Semantic structuring of video collections from speech: segmentation and hyperlinking

Anca-Roxana Simon

ANNÉE 2015

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

Ecole doctorale Matisse

présentée par

Anca-Roxana Şimon

préparée à l'unité de recherche IRISA
Institut de Recherche en Informatique et Systèmes Aléatoires
Université Rennes 1

Semantic structuring of video collections from speech : segmentation and hyperlinking

Thèse soutenue à Rennes
le 2 Décembre 2015

devant le jury composé de :

Brigitte Grau
Professeur, ENSIIE, LIMSI / Présidente

Sophie Rosset
Directeur de recherche, CNRS, LIMSI / Rapporteur

Marie-Francine Moens
Professor, KU Leuven / Rapporteur

Roeland Ordelman
Senior Researcher, Univ. of Twente / Rapporteur

Benoît Favre
Maître de conférences, Univ. Aix-Marseille, LIF / Examinateur

Pascale Sébillot
Professeur, INSA de Rennes, IRISA /Inria Rennes/ Directrice de thèse

Guillaume Gravier
Directeur de recherche, CNRS, IRISA /Inria Rennes/ Directeur de thèse

Inspiration exists, but it has to find you working.
Pablo Picasso

Acknowledgements

I am most grateful to my amazing PhD supervisors, Pascale Sébillot and Guillaume Gravier, for their patience, guidance and helpful advices. They were also the supervisors of my MSc thesis, so I knew before I started the PhD thesis that I will be in good hands. They are both great scientists and great persons. Thank you for pushing me to grow and exceed my limits, learn what research means, how to study the problems, to ask the right questions and search for answers. Thank you for everything!

I would also like to thank the members of the jury: Brigitte Grau and Benoît Favre and my evaluators: Sophie Rosset, Sien Moens and Roeland Ordelman for taking the time to evaluate the work I have done and for giving me valuable feedback.

Kind regards to my internship supervisor Sien Moens for giving me the opportunity to work with you and be part of your team. I am grateful for the valuable advices you have given me and for helping me improve my work and myself. Also, thank you for introducing me to the world of topic models. I also want to thank the members of the LIIR team, for making my time there a very nice experience. Especially, I want to thank Susana for being such a kind and helpful person. I enjoyed a lot our talks.

Thanks go also to all the current and former members of the LinkMedia team: Laurent, Vincent, Teddy, Hervé, Ewa, Simon, Christian, François, Aurélie, Raghavendran, Rémi, Petra, Ricardo, Ahmet, Bingquing, Vedran, Ronan, Andrei, Giorgos, Li, Gylfi, Jon, Bogdan and Camille. I will always remember my teammates for the nice moments we shared. I hope our paths will cross again. Vedran, thank you for the cool poster.

I want to thank also the Kerdata team members for all the nice moments we had. I learned a lot about conferences on BigData, Cloud and HPC from our lunches and coffees together. Thank you Gabriel for making me fell like I'm also part of your team. I want to especially thank Luc Bougé for the support he has given me. Thank you!

I want to thank also the person who taught me how to make the first steps in research, Octavian Creţ. Thank you for the support you have shown me along the way.

I would like to thank my beloved family: Didi, Mina, Ioan, Simina, Ileana and Radu, for all the support you have given me by being there for me at every step. Va mulţumesc foarte mult! Didi, thank you for always knowing what to say to push me to want to be and do better.

I would like to thank a couple of friends that have shown great support along this path. A special thanks goes to Lexi, for all the adventures we had visiting each other in the various countries our research work has lead us to. Thank you for the advices and support you have given me. Many thanks go also to Markov, Costin, Alex and Jon. I enjoyed our talks, both the scientific ones and the non-scientific ones. My oldest friends Iulia, Sebi, Oana, Razvan, Ovidiu, Mircea, Diana, Mihai, Claudia and Adina also deserve a thank you. I know you would always be there for me if I needed.

Finally, I would like to thank all the other people that had a direct or indirect contribution to this work and were not mentioned above. This thesis is the result of serendipitous and fortuitous encounters with people that influenced a lot the development of my academic career. Your help and support is appreciated.

Contents

0 Resumé **1**
0.1 Contexte . . . 1
0.2 Contributions . . . 5
0.3 Perspectives . . . 7

1 Introduction **9**
1.1 Context . . . 9
1.2 Contributions . . . 12
1.3 Organization of the Manuscript . . . 14

Part I — **Linear and hierarchical topical structuring of TV shows** **15**

2 Background: Automatic topic segmentation **17**
2.1 Fundamentals on topic segmentation . . . 17
2.1.1 Topic definition . . . 18
2.1.2 Features used for topic segmentation . . . 19
2.2 State of the art . . . 20
2.2.1 Linear topic segmentation . . . 21
2.2.2 Evaluating linear topic segmentation . . . 26
2.2.3 Hierarchical topic segmentation . . . 27
2.2.4 Evaluating hierarchical topic segmentation . . . 30
2.3 Automatic transcripts of TV shows . . . 30
2.3.1 Characteristics . . . 30
2.3.2 TV show corpus . . . 31
2.4 Overview . . . 33

3 Leveraging lexical cohesion and disruption for topic segmentation **35**
3.1 Our approach in a nutshell . . . 35
3.2 Combining lexical cohesion and disruption . . . 36
3.2.1 Probabilistic graph-based segmentation . . . 36
3.2.2 Introduction of the lexical disruption . . . 37
3.2.3 Segmentation algorithm . . . 38
3.3 Experiments . . . 40
3.3.1 Corpora . . . 40

3.3.2 Results . . . 41
3.4 Discussion . . . 45

4 Investigating hierarchical topic segmentation 47
4.1 Overview . . . 47
4.2 Is lexical cohesion enough for hierarchical topic segmentation? . . . 49
4.2.1 Classical measures for lexical cohesion . . . 49
4.2.2 Lexical cohesion through burst analysis . . . 53
4.2.3 Discussion . . . 56
4.3 When to stop applying a hierarchical topic segmentation algorithm? . . . 56
4.4 Hierarchical structure of topically focused fragments . . . 58
4.4.1 Algorithm . . . 58
4.4.2 Evaluation . . . 59
4.4.3 Comparison with a traditional dense segmentation . . . 60
4.4.4 Application-driven evaluation . . . 62
4.5 Discussion . . . 65

Part II — **Implications of the topical structure in video hyperlinking 69**

5 Video hyperlinking 71
5.1 Context . . . 71
5.2 SH and SAVA at MediaEval benchmark initiative . . . 72
5.2.1 Existing approaches . . . 74
5.2.2 Data . . . 75
5.2.3 Hyperlinking evaluation . . . 76
5.2.4 Anchoring evaluation . . . 79
5.3 Conclusion . . . 81

6 Investigating domain-independent techniques for precise anchor and target selection in video hyperlinking 83
6.1 Our approach in a nutshell . . . 83
6.2 Hyperlink creation . . . 84
6.2.1 Anchor selection . . . 84
6.2.2 Target selection . . . 85
6.2.3 Hyperlinking anchors and targets . . . 86
6.3 Experiments . . . 88
6.3.1 Anchoring evaluation . . . 88
6.3.2 Hyperlinking evaluation . . . 88
6.4 Conclusion . . . 92

7 Leveraging topic models to justify links and control diversity in video hyperlinking 93
7.1 Hierarchical topic models for language-based video hyperlinking . . . 93
7.2 Problem formulation: hyperlink creation . . . 95
7.2.1 Direct hyperlinks . . . 95
7.2.2 Indirect hyperlinks . . . 95

7.3 Leveraging topic models for indirect hyperlinks . . . 96
7.3.1 Building the topical structure . . . 96
7.3.2 Independent topic levels (IT) . . . 98
7.3.3 Hard links between topics (HLT) . . . 98
7.3.4 Hard & soft links between topics (HSLT) . . . 100
7.4 Evaluation in the context of the Search and Hyperlinking task . . . 100
7.4.1 Comparing direct and indirect links . . . 101
7.4.2 Hierarchical topical structures . . . 102
7.4.3 Analysis of the links . . . 103
7.5 Assessing serendipity and diversity in the links . . . 109
7.5.1 Survey . . . 109
7.5.2 Evaluation and results . . . 111
7.6 Conclusion . . . 119

8 Conclusions **121**
8.1 Thesis objectives . . . 121
8.2 Summary of the contributions . . . 122
8.3 A step further . . . 123

Appendix A Kleinberg's algorithm **137**

Appendix B Survey details **139**

Chapter 0

Resumé

Contents

0.1 Contexte **1**
0.2 Contributions **5**
0.3 Perspectives **7**

0.1 Contexte

Selon un récent rapport IDC [50], la taille des données électroniques mondiales, également connues sous l'appellation "univers numérique", double tous les deux ans et sera multipliée par dix à partir du 2013 jusq'en 2020, passant de 4 400 à 44 000 milliards de gigaoctets (soit plus de 5 200 gigaoctets par personne dans le monde). Pour donner une idée de l'échelle, si l'on considère chaque octet de données comme égal à un pouce (soit 2,54 cm), cela correspondrait à environ 1 million d'allers-retours entre la Terre et Pluton[1]. Cet univers numérique est constitué d'images, de vidéos générées par les utilisateurs, de contenus télévisuels, de collisions subatomiques enregistrées par le grand collisionneur de hadrons du CERN, de messages postés sur les réseaux sociaux, de sms, d'emails, etc. Le déluge de données a déjà commencé à transformer les entreprises qui cherchent à capitaliser les valeurs issues de grandes collections de données. Dans le même temps, il influence aussi le processus de découverte scientifique, en conduisant à ce qu'on appelle la Science des données [70]. En 2013, seuls 22% des données de l'univers numérique étaient considérés comme utiles. Pour que des données soient utiles, elles doivent être caractérisées ou étiquetées. Actuellement toutefois, moins de 5% de ces données pourraient être analysés. Le pronostic pour 2020 est que le pourcentage de données utiles pourrait croître jusqu'à plus de 35%

[1]Exemple donné par Yukun Harsono, directeur général de la Grande Asie pour Elsevier.

[50]. Pour atteindre cet objectif, un des principaux défis concerne les données non structurées. Aujourd'hui, ≈ 80% de l'univers numérique est encore non structuré, ce qui signifie que l'on sait peu de choses sur lui et sur les endroits où de la valeur peut être extraite.

Dans le cadre de notre travail, nous nous intéressons principalement aux contenus télévisuels. Bien que les données télévisuelles puissent avoir une certaine structure interne implicite, elles sont toujours considérées comme "non structurées" parce qu'elles ne peuvent être facilement organisées et ne sont pas aisément stockables en bases de données. Les contenus télévisuels disponibles représentent des volumes énormes et sont répandus à travers la planète. En France par exemple, l'Institut National de l'Audiovisuel (INA) archive les radios et chaînes audiovisuelles nationales. Il abrite plus de 5 millions d'heures de programmes. Depuis 2008, il recueille quotidiennement des données de 88 chaînes de télévision et de 20 stations de radio. Autre exemple d'archives de télévision, la BBC[2] détient 600 000 heures de contenus TV et 350 000 heures de radio. Elle produit du contenu pour 4 chaînes TV et 9 stations nationales, et propose une centaine d' heures d'actualités par jour. Toutes ces tendances montrent une augmentation continue de la diffusion de données télévisuelles, ce qui met l'accent sur de nombreux défis à relever.

Au cours des dernières années, de nouveaux challenges ont en effet émergé avec la transformation très significative du paysage audiovisuel due à l'émergence de la télévision sur Internet. Le changement visible dans le mode de consommation audiovisuelle, entre celle linéaire de la TV standard et celle de la télévision sur Internet, pousse l'écosystème de radiodiffusion traditionnel à s'adapter au paysage en-ligne (par exemple, la BBC fournit le service de *catch-up TV* BBC iPlayer). Les gens utilisent de plus en plus les services de TV connectée, de télévision de rattrapage ou de vidéos à la demande (VOD). Dans un rapport de l'Observatoire européen de l'audiovisuel [56], il est indiqué qu'au Royaume-Uni le pourcentage d'adultes utilisant des services de VOD a cru de 46% en 2010 à 59% en 2012, et à 67% en 2014. Aux États-Unis, 40% des ménages ont souscrit, en 2014, à des services de VOD en *streaming* tels que Netflix, Amazon Prime Instant Video, Hulu, etc. Plus d'un milliard d'heures d'émissions de télévision et de films sont diffusées par Netflix chaque mois. Netflix a commencé à étendre son emprise à divers pays de l'UE et une croissance de l'ordre de 20,7% du pourcentage de ménages en Europe souscrivant à la VOD est attendu pour 2020. Parmi les autres candidats forts du paysage audiovisuel en ligne, on peut aussi citer des sites comme YouTube, Dailymotion, Blinkx, etc. 300 heures de vidéos sont téléchargées sur YouTube chaque minute par plus d'un milliard d'utilisateurs. YouTube existe dans 75 pays et est disponible en 61 langues. Le paysage en-ligne change donc radicalement la façon dont les téléspectateurs consomment les données audiovisuelles. La décision de ce qui est regardé et dans quel ordre n'appartient plus à la chaîne de TV mais à l'utilisateur. Ceci se traduit par un modèle de diffusion de contenus télévisuels centré-utilisateur. Afin d'assurer une haute qualité de service, de nouveaux systèmes doivent par conséquent être mis en place.

Une première option consiste à étendre des outils traditionnels d'analyse et de gestion de données existants, mais ceci nécessiterait que les données soient structurées. Cependant, compte tenu de la croissance des contenus audiovisuels divers et non structurés, leur annotation manuelle, leur analyse et leur indexation sont des tâches coûteuses et consommatrices en temps. Par conséquent, le développement de nouvelles techniques automatiques de structuration des données est devenu une nécessité afin de faciliter l'accès à l'information

[2] http://www.bbc.co.uk/archive/tv_archive.shtml

audiovisuelle contenues dans les vidéos. Les utilisateurs doivent pouvoir trouver les informations recherchées rapidement et précisément. Plutôt que de devoir regarder l'intégralité d'une émission de télévision, ils devraient par exemple pouvoir accéder seulement aux parties qui les intéressent. Cela peut concerner la recherche d'un événement spécifique (par exemple, des sondages électoraux), d'un fragment de vidéo contenant une certaine personnalité, ou d'un fragment abordant un sujet particulier, etc. Par conséquent, les techniques de structuration devraient révéler l'organisation interne des vidéos. Cette organisation peut prendre la forme d'une table des matières ou encore d'un résumé pour permettre à un utilisateur de se faire rapidement une idée du contenu d'une vidéo et décider si cela vaut la peine de la regarder en intégralité, partiellement ou pas du tout. La structuration du contenu vidéo peut être définie comme "le processus de décomposition hiérarchique des vidéos en unités et la construction des relations qu'elles entretiennent" [143]. De la même façon que les textes peuvent être structurés en chapitres, paragraphes, phrases et mots, les vidéos peuvent être segmentées en unités telles que des scènes, des plans et des images-clés. Selon [143], une image-clé est celle qui représente le mieux le contenu d'un plan ou d'un sous-plan ; un sous-plan est un segment d'un plan qui correspond à un mouvement unique de la caméra ; un plan est un clip enregistré par une seule caméra de manière continue ; une scène est définie comme une collection de plans sémantiquement liés et temporellement adjacents, représentant un concept de haut niveau. La macro-segmentation du contenu télévisuel (c-à-d la segmentation en scènes) est à l'origine de nombreux nouveaux services de diffusion TV, en particulier du service de TV à la demande [11]. La segmentation en scènes offre la possibilité de produire la table des matières d'une vidéo. Une scène peut être formée à partir de différentes unités (par exemple, en liens temporels, thématiques) selon le type de vidéo considéré, et doit contenir des plans cohérents qui ont une signification pour le spectateur. Par conséquent, la définition d'une scène peut varier, ce qui a conduit à l'élaboration d'approches de structuration du contenu TV ayant des objectifs divers.

Les approches de structuration existantes peuvent être classées selon qu'elles utilisent ou non de la connaissance *a priori* sur le contenu des programmes TV. L'utilisation de connaissance préalable conduit habituellement au développement de systèmes spécifiques, pouvant structurer certains types de vidéos. Des systèmes de ce genre sont proposés dans [144], [81] et [33], où les auteurs utilisent des modèles de Markov cachés pour structurer les émissions de sport telles que des matchs de football et de tennis. D'autres systèmes spécifiques sont dédiés à la structuration des journaux TV en répartissant les plans vidéos en plusieurs classes (par exemple, présentateur, météo, reportage) [12, 35]. Certaines approches ne visent pas à structurer l'intégralité d'une vidéo, mais se focalisent sur la détection d'éléments structurels en son sein, tels que le présentateur [75, 120, 51] ou les buts des matchs de football [23, 6]. Elles demeurent toutefois spécifiques. Il existe également des approches génériques qui visent à extraire la structure entière ou certains éléments structurels sans aucune connaissance préalable. Généralement ces approches se fondent sur de la découverte de motifs dans le flux TV, cherchant à repérer des segments récurrents ayant une consistance sur un plan audiovisuel [10], ou utilisant une technique de *micro-clustering* pour regrouper des vecteurs de caractéristiques audio/visuelles similaires [11]. Des approches plus récentes reposent sur des techniques d'inférence grammaticale pour mettre en avant la structure sous-jacente de programmes récurrents [111, 110]. Les progrès réalisés au sein de la communauté du traitement automatique des langues ont conduit au développement de technologies applicables aux transcriptions manuelles ou automatiques de la parole contenue dans les émis-

sions TV afin d'en produire la structure thématique et donc une table des matières. Ces approches doivent déterminer les frontières entre les histoires individuelles présentes dans chaque émission et trouver des segments vidéos thématiquement cohérents. Dans [131], les auteurs proposent une méthode inspirée de la segmentation d'images combinée avec une indexation sémantique latente, technique employée en traitement automatique des langues, pour extraire la structure thématique d'un journal TV de CNN en se fondant sur sa transcription manuelle. Dans [60], les auteurs obtiennent la structure thématique d'émissions de télévision en adaptant une approche de segmentation thématique linéaire utilisée pour de l'écrit aux transcriptions automatiques. La précision de la segmentation est améliorée grâce à la prise en compte de relations sémantiques entre mots et celle des mesures de confiance fournies par le système de reconnaissance de la parole utilisé.

Dès lors qu'il est devenu possible d'extraire des caractéristiques multimodales des contenus audiovisuels, diverses approches ont donc émergé pour les structurer. Cependant structurer les contenus est inutile sauf si cela produit de la *Valeur*. On peut affirmer que la valeur est l'objectif principal de la révolution technologique actuelle. Par conséquent, l'étape naturelle suivante est d'exploiter les contenus structurés pour obtenir de la valeur. Par exemple, les entreprises exploitent les données de leurs clients pour apprendre leurs préférences et cherchent à fournir de nouvelles technologies multimédias offrant une large gamme de fonctionnalités aptes à répondre à leurs besoins. Dans le contexte des données audiovisuelles, un utilisateur peut vouloir suivre l'évolution d'un événement au fil du temps, ou connaître la façon dont il est présenté par différents chaînes pour savoir si une vidéo peut être regardée ou non par un jeune enfant ; il peut souhaiter découvrir des informations intéressantes et inattendues à partir d'un segment vidéo portant sur un sujet qui l'intéresse, etc. Des techniques notables ont été mises au point en ce sens par des chercheurs, en particulier dans le cadre de campagnes d'évaluation telles que MediaEval et TRECVid [105].

Parmi ces techniques, il existe des solutions pour la détection automatique de manipulations et d'utilisations abusives de contenus multimédias [15]. De telles solutions peuvent aider des professionnels qui cherchent à vérifier si une information est fiable ou non. D'autres techniques ont été développées pour indexer des personnes apparaissant dans de grandes archives TV, dans des conditions réelles (c-à-d sans liste préétablie de ces personnes) pour rendre ces archives interrogeables [107] et aptes à répondre à des questions telles que "Qui parle quand ?" et "Qui apparaît quand ?". Des solutions pour aider les utilisateurs à trouver des vidéos qui correspondent à leur humeur du moment, leur âge ou leurs préférences ont également été proposées dans le cadre de MediaEval [130]. La détection de l'impact émotionnel d'un film pourrait aider à améliorer les scénarios de recherche ou de recommandation. La détection de contenus violents pourrait permettre à des parents de choisir les contenus les plus appropriés pour leurs enfants. Au cours des dernières années dans le cadre de ces campagnes d'évaluation, les chercheurs se sont aussi intéressé à produire des solutions permettant la création d'hyperliens entre vidéos au sein de grandes collections [40, 105]. Disposer de tels hyperliens est un atout pour ces collections puisque cela peut encourager leur exploration et permettre la découverte d'information pertinente, intéressante ou inattendue. Toujours dans cette optique d'exploitation de la structure audiovisuelle, une autre initiative, Topic Detection and Tracking (TDT) [3], s'intéressait à la détection et au suivi de sujets. Plus précisément, elle visait à explorer les techniques de détection d'apparition de nouveaux sujets et de suivi de leur réapparition et leur évolution dans un flux de reportages.

Toutes les initiatives mentionnées précédemment montrent qu'une des futures tendances-clés consiste à fournir des technologies qui offrent des fonctionnalités diverses et complexes pour améliorer l'accès à l'information présente dans le déluge actuel de contenus audiovisuels. Ceci souligne l'intérêt et les défis de la structuration et de l'exploitation de ces contenus. Et ce sont les deux principaux aspects auxquels nous nous intéressons au sein de cette thèse. Des solutions pour relever ces challenges permettraient à des utilisateurs de profils divers, tels que des journalistes, des étudiants, des professionnels, des chercheurs, des archivistes ou des utilisateurs à domicile, de tirer profit des technologies telles que celles mentionnées. Ils auraient la possibilité de découvrir, de naviguer et de rechercher au sein de grandes collections de vidéos. Extraire de ces façons de la valeur de ces collections accroît par ailleurs leur valeur économique et/ou culturelle [38].

0.2 Contributions

Le premier objectif de cette thèse est de fournir des techniques automatiques et génériques pour la structuration thématique de données audiovisuelles. Le second objectif consiste à étudier les implications de la structure produite sur diverses tâches liées au traitement automatique des langues, telles que la création d'hyperliens entre vidéos, en sélectionnant des ancres et des cibles précises, ou la création de résumés automatiques. Une contrainte que nous nous imposons, en termes de structuration thématique, est de fournir des solutions automatiques et génériques, pouvant donc être appliquées à tout type de données audiovisuelles. En effet, étant donné les quantités impressionnantes de contenus audiovisuels, il devient difficile de créer des solutions spécifiques à chaque type d'émission TV. Des méthodes non supervisées, pouvant traiter des contenus télévisuels hétérogènes, sont donc nécessaires. Pour répondre à cette contrainte, nous fondons nos approches sur les transcriptions automatiques de la parole prononcée dans les émissions, ce qui nous permet d'être indépendante du genre de documents examinés. Ces transcriptions sont obtenues à l'aide d'un système de reconnaissance automatique de la parole.

Identifier la structure thématique d'une vidéo signifie découvrir son organisation sémantique globale. Deux formes de segmentation peuvent être distinguées : la segmentation linéaire et la segmentation hiérarchique. La segmentation thématique linéaire vise à structurer les données en thèmes consécutifs. Les techniques de segmentation thématique hiérarchique consistent, quant à elles, à diviser un sujet principal en sous-thèmes, qui peuvent à leur tour être subdivisés en sous-sous-thèmes, etc. Nous abordons ces deux types de structures thématiques, traitant ces deux points de vue sur l'organisation interne des données. Les techniques génériques de segmentation thématique (en particulier celles utilisées sur les données textuelles) exploitent habituellement la notion de cohésion lexicale, indépendante du type de documents textuels considérés et ne nécessitant pas de phase d'apprentissage. Segmenter thématiquement des données en se fondant sur la cohésion lexicale signifie analyser la distribution des mots afin d'identifier des changements importants dans le vocabulaire qui peuvent laisser supposer des changements de sujets.

Notre première contribution consiste en la proposition d'une nouvelle technique automatique et générique de segmentation thématique linéaire. Les méthodes de la littérature reposent généralement sur un critère de segmentation parmi deux possibles : soit sur la maximisation d'une mesure de cohésion lexicale au sein d'un segment, soit sur la détection de

ruptures lexicales. Notre solution combine ces deux critères de manière à obtenir le meilleur compromis entre cohésion et rupture. Nous donnons une formulation mathématique de ce nouveau critère de segmentation, et proposons un algorithme de segmentation l'exploitant. Des évaluations menées tant sur des corpus de textes écrits que sur des transcriptions automatiques de la parole d'émissions TV démontrent la pertinence de la combinaison des critères de cohésion et de rupture.

Nous nous penchons ensuite sur la segmentation thématique hiérarchique. Nous étudions tout d'abord jusqu'à quel point la segmentation hiérarchique peut tirer profit de la cohésion lexicale, et montrons que cette dernière est insuffisante pour reproduire des segmentations de référence (c-à-d des vérités-terrain manuelles) de flux télévisés. En conséquence, nous proposons une nouvelle façon de considérer ce genre de structure hiérarchique, passant d'une segmentation dense classique à une hiérarchie de fragments thématiquement concentrés. Afin de dépasser les limites imposées par un comptage global de récurrences lexicales dans les segments, cette nouvelle structure est obtenue en tirant partie de la répartition temporelle des récurrences de mots grâce à la recherche de *bursts*. Les mots saillants (*bursty words*) sont caractérisés par des intervalles entre apparitions longs suivis d'intervalles courts, alors que les mots non saillants présentent une variance plus faible. L'idée sous-jacente est que la présence de *bursts* lexicaux indique une forte concentration thématique. En nous fondant sur l'algorithme de Kleinberg [82] de détection de ces mots saillants, nous extrayons les idées importantes des données à différents niveaux de détail et nous les regroupons au sein d'une hiérarchie de fragments thématiquement concentrés. La nouvelle structure est évaluée à l'aide d'une comparaison qualitative à une segmentation dense classique sur divers jeux de données, mais également dans un contexte de production de résumés automatiques. Ces évaluations montrent la capacité de la structure à faire émerger l'information importante des données. La figure 1 propose une représentation générique des structures résultant de ces contributions. La figure 1(a) est la représentation d'une segmentation thématique linéaire dans laquelle la transcription d'une émission de télévision est divisée en thèmes principaux ; la figure 1(b) propose une segmentation thématique hiérarchique classique dans laquelle les principaux thèmes sont divisés en sous-thèmes, qui à leur tour peuvent être divisés. Contrairement à la façon habituelle de segmenter, l'idée présentée dans la figure 1(c) consiste à repérer des fragments thématiques ciblés, non nécessairement contigus, et à les organiser à différents niveaux d'une manière hiérarchique.

Nous étudions ensuite les implications des structures thématiques internes des données audiovisuelles, obtenues grâce à nos solutions précédentes, dans le cadre de la campagne d'évaluation MediaEval, et plus précisément, pour la tâche de *Search and Hyperlinking* (recherche et création d'hyperliens). L'objectif global est d'améliorer l'expérience d'utilisateurs navigant au sein d'une collection de vidéos grâce à des hyperliens. La génération automatique d'hyperliens dans les données vidéos est un sujet d'actualité, ayant pour objectif d'offrir des moyens de navigation en sus de la recherche d'information classique dans de grandes collections vidéos. Considérons un exemple : un utilisateur lance une recherche dans une collection de vidéos à l'aide de la requête "choses à voir à Londres". Une liste de vidéos lui est retournée et il commence à en visionner une. À un certain point, il se peut que la fameuse "cabine téléphonique rouge" britannique soit mentionnée. L'utilisateur peut souhaiter obtenir alors davantage de détails sur ce sujet, c-à-d quelque chose qu'il ne cherchait pas explicitement lors de la recherche initiale. Le fragment de vidéo

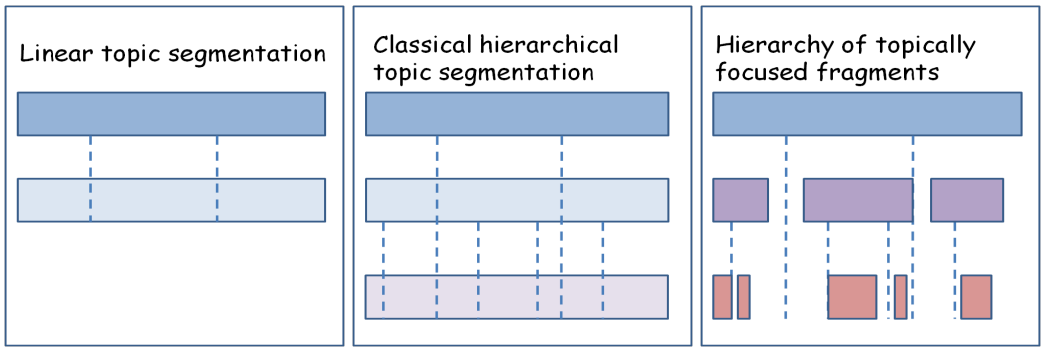


Figure 1: Représentations génériques de (a) la segmentation thématique linéaire (b) la segmentation thématique hiérarchique classique dense (c) la segmentation hiérarchique en fragments thématiquement concentrés. Les lignes verticales illustrent les frontières thématiques et sous-thématiques.

dans lequel la cabine téléphonique rouge est mentionnée sert alors de nouvelle "requête" et permet d'obtenir une nouvelle information sous la forme de fragments courts de vidéos liés à cette "requête". Idéalement, ces fragments-réponses doivent apporter des informations supplémentaires, par exemple dans notre cas, à propos de l'histoire de la cabine téléphonique, à propos de l'inventeur du téléphone, etc. Cet exemple illustre un scénario de recherche et de navigation à partir d'hyperliens. La nouvelle requête (la cabine téléphonique rouge) est appelée *ancre* ; les fragments courts de vidéos obtenus en réponse sont appelés *cibles*. L'objectif de la création d'hyperliens entre vidéos est donc de tisser des liens à partir d'une ancre, en sélectionnant automatiquement les fragments-cibles qui offrent des informations complémentaires. Dans nos approches, nous nous appuyons sur la structure thématique pour permettre, contrairement à la plupart des techniques existantes, d'identifier automatiquement des ancres et des cibles précises. Nous développons en particulier une nouvelle méthode qui s'intéresse à deux questions fondamentales de la création d'hyperliens entre vidéos : la diversité au sein des liens fournis et la caractérisation de ceux-ci. Pour ce faire, nous proposons différentes stratégies exploitant une hiérarchie de modèles de thèmes (*topic models*) comme représentation intermédiaire lors de la comparaison des transcriptions des segments vidéos. Ces représentations hiérarchiques offrent une base pour caractériser les hyperliens, grâce à la connaissance des thèmes ayant contribué à la création des liens, et pour produire des liens variés en choisissant de donner plus de poids à des thèmes soit généraux, soit spécifiques.

0.3 Perspectives

Plusieurs améliorations des solutions présentées dans cette thèse peuvent être envisagées. Pour ce qui concerne nos algorithmes de segmentation thématique, tant linéaire que hiérarchique, des informations supplémentaires pourraient être prises en compte pour tenter de pallier l'impact des erreurs de transcription qui conduisent, en particulier, à limiter certaines répétitions de mots. Pour préserver le caractère générique de nos approches, le potentiel des

mesures de confiance fournies par les systèmes de reconnaissance automatique de la parole et des relations sémantiques lexicales – qui ont déjà montré leur pertinence pour améliorer la segmentation thématique de transcriptions automatiques de reportages TV [60] – pourrait être exploré. Concernant la création d'hyperliens entre fragments de vidéos, une extension possible consisterait à prendre en compte toutes les modalités des données et à les combiner pour obtenir encore plus de diversité et de sérendipité dans les résultats. Au-delà d'une solution simple déjà existante, consistant à travailler sur chaque modalité indépendamment puis à combiner les résultats, une fusion précoce des modalités, à travers une traduction d'une modalité vers l'autre, serait une perspective plus intéressante que nous commençons à investiguer.

Pour répondre toujours mieux aux exigences, toujours plus diversifiées et personnalisées, des consommateurs, une attention croissante doit également être portée aux évaluations centrées utilisateurs. À l'instar des solutions centrées utilisateurs qui gagnent en importance, la prise en compte de l'évaluation subjective par l'utilisateur de la qualité de l'expérience devient un sujet de recherche important. Une évaluation capable de capturer la réponse d'un système à chaque demande d'un utilisateur par une mesure de la satisfaction de celui-ci pourrait être un indicateur de la qualité du système. En création d'hyperliens entre vidéos par exemple, une ancre peut être intéressante pour une personne soit d'un point de vue visuel ou du fait de ce qui y est dit. Par conséquent, un système qui pourrait fournir diverses fragments-cibles selon la modalité et laisser l'utilisateur choisir la cible à suivre, pourrait être intéressant à développer. Une évaluation globale de la satisfaction des utilisateurs, après avoir exploré une collection de vidéos via divers hyperliens, pourrait alors être une manière d'évaluer ce système. Construire des scénarios d'évaluations centrées utilisateurs est une tâche complexe. La création d'interfaces intelligentes, pouvant afficher les divers liens et cibles vers des points d'intérêt dans les vidéos, serait probablement une première étape à explorer.

Chapter 1

Introduction

Contents

1.1 Context 9
1.2 Contributions 12
1.3 Organization of the Manuscript 14

1.1 Context

According to a recent IDC report [50], the size of the world electronic data, also known as the «digital universe», is doubling every two years and will multiply 10-fold between 2013 and 2020: from 4.4 trillion gigabytes to 44 trillion gigabytes (more than 5,200 GB per person worldwide). To give a sense of scale, if you imagine each byte of data as equal to one inch, it would be approximately 1 million round trips between Earth and Pluto[1]. This digital universe is made up of images, user generated videos, digital movies populating the pixels of our high definition TVs, security footage, subatomic collisions recorded by the Large Hadron Collider at CERN, voice calls, texting, social media posts, emails, etc. This data deluge is already starting to transform businesses, which search to capitalize the values searched in large data collections. At the same time it also impacts the process of scientific discovery, moving towards what is called Data Science [70]. In 2013, only 22% of the data in the digital universe was considered to be useful. For data to be useful it needs to be characterized or tagged. Nevertheless, less than 5% of this data could actually be analyzed so far. The prognosis for 2020 is that the useful percentage could grow to more than 35% [50]. To reach this goal one of the key challenge to tackle is unstructured data. Today, $\approx$ 80% of the digital universe is still unstructured, which means little is known about it and where the value could be found.

[1] example given by Yukun Harsono, managing director of Greater Asia for Elsevier

In the context of our work we are mainly interested with television content. While television data may have some implied internal structure, it is still considered "unstructured", because they cannot be easily organized and don't fit neatly in a database. The available television content has a large volume and is widely spread all over the world. In France for example, the National Audiovisual Institute (INA) is a repository of all French radio and audiovisual archive. It contains over 5 millions hours of programs. Since 2008, it collects daily data from 88 TV channels and 20 radio stations. Another example of television archive is the BBC[2] one, containing 600,000 hours of TV content and 350,000 hours of radio. BBC produces material for 4 TV channels and 9 national network stations. It provides 100 hours of daily news coverage. These trends show that there is a continuous increase in broadcasting data, which brings forward many challenges to process it.

New challenges were brought as the audiovisual landscape has transformed significantly in the past few years with the emergence of Internet-based TV. There is a visible shift of audiovisual consumption from linear broadcast TV towards the Internet, which pushes the traditional broadcast ecosystem to adapt to the online landscape (e.g., BBC provides the catch-up TV service BBC iPlayer). People start to rely more on connected TV, catch-up TV or Video On Demand (VOD) services. In a report prepared by the European Audiovisual Observatory [56], it is stated that in the UK the percentage of adults accessing VOD increased from 46% in 2010 to 59% in 2012 and to 67% in 2014. In the U.S., 40% of households subscribed in 2014 to VOD streaming services like Netflix, Amazon Prime Instant Video, Hulu, etc. There are more than one billion hours of TV shows and movies streamed from Netflix per month. Netflix has started to expand its footprint to various countries in EU and it is expected to see a growth in the percentage of households in Europe subscribing to VOD of 20.7% by 2020. Other strong contestants to the audiovisual online landscape are sites like Youtube, Dailymotion, Blinkx, etc. Youtube has 300 hours of videos uploaded every minute with more than 1 billion users, it is localized in 75 countries and available in 61 languages. The online landscape drastically changes the way in which viewers are consuming audiovisual data. The decision for what to watch and in what order no longer belongs to the TV station but to the user. This translates to a user-centric model of TV streams. To ensure a high quality of service, new systems have to be put in place.

The first option is to build on existing traditional data management and analysis tools. This would require the data to be structured. Given the growth of diverse and unstructured audiovisual content, manual annotation, analysis and indexing are rather expensive and labour-intensive tasks. Therefore, the development of new automatic data structuring techniques has become a necessity in order to facilitate access to the audiovisual information contained in the videos. Users should be able to find the information fast and accurately. For example, instead of having to watch an entire TV show, they should be able to access just the parts that interest them. It can mean searching for a specific event (e.g., the election polls), a video fragment about a certain public person, or a fragment about a particular topic, etc. Therefore structuring techniques should prevail the internal organization of the videos. The organization can be in the form of a table of contents or summary to enable a user to quickly figure out the overview contents of a video and decide whether it is worth watching the whole video, only some part or nothing at all. Video content structuring can be defined as "the process of hierarchically decomposing videos into units and building their relationships" [143]. Similarly to how textual documents can be structured into chapters,

[2]http://www.bbc.co.uk/archive/tv_archive.shtml

paragraphs, sentences and words, videos can also be segmented into units such as: scenes, shots and keyframes. According to [143], a keyframe is the frame that best represents the content of a shot or a subshot; A subshot is a segment within a shot that corresponds to a unique camera motion; A shot is an uninterrupted clip recorded by a single camera; A scene is defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept. TV broadcast macro-segmentation (i.e., scene segmentation) is at the root of many novel services related to TV broadcast and in particular to the TV-on-Demand service [11]. Scene segmentation gives the possibility to obtain the table of contents for a video. A scene can be formed based on different units (e.g., temporal, topical), depending on the type of video considered, given that it should contain coherent shots that have a meaning for the viewer. Therefore, the definition of a scene can vary, which has lead to the development of approaches with different objectives when structuring the TV content.

The existing structuring approaches can be classified into approaches that use prior knowledge of the program content or not. Using prior information usually leads to the development of specific systems, that can structure certain types of videos. Such systems are proposed in [144], [81] and [33], where the authors use hidden Markov models to structure sports programs with soccer and tennis matches. Other specific systems are dedicated to structuring TV news by classifying the video shots into several classes (e.g., anchorman, news reports, weather forecast) [12, 35]. Differently from these approaches, that aim at structuring the entire video, several approaches focus on detecting structural elements in the videos, like anchorperson [75, 120, 51] or soccer goals [23, 6]. Still these approaches are specific. There exist also generic approaches that aim to extract the entire structure or some structural elements without any prior knowledge. Such approaches generally rely on pattern discovery in the TV stream, searching for recurrent segments exhibiting audiovisual consistency [10] or using a micro-clustering technique that groups similar audio/visual feature vectors [11]. Some more novel approaches rely on grammar inference techniques to evidentiate the underlying structure of recurrent programs [111, 110]. The progress made in the language processing community lead to the development of technologies that can be applied on automatic/manual transcripts of TV shows to obtain the topical structure, and therefore a table of contents for the shows. Such approaches need to determine the boundaries between individual stories in the broadcast and find topically coherent video segments. In [131], the authors propose a technique inspired from image segmentation combined with latent semantic indexing, a technique employed in natural language processing, to extract the topical structure of a CNN news show relying on the manual transcript of the show. In [60], the authors obtain the topical structure of TV shows, by adapting a linear topic segmentation approach used for written text to automatic transcripts. They add semantic relations and confidence measures for the words pronounced in the shows to improve the segmentation accuracy.

To sum up, as it becomes feasible to extract multi-modal features from audiovisual content, various approaches have emerged for structuring the content. However structuring the content is useless unless it produces *Value*. It can be safely stated that value is the primary goal of the current technological revolution. Thus the next natural step is to leverage the structured content to obtain value. For example, businesses mine for data patterns to learn their clients preferences and search to provide new multimedia technologies that offer a large spectrum of functionalities to answer users needs. In the context of audiovisual data a user might want to be able to follow the evolution of an event in time, or how it is presented

by different programs, to know if a video is suitable for a child to watch, to discover new interesting and unexpected information starting from a certain video segment on a topic of interest, etc. Remarkable techniques have been developed by researchers, with such objectives in mind, especially in benchmarking initiatives like MediaEval and TRECVid [105].

Among the developed techniques we can find solutions for automatic detection of manipulation and misuse of multimedia content [15]. Such solutions can assist professionals in the process of verifying if the information is trustworthy or not. Other techniques were developed to index people in large TV archives, under real-world conditions (i.e., with no pre-set list of people to index), in order to make the archives searchable [107] and able to answer questions like "who speaks when?" and "who appears when?". Solutions that can help users find videos that fit their particular mood, age or preferences have also been proposed in the context of MediaEval [130]. Detecting the emotional impact of a movie could help improve search or recommendation scenarios. While, the detection of violent content could help parents choose the materials that are more suitable for their children to watch. In recent years, researchers became interested also in providing solutions for video hyperlinking in large video collections in the context of these benchmarking initiatives [40, 105]. Hyperlinking videos is an important feature to have for a large collection since it can encourage further exploration and discoveries of relevant, interesting or unexpected information. Another initiative in the same direction of exploiting the audiovisual structure was the topic detection and tracking (TDT) study [3]. This study intended to explore techniques for detecting the appearance of new topics and tracking their reappearance and evolution in a stream of broadcast news stories.

All the previously mentioned initiatives indicate that a key future trend is to provide technologies that offer diverse and complex features to improve access to information in the current data deluge of audiovisual content. This brings forward the challenges of structuring and exploiting this content. And these are the two main aspects we are interested in addressing in this thesis. Solutions for tackling these challenges would allow users from a variety of backgrounds, such as: journalists, students, professionals, researchers, archivists, and home users, to benefit from technologies as the ones just mentioned. They would be offered the possibility to discover, navigate and search large video collections. More, extracting value from these collections, opens up libraries in a way that increases their economic and/or cultural value [38].

1.2 Contributions

The first goal of this thesis is to provide automatic and generic techniques for topical structuring of audiovisual data. Secondly, we aim at studying the implications of the produced structure for various NLP related tasks, such as video hyperlinking, video anchor detection and summarization. A necessity, in terms of topical structuring, is to provide solutions that are automatic and generic and therefore can be applied to any kind of audiovisual data. Indeed, given the impressive amounts of audiovisual data, it becomes difficult to create specific solutions for each type of TV show. Therefore it has become a necessity to design unsupervised approaches that can deal with heterogeneous TV content. With this requisite in mind, we base our approaches on the automatic transcripts of the speech pronounced in the programs, being independent of the type of documents considered. The automatic

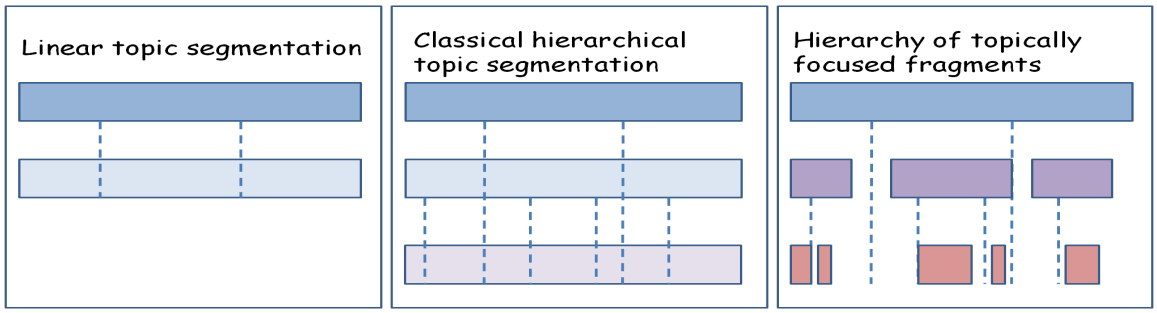


Figure 1.1: Generic respresentations of (a) linear topic segmentation (b) classical dense hierarchical topic segmentation vs (b) hierarchy of topically focused fragments. Vertical lines illustrate topic and sub-topic frontiers.

transcripts are obtained via an Automatic Speech Recognition (ASR) system.

Identifying the topical structure means finding the overall semantic organization of the video. Two different segmentation structures can be distinguished: linear and hierarchical. Linear topical segmentation aims to structure the data into successive topics. Hierarchical topics segmentation techniques consist in dividing a main topic into sub-topics, which in turn can be further divided into sub-sub-topics. We approach both types of topical structures, leveraging the two different views over the internal organization of the data. Generic techniques for topic segmentation usually exploit the lexical cohesion (especially the techniques used on textual data), which is independent of the type of textual documents considered and does not require a learning phase. Relying on the lexical cohesion means analyzing the distribution of words in order to identify significant changes in vocabulary which hint to changes in topic. We start by providing a new automatic and generic technique for linear topic segmentation. Then, we move towards hierarchical topic segmentation. We first investigate to which extent hierarchical segmentation can capitalize on lexical cohesion and show that the lexical cohesion is not sufficient to retrieve topical reference segmentations. As a result, we propose a new way of thinking about this kind of hierarchical structure, moving from classical dense segmentation to a hierarchy of topically focused fragments. We evaluate the new structure by a qualitative comparison to classical dense segmentation and in the context of automatic summarization. Figure 1.1 gives a generic representation of the structures resulting from these contributions. Figure 1.1(a) is the representation for linear topic segmentation, where a TV show transcript is divided into main topics, Figure 1.1(b) represents classical hierarchical topic segmentation, where the main topics are divided into sub-topics, which in turn can be divided. Departing from the traditional thinking, the idea in Figure 1.1(c) is to spot topically focused fragments that are not necessarily contiguous and organize the fragments at various levels in a hierarchical way.

After providing solutions for obtaining the topical structure of audiovisual data, we study the implications of these structures in the context of the MediaEval benchmarking initiative for the Search and Hyperlinking task. The purpose is to improve the user navigation experience in terms of video hyperlinking. Automatic generation of hyperlinks in video data is a subject with growing interest, offering information seeking and browsing capabilities in addition to search in large video collections. Consider the following example. A user starts

a search in a collection of videos with the query "things to see in London". A list of videos is returned and the user starts watching one. At some point, the British "red telephone box" is mentioned: Based on the video fragment about the telephone box, the user would like to find out more information about this topic, something he/she was not explicitly looking for at search time. The new information is returned in the form of short video fragments that are related to the new "query", i.e., to the video fragment where red telephone box is mentioned. Ideally, these fragments bring additional information, e.g., in our example, about the history of the phone box, about the inventor of the telephone, about the phone box used as a time traveling machine in "Doctor Who" TV series. This example illustrates a search and hyperlinking scenario. Typically the new "query" (red telephone box) is called an *anchor* and represents a segment for which the user requests other links (i.e., details on demand), while the short video fragments retrieved to create the links are called *targets*. Thus, the goal of video hyperlinking is to create links based on a given anchor, automatically selecting target segments that offer complementary information not found at search time. In our approaches, we rely on the topical structure for the automatic identification of precise targets and anchors. We then develop a novel approach that aims to address two essential aspects of hyperlinking, namely, diversity in the links and link justification. For this approach, we propose different strategies exploiting a hierarchy of topic models as an intermediate representation to compare the transcripts of video segments. These hierarchical representations offer a basis to characterize the hyperlinks, thanks to the knowledge of the topics which contributed to the creation of the links, and to induce diversity in the links by choosing to give more weights to either general or specific topics.

1.3 Organization of the Manuscript

The manuscript is organized in 2 parts.

The first part contains 3 chapters (Chapters 2 to 4) and presents the first set of contributions. We focus on the subject of automatic topical structuring of audiovisual data.The goal is to improve current segmentation strategies, to understand their limits and how we can alleviate them. We start with the general background on automatic topic segmentation, presenting the fundamental concepts and the state of the art. Chapter 3 presents a new technique for linear topic segmentation. In Chapter 4 we move towards hierarchical topic segmentation. We study the limits of current hierarchical segmentation approaches and provide a new way of structuring the data by constructing a hierarchy of topically focused fragments instead of a dense segmentation.

The second part includes the next 3 chapters (Chapters 5 to 7) and presents the second set of our contributions. We mainly focus on leveraging the techniques proposed in the first part for video hyperlinking. Chapter 5 provides the framework in which our approaches are tested, namely the MediaEval benchmarking initiative. The following chapters, each address important aspects of video hyperlinking: in Chapter 6 solutions for precise target selection and anchor detection are investigated; Chapter 7, focuses on the problem of diversity and link justification; Finally, Chapter 8 concludes this work and presents the contributions and the perspectives brought by our solutions.

Part I

Linear and hierarchical topical structuring of TV shows

Chapter 2

Background: Automatic topic segmentation

Contents

2.1 Fundamentals on topic segmentation 17
2.2 State of the art 20
2.3 Automatic transcripts of TV shows 30
2.4 Overview 33

This chapter introduces the grounds on which lies the work developed in the first part of this thesis. The first section of this chapter will provide an overview on topic segmentation, containing some fundamental theoretical notions and the features used for realizing topic segmentation. In Section 2.2, existing solutions for linear and hierarchical topic segmentation are presented. The third section focuses on the TV show transcripts peculiarities and contains also details regarding the TV show corpus on which we test our techniques. Section 2.4 provides a brief overview over the chapter.

2.1 Fundamentals on topic segmentation

Algorithms developed for evidentiating the topical structure of documents aim at automatically detecting frontiers that define topically coherent segments in a text. This is a difficult task, intensively studied and debated. In this section, we will provide the fundamental notions regarding the concept of topic segmentation. Subsection 2.1 focuses on defining the concept of topic, while, Subsection 2.1.2 presents the features exploited by the existing methods for topic segmentation.

2.1.1 Topic definition

Given our goal to detect the overall structure of TV shows in terms of topics, it is essential to provide a clear view of what a topic is. We will start by discussing the specific notion of topic and granularity at the level of topic and follow with the specific notion of topic in the context of TV shows.

The general concept of topic Defining the concept of topic precisely is not trivial and a large number of definitions have been given by linguists. Additionally, there is an absence of uniformity in the terminologies used (topic, theme, topos, subject, center, motive, etc.) which leads to confusion since some authors use these terms interchangeably and some make a clear distinction between them. As mentioned in [18], the notion of topic is problematic since there are two viewpoints to defining it: first there is the notion "sentential topic" [17], which refers to the sentence-internal "topic-comment" relationship being a purely syntactic function such as subject or object; second is the notion of "speaker's topic" [17], of what an individual speaker personally feels is being talked about at a given moment being more about the semantic content.

Among the definitions that can be found in the literature, for the concept of topic, are: a syntactic position, the starting point of an utterance (i.e., sentence or sequence of words), an important idea, an element ensuring the coherence of the discourse, the central point of a description, what the discourse is about, etc. While the concept of topic remains elusive, some researchers sustain that the definition remains intuitive [112], vague and mysterious [54] and that it seems to be a common shared knowledge allowing the use of the concept without having to explicitly define it [89]. Brown and Yule reinforce this paradoxical situation: "Yet the basis for the identification of «topic» is rarely explicit. In fact, «topic» could be described as the most frequent used, unexplained, term in the analysis of discourse.". Reinhart [113] suggests the 'what-about', 'as-for' and 'said-about' tests to identify topics. However, these tests do not seem to hold for all types of texts as criticized in [137]. Brown and Yule [17] discuss at length the difficulty of defining a topic and note: *"The notion of 'topic' is clearly an intuitively satisfactory way of describing the unifying principle which makes one stretch of discourse 'about' something and the next stretch 'about' something else, for it is appealed to very frequently in the discourse analysis literature. Yet the basis for the identification of 'topic' is rarely made explicit"*. To skirt the issue of defining a topic, Brown and Yule suggest to focus on topic-shift markers and to identify topic changes, what most current topic segmentation methods do.

The definition for a topic shift can be derived from [19]: *"Our data [...] suggest that as a speaker moves from focus to focus (or from thought to thought) there are certain points at which they may be a more or less radical change in space, time, character configuration, event structure, or even world [...] At points where all these change in a maximal way, an episode boundary is strongly present."*

Topic hierarchy While for the notion of topic various definitions can be found, being the objective of numerous studies, the concept of granularity at the level of topics has been poorly approached even though the structure of discourse is known to have a hierarchical form [57, 95]. A distinction between topics and sub-topics is proposed in [24]: a coherent thematic segment, considered as a sequence of sentences that are interdependent, characterizes a sub-topic if its interpretation is dependent of another thematic segment. In [112], the

difference between specific topic and generic topic can be seen as a representation of different granularity, in which the specific topic would be the sub-topic of the generic one. The definition of specific and generic topics is based on differential semantics. For differential semantics, the sense of a text emerges from the structuring of the sememes space, where sememes represent the words of the vocabulary. The sememes are defined one with respect to another through semes, which are semantic relations. Therefore any semantic unit can be broken down into semes [68]. A generic seme indicates that the sememe belongs to a semantic class. A specific seme distinguishes a sememe from all the other sememes of the same class. Based on these notions, Rastier defined the topic (sub-topic) as a stable structure of semes, generic (specific).

As mentioned previously most topic segmentation methods search for topic shifts. While, these are more prominently marked, fluctuations in topics at a fine grained level are more difficult to identify. A differentiation between topic and sub-topic changes can be inspired from [21], where the changes at the coarse level are seen as topic shifts and those at a fine-grained level as topic drifts. When the topic drifts smoothly from the information presented in the first span to the information presented in the second, similar elements are in focus in both textual units.

Topic and sub-topic in the context of TV shows Several pieces of work dealing with topic detection and tracking in streams of data (e.g., broadcast news) have relied on the definitions of topic and event proposed in the context of the *Topic Detection and Tracking* project (TDT), organized by NIST [147, 45, 2]. The focus of this project is to find segments, in broadcast news, that are topically correlated and consists of three major tasks: segmentation, detection and tracking. To evaluate the methods proposed for TDT, a guide for making a clear distinction between the notion of event and topic was created. An event in the TDT context refers to the subject of a story itself and is about the "who, what, where, when and why" in a story. An event happens at a specific place and time while a topic is considered to be an event together with all the events directly related to it [45] and becomes background among events. The main difference between event and topic is that the event is relatively short and evolves in time, while the topic is more stable and long. An event is considered to drift, while a topic does not. For a more clear view on the distinction between topic and event we retake the example given in [45] regarding the story of 'Kobe Japan quake':

The event contains reports on the damage, location, nature of the quake, rescue efforts, etc., while the topic is *'Kobe Japan quake'*. Other events on the same topic contain reports such as *'Emergency Work Continues After Earthquake in Japan'*, *'Death Toll Mounts in Japan Earthquake Zone' and 'U.S. Visitor Describes Tokyo Quake'*.

2.1.2 Features used for topic segmentation

The features of the data to segment should help identify cohesive segments, where cohesion as defined in [101] is a term for "sticking together". Cohesion manifests in the text through various relations between the elements in that text. These relations are: reference, ellipsis, substitution, conjunction and lexical cohesion. Most of the techniques for topic segmentation rely solely on the lexical cohesion, i.e., identifying segments with a consistent use of vocabulary, either based on words or on semantic relations between words, since it is domain independent and does not require any learning phase. More details on the notion of lexical

cohesion and other characteristics of the data that can be exploited for topic segmentation are given next.

Lexical cohesion Relying on the lexical cohesion means analyzing the distribution of words in order to identify significant changes in vocabulary that hint to changes in topic. Therefore topically cohesive segments can be found based on the words they contain and their positions. There are different ways for evidentiating the lexical cohesion, the most popular being reoccurrences of words or related words and lexical chains. Lexical chains can connect different occurrences of a word, other words semantically related, etc., as long as these occurrences are at a distance smaller than a threshold.

The measure of lexical cohesion is very sensitive to the errors found in TV show transcripts and the reduced number of words in these transcripts. Therefore, we can find in the literature some efforts like those presented in [59] to adapt the lexical cohesion to the peculiarities of the words pronounced in TV shows (e.g., use confidence measures for the words).

Linguistic markers Besides the lexical distribution of the words, various markers can be considered to highlight the topic changes: prosodic cues (e.g., pitch, pause, duration), discourse markers (e.g., first, accordingly, next), visual markers, etc. Visual markers are not available for transcripts, but in classical long texts which have a structure with chapters and sections, with titles and/or headings, some words may stand out through contrasting disposition (e.g., indentation) or typography (e.g., bold, italic) [31]. These visual markers are important since they are part of the writer's intentions.

Discourse markers, have been at the center of many studies in computational linguistics, having an important role in various discourse processing tasks. The role of these markers is detailed in [57] and the authors consider that "certain words and sentences and more subtle cues such as intonation or changes in tense and aspect" are "among the primary indicators of discourse segment boundaries". Indeed, they can indicate continuity and discontinuity in discourse.

Finally, prosodic cues are important for spoken communication and they can be used to identify the topical structure since they can reflect various features of the utterances as presented in [61]: the emotional state of the speaker; whether an utterance is a statement, a question or a command; whether the speaker is emphasizing, contrasting or focusing a particular item.

2.2 State of the art

Substantial efforts have been reported in the literature on the task of topic segmentation, most of them for structuring classical written texts. Two segmentation structures can be distinguished: linear and hierarchical. The first part of this section is reserved for presenting the existing work on linear topic segmentation and the second part is dedicated to hierarchical topic segmentation techniques.

2.2.1 Linear topic segmentation

Various methods for linear topic segmentation are described in the literature, most of them relying on the lexical cohesion. Still there are several approaches that exploit discourse markers to improve the detection of thematic frontiers in oral documents, e.g., [57] and [88]. In [71], the authors examined automatic topic segmentation based on prosodic cues for English broadcast news. In [61], the authors proposed the use of prosodic information to improve an ASR-based topic tracking system for French TV broadcast news. Linguistic markers are however often specific to a type of text and cannot be considered in a versatile approach as the one we are targeting. As mentioned in [127], the prosodic cues that work for TV news do not necessarily work for other TV shows (e.g., reports or debates). If for TV news the segmentation can benefit from additional information like pitch or intensity, or discourse markers that are used for introducing a new topic, for TV reports this kind of information can be misleading. The TV reports are characterized by outdoor investigations, where many people are interviewed. They have various accents, different ways of presenting the information, of emphasizing on topics, etc. It is difficult to find an agreement regarding these and the linguistic cues which could help distinguish between the different topics discussed. For this reason the use of linguistic markers for topic segmentation can make the technique specific.

Generic techniques usually exploit the lexical cohesion (especially the techniques used on textual data), which is independent of the type of textual documents considered and does not require a learning phase. We reiterate here the general idea behind lexical cohesion: A significant change in vocabulary is a sign of topic shift. This general idea translates into two families of methods with two radically different strategies. Global methods, where a measure of the *lexical cohesion* can be used to globally determine segments exhibiting coherence in their lexical distribution [114, 99, 136]. Local methods, where shifts in the use of vocabulary can be searched for to directly identify the segment frontiers by measuring the *lexical disruption* [66].

2.2.1.1 Local methods

Local methods [66, 41, 69, 29] locally compare adjacent fixed size regions, claiming a boundary when the similarity between the adjacent regions is small enough, thus identifying points of high lexical disruption. In the seminal work of Hearst [66], proposing the TextTiling algorithm, a fixed size window divided into two adjacent blocks is used, consecutively centred at each potential boundary. The content of the adjacent blocks is represented through a vector, each vector containing words that are weighted according to their frequency tf (i.e., term frequency):

$$w_{t,d} = \mathrm{tf}(t,d)$$

where $w_{t,d}$ is the weight associated to term[2] t in block d. A higher weight implies that the term is more relevant for the block. Different weighting schemes can be used, like tf-idf (term frequency-inverse document frequency), okapi (similar to tf-idf but takes better into account the lengths of the blocks), etc. After changing the representation space of the blocks

[2] *term* is used as a synonym for *word* and not in the context of terminology

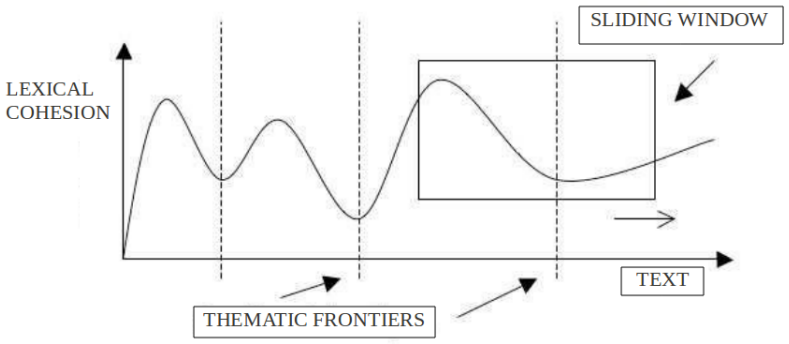


Figure 2.1: Local thematic segmentation based on sliding window.

in a vectorial one, the cosine similarity between two adjacent blocks is computed as

$$cos(b1, b2) = \frac{\sum_{t=1}^{n} w_{t,b1} \times w_{t,b2}}{\sqrt{(\sum_{t=1}^{n} w_{t,b1}^2) \times (\sum_{t=1}^{n} w_{t,b2}^2)}},$$

where t ranges over all the terms in the document. Blocks with similar content will have large cosine value, approaching 1, meaning that the angle between the vectors is close to 0. Similarly, high disrupting boundaries will correspond to low similarity measures. Similarity between the adjacent blocks is computed at each point, the resulting similarity profile being analyzed to find significant valleys which are considered as topic boundaries [66, 67, 65], as illustrated in Figure 2.1.

Other variations of the TextTiling algorithm can be found in the literature, where the authors tried to improve the results by changing the vectorial representation and the measure of similarity, as done in [69, 41]. In [41], the authors employ semantic relations to overcome the reduced number of repetitions due to the usage of synonyms in their data. In [29], the authors propose that instead of directly comparing the representations of adjacent blocks, to use an indirect comparison for evidentiating the semantic similarity between two blocks of utterances, even if they do not share common vocabulary. This technique is called *vectorization*. It has been introduced and implemented in [30], in a standard information retrieval (IR) scenario. The vectorization technique will find two blocks to be similar if they are similar to the same pivot documents (i.e., reference documents). It is an interesting technique for topic segmentation since it can overcome the drawbacks represented by poor repetition of vocabulary and by the presence of synonyms. An exploratory study on how to create and choose the pivot documents for topic segmentation is presented in [127].

As mentioned in Subsection 2.1.2, lexical cohesion can be evidentiated also through lexical chains. The advantage of using lexical chains is that they capture more than the repetitions of words, they capture also the locality of those repetitions. Morris and Hirst [101] were the first to use lexical chains in the context of linear topic segmentation. Their work shows that the ends and beginnings of the chains can be correlated with the topical structure of the text. The LCSeg algorithm, proposed in [46], relies on lexical chains based only on word repetitions and has been particularly influential in dialogue segmentation. In [133]

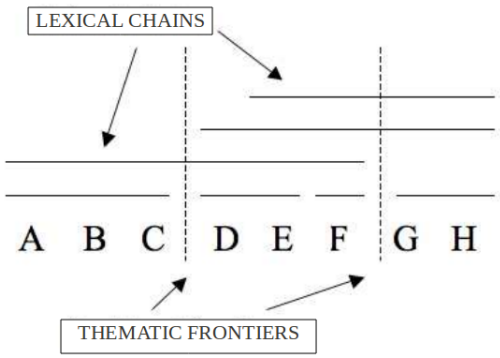


Figure 2.2: Local thematic segmentation based on lexical chains

	local methods
lexical cohesion	TextTiling [66]; [41]; [69]; Vectorization [29]
lexical chains	[101];[133];[129]; LCSeg [46]

Table 2.1: List of local methods for linear topic segmentation

the authors create lexical chains based on word repetitions and semantic relations from the WordNet thesaurus [96]. While in [129], the authors rely also on the syntactic information of the words to vary the importance of a chain in the text as opposed to the classical binary case when the chain is either active or not on each portion of the text. All these studies have integrated lexical chains in a TextTiling-like approach to perform the segmentation, by using the chains to represent the two blocks of the sliding window. The main principle is assigning a score for each potential topic frontier based on the chains starting, ending or crossing that frontier and also the strengths of these chains. The strength usually comes from the length of the chain and number of occurrences of the word the chain is build for. A frontier is proposed at places where few lexical chains are cut, as can be seen in Figure 2.2.

Table 2.1 gives a summary of the local methods for linear topical segmentation, previously discussed.

2.2.1.2 Global methods

Global methods seek to maximize globally on the text the value of the lexical cohesion on each segment resulting from the segmentation. Several approaches have been taken relying on self-similarity matrices [25, 77, 26], such as dot plots [114], or on graphs [92, 136, 98]. In [114], the author uses a similarity matrix to plot points which correspond to word repetitions on the dotplot as in Figure 2.3. For example if a word appears at position X and Y in a text, then four points will be plotted: (X,X), (X,Y), (Y,X) and (Y,Y). The topic boundaries could be identified visually as the regions along the diagonal (i.e., the line X=Y) that are darker than other regions. The author proposes an algorithm to minimize the regions not contained in the squares along the diagonal. The algorithm will select the boundaries trying to keep a low density in the outside region and will stop when either the outside density increases or a certain number of boundaries are added. This approach has been improved in [25], where the inter-sentence similarity matrix is replaced by a ranking scheme and the

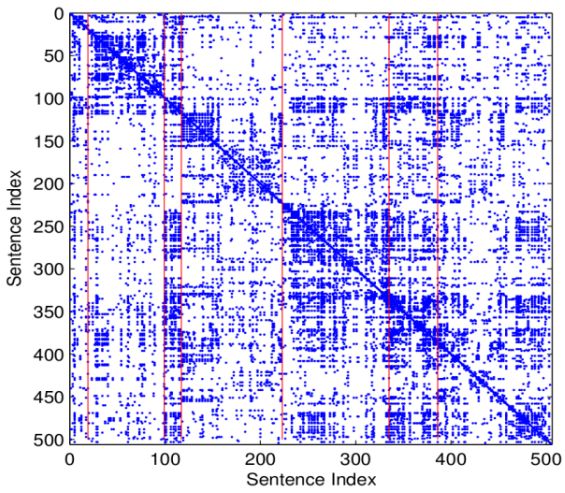


Figure 2.3: Dotplot representation for segmentation.

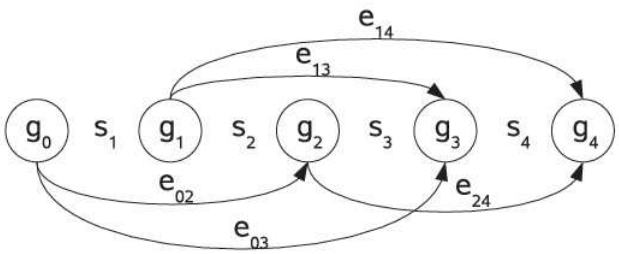


Figure 2.4: Graphic representation

cosine similarity measure. More, in [26], the authors showed that using a latent semantic analysis based metric for the similarity measure could improve the accuracy. Another approach that relies on similarity matrix is presented in [77], where the topic segmentation problem is converted into an image segmentation problem. The authors use a technique called anisotropic diffusion to the image representation of the similarity matrix to enhance the semantic cohesion of topical groups of sentences, while sharpening topic boundaries. In this last work the similarity matrix is build by computing the similarity between sentence pairs.

There exist also approaches that rely on graphs instead of matrices. A typical and state-of-the-art algorithm is that of Utiyama and Isahara [136] whose principle is to search globally for the best path in a graph representing all possible segmentations and where edges are valued according to the lexical cohesion measured in a probabilistic way. The lexical cohesion value of a segment is seen as the capacity of a language model learnt on that segment to predict the words in the segment. An illustration of the graphical representation of the text is given in Figure 2.4. The nodes of the graph represent potential frontiers and the edges thematic segments. This algorithm was generalized in [37] into a fully Bayesian version by marginalizing out the language model using an approach similar to Latent Dirichlet Allocation (LDA), which seems to improve the segmentation. However, the number of segments is

	global methods	
	no training	training
self-similarity matrices	C99 [25];CWM [26]; [77]; [114]	
graphs	MinCut[92];TextSeg [136];BayesSeg [37];	[98]; [146]

Table 2.2: List of global methods for linear topic segmentation

assumed to be provided as prior knowledge. Another global approach is the work proposed in [92], where the segment boundaries in the graph are found using the notion of normalized cut. However, this method assumes that the number of segments to find is known beforehand which makes it difficult for real-world usage.

As opposed to these approaches that require no training, there exist several attempts for identifying topic boundaries employing supervised or unsupervised training. In [98], the authors propose an extension of the algorithm in [136] by doing also topic labelling in addition to segmentation. For this, they use probabilistic topic modelling. First, they identify the latent topics in the document using LDA and modify the segmentation algorithm in [136] by associating with each edge in the graph a vector containing the probability of the segment, corresponding to that edge, given the latent topics detected. LDA is a generative model for documents and the main idea is that latent variables exist which determine how the words in documents might be generated. Training LDA on some documents means finding the best latent variables in order to explain the data. As a result, documents are observed as mixtures of latent topics and the topics are probability distributions over words. Therefore, by inferring the most likely topics from the observed words, the positions of the topical boundaries can be derived. In [146], the authors applied another generative model, hidden Markov model, to segment broadcast news. They used a segmented training data set to estimate the topic transition probability and the topic language models. The limitations of parametric topic models is represented by the difficulty to determine the number of topics for a data set, since computing the optimal number of topics is time-consuming and the optimal number varies for each data set.

Table 2.2 gives a summary of the global methods used for linear topical segmentation, previously discussed.

2.2.1.3 Challenges of current linear topic segmentation approaches

When the lengths of the respective topic segments in a text (or between two texts) are very different from one another, local methods are challenged. Finding out an appropriate window size and extracting boundaries become critical with segments of varying length, in particular when short segments are present. Short windows will render comparison of adjacent blocks difficult and unreliable while long windows cannot handle short segments. The lack of a global vision also makes it difficult to normalize properly the similarities between blocks and to deal with statistics on segment length. While global methods override these drawbacks, they face the problem of over-segmentation due to the fact that they mainly rely on the sole lexical cohesion. Short segments are therefore very likely to be coherent which calls for regularization introduced as priors on the segments length. However, knowing beforehand the number of segments or their duration is not a realistic assumption. And even then, it is hard to deal with varying segment lengths.

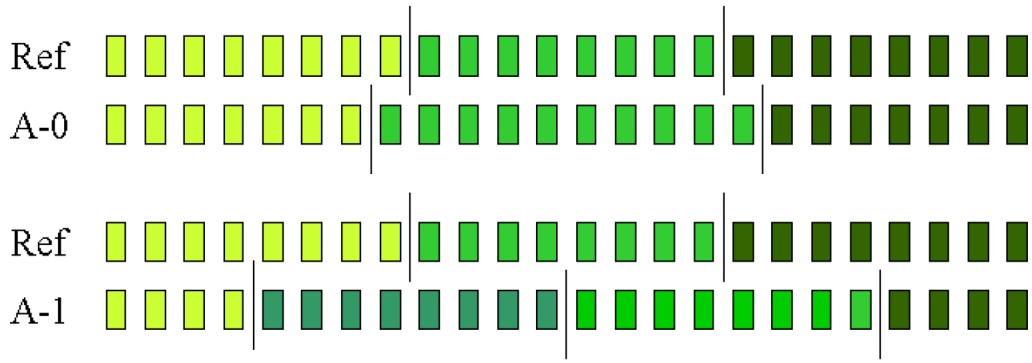


Figure 2.5: A reference segmentation *Ref* and two hypothesized segmentations *A-0* and *A-1*. The boxes indicate utterances and spaces between boxes indicate potential frontiers.

2.2.2 Evaluating linear topic segmentation

The evaluation of the quality of a linear topic segmentation is usually done either by comparing the segmentation with a reference one using a metric, or by evaluating the segmentation in the context of another application (e.g., summarization [7], information extraction [94]).

The metric used for evaluation can be inspired from the IR field, like precision, recall and F1-measure. These measures are defined as:

$$Recall = \frac{|H \cap R|}{|R|} \qquad Precision = \frac{|H \cap R|}{|H|}$$

$$F1 - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \ ,$$

where $|H|$ ($|R|$) is the number of frontiers in the hypothesized (reference) segmentation. *Recall* refers to the proportion of reference frontiers correctly detected and *Precision* corresponds to the ratio of hypothesized frontiers that belong to the reference segmentation. The F1-measure combines recall and precision in a single value. These measures are criticized in [9] and [106] and the following problems are raised: improving one causes the score for the other one to drop (i.e., adding more boundaries will tend to improve the recall and at the same time reduce the precision); the F1-measure is hard to interpret; precision and recall are not sensitive to near misses. If we take the example from [106] depicted in Figure 2.5, both segmentations *A-0* and *A-1* fail to match the reference segmentation *Ref* precisely. Thus the precision and recall score for both will be 0. However, the *A-0* segmentation is close to the reference one, while the *A-1* one is completely not. Thus it would be useful to be able to differentiate between the two and have a metric that penalizes one less harshly that the other. In [9], a new measure which should overcome the problems raised by precision and recall is proposed, called P_k. This solution is based on a sliding window of size k, which parses the reference segmentation and the hypothesized segmentation proposed by the system under evaluation. P_k consists in evaluating the similarity between the two segmentations inside

the window and is computed in the following manner:

$$P_k(r,h) = \frac{1}{N-k}\sum_{i=1}^{N-k}(\delta_r(i,i+k) \odot \delta_h(i,i+k)) \ ,$$

where the binary function $\delta_r(i,i+k)$ ($\delta_h(i,i+k)$) indicates whether the segment endpoints i and i+k belong to the same segment in the reference (hypothesized) segmentation and N is the number of atoms (i.e., sentences, utterances, words, depending at which level the segmentation was made) in the text. $\odot$ is the *XNOR* operator which translates to "both or neither". Several limitations of this measure are discussed in [106] such as: the sensitivity to the variations of segments dimensions; penalizing false negatives more than false positives; number of boundaries ignored, etc. Therefore, the authors of [106], have proposed a new error measure called $Window_{Diff}$, which instead of checking for a boundary in the window, counts the number of boundaries inside the window. This measure is considered more resistant to the dimensions of segments and is defined as:

$$Window_{Diff}(r,h) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|b(r_i,r_{i+k}) - b(h_i,h_{i+k})| > 0)| \ ,$$

where b(i,j) is the number of boundaries between positions i and j in the text. The segmentation algorithm is penalized if $b(r_i,r_{i+k}) \neq b(h_i,h_{i+k})$.

Discussion Because both *Pk* and *WindowDiff* employ the use of a sliding window, lower weights are given to the frontiers close to the beginning or ending of a document. In addition, the author of [22] considers that *WindowDiff* favours segmentations with fewer number of frontiers. In [102], a rigorous analytical explanation of the biases of *Pk* and *WindowDiff* is provided. The authors show that these measures are biased in favour of segmentations with fewer or multiple adjacent segment boundaries and that several topic segmentation algorithms benefit from this. For example, the segmentation algorithms in [37] and [46] benefit from clumped boundaries (i.e., boundaries placed close to one another). In [102], the authors use several measures to compare existing segmentation strategies and note that LCSeg from [46] and C99 from [25] are not significantly better than random, while BayesSeg [37] and TextSeg [136] are the only ones significantly better than the other algorithms, though the difference between the two is not significant. The evaluations are done on 25 meetings from ICSI meeting corpus. The authors of [102] propose using windowed variants of precision and recall when evaluating segmentation strategies, which is actually how they were used in the context of topic segmentation by allowing a tolerance between the reference and hypothesized frontiers [114]. Precision and recall are not sensitive to variations of segment length contrary to the *Pk* measure [9] and do not favour segmentations with a few number of frontiers as *WindowDiff* [106].

2.2.3 Hierarchical topic segmentation

In the case of hierarchical topic segmentation, a first approach is to apply a linear topic segmentation algorithm recursively [22, 59]. Several linear topic segmentation algorithms were cast in a hierarchical framework, among which TextSeg [136], LCSeg [46], BayesSeg [37], C99 [25] and CWM [26]. One of the main challenges is to decide when to stop. Additionally,

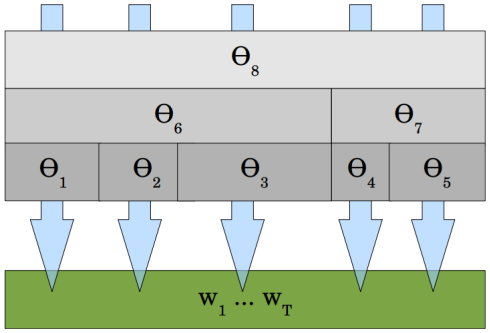


Figure 2.6: Pyramid of language models as illustrated in [36]. Each word w_t is drawn from a mixture of the language models located above t in the pyramid.

a segmentation error at a higher level in the hierarchy can be propagated towards the lower levels.

Most of the times the recursive application is done directly, without any adaptation to the inferior levels in the hierarchy. However words that have contributed at identifying the segments at one level should not have the same impact for other levels. One study in the context of segmenting TV shows transcripts has adapted the algorithm in [136] and modified it to reflect the distribution of the vocabulary at different levels in the hierarchy [59]. This work helped obtain a hierarchical topic segmentation for two levels, remaining a challenge to obtain it at all the levels in the hierarchy.

A few models have been proposed to explicitly model the hierarchical segment structure. HierBayes [36] is an unsupervised algorithm formalized in a Bayesian probabilistic framework. The underlying principle is that each word in a text is represented by a language model estimated on a portion, more or less important, of the text. A pyramid of language models, as shown in Figure 2.6, is build, having high-level language models explain words throughout large parts of the document, while low-level language models explain only a local set of words. To obtain the hierarchical segmentation the algorithm has the objective to maximize the lexical cohesion of the segments at each level in the hierarchy, imposing that the frontiers for a superior level in the hierarchy are aligned to those hypothesized at inferior levels. The probabilistic segmentation objective employed is similar to that used for TextSeg in [136] with several differences: TextSeg uses maximum a posteriori estimates of the language models, rather than marginalizing them out as done in HierBayes. Also, TextSeg relies on a minimum description length criterion to determine segmentation granularity, while HierBayes needs the expected segment durations, considering that the granularity of the segmentation is a user-defined characteristic.

In [80], the authors propose to use the hierarchical affinity propagation graphical model introduced in [55] to extract the hierarchical topic structure. This similarity-based approach searches for the best assignment of segments to form a topical tree. It takes as input a matrix of similarities between atomic units of text (i.e., sentences or paragraphs), the number of levels in the topical tree and a preference value that controls the granularity of segmentation (i.e., how many segments are to be identified at each level) and captures a priori belief about the segment centres. Indeed, each segment is characterized by a centre which best describes its content. The objective function is net similarity, the sum of similarities between all centres

and the data points which they exemplify. To compute the similarity a large sliding window is employed with a size of at least twice the anticipated average length.

Finally, in [99] the authors propose a bottom-up approach that consists in connecting the segments that have a hierarchical link to infer the cohesive structure of the text. The authors construct lexical chains for each important content term (i.e., terms remaining after removing stopwords), containing the term and its related terms (including resolved anaphors). They rely on generic heuristics to extract the main topic of each sentence, e.g. the position in the sentence, persistency with previous sentences, bound pronouns, etc. These heuristics are considered generic for several languages that primarily have a Subject-Verb-Object order, such as English, French and Dutch. The starts, interruptions and terminations of lexical chains are considered to give valuable information on topic boundaries. The lexical chains give several hypotheses of topical structures, having topically coherent passages hierarchically or sequentially grouped. Topic shifts, nested topics and sequential topics are detected by combining the information of the chains with the sentence topics.

Hierarchical topic segmentation can also be viewed as a clustering task. The agglomerative hierarchical clustering technique, HAC, proposed in [145] was designed to extract a hierarchical topical structure. The author applied rules to convert the hierarchical segmentation obtained into a linear segmentation to evaluate it. The difficulty lies in deciding the placement of the boundaries. In [131], the authors combine latent semantic indexing (LSI) with a technique used for signal processing (scale-space segmentation). LSI [28] is used to represent the documents to segment as term-sentence matrices: each sentence S_i from a text is represented by a vector corresponding to the i^{th} column of the matrix. LSI uses a mathematical technique called singular value decomposition (SVD) which allows reducing the number of dimensions leading to a low-dimensional representation of the matrix. After that the authors apply scale-space segmentation, which consists in smoothing each dimension of the vectors independently. This is done by using a Gaussian kernel associated with multiple values of scale σ. The importance of the thematic frontiers is defined for each vector by analyzing the smoothing at different levels of scale. The difference of importance between frontiers will determine the hierarchical aspect of the segmentation. This difference of importance is however difficult to set.

The techniques presented above were mostly applied to standard texts and still face several challenges that have not been addressed in the literature. In the case of recursively applying a linear segmentation algorithm, it is difficult to decide when to stop segmenting. Additionally, the errors in the segmentation from one level get propagated to another which makes it difficult to assess the real potential of the method. Dealing with the lack of words and high coherence between the segments, especially at lower levels in the hierarchy, has proven extremely challenging. Another important aspect that is usually not taken into account when linear segmentation algorithms are applied recursively is that words that contributed at segmenting one level in the hierarchy should have different importance for obtaining other levels. Some of the algorithms that extract the hierarchy directly need information about the granularity level and expected segment durations which are not available in a real scenario.

2.2.4 Evaluating hierarchical topic segmentation

The evaluation of hierarchical topic segmentation is even more challenging than for the linear one since the problem of subjectivity regarding the concept of topic is significantly accentuated. In [22], a couple of hierarchical topic segmentation strategies were evaluated on a Wikipedia corpus using an error metric designed specifically for this kind of segmentation, called E_{P_k} (with values in $[0,1]$, where 0 means perfect segmentation). The new measure extends the P_k and $Window_{Diff}$ error measures used for evaluating linear segmentation. The E_{P_k} error is the weighted average of P_k measurements over a series of linear segmentations and is defined as:

$$E_{P_k} = \frac{1}{|R|} \sum_{i=1} c_i P_k(R_i, H_i) \ ,$$

where R_i (H_i) is the set of all boundaries of level 1 (i.e., first level in the hierarchy) through i in the reference (hypothesized) segmentation and c_i is the number of reference boundaries at level i in the hierarchy. In the first step, only the highest boundaries are considered and each following step includes one more level of boundaries. A constraint is imposed at each step, respectively that $|H_i| = |R_i|$, in order to overcome under/over-segmentation. In case the number of hypothesized frontiers is smaller than the number of reference ones, frontiers from a lower level in the hierarchy will be included. This constraint makes the evaluation not able to characterize the behaviour of real segmentations, which usually give under/over-segmentation.

Others have evaluated their techniques indirectly by using the result of the segmentation for other tasks (e.g., summary generation at various levels of detail [99]). In [59, 36], the evaluation is done at each level separately using measures for linear topic segmentation, the drawback being that a global error cannot be computed (an error at a higher level in hierarchy influences the segmentations at a lower level).

2.3 Automatic transcripts of TV shows

In this thesis we are interested in TV shows, in particular automatic transcripts, obtained via an ASR system. Automatic transcripts have various characteristics that differentiate them from written text. We present next these characteristics following with details regarding the TV show corpus we use in this thesis.

2.3.1 Characteristics

The goal of an ASR system is to provide the automatic textual transcript of the words pronounced in an input audio signal. Modelled in a statistical framework the goal of the ASR translates to finding the most probable sequence of words from a vocabulary given a sequence of observed acoustic features in the input audio signal. The transcripts obtained do not respect the norms of written texts: they are not structured in sentences but in utterances (i.e., sequences of words often separated by breath intakes) that are only loosely syntactically motivated; they contain no punctuation signs or capital letters; the transcripts can contain also an important number of wrongly transcribed spoken words. These errors can be due to the quality of the recording, to the presence of noise, to the difference of speaking styles or

to the presence of words not contained in the dictionary of the ASR system. These problems that lead to errors in transcripts are very common in TV shows, where the recordings often take place in noisy environments that lead to spoken words difficult to hear. Also there are cases where the interviewed people speak a foreign language or have a strong accent, adding difficulty to the recognition process.

In Figure 2.7 we present an example extracted from [59] of an automatic transcript for a French TV journal and in the left side its reference transcript.

Reference transcript	**Automatic transcript**
Dix-neuf cent quatre vingt-deux, un évènement vient de se produire, il s'appelle Amandine. Trois kilos quatre, cinquante et un centimètres, le premier bébé éprouvette français est né. Ici, le bébé exploit qui a un an soufflera ce mois-ci ses vingt-cinq bougies.	dix neuf cent quatre-vingt-deux un événement vient de se produire il s'appelle *amman dina* trois kilos quatre cinquante-et-un centimètres le premier bébé *éprouvait* français est né ici le bébé *exploite y a* un an soufflera ce mois ci ses vingt-cinq bougies

Figure 2.7: reference and automatic transcript of a TV journal extracted from France 2, 7/02/2007. The words in italic correspond to transcript errors.

All these specificities of automatic transcripts challenge the lexical cohesion criterion on which generic approaches for topic segmentation rely on. Errors in the transcripts impact the analysis of the word distribution. A word will not be considered as appearing multiple times if its reoccurrences are not transcribed the same way. Also, some TV shows are characterized by a high employment of synonyms (e.g., TV news) which leads to a poor word repetition degrading the performance of lexical cohesion.

2.3.2 TV show corpus

To evaluate the techniques we propose for automatic topic segmentation two data sets of automatic transcripts are used: one for linear and one for hierarchical topic segmentation.

Corpus for linear topic segmentation This corpus consists of 56 news programs ($\approx$1/2 hour each), broadcasted in February and March 2007 on the French television channel France 2, and transcribed by two different automatic speech recognition (ASR) systems, namely IRENE [73] and LIMSI [52], with respective word error rates (WER) around 36 % and 30 %. There are also 7 shows that have reference transcripts available. Each news program consists of successive reports of short duration (2-3 min), possibly with consecutive reports on different facets of the same news. The reference topical segmentation was manually established and 1203 segments are obtained in total. The topics are defined similarly to how events are described in the context of TDT by associating a topic with each report, i.e., placing a boundary at the beginning of a report's introduction (and hence at the end of the closing remarks). This TV transcript data set, which corresponds to some real-world use cases in the multimedia field, is very challenging for several reasons. On the one hand, topical segments

level in the hierarchy	number of frontiers	average duration of segments	average number of repeated lemmas per segment	average number of lemmas per segment
first	26	32 min max:55 min, min:22 min	268	1567
second	246	3.4 min max:20 min, min:7 sec	24	180
third	722	1.6 min max:5 min, min:7 sec	5	54

Table 2.3: Comparison of different levels of granularity from Envoyé Spécial corpus.

are short, with an average of 107 lemmas in each segment and a reduced number of repetitions per segment, synonyms being frequently employed. Moreover, smooth topic shifts can be found, in particular at the beginning of each program with different reports dedicated to the headline.

Corpus for hierarchical topic segmentation This corpus contains 7 episodes of a report show, *Envoyé Spécial*. Each report has a duration of about 2 hours and was automatically transcribed with the IRENE ASR system. Manual transcripts for 4 reports are also available. The reference segmentations were obtained by manually dividing the TV reports into topics and sub-topics. These segmentations have 3 levels of hierarchy: the first contains 26 frontiers, the second one 246 and the third 722. A topic corresponds to a news report and is defined as a topic in TDT, a sub-topic corresponds to different points of view or different aspects presented in the report, therefore is defined as an event in TDT. For the third level in the topical hierarchy of TV reports a sub-sub-topic is more difficult to define. For this work we will consider sub-sub-topics as different points of view or comments on the event. If we take again the example with the 'Kobe Japan quake' (Section 2.1), the sub-sub-topics of the event will be: the part on damage, the part on location, the part on the nature of the quake, etc.

In Table 2.3, a comparison between these levels is provided. The segments of the first level in the hierarchy can be characterized as long and relatively stable in size, the ones at lower levels are short and have few lemmas repetitions. Another characteristic of these TV reports is that they favour outdoor investigations, which lead to a high word error rate (WER) in the transcripts obtained from the ASR system.

In what follows we give an example of a topical hierarchical structure, for a manual transcript corresponding to a TV show of Envoyé Spécial from 01/22/2009:

level1, first 2 topics:
"Private home schooling, success or failure?",
"The flowers of discord".

level2, first 4 sub-topics of topic 1 from level1:
"a mother who wants to see how private home schooling works and how her children would be taught",
"general statistics over the private home schooling in France and the level of the teachers",
"Acadomia, an institution that recruits teachers for home schooling",
"Completude, another institution that recruits teachers".

level3, the 5 sub-sub-topics of sub-topic 1 from level 2:
"Private home schooling in general",
"A mother turns to private homeschooling and expresses the wish to evaluate the teachers",
"The mother spies on her daughter's English lesson",
"The math teacher for her son does not show up and does not announce",
"The mother analyses the experience and concludes that she would not pay for private homeschooling".

In this example, at the first level we distinguish 2 different topics. The sub-topics (events) presented are different points of view over the same fact (first topic): private home schooling. Each event contains different aspects of the subject addressed that translate to sub-sub-topics.

2.4 Overview

In this chapter we have introduced several existing approaches for obtaining a topical structure of documents, linear or hierarchical. The generic approaches, relying on the notion of lexical cohesion, can be applied to any text-like data, including automatic transcripts of TV shows. Generic techniques raise more interest since they can offer the possibility to topically structure any kind of TV show (e.g., documentaries, reports, news, series, etc.). The approaches presented search for topic shifts and drifts, assuming that the distribution of vocabulary is a good indicator for them. The linear segmentation methods, local or global, face several limitations. Local methods can't cope with segments of variable length and need parameters to define the window width and cut-off values to decide where to place boundaries. While, global methods face the problem of over-segmentation that, up to now, can only be solved by providing prior information regrading the distribution of segment length or the expected number of segments. When it comes to hierarchical topic segmentation, the challenges that arise are related to knowing when to stop the segmentation, dealing with the lack of words at lower levels in the hierarchy, dealing with the propagation of errors and being able to propose segments of variable lengths. Additionally, several existing approaches need the segment duration and granularity level beforehand. In the following chapters we will address these challenges and propose new approaches to overcome them. We will start with linear topic segmentation and continue with the hierarchical one.

Chapter 3

Leveraging lexical cohesion and disruption for topic segmentation

Contents

3.1 Our approach in a nutshell 35
3.2 Combining lexical cohesion and disruption 36
3.3 Experiments 40
3.4 Discussion 45

3.1 Our approach in a nutshell

In Chapter 2 of this thesis, we presented the existing work done on topic segmentation, linear and hierarchical and discussed the limitations of the existing techniques. Our observations regarding linear topic segmentation emphasized the shortcomings of local and global methods, as both come with different advantages and disadvantages. These considerations naturally lead to the idea of methods combining lexical cohesion and disruption to make the best of both worlds. While the two criteria rely on the same underlying principle of lexical coherence [58] and might appear as redundant, the resulting algorithms are quite different in their philosophy. A first (and, to the best of our knowledge, unique) attempt at capturing a global view of the local dissimilarities is described in Malioutov and Barzilay [92]. The authors cast text segmentation in a graph-based framework, abstracting text into a weighted undirected graph. The nodes denote adjacent sentences and the edge weights define a measure of similarity between pairs of sentences. Higher similarity means higher lexical similarity between sentences. The edges between sentences exceeding a certain threshold distance are discarded. The threshold value was tuned on a held-out development set. The segmentation corresponds to a graph partitioning problem that aims to optimize the normalized-cut

criterion. However, this method assumes that the number of segments to find is known beforehand which makes it difficult for real-world usage.

In this chapter, we propose a segmentation criterion combining both cohesion and disruption along with the corresponding algorithm for topic segmentation. Such a criterion ensures a coherent use of vocabulary within each resulting segment, as well as a significant difference of vocabulary between neighbouring segments. Moreover, the combination of these two strategies enables regularizing the number of segments found without resorting to prior knowledge.

The starting point of our approach is the algorithm of Utiyama and Isahara [136], a versatile and performing topic segmentation algorithm cast in a statistical framework. The benefits of this algorithm were presented in Chapter 2 and we briefly remind them here: independency to any particular domain and ability to cope with thematic segments of highly varying lengths. These features are extremely interesting to obtain a generic solution for the problem of topic segmentation. Moreover, the algorithm has proven to be up to the state of the art in several studies, with no need for prior information about the number of segments (contrary to algorithms in [92, 37] that can attain a higher segmentation accuracy). It also provides an efficient graph-based implementation of which we take advantage.

To account both for cohesion and disruption, we extend the formalism of Isahara and Utiyama using a Markovian assumption between segments in place of the independence assumption of the original algorithm. Keeping unchanged their probabilistic measure of lexical cohesion, the Markovian assumption enables to introduce the disruption between two consecutive segments. We propose an extended graph-based decoding strategy, which is both optimal and efficient, exploiting the notion of generalized segment model or semi Markov models. Clearly, other formalisms than the graph-based one could have been considered. However, graph-based probabilistic topic segmentation has proven very accurate and versatile, relying on very minimal prior knowledge on the texts to segment. Good results at the state-of-the-art have also been reported in difficult conditions with this approach [98, 29, 60]. The seminal idea of this work was partially published in [126] in the French language and was extended afterwards in [125] with a more detailed description of the algorithm and additional contrastive experiments including more data sets. In particular, new experiments clearly demonstrate the benefit of the method in a realistic setting with statistically significant gains.

The organization of the rest of this chapter is as follows: Section 3.2 details the baseline method of Utiyama and Isahara before introducing our algorithm. Experimental protocol and results are given in Section 6.3. Section 3.4 summarizes the findings and concludes with a discussion.

3.2 Combining lexical cohesion and disruption

3.2.1 Probabilistic graph-based segmentation

The idea of the probabilistic graph-based segmentation algorithm is to find the segmentation into the most coherent segments constrained by a prior distribution on segments length. This problem is cast into finding the most probable segmentation of a sequence of t basic units (i.e., sentences or utterances composed of words) $W = u_1^t$ among all possible segmentations,

i.e.,

$$\hat{S} = \arg\max_{S} P[W|S]P[S] \ . \tag{3.1}$$

Assuming that segments are mutually independent and assuming that basic units within a segment are also independent, the probability of a text W for a segmentation $S = S_1^m$ is given by

$$P[W|S_1^m] = \prod_{i=1}^{m}\prod_{j=1}^{n_i} P[w_j^i|S_i] \ , \tag{3.2}$$

where n_i is the number of words in the segment S_i, w_j^i is the j^{th} word in S_i and m the number of segments. The probability $P[w_j^i|S_i]$ is given by a Laplace law where the parameters are estimated on S_i, i.e.,

$$P[w_j^i|S_i] = \frac{f_i(w_j^i)+1}{n_i+k} \ , \tag{3.3}$$

where $f_i(w_j^i)$ is the number of occurrences of w_j^i in S_i and k is the total number of distinct words in W, i.e., the size of the vocabulary $\mathcal{V}$. This probability favors segments that are homogeneous, increasing when words are repeated and decreasing consistently when they are different. The prior distribution on segment length is given by a simple model, $P[S_1^m] = n^{-m}$, where n is the total number of words, exhibiting a large value for a small number of segments and conversely.

The optimization of Eq. 3.1 can be efficiently implemented as the search for the best path in a weighted graph representing all possible segmentations, taking advantage of the left-right topology of the graph. Each node in the graph corresponds to a possible frontier placed between two utterances (i.e., we have a node between each pair of utterances), the arc between nodes i and j representing a segment containing utterances u_{i+1} to u_j. The corresponding arc weight is computed according to

$$v(i,j) = \sum_{k=i+1}^{j} ln(P[u_k|S_i]) - \alpha ln(n)$$

where the first term corresponds to the generalized probability of the words within segment S_i. ($P[u_k|S_i]$ is given as in Eq. 3.3. The factor α is introduced to control the trade-off between the segments length and the lexical cohesion.

3.2.2 Introduction of the lexical disruption

Eq. 3.2 derives from the assumption that each segment S_i is independent from the others, which makes it impossible to consider disruption between two consecutive segments. To do so, the weight of an arc corresponding to a segment S_i should take into account how different this segment is from S_{i-1}. This is typically handled using a Markovian assumption of order 1. Under this assumption, Eq. 3.2 is reformulated as

$$P[W|S_1^m] = P[W|S_1]\prod_{i=2}^{m} P[W|S_i, S_{i-1}] \ ,$$

where the notion of disruption can be embedded in the term $P[W|S_i, S_{i-1}]$ which explicitly mentions both segments. Formally, $P[W|S_i, S_{i-1}]$ is defined as a probability. However, arbitrary scores which do not correspond to probabilities can be used instead as the search for

the best path in the graph of possible segmentations makes no use of probability theory. In this study, we define the score of a segment S_i given S_{i-1} as

$$\ln P[W|S_i, S_{i-1}] = \ln P[W_i|S_i] - \lambda\Delta(W_i, W_{i-1}) \tag{3.6}$$

where W_i designates the set of utterances in S_i and the rightmost part reflects the disruption between the content of S_i and of S_{i-1}. Eq. 3.6 clearly combines the measure of lexical cohesion with a measure of the disruption between consecutive segments: $\Delta(W_i, W_{i-1}) > 0$ measures the coherence between S_i and S_{i-1}, the substraction thus accounting for disruption by penalizing consecutive coherent segments. The underlying assumption is that the bigger $\Delta(W_i, W_{i-1})$, the weaker the disruption between the two segments. Parameter λ controls the respective contributions of cohesion and disruption.

We initially adopted a probabilistic measure of disruption based on cross probabilities, i.e., $P[W_i|S_{i-1}]$ and $P[W_{i-1}|S_i]$, which proved to have limited impact on the segmentation. We therefore prefer to rely on a cosine similarity measure between the word vectors representing two adjacent segments, building upon a classical strategy for local methods such as TextTiling [66]. The cosine similarity measure is calculated between vectors representing the content of respectively S_i and S_{i-1}, denoted $\mathbf{v}_i$ and $\mathbf{v}_{i-1}$, where $\mathbf{v}_i$ is a vector containing the (tf-idf) weight of each term of $\mathcal{V}$ in S_i. $\Delta(W_i, W_{i-1})$ is calculated from the cosine similarity measure as

$$\Delta(W_i, W_{i-1}) = (1 - \cos(\mathbf{v}_{i-1}, \mathbf{v}_i))^{-1} \quad , \tag{3.7}$$

thus yielding a small penalty in Eq. 3.6 for highly disrupting boundaries, i.e., corresponding to low similarity measure.

Given the quantities defined above, the algorithm boils down to finding the best scoring segmentation as given by

$$\hat{S} \quad = \quad \arg\max_S \sum_{i=1}^{m} ln(P[W_i|S_i]) - \lambda \sum_{i=2}^{m} \Delta(W_i, W_{i-1}) - \alpha m ln(n) \quad .$$

3.2.3 Segmentation algorithm

Translating Eq. 3.8 into an efficient algorithm is not straightforward since all possible combinations of adjacent segments need be considered. To do so in a graph-based approach, one needs to keep separated the paths of different lengths ending in a given node. In other words, only paths of the same length ending at a given point, with different predecessors, should be recombined so that disruption can be considered properly in subsequent steps of the algorithm. Note that, in standard decoding as in Utiyama and Isahara's algorithm, only one of such paths, the best scoring one, would be retained. We employ a strategy inspired from the decoding strategy of segment models or semi-hidden Markov model with explicit duration model [104, 34].

Search is performed through a lattice $L = \{V, E\}$, with V the set of nodes representing potential boundaries and E the set of edges representing segments, i.e., a set of consecutive utterances. The set V is defined as

$$V = \{n_{ij} | 0 \leq i, j \leq N\} \quad ,$$

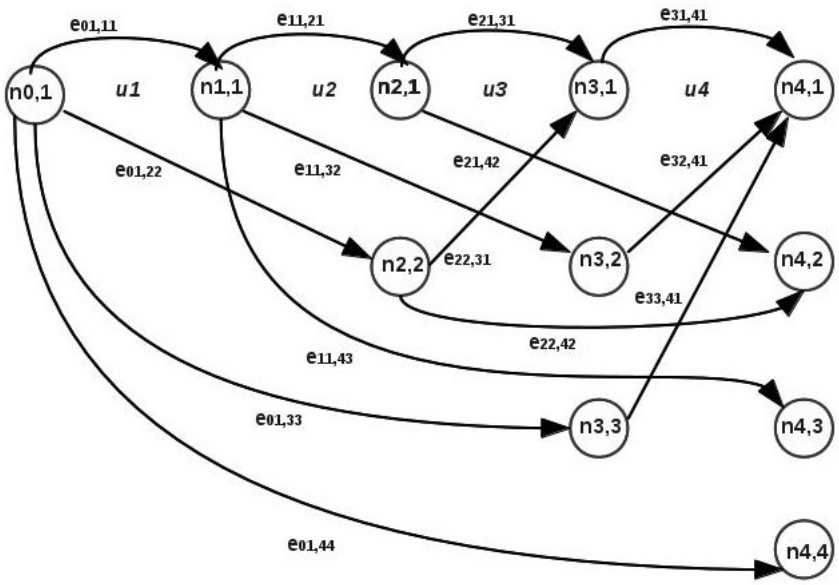


Figure 3.1: An example of a lattice L.

where n_{ij} represents a boundary after utterance u_i reached by a segment of length j utterances and $N = t + 1$. In the lattice example of Fig. 3.1, it is trivial to see that for a given node, all incoming edges cover the same segment. For example, the node n_{42} is positioned after $u4$ and all incoming segments contain the two utterances $u3$ and $u4$. Edges are defined as

$$E = \{e_{ip,jl} | 0 \leq i, p, j, l \leq N; \tag{3.9}$$

$$i < j; i = j - l; L_{\min} \leq l \leq L_{\max}\} \ , \tag{3.10}$$

where $e_{ip,jl}$ connects n_{ip} and n_{jl} with the constraint that $l = j - i$ and $L_{\min} \leq l \leq L_{\max}$. Thus, an edge $e_{ip,jl}$ represents a segment of length l containing utterances from u_{i+1} to u_j, denoted $S_{i \to j}$. In Fig. 3.1, $e_{01,33}$ represents a segment of length 3 from n_{01} to n_{33}, covering utterances u_1 to u_3. To avoid explosion of the lattice, a maximum segment length $L_{\max}$ is defined. Symmetrically, a minimum segment size can be used.

The property of this lattice where, by construction, all edges out of a node have the same segment as a predecessor, makes it possible to weight each edge in the lattice according to Eq. 3.6. Consider a node n_{ij} for which all incoming edges encompass utterances u_{i-j} to u_i. For each edge out of n_{ij}, whatever the target node (i.e., the edge length), one can therefore easily determine the lexical cohesion as defined by the generalized probability of Eq. 3.3 and the disruption with respect to the previous segment as defined by Eq. 3.7.

Given the weighted decoding graph, the solution to Eq. 3.8 is obtained by finding out the best path in the decoding lattice, which can be done straightforwardly by scanning nodes in topological order. The decoding algorithm is summarized in Algorithm 2 with an efficient implementation in $o(NL^2_{\max})$ that does not require explicit construction of the lattice.

Algorithm 1 Maximum probability segmentation

Step 0. Initialization
$q[0][j] = 0 \quad \forall j \in [L_{\min}, L_{\max}]$
$q[i][j] = -\infty \quad \forall i \in [1, N], j \in [L_{\min}, L_{\max}]$

Step 1. Assign best score to each node
for $i = 0 \rightarrow t$ **do**
  for $j = L_{\min} \rightarrow L_{\max}$ **do**
    for $k = L_{\min} \rightarrow L_{\max}$ **do**
      /*extend path ending after u_i with a segment of length j with an arc of length k*/

$$q[i+k][k] = \max \begin{cases} q[i+k][k], \\ q[i][j] + \\ \text{Cohesion}(u_{i+1} \rightarrow u_{i+k}) - \\ \lambda \Delta(u_{i-j} \rightarrow u_i; u_{i+1} \rightarrow u_{i+k}) \end{cases}$$

    end for
  end for
end for

Step 2. Backtrack from n_{Nj} with best score $q[N][j]$ over all j

$z =$	3–11	3–5	6–8	9–11
# samples	400	100	100	100

Table 3.1: Number of documents in Choi's corpus [25].

3.3 Experiments

The technique we propose can be applied to any kind of text-like data, not only transcripts, so we provide results also on the written texts usually employed in the literature in the context of topic segmentation. The experiments are performed on three distinct corpora that exhibit different characteristics, two containing textual data and one spoken data. We first describe the corpora before presenting and discussing results on each.

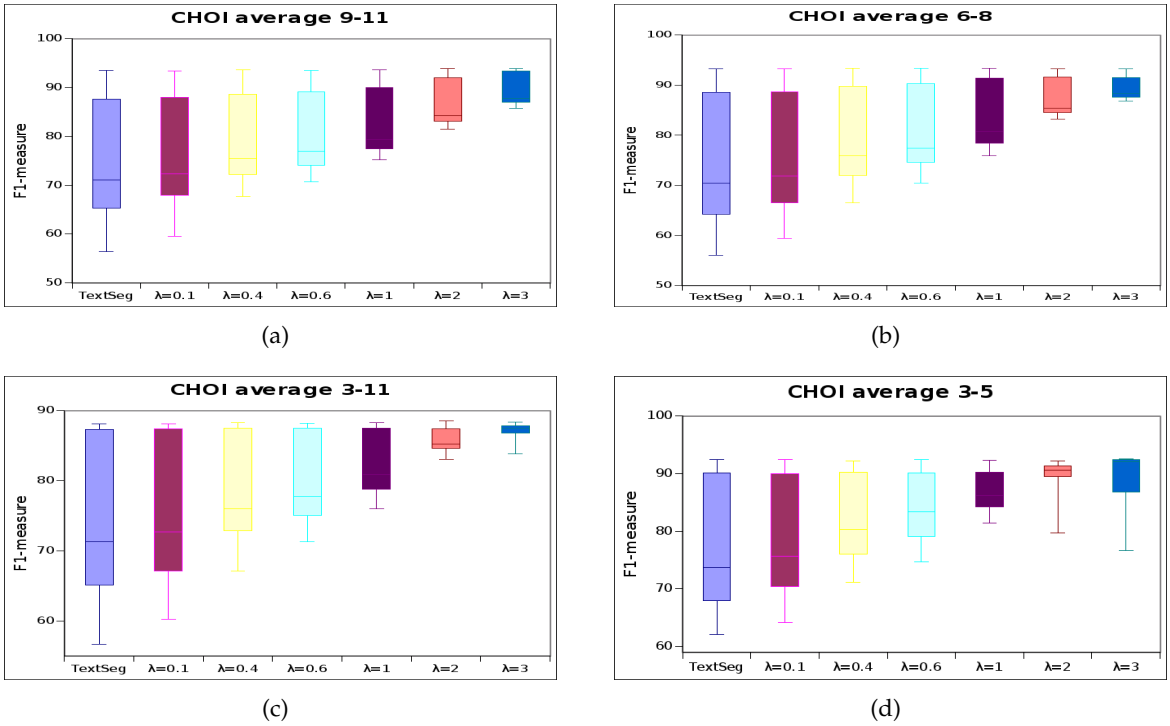
3.3.1 Corpora

The artificial data set of Choi [25] is widely used in the literature and enables comparison of a new segmentation method with existing ones. Choi's data set consist of 700 documents, each created by concatenating the first z sentences of 10 articles randomly chosen from the Brown corpus, assuming each article is on a different topic. Table 3.1 provides the corpus statistics, where z=3–11 means z is randomly chosen in the range $[3, 11]$. Hence, Choi's corpus is adapted to test the ability of our model to deal with variable segments length, z=3–11 being the most difficult condition. Moreover, Choi's corpus provides a direct comparison with results reported in the literature.

One of the main criticism of Choi's data set is the presence of abrupt topic changes due to the artificial construction of the corpus. We therefore report results on a textual corpus with more natural topic changes, also used in [37]. The data set consists of 277 chapters selected from [140], a medical textbook, where each chapter—considered here as a document—was divided by its author into thematically coherent sections. The data set has a total of 1,136 segments with an average of 5 segments per document and an average of 28 sentences per segment. This data set is used to study the impact of smooth, natural, topic changes. Finally, results are reported on a corpus of automatic transcripts of TV news spoken data. This data set was previously presented in Chapter 2.

All data were preprocessed in the same way: Words were tagged and lemmatized with TreeTagger[1] and only the nouns, non modal verbs and adjectives were retained for segmentation. Inverse document frequencies used to measure similarity in Eq. 3.7 are obtained on a per document basis, referring to the number of sentences in textual data and of utterances in spoken data.

3.3.2 Results



(a) (b) (c) (d)

Figure 3.2: F1-measure variation obtained on Choi's corpus. In each graphic, the leftmost boxplot TextSeg corresponds to results obtained by using the sole lexical cohesion (baseline), while the λ value is the importance given to the lexical disruption in our approach. Results are provided for the same range of variation of factor α, allowing a tolerance of 1 sentence between the hypothesized and reference frontiers.

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

z	τ	F1 gain	Confidence interval 95 % TextSeg	Combined
3-5	0	-0.2	[66.6,74.26]	[75.23,78.08]
3-5	1	0.7	[72.25,83.4]	[87.88,92.13]
3-11	1	0.23	[68.5,79.3]	[86.6,87.43]
6-8	1	0.4	[68.48,80.99]	[76.9,85.17]
9-11	0	1.6	[64.35,75.16]	[81.31,84.86]
9-11	1	1.4	[68.39,80.39]	[84.37,88.9]

Table 3.2: Gain in F1-measure for Choi's corpus when using lexical cohesion and disruption, and the corresponding 95 % confidence intervals for the F1-measure. Results are reported for different tolerance τ. TextSeg denotes the baseline and Combined the proposed model.

Performance is measured by comparison of hypothesized frontiers with reference ones. The alignment between the reference and hypothesized frontiers assumes a tolerance of 1 sentence on texts and of 10 seconds on transcripts, which corresponds to standard values in the literature. Results are reported using recall, precision and F1-measure as defined in Chapter 2. In Eq. 3.8, the parameter α, which controls the contribution of the prior model with respect to the lexical cohesion and disruption, allows for different trade-offs between precision and recall. For any given value of λ, α is thus varied, providing the range of recall/precision values attainable. Results are compared to a baseline system corresponding to the application of the original algorithm of Utiyama and Isahara (i.e., setting $\lambda = 0$). This baseline has been shown to be a high-performance algorithm, in particular with respect to local methods that exploit lexical disruption. Differences in F1-measure between this baseline and our system presented below are all statistically significant at the level of $p < 0.01$ (paired t-test).

Choi's corpus. Figure 3.2 reports results obtained on Choi's data set, each graphic corresponding to a specific variation in the size of the thematic segments forming the documents (e.g., 9 to 11 sentences for the top left graphic). Results are provided for different values of λ in terms of F1-measure boxplots, i.e., variations of the F1-measure when α varies (same range of variation for α considered for each plot), where the leftmost boxplot, denoted by *TextSeg*, corresponds to the baseline. Box and whisker plots graphically depicts the distribution of the F1-measures that can be attained by varying α, plotting the median value, the first and third quartile and the extrema.

Figure 3.2 shows that, whatever the segments length, results globally improve as the weight given to the disruption (λ variable) increases. Moreover, the variation in F1-measure diminishes when disruption is considered, thus indicating the influence of the prior model diminishes. When the segments size decreases (see Figs. 3.2(b), 3.2(c), 3.2(d)), the difference in the maximum F1-measure between our results and that of the baseline lowers, however still in favor of our model. This can be explained by the fact that our approach is based on the distribution of words, thus more words better help discriminate between potential thematic frontiers. Finally, using too large values for λ can lead to under-segmentation, as can be seen in Fig. 3.2(d) where, for $\lambda = 3$, the variation of F1-measure increases and the distribution becomes negatively skewed (i.e., the median is closer to the third quartile than to the first).

These results are confirmed by Table 3.2 where we report the gain in F1-measure (i.e., the difference between the highest F1-measure obtained when combining lexical cohesion and disruption and the highest value for the baseline) for each of the four sets of documents in Choi's corpus, together with the 95 % confidence intervals. The effect of using the disruption is higher when segment size is longer, whether evaluation allows or not for a tolerance τ between the hypothesized frontiers and the reference ones. A qualitative analysis of the segmentations obtained confirmed that employing disruption helps eliminate wrong hypothesis and shift hypothesized frontiers closer to the reference ones (explaining the higher gain at tolerance 0 for the 9-11 data set). When smaller segments—thus few word repetitions—and no tolerance are considered (e.g., 3–5), disruption cannot improve segmentation. Our model is globally stable with respect to segment length, with relatively similar gain for 3–11 and 6–8 data sets in which the average number of words (distinct or repeated) is close.

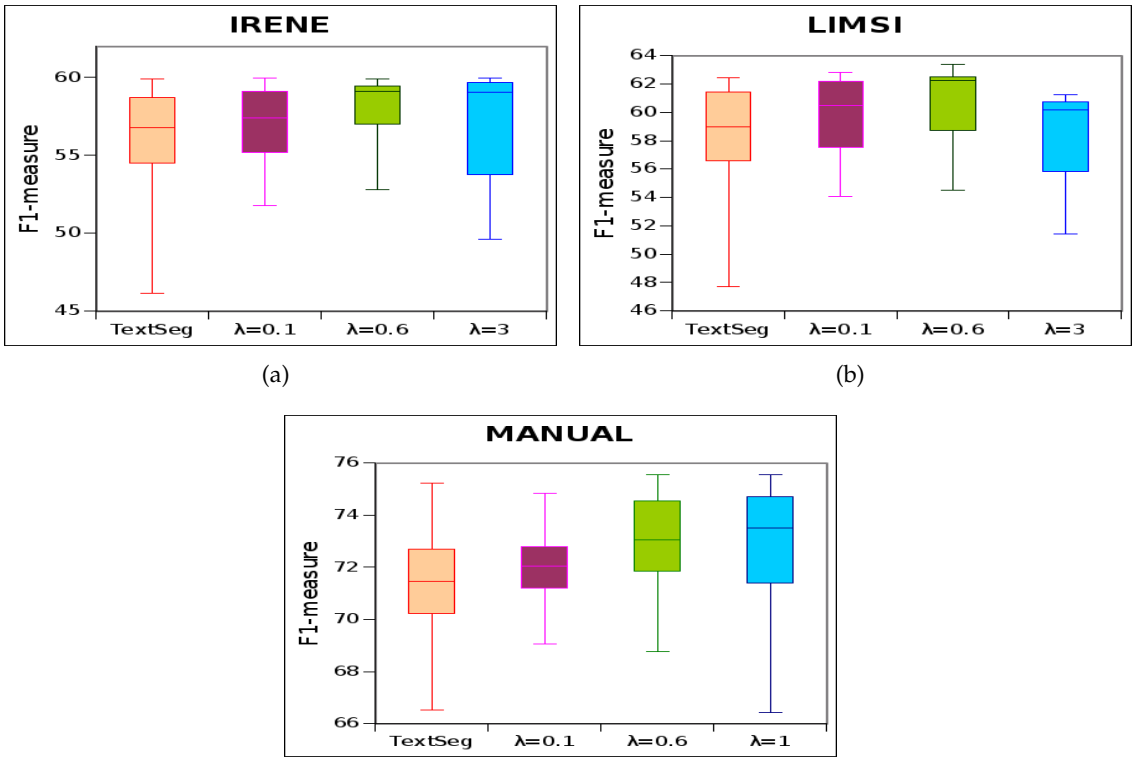


Figure 3.3: Boxplots showing F1-measure variation on transcripts obtained using IRENE and LIMSI automatic speech recognition system and on the manual transcripts.

Results discussed up to now are optimistic as they correspond to the best F1 value attainable computed a posteriori. Stability of the results was confirmed using cross-validation with 5 folds (10 folds for z=3–11): Parameters λ and α maximizing the F1-measure are determined on all but one fold, this last fold being used for evaluation. Results, averaged over all folds, are reported in Table 3.3 for the baseline and the method combining cohesion and disruption.

$z =$	3–5	3–11	6–8	9–11
TextSeg	91.9	87.0	93.1	92.8
Combined	92.9	87.5	93.5	94.0

Table 3.3: F1 results using cross-validation on Choi's data set.

Medical textbook corpus. The medical textbook corpus was previously used for topic segmentation by Eisenstein and Barzilay [37] with their algorithm BayesSeg[2]. We thus compare our results with those obtained by BayesSeg and by the baseline. When considering the best F1-measure (i.e., the best F1-measure which can be achieved by varying α and λ), we achieved an improvement of 2.2 with respect to BayesSeg when no tolerance is allowed, and of 0.5 when the tolerance is of 1 sentence. The corresponding figures with respect to the baseline are 0.6 and 0.4. When considering the F1-measure value for which the number of hypothesized frontiers is the closest to the number of reference boundaries, improvement is of resp. 1.5 and 0.5 with respect to BayesSeg, -0.1 and 0.4 with respect to the baseline. These results show that our model combining lexical cohesion and disruption is also able to deal with topic segmentation of corpora from a homogeneous domain, with smooth topic changes and segments of regular size.

One can argue that the higher number of free parameters in our method explains most of the gain with respect to BayesSeg. While BayesSeg has only one free parameter (as opposed to two in our case), the number of segments is assumed to be provided as prior knowledge. This assumption can be seen as an additional free parameter, i.e., the number of segments, and is a much stronger constraint than the ones we are using. Moreover, cross-validation experiments on the Choi data set show that improvement is not due to over-fitting of the development data thanks to an additional parameter. Gains on development set with parameters tuned on the development set itself and with parameters tuned on a held-out set in cross-validation experiments are in the same range.

TV news transcripts corpus Figure 3.3 provides results, in terms of F1-measure variation, for TV news transcripts obtained with the two ASR systems and on the manual transcripts. On this highly challenging corpus, with short segments, wrongly transcribed spoken words, and thus few word repetitions, the capabilities of our model to overcome the baseline system are reduced. Yet, an improvement of the quality of the segmentation of these noisy data is still observed, and general conclusions are quite similar—though a bit weaker—to those already made for Choi's corpus. Results are confirmed in Table 3.4 which presents the gain in F1-measure of our model together with the 95 % confidence interval, where F1-measure values correspond to that of segmentations obtained with a number of hypothesized frontiers as close as possible to the reference. The two first lines show that the gain is smaller for IRENE transcripts which have fewer correct words available to discriminate between segments belonging to different topics. The impact of transcription errors is illustrated in the last three lines, when segmenting six TV news for which reference transcripts are available (line 3), where the higher the WER, the smaller the F1-measure gain.

[2]The code and the data set are available at http://groups.csail.mit.edu/rbg/code/bayesseg/

Corpus	F1 gain	Confidence interval 95 % TextSeg	Combined
IRENE	0.3	[54.4,57.6]	[56.92,59]
LIMSI	0.86	[56.7,60.2]	[59.44,61.95]
MANUAL (6)	0.77	[70.39,72.29]	[71.7,73.29]
IRENE (6)	0.2	[56.81,60.94]	[59.51,63.43]
LIMSI (6)	0.5	[64.27,68.64]	[67.7,71.56]

Table 3.4: Gain in F1-measure for TV news corpus automatic and manual transcripts when using lexical cohesion and disruption, and the corresponding 95 % confidence intervals. Last three rows report results on only 6 shows for which manual reference transcripts are available.

3.4 Discussion

In this chapter we have presented a new approach for linear topic segmentation that combines lexical cohesion and disruption. Experimental results on various data sets with various characteristics have demonstrated the impact of taking into account disruption in addition to lexical cohesion. We observed gains both on data sets with segments of regular length and on data sets exhibiting segments of highly varying length within a document. Unsurprisingly, bigger gains were observed on documents containing relatively long segments. However, the segmentation algorithm has proven to be robust on automatic transcripts with short segments and limited vocabulary reoccurrences. Finally, we tested both abrupt topic changes and smooth ones with good results on both.

With this approach we overcame several challenges characteristic to local and global methods. The approach can cope with segments of variable length, a challenge for local methods. Accounting for the lexical disruption leads to a regularization of the segments found and reduces the over-segmentation characteristic of global methods. We envision several ways that can further improve the new approach we proposed. Additional information could be added to overcome the presence of errors in the transcripts and to overcome the reduced number of word repetitions. This information should maintain the generic character of the approach and therefore be independent of the type of data. Confidence measures provided by the ASR system for spoken words and semantic relations have already been proven to improve topic segmentation for automatic transcripts of TV shows [60]. Thus we expect such additional information to bring gains to our new algorithm also. Other speech related features such as speech turn, speaker id and jingles have been studied for correlation with topic segments. In [74], it has been observed that the change between female and male speaker is correlated with topic changes in radio-phonic journals, while pauses between breath-groups are not. However, the correlation of speaker gender-based turn information with topic changes for any kind of TV show needs to be investigated before considering it in the context of generic approaches.

Another idea is to change the representation of the data from the bag-of-words model to a more complex model. The bag-of-words model looses semantic information, it assumes term independence and homogeneity, i.e., the words in a document are assumed to be distributed homogeneously [121]. The advantage of this representation is that it simplifies further pro-

cessing. However, we believe that using a model that better captures the distribution of words by taking into account the positional information could improve the segmentation. The assumption that this positional information does not provide any extra leverage to the performance of NLP and IR approaches has been proven to be wrong in certain applications [44]. With such a model in mind, we propose to move from the linear landscape to the hierarchical one.

Chapter 4

Investigating hierarchical topic segmentation

Contents

4.1 Overview 47
4.2 Is lexical cohesion enough for hierarchical topic segmentation? 49
4.3 When to stop applying a hierarchical topic segmentation algorithm? . . . 56
4.4 Hierarchical structure of topically focused fragments 58
4.5 Discussion 65

4.1 Overview

Most approaches for automatic topic segmentation are designed for linear segmentation even though many documents exhibit a hierarchical structure. Such a structure is far more complex to apprehend than a linear one. A linear segmentation will structure the data into successive topics and provide a general overview of the data. In the case of a hierarchical segmentation, the internal structure of the data is put forward, shaping the information contained at different granularity levels. The hierarchy offers both a global view over the subjects approached and the possibility to zoom over their different aspects, various points of view, etc. Identifying this kind of structure is an essential step for various natural language processing tasks, such as question answering, information retrieval, information visualization, summarization, etc.

The focus of this chapter is on hierarchical topic segmentation. We are interested in solutions relying on lexical cohesion, since they can be applied on any kind of TV show. We propose to investigate the limitations of existing approaches and search for ways to alleviate

them. One of the challenges that current approaches have to deal with is the fact that segments down in the hierarchy can be short and have high cohesion between them, which can be difficult for lexical cohesion. Thus, we start our investigation by addressing the following question: *Is lexical cohesion enough for hierarchical topic segmentation?* To answer this question, we propose to evaluate if the current lexical cohesion measures are sufficient or not to grasp a "true" (i.e., reference) segmentation. We will show that they are not sufficient and therefore propose to test another way to measure lexical cohesion, through burstiness analysis.

The term "burstiness" is used to describe the phenomenon that words are likely to occur again in a text after they have occurred once, as opposed to being emitted independently [27, 78]. Such an analysis considers the positions of a word in a text and the intervals between its reoccurrences. This means it considers whether a word occurred in the beginning, middle or end of a document and if it occurs frequently in close succession or rather uniformly throughout the document [121]. Bursty words tend to be characterized by long inter-arrival times followed by short inter-arrival times. While, non-bursty words exhibit inter-arrival times with smaller variance [86]. In [4], the authors emphasize that bursty words characterize better discourse topics. Also, in [82], the author upholds this idea and hints that a topic should be characterized by a "burst of activity". To capture the bursts we rely on Kleinberg's algorithm [82], which identifies for each word whether it has bursts, yielding a nested representation of the set of bursts (i.e., a hierarchy of burst intervals).

We therefore perform a similar test to the one done with current measures for lexical cohesion and analyze the distribution of words in the reference segmentation from the perspective of bursts. The analysis on words bursts reveals interesting characteristics of the data, like the fact that some fragments of the data bear important ideas while others are simple fillers, i.e., they do no bring additional important information. This means that the important ideas in the data can still be extracted, however burst analysis cannot fully grasp the reference segmentation. Thus, just including it in traditional algorithms is not enough to solve the problem of hierarchical topic segmentation. Still, burst analysis may help to know when the traditional algorithms should stop being applied. This idea leads to the second part of our investigation which addresses another challenge of hierarchical topic segmentation: *the absence of a criterion to know when to apply or stop applying a hierarchical topic segmentation algorithm*. We show that burst analysis is a good tool to bring answers to this problem.

The end of our investigation consists in proposing a novel organization of the topical structure of textual content, that overcomes several challenges that hierarchical topic segmentation strategies have to deal with. Rather than searching for topic shifts to yield classical dense segmentation, we extract topically focused fragments organized in a hierarchical manner, leveraging the burstiness phenomenon. A fragment is seen as a part of text that can be detached from the whole text, while a segment is a part into which the text can be divided. We rely on the fact that the presence of lexical bursts indicates a strong topical focus. The result is an algorithm that extracts a hierarchy of topically focused fragments. Comparison to a reference dense segmentation on varied datasets indicates that we can achieve a better topic focus while retrieving all of the important aspects of a text. We propose also a task driven evaluation by analyzing the impact of the new structure in the context of automatic summarization.

The organization of the rest of this chapter is as follows: Section 4.2 presents the first part of our investigation that aims to study if lexical cohesion is enough for hierarchical topic segmentation. Section 4.3 is dedicated to the second part of the investigation and introduces

the criterion based on burst analysis to decide when to continue or to stop the segmentation. Section 4.4 presents our new take on the problem of hierarchical topic segmentation and describes the algorithm to build the hierarchy of topically focused fragments together with the evaluation. Section 4.5 summarizes our findings.

4.2 Is lexical cohesion enough for hierarchical topic segmentation?

4.2.1 Classical measures for lexical cohesion

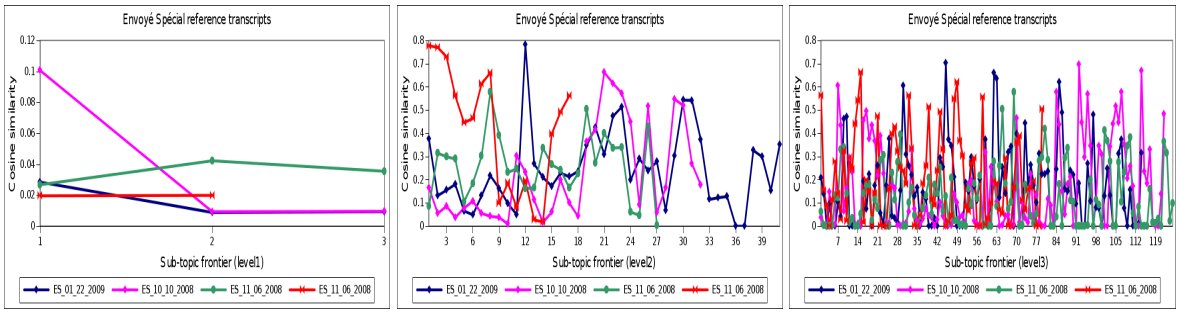
We propose to study the behavior of two commonly-used measures of lexical cohesion (cos-similarity measure and probabilistic lexical cohesion measure) to see if they can explain the reference hierarchical segmentation. The study is done on three different datasets.

4.2.1.1 Corpora

Three datasets, previously used in the context of hierarchical segmentation, are used in our experiment: a medical textbook [36]; Wikipedia articles [22]; reference and automatic French TV show transcripts [59]. All the datasets are preprocessed in the same way: Words are tagged and lemmatized with TreeTagger and only the nouns, non modal verbs and adjectives are retained. The Wikipedia corpus contains 66 articles with a hierarchy of up to 4 levels. The reference segmentation is obtained from the structures given by the author of each article. The medical textbook and the transcripts data-sets have already been presented (page 40 and page 32). Throughout this chapter the highest level in the hierarchy will be denoted level 0 and represents an entire Wikipedia article/part of the medical textbook/transcript of a TV show and the lowest level will correspond to level 4/2/3 respectively.

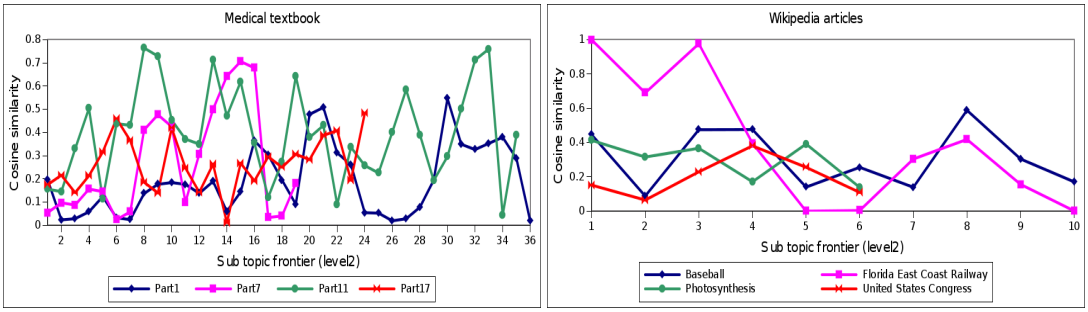
4.2.1.2 Cos-similarity measure

The first measure considered is the similarity-based approach for which a cosine measure is computed between vectors representing the content of adjacent segments. The cosine measure was previously introduced in Chapter 2, page 22. Figure 4.1 reports the evolution of the cosine similarity measure between consecutive sub-topics, over all segments at each level in the hierarchy, for the reference transcripts of the TV show data. The plots correspond from left to right to the values obtained at the first level, the second level and finally the third level. As it can be observed, there is a high variability in the similarity between consecutive segments within a document for the second and third level. Thus, it is difficult to define a threshold for segmentation purposes, which impacts the capacity of the segmentation algorithms to find topic frontiers at these levels. Figure 4.2 reports the values obtained with the cosine measure on the medical textbook (Fig. 4.2(a)) and Wikipedia (Fig. 4.2(b)) data sets. For these two data sets the values are given on 4 samples for better visibility. We report here only the values obtained for the second level in the hierarchy, for brevity, since similar trends were observed at the other levels in the hierarchy.



(a) (b) (c)

Figure 4.1: Cosine similarity measure between consecutive sub-topics, at all levels in the hierarchy, are given for the TV show reference transcripts.



(a) (b)

Figure 4.2: Cosine similarity measure between consecutive sub-topics, at the second level, for the medical textbook and Wikipedia articles data sets. Only a fraction of the results are presented for legibility reasons.

4.2.1.3 Probabilistic measure

The second measure considered is a probabilistic one where lexical cohesion for a segment S_i is computed using a Laplace law as in [136]. This measure was introduced in Chapter 3, page 37, Eq. 3.3 as $P[w_j^i|S_i]$. For simplicity we will denote it here as $C(S_i)$. Note that the two measures are complementary: The cos-similarity measure considers adjacent segments to identify topic shifts, while the probabilistic one intrinsically measures the cohesion of a segment.

Figure 4.3 reports the evolution of the probabilistic lexical cohesion measure over all segments of the second level in the reference topic hierarchy as well as global statistics for $C(S_i)$. Each row corresponds to a different dataset: First, the TV show transcripts (Fig. 4.3(a), 4.3(b)), second, the medical textbook (Fig. 4.3(c), 4.3(d)) and third the Wikipedia articles (Fig. 4.3(e), 4.3(f)). Figures on the first column show the cohesion values obtained with the probabilistic measure for each sub-topic in the reference segmentation. The figures on the second column show general statistics (average, min and max values) for the same measure on the entire data sets. Once again, for the Wikipedia and medical textbook corpora the general statistics values reported only on 4 samples for a better visibility. As it can be

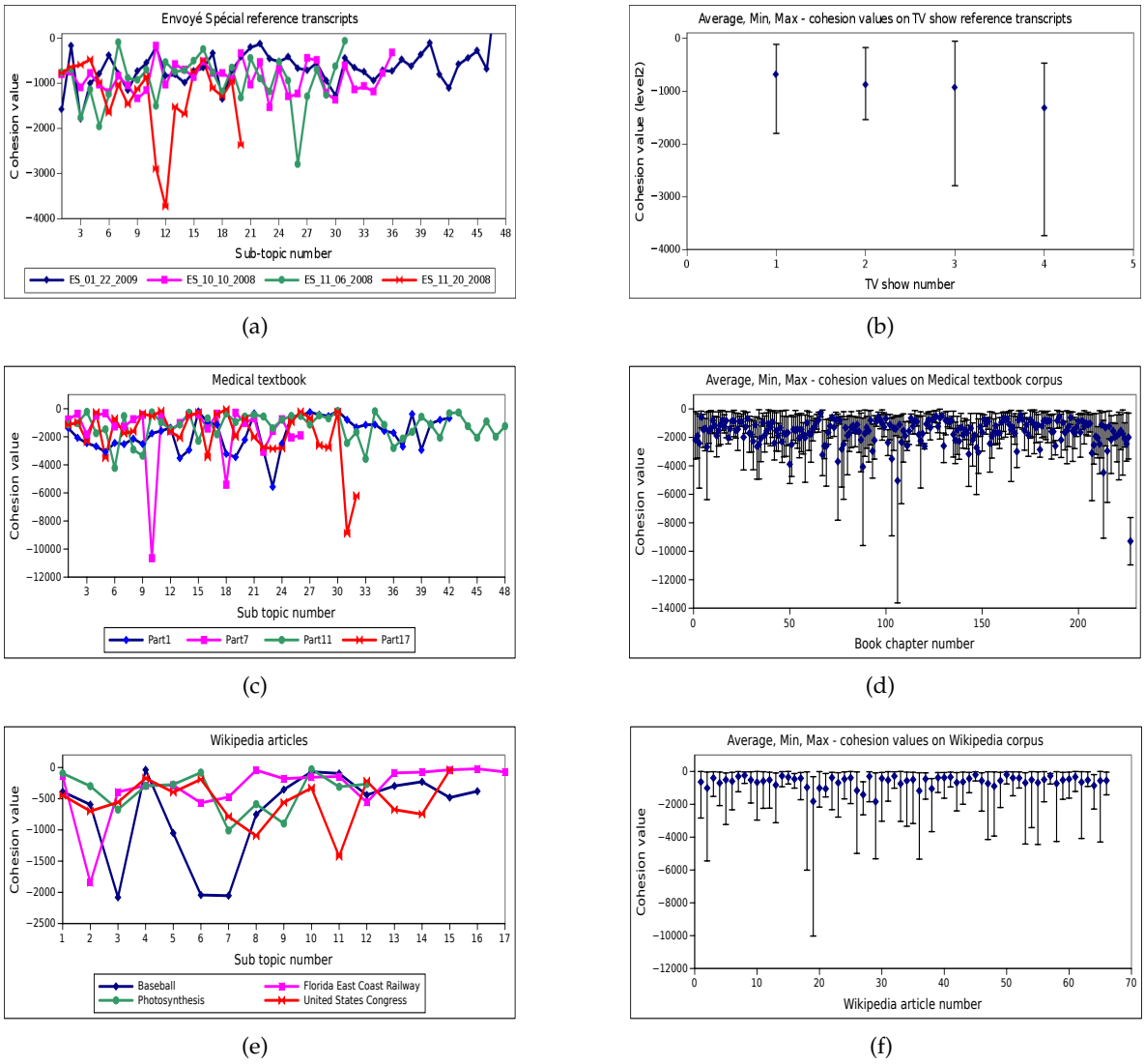


Figure 4.3: Probabilistic lexical cohesion measure for each dataset. Each row corresponds to a dataset, from top to bottom: TV shows, medical textbook, Wikipedia articles. Columns correspond to, from left to right: $C(S_i)$, distribution of $C(S_i)$ per document. Only a fraction of the results are presented for the textbook or Wikipedia for legibility reasons.

observed, there is a high variability in the cohesion values across sub-topics segments, as well as across documents (Fig. 4.3(b),4.3(d),4.3(f)). The values show that not all segments in the reference segmentation are characterized by a high cohesion value. So, having as objective to retrieve segments that maximize the lexical coherence will not necessarily explain a reference segmentation.

To sum up, the currently employed lexical cohesion measures for topic segmentation are not sufficient to grasp a reference segmentation. There exist several approaches in the literature, where the authors try to improve the lexical cohesion measures in the context of hierarchical segmentation by looking at the distributions of words. They try to adapt the

measures of lexical cohesion at lower levels in the hierarchy. The basic idea is that words that have contributed at the segmentation of a superior level should have less impact for the segmentation at lower levels in the hierarchy. In [59], an exploratory study for hierarchical topic segmentation is proposed, that aims to include this idea of adapting the lexical cohesion measure to the segmentation level. The authors recursively apply the TextSeg linear segmentation algorithm [136] to obtain a hierarchy of two levels. They propose first to modify the probabilistic measure for lexical cohesion by giving more weight to a word appearing only in a part of the text, compared to a word appearing everywhere in the text. Then, they propose to compare the probability distribution of a word appearing in part of the text with the distribution of the word appearing in a larger part of the text. However, these approaches cannot capture the possibility that a word that is highly characteristic of a specific part of a document can appear in other parts of the document as well. To overcome this, the authors propose to use lexical chains to account for the position of the words. To construct the chains they use semantic relations and have to define a parameter to determine the maximum length of a chain. Still, the lexical cohesion measures are based on the bag-of-words assumption, i.e., every word can occur at any position in a text with an equal probability. Several studies have pointed out that the bag-of-word model is a poor descriptor of word occurrences [121, 91, 86].

Words have different behaviour in language even at the word frequency level. Existing work for statistical laws in language have proposed burst detection models that analyze the distributional pattern of words [121, 91], to overcome the deficiencies of the multinomial model. The quest for these models has been driven by various applications like: keyword extraction [100] or style investigation [122]. In [36], the author incorporates, in the context of hierarchical topic segmentation, the Dirichlet compound multinomial model as proposed in [91] for modeling word burstiness. They claim this model is a better alternative to the multinomial model, which is considered appropriate only for common words. The explanation is that common words are more likely to satisfy the independence assumption since many of them are non-content function words. With the Dirichlet compound multinomial model, the burstiness of information-carrying words is captured. However the approach proposed in [36], relying on this model, requires priors on the number of segments at each level and the number of levels in the hierarchy, which is not a realistic scenario.

In addition to the previous models, other models have been proposed in the literature, that capture the burstiness phenomenon in words by modelling the gaps between successive occurrences of a word and the positions of its occurrence [121, 82]. Such a model is dedicated to a single word at a time. To our knowledge, such a model has not been used in the context of hierarchical topic segmentation before. In order to find if word bursts is a better alternative to justify the reference segmentation, we propose to study the presence of bursts of activity in the reference segmentation. The intuition here is that word bursts are good indicators both of lexical cohesion and disruption, while taking into account more information than previous models by considering the positional information of words. A burst word can be associated with a burst interval. This interval corresponds to a period where the word occurs with increased frequency with respect to normal behaviour. Thus a burst interval signals both the existence of lexical disruption and of fragments of text that are cohesive: A fragment with one or more words bursts has a more consistent use of vocabulary, with concepts repeated locally in the fragment, apart from the rest of the text; also a fragment with bursty words can be differentiated from other fragments in the text since the burst of a word signals a high

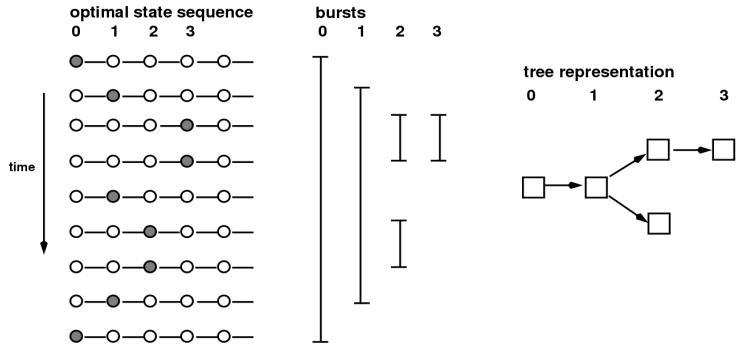


Figure 4.4: Example (taken from http://vw.indiana.edu/sackler03/ppts/Kleinberg.pdf) of an optimal state sequence with the corresponding bursts and tree representation.

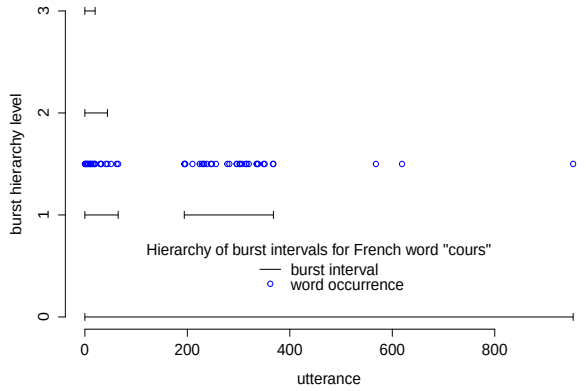


Figure 4.5: Sample output of Kleinberg's algorithm: The y-axis depicts the burstiness level while utterance number are on the x-axis; Circles indicate occurrences of the word considered. This example shows two bursts of level 1, the first one coming along with a burst of level 2 for a fraction of its time and another smaller burst of level 3.

frequency of that word in a restricted interval and therefore increases the disruption with adjacent fragments.

4.2.2 Lexical cohesion through burst analysis

4.2.2.1 Kleinberg's algorithm

At the core of the analysis of the burstiness phenomenon, we rely on Kleinberg's algorithm [82] (c.f. Appendix A) to identify word bursts, together with the intervals where they occur[1]. The algorithm relies on an infinite-state automaton where the states $i \in \mathbb{N}^+$ correspond to the frequency at which an individual word repeats. Arbitrarily, state 0 accounts for normal behaviour while increasing values of i correspond to increasing levels of burstiness. State transitions thus correspond to points in time when there is a significant change in the occurrence frequency of a word. The algorithm outputs a hierarchy of burst intervals for

[1]We use Jeff Binder's open-source implementation, available at http://cran.r-project.org/web/packages/bursts.

each word, taking one word at a time, by searching for the state sequence that minimizes a cost function. The interval of a burst at level j in the hierarchy of bursts is the maximal interval during which the optimal state sequence is in state j or higher, i.e., $k > j$, thus forming a hierarchical organization of burst intervals. A representation of an optimal state sequence and the corresponding bursts and their tree representation is provided in Figure 4.4. In other words, a word considered bursty on a time interval $[a, b]$ with a burstiness level of i is simultaneously considered as bursty at a level $i - 1$ on an interval $[c, d]$, with $[a, b] \subset [c, d]$. This hierarchy is illustrated in Figure 4.5 for one word: The word occurs with a burstiness level of 1 on the first utterances (i.e., sentences or sequences of words separated by breath intakes for automatic transcripts), with a significant amount of occurrences at the very beginning yielding two short intervals at level 2 and 3, both included in the interval at level 1. Long bursts intensifying into briefer ones can be seen as imposing a fine-grain organization within the text according to a natural tree structure.

4.2.2.2 A case analysis of bursts

We conducted a case-study to address the following question: *Can burst analysis explain the reference segmentation?* First, for each segment at each level of the reference topic segmentation, a hierarchy of burst intervals as the one illustrated in Figure 4.5 is computed for each word. Then, given the set of burst intervals, we count for each utterance the number of words within the utterance which appear as bursty at that position. We expect that local minima in the plot, i.e., utterances that contain few bursty words, are indicators of topic shifts. Figure 4.6 presents the counts for bursts computed at two levels (level 0 and level 1) in the reference hierarchical topic segmentation for a sample from the TV show transcripts (4.6(a) and 4.6(b)), for a Wikipedia article (4.6(c) and 4.6(d)), and for a book chapter(4.6(e) and 4.6(f)). The reference frontiers are marked with vertical lines.

In Figure 4.6(a), the counts concern the entire TV show transcript (level 0). Clearly, local minima in the plot can be associated with the reference frontiers: The number of bursts shared between the utterances at these points are considerably fewer than at any other point. Thus, at this level, the intuition behind lexical cohesion that a significant change in vocabulary signals a topic change still holds and the topical segments can be easily identified relying on bursts information. The same analysis for level 1 shows that local minima are neither easy to identify in this case, nor do they correspond with reference frontiers (see, Figure 4.6(b)). Results on a Wikipedia article in Figures 4.6(c) and 4.6(d) show that in this type of documents the topic shifts are not as obvious to identify as in the case of the TV show at level 0. For the medical textbook corpus the topic frontiers could be correlated with the local minima points in the histogram for level 0 (Figure 4.6(e)), while for the second level this is not true (Figure 4.6(f)).

By looking specifically at each segment and analyzing the bursts in one segment at a time, two types of bursts can be distinguished: Bursts that are specific to each of the sub-segments contained in the segment and bursts that are shared between the sub-segments contained in the segment. The number of specific bursts for a sub-segment is given by the number of burst intervals contained between the boundaries of that sub-segment, while the number of bursts shared between sub-segments is given by the number of burst intervals crossing over the frontier between the sub-segments. If at level 0 the average number of specific bursts is significantly higher than of those shared, at lower levels this does not hold. For example, the

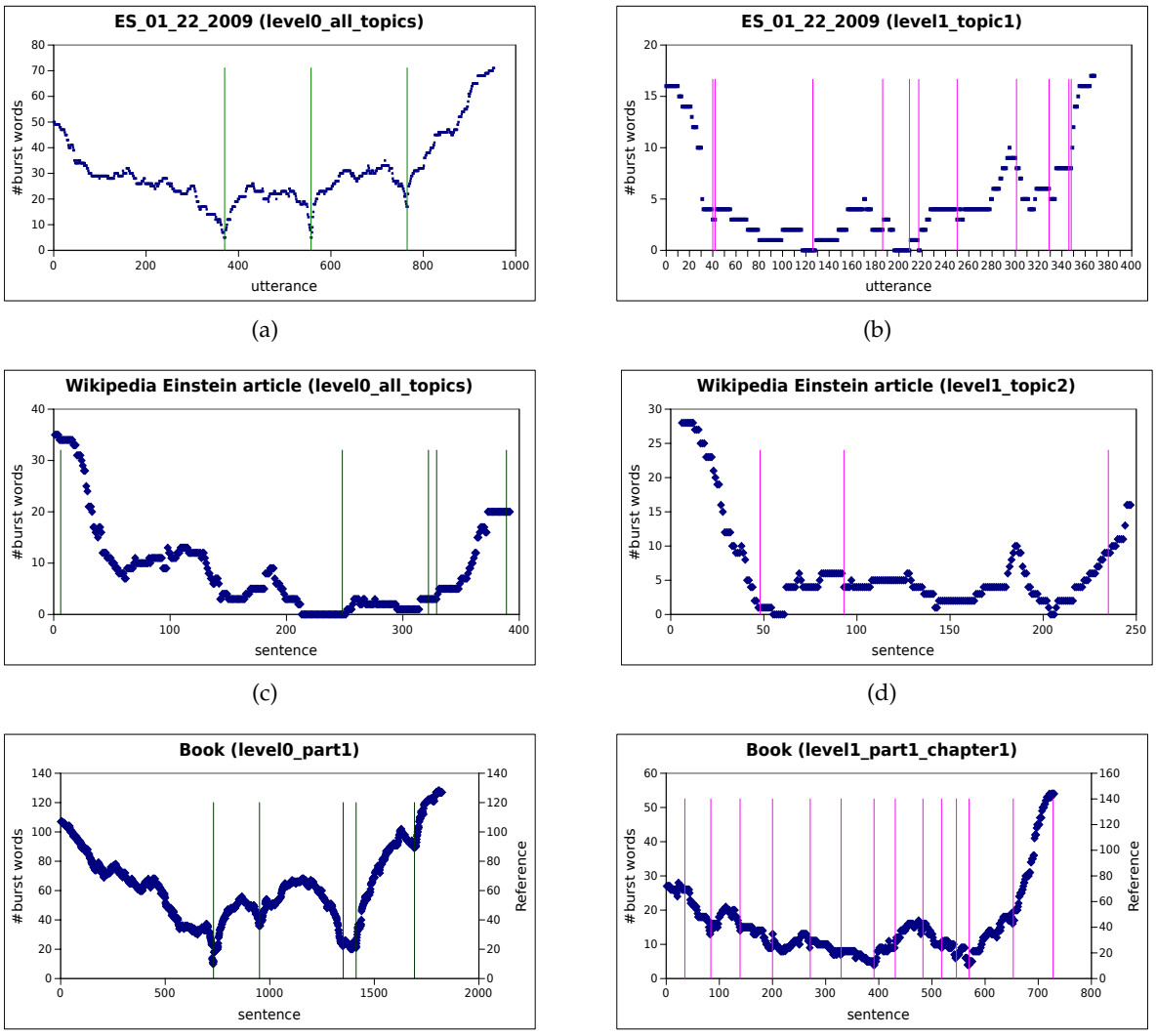


Figure 4.6: Number of bursty words for each utterance on a TV show (top) and on a Wikipedia article (bottom). Burst intervals are computed either from dense topic segments taken at level 0 (left), or from the level 1 subtopics of the first level-0 topic (right). Vertical lines indicate reference segment boundaries.

French TV show has an average number of specific (resp. shared) bursts of 51 (resp. 6.75) at level 0 while the figures decrease to resp. 2.91 and 1.58 at level 1. Thus similar observations as the ones drawn from the counts of bursts (Figure 4.6) can be made. As one would expect, if there is a significant number of bursts shared between segments as compared to the number of bursts specific to each segment, the similarity between the segments increases. Adjacent segments with a smooth change in vocabulary having no prominent concept to differentiate them one from another cannot thus be separated.

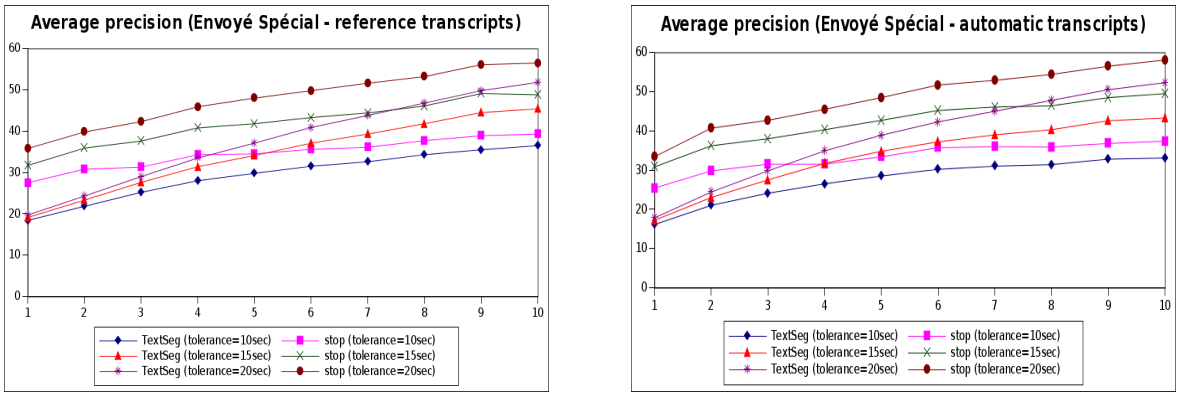
4.2.3 Discussion

This entire section was dedicated to answering the question whether lexical cohesion is enough for hierarchical topic segmentation. The analysis done indicates that the answer is *no*. The traditional measures used in hierarchical topic segmentation algorithms cannot grasp a hierarchical reference segmentation. These measures do not take into account the different importance of words in a text, that depends on their position. Several attempts have been made in the literature to modify these measures and account for this difference in importance. However, each attempt comes with several disadvantages. Given that the words that are important in the process of topic segmentation are those with increased frequency for a particular portion and with insignificant appearances in the rest of the text, they can be captured through burst analysis. Therefore we have proposed a case study to find whether burst analysis can explain a reference topic segmentation. The outcome of this study is that bursts are still not enough to fully grasp the reference segmentation. Unsurprisingly, just including bursts in traditional algorithms is not enough to solve the problem of hierarchical topic segmentation. During an internship at KULeuven, an exploratory study tackling the problem of finding a way in which the burstiness phenomenon in words could help hierarchical topic segmentation showed that the impact of using bursts is not significant. The experiments conducted in the study were done before studying bursts on the reference segmentation. The results we obtained in this section of the thesis, when studying the reference segmentations, explain why we had an unconvincing result in our exploratory study of using bursts to help hierarchical topic segmentation. Still, the case study we did on the burstiness phenomenon in the data leads to several important observations: Frontiers can be identified when there are few bursts across a position and many before/after that position; words that are bursty at one level in the topic hierarchy (i.e., specific at this level) can become general for lower levels in the hierarchy; when going to lower levels in the topic hierarchy, the number of bursts decreases; there are segments with no bursty words. Thus we consider that burst analysis may help to know when the traditional algorithms should stop being applied. This idea leads to the second part of our investigation which we address next.

4.3 When to stop applying a hierarchical topic segmentation algorithm?

The entire hierarchical topic structure of a text is difficult to fully determine: Some ideas in the data are more important than others, i.e., more salient, while others are simple fillers. Burst modeling has the effect of exposing salient words (i.e., keywords) and thus reveal the important ideas in the data [27]. Exploiting this idea, we propose to investigate whether we can find when a segmentation can be applied and when not. The main goal of this investigation is to overcome the fact that with recursive hierarchical topic segmentation, it is difficult to know when to stop. Additionally, the segments that have no sub-segments should be somehow identified, so that the segmentation is not applied on them anymore. Two ideas govern the investigation:

1. If there is no burst in a segment, we assume sub-segments cannot be identified by looking just at the lexical cohesion.



(a) (b)

Figure 4.7: Average precision curve for the French TV show dataset with the segmentation criterion applied(stop) and without(TextSeg). The segmentation is done at the second level in the hierarchy to obtain the third level. A tolerance of 10, 15 and 20 seconds is considered between the hypothesis and the reference.

2. Whenever there is a burst in a segment, we consider that at least one sub-segment can be extracted containing that burst. Frontiers are generally identified after a burst or at the beginning of a burst.

We can translate the ideas proposed in a stopping/continuing criterion, to decide if we stop or continue the segmentation. A segmentation can be performed on the segments which contain meaningful information compared to the rest of the text, i.e., segments that enclose salient words within their boundaries. These segments can be characterized as salient and can prevail at different levels of detail highlighting the prominent ideas in the text as they emerge. The ideas we defined can be integrated in various segmentation strategies like the recursive application of a linear segmentation, to decide whether a segment is worth segmenting.

Experimental evaluation was performed segmenting the automatic and reference transcripts of the French TV show Envoyé Spécial. Figure 4.7 depicts the average precision obtained on this data set, both for reference 4.7(a) and automatic 4.7(b) transcripts. To compute the precision, we perform a segmentation to obtain the third level using the TextSeg algorithm, with and without the stopping criterion. The values are obtained by varying the α parameter in the TextSeg algorithm [136]. Tolerance of 10, 15 and 20 seconds, between the hypothesis and reference segmentation, is considered. As it can be observed the precision values obtained with the criterion are higher than without the criterion. This observation holds both for automatic and manual transcripts of the TV show. The difference in precision between the values obtained with the criterion and without is increasing as more tolerance is allowed between the hypothesized and reference frontiers. This means that frontiers proposed for the segments with bursts are rather found as true positives, when more tolerance is allowed, than those proposed for segments with no burst inside. When analyzing the segments that have no sub-segments we observed that, both on the automatic and manual transcripts, those segments also have no burst. When using the criterion such segments are not segmented. However, the TextSeg algorithm also segments the segments that do not

have sub-segments according to the reference segmentation, which leads to an increase of the number of false positives.

Using the criterion for hierarchical topic segmentation has several advantages. First, the risk of proposing false positives can be diminished. Secondly, we can overcome the fact that with recursive hierarchical topic segmentation it is difficult to know when to stop. Thirdly, with this criterion, the risk of segmenting the segments that do not have any sub-segments in the reference segmentation diminishes.

Burst analysis proves to be relevant in the context of hierarchical topic segmentation. It can point out the salient segments based on the hierarchy of bursts and can help decide when to stop applying a hierarchical topic segmentation algorithm. However, an appropriate way to exploit it has to be proposed. We argue for a change of representation for the hierarchy of topics that overcomes the challenges hierarchical topic segmentation strategies have to face to obtain the classical representation. We address this open issue in the following section.

4.4 Hierarchical structure of topically focused fragments

In this section we investigate a different way of organizing the topical structure of textual content, leveraging the burstiness of words. As an alternative to classical dense (i.e., contiguous) hierarchical topic segmentation, we propose to derive a hierarchy of topically focused fragments. In classical hierarchical topic segmentation the main topics are divided into subtopics, which in turn can be divided. Departing from this model, the idea we pursue is spotting topically focused fragments that are not necessarily contiguous and organize the fragments at various levels in a hierarchical way. Exploiting Kleinberg's algorithm [82] to provide a hierarchy of bursty fragments for each word, we propose an algorithm that does an agglomerative clustering of burst fragments to build a topical organization of a document. Obtaining this structuring of the data brings several advantages: It is a representation of the entire document; It is highly informative since the words included are the most informative ones in the document; The bursty words present in the resulting fragments offer an accurate approximation of what the document is about and facilitate its understanding; Relevant information is given at various levels of detail. As a proof of concept, evaluations are first performed by qualitative and quantitative comparison to the traditional dense segmentation for which hierarchical reference segmentation exists. Then, we evaluate the structure in the context of automatic summarization.

4.4.1 Algorithm

Our clustering algorithm (denote it HTFF) exploits the output of Kleinberg's burst detection algorithm which provides for each word a hierarchy of burst intervals. The key idea of the algorithm is to iteratively group together burst intervals from distinct words at each level of the hierarchy of bursts based on their overlaps, thus yielding a nested set of clusters. We first group each two overlapping intervals to form a new interval (or fragment) and proceed until no more overlapping intervals appear. Details are given in Algorithm 2. For each level $l \in [1, L]$ in the hierarchy of bursts H, the burst intervals contained at this level for each word w form a collection of intervals I_{l_w}. Each interval $I_{l_w}(i)$ in the collection has a start $S_{l_w}(i)$ and an end $E_{l_w}(i)$ point. An exhaustive comparison between the intervals in H is done

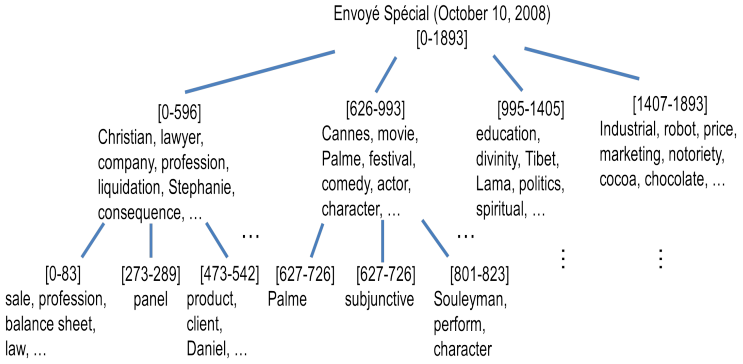


Figure 4.8: An example of a two-level hierarchy of topically focused fragments obtained with a Envoyé Spécial TV show. At each level, fragments are represented by their limits in terms of utterance number (in brackets) and characterized with the bursty words (translated from French) that helped form the fragments.

independently for each level. If two burst intervals ($I_{l_u}(i)$, $I_{l_v}(j)$) overlap, they are merged together and a new interval is obtained ($I_{l_{u,v}}(t)$) and added to the collection. This step is done until there are no more overlapping intervals. At the end the fragments corresponding to the final intervals are extracted to represent the salient fragments at level l. The hierarchy of topically focused fragments is created using a mapping across levels of the fragments obtained. An example of such a hierarchy, is presented in Figure 4.8 for two levels. The limits of the fragments formed are given by the starting and ending utterance/sentence positions and their content is represented by a sample of the bursty words that contributed in forming them. The solution we propose to create the hierarchy of topically focused fragments has the advantage of deriving the hierarchy directly, without any prior on the duration of fragments (segments in case of traditional segmentation) and number of levels in the hierarchy, unlike traditional hierarchical topic segmentation strategies.

4.4.2 Evaluation

Currently, there is no metric to evaluate the structure resulting from the algorithm above. The measures traditionally used for hierarchical topic segmentation being inappropriate for at least two reasons:

1. The structure that our algorithm outputs is a hierarchy of topically focused fragments and not a dense hierarchy of segments;

2. There is no groundtruth for this kind of hierarchy of topically focused fragments, which is required for the metrics used to evaluate traditional segmentations.

Moreover building such a groundtruth is not an easy task: The topically focused fragments are obtained in a data-driven, bottom-up, manner that does not necessarily reflect a prior organization as would be provided by human experts. In addition of being costly, annotating new data requires that clear, shared, annotation guidelines be defined first. This last point requires a good understanding and characterization of what our approach can yield, which is exactly our aim in this evaluation. To prove the relevance of our approach and provide

Algorithm 2 Create a hierarchy of topically focused segments.

for each level l **do**
  Step 0. Initialize segment clusters
  for all word w **do**
  $I_{l_w} = \{I_{l_w}(1), I_{l_w}(2), ...I_{l_w}(n_{l_w})\}$
  where $I_{l_w}(i) = [S_{l_w}(i), E_{l_w}(i)]$
  end for
  Step 1. Agglomerative clustering
  repeat
    for all $I_{l_u}(i), I_{l_v}(j) \in I_{l_w}, \forall u, v, \forall i, j, i \neq j$ **do**
      if $I_{l_u}(i) \cap I_{l_v}(j) \neq \emptyset$ **then**
  $I_{l_{u,v}}(t) = [min(S_{l_u}(i), S_{l_v}(j)), max(E_{l_u}(i), E_{l_v}(j))]$
  $add(I_{l_{u,v}}(t), I_{l_w})$
  $remove(I_{l_u}(i), I_{l_w})$
  $remove(I_{l_v}(j), I_{l_w})$
      end if
    end for
  until convergence
end for
Step 2. Mapping across levels
for $l = L \rightarrow 1$ **do**
  $I_{l_w}(i)$ mapped to $I_{l-1_w}(j)$ such that $I_{l_w}(i) \subset I_{l-1_w}(j)$
end for

a good insight into the hierarchical fragments that it outputs, we first propose to see how focused fragments compare with traditional dense segmentation. Then, we move further into applying it for other tasks.

4.4.3 Comparison with a traditional dense segmentation

We thus report here a number of measures relying on the reference dense segmentations: At each level, hypothesized fragments are compared to their counterpart in the reference dense segmentation. Conversely, reference dense segments are mapped to hypothesized fragments. We compare here both our new approach and the state of the art hierarchical topic segmentation algorithm HierBayes [36], with the reference dense segmentation. The aim is to test if with HTFF we can provide a better topical focus while covering the main topics in the data. Two measures are defined: $M1$, the proportion of topically focused fragments belonging to a unique reference segment; $M2$, the percentage of reference segments which have at least one matching topically focused fragment. The formula to compute $M1$ is defined as:

$$M1(TFF, R) = \frac{\sum_{i,j}(\delta(TFF_{i,j}, R_{k,l}))}{|TFF|} \ ,$$

where $|TFF|$ is the total number of topically focused fragments proposed, $TFF_{i,j}$ is a fragment starting with utterance i and ending with utterance j, and $R_{k,l}$ is a reference segment starting

		HTFF		HierBayes	
Data-set	level	M1	M2	M1	M2
ES(manual)	1	0.75	1	0.51	1
	2	0.56	0.74	0.15	1
	3	0.47	0.17	–	–
ES(auto)	1	0.73	1	0.48	1
	2	0.46	0.62	0.1	1
	3	0.51	0.11	–	–
Textbook	1	0.89	1	0.22	1
	2	0.72	0.64	0.06	1
Wikipedia	1	0.22	0.97	0.29	1
	2	0.62	0.66	0.42	1
	3	0.69	0.29	–	–
	4	0.49	0.06	–	–

Table 4.1: The values obtained with M1 and M2 measures on three datasets after applying HierBayes and HTFF.

with utterance k and ending with utterance l. The function $\delta(TFF_{i,j}, R_{k,l})$ is defined as:

$$\begin{aligned} \delta(TFF_{i,j}, R_{k,l}) &= 1, TFF_{i,j} \subseteq R_{k,l}, k \leq i, l \geq j \\ & \quad 0, otherwise \ . \end{aligned} \tag{4.2}$$

The formula to compute $M2$ is:

$$M2(R, TFF) = 1 - \frac{\sum_{k,l}(\psi(R_{k,l}, TFF_{i,j})}{|R|} \ ,$$

where $\psi(TFF_{i,j}, R_{k,l})$ is defined as:

$$\begin{aligned} \psi(R_{k,l}, TFF_{i,j}) &= 1, \nexists TFF_{i,j} \subseteq R_{k,l} \forall k, l \\ & \quad 0, otherwise \ . \end{aligned} \tag{4.4}$$

The values obtained with these measures both for a dense segmentation resulting from applying HierBayes and a hierarchy of topically focused fragments (HTFF) are reported in Table 4.1 on the three datasets described in the previous chapter: manual and automatic transcripts of TV shows, Wikipedia articles and the medical textbook. For HierBayes we report only the results at two levels since trying to obtain more levels worsened the segmentation, resulting in the same segments at all levels. As going to lower levels with HTFF it is expected to have such a small coverage of the reference topics ($M2$) since their number is considerably high and the average number of bursts is ≈ 1. Results obtained with the $M1$ measure demonstrate that the fragments we extract in a bottom-up manner usually have an equivalent in a dense segmentation and have a stronger focus than their counterpart. Indeed, even at lower levels, at least half of the fragments ($M1$) have a unique counterpart among the reference topics. When looking at the results obtained with HierBayes, we can observe that even at the first level in the hierarchy that less than half of the segments do not belong to a unique topic. Next, we propose an application driven evaluation of the hierarchy of topically focused fragments.

4.4.4 Application-driven evaluation

Following the steps of [7] and [80] we also consider evaluating our approach for topic structuring in the context of automatic summarization. With our new hierarchical structure of data, being not contiguous, part of the content at various levels of detail is eliminated. Therefore, before using the structure for summarization, we propose first to measure how much of the data is compressed for each data set (i.e., manual/automatic transcripts, medical textbook and Wikipedia articles), to understand if there is a variation based on the type of data. Next, we propose to analyze the compression and evaluate whether it keeps the important information from the initial data or not. For this evaluation we rely on a new corpus[2] proposed for automatic summarization in [79]. We start by applying our HTFF algorithm to obtain the hierarchy of topically focused fragments on this corpus. Next, we compute the percentage of sentences wrongly eliminated by HTFF according to the four annotator's groundthruth summaries created for this data set. The aim of this evaluation is to ensure that we do not eliminate a large part of what is important in the data since the important information is what should appear in the summary of the data. The result of the evaluation is promising and therefore we continue with the evaluation of HTFF in the context of automatic summarization.

4.4.4.1 Data compression

To measure how much of the data is compressed (i.e., kept in the final hierarchical structure) we first apply HTFF on all the data sets used for hierarchical topic segmentation. Then we measure the average compression percentage with respect to the full text (100%), at each level in the hierarchy of topically focused fragments. The results obtained are reported in Figure 4.9. The first level in the hierarchy keeps a large percentage of the initial data for the TV shows data set, compared to the Wikipedia articles and the medical textbook corpus. This is expected since each TV show contains several reports on different subjects, which will be identified through large burst intervals at the first level in the burst hierarchy. While the Wikipedia articles and the medical textbook corpora are focused on one topic alone. This results in higher compression for this corpora than the compression obtained on the TV show transcripts. On average only 44% of the data is kept for the medical textbook at the first level and 46% for Wikipedia articles, while for the manual transcripts 98% is kept at the first level. Therefore, the new structure keeps as representative different amounts of data for different types of data.

When creating summaries for textual data the aim is at keeping the relevant information in the texts. Thus, before using HTFF for automatic summarization, we propose an experiment to check if the data eliminated with HTFF contained relevant information or not. We analyze if the information we eliminate with HTFF on a text is usually kept as relevant or not in the text's summary. For this experiment, we rely on the corpus proposed for automatic summarization in [79]. It contains 20 chapters from several novels from the XIX–early XX century which have been split into two groups of 10 chapters, G1 and G2. The chapters in each group have been manually annotated at the sentence level, according to a set of instructions[3], by three different people, plus one annotator in common for both groups (i.e.,

[2]The corpus is available at http://www.eecs.uotawa.ca/ ankazant

[3]The guidelines for the annotations are available at http://www.site.uottawa.ca/ ankazant/instructions.zip

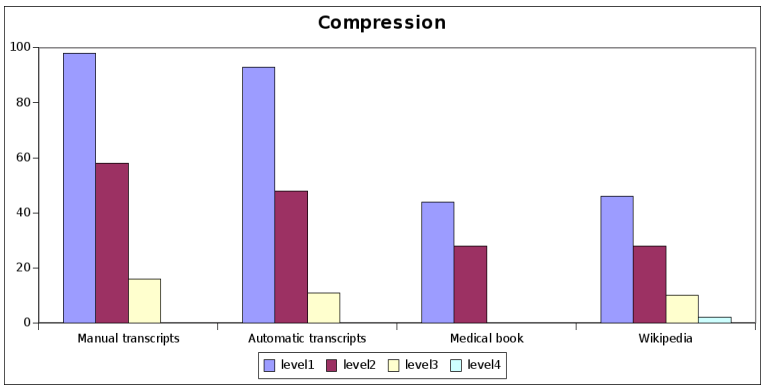


Figure 4.9: Compression statistic

Data-set		average % of sentences eliminated from the original texts	wrongly eliminated from those eliminated A1	A2	A3	A4
level 1	G1	44.5%	3.1%	2.1%	2.3%	3%
	G2	44%	5.2%	3%	0.7%	3.3%
level 2	G1	85%	5.5%	5.5%	5.6%	5.8%
	G2	89%	5.9%	5.6%	5.1%	6.1%

Table 4.2: The average percentage of sentences eliminated from the initial texts and the average percentage of sentences that should not have been eliminated according to the 4 annotators summaries.

the author of [79]). This results in four manual summaries, for each chapter in each group.

We apply our algorithm to create a hierarchy of topically focused fragments for each novel. Table 4.2 presents both the average percentage of the overall eliminated sentences and the average percentage of relevant sentences that shouldn't have been eliminated from the original texts with HTFF, at level 1 at 2. The percentage of relevant sentences eliminated is computed based on the annotations, for all texts. With the first level in the hierarchy, we eliminate around 45% of the sentences in the original texts for each group. Out of these eliminated sentences between 2.1-3.1% are wrongly eliminated for the first group and between 0.7-5.2% for the second group. At the second level we eliminate 85% for the first group and 89% for the second one, while the amount of wrongly eliminated sentences increases by about 3 for the first group, while the second group has a percentage of wrongly eliminated sentences between 5.1-6.1%. This means that the hierarchy of topically focused fragments manages to keep the most important parts of the data according to what is usually found in the summaries of this data. We propose next to evaluate the impact of using the new structure when generating automatic summaries.

4.4.4.2 Automatic summarization

Given our new structure we can tackle the task of automatic summarization from two angles. On the one hand, we can evaluate whether a state-of-the-art automatic summary generator can benefit from the compression of the texts achieved with HTFF or not. On the

							Stemmer					
	Original			Compressed			Original			Compressed		
measures	R	P	F	R	P	F	R	P	F	R	P	F
Rouge-1	0.451	0.442	0.446	0.456	0.448	0.452	0.467	0.458	0.463	0.472	0.464	0.468
Rouge-2	0.175	0.171	0.173	0.188	0.184	0.186	0.177	0.174	0.176	0.19	0.187	0.188
Rouge-3	0.127	0.124	0.125	0.144	0.142	0.143	0.127	0.125	0.126	0.145	0.142	0.143
Rouge-4	0.115	0.112	0.113	0.134	0.131	0.132	0.115	0.112	0.113	0.134	0.131	0.133
Rouge-L	0.429	0.421	0.425	0.434	0.426	0.43	0.443	0.434	0.438	0.447	0.439	0.443
Rouge-W	0.128	0.239	0.167	0.19	0.191	0.191	0.131	0.245	0.171	0.135	0.253	0.176

Table 4.3: Recall, Precision and F1-measure for Rouge 1-4, Rouge-L and Rouge-W, measures obtained on the novels corpus, original and compressed versions, with ILP-sum. Results for the stemmed version of the input are also given.

other hand, we can create summmaries directly, leveraging HTFF to retrieve the summary at different levels of details. This approach would require a groudthruth of summaries at different levels of details to assess the quality of the entire hierarchy. Since we do not have such a groudthruth, we could limit at one level in the hierarchy. Still, a limited sized summary needs to be proposed, in order to be able to compare it with another method. This is necessary because the measures used for summary evaluation would be biased towards longer summaries. We will focus thus on the first angle, an appropriate way to directly use the hierarchy for summary generation remains to be found.

In order to test if a summarizer can benefit from the compression of the text, our first experiment consists in running the automatic summary generator proposed in [53] (denote it ILP-sum[4]) both on the original texts in the novels corpus and on the compressed texts (i.e, level 1 in the hierarchy of topically focused fragments). The quality of the summaries is classically assessed using Rouge metrics [87]. The Precision, Recall and F1-measure scores obtained for Rouge-1, Rouge-2, Rouge-3, Rouge-4, Rouge-L and Rouge-W are given in Table 4.3. Rouge-n compares the n-grams contained in the generated summary with those in the groundtruth summaries. Rouge-L takes into account the longest common subsequence (LCS) between the generated summary and the groundtruth, while Rouge-W gives consecutive matches of length L in a LCS a weight of L^{weight} instead of just L. For our evaluation the traditional weight of 1.2 is chosen. *Compressed* in the table refers to the summaries obtained on the compressed data using the hierarchy of topically focused fragments. All the results on the compressed data are higher than those obtained when the entire data is considered. This can be explained by the fact that the new structure can help bring to surface other information that would not be included in the summarizer when an entire text is considered. With the compression, parts of the data where the important words tend to fade away are disregarded, while, when considering the entire text they are still seen as highly informative.

In [90], the authors propose several evaluation strategies to automatically assess machine summary content without a gold standard. They show that by quantifying the similarity between the source text and its summary with appropriately chosen measures, they can replicate human assessments accurately. Therefore we consider using their proposed measures to evaluate the quality of the summaries produced on the TV shows corpus. Similar to our previous setting, we create summaries using ILP-sum both on the original input and on the compressed version using HTFF. The quality of the summaries is evaluated based on the distribution of terms in the input and the summaries. The intuition motivating this evalua-

[4]https://github.com/boudinfl/sume

	TV shows (automatic)		TV shows (manual)		Novels	
systems	Avg JSd	Avg JSd smooth	Avg JSd	Avg JSd smooth	Avg JSd	Avg JSd smooth
Input–Original	0.54	0.46	0.53	0.45	0.36	0.32
Input–Compressed	0.53	0.46	0.53	0.46	0.37	0.33
Input–Groundtruth	–	–	–	–	0.4	0.35
Groundtruth–Original	–	–	–	–	0.47	0.45
Groundtruth–Compressed	–	–	–	–	0.46	0.44

Table 4.4: JSd scores obtained on the TV shows and novels corpora.

tion in [90] is that good extractive summaries will tend to be similar to the input in terms of content. Among the measures proposed in [90], Jensen Shannon divergence (JSd) obtains the best correlations between manual and automatic scores. The intuition behind this measure is that the distance between two distributions cannot be very different from the average of distances from their mean distribution. JSd is defined as:

$$JSd(P||Q) = \frac{1}{2}[D(P||A) + D[Q||A]] \ ,$$

where P and Q are the input and summary word distributions, $A = \frac{P+Q}{2}$ is the mean distribution of P and Q. The divergence between two probability distributions P and A is computed as:

$$D(P||A) = \sum_{w} p_P(w) log_2 \frac{p_P(w)}{p_A(w)} \ .$$

We report in Table 4.4 the results obtained for JSd for the summaries obtained on the TV shows and novels data sets. We use the tool made available by the authors of [90][5]. In this table *Input* refers to the original text to be summarized. *Original* is the summary obtained with ILP-sum applied on the entire text and *Compressed* are summaries obtained with ILP-sum on the compressed version of the text. Both smoothed and unsmoothed versions of JSd are reported. For the novels data set we report also the JSd scores between input and groundtruth (i.e., manual annotated summaries), groundtruth and original and groundtruth and compressed summaries. Lower divergence scores are better. As it can be observed the differences between the methods are small or non-existent in terms of JSd measure. All the evaluation strategies proposed so far show that with the new proposed structure of topically focused fragments, we succeed in keeping the important information from the initial data.

4.5 Discussion

In this chapter we have done an investigation of current hierarchical topic segmentation strategies and focused on addressing their limits. The first part of this chapter has been dedicated to analyzing the limits of lexical cohesion in the context of hierarchical topic segmentation. We showed that global measures of lexical re-occurrence are not adequate to detect topic shifts, while the temporal distribution of word re-occurrences, i.e., burst analysis, provides strong cues. Burst analysis helps extract the important ideas in the data however it cannot fully grasp the reference segmentation. Thus, just including it in traditional algorithms is not enough to solve the problem of hierarchical topic segmentation. Still, burst analysis helps to know when the traditional algorithms should stop being applied. This

[5]http://homepages.inf.ed.ac.uk/alouis/IEval2.html

addresses a major challenge hierarchical topic segmentation approaches have to face. The advantages shown by burst analysis, i.e., exposing salient segments based on the hierarchy of bursts and helping decide when to stop the segmentation, prove it is relevant in the context of hierarchical topic segmentation. As a consequence, we have proposed an algorithm to extract a hierarchy of topically focused fragments using agglomerative clustering of burst intervals. Comparison of this novel structure to a reference dense segmentation on several data sets has indicated that a better topic focus can be achieved than the one provided by the reference dense segmentation while retrieving all of the important aspects of a text. Additionally, a task driven evaluation in the context of automatic summarization has proven that using this representation for the data to summarize does not affect negatively the outcome of the summarization. The results validate the seminal idea of the salient fragment paradigm and justifies further work on the design of annotation guidelines and on the subsequent annotation of data.

The investigation done in this chapter opens several interesting paths to address the problem of hierarchical topic segmentation. On the one hand, if we choose to remain in the classical dense segmentation space, new concepts need to be leveraged and most probably they need to be characteristic for the data. One way could be to exploit rethorical relations [93] at lower levels. Topic frontiers at lower levels could rather be justified by relations such as: elaboration, justification, cause, etc., than by sudden changes in vocabulary. However, there is no generic automatic solution, yet, for detecting such relations [8]. Also, correlations between topic/sub-topic frontiers and such relations need to be studied in depth. Some possible correlations have been suggested [72] and an initial attempt was proposed in [20]. On the other hand, we could continue investigating the hierarchy of topically focused fragments and work on improving it. First, the burst detection model could be improved by adding semantic relations, or more complex models for analyzing the distribution of a word. Since Kleinberg's algorithm in 2002, other more complex models have emerged [109, 121, 91, 4]. A way to include them in Kleinberg's algorithm would be ideal since it has the burst hierarchy feature. For creating the topical hierarchy other approaches could be envisioned. Instead of clustering together all overlapping segments, we could analyze the overlap percentage before deciding what to merge together. Not all burst intervals that overlap should necessarily be combined into one fragment. Regarding the application driven evaluation, the next step would be to extract the summary directly from the hierarchy and perform evaluations on larger data sets (e.g., data from DUC and TAC summarization tracks).

This chapter ends the first part of this thesis, focused on automatic topical structuring of TV shows in particular. As we argued in the introduction of this thesis, having a structured representation of television content can help extract value from it. Various benchmarking initiatives emerged to foster the interest of the multimedia community to process, analyze and derive value from various structured and unstructured data such as text, speech, audio, video, image, multimedia. One such benchmark is the MediaEval initiative proposing various tasks such as Search and Hyperlinking, Placing: Multimodal Geo-location, Multimodal Person Discovery in Broadcast TV, etc. Relying on language processing techniques has proven to be very successful in these tasks. Therefore, we consider evaluating the potential of the solutions we proposed for structuring audiovisual content in the context of a task at MediaEval. We choose the Search and Hyperlinking task since it gives us the possibility to test the implications of the topical structures in linking video content based on topics. Video

hyperlinking consists in creating links that originate from parts of video material and point to other relevant content. Thus a prerequisite for creating hyperlinks is to segment the video into linkable content, i.e., meaningful pieces of information. For humans this task comes naturally. However doing it by hand becomes a tedious task, while developing a method to do this automatically is a complex problem. Video hyperlinking will thus represent the focus of the second part of the thesis. We will start the next part by first describing the task of video hyperlinking as proposed in the MediaEval benchmark, giving details on the context, approaches proposed so far, data and evaluation protocol. The following chapters will present our approaches for video hyperlinking, relying on the structuring techniques proposed in the first part of the thesis.

Part II

Implications of the topical structure in video hyperlinking

Chapter 5

Video hyperlinking

Contents

5.1 Context 71
5.2 SH and SAVA at MediaEval benchmark initiative 72
5.3 Conclusion 81

5.1 Context

Previous work on hyperlink generation falls into one of two categories: Hypermedia system modeling, whose goal is to develop models for hyperlinks in multimedia content; Link generation with the goal of dynamically creating links between both text and multimedia documents [38]. Hypermedia system modeling defines how individual pieces of information relate to each other at different levels [63], focusing on how the data is stored, the navigation capabilities of the system, the link representation and traversal as well as on user adaptation. On the contrary, link generation places emphasis on the creation of the links from a content-based analysis perspective. In particular, link generation usually targets alternate ways of searching information in large collections of multimedia data, providing information seeking and browsing capabilities in addition to search.

Content-based link creation has been initially addressed in the hypertext community with the goal of enriching texts with hyperlinks [1, 141]. Hypertext authoring has so far mainly been considered for well-structured documents (e.g., mails, Wikipedia articles) or in limited collections, typically to browse among documents retrieved as a response to a query. The idea of organizing in threads the result of multimedia search is also exploited in [117] for videos. Extending the idea of hypertext authoring, seminal work on topic threading in the broadcast news domain have considered time-aware collections [76, 142], addressing the temporal issue in an ad hoc way. The Search and Hyperlinking evaluation at MediaEval,

and more recently at TRECVid, further introduces the notion of selecting the targets of a link in a TV stream, as pointed out in [119, 62, 38, 40, 105]. Globally, the idea is that of creating hyperlinks within video data based on content analysis and comparison, where links might reflect various types of relations between the source (i.e., anchor) and target fragments of the link. For instance, in multimedia data, links can reflect the presence of similar entities, e.g., images, locations or speakers, in the two fragments, or a semantic proximity, e.g., related to a topic or an event. Therefore, hyperlink generation does not only require to assess the relevance between two content items but also to identify said items, i.e., to find the boundaries of an hyperlink source and target segments. This requirement raises an interesting question regarding the granularity level for decomposing the video content, i.e., how to structure it. We believe that this aspect of hyperlinking allows us to test our structuring approaches in a more realistic scenario. Therefore, in Chapter 6 we propose an exploratory study on how the topical structure of videos can help extract precise anchor and target segments.

After structuring the videos, the next important aspect of hyperlinking is linking the anchors with the targets. An important characteristic of the links created is that they should offer diversity. We believe that the main purpose of hyperlinking is to provide complementary information that would not be found at search time. By offering diversity in the links we can understand better the user needs and help them maximize their ability to explore a collection encouraging serendipitous encounters, differentiating the task from a typical search scenario. The notion of serendipity has various definitions in the literature, such as: pleasant surprise [97], accidental discovery [115], unexpected relevance [134], unexpected encounters that are semantically cohesive, i.e., relevant to some information need of the user [16], etc. In the past years, several attempts have been made to introduce serendipity into browsing systems, in various contexts. Examples include TweetMotif [83], for serendipitous tweets recommendation, Auralist [148], for serendipitous music recommendation, Googles' attempt for a serendipitous search engine [1], Wikipedia articles in StumbleUpon, for serendipitous Wikipedia articles [64], etc. As mentioned in [134], the importance of serendipity has been long recognized. It has been proven to improve user satisfaction [148], encouraging either an existing direction or a new direction in information seeking [43]. The authors of [134], stress the need for a principled model of serendipity and a systematic way of identifying serendipitous information. Therefore, after testing our approaches for structuring videos to extract anchors and targets we will investigate the problem of serendipity in the creation of links in Chapter 7.

All our experiments concerning video hyperlinking will be conducted in the framework of the Search and Hyperlinking (SH) and Search and Anchoring in Video Archives (SAVA) tasks at MediaEval, that precisely aims at developing hyperlink generation in broadcast videos, as a complement to a search engine [38]. We present next the tasks, giving details about existing approaches, the data and evaluation protocol.

5.2 SH and SAVA at MediaEval benchmark initiative

Implemented since 2012 in the framework of the MediaEval benchmark initiative, in 2015, the Search and Hyperlinking challenge transformed into the Search and Anchoring in Video Archives [39] task at MediaEval, while, the Hyperlinking sub-task became a new task in the

[1] http://techcrunch.com/2010/09/28/eric-schmidt-future-of-search/

Figure 5.1: Terminology used for describing the hyperlinking subtask

TRECVid initiative. This resulted in three different sub-tasks: the *Search* sub-task aims at returning a ranked list of video segments that are relevant to a textual user query. The *Anchoring* sub-task focuses on the automatic selection of video segments, from a list of videos, that can be used as anchors to encourage further exploration within the archive. The *Hyperlinking* sub-task aims a creating links between anchor segments and short video segments, called targets, which should offer complementary information to that found in the anchors.

To sum up, the search, anchoring and hyperlinking tasks represent key elements of an audio-visual archive exploration scenario. It starts with archive search assuming an information need. Then, the search becomes serendipitous or exploratory by following hyperlinks. Figure[2] 5.1 introduces the terminology used for the anchoring and hyperlinking sub-tasks together with an example. In the example, the anchor segment is represented by a person standing in front of a Fish&Chips restaurant in London, talking about things to do in London. A relevant target for this anchor is a video segment containing information on how to cook Fish&Chips. Other relevant targets can be envisioned, such as video segments about other restaurants, about how Fish&Chips became a famous dish in England, about other traditional dishes in England or other countries, etc. Within the hyperlinking scenario the aim is to explore multimodal access over multimedia content, and linking of video content to support potential user interests or needs. The following subsections consist in presenting the existing approaches developed in the context of the hyperlinking and anchoring tasks, the data sets we will also use in our experiments and the evaluation protocols employed.

[2] The figure is taken from the task presentation which can be found at http://www.slideshare.net/robinaly/me13sh-task-overview

5.2.1 Existing approaches

The hyperlinking task at MediaEval has been implemented for three years, leading to various approaches being proposed. Ideally, the anchor segments for hyperlinking should be automatically identified. However, this is rather complex and in order to simplify the hyperlinking task, the task organizers have given manually predefined anchor segments. Thus, the participants had to propose relevant targets for predefined anchors. An exploratory study for automatic anchor selection has been recently proposed at MediaEval in the context of the SAVA task which lead to a couple of approaches being proposed. We will present next the existing approaches for both tasks.

Hyperlinking with predefined anchors. The hyperlinking task has been mostly handled as an information retrieval task. Indeed, the approaches proposed for the search task have been adapted for hyperlinking by considering the anchor segment as the query. The existing approaches for hyperlinking, need first to structure the video to generate a list of potential target segments from the collection. Next, these targets are ranked based on how relevant they are for a given anchor segment.

The first step is usually done using fixed-length segmentation [49] or topic segmentation strategies [128, 108, 13, 47] or relies on video shots [138, 84] or the utterances (i.e., sequences of words separated by breath intakes) in the automatic transcripts [123]. Using fixed-length segmentation seems to outperform most segmentation approaches [38]. However, this approach has the disadvantage of having to set the window size and deal with overlapping segments. In [123], the authors note other limitations of fixed-length segmentation respectively that the evidence for a relevant passage can be divided between two segments and that the segments are too long (i.e., they do not contain only the relevant information).

For the target selection step, most approaches rely on pairwise content-based proximity exploiting subtitles, automatic transcripts or visual content. Some authors enrich text and visual content with additional information, e.g., named entities [32, 128], metadata(e.g., title, description, synopsis) [108, 13], prosodic information [49], or OCR [84]. In most cases, a vector space model is used to represent the content of the anchor and the target along with standard similarity measures. Several works combine textual and visual information via a weighting scheme between separated approaches [49], or sequence the process to use one as a prefiltering step to the other method [13, 84]. In 2013, target segments from the same video as the anchor segment were allowed, while in 2014 they were not. Still, the target segments are allowed to overlap with previously returned segments. Human-based evaluations done within MediaEval hints that the best systems were those that proposed targets very similar to the anchor.

Automatic anchor selection. Three different approaches have been proposed for the anchor selection task [124, 48, 139]. We will discuss here two of them, since we will present our approach in Chapter 6. In [139], the authors propose to use social activity on Twitter to find topics in the videos on which people have questions. The more Twitter questions are associated to a topic discussed in a certain shot, the greater the likelihood that the corresponding part of the video represents an anchor. The relation between the questions on Twitter and the topics is reflected through the number of keyphrases (i.e., noun phrases in their work) shared. For their approach the authors rely on subtitles to extract noun phrases. A final

list of keyphrases is build for crawling, after being reduced using several heursistics on the content of the keyphrases (e.g., at least one capital letter).

The approach proposed in [48] also relies on the subtitles. First, they segment each video using either fixed-length segmentation or a machine learning-based method to obtain a list of all possible anchor segments. Then, they rank the anchor segements using two different strategies. The first strategy consists in placing the task in an information retrieval scenario. For this they create queries using the metadata (i.e., program name and short description) of the video from which they want to select anchors and compute similarity scores between the query and each potential anchor. The second strategy consists in ranking the segments according to the frequency of numbers and proper names contained.

5.2.2 Data

Two data sets are considered in this work, corresponding to the data used for the SH task in the MediaEval benchmark in 2013 and 2014. The data set used for evaluating the anchors selection from videos, corresponds to a subset (i.e., 33 videos) of the 2014 video collection. The entire data contains a collection of videos provided by the BBC of approximately 4,000 hours of videos with an average length per video of 45 minutes. The videos were broadcast between 01.04.2008 and 31.07.2008 and they are very diverse. They contain news, TV series, documentaries, children shows, sports, entertainment shows, etc. In addition to the video content, organizers also released reference and automatic transcripts, metadata (e.g., cast or synopsis from the BBC website) and visual information such as shot boundaries, concept detection and face detection provided by partners of the AXES EU-project. All videos were transcribed by human experts and by several ASR systems, such as LIMSI [52] and LIUM [118]. Regarding the visual information, each video is represented as a set of keyframes for which visual concepts scores are available. There are 1,537 visual concepts (i.e., text captioning for an image), composed of the 1,000 classes of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2010 plus a new set of images of 537 classes indicated as "popular" by ImageNet [135]. Figure 5.2 illustrates how a visual concept is defined and associated with a keyframe. The example is given for a keyframe from the Top Gear TV show, associated with the visual concept *car* (where wnid, is the id of the visual concept).

In the context hyperlinking, for the 2013 (resp. 2014) data set 29 (resp. 28) users with age between 18 and 30 were asked to define realistic anchors. The anchors defined are segments of variable lengths which the users found interesting or relevant watching a subset of the video collection. Each user was required to provide a description of what they wish to see, given the anchor they define. For example for the anchor extracted from a video dealing with evolution in football, the user added the following description: "I want to see more videos about a comparison on how football has changed in 50 years". This description is only intended to help in the evaluation process and it is not available to use for the hyperlink generation.

For the 2013 and 2014 evaluations within MediaEval, 30 anchors among those predefined by users were chosen by the task organizers for each data set. Table 5.1 reports the average anchor duration with the 95 % confidence interval, on the two data sets: The duration of the anchor segments defined for the 2014 campaign is reduced compared to that defined for the 2013 one. Comparing the results between 2013 and 2014, it is clear that changes made to the

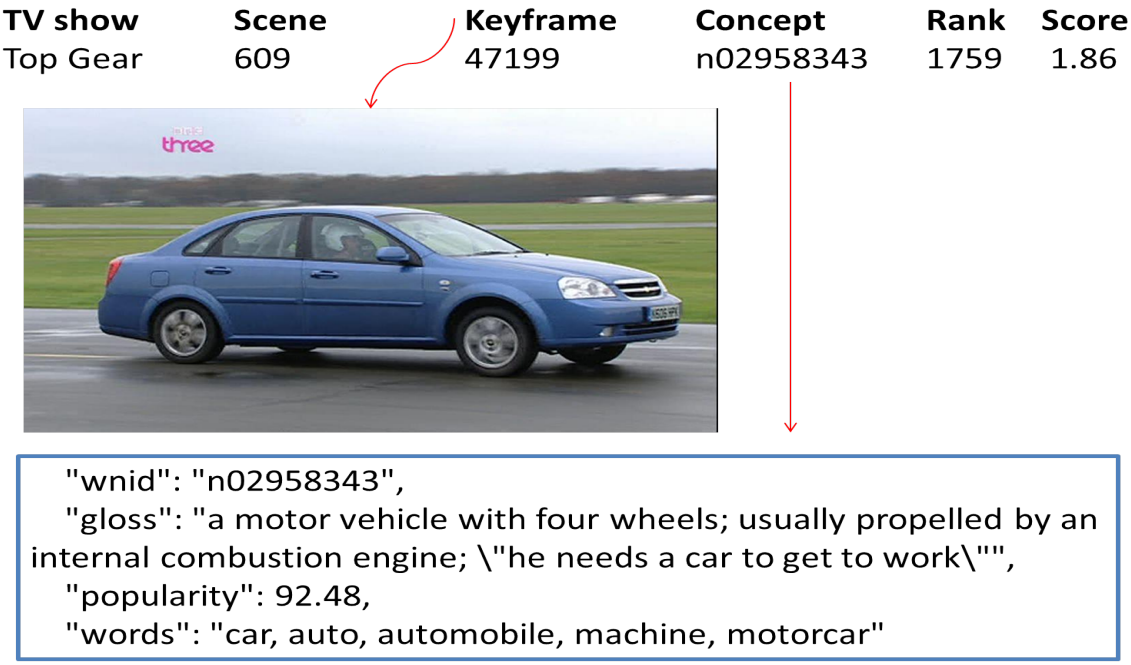


Figure 5.2: Example with the visual concept 'car' associated with a keyframe from the Top Gear TV show.

data-set	average anchor duration	confidence interval 95 %
2013	32.2 sec.	[13.4, 51]
2014	22.9 sec.	[11.1, 34.8]

Table 5.1: Average anchor segment duration with the corresponding 95 % confidence intervals, on both data sets.

task definition in 2014 (shorter anchor segments and no context information[3]) made the task more difficult. The systems that participated in the task proposed a total of 9,973 targets for the 2013 data set among which 29.9 % were judged relevant by assessors, and 12,340 for the 2014 one, with 15.3 % judged relevant. Thus, there are more non relevant targets proposed by the systems participating in 2014.

5.2.3 Hyperlinking evaluation

Performance evaluation was done through the crowd-sourcing Amazon Mechanical Turk (AMT) platform. This crowd-sourcing evaluation aims at analyzing the hyperlinks/anchors provided by task participants in real-life scenarios, more focused on user needs than an evaluation based on a ground truth defined *a priori*.

[3]The context here refers to the video content surrounding the anchor segment

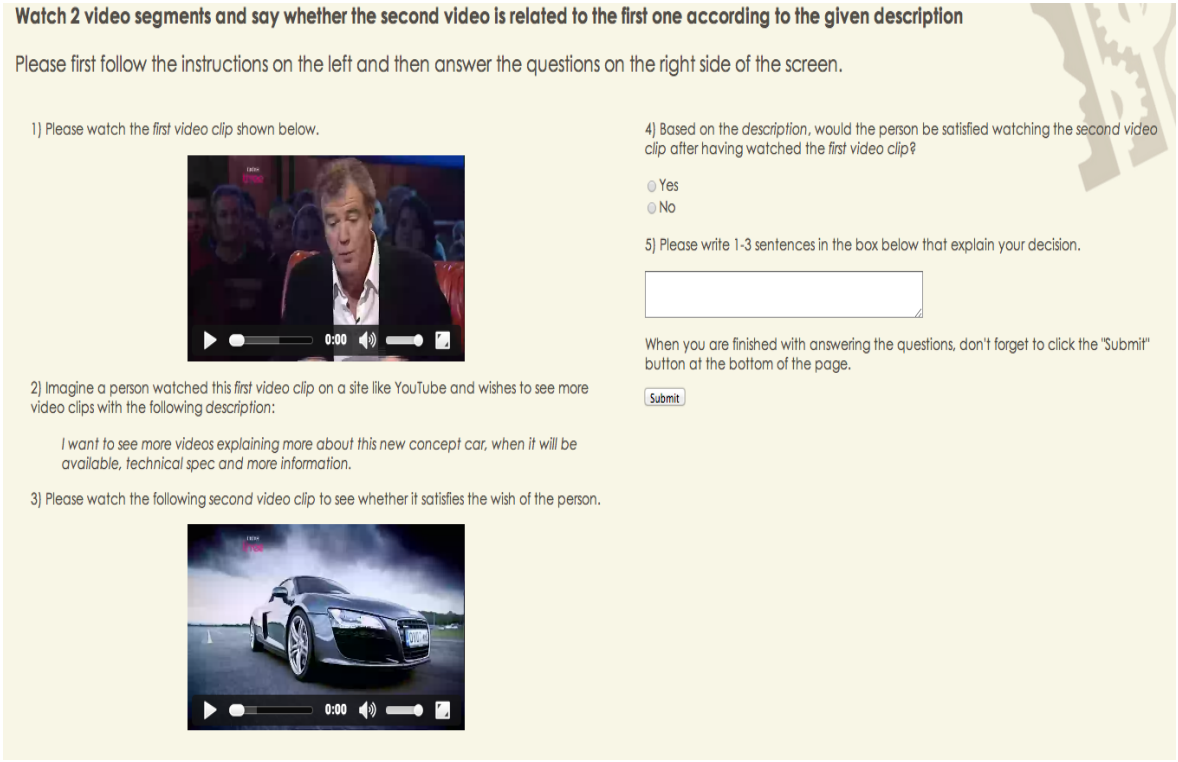


Figure 5.3: Evaluation scenario with AMT from the 2013 challenge.

5.2.3.1 Scenario

The top 10 hyperlinks targets returned for each 30 anchors by each system proposed by the participants were judged for relevance using the AMT platform. For each anchor-target pair there is only one relevance assessment made, due to the fact that having more judgements becomes rather expensive. Figure[4] 5.3 depicts an example of evaluation scenario done with AMT, from 2013. For each link, the turkers were asked to judge and explain whether the target is related to the anchor and satisfies the wishes of the person who defined the anchor (e.g., "The second video does not contain any information on change in football as the user requested. So the user will not be satisfied watching the second video after watching the first one."). Asking the turkers only if the person that defined the anchor would be satisfied with the target does not capture the diversity of information in the targets proposed. Both very similar targets and targets on related topics will be judged the same.

As evaluation was done by creating a pool containing the top 10 results from each system for each anchor, not all targets got to be evaluated. Thus, in order to estimate the performance of every system developed for the task, runs that were not judged by turkers were evaluated through the judgments obtained from turkers for the top 10 results across all the systems. Those runs are therefore only partially evaluated since some hyperlinks targets that were not returned by any other system in top 10 could not be evaluated.

[4] The figure is taken from the task presentation which can be found at http://www.slideshare.net/robinaly/me13sh-task-overview

5.2.3.2 Relevance assessment

Relevance of the hyperlinks targets is measured traditionally by means of their precision at 10. However, as mentioned in [5], this measure does not always reflect the effectiveness of a system and ignores the diversity of the results, which we consider important in a hyperlinking scenario. For instance, if a system proposes several targets in a small time window that corresponds to a video segment considered as relevant by turkers, then all the targets are considered as relevant, giving a high value for precision at 10. To avoid this issue, task organizers have proposed three different options for computing the precision at 10:

1. the overlap relevance, noted *P_10*, where hyperlinks targets are considered as relevant if they overlap with a relevant segment (from the pool of segments judged relevant with AMT),

2. the binned relevance, noted *P_10_bin*, where hyperlinks targets are included in a 5 minutes long time window. If there is a relevance judgment within the window, then all targets in that window are considered as relevant,

3. the tolerance to irrelevance, noted *P_10_tol*, where a hyperlink target is considered as irrelevant after 15 seconds of viewing non-relevant video content.

Figures 5.4, 5.5 and 5.6[5] illustrate the three metrics as depicted in [5]. The judgement for the results (Result 1, Result 2, Result 3, Result 4), having 0 denote not-relevant and 1 relevant, is as follows: in Figure 5.4 0,1,0,1; in Figure 5.5 0,1; in Figure 5.6 0,1,s,S. For the binned relevance, Result 1 and Result 3 and similarly Result 2 and Result 4 are merged into bin 1 and bin 2 respectively. A bin, i.e., a window of 5 minutes, is considered relevant if it contains at least one passage of relevant content (e.g., bin 2 and 3). For the tolerance to irrelevance, Result 4 is judged non relevant, because it has already been seen through Result 2. Similarly Result 3 has been seen through Result 1.

Additionally the task organizers have done a relevance assessment by verifying manually or automatically the crowd-sourcing judgments. For the 2013 evaluation, approximately one third of the judgments was checked manually (3,662 out of 9,973), and approximately 10% of them (328) contained errors (i.e., the relevance/irrelevance decision was incorrect, missing, or both appeared). This shows that some of the turkers evaluating the task have not understood the task or haven't done it properly. Verifying all judgments is costly, but maybe if the turkers are presented a short demo to understand better the task could help. Such a demo could present anchor-target pairs to the turkers for judging and after a judgment is made the correct answer can be shown.

If the absence of ground truth reflects real life, where user needs are not known in advance, it also makes it more difficult to figure out what kind of related content to return, which led us to make decisions and adopt strategies, typically trying to introduce diversity in our results. However the evaluation does not necessarily reflect this. For example, hyperlinks targets extracted from the same video as the anchor were considered relevant. Segments, overlapping with previously retrieved targets are allowed. Also, only one relevance judgement per anchor-target pair which means that the evaluation is highly subjective. Regarding the judgement given by the turkers, whether a target is relevant or not for an anchor,

[5]The figures are taken from [5].

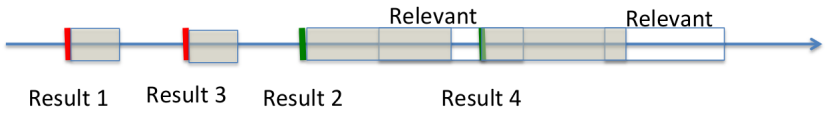


Figure 5.4: Overlap Relevance: segments are relevant if they overlap with a relevant segment. The relevance assessment for the example is $rel = 0101$, where 0 signifies relevant and 0 non-relevant, for the results: Result 1, Result 2, Result 3, and Result 4.

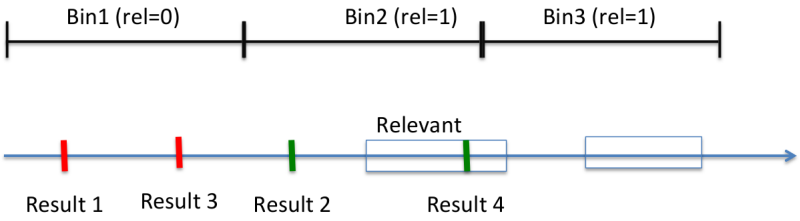


Figure 5.5: Binned Relevance: relevant segments are put into bins; A segment is relevant if there is a relevance assessment in the bin the start time of the segment fits into. The relevance assessment for the example is $rel = 01$.

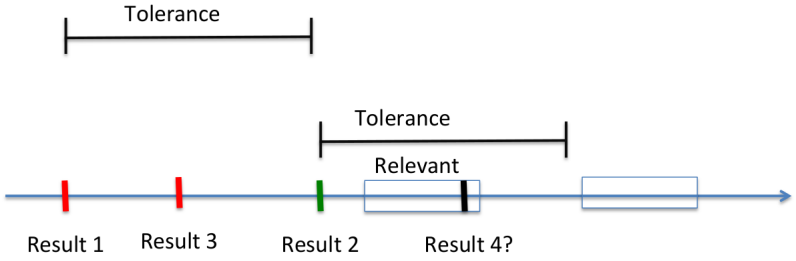


Figure 5.6: Tolerance to irrelevance: only the start times of segments are considered. The relevance assessment for the example is $rel = 01sS$, where s (S) means the result has already been seen through another non-relevant (relevant) one.

is not enough to evaluate diversity and even more serendipity. Serendipitous encounters consist in more than just being relevant, they need to be unexpected and interesting. Despite the proposed evaluation method, we choose to go in the diverse and serendipitous direction and explore in Chapter 7 ways to induce serendipity by diversifying the links in a controlled manner.

5.2.4 Anchoring evaluation

To evaluate the submissions from all participants, the top 25 anchors proposed for each video in the data set, by all participants, were judged by AMT workers. All overlapping anchor segments were combined. The workers gave their opinion on the anchor segments

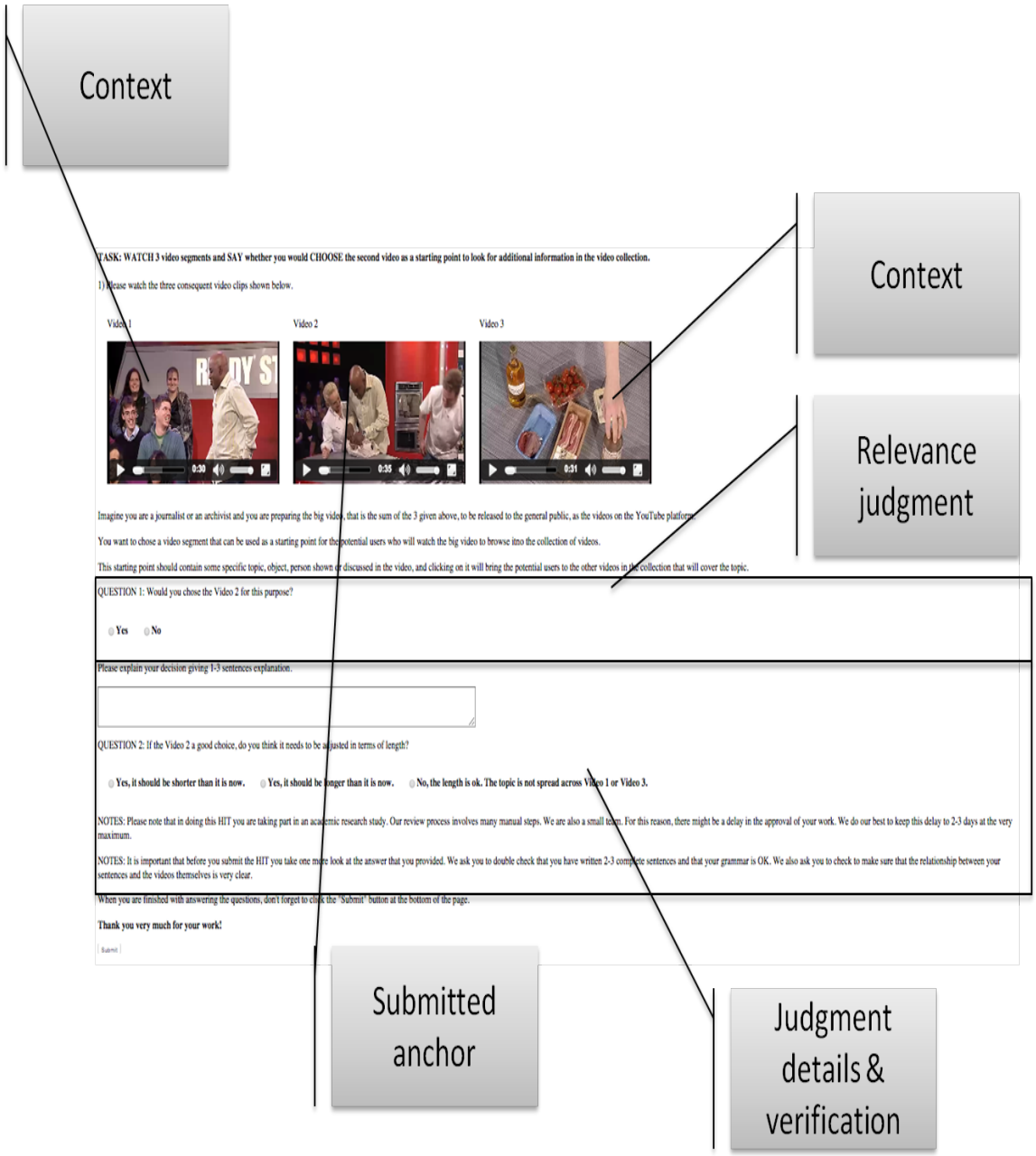


Figure 5.7: Evaluation scenario with AMT for the SAVA task.

taken from the context of the videos as illustrated in Figure 5.7[6]. Precision@10, Recall and mean reciprocal rank (MRR) metrics are used to score the runs of the participants. The MRR measure is calculated as the reciprocal value of the rank of the first correctly retrieved anchor.

[6]Figure taken from the task overview in 2015

5.3 Conclusion

One of the main challenges of hyperlinking is the creation of links between anchors and targets based on related but diverse content, without knowing the users interests. Existing approaches seem to provide targets with very similar content to that of the anchor. We believe that diversity in the links is a necessary characteristic when selecting the targets. At the basis of good links is the creation of potential anchor and target segments. Therefore, in Chapter 6 we focus on providing solutions for precise (i.e., with precise jump-in points) anchor and target selection. We will use domain-independent techniques and search to create and link, anchors and targets, that are topically coherent and thus rely on the topical structure of the videos. Then, in Chapter 7, we will be interested in solutions that offer diversity in the links favouring serendipitous encounters and have the potential to explain the nature of a link (i.e., why is this link proposed?).

Chapter 6

Investigating domain-independent techniques for precise anchor and target selection in video hyperlinking

Contents

6.1 Our approach in a nutshell 83
6.2 Hyperlink creation 84
6.3 Experiments 88
6.4 Conclusion 92

6.1 Our approach in a nutshell

The key challenges of video hyperlinking are identifying anchors and targets for potential links, selecting targets with relevant content for the anchors and dealing with the multimodal nature of the anchors and targets. In [103], the authors note that after doing a user study for manually defining anchors, the users referred primarily to spoken content and whole scenes. This motivates us to continue exploiting the spoken data obtained from automatic speech transcripts, favouring semantic links as opposed to similar visual content. We will investigate generic approaches for the selection of precise hyperlink anchor and target segments. We believe that precise anchor and target selection is a crucial step: Wrong timestamps within semantically related videos can make the result useless even though the video is per se relevant.

The anchor selection part consists in automatically extracting video segments for which users could require additional information to explore a video archive. Our approach con-

sists in structuring each video as a *hierarchy of topically focused segments* using the algorithm we proposed in Chapter 4, i.e., HTFF. This structure helps to extract segments with precise jump-in points and at various levels of details. Once extracted, the segments will be used to select anchor segments for the videos. The advantage of using HTFF is that it helps to identify the salient information in the videos, skipping irrelevant information. Moreover, having a hierarchical representation, the segments we provide as results can be at different granularity, i.e., more specific or more general. Anchors that cover a more general topic or different points of view on some topic can be selected. We believe that such an approach brings focus to what is extracted from the videos. The evaluation is carried out in the context of the SAVA challenge at MediaEval 2015 [39].

For the target selection part we experimentally compare two methods that we propose. The first method relies on the TextSeg linear topic segmentation [136], presented in Chapter 2, while the second method relies on the new linear segmentation algorithm, MSeg, we have proposed in Chapter 3, cast in a hierarchical setting. The goal is to compare linear and hierarchical segmentation strategies, where hierarchical methods are likely to give shorter and equally accurate targets. The evaluation for the target selection part is done within the Search and Hyperlinking task at MediaEval 2014 [40].

The rest of the chapter is organized as follows: Section 6.2 describes our hyperlinking solution, detailing the anchor and target segments selection and the link creation. Experimental results are reported in Section 6.3. Section 6.4 concludes the chapter.

6.2 Hyperlink creation

We present here our approach for creating hyperlinks tackling first the anchor selection part and continue with the target selection and hyperlink generation.

6.2.1 Anchor selection

The aim of our approach is first to find precise jump-in points to the salient segments in the videos, at various levels of details. These segments are obtained by applying the HTFF algorithm, which outputs a hierarchy of topically focused fragments for each video. HTFF relies on text-like data. Therefore, we exploit spoken data obtained from automatic transcripts and manual subtitles [52]. An example of the representation obtained for a video in the collection is given in Figure 6.1, with some keyframes found in the segments formed at the lowest level in the hierarchy. After obtaining the topically focused fragments we perform content analysis to propose the top anchor segments for each video. Thus, for every video for which anchors need to be extracted from, we compute the probabilistic cohesion measure $C(S_i)$ (Eq. 3.3) to rank the fragments in the hierarchy, where S_i is a fragment in the hierarchy. This measure will favour short and cohesive segments. Basically, we rank each of the fragments in the hierarchy from all levels using the cohesion measure. The fragments that are longer than 2 minutes are eliminated, to favor more precise fragments. Using HTFF for anchor detection does not ensure a minimum number of anchor segments to be found for a video. The hierarchy will prevail only the most important information and this results in a fixed number of fragments in total (summed over all levels). The average number of anchor segments per video obtained when exploiting subtitles is 18.9 with the confidence interval

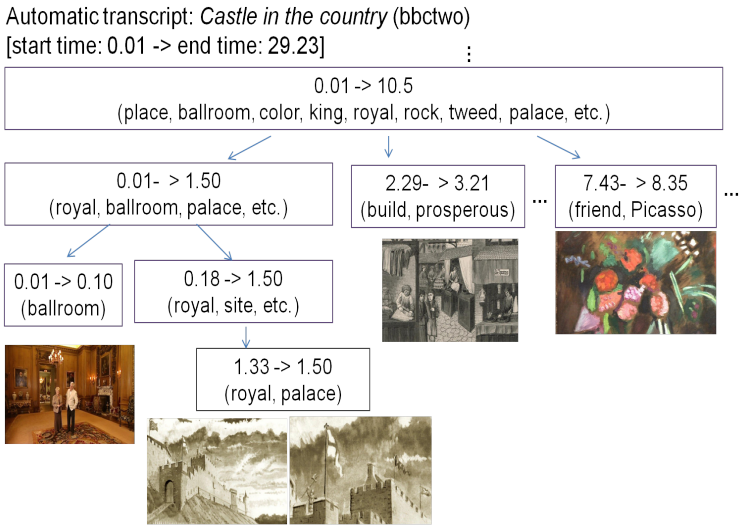


Figure 6.1: Example of hierarchy of topically focused fragment obtained for a video in the collection.

at 95% of [18.56, 19.25]. While, when automatic transcripts are used the average number of anchor segments is 19.03 with the confidence interval at 95% of [18.4, 19.65]. On subtitles there are less anchors proposed per video and this can be due to the fact that longer topically focused fragments, based on longer burst intervals, are formed when there are no erroneous words. Therefore, some videos might have more or less anchors proposed than others. This is realistic, since the number of anchors that can be found in a video depends on the salient information contained. Our focus is to propose a system that targets high precision rather than high recall. Next, we detail our approaches for the target selection step.

6.2.2 Target selection

In the absence of prior knowledge or experience on what users—human assessors in the framework of a comparative evaluation—are expecting, we posit that good fragments to be selected as targets for hyperlinks with the anchor as the source should verify the following characteristics: They should be short enough so as to be focused on a single semantic aspect; They should be semantically related to the anchor from a topic point of view; They should not be exactly redundant with the information provided by the anchor. Interestingly, the two last characteristics are conflicting, calling for a trade-off between exact repetition and related content. These three characteristics call for target selection methods which heavily rely on semantic characterization, possibly at a higher level than the mere repetition of words.

We approach the problem with the following strategy: structure the videos into topics, linearly or hierarchically, to create potential target segments and enforce the coherence of the targets. The leading idea that we pursue is to perform topic segmentation both at a general and at a specific topic level in a hierarchical manner so as to identify short and accurate target fragments. Linear topic segmentation provides a rough structure where a homogeneous segment can in fact approach various aspects (sub-topics) of a main topic. Having a more detailed organization can help provide precisely related segments of shorter length than the

ones obtained linearly. For this reason hierarchical topic segmentation was considered.

To obtain the linear segmentation we rely on the TextSeg algorithm described in [136] which is domain independent and has proven performant on speech transcripts and on segments of highly varying length. It was described in the first part of this thesis, in Chapter 3. This algorithm is known to face a problem of over-segmentation, but in our scenario, over-segmentation is an advantage since there is a constraint on the maximum length of the segments that must be retrieved, which need to be smaller than 2 minutes. The hierarchical topic segmentation is obtained by resegmenting independently each segment resulting from the linear topic segmentation. For resegmentation, our new algorithm proposed for linear topic segmentation, MSeg, in Chapter 3, which is adapted to very short segments is used. The interest of such an approach in the hyperlinking scenario is to obtain short and focused targets. We detail next, how the links are created based on the potential target segments resulting from the two structuring techniques, linear and hierarchical.

6.2.3 Hyperlinking anchors and targets

Based on speech transcripts or subtitles, hyperlinking consists in finding in a video collection fragments whose words are semantically related to words in the anchor. We used a two step approach to this end, applied independently for each of the predefined anchors, as illustrated Figure 6.2. The first step consists in retrieving a shortlist of videos semantically related to the anchor within the collection, considering the video as an atomic entity, with the goal of establishing a link between the anchor and a fragment of each of the videos in the shortlist. The second step aims at selecting the target fragment within each video of the shortlist, searching for fragments that are relatively short and relevant and that present diversity in the result. Different measures of the semantic resemblance between the anchor and topic segments are explored, offering different trade-offs between similarity and diversity.

The first step, i.e., the shortlist selection follows a classical textual information retrieval framework with a cosine distance computed between weighted vectors representing resp. the anchor and each video of the collection. Each vector is composed by nouns, adjectives and non modal verbs associated with a BM25 score [116]. The cosine distance is computed to obtain a score for each couple anchor-video and to create a list of results (ranked in decreasing order) for each anchor. As we want diversity, i.e., providing users with hyperlinks targets that cover various aspects or point of views related to the anchor, we do not consider in the ranking the video from where the anchor is extracted and possible rebroadcasted versions[1]. A shortlist of the 50 most related videos within the collection is established and further processed to find precise link targets according to different strategies discussed hereunder.

For the second step, three different system settings are considered: two settings with linear segmentation but different representations for the similarity computation and one with hierarchical segmentation. All settings rely on a similarity measure to compare the content of the anchor and the target, using either a bag of words (BoW) representation or bags of ngrams alignments. When BoW are used, a cosine similarity measure is employed. In the case of ngrams, similarity is computed between bags of unigrams, bags of bigrams and bags of trigrams separately and the scores obtained are combined with weights 0.2, 0.3

[1]Some videos in the collection correspond to the exact same program rebroadcasted later the same day or during the week.

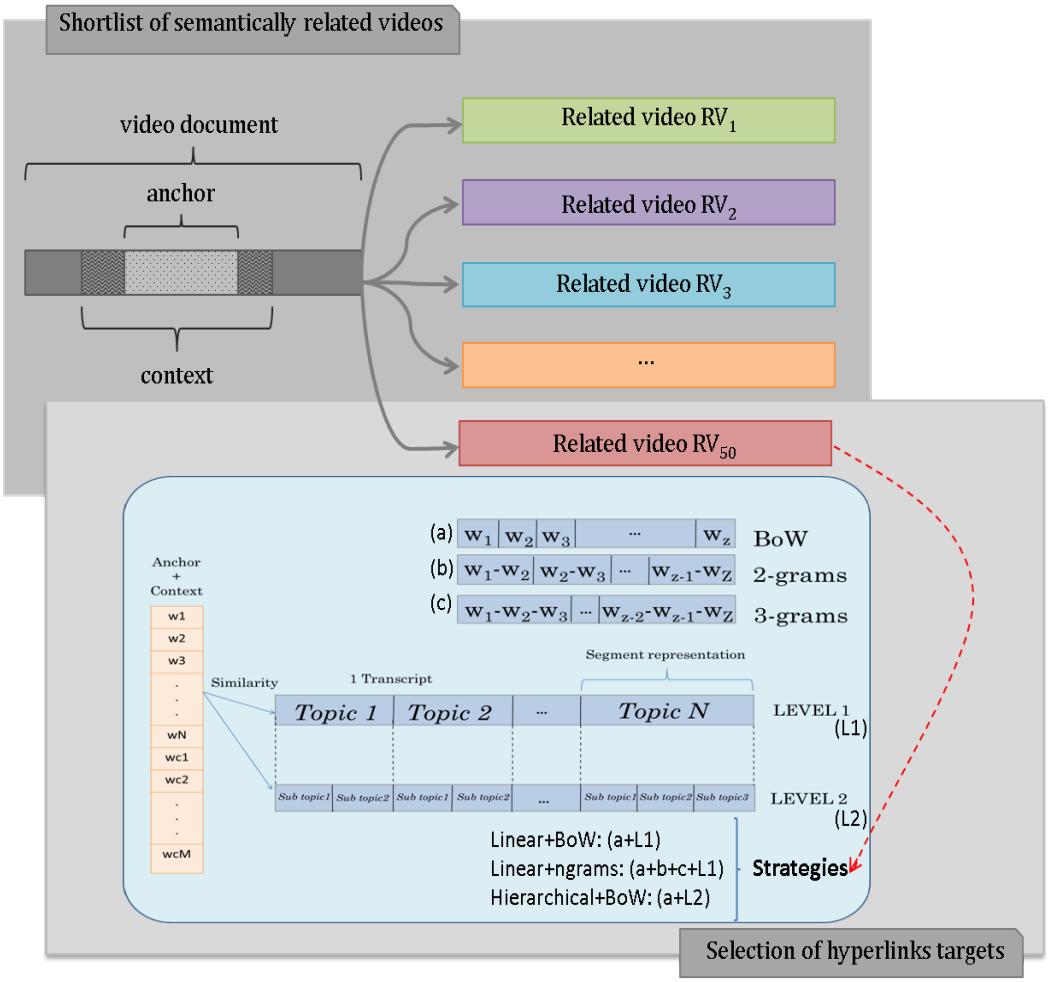


Figure 6.2: Global architecture of the two-step hyperlink generation approach.

and 0.5 respectively. The weights were chosen empirically with the idea of emphasizing precise alignments to the expense of diversity. The ngrams alignments potentially capture high similarity between the anchor and the segments while the BoW potentially allows better account for serendipity. The description of the anchor is established considering the context in which it appears, i.e., taking into accounts words and ngrams in its neighbourhood. For the linear segmentation approach the two different representations are considered for the similarity computation between the anchor and the target (i.e., BoW or ngrams). For the hierarchical approach, as segments are short, the use of ngrams is of limited interest and we limit content comparison to bag of words.

Finally, in all cases, the boundaries of the topic segment selected as the closest to the anchor according to the similarity measure chosen are refined to match length constraints imposed by the evaluation protocol:

- if the length exceeds 2 minutes, a sliding window of 2 minutes is used inside the segment to find the best matching sub-segment;
- if the length is below 10 seconds, the segment is combined with neighboring segments

(chosen based on highest similarity principle) until the new formed segment is longer than 10 seconds.

6.3 Experiments

6.3.1 Anchoring evaluation

The results obtained for all the participants at the task are given in Table 6.1. Our results are named IRISA and we were the only ones doing experiments also on automatic transcripts. As it can be observed IRISA and TUD–MMC obtain the highest precision scores. We obtain the highest score on automatic transcripts while TUD–MMC on subtitles. We observe a decrease when using manual subtitles with our approach. We believe this is due to the fact that the anchor segments obtained when using subtitles are shorter in duration. The average duration is of 11.46 seconds, with the confidence interval at 95% of [11.5, 11.86]. While, with the automatic transcripts, the anchors obtained have 22.92 seconds on average with the confidence interval at 95% of [21.48, 24.35]. Additionally, there are fewer anchors proposed when using subtitles than with automatic transcripts. We believe this is due to the fact that there are more word reoccurrences in the subtitles than in the automatic transcripts, which have also erroneous words. In terms of recall the best system is the TUD-MMC one. Our approach focuses on high precision and we do not propose more than 20 anchors per video. While, the evaluation considered the top 25 ranks for all submissions to fully evaluate, the rest of the anchors in the ranks being assigned a judgement if they were found in the top 25 of other runs. This results in a lower recall value for IRISA. The mean reciprocal rank value is lower than for other participants, which means their first relevant anchor appears higher in the rank than in our case.

6.3.2 Hyperlinking evaluation

The data set corresponding to the MediaEval 2013 evaluation was already presented in Chapter 5 along with the evaluation protocol that was implemented based on crowdsourcing. In total we propose three systems, two that use linear topic segmentation and two different representations for the content of the anchors and the targets: *Linear+ngrams* and *Linear+BoW*, and one that uses hierarchical topic segmentation and BoW representation: *Hierarchical+BoW*. For our experiments all transcripts are lemmatized with TreeTagger and only nouns, non modal verbs and adjectives are kept. Results from the evaluation are given and discussed .

Among our submitted runs, the organizers of the task have selected the *Linear+ngrams* for full evaluation of top 10 results for each anchor. This run corresponds to the linear segmentation with bags of ngrams and was performed on subtitles and automatic transcripts from LIMSI and LIUM. However, the number of targets that were judged (and thus considered) for the other runs, as reported in Table 6.2, is significantly lower than the number of targets judged for *Linear+ngrams*. For example, only one third of the hyperlinks targets of run *Hierarchical+BoW LIUM* are judged. This partial evaluation does not allow us to objectively compare the various runs that we submitted and results must be analyzed with caution. However, some conclusions, most qualitative, can still be drawn from the evaluation process performed through AMT.

system	data type	Precision@10	Recall	MRR
IRISA	Manual subtitles	0.469	0.38	0.73
IRISA	LIMSI transcripts	**0.557**	0.435	0.77
CUNI	Manual subtitles	0.31	0.27	0.83
TUD–MMC	Manual subtitles	**0.557**	**0.474**	**0.87**

Table 6.1: Precision, recall and MRR run results for all systems.

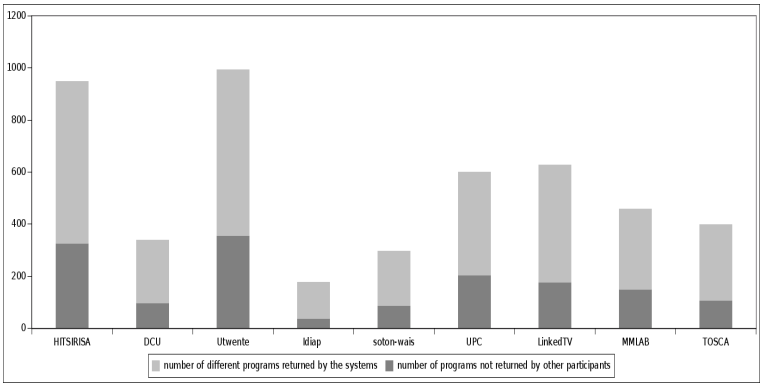


Figure 6.3: The number of videos returned for each participant's system at the Hyperlinking task.

First, in order to estimate the performance of the first step of our system, i.e., the computation of the semantically related videos shortlist, we can compare the number of different programs returned by our system with the number of different programs returned by the other participants of the task, as well as the proportion of programs that are not found in the relevant results of other participants' system[2]. In Figure 6.3, light grey rectangles represent the number of different programs returned by the systems (our system is on the left, named HITSIRISA) and the dark grey rectangles correspond to the number of videos that are not found in the relevant results of other participants. From this figure, we can see that our approach tends to give much more different programs than the majority of other participants. Moreover, our system returns more than five times more different programs than the participant that gives the lowest number of different videos (949 vs. 176) while the proportion of programs that are not found as relevant by the other participants is comparable. Therefore, we can conclude from this figure that our system gives more diversity concerning the programs the targets are extracted from. However, this does not ensure a diversity in the information contained in the targets.

Concerning the crucial target selection step, it can be seen from Table 6.2 that P_10 and P_10_bin measures do not vary much for our runs since all the hyperlinks targets in a run are extracted from different videos. Table 6.2 also shows that the P_10_tol measure, which favours a more precise identification of the starting point of targets, is a bit lower than the two other measures, which means that the beginning of our returned hyperlinks targets is usually not very precise. The run *Linear+ngrams*, applied on subtitles, has, however, almost the same value for each measure meaning that the boundaries of the hyperlinks targets returned by this approach are precise.

[2]The fact that a program is not found in the relevant results of other participants does not necessarily mean

	P_10	P_10_bin	P_10_tol	judged_10	judged_10_bin	judged_10_tol
Linear+BoW LIMSI	0.2	0.24	0.14	0.48	0.56	0.44
Linear+BoW LIUM	0.19	0.2	0.15	0.44	0.49	0.42
Linear+BoW MAN	0.31	0.31	0.25	0.58	0.61	0.53
Linear+ngrams LIMSI	0.33	0.35	0.3	1	0.98	1
Linear+ngrams LIUM	0.34	0.33	0.3	1	0.98	1
Linear+ngrams MANUAL	0.42	0.41	0.41	1	0.98	1
Hierarchical+BoW LIMSI	0.19	0.23	0.17	0.42	0.53	0.42
Hierarchical+BoW LIUM	0.16	0.18	0.14	0.37	0.48	0.36
Hierarchical+BoW MANUAL	0.26	0.28	0.26	0.49	0.59	0.49

Table 6.2: Precision values obtained using the Amazon Mechanical Turk platform. For each value an estimate of the proportion of hyperlinks that were actually evaluated is reported in columns *judged_10, judged_10_bin, judged_10_tol*.

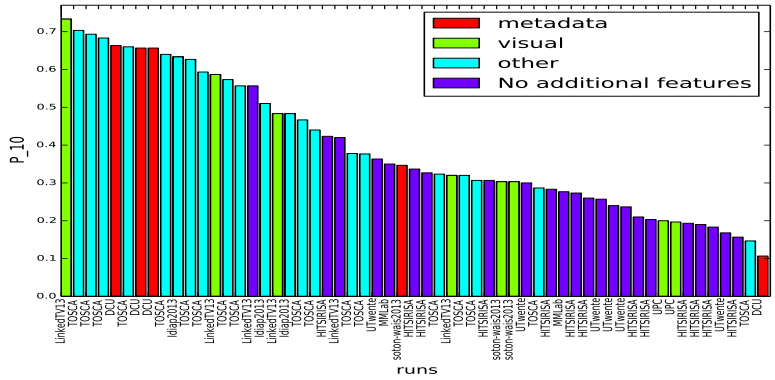
	judged											
	(overlap)				(binned)				(tolerance)			
Transcript	different	same	not	only one	different	same	not	only one	different	same	not	only one
LIMSI	1	123	154	22	1	223	68	8	1	136	143	20
LIUM	3	103	163	31	1	231	62	6	3	113	156	28
MANUAL	2	145	127	26	5	232	58	5	2	158	120	20

Table 6.3: Number of target segments, obtained with Linear+BoW and Hierarchical+BoW, that were judged differently, the same, not judged and judged only for one method. These numbers are compared for different relevance assessment of targets: overlap, binned and tolerance to irrelevance.

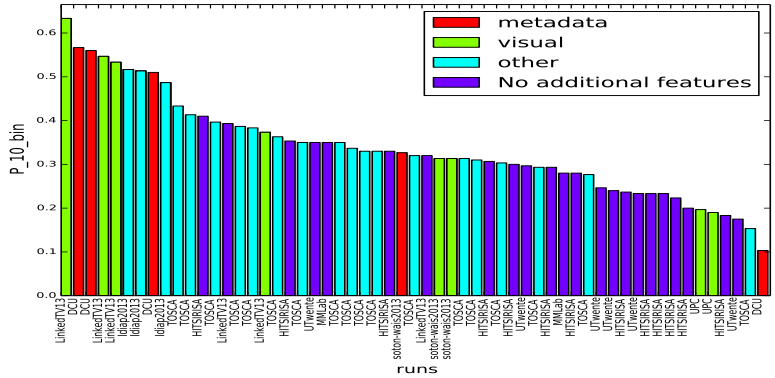
As the judgement for hierarchical topic segmentation based approaches is only partial, we compared the judged targets for the *Linear+BoW* and *Hierarchical+BoW* methods in order to evaluate the capability of providing more precise segment extraction, on all transcripts. As it can be observed in Table 3.2, the hyperlinks that were actually evaluated (ca. 50 %) in the *Linear+BoW* and *Hierarchical+ BoW* methods were thus compared to find out whether they agreed or not. With the overlap relevance, for the subtitle transcript, out of the 173 hyperlinks actually judged, 145 were found to be in agreement. Differences in the judgement were observed in 2 cases while the 26 remaining cases correspond to the situation were a hyperlink was found only by one of the methods. With the binned relevance and the tolerance to irrelevance judgements, more targets get judged. With the binned metric results get merged together if they belong to the same bin and there will be only one judgment. With the tolerance, if one segment has already been seen through another one it it will be judged as already seen. Therefore the number of not judged targets decreases, as can be observed in Table 6.3 (column *not* for binned and tolerance). Similar trends were observed on ASR transcripts. This last observation clearly indicates that hierarchical topic segmentation is efficient in selecting relevant targets which are more precise and smaller than the one obtained by linear topic segmentation.

Finally, compared to other participants, our best system *Linear+ngrams* is the second among the systems based on lexical cohesion, according to the P_10_bin and P_10_tol measures. More, as can be seen fom Figure 6.4, our approach performs the best among all the systems that do not use additional features (metadata, visual, etc.) according to P_10_bin

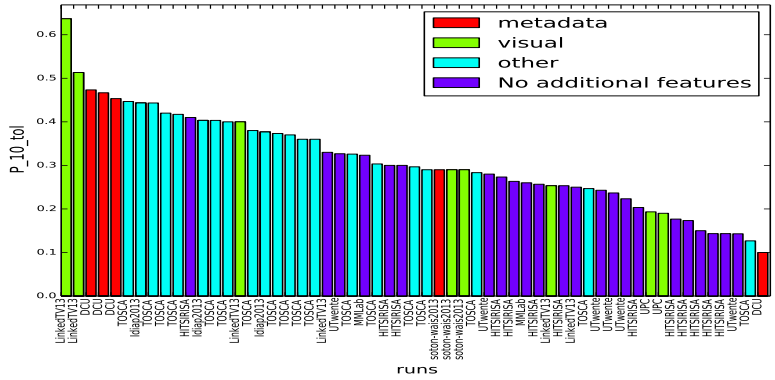
that it is irrelevant. It can also mean that the video was not returned by any other participant.



(a)



(b)



(c)

Figure 6.4: Comparison between the results proposed by all participants with P_10 (Fig 6.4(a)), P_10_bin (Fig 6.4(b)) and P_10_tol (Fig 6.4(c)). The identifier for our runs is HITSIRISA.

and P_10_tol measures, and the second one based on P_10.

6.4 Conclusion

Automatic topic structuring was approached in this chapter for anchor and target selection in video hyperlinking, exploiting language data only. The anchor selection was done using the hierarchy of topically focused fragments. The results show that we can achieve with automatic transcripts the highest precision score equivalent to what another system obtained on manual transcripts. When we use the same approach on subtitles the precision decreases. We believe this is due to the smaller number of anchor segments proposed and to the reduced duration of these anchors. A way to alleviate this is by improving the anchor relevance assessment and propose as top anchors longer segments. For the target selection we compared various strategies to obtain precise fragments to link to a given anchor. While objective comparison is difficult because of incomplete evaluations by human assessors, some conclusions can be drawn. In particular, it was shown that, on this dataset, the two step approach consisting in a preselection of relevant videos followed by fragment selection within each preselected video, offers a diversity of sources from which the targets are selected from. The comparison between linear and hierarchical topic segmentation also demonstrated that precise target selection was possible using fine-grain hierarchical topic segmentation. Finally, good results obtained with ngram comparison hint that assessors judged as relevant content very similar to the anchor, not rewarding serendipity. This was confirmed by the analysis of the whole set of results of the MediaEval 2013 benchmark. The target selection could be improved by using the hierarchy of topically focused fragments previously used for anchor selection. As we demonstrated in Chapter 4 it is easier to find salient information at various levels of detail than topic and sub-topic boundaries.

In the following chapter we will address the problem of diversity in the links. We believe that adding a control over the links, will allow us to explain why we linked two fragments, and would help in improving diversity, favouring serendipitous encounters, while maintaining link acceptability by users at high standards. For this we will propose also a new evaluation scenario that aims at capturing if the users find the links serendipitous or not.

Chapter 7

Leveraging topic models to justify links and control diversity in video hyperlinking

Contents

7.1 Hierarchical topic models for language-based video hyperlinking 93
7.2 Problem formulation: hyperlink creation 95
7.3 Leveraging topic models for indirect hyperlinks 96
7.4 Evaluation in the context of the Search and Hyperlinking task 100
7.5 Assessing serendipity and diversity in the links 109
7.6 Conclusion 119

7.1 Hierarchical topic models for language-based video hyperlinking

As mentioned in the previous chapter, most of the existing approaches for video hyperlinking are being proposed in the context of the MediaEval Search and Hyperlinking benchmark initiative. The participants' focus for video hyperlinking is set mainly on the extraction of targets and on proposing new ways to analyze content similarity using one or more modalities. However, we believe that another crucial aspect of hyperlinking is to offer diversity, favouring serendipity. Indeed, when offering diverse links, unexpected but still relevant targets are more likely to appear, than when the focus is on targets with content very similar to the anchors. Additionally, being able to explain why two video segments are linked can improve the acceptance of serendipitous targets by the users of a hyperlinking system.

Thus, our approach is to investigate how we can control the diversity in the links and to add a characterization of the links, i.e., justify the links created, while maintaining users link acceptability at high standards. For controlling diversity, we would like to be able to choose which kind of topical relation generates the selection of a certain target. Having a topical diversity in the links, for the characterization of the links we could rely on the topical connection between an anchor and a target. This means that we would be able to say that a target is on the same topic *X* as the target or on a different point of view on the topic *X*, etc.

As a potential solution to these aspects, we investigate transcript-based indirect content comparison mediated via a hierarchical topical structure. The key idea is to have a fine-grain control on the topics that are highlighted in the targets proposed for a given anchor. The topical structure is composed of topics at different levels of granularity, from general to specific, generated by iterative application of the latent Dirichlet allocation (LDA) model. A first advantage of this structure over the direct bag-of-words representation is the ability to link related anchor-target pairs that do not share a consistent part of vocabulary. Additionally, the hierarchical topic structure of the model, linking fine-grain (e.g., election results in France) topics with coarse-grain themes (e.g., politics), allows for an increased control over diversity and has the potential to explain the nature of the link (i.e., why is this linked proposed?).

In this chapter, we disregard the problem of extracting target segments to focus on the comparison of existing anchor-target pairs, proposed by MediaEval participants in 2013 and 2014, with the proposed topical structure. The first contribution of this work is the new indirect structure used to connect anchors and targets. Using this structure brings along several advantages which we capitalize on. We will look at the problem as a target ranking task and show that with the new approach we achieve link precision comparable to direct text comparison and provide improved capabilities for serendipity and link justification. For this, we rely on the relevance judgements done within the task evaluation. When judging the anchor-target pairs in the context of the MediaEval Search and Hyperlinking benchmark initiative via AMT crowdsourcing, the turkers have a description of what the targets should be about. This description is given by the users that defined the anchors in the videos and correspond to some information needs that should guide turkers for the relevance assessment. However, being relevant is not directly comparable to being serendipitous. Thus, the unexpectedness part of serendipity is not accounted for. The second contribution is the definition of a new evaluation scenario that aims at capturing the important aspect of diversity and serendipity in the links. We thus evaluate our approach also in the context of this new scenario. We believe that the work done opens new perspectives and opportunities for video hyperlinking, placing more focus on user needs.

The organization of the rest of this chapter is as follows: Section 7.2 formulates the problem addressed in this work regarding the control of serendipity and link justification in the context of link generation. The creation of the topical structure that offers the basis to answer this problem is detailed in Section 7.3. Section 7.4 reports the results together with their analysis using the relevance assessment scenario form MediaEval. Section 7.5 introduces the new evaluation scenario and reports the results obtained employing it. Section 7.6 concludes the chapter.

7.2 Problem formulation: hyperlink creation

In this section, the problem that we address goes beyond finding anchor-targets pairs on similar topics as in a traditional information seeking scenario. Instead, we want to have diversity in the links proposed hinting to a serendipitous approach, allowing users to acquire information they never searched for and for which they might not have had predilection. Therefore, we investigate how hyperlinks can be created in order to reveal hidden connections (or "hidden analogies" [43]) between video fragments, while controlling the basis of the connections, i.e., control the diversity. Additionally, we want to understand why an anchor is connected to a particular target, i.e., understanding the hidden connections, in order to understand the users need and judgments. With this purpose in mind the starting point is the analysis of the types of hyperlinks that can be created and how can they offer the basis to understand and control the linking.

7.2.1 Direct hyperlinks

Direct hyperlinks refer to the links created when the content of the anchor is compared directly with the content of the target. A vectorial representation of the content can be used and the similarity can be computed using the cosine similarity measure as in [62]. Segments with similar content will have a large cosine value, meaning that the angle between the vectors representing the segments is close to zero. This kind of links, based on bag of words, does not favour diversity but rather targets with very similar content. In the evaluation of the systems that we propose we will use this kind of links as a baseline.

7.2.2 Indirect hyperlinks

There exist segments that speak about the same things, or related ones, without sharing much vocabulary. One way to link such segments is to change the representation space and use an intermediate structure to compute their similarity. One technique to do so is vectorization, introduced in [30] in a standard information retrieval scenario. The idea is that instead of directly comparing two documents d_1 and d_2, first, each document is compared with the same m pivot or, reference, documents using a proximity score ∇_i^j, e.g., cosine, as: $\nabla_j^i = cos(d_i, PD_j)$, for each document d_i with pivot documents PD_j, $j = 1 \rightarrow m$; For each document, the m scores obtained are gathered to form a vector representing the document; Thus document d_1 will be represented as $\nabla_1^1, \nabla_2^1, ..., \nabla_m^1$; The comparison between two documents is indirectly performed by comparing their associated vectors using, e.g., the Euclidian L2 distance as: $L2(d_1, d_2) = \sqrt{\sum_{t=1}^{m}(\nabla_t^1 - \nabla_t^2)^2}$. The most interesting property of this technique is that two documents with limited common vocabulary can be deemed similar if they are similar to the same pivot documents.

Building on the idea of vectorization in video hyperlinking, two video segments can be compared indirectly by measuring how close their respective decompositions on some intermediate structure—the space of pivot documents in the case of standard vectorization—are. This approach sounds appealing in our case for several reasons. First, our goal is to be able to find information that would not be found at search time, information that the user is not specifically looking for: Such links cannot be discovered just by searching for very

similar content and hence by directly comparing content. Second, we want to be able to have control over how similar the target is with the anchor, in particular in terms of topics, and to explain the nature of the relation between the anchor and target segments. Our interest is the topical relation since we believe it can offer a good characterization of a link and also because in [103] the authors noticed that users favoured semantic links as opposed to similar visual content. These last two requirements call for an intermediate structure which should reflect a topical organization, to change the representation space of the video segments into a semantically organized space. Such an intermediate structure enables the identification of the topics that can explain the relation between two segments. More, creating relations between the topics, such as meronymy, hyperonymy, can help diversify the links created.

7.3 Leveraging topic models for indirect hyperlinks

Instead of using pivot documents as in [30], we decompose documents in topics at multiple levels of granularity, from general to specific topics, possibly considering a hierarchical organization between adjacent levels, i.e., topic-subtopics relations. Each topic is characterized by a probability distribution function over the set of words. Using this topical decomposition, similarity between two segments is performed by analyzing the distributions of words in the segments given the topics, i.e., by computing the probability of the words in the segments given the word distributions of the topics. For each comparison between the segments, different topics can be accounted for, with a certain weight associated, enabling serendipitous links. For example, a connection between an anchor and a target can be justified by one or more common general topics (e.g., animals) or the same specific topics (e.g., tigers eating in the wild), or related topics (e.g., elephants in the wild), etc. An important aspect when building the topical structure is to make it informative so that each link can be justified by the content of the structure that contributed to its creation. We now discuss how to build a hierarchy of topics in a data-driven manner before detailing different strategies to compare segments using the topical structure constructed.

7.3.1 Building the topical structure

The topical structure is built using latent Dirichlet allocation probabilistic topic models [14] learned on the transcripts of a collection of videos. In this model, each transcript is represented as a mixture of K latent topics, where each latent topic is characterized by a probability distribution over the set of words in the transcript (the vocabulary). The advantage of this representation is that, contrary to bag of words, co-occurring terms with semantic similarity are clustered. LDA models were estimated using Gibbs sampling with standard values for the hyperparameters $\alpha = 50/K$ and $\beta = 0.01$ [132]. To define the various levels of the topic hierarchy, we trained model for different numbers of latent topics, namely $K \in \{50, 100, 150, 200, 300, 500, 700, 1000, 1500, 1700\}$. This range of values for the number of topics was chosen to obtain topics that go from being general to highly specific and to have a large number of granularity levels for a better control of link creation. The hierarchical Pachinko topic model (PTM) [85] was not considered suitable for building the topical structure since it produces only a four-level hierarchy consisting of a root, a set of super-topics, a set of sub-topics and the set of words in the vocabulary. This translates into a 2-levels hierarchy of

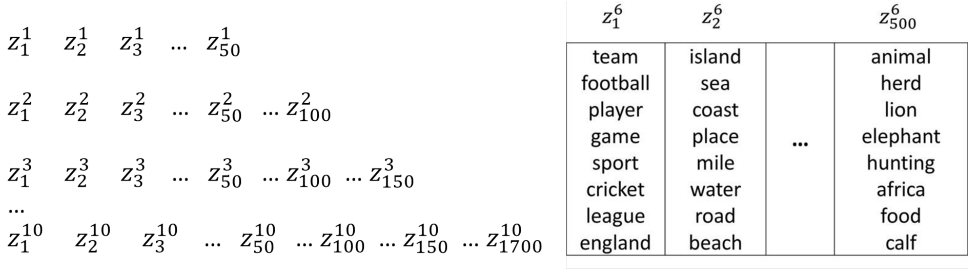


Figure 7.1: Representation of the topical structure obtained from topic models. On the left side is the representation of the structure for $K = 50 \rightarrow 1700$. z_j^l corresponds to topic j, $j = 1 \rightarrow K$ at level l. On the right side is an example of topics learned with $K = 500$, i.e., level 6.

topics. We consider necessary for the purpose of this work to have more levels in the topics, since these levels will be the key point for controlling serendipity.

At each level l, a word distribution z_i^l is obtained for each topic $i \in [1, K_l]$, where K_l is the number of latent variables at level l ($K_1 = 50, \ldots, K_{10} = 1700$). As all vocabulary words belong to each topic, regardless of the level, the difference is that words have a different probability for each topic. Therefore a word that has a lower probability in a more general topic can have a higher probability in a more specific topic. The result of the process of building LDA topic models at 10 levels is illustrated in Figure 7.1, where the most likely words for some topics obtained with $K_6 = 500$ are given on the right side. Clearly, the first topic is about sport while the second one is about the sea.

As mentioned in Section 7.2.2, the topical structure is used to change the representation space of the anchor and target segments. Thus, given a segment x, which can represent either an anchor or a target, the word distribution z_i^l for the i-th topic at level l enables the computation of the probability that x was obtained from z_i^l according to

$$p(x|z_i^l) = \sqrt[n_x]{\prod_{j=1}^{n_x} p(w_j|z_i^l)} \ , \tag{7.1}$$

where n_x is the size of the vocabulary in x and w_j is the j-th word in x. The word probabilities are given by

$$p(w_j|z_i^l) = \frac{n(z_i^l, w_j) + \beta}{\sum_{k=1}^{n} n(z_i^l, w_k) + \beta|V|} \ . \tag{7.2}$$

These probabilities are estimated on the entire collection, with $n(z_i^l, w_j)$ being the number of times topic z_i^l was assigned to word w_j occurring at a certain position in the training documents. The denominator thus corresponds to the total number of words assigned to topic z_i^l. V represents the number of distinct words in the entire vocabulary and β is the Dirichlet prior.

This topical structure, i.e., the set of values $p(x|z_j^l)$ for all the topics, is the basis for the comparison between anchors and targets. Three variations are proposed, with the goal of investigating how the topics should be accounted for. The goal is to structure the different

levels in a hierarchical manner. Thus, different relations will be considered between the topics in the structure to compare and anchor with a target segment.

7.3.2 Independent topic levels (IT)

The simplest structure considers the topics obtained with different K values as independent, meaning there is no relation between the topics at different levels. In this case, for each K and for each anchor-target pair, two vectors are obtained having at each position the probability of topic z_j^l given the words contained in the segment. This probability is an approximation of the posterior, considering uniform distribution of the topics in the document collection. This assumption is realistic in a typical setting where we do not know if some latent topics are specific to that collection. Thus, the new representation of a segment is given at level l by the vector gathering topic-wise probabilities of x, i.e.,

$$x_l = (p(x|z_1^l), p(x|z_2^l), ..., p(x|z_{K_l}^l)) \ . \tag{7.3}$$

For efficiency reasons, we use a sparse version of x_l, zeroing all but the 10 top-scoring topics. The discarded probability mass is redistributed evenly on the top-10 topics.

Comparing two segments x and y is done via the respective representations x_l and y_l according to

$$S_1(x,y) = -\sum_l \alpha_l \log\left(x_l' y_l'\right) \ . \tag{7.4}$$

We use the logarithm of the dot product as a distance computation between the two vectors representing an anchor and a target. The weights α_l allows to control the relative weights of the topic levels, for instance, to select one single level or to emphasize fine-grain levels over general topics. We compare three weighting variants: equal importance to all topics ($\text{IT}_{\text{Comb=}}$), increasing importance ($\text{IT}_{\text{Comb<}}$) as going from general topics to specific ones and conversely ($\text{IT}_{\text{Comb>}}$).

7.3.3 Hard links between topics (HLT)

Exploiting explicit links between topics at different levels of theprobability of topic z_j^l hierarchy—e.g., meronymy, hyperonymy—appears as appealing for a better control of the diversity of the targets and of the relation between anchor and target. We thus propose two strategies to turn the independent 10 levels of LDA models into a tree structure. A straightforward way to build a tree structure exploits the similarity between topics at two consecutive levels, where the similarity between topic i at level l and topic j at $l+1$ is given by $-\log\left(z_i^l z_j^{l+1}\right)$. The tree is obtained by connecting a topic to the most similar topic at the previous level. Formally, z_j^{l+1} is linked to z_k^l such that $k = \text{argmin}_i \log\left(z_i^l z_j^{l+1}\right)$. We call such links 'hard' links, meaning that every node has a unique parent (except at $l = 1$) but not necessarily a sibling or a child.

A generic representation of such a hierarchy is given in Figure 7.2. By construction, nodes at the lowest level will all have a path to a node at the first level, but not the contrary. Figure 7.3 depicts an example of two topics, one from level 7 and one from level 1, connected by a path in the topical structure. The topics are represented by their top-words.

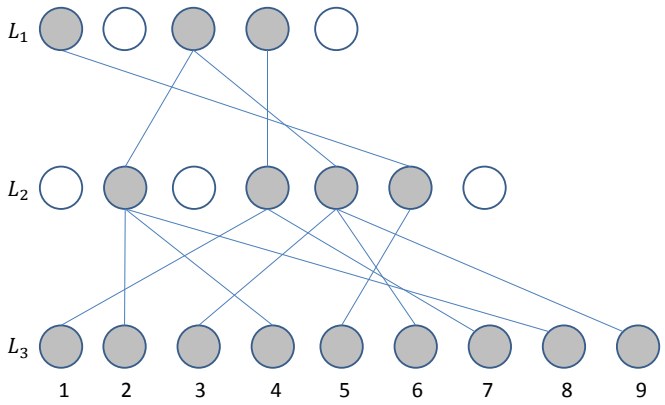


Figure 7.2: Generic representation of a hierarchy of 3 levels. L_i denotes the level in the hierarchy. The circles represent nodes and an arc is formed when there is a parent-child relation between two nodes. A node that is the parent or the child of another node is colored.

Specific topic	General topic
referendum	people
minister	government
scotland	tax
independence	minister
alexander	party

Figure 7.3: Representation of two topics, one general and another that is more specific, connected in the tree hierarchy. The 5 top-words (i.e., words that have the highest probability in the topic) for both topics are given. The specific topic is chosen from level 7 and the general one from level 1.

The 'hard link' tree-structured (HLT) hierarchy of topics is used to define a new representation of an anchor x depicting the path in the tree that ends at $l = 10$ with the best matching fine-grain topic. For an anchor segment x, we first identify the best matching topic at the lowest level, i.e., $k = \arg\max_j p(x|z_j^{10})$. By construction of the tree structure, this node has a unique parent and we follow the path from z_k^{10} to the first level in the tree. This path corresponds to a sequence of topics $\mathbf{t}^x = \{t_1^x, \ldots, t_{10}^x\}$, where $t_{10}^x = z_k^{10}$, and $t_l^x = z^l_{\text{parent}(t_{l+1}^x)}$ for $l = 9$ to 1. Given $\mathbf{t}^x$, the similarity between a target segment y and the anchor x is defined as

$$S_2(x, y) = \sum_{l=1}^{10} \alpha_l p(y|t_l^x) \ . \tag{7.5}$$

The interest of using a path to change the representation of the segments is that it creates a connection with the anchor at different levels of topics from one perspective and allows a control over the importance of each topic in the path with respect to its sub-topics (fine-grained) or super-topics (coarse-grained). Such a path will help justify the relation between the anchors and targets, considering the relations between the topics and not only the top-words of the topic that influenced most the link creation.

7.3.4 Hard & soft links between topics (HSLT)

The 'hard link' tree structure is rather simple and, by construction, some nodes might be unreachable from the lower level. We thus propose another tree construction algorithm where we enforce a more complex (and balanced) structure where each node have at least two children. The resulting tree-structure guarantees that no topic will be left aside, and allows the use of richer relations between nodes. Integer linear programming (ILP) is employed to obtain an optimal structure[1], maximizing the weight of the links created. More formally, for link creation between levels l and $l+1$, the ILP optimization consists of maximizing

$$\sum_{i \in [1,K_l], j \in [1,K_{l+1}]} \text{sim}(i,j)\ \text{link}(i,j) \tag{7.6}$$

subject to

$$\sum_{i \in [1,K_l]} \text{link}(i,j) = 1 \quad \forall j \in [1,K_{l+1}] \tag{7.7}$$

and

$$\sum_{j \in [1,K_{l+1}]} \text{link}(i,j) \geq 2 \quad \forall i \in [1,K_l]\ , \tag{7.8}$$

where $\text{link}(i,j) = 1, i \in [1,K_l]$ and $j \in [1,K_{l+1}]$, if a link is created between topic i at level l and topic j at level $l+1$, 0 otherwise, and where $\text{sim}(i,j)$ is the cosine similarity between the two topics. : Because every topic is represented as a distribution over the words in the vocabulary, the similarity between two topics corresponds to a simple cosine between their sets of words, where each word is weighted by the probability in the respective topic. Eq. 7.7 ensures that every node has only one parent while Eq.7.8 ensures that each parent has at least two children. At hyperlinking time, the ILP tree-structure is used as the HLT one to generate a path from the best matching node at the lowest level to the coarsest level.

7.4 Evaluation in the context of the Search and Hyperlinking task

For this evaluation we rely on the results obtained by the systems participating at the MediaEval 2013 and 2014 campaigns. Systems not only focused on textual data but also on visual content for some of them, possibly enriching the data representation with prosodic information, metadata, context for the anchors, or named entities. As a result, a wide variety of targets were proposed, with links established from multiple modalities and cues. In our experiments[2], we leverage all the targets that were proposed, regardless of the system. Each anchor-target pair proposed by these systems was evaluated via crowdsourcing on Amazon Mechanical Turk, thus enabling to divide targets into relevant and non relevant ones (according to turkers) for each anchor. Target ranking is thus evaluated using precision-based metrics.

Several strategies are considered for evaluating the three link authoring approaches proposed (IT, HLT, HSLT). The aim of these strategies is to show that we can have an insight on how to control the diversity in the links, by giving more or less weight to general or

[1] We used https://www.gnu.org/software/glpk as solver

[2] The experiments were done in collaboration with Rémi Bois, PhD student in LinkMedia team at IRISA, within the framework of the project "Linking the Media in Acceptable Hypergraphs".

	2013			2014		
method	P_10	P_10_bin	P_10_tol	P_10	P_10_bin	P_10_tol
DirectH	0.61	0.51	0.25	0.41	0.33	0.19
IT_{50}	0.65	0.63*	0.44*	0.26	0.23	0.18
IT_{150}	0.57	0.51	0.34*	0.37	0.34	0.25*
IT_{300}	0.61	0.54	0.35*	0.34	0.3	0.26*
IT_{700}	0.64	0.53	0.34*	0.31	0.28	0.21
IT_{1500}	0.59	0.54	0.32*	0.32	0.32	0.24
$IT_{Comb=}$	0.66	0.58	0.35*	0.27	0.33	0.22
$IT_{Comb<}$	0.67	0.57	0.37*	0.27	0.33	0.21
$IT_{Comb>}$	0.65	0.57	0.35*	0.29	0.35	0.22

Table 7.1: Comparison between direct and indirect hyperlinking. Precision values are reported for each method, on each data set. DirectH denotes the method that creates direct links between the anchors and the targets. The indirect hyperlinking is tested using the IT either by employing topics from one level at a time or by combining topics from different levels. Statistical significant values (paired t-test, $p<0.05$) obtained when IT-based methods are compared with DirectH, are marked with *.

specific topics. Also the links created can be justified by checking the top-words of the topics that contributed the most. Additionally, the cosine similarities between the anchors and the relevant targets proposed with our systems are computed. These cosine values will be compared to those obtained when the targets are actually proposed using direct cosine similarity. The goal is to show that topic models bring forward relevant segments that word-level comparisons would not.

7.4.1 Comparing direct and indirect links

The first experiment consists in comparing direct with indirect hyperlinking. To create direct hyperlinks, we use the cosine measure to compute the similarities between the anchor-target pairs and then rank the targets for each anchor based on the values obtained. For the indirect hyperlinking, variations of the IT system were employed. These variations lie in how the topics from different levels (i.e., learned with different values of K) are accounted for: either considering only topics from one level or combining topics from different levels. The goal of this comparison is that we want to show that similar results can be obtained with both methods.

Results are given in Table 7.1 for the 2013 and 2014 datasets, reporting only results for the most representative values of K. The method creating direct hyperlinks between the anchors and the targets via cosine similarity is denoted DirectH. For indirect hyperlinking, we provide the results obtained both when topics from a certain level are used (IT_K) and when a combination of topics from different levels is used. The combination of topics is done by giving different weights α_l to the similarity (7.4) computed at different levels. Three weighting variations are considered: equal importance to all topics ($IT_{Comb=}$), using $S1(x, y)$ for the anchor target comparison; increasing importance ($IT_{Comb<}$ with weights 0.1, 0.15, 0.2, 0.25, 0.3) from coarse to fine grain levels and vice-versa ($IT_{Comb>}$ with weights 0.3, 0.25, 0.2, 0.15, 0.1). The weights were chosen empirically. The purpose of providing all these variations is to show the flexibility of the method in changing the way links are created.

data-set	average target duration	confidence interval 95 %
2013	83.38 sec.	[82.58, 84.18]
2014	58.85 sec.	[58.12, 59.58]

Table 7.2: Average target segment duration with the corresponding 95 % confidence intervals, on both data sets.

As it can be observed there is a considerable drop in precision from the 2013 data set to the 2014 one, justified by the fact that the task was more challenging in 2014. We believe it was challenging for several reasons: anchors were shorter in duration on the 2014 data set compared to 2013, which made it difficult for the systems proposed to find many relevant targets, the proportion of relevant vs. not relevant targets proposed for the 2014 data set being smaller than the same proportion on the 2013 data set (0.18 and 0.42 respectively). Looking at the duration of the target segments, reported in Table 7.2, we also observe a decrease in length from the 2013 data set compared to the 2014 data set. In 2014, the anchors had no context information (i.e., video content surrounding the anchor segment), which proved to help in 2013, having the best results obtained with systems accounting also for the context of the anchors. All these different characteristics for the 2014 data set compared to the 2013 data set are detrimental for direct content comparison and benefits topic-based matching.

We observed that there is a statistically significant[3] increase for the P_10_tol measure with indirect hyperlinking compared to the direct hyperlinking. This means that with indirect hyperlinking we rank higher targets with precise starting points and do not propose targets that have already been seen through another target. The fact that with the other measures (except for the P_10_bin on the 2013 data set when comparing IT_50 with DirectH) there is no statistically significant increase compared to direct hyperlinking is not discouraging. As mentioned before, the purpose of this technique is to provide a basis for justification and control over the links proposed. Indeed, using the topical structure, we can observe and analyze also what users did not find relevant by looking at the topics that connected the anchors to those targets. Given the results in Table 7.1, there is no clear trend to observe between the precision values obtained with general or specific topics. This means both general and specific topics can help find relevant targets.

As it can be observed from the outcome of this experiment using the topical structure offers an alternative for creating the links and offers a basis to understand them. Therefore, we considered exploring new ways of employing the topical structure, respectively using the two systems described in Section 7.2.2: HLT and HSLT.

7.4.2 Hierarchical topical structures

We follow the exact same experimental conditions as in the previous section to compare indirect linking strategies based on a topical tree structure with direct hyperlinking, i.e., DirectH. We do not reiterate the results obtained with DirectH, which are given in Table 7.1. Results are reported in Table 7.3 where two different paths are considered for the HLT system: HLT_1 contains topics learned with $K = 50, 10, 300, 700, 1500$ and HLT_2 with $K = 50, 150, 300, 700$. The second path is considered to enable the comparison between HLT and HSLT as con-

[3]We perform also significance test at level of $p<0.05$ (paired t-test).

	2013			2014		
method	P_10	P_10_bin	P_10_tol	P_10	P_10_bin	P_10_tol
HLT_1	0.37	0.39	0.3	0.33	0.32	0.25
HLT_2	0.41	0.42	0.33	0.29	0.31	0.23
$HSLT_2$	0.39	0.38	0.32	0.32	0.33	0.23
$HSLT_S$	0.41	0.4	0.32	0.29	0.29	0.22

Table 7.3: Comparison between the two ways of creating the hierarchy of the topical structure. Precision-based values are reported on each data set. HLT_1 corresponds to the path of topics learned with $K = 50, 150, 300, 700, 1500$, while HLT_2, $HSLT_2$ and $HSLT_S$ with $K = 50, 150, 300, 700$. $HSLT_S$ selects a sibling at the most specific level.

structing HSLT is computationally intensive when there is a large number of topics to consider: We therefore limited the hierarchy to four levels.

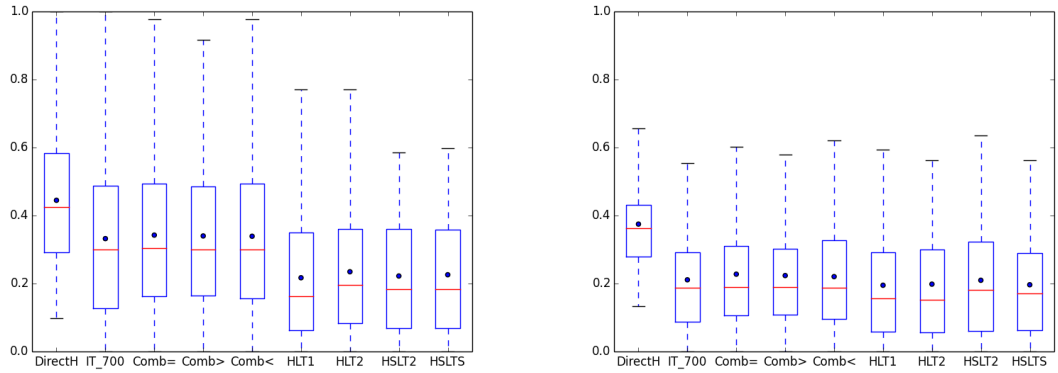
The results for all precision-oriented measures, obtained with the hierarchy, are comparable to those obtained with direct hyperlinking on the 2014 data set. However on the 2013 one the results with the hierarchy are lower than with DirectH, except for P_10_tol values which are comparable. We believe that the P_10_tol is the most representative measure for the capability of the systems, since it follows the real behaviour of users that would give up after watching a certain amount of non-relevant content. The starting point of the segment being more relevant than the end point.

The advantage of using the hierarchy is that it gives more insight than the IT on the topics that contribute to the creation of the links. The variations proposed to account for the hierarchy have the purpose to connect anchors and targets through links that are topically motivated by parent-child or sibling relations. Links can be created between anchors and targets having the same specific or general topic (parent or child relation), or a different aspect on the same topic (sibling relation). There is no significant difference in terms of precision between the variations of the hierarchy model.

7.4.3 Analysis of the links

7.4.3.1 Comparing methods

One advantage of using the topical structure is that it helps to create links between segments that would not be identified by direct content comparisons, i.e., bring forward anchor-target pairs that do not share much vocabulary. To support this claim, we studied the distribution of the cosine similarity between an anchor and the top 20 relevant targets proposed by the various methods. Figure 7.4 reports results obtained on the two data sets. Box and whisker plots graphically depict the distribution of the cosine similarity measures that can be attained, plotting the median value, the mean value, the first and third quartile and the extrema. As the topic structure gets more complex, from independent topics to tree-structures, the median cosine similarity between anchor and targets gets lower, particularly on the 2013 data. This fact highlights the potential interest of topic-based hyperlinking to provide links between segments that share little vocabulary and potentially exhibit serendipity. Also, there exist high differences between the anchor-target pairs proposed by the systems. For brevity, in Table 7.4, we provide only some of the comparisons between the systems, in terms of



(a) (b)

Figure 7.4: Boxplots showing cosine similarity measure variation between the anchor-target pairs judged relevant. Each boxplot corresponds to one of the systems proposed in this work. Figure 7.4(a) presents the results obtained on the 2013 data set and Figure 7.4(b) those obtained on the 2014 data set.

		% difference	
System 1	System 2	2013	2014
IT_{700}	DirectH	93	86
IT_{700}	$IT_{Comb>}$	82	90
IT_{700}	HLT_2	98	93
HLT_2	$HSLT_S$	29	43
$IT_{Comb=}$	$HSLT_2$	94	95

Table 7.4: Percentage of anchor/target pairs proposed and that differ between two runs.

shared anchor-target pairs. For this comparison we looked at the top 20 relevant targets obtained by the systems. While all systems exhibit comparable precision scores, the pairwise comparison shows that a large proportion of the links proposed differs between two systems. As expected, between the two types of hierarchies, the differences are smaller, since part of the path selected for the anchor is sometimes the same. This proves again that the different strategies proposed here are complementary and hints that all those techniques can be leveraged to propose a wider variety of links than those offered by direct content comparison.

7.4.3.2 Towards explaining links

We have mentioned throughout this chapter that having a meaningful structure can help understand the links created and justify them. In what follows, we provide an example of a link created with $HSLT_S$ and show how it can be explained using the topical structure. This link was not proposed with the DirectH method, having a low cosine similarity score of 0.062, between the anchor and the target. In Figure 7.5, the content of the anchor and of two targets proposed with DirectH and $HSLT_S$ is given. For this anchor the users stated: *I want to see more videos on old castles in the UK*. The top 40 targets proposed with DirectH for the chosen

anchor are from a re-broadcast of the show from which the anchor was extracted from. Out of these 40 targets, 39 are overlapping with the anchor segment and 2 were judged as not being relevant by the same turker. The turkers are supposed to motivate their judgement. For these two targets judged not relevant the motivation was the following: *The person wishes to see more videos about old castles in the UK. So, he will not be satisfied watching the second video clip*. For the rest of the targets that overlap with the anchor segment and were judged relevant, we can find motivations given by turkers such as: *These videos are nearly the same*, *Both clips are from the same show and second clip mentions a bit about castles*, *It completes the thoughts of the first clip and shows interesting images of castles. It also has more content*. Note that the first clip refers to the anchor segment and the second clip to the target segment. In the example given in Figure 7.5, the target for DirectH is the one ranked highest without overlapping with the anchor. It was ranked 17, and is the video segment immediately continuing the anchor segment. The motivation given for this target when judged as relevant by a turker was: *Both are castle related videos same show*. For the target example given for the HSLT$_S$ system we chose the first ranked target. This is not from the same show as the anchor segment.

As it can be observed, the anchor segment speaks about the symbolism behind castles and what they stand for. Meanwhile, the target segment proposed by HSLT$_S$ discusses the circumstances around the construction of another English Castle (e.g., its builder, the links with the king).

Anchor	
William the Conqueror. He parcelled out the country to the leading families who had fought for him. To control their enormous estates, they built the first stone castles in England. They were the power bases of the second order of society, the military aristocracy. The medieval world was studded with castles, hundreds of them. "The bones of the kingdom", as one contemporary called them. They were built to be high, to act as giant watchtowers over the surrounding countryside. To see, and to be seen.	
Target (DirectH)	**Target (HSLT$_S$)**
A stone castle like this would be the biggest, most expensive and most threatening building you'd be likely to see in your life. It was a symbol of the power of the aristocracy, the centre of their great estates and the foundation of their military might.	I'm on my way to the site of the biggest castle in England. It must also rank as one of the very oddest in the whole of medieval Britain. It was built around 1313 by a colorful character called Thomas of Lancaster. In his day, Thomas was talked about even more than his cousin, who happened to be none other than the King of England, Edward II. Thomas fell out spectacularly with the king when he murdered one of Edward's closest friends. It was then that Thomas built this Dunstanburgh Castle.

Figure 7.5: Anchor and targets links example that were proposed with DirectH and HSLT$_s$ and were judged as relevant.

With the HSLT$_S$ system, the link between the two segments was created based on a path

in the hierarchy that starts from the sibling of the topic that maximizes the probability of the words in the anchor given its distribution over the words in the vocabulary. Table 7.5 shows the top-words for three of the topics that contributed in creating the link. These topics are: the best specific topic, the sibling topic and the general topic. The best specific topic is the topic that described best the anchor among the topics learned with $K = 700$. Looking at the distributions over the words, the specific topic can be presented as the *countryside wonders*. Looking a bit further in the topic than the top-words in the table, there are words like: dream, magic, etc. The best sibling for this topic is about *constructions*. Again, if we look at more than just the top 10 words, we find words like: tower, pantheon, kingdom, etc. Meanwhile, the general topic in the path starting from the sibling topic is about the *history of Britain*. To sum up, the link can be justified by the fact that the target shows a different aspect on the subject of castles in Brittany throughout history. If the anchor is more about what the castles represent as a marvel, the target speaks about the construction of another English castle.

Best specific topic	Sibling topic	General topic
city	great	people
people	city	world
place	empire	war
good	roman	city
countryside	world	british
heart	christian	britain
centre	building	life
nation	living	great
visit	light	work
capital	modern	history

Table 7.5: The 10 top-words for 3 of the topics that contributed at linking an anchor and a target in the data set from 2013.

It was interesting to note that target segments from the same video as the anchor, in the example given above, were not necessarily found relevant, even though they were on the same subject. Some turkers respect literally the description given by the user defining the anchor as to what this user would want to see. Other turkers consider that targets that do not speak/show about other castle in UK, but are rather on related topics, are still relevant. Figure 7.6 gives some examples of motivations and judgements of turkers given for targets proposed from the same video as the anchor presented in the example above. First, it can be observed that while some turkers find targets not relevant if they are from the same show, others find relevant targets even if they overlap with the anchor. Second, some turkers appreciate as relevant targets that bring more information related to the subject discussed in the anchor, while others are more strict and if the target does not provide meaningful information about castles is judged not relevant.

This discussion points out some minuses in the evaluation done within the Search and Hyperlinking task at MediaEval. First, having only one judgement per anchor-target pair, makes the evaluation biased. Links that are motivated by the same connection might be judged differently based on how strict the turkers are and how they interpret the task. A way to overcome this is by having more judgements per anchor-target pair. This is rather expensive and doing it at large scale (as would be necessary in the context of the task) seems not feasible for the moment. However, we believe that it would be better to have more judge-

ments per link than judging more links. At least, this way, the links judged are more likely to be fairly judged. Second, the diversity in the links is not captured. Usually target very similar to the anchors get judged relevant rather than targets on related topics. Thus a system that aims for diverse targets is disfavoured compared with one that proposes very similar targets. More, serendipity is not evaluated, while it represents a scope within the task. The description given by the user defining the anchor segment limits the unexpectedness characteristic of serendipitous links. We believe that judgements that go beyond answering yes or no to the following question, addressed when evaluating with ATM, should be considered: *Based on the description, would the person be satisfied watching the second video clip after having watched the first video clip?*. Indeed, it is more time consuming to address more questions than only one, but we find it necessary in order to understand better the user needs. While in a real-life scenario what is serendipitous for one user is not for another and thus what is relevant for one is not necessarily for another, having more judgements for each link and questioning the link from different perspectives to capture diversity could help build systems with serendipitous capabilities in hyperlinking.

We propose next to address the minuses identified in the evaluation as done in the Search and Hyperlinking task, and define a new evaluation scenario. We believe that the benchmarking initiative for hyperlinking is an important one and the purpose of the new evaluation scenario is to complement this initiative. We want to test if it is feasible and interesting to have more judgements for each anchor-target pair and to have each person evaluating the links judge diverse links, motivated by different aspects of the information contained. Another goal is to assess if the links are serendipitous.

Judgement	Motivation given by the turker	Overlap with anchor
Relevant	The second video is about old castles in the UK	Yes
Not relevant	The person wishes to see more videos about old castles in the UK. So, he will not be satisfied watching the second video clip	Yes
Relevant	The videos both have historical importance	No
Relevant	The second video is the continuation of the first one	No
Relevant	The second video has more information about many other topics instead of just stone castles of England	No
Relevant	The second video tells the story of John who lived in a castle in UK	No
Not relevant	Both videos are from the same show	No
Not relevant	The second video provides information about the time period in England when castles were used; however it does not provide any meaningful details on the actual castles. Instead it focuses on the time period	No
Not Relevant	Both videos deal with medieval England/Europe, but differently. The second video mentions castles in passing but is mostly about the nature of kingship	No
Not Relevant	The second video was more about the life and times of people in the medieval world[...] It talked about the class system existing in the medieval times	No
Not relevant	Second clip is about kings not castles	No
Not relevant	Person is looking for castle not for history lesson	No

Figure 7.6: Examples of judgements for targets proposed for the anchor in the previous example. The targets are in top 40 and are extracted from the same video as the anchor segment. The motivations given by the turkers for the relevance judgements are also given.

7.5 Assessing serendipity and diversity in the links

The new evaluation scenario aims at reevaluating the targets used in the previous experiments, searching to assess diversity and serendipity in the links. For this evaluation we have created an online survey[4]. The survey contains two parts. The first part introduces the users to the task and requests general information to be filled in (e.g., age, field of study). The second part is the actual judging of the hyperlinks after watching the anchor and target video segments. We will present next the survey in more details and afterwards statistics on the evaluation and the results obtained.

7.5.1 Survey

Figure 7.7 illustrates the first part of the survey. This part deals with gathering information about the participants that are completing the survey and describing the general idea of the task. We did not have a way to finance participants as done with AMT, so we made a call for evaluation to several research groups. The goal of gathering this information is to observe if there is variation between the participant's profiles. The strategies we test for hyperlinking are not designed for a targeted audience but rather to anyone, no matter the age, gender, studies done, etc.

The second part of the survey is illustrated in Figure 7.8. First we describe the scenario, so that participants understand the task and what a target video segment should bring. A target should bring diverse information, giving a global picture of the information contained in the anchor, or a different point of view, or more details on the subject, etc. We propose a series of questions and answers participants can choose from. What we want to do is first establish if there is a topical connection between the anchor clip and the target clip, and then ask whether or not the connection was something the person would have thought of themselves. The goal is to assess if the target is serendipitous and if the participant can find the connection between the anchor and the target. We propose two targets in parallel for each anchor with the goal of assessing which one is found to be more interesting for further exploration.

Given our approaches using the topical structure, we wished to incorporate also questions regarding the topic granularity at which an anchor relates to the target. However, after several preliminary trials we noticed that it is difficult for participants to identify topic granularity. The notion of general and specific topics is not easy to comprehend. Asking a participant if the link is justified by the fact that the anchor and target share the same general/specific target turns out confusing for the participant. We could envision a potential manner to incorporate this aspect of granularity by specifically asking if a certain topic (selected from the topical structure) can explain the link. Still, a way to formulate the topic for participants to understand needs to be found. Thus, the second part of the survey is organized around three questions. The first one questions the relationship between the anchor and each of the two targets proposed. The two targets represent a combination between the possible types of targets. The aim is to find if the participant can differentiate between the types of targets and judge them as related from five various points of view or not related. The target can be related if it is from the same program or series, or the same program or

[4]The survey is available at http://limah.irisa.fr/hyperlinking_scenario/hyperlink.php We would like to thank Martha Larson for her valuable input in realizing this survey.

series and on the same topic, or it is on related topics, or they seem to be related even though it is difficult to tell. The second question aims at capturing the unexpected character of the targets. From the different answers proposed a participant can choose whether such a target would have occurred to him/her, or it was possible to occur, or it would never have occur to him/her. The third question focuses on how interesting the targets are. The aim is to compare the targets to see what would participants find more interesting for further exploration. They can choose both targets, none or one of them.

Differently than how is done in the MediaEval framework, where there is only one evaluation per anchor-target pair and each turker can do any number of evaluations, we imposed at least 3 evaluations per anchor-target pair and each participant was asked to do 5 evaluations. Due to limited resources for doing this survey, we have chosen to evaluate only some anchor-target pairs that would allow us to assess the potential of indirect hyperlinking. For this, we have randomly selected 5 anchors and the first top target proposed by the following systems: DirectH, $\text{IT}_{\text{Comb<}}$ and $\text{IT}_{\text{Comb>}}$. We chose these systems to understand if the participants can distinguish between very similar anchor-target pairs and pairs similar from a general/specific point of view. Irrelevant targets were also proposed by randomly selecting 1 minute length video segments from the collection, that have 0 cosine similarity score with the anchors. These targets are chosen as a safety to check if the judgement was correctly done. In case such a target was found relevant we do a verification of the evaluation. Each anchor has four targets associated: very similar(*DirectH*), general($IT_{Comb>}$), specific($IT_{Comb<}$), unrelated (*Unrelated*). Each participant was asked to do do 5 evaluations that consist of different combinations between the types of targets (i.e., *DirectH–*$IT_{Comb>}$, *DirectH–*$IT_{Comb<}$, $IT_{Comb>}$*–*$IT_{Comb<}$, *DirectH–Unrelated,* $IT_{Comb>}$*–Unrelated,* $IT_{Comb<}$*–Unrelated*). The combinations selected for each participant are prioritized so that those that are for different anchors (not already seen by the participant) and with different combination type (not already done by the participant) are proposed first. Imposing 5 evaluations per participant gives insight whether the participants distinguish between the combinations tested for anchor-target pairs or not.

Concerning the evaluation, only the first relevant target proposed by each system is selected for each anchor. Thus, each of the 5 anchors will have associated 6 combinations between target types. Each combination should be evaluated at least 3 times by different participants. This results in $5x6x3 = 90$ combination to test in total. Having 5 evaluations done by each participants, it means we need at least 18 participants. After sending the survey to the community, in three days all the necessary evaluations were completed. There were 23 out of all participants that have done all 5 evaluations requested. To sum up there are $5x6 = 30$ total combinations done at least 3 times, with 25 out of these combinations done 4 times. We detail next the results obtained with the new evaluation scenario.

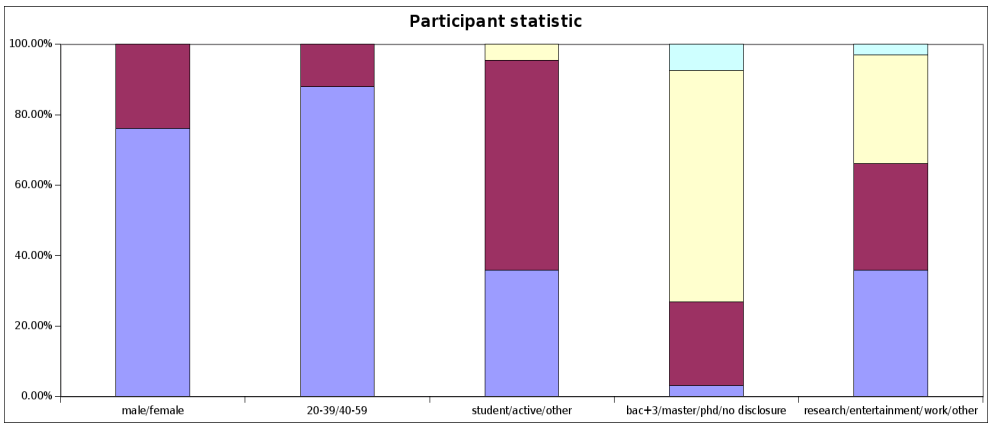


Figure 7.9: Statistics on the participants to the survey

7.5.2 Evaluation and results

7.5.2.1 Participants

There were 72 participants that managed to submit at least one response. An overview over the participants profile is given in Figure 7.9. Each column corresponds to the percentage of participants in the categories from the x-axis. These categories correspond to the answers given in the first part of the survey. The first one is for the gender question and we can observe that there are $\approx$ 80% male participants and 20% female. It can also be observed we have participants of various ages (second column) and interests(fifth column). Since the targeted audience was in research, it is expected to have high study levels with most positions occupied in the range of: engineers, PhD students, post-doctoral researchers, (assistant) professors. These results show we do have diversity in the participant's profile, which is important in such an evaluation that does not target a certain audience.

7.5.2.2 Hyperlinking evaluation

To evaluate the responses to the survey we proposed three strategies: First, we examine if the participants managed to differentiate the relations between anchors and targets (7.5.2.3). Second, we compute the user agreement on the answers provided for the questions in the survey(7.5.2.4). Third, we investigate if the participants are consistent in their answers when assessing different types of targets (7.5.2.5).

7.5.2.3 Overall differentiation between target types

We first want to understand what types of targets were associated to each kind of possible answer for the three questions asked in the survey. Figure 7.10 gives an overview of the types of targets being associated to he possible answers for the three questions, asked in the survey, for all anchors (Question1 7.10(a), Question2 7.10(b), Question3 7.10(c)). Detailed statistics for each anchor individually are given in Appendix B.

Question 1: ***Which answer best describes your thoughts on the relationship between A and B/C?*** Figure 7.10(a) reports the results obtained for this question. It can be observed that on average $\approx$ 25.2% of the general targets are considered from the same program or series as the anchor, while $\approx$ 16.5% are considered not related. For the specific targets it can be observed that $\approx$ 8.6% are considered from the same program or series as the anchors and $\approx$ 21.4% are considered not related. A large proportion of specific targets is considered from the same topic as the anchor $\approx$ 24.28%. Regarding the very similar targets $\approx$ 33.5% are judged as being from the same program/series and same topic. This value is higher than those obtained for the other possible answers to this question for the very similar targets. The second larges proportion of answers for very similar targets corresponds to same program or series targets and is $\approx$ 30.8%. The difference in judgements between the types of targets shows that participants identify that general and specific targets are more often on related topics with the anchors than the very similar targets are (19% and 20% compared to 11%). These results show that as expected "very similar" targets are found more similar than general or specific target, representing near duplicated and timeline events rather then diverse. The most diversity is offered by the specific topics. The fact that the number of specific targets is reduced for the same program or series category compared to the other categories, is an interesting result, hinting towards diversity in the links. This is expected since the core of the methods that propose general and specific targets aim at going further than hyperlinking on the exact same topic. There are also 3% of unrelated targets that were considered as seemingly related but difficult to tell. We have verified these unrelated targets and they do not seem to have anything in common with the anchors.

Question 2: ***If you had been searching online for material related to A, do you think that you would have explicitly been searching for something like B/C?*** The results obtained for this question are reported in Figure 7.10(b). The goal here is to observe the unexpectedness character of the targets proposed. The very similar targets are more intuitive than the others, since a higher percentage of targets have been put into the category *'Yes, it would have occurred to me.'*. Contrarily, the specific targets are less intuitive compared to the general and very similar ones. This shows an advantage for using the hierarchy of topics to propose targets that are on different points of view. The participants would not immediately expect these targets compared to very similar ones. As for the unrelated targets it is expected to have them categorized as *'Never would have occurred to me'*. Looking at both the answers from Question 1 and Question 2 we observe that $\approx$ 89% of the targets judged as in the same program or series and the same topic as the anchor are also judged as expected. Following the same analysis it results that $\approx$ 71% of the targets judged as same program or series are judged as expected, while $approx$27% are put into the category *'Possibly, it would have occurred to me.'*. Looking at the targets judged as having the same topic as the anchor, $\approx$ 58% are found expected, while $\approx$ 42% as possibly. The targets that are judged as on related targets are $\approx$ 29% expected and $\approx$ 62% possibly. All these values show that targets that are near duplicates (same program or series and same topic) or timeline events (same program or series) are more expected than targets on the same topic or related one. These results are encouraging, showing that the new approaches offer more diversity compared to classical techniques that propose very similar targets.

Question 3: ***If you were interested in clip A which of the 2 clips B and C would you choose as a starting point for further exploration?*** Figure 7.10(c) illustrates the results obtained for this question. The main objective is to capture the interestingness character of the targets for the participants. Participants seem to prefer to explore both very similar and general targets or very similar and specific targets than further exploring only one of them. Keeping in mind that the participants that answer this question are not the ones that have chosen those anchor video segments as interesting videos for them, it is difficult to evaluate this question at its full potential since we are asking participants to act as if they were interested in the anchors. We believe this is the main reason for preferring the two targets, in each combination tested, as starting points, (in case they are both relevant for the anchor) rather than only one. Also, general targets are preferred for further exploration rather than specific ones, giving a larger spectrum for further exploration.

7.5.2.4 Agreement among participants

The second set of evaluations consists in analyzing the participants agreement. This means looking at the answers chosen for each combination of targets, for each participant and measure the agreement in the answer selected for each question. To compute the degree of agreement between participants we rely on Fleiss' kappa statistical measure [42]. This measure shows how much consensus there is in the answers provided. There is no generally agreed-upon measure of significance, although there exist some guidelines. The kappa value is defined as: $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$, where $\bar{P}$ is the observed level of agreement and $\bar{P}_e$ is the value expected if the raters were totally independent, $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance and $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance. If the participants are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$. Table 7.6 gives the κ values obtained to represent the degree of agreement for the answers given to the first question for each anchor and combination tested. As it can be observed participants mostly agree when the combination contains an unrelated target. For anchor 3 and 5, the agreement between the participants is higher compared to the others anchors. This is probably due to the fact that the targets in the combinations for these two anchors are easier to distinguish. Table 7.7 gives the κ values obtained for the second question. Consistent with the previous results, there is a lot of disagreement between participants. This is consistent to real life. Participants have different interests and views and this reflects in a task that is based on promoting serendipitous encounters. This disagreement persists also at the third question. This study shows how difficult it is to assess the potential of the proposed systems. While having only one judgement for each anchor-target pair does not allow to evaluate correctly a system, having more judgements results in disagreement. The lesson to learn from this experiment is that if we want to make a system that provides serendipitous encounters for participants, without having any profile on their preferences, we should expect high disagreement about why something is relevant and whether is unexpected and interesting.

7.5.2.5 Per user differentiation between target types

The next set of evaluations aims at identifying if the users are constant in their judgements towards the type of targets. In other words, when they test different combinations of tar-

anchor	general specific	general very similar	general unrelated	specific very similar	specific unrelated	very similar unrelated
1	-0.17	-0.09	-0.14	0.11	0.05	0.27
2	-0.24	-0.33	0.52	-0.22	0.45	0.47
3	-0.09	0.27	0.65	0.23	0.47	0.58
4	-0.33	-0.2	-0.33	-0.25	0.48	0.27
5	0.44	-0.33	0.36	0.58	-0.14	0.47

Table 7.6: Fleiss' κ measure of agreement between all answers given for each existing combination of targets for each anchor (Question 1).

anchor	general specific	general very similar	general unrelated	specific very similar	specific unrelated	very similar unrelated
1	0.16	-0.24	0.11	0.06	0.13	0.24
2	-0.33	-0.33	0.56	-0.25	0.65	0.58
3	-0.06	-0.06	0.65	0.4	0.58	0.58
4	-1.2	-0.25	-0.33	-0.24	0.56	0.58
5	0.56	0	0.47	0.58	-0.14	1

Table 7.7: Fleiss' κ measure of agreement between all answers given for each existing combination of targets for each anchor (Question 2).

gets, do they assign the same answer to the same type of target or not? In Figure 7.11, the agreement (computed with Fleiss' κ) between the answers given for the same type of target by the same participant is presented. As it can be observed there is not much agreement over the answers for the first question compared to those for the second question. It means it is difficult for participants to tell where the different types of targets should fit, from the answers provided. However they manage to differentiate between the types of targets for the second question and are usually consistent in their choice of what they find unexpected (general, specific or very similar).

An interesting further step would be integrating the link justification via the topics to guide participants during the evaluation. Such a guidance could help the evaluation. If the justification of a link fits or not according to an assessor it could be easier to judge. Also, having the justification of why a certain link was proposed can improve the acceptance of serendipitous targets. Given that some of the targets that were judged as relevant with AMT have not been found relevant to the anchors by the participants to the survey ($\approx$ 16.56 general targets, $\approx$ 21.42 specific targets, $\approx$ 2.6 very similar targets), or that some links were judged as "seem related but difficult to tell"($\approx$ 6.17 general targets, $\approx$ 21.42 specific targets, $\approx$ 4 very similar targets), having an explanation for these links might have helped the participants understand the connection between the anchor and the target that other evaluators have made.

Welcome to video hyperlinking!

Task description:

We are working on algorithms to automatically create links between video fragments. These links are typically used for recommendation, to provide additional information and to facilitate knowledge and content discovery. For a given video fragment, links should point to video fragments bearing a relation with the initial fragment (e.g., same program, or topic, or person, or place, etc.), possibly exhibiting diversity, e.g., to capture different aspects of a subject.

The survey will present you with a fragment of video (Clip A) along with two linked fragments of 1 minute each. You will be asked to judge the relevance of each linked fragment and to rank them. After filling in general information, you will have a total of 5 tests to perform, which should take you about 10-15 minutes.

We thank you for your collaboration.

General questions

*You are: ○Male ○Female ○I choose not to disclose
*Age: ○<20 ○20-39 ○40-59 ○>60
Status: ○Student ○Active ○Other
*Level: ○<Bachelor degree (Bac) ○<Licence degree (Bac+3) ○<Master degree (Bac+5) ○>Master degree (PhD)
○I choose not to disclose
In case you are a student, what is your field of study?
In case you are employed, what is your profession?
*What is your main scope when using the internet? (Multiple choices are possible.)
☐Research ☐Entertainment ☐Work ☐Other

By doing this survey, you agree that we may publish parts of your answers as part of our research study. We will NOT publish any information that could be linked to you. We do NOT use your worker ID, or any other information that links to you, during data analysis or storage. Your answers are used only by researchers for the purposes of gaining insight into general opinions concerning video hyperlinking. Beyond the people who are doing research in this area, no other parties are allowed to use your answers.

Start the survey

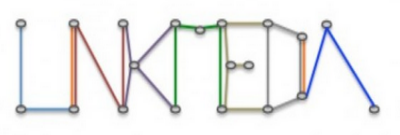


Figure 7.7: New survey for hyperlinks evaluation (page 1)

Scenario:
Imagine you are watching Clip A, a video fragment on a topic you are interested in and that caught your attention. You want to continue exploring in that direction and gain knowledge on the subject. For instance, you might want to see something that gives a global picture, a different point of view, or more details on the subject, etc. Any type of relation between clip A and the proposed suggestions might be of interest. To guide you in your exploration, we propose the two clips B and C below.

Clip A

Two video clips (B and C) that could be linked to video A are recommended to you that should encourage this further exploration. Please watch the two videos and answer the questions.

Clip B

Clip C

Choose the answer that best describes your thoughts on the relationship between A and B.
- ○ It's easy to tell that they are related, they are from the same program or series.
- ○ It's easy to tell that they are related, they are on the same topic.
- ○ It's easy to tell that they are related, they are from the same program or series and on the same topic.
- ○ It's easy to tell that they are related, they are on related topics.
- ○ It's difficult to tell that they are related, but it seems that they are.
- ○ A and B do not seem to be at all related.

Choose the answer that best describes your thoughts on the relationship between A and C.
- ○ It's easy to tell that they are related, they are from the same program or series.
- ○ It's easy to tell that they are related, they are on the same topic.
- ○ It's easy to tell that they are related, they are from the same program or series and on the same topic.
- ○ It's easy to tell that they are related, they are on related topics.
- ○ It's difficult to tell that they are related, but it seems that they are.
- ○ A and C do not seem to be at all related.

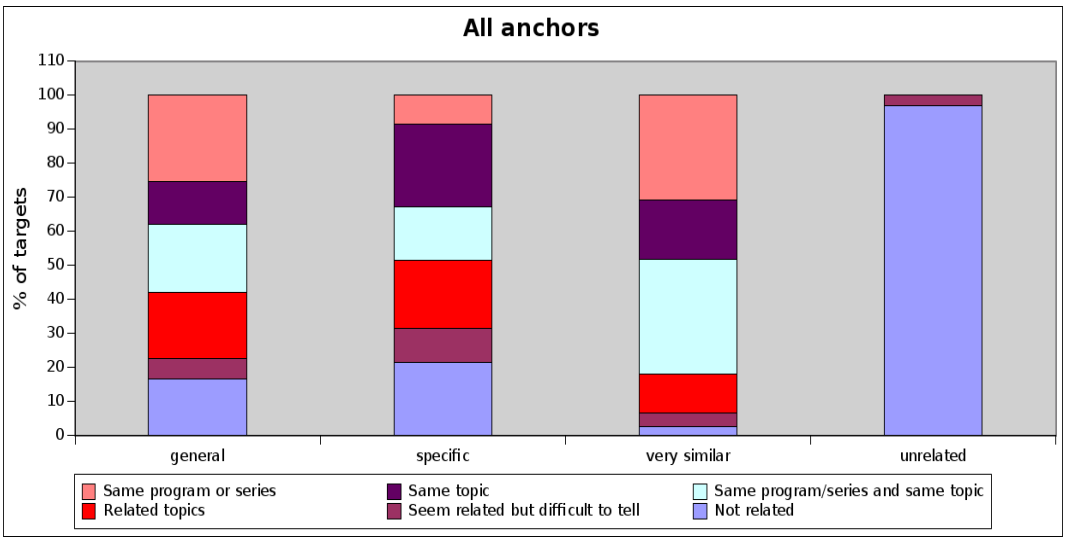
If you had been searching online for material related to A, do you think that you would have explicitly been searching for something like B?
- ○ Yes, it would have occured to me.
- ○ Possibly, it would have occured to me.
- ○ Never would have occured to me.

If you had been searching online for material related to A, do you think that you would have explicitly been searching for something like C?
- ○ Yes, it would have occured to me.
- ○ Possibly, it would have occured to me.
- ○ Never would have occured to me.

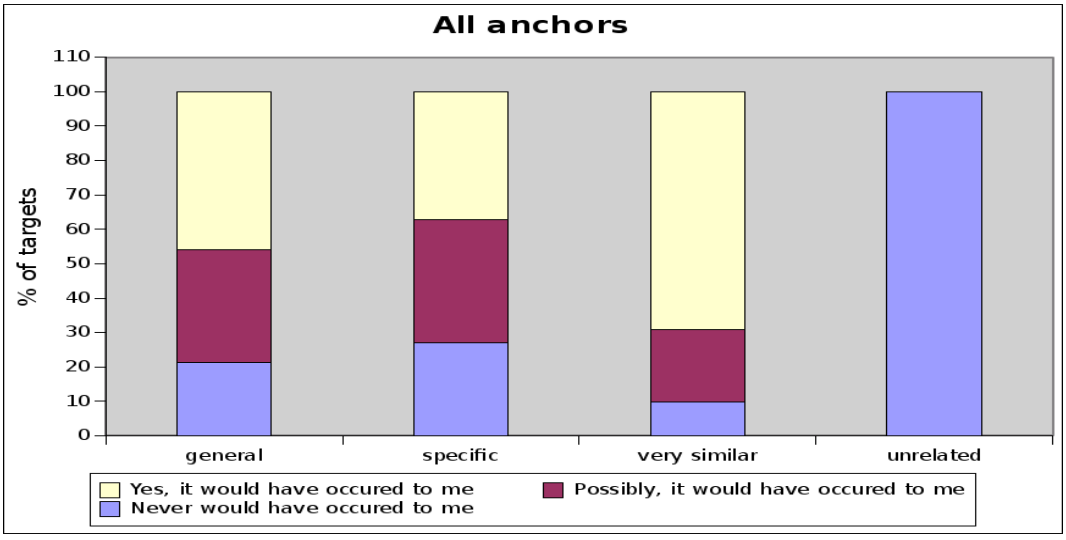
If you were interested in clip A which of the 2 clips B and C would you choose as a starting point for further exploration?
○ B ○ C ○ None ○ Both

Submit
(count: 0)

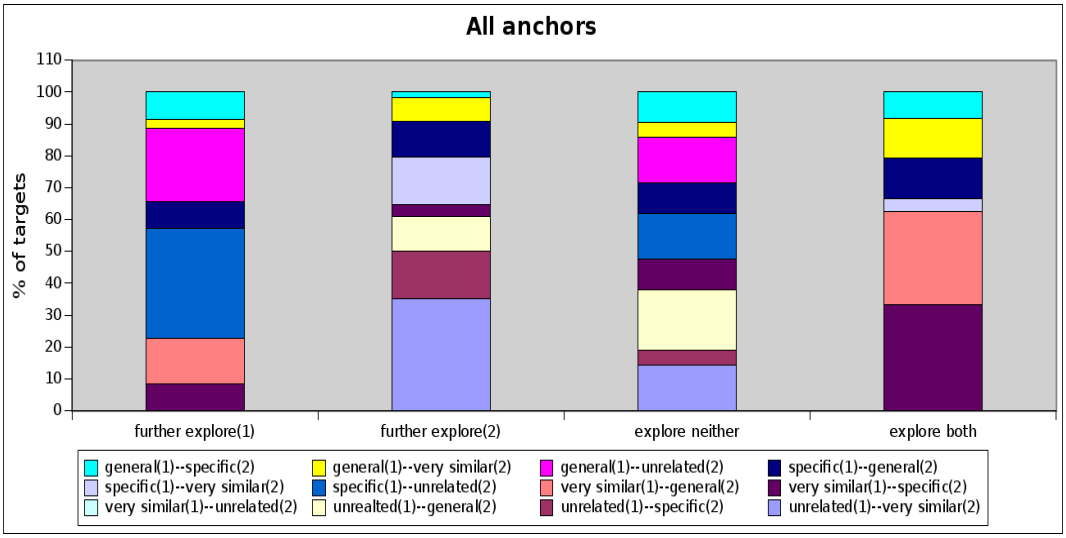
Figure 7.8: New survey for hyperlinks evaluation (page 2)



(a)

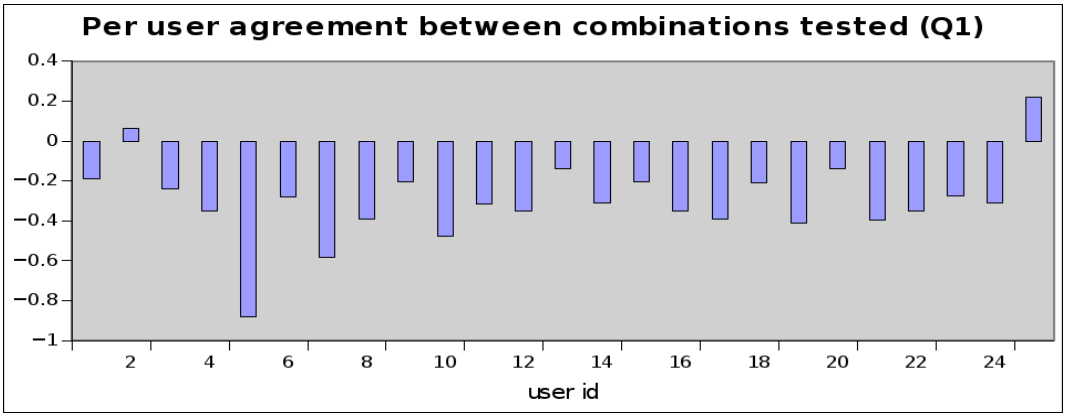


(b)

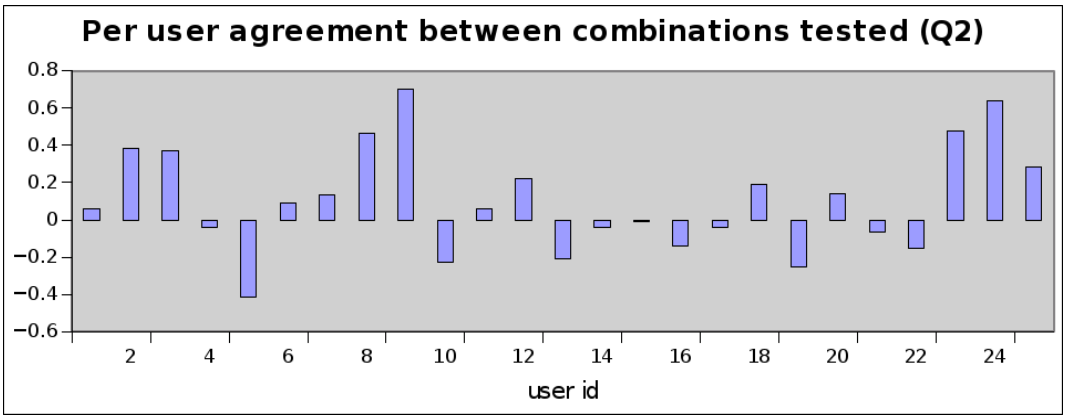


(c)

Figure 7.10: Average percentage of targets, for each type of target, with any possible answer associated to. Figure 7.10(a) contains the possible answers to first question in the survey. Figure 7.10(b) and 7.10(c) contain the answers for the second and third questions, respectively.



(a)



(b)

Figure 7.11: Agreement per user between the answers given for the same type of target for the first question 7.11(a) and for the second question 7.11(b)

7.6 Conclusion

The work presented in this chapter addresses two important aspects of hyperlinking, namely serendipity in the links and link justification. The experiments reported in the first part of this chapter show that indirect hyperlinking via the topical structure attains comparable results to direct hyperlinking, with the advantage of enabling serendipity and link justification. Indeed, thanks to this informative structure a link can be justified by the topics that contributed to its creation. More, the topics in the structure have different levels of specificity which can offer diversity in the links. However, the evaluation as done in the context of the Search and Hyperlinking task does not capture these aspects and has several disadvantages. Among them one single judgement per anchor-target pair, only a yes no question to assess if a target is relevant or not and the judgement is accompanied by a description on what information the target should contain. Thus in the second part of the chapter we propose a new evaluation scenario, that aims to assess the relationships between anchor-target pairs, if the links are unexpected and informative, while imposing at least 3 judgements per anchor-target pair. The results offer new opportunities for link authoring both form a technical point of view and from a user perspective. New ways of creating the structure and accounting the topics can uncover more diversified analogies between video fragments. Additionally, by inducing serendipity in a controlled manner, we can understand the users needs and help maximize their ability to explore a collection for serendipitous information encounters. The most important outcome of this work is that all the experiments on agreement question the evaluation of hyperlinking. New ways to evaluate it need to be considered. A possible direction might be the evaluation through a user interface. Giving users the possibility to select targets to follow on based on some relation to the anchor. This way we can understand what users prefer and we could focus on perfecting the selection of targets to grasp the required relations. It would mean to move the focus from providing relevant targets given an anchor, for everyone, to providing targets motivated by explicit relations. For example some targets could be visually based so for example, if the anchor is about a certain person, a set of targets that provide relevant/unexpected/interesting information about that person, can be proposed. Other targets could be on different points of view on the topic discussed in the anchors, etc.

Chapter 8

Conclusions

Contents

8.1 Thesis objectives **121**
8.2 Summary of the contributions **122**
8.3 A step further **123**

In this chapter we first look back at the initial objectives of this thesis and then sum up the work done to achieve these objectives. Some future extensions that come naturally from the technical content of the work done have already been presented at the end of each chapter dedicated to a contribution (Chapters 3, 4, 6, 7). Thus in the end of this chapter we will consider research ideas that are further along the line as a continuation of the underlying goals of this thesis.

8.1 Thesis objectives

One of the critical scientific challenges nowadays is how to enable the era of data science. As the volume, complexity and heterogeneity of data grows, new methods and models are required to structure the data, with the ultimate goal of extracting value to drive data science. A large amount of the data deluge is represented by audiovisual television content. It is the landscape of data science focused on television content we were interested in this thesis and the interest was twofold. The first objective consisted in proposing new solutions for *generic and automatic topical structuring of TV shows*. This objective is justified by the over increasing amount of heterogeneous television data that has to be structured to extract *Value* from it. And a valuable information in this data is the overall semantic organization, that can be captured through the identification of the topical structure. The second objective consisted in studying the implications of these solutions for obtaining the topical structure of television data in the context of *video hyperlinking*. The purpose of video hyperlinking is to improve the

exploitation of large video collections by users with various backgrounds and interests. With these two objectives in mind we sum up next the contributions made to tackle the existing challenges.

8.2 Summary of the contributions

For the first objective, focused on automatic generic structuring of television content, we managed to introduce several structuring techniques, described in Chapter 3 and Chapter 4. Their aim was to address several limitations of current topical structuring approaches, linear or hierarchical. Linear topic segmentation classically relies on one of two criteria, either finding areas with coherent vocabulary use or detecting discontinuities. Techniques relying on the first criterion face the problem of over-segmentation which, up to now, can only be solved by providing prior information regarding the distribution of segment lengths or the expected number of segments. The techniques relying on the second criterion cannot cope with segments of variable lengths and need several parameters to detect and decide where to place frontiers. These considerations naturally lead us to the idea of a method combining lexical cohesion and disruption to make the best of both worlds. Therefore, we proposed a segmentation criterion combining both criteria, enabling a trade-off between the two. We provided the mathematical formulation of the criterion and an efficient graph-based decoding algorithm for topic segmentation. Experimental results on standard textual data sets and on a more challenging corpus of automatically transcribed broadcast news shows demonstrate the benefit of such a combination. This new approach represents our first contribution for obtaining a linear topical structure for TV shows. The second contribution consists in moving from a linear structure of topics to a hierarchical one. We started with an investigation following some basic questions such as: Can we find when it is worth segmenting the data or not? Are the measures currently employed enough to justify a reference segmentation? Given that the identification of subtle topic changes is difficult to identify, can we find the salient fragments in the text that correspond to the topics discussed? And can these fragments be at different levels of details? All these questions were addressed in an exploratory study which lead us to propose a new hierarchical topical structure. This new structure consists of a hierarchy of topically focused fragments. It was obtained by leveraging the temporal distribution of word reoccurrences, searching for bursts, to skirt the limits imposed by a global counting of lexical reoccurrences within segments. The underlying idea is that the presence of lexical bursts indicates a strong topical focus. Relying on Kleinberg's algorithm for detecting the words burst allowed us to extract the important ideas in the data at various levels of details and cluster them into a hierarchy of topically focused fragments. To evaluate the hierarchy obtained we proposed first to compare it to a reference dense segmentation on varied data sets. This comparison indicated that we can achieve a better topic focus while retrieving the important aspects of a text. Another evaluation was done in the context of automatic summarization, where summaries are usually limited by length. One of the proven benefit of using this new structure for summarization is that we manage to keep the important information from the initial data. Additionally, it helps bring to surface other information by disregarding parts of data where the important words tend to fade away.

For the second objective of this thesis we have used the structuring solutions previously mentioned to segment videos into linkable content, to be used further for video content hy-

perlinking. Existing approaches for hyperlinking mostly focus on linking anchors with relevant videos but do not pay much attention to precise target selection and anchor detection, exploiting domain-independent techniques. We proposed two approaches, one for target selection and one for anchor detection. The target selection approach exploited explicit topic segmentation, whether hierarchical or not. For the anchor detection we leveraged the hierarchy of topically focused fragments. The evaluation was performed on a data set with videos from BBC, in the context of the hyperlinking sub-task at MediaEval 2013 and anchoring in video archives sub-task at MediaEval 2015. We obtained positive results both for target selection and anchor detection. The next step was to focus on understanding the users needs in terms of the links created. For this, we leveraged a hierarchical topical structure to address two essential aspects of hyperlinking, namely, serendipitous link creation and link justification. We proposed different approaches exploiting a hierarchy of topic models as an intermediate representation to compare the transcripts of video segments. These hierarchical representations offer a basis to characterize the hyperlinks, thanks to the knowledge of the topics which contributed to the creation of the links, and to offer serendipity by choosing to give more weights to either general or specific topics. First round of experiments are performed on BBC videos from the Search and Hyperlinking task at MediaEval 2013 and 2014, achieving link precision comparable to direct text comparison. This evaluation was done via crowdsourcing using the Amazon mechanical turk platform. We believe this evaluation has several minuses and cannot capture some important aspects of hyperlinking. Therefore we proposed a new evaluation scenario that still consists of user-based evaluations and aims at complementing the previous evaluation done for MediaEval. We did a second round of experiments within this new scenario, that allowed us to have more judgements for each anchor-target pair and relevance assessment for the links in terms of unexpectedness and interestigness. The results obtained showed there is a high disagreement among the participants doing the evaluation which questions the evaluation of hyperlinking. Participants have different interests and views and this reflects in a task that is based on promoting serendipitous encounters.

8.3 A step further

Several future directions have been already suggested to improve the solutions presented in this thesis. Of course the suggested improvements are not exhaustive and new ways of integrating concepts, analyzing the problems, combining approaches etc. can be envisioned. To keep up with the consumers requirements which get more diversified and personalized we should pay more attention to user-based evaluations. As user centric solutions are gaining momentum handling the user subjective assessment of the quality of experience becomes an important part of further research. If we think about the study we have done for video hyperlinking, the high disagreement between participants taking part at the evaluation proves the need of new ways to evaluate systems. What we envision is first solutions that can offer diverse results that would meet diverse requests from users. Then, an evaluation that can capture the system's response to each user request through a measure of user satisfaction could be an indicator for system's quality. For example, in video hyperlinking, an anchor might be interesting to a user from the visual point of view or from the spoken content. Thus, a system that could provide diverse targets based on each modality and let the user choose which target to follow on would give more insight into its capabilities. An overall

satisfaction assessment after exploring a video collection via diverse hyperlinks could be a way to evaluate the system. Building user centric evaluation scenarios is a complex task in itself. Probably the creation of smart interfaces that can display the diverse links and targets to points of interest in videos is the way to go.

Another interesting step further we have not discussed in the context of hyperlinking would be to explore all modalities in the data and combine them to offer even more diversity and serendipity in the results. One simple existing solution is to work on each modality independently and then combine the results. A more interesting solution we have started to investigate is doing an early fusion of modalities through a translation from one modality to another one. The goal is to diversify more the targets by creating links that *show* related information to what it is *spoken* about in the anchors and vice versa. Still the challenging part remains the evaluation of such links. Having them evaluated together with other types of links is not enough.

Bibliography

[1] Maristella Agosti and James Allan. "Introduction to the special issue on methods and tools for the automatic construction of hypertext". In: *Information Processing and Management* 33.2 (1997), pp. 129 –131.

[2] James Allan, ed. *Topic Detection and Tracking: event-based information organization*. Kluwer Academic Publishers, 2002.

[3] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. "Topic Detection and Tracking Pilot Study Final Report". In: *DARPA broadcast news transcription and understanding workshop*. 1998, pp. 194–218.

[4] Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. "Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words". In: *PLoS ONE* 4.11 (Nov. 2009).

[5] Robin Aly, Dolf Trieschnigg, Kevin McGuinness, Noel E. O'Connor, and Franciska De Jong. "Average precision: Good guide or false friend to multimedia search effectiveness?" In: *Intl. Conf. on Multimedia Modeling*. 2014.

[6] Nicola Ancona, Grazia Cicirelli, Antonella Branca, and Arcangelo Distante. "Goal detection in football by using support vector machines for classification". In: *International Joint Conference on Neural Networks*. Vol. 1. 2001, pp. 611–616.

[7] Roxana Angheluta, Rik De Busser, and Marie-Francine Moens. "The Use of Topic Segmentation for Automatic Summarization". In: *In Workshop on Text Summarization in Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization*. 2002, pp. 11–12.

[8] Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim. "An Investigation on the Influence of Genres and Textual Organisation on the Use of Discourse Relations". English. In: Lecture Notes in Computer Science 8403 (2014), pp. 454–468.

[9] Doug Beeferman, Adam Berger, and John Lafferty. "Text segmentation using exponential models". In: *2nd Conference on Empirical Methods in Natural Language Processing*. 1997, pp. 35–46.

[10] Mathieu Ben and Guillaume Gravier. "Unsupervised mining of audiovisually consistent segments in videos with application to structure analysis". In: *IEEE International Conference on Multimedia and Exhibition* (2011), pp. 1–6.

[11] Sid-Ahmed Berrani, Gaël Manson, and Patrick Lechat. "A Non-supervised Approach for Repeated Sequence Detection in TV Broadcast Streams". In: *Image Commun.* 23.7 (Aug. 2008), pp. 525–537.

[12] Marco Bertini, Alberto Del Bimbo, and Pietro Pala. "Content-based Indexing and Retrieval of TV News". In: *Pattern Recognition Letters* 22.5 (2001), pp. 503–516.

[13] Chidansh Bhatt, Nikolaos Pappas, Maryam Habibi, and Andrei Popescu-Belis. "Idiap at MediaEval 2013: Search and Hyperlinking Task". In: *Working Notes Proc. of the MediaEval Workshop*. 2013.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[15] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. "Verifying Multimedia Use at MediaEval 2015". In: *Working Notes Proc. of the MediaEval Workshop*. 2015.

[16] Ilaria Bordino, Yelena Mejova, and Mounia Lalmas. "Penguins in Sweaters, or Serendipitous Entity Search on User-generated Content". In: *22Nd ACM International Conference on Information and Knowledge Management*. 2013, pp. 109–118.

[17] Gillian Brown and George Yule. "Discourse analysis". In: *Cambridge University Press* (1983).

[18] Pierre Cadiot and Bernard Fradin. "Présentation : une crise en thème?" In: *Langue française* 78.1 (1988), pp. 3–8.

[19] Wallace L. Cafe. "The flow of thought and the flow of language". In: *Syntax and Semantics: Discourse and Syntax*. Vol. 12. 1979, pp. 159–182.

[20] Paula C. F. Cardoso, Maite Taboada, and Thiago A. S. Pardo. "On the contribution of discourse structure to topic segmentation". In: *Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 2013.

[21] Lynn Carlson, Daniel Marcu, and Mary E. Okurowski. "Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory". In: *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 2001, pp. 1–10.

[22] Lucien Carroll. "Evaluating hierarchical discourse segmentation". In: *11th International Conference of the North American Chapter of the Association for Computational Linguistics*. 2010, pp. 993–1001.

[23] Shu-Ching Chen, Mei-Ling Shyu, Min Chen, and Chengcui Zhang. "A Decision Tree-based Multimodal Data Mining Framework for Soccer Goal Detection". In: *Proc. of IEEE International Conference on Multimedia and Expo*. 2004, pp. 265–268.

[24] Freddy Y. Y. Choi. "A speech interface for rapid reading". In: *IEE colloquium: Speech and Language Processing for Disabled and Elderly People*. 2000, pp. 1–4.

[25] Freddy Y. Y. Choi. "Advances in domain independent linear text segmentation". In: *1st International Conference of the North American Chapter of the Association for Computational Linguistics*. 2000, pp. 26–33.

[26] Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. "Latent Semantic Analysis for Text Segmentation". In: *Proceedings of Empirical Methods in Natural Language Processing*. 2001, pp. 109–117.

[27] Kenneth W. Church and William A. Gale. "Poisson Mixtures". In: *Natural Language Engineering* 1 (1995), pp. 163–190.

[28] Vincent Claveau. "Acquisition automatique de lexiques sémantiques pour la recherche d'information". PhD thesis. École doctorale: MATISSE, University of Rennes 1, 2003.

[29] Vincent Claveau and Sébastien Lefèvre. "Topic segmentation of TV-streams by mathematical morphology and vectorization". In: *12th International Conference of the International Speech Communication Association*. 2011, pp. 1105–1108.

[30] Vincent Claveau, Romain Tavenard, and Laurent Amsaleg. "Vectorisation des processus d'appariement document-requête". In: *Conférence en Recherche d'Informations et Applications*. 2010.

[31] Anne Condamines and Marie-Paule Péry-Woodley. *Linguistic markers of semantic and textual relations*. 2007, pp. 3–16.

[32] Tom De Nies, Wesley De Neve, Erik Mannens, and Rik Van de Walle. "Ghent University-iMinds at MediaEval 2013: an unsupervised named entity-based similarity measure for search and hyperlinking". In: *Working Notes Proc. of the MediaEval Workshop*. 2013.

[33] Manolis Delakis. "Multimodal Tennis Video Structure Analysis with Segment Models". PhD thesis. Université de Rennes 1, France, 2006.

[34] Manolis Delakis, Guillaume Gravier, and Patrick Gros. "Audiovisual integration with segment models for tennis video parsing". In: *Computer Vision and Image Understanding* 111.2 (2008), pp. 142–154.

[35] Stefan Eickeler and Stefan Muller. "Content-based video indexing of TV broadcast news using hidden Markov models". In: *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. Vol. 6. IEEE. 1999, pp. 2997–3000.

[36] Jacob Eisenstein. "Hierarchical text segmentation from multi-scale lexical cohesion". In: *10th International Conference of the North American Chapter of the Association for Computational Linguistics*. 2009, pp. 353–361.

[37] Jacob Eisenstein and Regina Barzilay. "Bayesian unsupervised topic segmentation". In: *Conference on Empirical Methods in Natural Language Processing*. 2008, pp. 334–343.

[38] Maria Eskevich, G. J. F. Jones, Robin Aly, and et al. "Multimedia information seeking through search and hyperlinking". In: *ACM Intl. Conf. on Multimedia Retrieval*. 2013.

[39] Maria Eskevich, Robin Aly, Roeland Ordelman, David N. Racca, Shu Chen, and G. J. F. Jones. "SAVA at MediaEval 2015: Search and Anchoring in Video Archives". In: *Proceedings of the MediaEval 2015 Workshop*. 2015.

[40] Maria Eskevich, Robin Aly, David N. Racca, Roeland Ordelman, Shu Chen, and Gareth J. F. Jones. "The Search and Hyperlinking task at MediaEval 2014". In: *Working Notes Proc. of the MediaEval Workshop*. 2014.

[41] Olivier Ferret, Brigitte Grau, and Nicolas Masson. "Thematic segmentation of texts: Two methods for two kinds of texts". In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. 1998, pp. 392–396.

[42] Joseph L. Fleiss. "Measuring nominal scale agreement among many raters". In: *Psychological Bulletin* 76.5 (1971), pp. 378–382.

[43] Allen Foster and Nigel Ford. "Serendipity and information seeking: an empirical study". In: *Journal of Documentation* 59.3 (2003), pp. 321–340.

[44] Alexander Franz. "Independence Assumptions Considered Harmful". In: *In Proceedings of the 35th Annual Meeting of the ACL, and 8th Conference of the EACL*. 1997, pp. 182–189.

[45] Fumiyo Fukumoto and Yoshimi Suzuki. "Event tracking based on domain dependency". In: *SIGIR*. 2000, pp. 57–64.

[46] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. "Discourse segmentation of multi-party conversation". In: *41st Annual Meeting of the Association for Computational Linguistics*. 2003, pp. 562–569.

[47] Petra Galuščáková and Pavel Pecina. "CUNI at MediaEval 2013 Search and Hyperlinking Task". In: *Working Notes Proceedings of the MediaEval Workshop*. 2013.

[48] Petra Galuščáková and Pavel Pecina. "CUNI at MediaEval 2015 Search and Anchoring in Video Archives: Anchoring via Information retrieval". In: *Proceedings of the MediaEval 2015 Workshop*. 2015.

[49] Petra Galuščáková, Martin Krulis, Jakub Lokoc, and Pavel Pecina. "CUNI at MediaEval 2014 Search and Hyperlinking Task: Visual and Prosodic Features in Hyperlinking". In: *Working Notes Proc. of the MediaEval Workshop*. 2014.

[50] John Gantz and David Reinsel. *The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. Tech. rep. Internet Data Center(IDC), 2012.

[51] Xinbo Gao and Xiaoou Tang. "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing". In: *Circuits and Systems for Video Technology, IEEE Transactions on* 12.9 (2002), pp. 765–776.

[52] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. "The LIMSI broadcast news transcription system". In: *Speech Communication* 37.1–2 (2002), pp. 89–108.

[53] Dan Gillick and Benoît Favre. "A Scalable Global Model for Summarization". In: *Workshop on Integer Linear Programming for Natural Langauge Processing*. 2009, pp. 10–18.

[54] Talmy Givòn. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Vol. 63. 1. Linguistic Society of America, 1987, pp. 160–164.

[55] Inmar E. Givoni, Clement Chung, and Brendan J. Frey. "Hierarchical Affinity Propagation". In: *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*. 2011, pp. 238–246.

[56] Christian Grece, André Lange, Agnes Schneeberger, and Sophie Valais. *The development of the European market for on-demand audiovisual services*. Tech. rep. European Audiovisual Observatory, 2015. URL: https://ec.europa.eu/digital-agenda/en/news/development-european-market-demand-audiovisual-services.

[57] Barbara J. Grosz and Candace L. Sidner. "Attention, intentions, and the structure of discourse". In: *Computational Linguistics* 12.3 (1986), pp. 175–204.

[58] Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. "Centering: a framework for modeling the local coherence of discourse". In: *Computational Linguistics* 21.2 (1995), pp. 203–225.

[59] Camille Guinaudeau. "Structuration automatique de flux télévisuels". PhD thesis. INSA de Rennes, 2011.

[60] Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. "Enhancing Lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation". In: *Computer Speech and Language* 26.2 (2012), pp. 90–104.

[61] Camille Guinaudeau and Julia Hirschberg. "Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news". In: *In 12th Annual Conference of the International Speech Communication Association* (2011), pp. 1401–1404.

[62] Camille Guinaudeau, Anca-Roxana Şimon, Guillaume Gravier, and Pascale Sébillot. "HITS and IRISA at MediaEval 2013: Search and Hyperlinking Task". In: *Working Notes Proc. of the MediaEval Workshop*. 2013.

[63] Lynda Hardman, Dick C. A. Bulterman, and Guido van Rossum. "The Amsterdam Hypermedia Model: Adding Time and Context to the Dexter Model". In: *Communications of the ACM* 37.2 (Feb. 1994), pp. 50–62.

[64] Claudia Hauff and Geert-Jan Houben. "Serendipitous Browsing: Stumbling through Wikipedia". In: (2012).

[65] Marti A. Hearst. "Multi-paragraph segmentation of expository texts". In: *32nd Annual meeting of the Association for Computational Linguistics*. 1994, pp. 9–16.

[66] Marti A. Hearst. "TextTiling: Segmenting text into multi-paragraph subtopic passages". In: *Computational Linguistics* 23.1 (1997), pp. 33–64.

[67] Marti A. Hearst and Christian Plaunt. "Subtopic structuring for full-length document access". In: *Special Interest Group on Information Retrieval*. 1993.

[68] Louis Hébert. *Tools for Text and Image Analysis: An Introduction to Applied Semiotics*. http://www.revue-texto.net/Parutions/Livres-E/Hebert_AS/Hebert_Tools.html. 2006.

[69] Nicolas Hernandez and Brigitte Grau. "Analyse thématique du discours : segmentation, structuration, description et représentation". In: *5e colloque international sur le document électronique*. 2002, pp. 277–285.

[70] Tony Hey, Stewart Tansley, and Kristin M. Tolle, eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[71] Julia Hirschberg and Christine H. Nakatani. "Acoustic indicators of topic segmentation". In: *5th International Conference on Spoken Language Processing*. 1998, pp. 976–979.

[72] Eduard Hovy and Chin-Yew Lin. "Automated text summarization and the summarist system". In: *TIPSTER Text Program*. 1998.

[73] Stéphane Huet, Guillaume Gravier, and Pascale Sébillot. "Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition". In: *Computer Speech and Language* 24.4 (2010), pp. 663–684.

[74] Stéphane Huet, Guillaume Gravier, and Pascale Sébillot. "Un modèle multi-sources pour la segmentation en sujets de journaux radiophoniques". In: *15e conférence sur le traitement automatique des langues naturelles, TALN'08*. Avignon, France, 2008, pp. 49–58.

[75] Ichiro Ide, Koji Yamamoto, Reiko Hamada, and Hidehiko Tanaka. "An Automatic Video Indexing Method Based on Shot Classification". In: *Systems and Computers in Japan*. 2001.

[76] Ichiro Ide, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh. "Topic threading for structuring a large-scale news video archive". In: *International Conference on Image and Video Retrieval*. 2004.

[77] Xiang Ji and Hongyuan Zha. "Domain-independent text segmentation using anisotropic diffusion and dynamic programming". In: *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2003, pp. 322–329.

[78] Slava M. Katz. "Distribution of Content Words and Phrases in Text and Language Modelling". In: *Nat. Lang. Eng.* 2.1 (Mar. 1996), pp. 15–59. ISSN: 1351-3249.

[79] Anna Kazantseva. "Automatic Summarization of Short Fiction". MA thesis. University of Ottawa, 2006.

[80] Anna Kazantseva and Stan Szpakowicz. "Hierarchical Topical Segmentation with Affinity Propagation". In: *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 37–47.

[81] Ewa Kijak, Guillaume Gravier, Lionel Oisel, and Patrick Gros. "Audiovisual integration for sport broadcast structuring". In: *Multimedia Tools and Applications* 30.3 (2006), pp. 289–312.

[82] Jon Kleinberg. "Bursty and Hierarchical Structure in Streams". In: *8th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*. 2002, pp. 91–101.

[83] Michel Krieger and David Ahn. "TweetMotif: Exploratory Search and Topic Summarization for Twitter". In: *4th International AAAI Conference on Weblogs and Social Media*. 2010.

[84] Hoang An Le, Quoc-Minh Bui, Benoît Huet, Barbora Cervenková, Jan Bouchner, E. Evlampios Apostolidis, Fotini Markatopoulou, Alexandros Pournaras, Vasileios Mezaris, Daniel Stein, Stefan Eickeler, and Michael Stadtschnitzer. "LinkedTV at MediaEval 2014 search and hyperlinking task". In: *Working Notes Proc. of the MediaEval Workshop*. 2014.

[85] Wei Li and Andrew McCallum. "Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations". In: *International Conference on Machine Learning*. 2006.

[86] Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. "Analyzing Word Frequencies in Large Text Corpora Using Inter-arrival Times and Bootstrapping". In: *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*. 2011, pp. 341–357.

[87] Chin-Yew Lin. “Rouge: a package for automatic evaluation of summaries”. In: *Text Summarization Branches Out, ACL Workshop*. 2004, pp. 25–26.

[88] Diane J. Litman and Rebecca J. Passonneau. “Combining multiple knowledge sources for discourse segmentation”. In: *33rd Annual Meeting of the Association for Computational Linguistics*. 1995, pp. 108–115.

[89] Laurence Longo. “Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l’identification automatique des chaînes de référence”. PhD thesis. Strasbourg, 2013.

[90] Annie Louis and Ani Nenkova. “Automatically Assessing Machine Summary Content Without a Gold-Standard”. In: *Computational Linguistics* 39.2 (2013), pp. 267–300.

[91] Rasmus E. Madsen, David Kauchak, and Charles Elkan. “Modeling Word Burstiness Using the Dirichlet Distribution”. In: *22nd International Conference on Machine Learning*. Bonn, Germany, 2005, pp. 545–552.

[92] Igor Malioutov and Regina Barzilay. “Minimum cut model for spoken lecture segmentation”. In: *21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. 2006, pp. 25–32.

[93] William C. Mann and Sandra A. Thompson. “Rethorical Structure Theory: toward a Functional Theory of Text Organization”. In: *Text* 8 (1988), pp. 243–281.

[94] Christopher D. Manning. *Rethinking Text Segmentation Models: An Information Extraction Case Study*. Tech. rep. University of Sydney, 1998.

[95] Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, 2000.

[96] George A. Miller. “WordNet: A Lexical Database for English”. In: *Commun. ACM* 38.11 (1995), pp. 39–41.

[97] Golin Millton. “Serendipity-big word in medical progress”. In: *Journal of the American Medical Association* 165.16 (1957), pp. 2084–2087.

[98] Hemant Misra, François Yvon, Joemon M. Jose, and Olivier Cappe. “Text segmentation via topic modeling: an analytical study”. In: *Proc. ACM conference on Information and knowledge management*. 2009, pp. 1553–1556.

[99] Marie-Francine Moens and Rik De Busser. “Generic topic segmentation of document texts”. In: *24th International Conference on Research and Developement in Information Retrieval*. 2001, pp. 418–419.

[100] Paola Monachesi, Lothar Lemnitzer, and Kiril Simov. “Language Technology for eLearning”. English. In: *Innovative Approaches for Learning and Knowledge Sharing*. Ed. by Wolfgang Nejdl and Klaus Tochtermann. Vol. 4227. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, pp. 667–672.

[101] Jane Morris and Graeme Hirst. “Lexical cohesion computed by thesaural relations as an indicator of the structure of text”. In: *Computational Linguistics* (1991), 17:21–48.

[102] John Niekrasz and Johanna D. Moore. “Unbiased discourse segmentation evaluation”. In: *Spoken Language Technology*. 2010, pp. 43–48.

[103] Roeland J.F Ordelman, Maria Eskevich, Robin Aly, Benoit Huet, and Gareth J.F Jones. "Defining and evaluating video hyperlinking for navigating multimedia archives". In: *3rd International Workshop on Linked Media, co-located with WWW*. 2015.

[104] Mari Ostendorf, Vassilios V. Digalakis, and Owen A. Kimball. "From HMM's to segment models: a unified view of stochastic modeling for speech recognition". In: *IEEE Transactions on Speech and Audio Processing* 4.5 (1996), pp. 360–378.

[105] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, Georges Quéenot, and Roeland Ordelman. "TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics". In: *Proceedings of TRECVID 2015*. NIST, USA. 2015.

[106] Lev Pevzner and Marti A. Hearst. "A critique and improvement of an evaluation metric for text segmentation". In: *Computational Linguistics* 28 (2002), pp. 19–36.

[107] Johann Poignant, Hervé Bredin, and Claude Barras. "Multimodal Person Discovery in Broadcast TV at MediaEval 2015". In: *Working Notes Proc. of the MediaEval Workshop*. 2015.

[108] John Preston, Jonathon Hare, Sina Samangooei, Jamie Davies, Neha Jain, David Dupplaw, and Paul H. Lewis. "A Unified, Modular and Multimodal Approach to Search and Hyperlinking Video". In: *Working Notes Proc. of the MediaEval Workshop*. 2013.

[109] Gabriel Pui, Cheong Fung, Jeffrey Xu, Yu Philip, S. Yu, and Hongjun Lu. "Parameter Free Bursty Events Detection in Text Streams". In: *31st International Conference on Very Large Data Bases*. 2005, pp. 181–192.

[110] Bingqing Qu, Félicien Vallet, Jean Carrive, and Guillaume Gravier. "Content-Based Discovery of Multiple Structures from Episodes of Recurrent TV Programs Based on Grammatical Inference". In: *MultiMedia Modeling - 21st International Conference*. 2015, pp. 140–154.

[111] Bingqing Qu, Félicien Vallet, Jean Carrive, and Guillaume Gravier. "Content-based inference of hierarchical structural grammar for recurrent TV programs using multiple sequence alignment". In: *IEEE International Conference on Multimedia and Expo*. 2014, pp. 1–6.

[112] François Rastier. "Sémantique interprétative". In: *Presses universitaires de France* (1987).

[113] Tanya Reinhart. "Pragmatics and Linguistics: An Analysis of Sentence Topics". In: *Philosophica* 27 (1981).

[114] Jeffrey C. Reynar. "An automatic method of finding topic boundaries". In: *32nd Annual Meeting on Association for Computational Linguistics*. 1994, pp. 331–333.

[115] Royston M. Roberts. *Serendipity: Accidental Discoveries in Science*. 1989.

[116] Stephen E Robertson and Steve Walker. "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval". In: *Conf. on Research and Development in Information Retrieval*. 1994.

[117] Ork de Rooij and Marcel Worring. "Browsing Video Along Multiple Threads". In: *IEEE Transactions on Multimedia* 12.2 (2010), pp. 121–130.

[118] Anthony Rousseau, Paul Deléglise, and Yannick Estève. "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2014, pp. 3935–3939.

[119] Mathilde Sahuguet, Benoît Huet, Barbora Červenková, Evlampios Apostolidis, Vasileios Mezaris, Daniel Stein, Stefan Eickeler, Jose Luis Redondo Garcia, and Lukáš Pikora. "LinkedTV at MediaEval2013 Search and Hyperlinking Task". In: *Working Notes Proceedings of the MediaEval Workshop*. 2013.

[120] M. De Santo, P. Foggia, C. Sansone, G. Percannella, and M. Vento. "An Unsupervised Algorithm for Anchor Shot Detection". In: *Pattern Recognition, International Conference on* 2 (2006), pp. 1238–1241.

[121] Avik Sarkar, Paul H. Garthwaite, and Anne De Roeck. "A Bayesian Mixture Model for Term Re-occurrence and Burstiness". In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. CONLL '05. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 48–55.

[122] Avik Sarkar, Anne De Roeck, and Paul Garthwaite. "Team re-occurrence measures for analyzing style". In: *Proceedings of the SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access.* Ed. by S. Argamon, J. Karlgren, and J.G. Shanahan. ACM Press, 2005, pp. 28–36.

[123] Kim Schouten, Robin Aly, and R. Ordelman. "Searching and Hyperlinking using Word Importance Segment Boundaries in MediaEval 2013". In: *Working Notes Proc. of the MediaEval Workshop*. 2013.

[124] Anca Simon, Guillaume Gravier, and Pascale Sébillot. "IRISA at MediaEval 2015: Search and Anchoring in Video Archives". In: *Proceedings of the MediaEval 2015 Workshop*. 2015.

[125] Anca Simon, Guillaume Gravier, and Pascale Sébillot. "Leveraging lexical cohesion and disruption for topic segmentation". In: *Proceedings of Empirical Methods in Natural Language Processing*. 2013, pp. 1314–1324.

[126] Anca Simon, Guillaume Gravier, and Pascale Sébillot. "Un modèle segmental probabiliste combinant cohésion lexicale et rupture lexicale pour la segmentation thématique". In: *20e conférence Traitement Automatique des Langues Naturelles*. 2013, pp. 202–214.

[127] Anca Simon, Pascale Sébillot, and Guillaume Gravier. "Hierarchical topic segmentation of TV shows automatic transcripts". MA thesis. INSA de Rennes, 2012.

[128] Anca-Roxana Simon, Guillaume Gravier, Pascale Sébillot, and Marie-Francine Moens. "IRISA and KUL at MediaEval 2014: Search and Hyperlinking Task". In: *Working Notes Proc. of the MediaEval Workshop*. 2014.

[129] Laurianne Sitbon and Patrice Bellot. "Topic Segmentation Using Weighted Lexical Links (WLL)". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2007, pp. 737–738.

[130] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. "The MediaEval 2015 Affective Impact of Movies Task". In: *Working Notes Proc. of the MediaEval Workshop*. 2015.

[131] Malcolm Slaney and Dulce Ponceleon. "Hierarchical segmentation : Finding changes in a text signal". In: *1st International Conference of the Society for Industrial and Applied Mathematics-Text Mining Workshop*. 2001, pp. 6–13.

[132] Mark Steyvers and Tom Griffiths. "Probabilistic topic models". In: *Handbook of Latent Semantic Analysis* 427.7 (2007), pp. 424–440.

[133] Nicola Stokes, Joe Carthy, and Alan F. Smeaton. "Segmenting broadcast news streams using lexical chains". In: *1st Starting AI Researchers Symposium*. 2002, pp. 145–154.

[134] Tao Sun, Ming Zhang, and Qiaozhu Mei. "Unexpected Relevance: An Empirical Study of Serendipity in Retweets". In: *International AAAI Conference on Web and Social Media*. 2013.

[135] T. Tommasi, R. B. N. Aly, K. McGuinness, K. Chatfield, and et al. "Beyond metadata: searching your archive based on its audio-visual content". In: *International Broadcasting Convention*. 2014.

[136] Masao Utiyama and Hitoshi Isahara. "A statistical model for domain-independent text segmentation". In: *39th Annual Meeting on the Association for Computational Linguistics*. 2001, pp. 499–506.

[137] Enric Vallduvi. *The Informational Component*. Tech. rep. 1992.

[138] Carles Ventura, Marcel Tella-Amo, and Xavier Girói Nieto. "UPC at MediaEval 2013 Hyperlinking Task". In: *Working Notes Proc. of the MediaEval Workshop*. 2013.

[139] Raynor Vliegendhart, Cynthia C. S. Liem, and Martha Larson. "Exploring microblogs activity for the prediction of hyperlink anchors in television broadcasts". In: *Proceedings of the MediaEval 2015 Workshop*. 2015.

[140] Kenneth H. Walker, Dallas W. Hall, and Willis J. Hurst. *Clinical Methods: The History, Physical, and Laboratory Examinations*. Butterworths, 1990.

[141] R. Wilkinson and A. Smeaton. "Automatic link generation". In: *ACM Computing Surveys* 31.4 (1999).

[142] Xiao Wu, Chong-Wah Ngo, and Qing Li. "Threading and autodocumenting news videos: a promising solution to rapidly browse news topics". In: *IEEE Signal Processing Magazine* 23.2 (2006), pp. 59–68.

[143] Hua Xian-Sheng and Wang Meng. "Video Content Structure". In: (2009), pp. 3281–3286.

[144] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. "Structure analysis of soccer video with domain knowledge and hidden Markov models". In: *Pattern Recognition Letters* 25.7 (2004), pp. 767 –775.

[145] Yaakov Yaari. "Segmentation of expository texts by hierarchical agglomerative clustering". In: *In Proceedings of the 2nd International Conference on the Recent Advances in Natural Language Processing* (1997).

[146] J. Yamron, I. Carp, L. Gillick, S. Lowe, and van Mulbregt P. "A hidden Markov model approach to text segmentation and event tracking". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1998, pp. 333–336.

[147] Yiming Yang, Tom Ault, Thomas Pierce, and Charles W. Lattimer. "Improving text categorization methods for event tracking". In: *23rd ACM International Conference on Research and Development in Information Retrieval*. 2000, pp. 65–72.

[148] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. "Auralist: Introducing Serendipity into Music Recommendation". In: *Proceedings of the 5th ACM Conference on Web Search and Data Mining*. 2012.

Appendix A

Kleinberg's algorithm

In this Appendix we detail Kleinberg's algorithm [82] for burst detection. Kleinberg originally applied the algorithm to model a stream of email messages in order to identify "bursts of activity" in the topics of the messages. He proposes a formal approach for modeling such "bursts" using an infinite-state automaton A, which at any point in time can be in any of its states and emits messages at different rates depending on its state. A cost is associated to each state transition and the goal is to find the optimal state sequence that minimizes the cost. A representation of the infinite-state model is given in Figure A.1.

The "base state" q_0 of the automaton has an associated exponential density function f_0 with rate $\alpha_0 = g^{-1} = n/T$, where g is the gap size between $n+1$ messages that arrive over a period of time of length T. This is consistent with uniform message arrivals (i.e., messages spaced evenly over time interval T). For each state q_i, $i > 0$, an exponential density function f_i is associated having rate $\alpha_i = g^{-1}s^i$, where $s > 1$ is a scaling parameter. In other words, the infinite state sequence $q_0, q_1, ...$ models inter-arrival gaps that decrease geometrically from g. For every i and j there is a cost $\tau(i,j)$ associated with the state transition from q_i to q_j. The cost function is defined so that the cost of moving from a lower-intensity burst state to a higher-intensity one is proportional to the number of intervening states, but the cost is 0 to end a higher-intensity burst and drop down to a lower-intensity one. Thus, moving from q_i to q_j, when $j > i$, has a cost of $(j-i)\gamma ln(n)$ and when $j < i$ the cost is 0. Given a sequence of positive gaps $x = (x_1, x_2, .., x_n)$ between message arrivals, the goal is to find a state sequence $q = (q_{i_1}, ...q_{i_n})$ that minimizes the cost function:

$$c(q|x) = (\sum_{t=0}^{n-1} \tau(i_t, i_{t+1})) + (\sum_{t=1}^{n} -\ln f_{i_t}(x_t))) \ ,$$

where $i_0 = 0$ in order to start in state q_0 and $f_{i_t}(x_t) = \alpha_{i_t} e^{-\alpha_{i_t} x_t}$ is the density function according to which messages are being emitted independently. Minimizing the first term is consistent with having few state transitions. Minimizing the second term is consistent with passing through states whose rates agree closely with the inter-arrival gaps. To find the

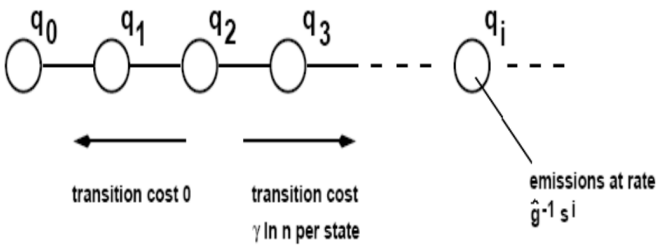


Figure A.1: The infinite-state automaton $A^*_{s,\gamma}$; in state q_i, messages are emitted at a spacing in time that is distributed according to $f(x) = \alpha_i e^{-\alpha_i x}$, where $\alpha_i = g^{-1}s^i$ and γ is a parameter that controls the ease with which the automaton can change states.

optimal state sequence the author proposes to adapt the standard forward dynamic programming algorithm used for hidden Markov models.

Appendix B

Survey details

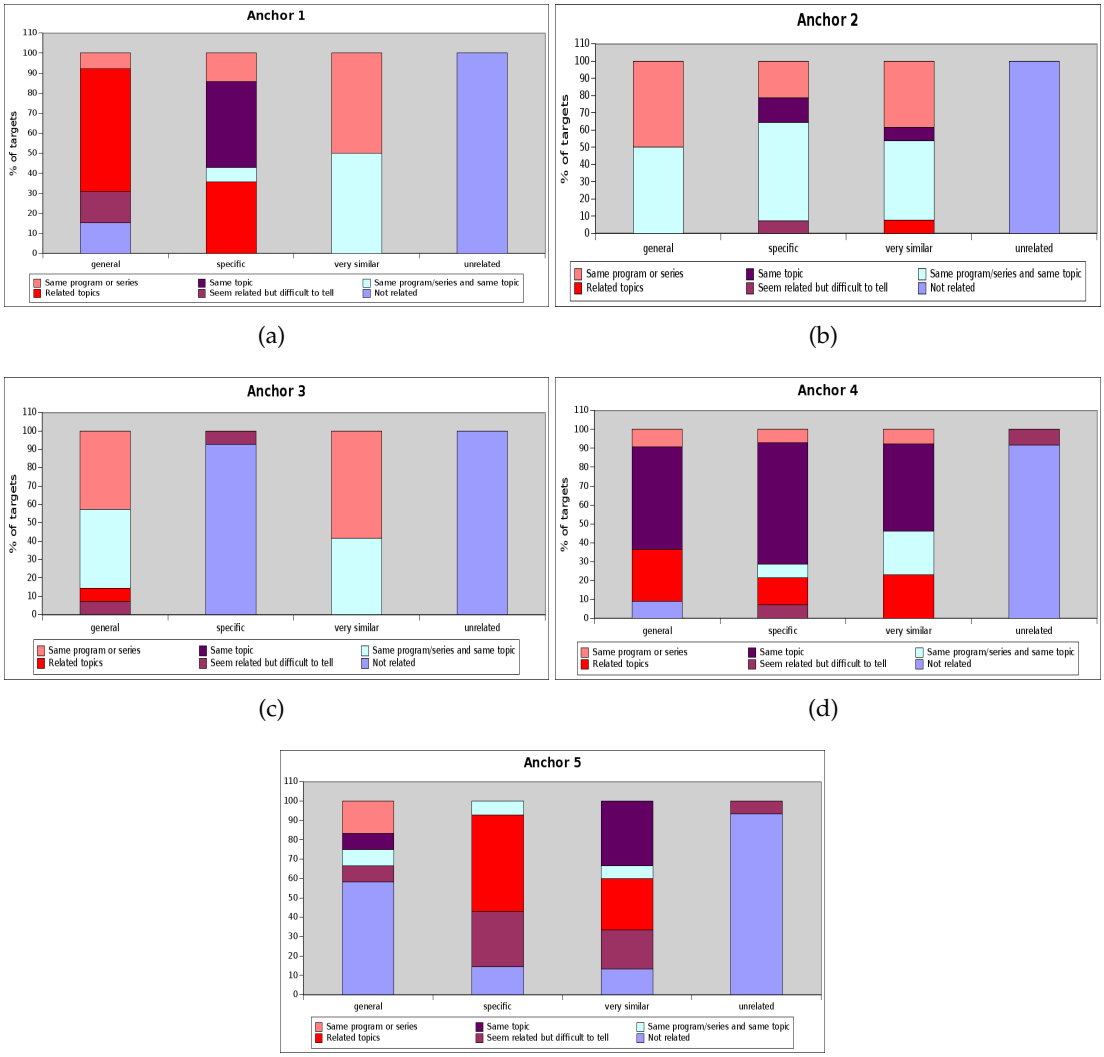


Figure B.1: Percentage of targets, for every type of target for each anchor, with any possible answer associated to the first question.

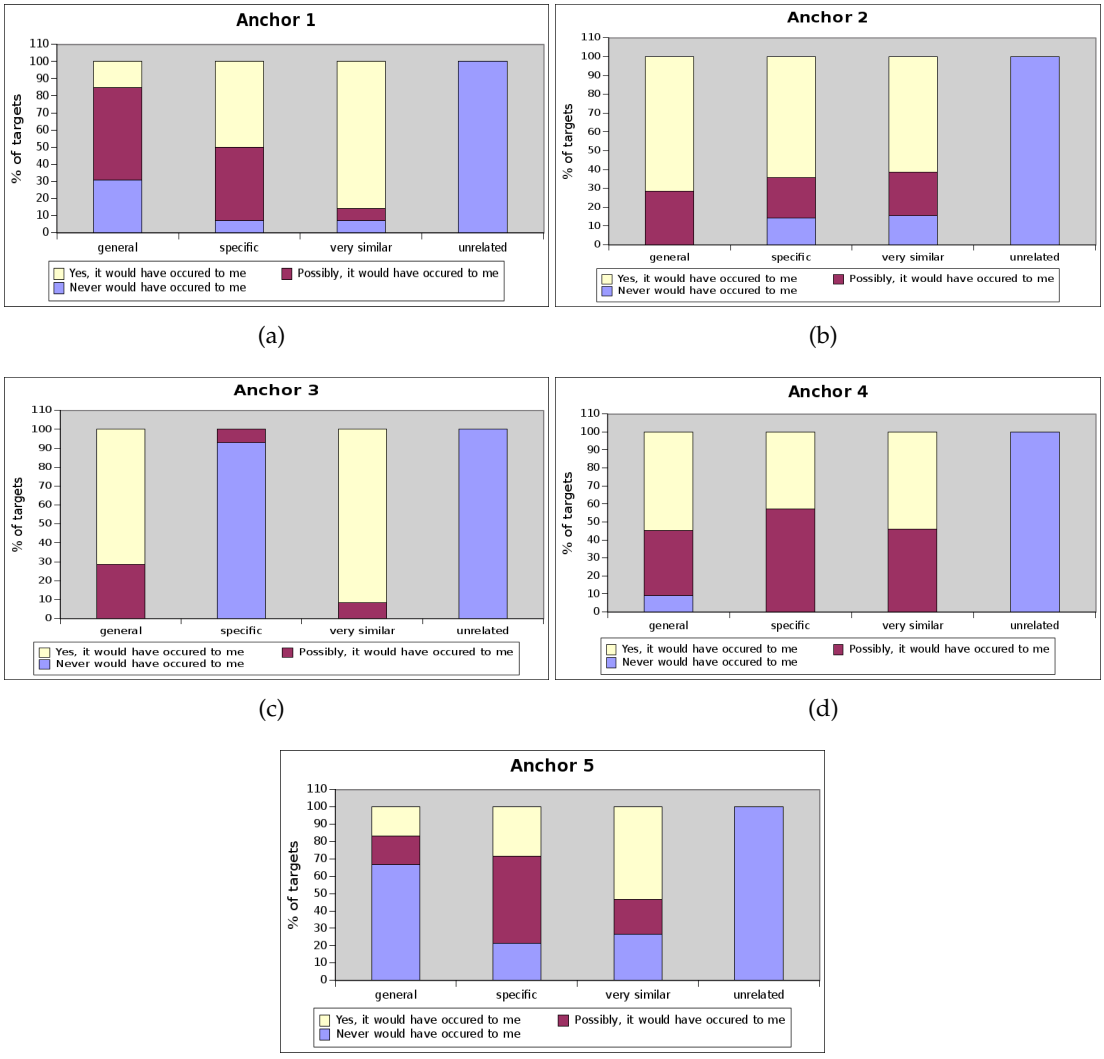


Figure B.2: Percentage of targets, for every type of target for each anchor, with any possible answer associated to the second question.

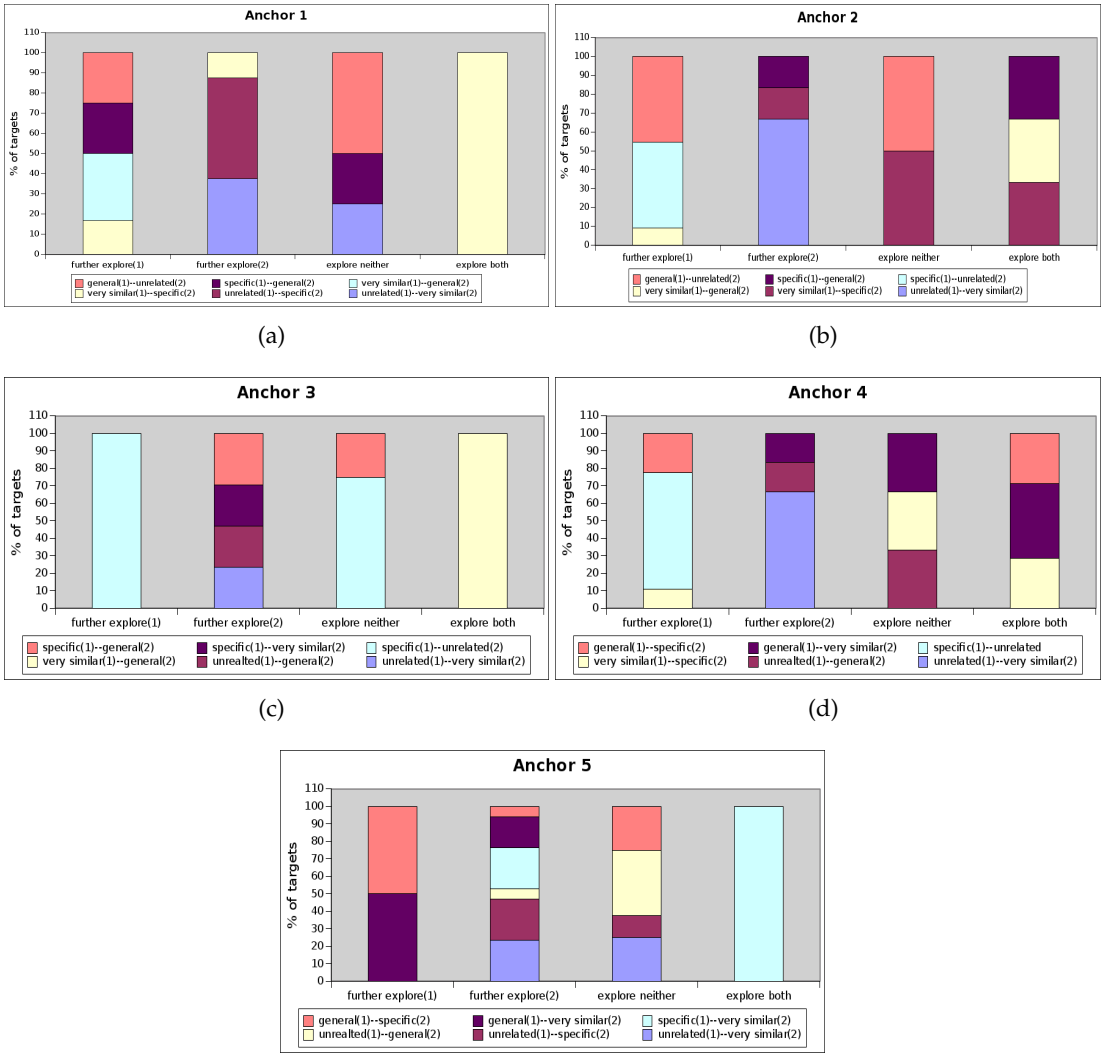


Figure B.3: Percentage of targets, for every type of target for each anchor, with any possible answer associated to the third question.