of the LM has been set to 1 for all metrics.

<sup>5</sup>Performance of  $\mathcal{L}_\infty$  in some configurations is surprising as  $\mathcal{L}_\infty$  is extremely selective as it outputs  $\max_i |p_i - q_i|$ . Considering the size of the vocabulary and the sparsity of the output of the LM, one could expect that the LM will output the probability of the word that is most likely in one sentence and not likely at all in the other. This appears to be a good heuristic to measure the performance of some abstractive systems.



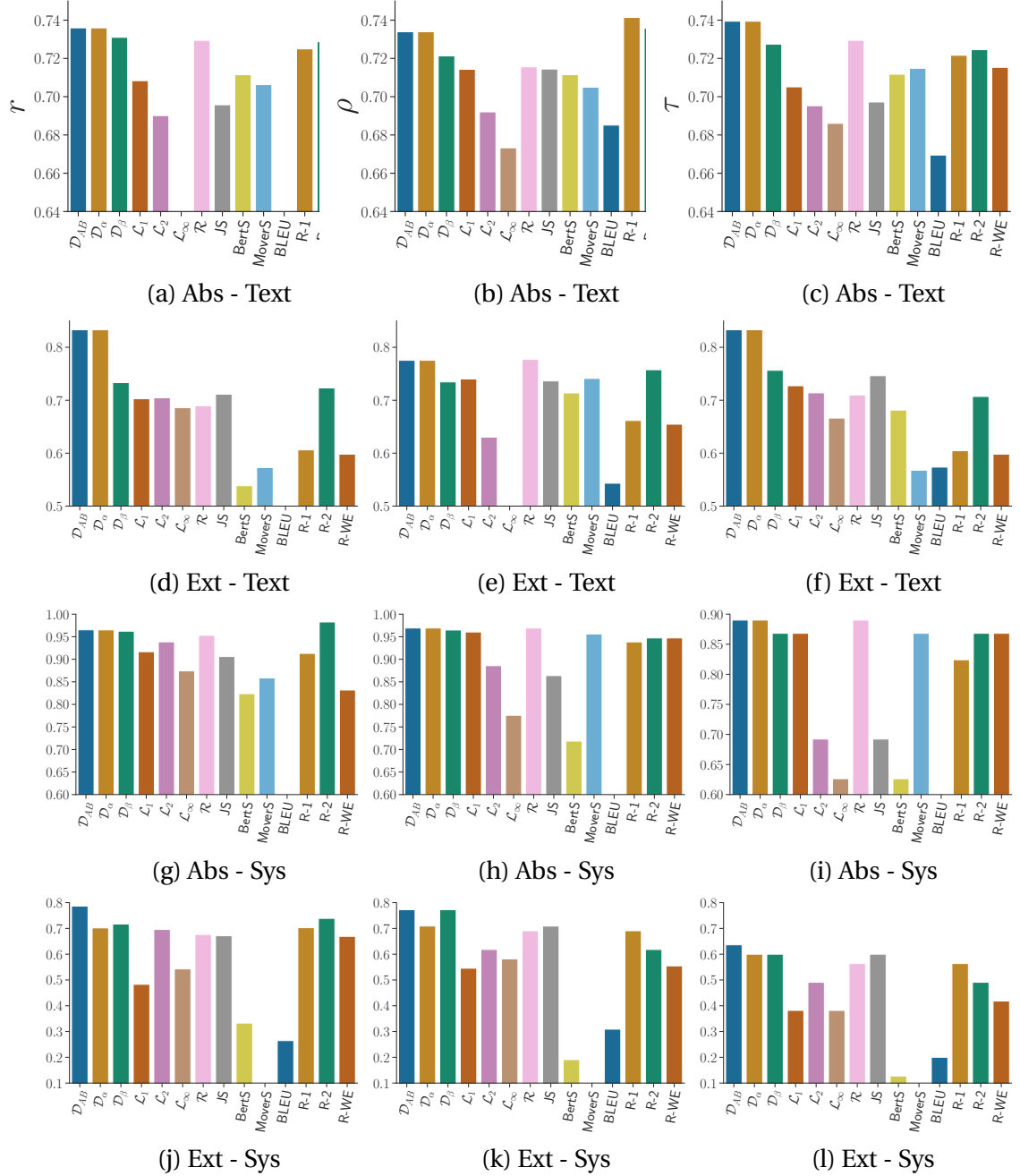


Figure 8.1 – Results of the correlation between metrics and human judgement on the CNN dataset. First column reports the report correlations as measured by the Person ( $r$ ), second column reports Spearman ( $\rho$ ) and third column reports Kendall ( $\tau$ ) coefficient.

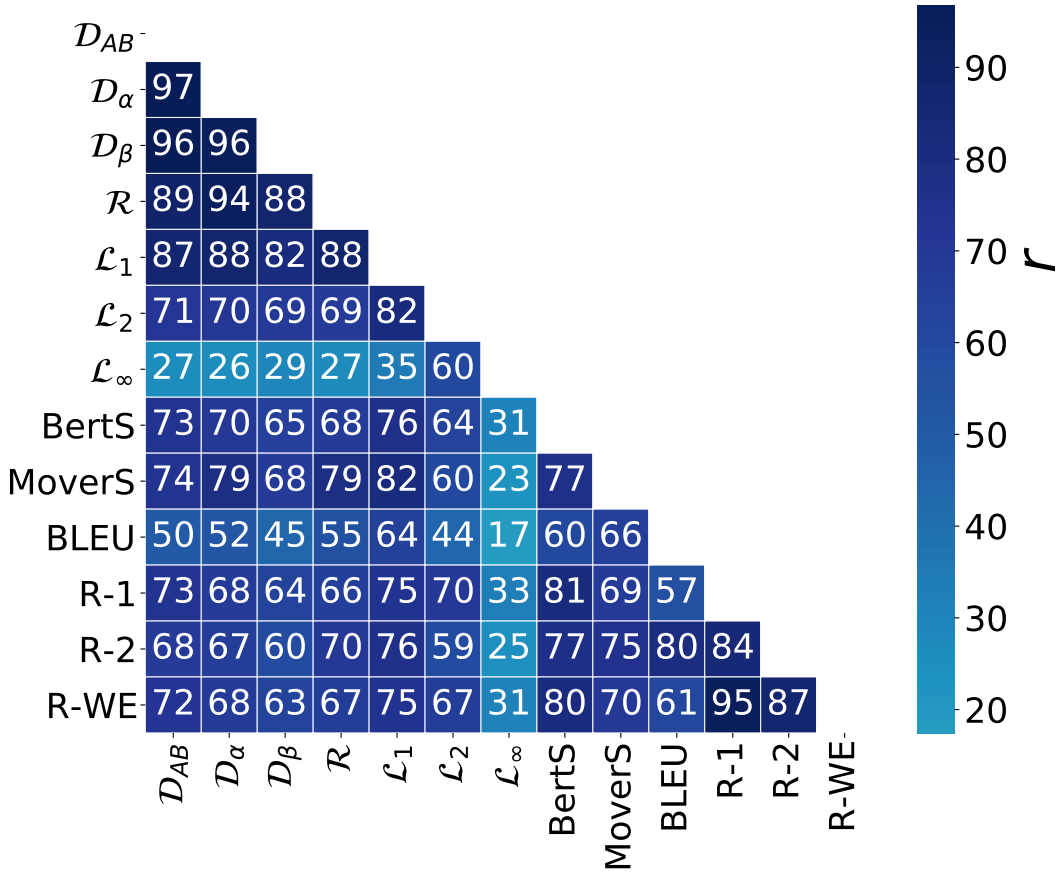


Figure 8.2 – Pearson correlation at the system level between metrics when considering abstractive system outputs.

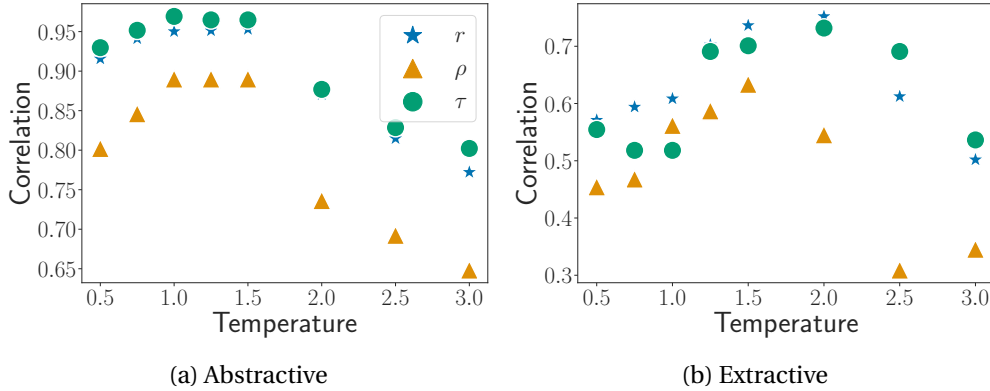


Figure 8.3 – Impact of Calibration on system-level correlation (with Pearson ( $r$ ), Spearman ( $\rho$ ) or Kendall ( $\tau$ )) for CNN. The chosen measure is Rao as it is parameter-free. Calibration is changed using temperature scaling.

### 8.4.2 Correlation analysis between metrics

In this experiment, we complete our global analysis by comparing the scores obtained by the different metrics with each other. We want to increase our understanding of how different our metric is from other metrics and how the choice of information geometric measure affects the predictions. Figure 8.2 gathers the results of the experiment. We observe a strong correlation ( $r > 88$ ) between  $\mathcal{D}_{\alpha}$ ,  $\mathcal{D}_{\beta}$ ,  $\mathcal{D}_{AB}$  and



$\mathcal{R}$ <sup>6</sup>. Interestingly, we observe that these groups are moderately correlated ( $r \approx 70$  with BERT score and n-gram matching metrics, e.g., ROUGE) whereas BERT score achieves a stronger correlation with ROUGE ( $r \approx 80$ ).

**Takeaways:** InfoLM metrics (using scores from  $\mathcal{D}_\alpha$ ,  $\mathcal{D}_\beta$ ,  $\mathcal{D}_{AB}$  and  $\mathcal{R}$ ) are moderately correlated with both n-gram matching metrics and BERT confirming that InfoLM metrics capture a different notion of similarity.

### 8.4.3 Role of calibration

Here we study the impact of calibration on system-level correlation as measured by the different correlation coefficients. For space constraint, we limit our study to the Fisher-Rao distance<sup>7</sup>, denoted by  $\mathcal{R}$ , as it is parameter-free. Figure 8.3 gathers results of the experiments. When changing the temperature we observe a smooth change in correlation.

**Takeaways:** Optimal temperature  $T$  seems to be reached for  $T \in [1, 2]$  which suggests that InfoLM benefits from a language model that is not too selective (case  $T \ll 1$ ). For specific application the temperature of InfoLM can be tuned to improve correlation.

### 8.4.4 Choice of $\alpha$ and $\beta$

In this experiment, we aim at quantifying the sensitivity of  $\mathcal{D}_{AB}$  to the choice of  $\alpha$  and  $\beta$ . Figure 8.4 gathers the results of the analysis. We observe that a change in  $\beta$  induces a stronger change in the metric. Additionally the lower  $\beta$  the better result we obtain. We can also note that the variation of both parameters is smooth. Interestingly for abstractive systems, a low value of  $\alpha$  should be chosen where for extractive higher is better. It suggests that for evaluating abstractive systems the metric should focus on low values of  $p_i/q_i$  (words that are probable for both candidate and reference text) where for extractive systems the attention should be focused on high values of  $p_i/q_i$  (words that are likely only in one text).

**Takeaways.** Low values of  $\beta$  leads to better results, optimal value of  $\alpha$  is 1.25 for abstractive and 3 for extractive. Good parameter combinations achieve consistently high performance when using different correlation coefficient.

<sup>6</sup>Note that these metrics consider the product of  $p_i$  and  $q_i$ .

<sup>7</sup>As calibration that will increase specific words probability Rao is particularly suitable to study the calibration because it only considers product  $p_i \times q_i$ .

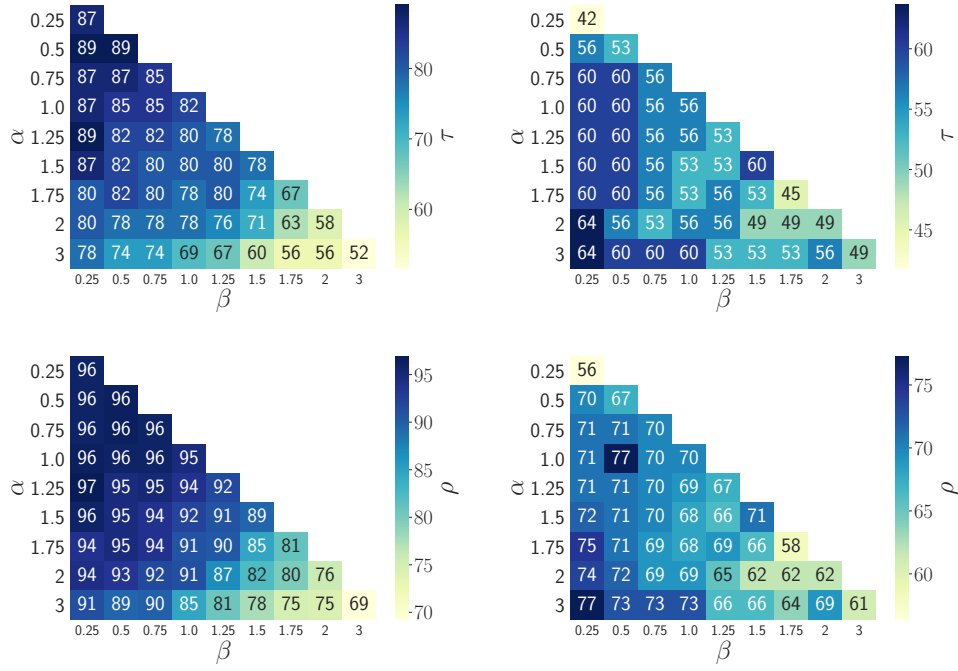


Figure 8.4 – Impact of change in  $\alpha$  and  $\beta$  for  $\mathcal{D}_{AB}$ . System level correlation, as measured by Pearson ( $r$ ) or Spearman ( $\rho$ ), is presented on abstractive (first column) and extractive system (second column).

## 8.5 Numerical Results on Data2Text

In this section, we evaluate our metric InfoLM on a data2text task. In this challenge, automatic metrics are used to compare different systems. We first present a correlation analysis with the human judgement that we complete with statistical analysis to answer the following question: is InfoLM significantly better than the metrics used for comparing systems?

### 8.5.1 Correlation analysis with human judgement

**Global Analysis:** Table 8.1 gathers results of the correlation analysis of the metric with human judgement following the five different axes. We observe that the five axes of annotation are not independent: text structure and fluency achieve a strong correlation coefficient ( $> 98$ ). Additionally, all metrics achieve similar results when the correlation is computed on these two criteria. We observe that the best performing group of metric is based on InfoLM followed by metrics based on continuous representation from BERT (*i.e.*, MoverScore and BertScore) followed by n-gram matching metrics. Regarding correctness, data coverage and relevance, we observe that both  $\mathcal{D}_{AB}$  and  $\mathcal{D}_{\alpha}$  achieve the best results along almost all correlation coefficients. On data coverage InfoLM achieves improvement up to 17 points in correlation compared to both Bert based or n-gram matching metrics. Regarding fluency and text structure, Fisher-Rao distance works better and slightly outperforms the second-best performing metric, namely Best Score.

**Choice of Information Geometry Measure.** Similar to summarisation, we observe very low correlation for  $L_p$ ,  $p \in \{1, 2, \dots, +\infty\}$ . We also observe that  $\beta$ -divergences achieve lower results than both  $\alpha$  and AB divergences suggesting that, as noticed for

Metric	Correctness			Data Coverage			Fluency			Relevance			Text Structure		
	$r$	$\rho$	$\tau$	$r$	$\rho$	$\tau$	$r$	$\rho$	$\tau$	$r$	$\rho$	$\tau$	$r$	$\rho$	$\tau$
Correct	100.0	100.0	100.0	97.6	85.2	73.3	80.0	81.1	61.6	99.1	89.7	75.0	80.1	80.8	60.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	71.8	51.7	38.3	96.0	93.8	81.6	71.6	51.4	36.6
Fluency	81.1	80.0	61.6	71.8	51.7	38.3	100.0	100.0	100.0	77.0	61.4	46.6	99.5	99.7	98.3
Relev	89.7	99.1	75.0	96.0	93.8	81.6	77.0	61.4	46.6	100.0	100.0	100.0	77.2	61.1	45.0
TextS	80.8	80.1	60.0	71.6	51.4	36.6	99.5	99.7	98.3	77.2	61.1	45.0	100.0	100.0	100.0
$\mathcal{D}_{AB}$	88.8	<b><u>89.3</u></b>	<b><u>76.6</u></b>	<b><u>81.8</u></b>	<b><u>82.6</u></b>	<b><u>70.0</u></b>	86.6	92.0	76.6	<b><u>89.8</u></b>	<b><u>87.9</u></b>	<b><u>73.3</u></b>	86.6	91.4	75.0
$\mathcal{D}_\alpha$	88.8	<b><u>89.3</u></b>	<b><u>76.6</u></b>	<b><u>81.8</u></b>	<b><u>82.6</u></b>	<b><u>70.0</u></b>	86.6	92.0	76.6	<b><u>89.8</u></b>	<b><u>87.9</u></b>	<b><u>73.3</u></b>	86.6	91.4	75.0
$\mathcal{D}_\beta$	81.4	50.0	71.6	48.4	79.7	65.0	44.8	84.7	76.6	49.3	72.3	60.0	48.0	83.8	75.0
$\mathcal{L}_1$	75.2	33.8	61.6	32.4	53.8	40.0	22.7	83.5	73.3	32.2	57.9	45.0	25.6	83.2	71.6
$\mathcal{L}_2$	67.0	21.9	56.6	21.6	37.9	33.3	11.9	75.2	58.3	20.1	43.8	38.3	14.8	75.5	60.0
$\mathcal{L}_\infty$	63.2	33.0	46.6	30.4	36.4	26.6	67.6	65.0	46.6	29.1	49.1	35.0	67.2	65.2	46.6
$\mathcal{R}$	<b><u>89.7</u></b>	86.0	75.0	78.7	70.5	51.6	<b><u>93.3</u></b>	<b><u>95.7</u></b>	<b><u>85.3</u></b>	87.6	84.4	70.0	<b><u>92.4</u></b>	93.8	<b><u>81.6</u></b>
JS	79.4	81.1	70.0	69.3	75.5	60.0	89.4	91.4	75.0	81.7	70.5	60.0	91.9	91.1	73.3
BertS	85.5	83.4	73.3	74.7	68.2	53.3	92.3	95.5	85.0	83.3	79.4	65.0	91.9	<b><u>95.0</u></b>	<b><u>83.3</u></b>
MoverS	84.1	<b><u>84.1</u></b>	<b><u>73.3</u></b>	<b><u>78.7</u></b>	66.2	53.3	91.2	92.1	78.3	82.1	77.4	65.0	90.1	91.4	76.3
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	<b><u>89.4</u></b>	90.5	78.3	63.0	65.2	51.6	88.5	89.1	76.6
R-1	80.6	65.0	65.0	61.1	<b><u>59.6</u></b>	<b><u>48.3</u></b>	76.5	76.3	60.3	64.3	<b><u>69.2</u></b>	56.7	75.9	77.5	58.3
R-2	73.6	63.3	58.3	54.7	43.1	35.0	86.4	81.9	63.4	62.0	60.8	46.7	86.5	80.5	61.7
R-WE	60.9	73.4	60.0	40.2	58.2	40.1	61.4	84.7	61.3	49.9	64.1	48.3	60.2	85.9	60.0
METEOR	<b><u>86.5</u></b>	<b><u>66.3</u></b>	<b><u>70.0</u></b>	<b><u>77.3</u></b>	50.2	46.6	86.7	90.5	78.3	<b><u>82.1</u></b>	65.2	58.6	86.2	89.1	76.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	89.1	<b><u>93.5</u></b>	<b><u>80.0</u></b>	75.0	70.2	<b><u>77.6</u></b>	<b><u>89.5</u></b>	<b><u>91.1</u></b>	<b><u>78.6</u></b>

Table 8.1 – Correlation at the system level with human judgement along five different axis: correctness, data coverage, fluency, relevance and text structure for the WebNLG task. Best results by group are underlined, overall best results are bolted.

summarisation, robustness to unlikely words (*i.e.*, outliers) is less relevant for the task.

### 8.5.2 Statistical analysis

Automatic metrics are used in the WebNLG challenge to compare the different systems. In order to evaluate whether observed improvement in correlation is significant, we report the results of William’s Significance test in Figure 8.5.

**Takeaways:** (i) Regarding correctness and relevance  $\mathcal{D}_{AB}$ , is a suitable choice that is significantly better than other metrics; (ii) Regarding text structure,  $\mathcal{R}$  is significantly better and compare favourably against all metrics except MoverScore for automatic fluency evaluation; (iii) Regarding data coverage, METEOR achieves good result however significance difference is only observed with BertScore.

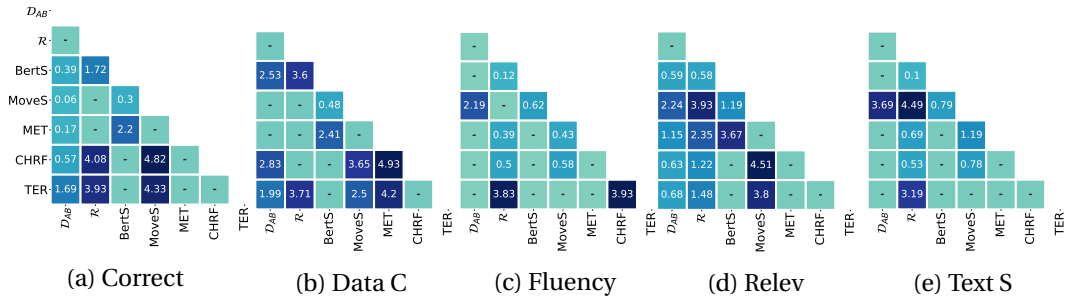


Figure 8.5 – Results of William’s Significance Test: the tested hypothesis is: “is the increase of correlation significant”. For clarity and due to space constraints the p-values are truncated and multiply par 100. Only p-values that are lower than 5.00 are displayed.

### Chapter 8 Conclusion

In this chapter, we introduced InfoLM that is a novel family of metrics motivated by the use of a measure of information and based on pre-trained language models. InfoLM does not require training and it is among the first metrics computing the similarity between two probability distributions over the vocabulary (which is similar to string-based metrics) but also leverages the recent advance in contextualised embedding thanks to the deep language model. Our experiments on both summarization and data2text generation demonstrate the validity of our approach. Among available contrast measures, the Fisher-Rao distance is parameter-free and thus, it is easy to use in practice while the AB-Divergence achieves better results but requires to choose  $\alpha$  and  $\beta$ . The performances of AB-Divergence when varying the values of  $\alpha$  and  $\beta$  could also help to gain intuition on the role of rare and frequent words and drive future research to design better metrics or new training losses.

## 8.6 References

- AUER, S., C. BIZER, G. KOBILAROV, J. LEHMANN, R. CYGANIAK and Z. IVES. 2007, «Dbpedia: A nucleus for a web of open data», in *The semantic web*, Springer, p. 722–735. 159
- BANERJEE, S. and A. LAVIE. 2005, «Meteor: An automatic metric for mt evaluation with improved correlation with human judgments», in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, p. 65–72. 158, 159
- BELZ, A. and E. REITER. 2006, «Comparing automatic and human evaluation of nlg systems», in *11th Conference of the European Chapter of the Association for Computational Linguistics*. 155
- BHANDARI, M., P. GOUR, A. ASHFAQ, P. LIU and G. NEUBIG. 2020, «Re-evaluating evaluation in text summarization», *arXiv preprint arXiv:2010.07100*. 158, 160
- BRAVERMAN, M., X. CHEN, S. KAKADE, K. NARASIMHAN, C. ZHANG and Y. ZHANG. 2020, «Calibration, entropy rates, and memory in language models», in *Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol. 119, édité par H. D. III and A. Singh, PMLR, p. 1089–1099. URL <http://proceedings.mlr.press/v119/braverman20a.html>. 158
- BROWN, T. B., B. MANN, N. RYDER, M. SUBBIAH, J. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, P. SHYAM, G. SASTRY, A. ASKELL and collab.. 2020, «Language models are few-shot learners», *arXiv preprint arXiv:2005.14165*. 157
- CHEN, Y.-C. and M. BANSAL. 2018, «Fast abstractive summarization with reinforce-selected sentence rewriting», *arXiv preprint arXiv:1805.11080*. 158
- DANG, H. T. and K. OWCZARZAK. 2008, «Overview of the tac 2008 update summarization task.», in *TAC*. 158

- DENKOWSKI, M. and A. LAVIE. 2014, «Meteor universal: Language specific translation evaluation for any target language», in *Proceedings of the ninth workshop on statistical machine translation*, p. 376–380. 159
- DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805*. 157
- DONG, L., N. YANG, W. WANG, F. WEI, X. LIU, Y. WANG, J. GAO, M. ZHOU and H.-W. HON. 2019, «Unified language model pre-training for natural language understanding and generation», *arXiv preprint arXiv:1905.03197*. 158
- DUŠEK, O., J. NOVIKOVA and V. RIESER. 2020, «Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge», *Computer Speech & Language*, vol. 59, p. 123–156. 155
- FABBRI, A. R., W. KRYŚCIŃSKI, B. MCCANN, C. XIONG, R. SOCHER and D. RADEV. 2020, «Summeval: Re-evaluating summarization evaluation», *arXiv preprint arXiv:2007.12626*. 158
- FERREIRA, T., C. GARDENT, N. ILINYKH, C. VAN DER LEE, S. MILLE, D. MOUSSALLEM and A. SHIMORINA. 2020, «The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020)», in *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*. 159
- FERREIRA, T. C., D. MOUSSALLEM, E. KRAHMER and S. WUBBEN. 2018, «Enriching the webnlg corpus», in *Proceedings of the 11th International Conference on Natural Language Generation*, p. 171–176. 159
- GANESAN, K. 2018, «Rouge 2.0: Updated and improved measures for evaluation of summarization tasks», *arXiv preprint arXiv:1803.01937*. 159
- GARDENT, C., A. SHIMORINA, S. NARAYAN and L. PEREZ-BELTRACHINI. 2017, «Creating training corpora for nlg micro-planning», in *55th annual meeting of the Association for Computational Linguistics (ACL)*. 155, 159
- GEHRMANN, S., Y. DENG and A. RUSH. 2018, «Bottom-up abstractive summarization», in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, p. 4098–4109, doi: 10.18653/v1/D18-1443. URL <https://www.aclweb.org/anthology/D18-1443>. 158
- GUO, C., G. PLEISS, Y. SUN and K. Q. WEINBERGER. 2017, «On calibration of modern neural networks», in *International Conference on Machine Learning*, PMLR, p. 1321–1330. 158
- GUO, Y. and J. HU. 2019, «Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation», in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Association for Computational Linguistics, Florence, Italy, p. 501–506, doi: 10.18653/v1/W19-5357. URL <https://www.aclweb.org/anthology/W19-5357>. 159

- HE, T., B. MCCANN, C. XIONG and E. HOSSEINI-ASL. 2021, «Joint energy-based model training for better calibrated natural language understanding models», *arXiv preprint arXiv:2101.06829*. 158
- HERMANN, K. M., T. KOČISKÝ, E. GREFFENSTETTE, L. ESPEHOLT, W. KAY, M. SULEYMAN and P. BLUNSOM. 2015, «Teaching machines to read and comprehend», *arXiv preprint arXiv:1506.03340*. 158
- HOVY, E. H. 1999, «Toward finely differentiated evaluation metrics for machine translation», in *Proceedings of the EAGLES Workshop on Standards and Evaluation Pisa, Italy, 1999*. 155
- KEDZIE, C., K. MCKEOWN and H. DAUME III. 2018, «Content selection in deep learning models of summarization», *arXiv preprint arXiv:1810.12343*. 158
- KUSNER, M., Y. SUN, N. KOLKIN and K. WEINBERGER. 2015, «From word embeddings to document distances», in *International conference on machine learning*, PMLR, p. 957–966. 156, 158
- LAN, Z., M. CHEN, S. GOODMAN, K. GIMPEL, P. SHARMA and R. SORICUT. 2019, «Albert: A lite bert for self-supervised learning of language representations», *arXiv preprint arXiv:1909.11942*. 157
- LEWIS, M., Y. LIU, N. GOYAL, M. GHAZVININEJAD, A. MOHAMED, O. LEVY, V. STOYANOV and L. ZETTLEMOYER. 2019, «Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension», *arXiv preprint arXiv:1910.13461*. 158
- LIN, C.-Y. 2004, «ROUGE: A package for automatic evaluation of summaries», in *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, p. 74–81. URL <https://www.aclweb.org/anthology/W04-1013>. 156, 159
- LITA, L. V., M. ROGATI and A. LAVIE. 2005, «Blanc: Learning evaluation metrics for mt», in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 740–747. 156
- LIU, Y. and M. LAPATA. 2019, «Text summarization with pretrained encoders», *arXiv preprint arXiv:1908.08345*. 158
- LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER and V. STOYANOV. 2019, «Roberta: A robustly optimized bert pretraining approach», *arXiv preprint arXiv:1907.11692*. 157
- MA, Q., Y. GRAHAM, S. WANG and Q. LIU. 2017, «Blend: a novel combined mt metric based on direct assessment—casict-dcu submission to wmt17 metrics task», in *Proceedings of the second conference on machine translation*, p. 598–603. 156
- MAIRESSE, F., M. GASIC, F. JURCICEK, S. KEIZER, B. THOMSON, K. YU and S. YOUNG. 2010, «Phrase-based statistical language generation using graphical models and active learning», in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1552–1561. 159
- MANI, I. 2001, *Automatic summarization*, vol. 3, John Benjamins Publishing. 155



- MCNAMEE, P. and H. T. DANG. 2009, «Overview of the tac 2009 knowledge base population track», in *Text Analysis Conference (TAC)*, vol. 17, p. 111–113. 158
- MELLISH, C. and R. DALE. 1998, «Evaluation in the context of natural language generation», *Computer Speech & Language*, vol. 12, n° 4, p. 349–373. 155
- NALLAPATI, R., B. ZHOU, C. GULCEHRE, B. XIANG and collab.. 2016, «Abstractive text summarization using sequence-to-sequence rnns and beyond», *arXiv preprint arXiv:1602.06023*. 158
- NARAYAN, S., S. B. COHEN and M. LAPATA. 2018, «Ranking sentences for extractive summarization with reinforcement learning», *arXiv preprint arXiv:1802.08636*. 158
- NEMA, P. and M. M. KHAPRA. 2018, «Towards a better metric for evaluating question generation systems», *arXiv preprint arXiv:1808.10192*. 156
- NENKOVA, A., R. PASSONNEAU and K. MCKEOWN. 2007, «The pyramid method: Incorporating human content selection variation in summarization evaluation», *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, n° 2, p. 4–es. 158
- NENKOVA, A. and R. J. PASSONNEAU. 2004, «Evaluating content selection in summarization: The pyramid method», in *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, p. 145–152. 158
- NG, J.-P. and V. ABRECHT. 2015, «Better summarization evaluation with word embeddings for rouge», *arXiv preprint arXiv:1508.06034*. 159
- PAPINENI, K., S. ROUKOS, T. WARD and W.-J. ZHU. 2002, «Bleu: a method for automatic evaluation of machine translation», in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 311–318, doi: 10.3115/1073083.1073135. URL <https://www.aclweb.org/anthology/P02-1040>. 156
- PEREZ-BELTRACHINI, L., R. SAYED and C. GARDENT. 2016, «Building rdf content for data-to-text generation», in *The 26th International Conference on Computational Linguistics (COLING 2016)*. 159
- PEYRARD, M. 2019, «Studying summarization evaluation metrics in the appropriate scoring range», in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5093–5100. 158
- PLATT, J. and collab.. 1999, «Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods», *Advances in large margin classifiers*, vol. 10, n° 3, p. 61–74. 158
- RADFORD, A., K. NARASIMHAN, T. SALIMANS and I. SUTSKEVER. 2018, «Improving language understanding by generative pre-training», . 157
- RADFORD, A., J. WU, R. CHILD, D. LUAN, D. AMODEI and I. SUTSKEVER. 2019, «Language Models are Unsupervised Multitask Learners», *OpenAI Blog*, vol. 1, n° 8. URL <https://api.semanticscholar.org/CorpusID:160025533>. 157

- RAFFEL, C., N. SHAZEER, A. ROBERTS, K. LEE, S. NARANG, M. MATENA, Y. ZHOU, W. LI and P. J. LIU. 2019, «Exploring the limits of transfer learning with a unified text-to-text transformer», *arXiv preprint arXiv:1910.10683*. 158
- RANKEL, P. A., J. CONROY, H. T. DANG and A. NENKOVA. 2013, «A decade of automatic content evaluation of news summaries: Reassessing the state of the art», in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 131–136. 158
- REITER, E. and A. BELZ. 2009, «An investigation into the validity of some metrics for automatically evaluating natural language generation systems», *Computational Linguistics*, vol. 35, n° 4, p. 529–558. 156
- SEE, A., P. J. LIU and C. D. MANNING. 2017, «Get to the point: Summarization with pointer-generator networks», *arXiv preprint arXiv:1704.04368*. 158
- SHAFIEIBAVANI, E., M. EBRAHIMI, R. WONG and F. CHEN. 2018, «A graph-theoretic summary evaluation for rouge», in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 762–767. 159
- SHAPIRA, O., D. GABAY, Y. GAO, H. RONEN, R. PASUNURU, M. BANSAL, Y. AMSTERDAMER and I. DAGAN. 2019, «Crowdsourcing lightweight pyramids for manual summary evaluation», *arXiv preprint arXiv:1904.05929*. 158
- SHIMANAKA, H., T. KAJIWARA and M. KOMACHI. 2018, «Ruse: Regressor using sentence embeddings for automatic machine translation evaluation», in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, p. 751–758. 156
- SPECIA, L., C. SCARTON and G. H. PAETZOLD. 2018, «Quality estimation for machine translation», *Synthesis Lectures on Human Language Technologies*, vol. 11, n° 1, p. 1–162. 155
- STANOJEVIĆ, M. and K. SIMA'AN. 2014, «BEER: BEtter evaluation as ranking», in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Baltimore, Maryland, USA, p. 414–419, doi: 10.3115/v1/W14-3354. URL <https://www.aclweb.org/anthology/W14-3354>. 156
- TIAN, R., S. NARAYAN, T. SELLAM and A. P. PARIKH. 2019, «Sticking to the facts: Confident decoding for faithful data-to-text generation», *arXiv preprint arXiv:1910.08684*. 155
- VEDANTAM, R., C. LAWRENCE ZITNICK and D. PARIKH. 2015, «Cider: Consensus-based image description evaluation», in *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4566–4575. 158
- WANG, D., P. LIU, Y. ZHENG, X. QIU and X. HUANG. 2020, «Heterogeneous graph neural networks for extractive document summarization», *arXiv preprint arXiv:2004.12393*. 158
- WEN, T.-H., M. GASIC, N. MRKSIC, P.-H. SU, D. VANDYKE and S. YOUNG. 2015, «Semantically conditioned lstm-based natural language generation for spoken dialogue systems», *arXiv preprint arXiv:1508.01745*. 159



- WHITE, J. S., T. A. O'CONNELL and F. E. O'MARA. 1994, «The arpa mt evaluation methodologies: evolution, lessons, and future approaches», in *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. 155
- WIETING, J., T. BERG-KIRKPATRICK, K. GIMPEL and G. NEUBIG. 2019, «Beyond bleu: Training neural machine translation with semantic similarity», *arXiv preprint arXiv:1909.06694*. 156
- YANG, Z., Z. DAI, Y. YANG, J. CARBONELL, R. R. SALAKHUTDINOV and Q. V. LE. 2019, «Xlnet: Generalized autoregressive pretraining for language understanding», in *Advances in neural information processing systems*, p. 5754–5764. 157
- YOON, W., Y. S. YEO, M. JEONG, B.-J. YI and J. KANG. 2020, «Learning by semantic similarity makes abstractive summarization better», *arXiv preprint arXiv:2002.07767*. 158
- ZHANG, T., V. KISHORE, F. WU, K. Q. WEINBERGER and Y. ARTZI. 2019a, «Bertscore: Evaluating text generation with bert», *arXiv preprint arXiv:1904.09675*. 156, 157, 158, 159
- ZHANG, X., F. WEI and M. ZHOU. 2019b, «Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization», *arXiv preprint arXiv:1905.06566*. 157
- ZHAO, T. Z., E. WALLACE, S. FENG, D. KLEIN and S. SINGH. 2021, «Calibrate before use: Improving few-shot performance of language models», *arXiv preprint arXiv:2102.09690*. 158
- ZHAO, W., M. PEYRARD, F. LIU, Y. GAO, C. M. MEYER and S. EGER. 2019, «Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance», *arXiv preprint arXiv:1909.02622*. 157, 158, 159
- ZHONG, M., P. LIU, Y. CHEN, D. WANG, X. QIU and X. HUANG. 2020, «Extractive summarization as text matching», *arXiv preprint arXiv:2004.08795*. 158
- ZHONG, M., P. LIU, D. WANG, X. QIU and X. HUANG. 2019, «Searching for effective neural extractive summarization: What works and what's next», *arXiv preprint arXiv:1907.03491*. 158
- ZHOU, Q., N. YANG, F. WEI, S. HUANG, M. ZHOU and T. ZHAO. 2018, «Neural document summarization by jointly learning to score and select sentences», *arXiv preprint arXiv:1807.02305*. 158



## Chapter 9

# Conclusions, Limitations and Future Work

### 9.1 Conclusions

This manuscript presents the research conducted over the last three years on integrating measures of information to solve various problems related to NLU and NLG.

Regarding RQ1, the contribution of [Part II](#) is a new framework to include conversational and multimodal dimension in transcript representations. In [Chapter 5](#), we focused on the conversational part and we collected and processed a large number of conversations from OpenSubtitles and propose new hierarchical losses composed of two terms. The first term focuses on the lowest level (*i.e* it masks words and thus models intra-utterance dependencies), while the second term focuses on the dynamics at the utterance level (*i.e* it masks sequence and thus model inter-utterances dependencies). We also highlight how these new objectives relate to the MI. For the evaluation of the learnt transcript representations we also gathered for the community a new evaluation benchmark named SILICONE. To include the multimodal dimension we focused, in [Chapter 7](#), on the use of multivariate dependency measures to better fusion different modalities. Our losses which do not require any modification of the architecture can be combined with both pretrained and randomly initialized representation and improve the classification performance and provide more robust representations to modality drop. A by-product of our approach includes a tool to explain the learned representations.

Regarding RQ2, we demonstrate the usefulness of measures of information in [Part III](#). In [Chapter 7](#), we presented a new estimate of the MI information to learn better-disentangled representations. We exhaustively tested this estimator for both fair classification and sentence generation tasks and experimentally showed that it does not suffer from the limitations of current estimates. Additionally, we illustrate the different trade-offs while learning disentangled representation for different conditional sentence generation settings, including style transfer. Last, in [Chapter 8](#), we developed a new class of metrics called InfoLM. InfoLM can be viewed as combining the best of two categories of metrics: string-based metrics and embedding based metrics. Its ability to work with strings allows interpreting the result of the metrics (contrarily to other metrics based on continuous representation), in that sens InfoLM can be viewed as a string-based metric. On the other hand, it relies on a language model based on continuous representation, thus does not suffer from the problem

of existing string matching metrics (*e.g* BLUE score will assign a low score in case of paraphrase). Last, InfoLM uses discrete measures of information and uses all the layers of the language model instead of adding heuristic-based operations.

## 9.2 Limitation and Future Work

In this section, we put our contributions into perspective, discuss some of the limitations of our work as well as the numerous perspectives for the different parts of my work.

### 9.2.1 Future Research Directions Related to RQ1

Results from [Part II](#) open new future research directions:

- **Learning multimodal embeddings of transcripts at scale.** In [Part II](#), we have addressed the two dimensions of conversation separately. In [Chapter 5](#) we focused on the conversational dimension and in [Chapter 6](#) we focused the multimodal aspect. It would be interesting to combine both aspects and replace the pretrained BERT or XLNET used in [Chapter 7](#) with our hierarchical pretrained models on OpenSubtitles and measure the induced improvement.
- **Learning multilingual embeddings of transcripts.** As AI increasingly blends into everyday life across the globe, new applications are emerging [[JOSHI and collab., 2020b](#); [RUDER and collab., 2019](#)] and researchers are investigating the development of a dialogue system [[IPSIC and collab., 1999](#)] that would be able to handle the 7,000 languages spoken around the world. One of the key steps to developing such a multilingual system is to build generic cross-lingual dialogue representations that model contextual dependencies across multiple consecutive turns [[MITKOV, 2014](#); [WILLIAMS and collab., 2014](#)]. Similar techniques as the one described in [Chapter 5](#) could be leveraged to developing generic multilingual spoken dialogue representations using pre-training. Additionally, it would be the opportunity to collect a multilingual equivalent of SILICONE. This future research direction has been explored in [E. CHAPUIS \[2021\]](#).
- **Learning fair and debiased representations.** Pretrained representations on large corpora are known to be biased [[BENDER and collab., 2021](#)]. We believe that is particularly the case for the model presented in [Chapter 5](#), as it is pretrained on film corpora and thus particularly exposed to gender stereotypes [[GÁLVEZ and collab., 2019](#)] to only cite one [[SCHWEINITZ, 2010](#)]. Thus exploring technics based on MI estimation or involving other measures of information to learning fair and unbiased representations would be an interesting research direction.
- **Theoretical extension of the framework from [KONG and collab. \[2019\]](#).** As discussed in [Chapter 3](#) the framework described in [KONG and collab. \[2019\]](#) unifies different pretraining objectives. However, discrete is little theoretical comprehension of the role of each of the different freedom parameters (*e.g* role and relative importance of  $f_{\theta, \hat{\mathcal{B}}, q}$ ). New literature arises to propose new forms of noise [[JOSHI and collab., 2020a](#); [ROZIERE and collab., 2021](#)] we believe

that developing a better comprehension of [Equation 3.2](#) could allow proposing novel forms of noise to further improve the learning on massive corpora.

## 9.2.2 Future Research Directions Related to RQ2

Results from [Part III](#) open new future research directions:

- **Transferring other types of style.** The proposed method can replace the adversary in any kind of algorithms with no modifications [[FU and collab., 2017](#); [TIKHONOV and collab., 2019](#)]. Future work includes testing with other types of labels such as dialog acts [[COLOMBO and collab., 2020](#)], emotions [[WITON and collab., 2018](#)] or a speaker stance, or levels of a speaker confidence [[DINKAR and collab., 2020](#)]. Since our model allows more fine-grained control over the amount of disentanglement, we expect it to be easier to tune when combined with more complex models. Additionally, it would be the opportunity to collect and release an appropriate dataset for this task.
- **Developing new estimators of MI and other measures of information.** Estimation of mutual information has been pursued with significant efforts by the machine learning community. However, the estimation of the closely related of estimating differential entropy, fundamental concepts that began as an attempt by Shannon to extend the notion of entropy to continuous probability distributions, has received little attention in the last decades. We believe that estimating differential entropy could allow us to also construct a new family of estimators for different applications and we believe it could pave the way towards new applications of differential entropy estimation in its own right.
- **Extension of InfoLM settings.** Future work includes extending InfoLM to new tasks such as machine translation [[SHAH and collab., 2016](#)], image captioning and paraphrase detection. Our early results on the current version of *InfoLM* show that InfoLM does not achieve competitive results on machine translation task: *e.g* predictions obtain low correlations with human scores. We are particularly interested in reference-free MT evaluation as it is one of the most challenging probing tasks for encoders and it remains overlooked [[ZHAO and collab., 2020](#)]. We see multiple limitations of InfoLM in its current form:
  - InfoLM adopts a bag-of-distribution thus the summation does not take into account the orders of the unigrams.
  - InfoLM relies on unigram probability only where metrics such as BLUE can be generalized to n-grams.
  - The considered measures of informations might no be suited for other tasks, thus InfoLM could be extended to different information geometric measures (*e.g.*, Wasserstein distance, or other types of divergences) to obtain higher correlations. We also would like to consider using data depths [[STAERMAN and collab., 2021](#)] and its robust version [[STAERMAN and collab., 2020](#)].

Finally, we believe that ideas from this work can be adapted to different settings and augment metrics such as PRISM [[THOMPSON and POST, 2020](#)] that explore only a specific measure of information (*e.g* Entropy) and. We would like to

study the relationship with Markov Chains AZERAF and collab. [2020a, 2021a, 2020b, 2021b,c]; GORYNIN and collab. [2016].

## 9.3 References

- AZERAF, E., E. MONFRINI and W. PIECZYNSKI. 2020a, «Using the naive bayes as a discriminative classifier», *arXiv preprint arXiv:2012.13572*. 176
- AZERAF, E., E. MONFRINI and W. PIECZYNSKI. 2021a, «On equivalence between linear-chain conditional random fields and hidden markov chains», *arXiv preprint arXiv:2111.07376*. 176
- AZERAF, E., E. MONFRINI, E. VIGNON and W. PIECZYNSKI. 2020b, «Hidden markov chains, entropic forward-backward, and part-of-speech tagging», *arXiv preprint arXiv:2005.10629*. 176
- AZERAF, E., E. MONFRINI, E. VIGNON and W. PIECZYNSKI. 2021b, «Highly fast text segmentation with pairwise markov chains», in *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, IEEE, p. 361–366. 176
- AZERAF, E., E. MONFRINI, E. VIGNON and W. PIECZYNSKI. 2021c, «Introducing the hidden neural markov chain framework», *arXiv preprint arXiv:2102.11038*. 176
- BENDER, E. M., T. GEBRU, A. MCMILLAN-MAJOR and S. SHMITCHELL. 2021, «On the dangers of stochastic parrots: Can language models be too big?», in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, p. 610–623. 174
- COLOMBO, P., E. CHAPUIS, M. MANICA, E. VIGNON, G. VARNI and C. CLAVEL. 2020, «Guiding attention in sequence-to-sequence models for dialogue act prediction.», in *AAAI*, p. 7594–7601. 175
- DINKAR, T., P. COLOMBO, M. LABEAU and C. CLAVEL. 2020, «The importance of fillers for text representations of speech transcripts», in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, édité par B. Webber, T. Cohn, Y. He and Y. Liu, Association for Computational Linguistics, p. 7985–7993, doi: 10.18653/v1/2020.emnlp-main.641. URL <https://doi.org/10.18653/v1/2020.emnlp-main.641>. 175
- E. CHAPUIS, M. L. C. C., P. COLOMBO. 2021, «Cross-lingual pretraining methods for spoken dialog», . 174
- FU, Z., X. TAN, N. PENG, D. ZHAO and R. YAN. 2017, «Style transfer in text: Exploration and evaluation», *arXiv preprint arXiv:1711.06861*. 175
- GÁLVEZ, R. H., V. TIFFENBERG and E. ALTSZYLER. 2019, «Half a century of stereotyping associations between gender and intellectual ability in films», *Sex Roles*, vol. 81, n° 9, p. 643–654. 174
- GORYNIN, I., E. AZERAF, W. SABBAGH, E. MONFRINI and W. PIECZYNSKI. 2016, «Optimal filtering in hidden and pairwise gaussian markov systems», *International Journal of Mathematical and Computational Methods*, vol. 1. 176

- IPSIC, I., N. PAVESIC, F. MIHELIC and E. NOTH. 1999, «Multilingual spoken dialog system», in *ISIE'99. Proceedings of the IEEE International Symposium on Industrial Electronics (Cat. No. 99TH8465)*, vol. 1, IEEE, p. 183–187. [174](#)
- JOSHI, M., D. CHEN, Y. LIU, D. S. WELD, L. ZETTLEMOYER and O. LEVY. 2020a, «Spanbert: Improving pre-training by representing and predicting spans», *Transactions of the Association for Computational Linguistics*, vol. 8, p. 64–77. [174](#)
- JOSHI, P., S. SANTY, A. BUDHIRAJA, K. BALI and M. CHOUDHURY. 2020b, «The state and fate of linguistic diversity and inclusion in the nlp world», *arXiv preprint arXiv:2004.09095*. [174](#)
- KONG, L., C. D. M. D'AUTUME, W. LING, L. YU, Z. DAI and D. YOGATAMA. 2019, «A mutual information maximization perspective of language representation learning», *arXiv preprint arXiv:1910.08350*. [174](#)
- MITKOV, R. 2014, *Anaphora resolution*, Routledge. [174](#)
- ROZIERE, B., M.-A. LACHAUX, M. SZAFRANIEC and G. LAMPLE. 2021, «Dobf: A de-obfuscation pre-training objective for programming languages», *arXiv preprint arXiv:2102.07492*. [174](#)
- RUDER, S., I. VULIĆ and A. SØGAARD. 2019, «A survey of cross-lingual word embedding models», *Journal of Artificial Intelligence Research*, vol. 65, p. 569–631. [174](#)
- SCHWEINITZ, J. 2010, «Stereotypes and the narratological analysis of film characters», *Characters in Fictional Worlds: Understanding Imaginary Beings in Literature, Film, and Other Media*, vol. 3, p. 276. [174](#)
- SHAH, K., F. BOUGARES, L. BARRAULT and L. SPECIA. 2016, «Shelfium-nn: Sentence level quality estimation with neural network features», in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, p. 838–842. [175](#)
- STAERMAN, G., P. MOZHAROVSKIY, S. CLÉMEN and collab.. 2020, «The area of the convex hull of sampled curves: a robust functional statistical depth measure», in *International Conference on Artificial Intelligence and Statistics*, PMLR, p. 570–579. [175](#)
- STAERMAN, G., P. MOZHAROVSKIY, S. CLÉMENÇON and F. D'ALCHÉ BUC. 2021, «Depth-based pseudo-metrics between probability distributions», *arXiv preprint arXiv:2103.12711*. [175](#)
- THOMPSON, B. and M. POST. 2020, «Automatic machine translation evaluation in many languages via zero-shot paraphrasing», *arXiv preprint arXiv:2004.14564*. [175](#)
- TIKHONOV, A., V. SHIBAEV, A. NAGAEV, A. NUGMANOVA and I. P. YAMSHCHIKOV. 2019, «Style transfer for texts: Retrain, report errors, compare with rewrites», in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3927–3936. [175](#)



- WILLIAMS, J. D., M. HENDERSON, A. RAUX, B. THOMSON, A. BLACK and D. RAMACHANDRAN. 2014, «The dialog state tracking challenge series», *AI Magazine*, vol. 35, n° 4, p. 121–124. [174](#)
- WITON, W., P. COLOMBO, A. MODI and M. KAPADIA. 2018, «Disney at IEST 2018: Predicting emotions using an ensemble», in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018, Brussels, Belgium, October 31, 2018*, édité par A. Balahur, S. M. Mohammad, V. Hoste and R. Klinger, Association for Computational Linguistics, p. 248–253, doi: 10.18653/v1/w18-6236. URL <https://doi.org/10.18653/v1/w18-6236>. [175](#)
- ZHAO, W., G. GLAVAŠ, M. PEYRARD, Y. GAO, R. WEST and S. EGER. 2020, «On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation», *arXiv preprint arXiv:2005.01196*. [175](#)

# Appendix A

## Annexes

### A.1 Proofs of Chapter 7

#### A.1.1 Proof of Inequality Equation 7.4

In this section, we provide a formal proof of the Equation 7.4. Let  $(Z, Y)$  be an arbitrary pair of RVs with  $(Z, Y) \sim p_{ZY}$  according to some underlying pdf, and let  $q_{\hat{Y}|Z}$  be a conditional variational probability distribution on the discrete attributes satisfying  $p_{ZY} \ll p_Z \cdot q_{\hat{Y}|Z}$ , i.e., absolutely continuous.

$$I(Z; Y) \geq H(Y) - \text{CE}(\hat{Y}|Z). \quad (\text{A.1})$$

*Proof:* We start by the definition of the MI and use the fact that the maximum entropy distribution is reached for the uniform law in the case of a discrete variable (see ?).

$$I(Z; Y) = H(Y) - H(Y|Z) \quad (\text{A.2})$$

$$= \text{Const} - H(Y|Z). \quad (\text{A.3})$$

We then need to find the relationship between the cross-entropy and the conditional entropy.

$$\begin{aligned} & \text{KL}(p_{YZ} \| q_{\hat{Y}|Z}) \\ &= E_{YZ} \left[ \log \frac{p_{Y|Z}(Y|Z)}{q_{\hat{Y}|Z}(Y|Z)} \right] \\ &= E_{YZ} [\log p_{Y|Z}(Y|Z)] - E_{YZ} [\log q_{\hat{Y}|Z}(Y|Z)] \\ &= -H(Y|Z) + \text{CE}(\hat{Y}|Z). \end{aligned} \quad (\text{A.4})$$

We know that  $\text{KL}(p_{YZ} \| q_{\hat{Y}|Z}) \geq 0$ , thus  $\text{CE}(\hat{Y}|Z) \geq H(Y|Z)$  which gives the result.

The underlying hypothesis made by approximating the MI with an adversarial loss is that the contribution of gradient from  $\text{KL}(p_{YZ} \| q_{\hat{Y}|Z})$  to the bound is negligible.

#### A.1.2 Proof of Theorem 2

Let  $(Z, Y)$  be an arbitrary pair of RVs with  $(Z, Y) \sim p_{ZY}$  according to some underlying pdf, and let  $q_{\hat{Y}|Z}$  be a conditional variational probability distribution satisfying  $p_{ZY} \ll$

$p_Z \cdot q_{\hat{Y}|Z}$ , i.e., absolutely continuous. To obtain an upper bound on the MI we need to upper bound the entropy  $H(Y)$  and to lower bound the conditional entropy  $H(Y|Z)$ .

**Upper bound on  $H(Y)$ .** Since the KL divergence is non-negative, we have

$$H(Y) \leq \mathbb{E}_Y [-\log q_Y(Y)] \quad (\text{A.5})$$

$$= \mathbb{E}_Y \left[ -\log \int q_{\hat{Y}|Z}(Y|z) p_Z(z) dz \right]. \quad (\text{A.6})$$

**Lower bounds on  $H(Y|Z)$ .** We have the following inequalities:

$$H(Y|Z) = \mathbb{E}_{YZ} \left[ -\log q_{\hat{Y}|Z}(Y|Z) \right] - \text{KL}(p_{YZ} \| p_Z \cdot q_{\hat{Y}|Z}), \quad (\text{A.7})$$

where  $\text{KL}(p_{YZ} \| p_Z \cdot q_{\hat{Y}|Z})$  denotes the KL divergence. Furthermore, for arbitrary values  $\alpha > 1$ ,

$$H(Y|Z) \leq \mathbb{E}_{YZ} \left[ -\log q_{\hat{Y}|Z}(Y|Z) \right] - D_\alpha(p_{YZ} \| p_Z \cdot q_{\hat{Y}|Z}), \quad (\text{A.8})$$

where  $D_\alpha(p_{YZ} \| p_Z \cdot q_{\hat{Y}|Z}) =$

$$\frac{1}{\alpha - 1} \log \mathbb{E}_{ZY} [R^{\alpha-1}(Z, Y)]$$

is the Renyi divergence with

$$R(y, z) = \frac{p_{Y|Z}(y|z)}{q_{\hat{Y}|Z}(y|z)}.$$

The proof of Equation A.7 is given in subsection A.1.1. In order to show Equation A.8, we remark that Renyi divergence is non-decreasing function  $\alpha \mapsto D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z})$  in  $\alpha \in [0, +\infty)$ . Thus, we have  $\forall \alpha > 1$ ,

$$\text{KL}(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}) \leq D_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}). \quad (\text{A.9})$$

Therefore, from expression Equation A.7 we obtain the desired result.

### A.1.3 Optimization of the Surrogates on MI

In this section, we give details to facilitate the practical implementation of our methods.

**Computing the entropy  $H(Y)$**

$$\begin{aligned} H(Y) &\leq \mathbb{E}_Y \left[ -\log \int q_{\hat{Y}|Z}(Y|z) p_Z(z) dz \right] \\ &\approx \mathbb{E}_Y \left[ -\log \sum_{i=1}^n q_{\hat{Y}|Z}(Y|z_i) \right] + \text{const.} \\ &\approx -\frac{1}{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{Y}|} \log \sum_{i=1}^n C_{\theta_c}(z_i)_{y_j} + \text{const.} \end{aligned} \quad (\text{A.10})$$

where  $C_{\theta_c}(z_i)_{y_j}$  is the  $y_j$ -th component of the normalised output of the classifier  $C_{\theta_c}$ .

### Computing the lower bound on $H(Y|Z)$

The upper bound holds for  $\alpha > 1$ ,

$$\begin{aligned} H(Y|Z) &\approx \text{CE}(Y|Z) - \hat{D}_\alpha(p_{ZY} \| p_Z \cdot q_{\hat{Y}|Z}) \\ &\approx -\frac{1}{n} \sum_{i=1}^n \log q_{\hat{Y}|Z}(y_i|z_i) - \\ &\quad \frac{1}{\alpha-1} \log \sum_{i=1}^n R^{\alpha-1}(z_i, y_i). \end{aligned} \quad (\text{A.11})$$

**Estimating the density-ratio**  $R(z, y)$  In what follows we apply the so-called density-ratio trick to our specific setup. Suppose we have a balanced dataset  $\{(y_i^p, z_i^p)\} \sim p_{YZ}$  and  $\{(y_i^q, z_i^q)\} \sim q_{\hat{Y}|Z} p_Z$  with  $i \in [1, K]$ . The density-ratio trick consists in training a classifier  $C_{\theta_R}$  to distinguish between these two distributions. Samples coming from  $p$  are labelled  $u = 1$ , samples coming from  $q$  are labelled  $u = 0$ . Thus, we can rewrite  $R(z, y)$  as

$$R(z, y) = \frac{p_{Y|Z}(y, z)}{q_{\hat{Y}|Z}(y, z)} \quad (\text{A.12})$$

$$= \frac{p_{YZ|U}(y, z|u=0)}{p_{YZ|U}(y, z|u=1)} \quad (\text{A.13})$$

$$= \frac{p_{U|YZ}(u=0|y, z)}{p_{U|YZ}(u=1|y, z)} \frac{p_U(u=1)}{p_U(u=0)} \quad (\text{A.14})$$

$$= \frac{p_{U|YZ}(u=0|y, z)}{p_{U|YZ}(u=1|y, z)} \quad (\text{A.15})$$

$$= \frac{p_{U|YZ}(u=0|y, z)}{1 - p_{U|YZ}(u=0|y, z)}. \quad (\text{A.16})$$

Obviously, the true posterior distribution  $p_{U|YZ}$  is unknown. However, if  $C_{\theta_R}$  is well trained, then  $p_{U|YZ}(u=0|y, z) \approx \sigma(C_{\theta_R}(y, z))$ , where  $\sigma(\cdot)$  denotes the sigmoid function. A detailed procedure for training is given in Algorithm 2.

**Titre :** Apprendre à représenter et à générer du texte en utilisant des mesures d'information

**Mots clés :** Apprentissage Profond, Traitement du Langage Naturel, Théorie de L'information.

**Résumé :**

Le traitement du langage naturel (NLP) permet la compréhension et la génération automatiques du langage naturel. Le traitement du langage naturel a récemment fait l'objet d'un intérêt croissant de la part de l'industrie et des chercheurs, car l'apprentissage profond (AD) a exploité la quantité stupéfiante de textes disponibles (*e.g* web, youtube, médias sociaux) et a atteint des performances similaires à celles de l'homme dans plusieurs tâches (*e.g* traduction, classification de textes). Par ailleurs, la théorie de l'information (TI) et la DL ont développé un partenariat de longue date. En effet, l'informatique a favorisé l'adoption des réseaux neuronaux profonds grâce à des principes célèbres tels que la longueur minimale de description (LMD), le goulot d'étranglement de l'information (GIO) ou le célèbre principe InfoMax. Dans tous ces principes, différentes mesures de l'information (*e.g* entropie, MI, divergences) sont l'un des concepts fondamentaux.

Dans cette thèse, nous abordons l'interaction entre le NLP et les mesures d'information. Nos contributions se concentrent sur deux types de problèmes PNL : la compréhension du langage naturel (NLU) et la génération du langage naturel (NLG). La NLU vise à comprendre et à extraire automatiquement des informations sémantiques d'un texte d'entrée, tandis que la NLG vise à produire un langage naturel à la fois

bien formé (*i.e* grammaticalement correct, cohérent) et informatif.

La construction d'agents conversationnels parlés est un défi et le traitement des données conversationnelles parlées reste un problème difficile et négligé. Ainsi, nos premières contributions sont tournées vers l'UAL et nous nous concentrons sur l'apprentissage de représentations de transcriptions. Notre contribution se concentre sur l'apprentissage de meilleures représentations de transcriptions qui incluent deux caractéristiques importantes des conversations humaines parlées : la dimension conversationnelle et la dimension multimodale. Pour ce faire, nous nous appuyons sur diverses mesures d'information et nous tirons parti du principe de maximisation de l'information mutuelle. Le deuxième groupe de contributions aborde les problèmes liés au NLG. Cette thèse se concentre spécifiquement sur deux problèmes centraux. Premièrement, nous proposons une nouvelle limite supérieure de l'information mutuelle pour aborder le problème de la génération contrôlée via l'apprentissage de la représentation démêlée (transfert de style *i.e* et génération de phrases conditionnelles). Deuxièmement, nous abordons le problème de l'évaluation automatique des textes générés en développant une nouvelle famille de métriques utilisant diverses mesures d'information.

**Title :** Learning to Represent and Generate Text using Information Measures

**Keywords :** Deep Learning, Natural Language Processing, Information Theory

**Abstract :** Natural language processing (NLP) allows for the automatic understanding and generation of natural language. NLP has recently received growing interest from both industry and researchers as deep learning (DL) has leveraged the staggering amount of available text (*e.g* web, youtube, social media) and reached human-like performance in several tasks (*e.g* translation, text classification). Besides, Information theory (IT) and DL have developed a long-lasting partnership. Indeed, IT has fueled the adoption of deep neural networks with famous principles such as Minimum Description Length (MDL), Information Bottleneck (IB) or the celebrated InfoMax principle. In all these principles, different measures of information (*e.g* entropy, MI, divergences) are one of the core concepts.

In this thesis, we address the interplay between NLP and measures of information. Our contributions focus on two types of NLP problems : natural language understanding (NLU) and natural language generation (NLG). NLU aims at automatically understand and extract semantic information from an input text where NLG aims at producing natural language that is both

well-formed (*i.e* grammatically correct, coherent) and informative.

Building spoken conversational agents is a challenging issue and dealing with spoken conversational data remains a difficult and overlooked problem. Thus, our first contributions, are turned towards NLU and we focus on learning transcript representations. Our contribution focuses on learning better transcript representations that include two important characteristics of spoken human conversations : namely the conversational and the multi-modal dimension. To do so, we rely on various measures of information and leverage the mutual information maximization principle. The second group of contributions addresses problems related to NLG. This thesis specifically focuses on two core problems. First, we propose a new upper bound on mutual information to tackle the problem of controlled generation via the learning of disentangled representation (*i.e* style transfer and conditional sentence generation). Secondly, we address the problem of automatic evaluation of generated texts by developing a new family of metrics using various measures of information.