Détection de dérivation de texte

Fabien POULARD

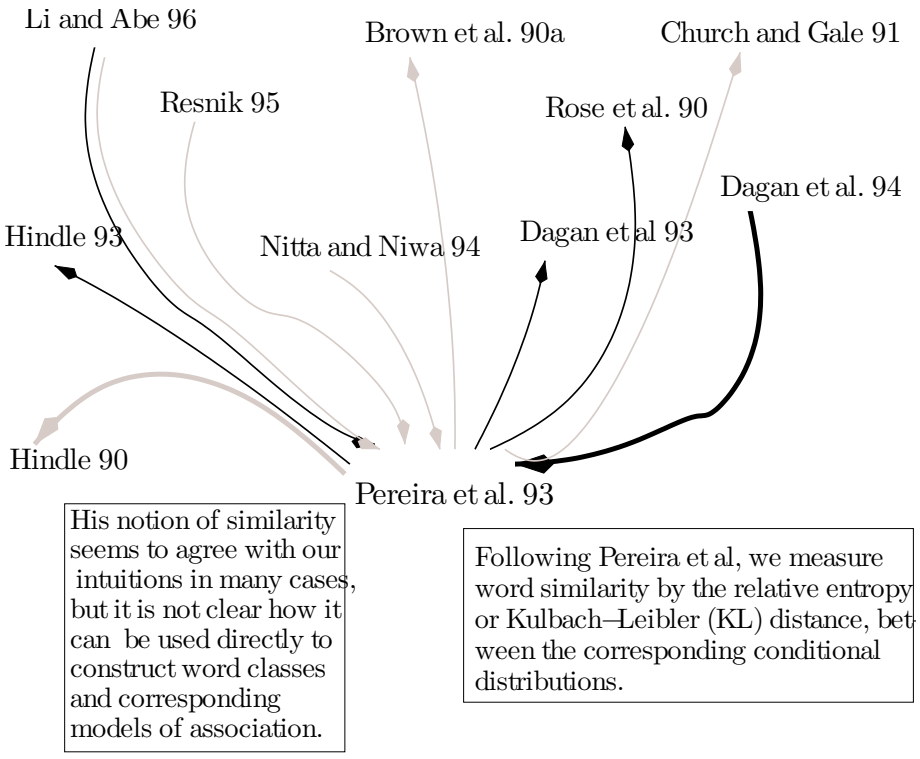
Université de Nantes — LINA (CNRS - UMR 6241)
Encadré par Béatrice DAILLE et Nicolas HERNANDEZ

24 Mars 2011

UNIVERSITÉ DE NANTES

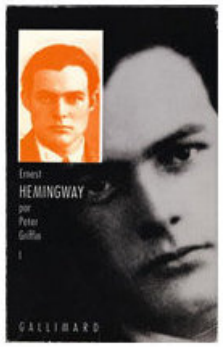
LABORATOIRE D'INFORMATIQUE
DE NANTES ATLANTIQUE

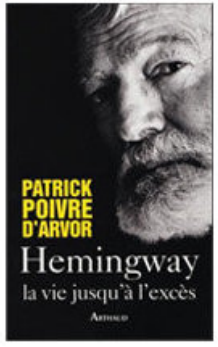
Identification de relations entre les textes



- Citations et références [Teufel et al., 2006]
- Copie illicite, contrefaçon (plagiat) [Brin et al., 1995, Heintze, 1996]
- Évolution d'un flux d'information [Metzler et al., 2005]
- Plagiat dans un contexte académique [Lyon et al., 2006]
- Révisions [Bourdaillet, 2007]

Identification de relations entre les textes





L'original

Hemingway, au fil de sa jeunesse, par Peter Griffin (Gallimard 1989)

Page 18

il revint à Dyersville. Dès qu'il le put, il s'engagea dans la First Iowa Cavalry et acheta un bon cheval. Grâce au cheval il fut promu caporal.

Page 19

A quatorze ans, Grace contracta la chorée, la danse de Saint-Gui. Durant ses six mois de convalescence elle grandit d'environ quinze centimètres ; aucun de ses vêtements ne lui allait plus et sa mère, qui mesurait quant à elle un mètre cinquante, s'alarma.

Page 36

Quant à Ernest, il portait des culottes courtes, de longs bas noirs, des bottines et une casquette à visière. Sa sœur, avec son mètre soixante-quatorze environ, avait une demi-tête de plus que lui.

Page 69

il cou-

La copie

Hemingway, la vie jusqu'à l'excès, par Patrick Poivre d'Arvor (VERSION INITIALE)

Page 26

Il revient à Dyersville, s'engage dans la First Iowa Cavalry et, comme il possède un cheval, il sert comme caporal.

Page 27

Et voici qu'à quatorze ans, elle contracte la chorée ou danse de Saint-Gui, et grandit de quinze centimètres durant les six mois de sa convalescence, une croissance record qui fait craquer tous ses vêtements et donne des sueurs froides à sa mère qui, du haut de son mètre cinquante, domine très mal la situation.

Page 42

Ernest, pour l'heure condamné à porter des culottes courtes et des bas noirs, fait pâle figure sous sa casquette à visière, surtout à côté de sa sœur qui le domine d'une demi-tête.

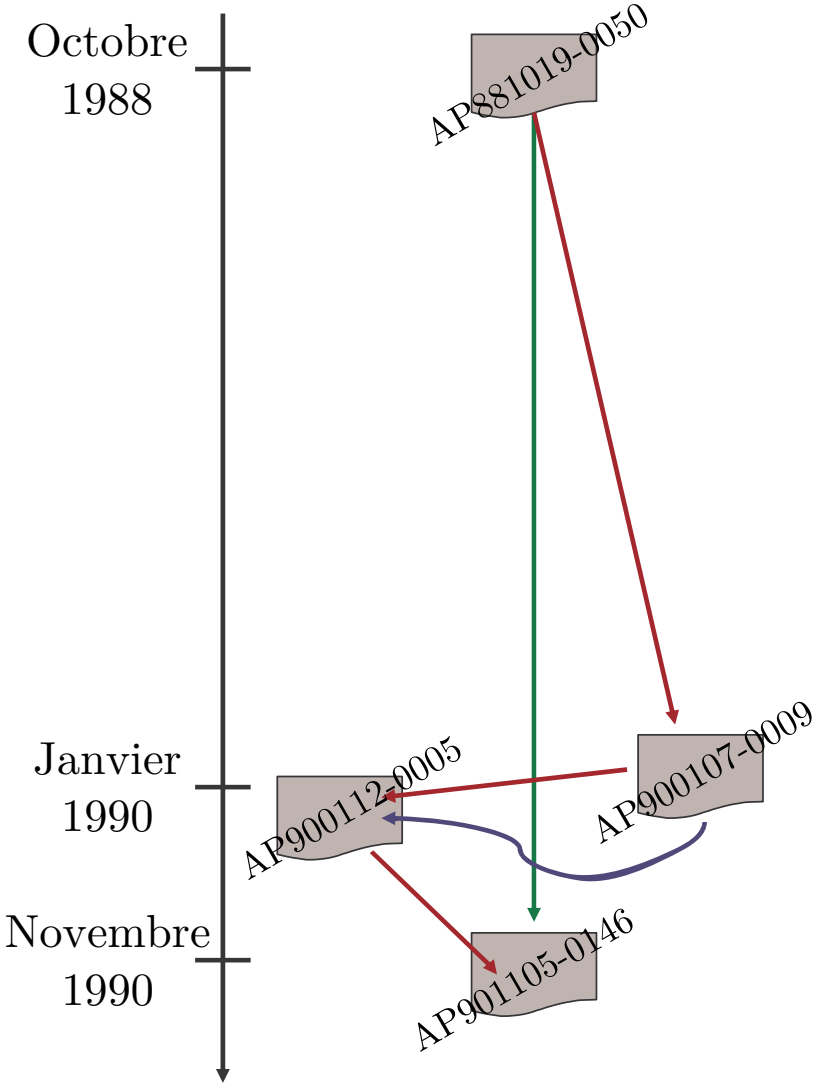
Page 57

Il explore les salles de billard, les dancings, les établisse-

© L'Express (lexpress.fr)

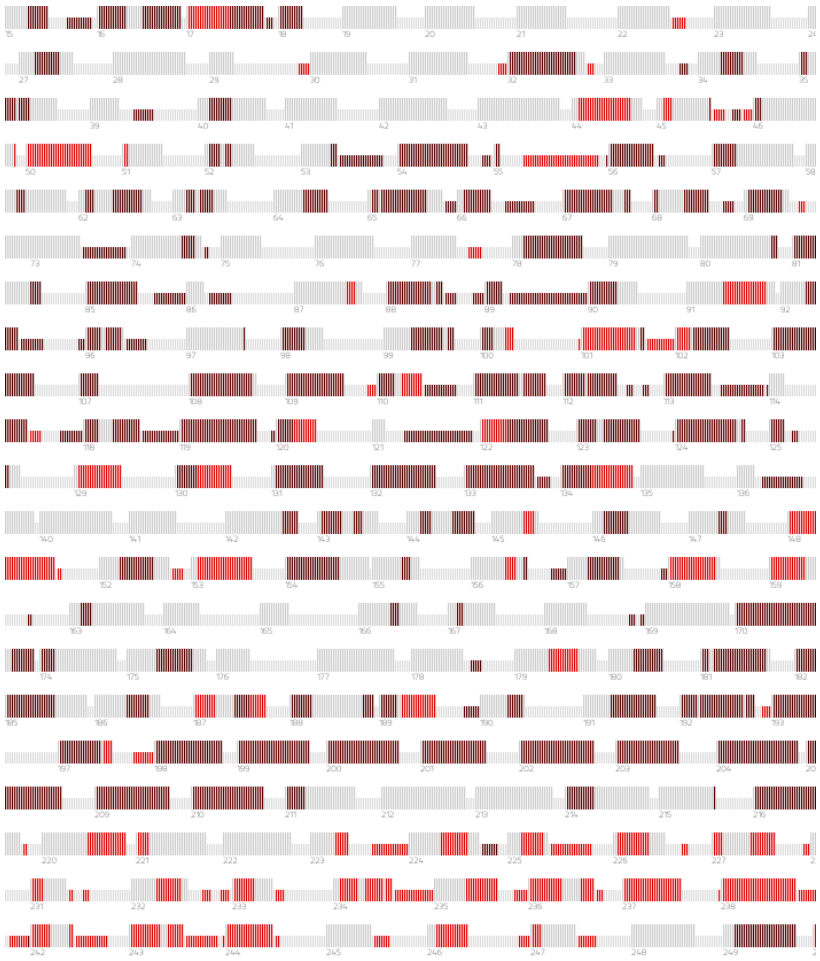
- Citations et références [Teufel et al., 2006]
- Copie illicite, contrefaçon (plagiat) [Brin et al., 1995, Heintze, 1996]
- Évolution d'un flux d'information [Metzler et al., 2005]
- Plagiat dans un contexte académique [Lyon et al., 2006]
- Révisions [Bourdaillet, 2007]

Identification de relations entre les textes



- Citations et références [Teufel et al., 2006]
- Copie illicite, contrefaçon (plagiat) [Brin et al., 1995, Heintze, 1996]
- Évolution d'un flux d'information [Metzler et al., 2005]
- Plagiat dans un contexte académique [Lyon et al., 2006]
- Révisions [Bourdaillet, 2007]

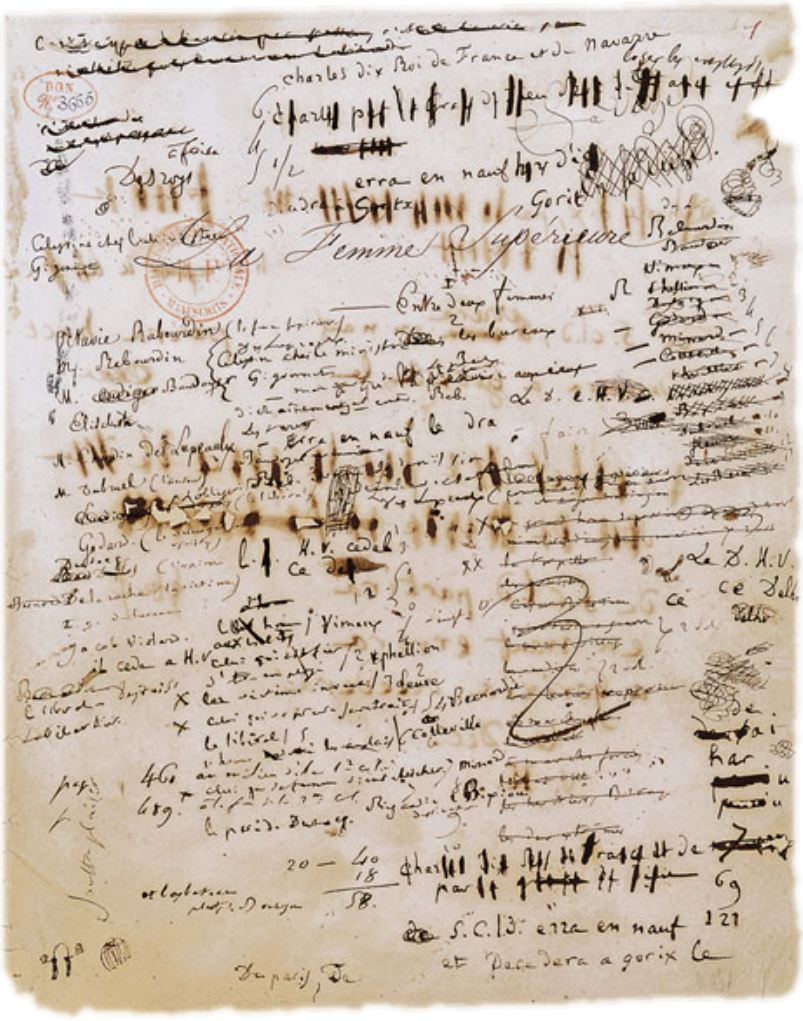
Identification de relations entre les textes



© Guttenplag-Wiki

- Citations et références [Teufel et al., 2006]
- Copie illicite, contrefaçon (plagiat) [Brin et al., 1995, Heintze, 1996]
- Évolution d'un flux d'information [Metzler et al., 2005]
- Plagiat dans un contexte académique [Lyon et al., 2006]
- Révisions [Bourdaillet, 2007]

Identification de relations entre les textes



Honoré de Balzac, La Femme supérieure © BNF

- Citations et références [Teufel et al., 2006]
- Copie illicite, contrefaçon (plagiat) [Brin et al., 1995, Heintze, 1996]
- Évolution d'un flux d'information [Metzler et al., 2005]
- Plagiat dans un contexte académique [Lyon et al., 2006]
- Révisions [Bourdaillet, 2007]

Problématique

Point commun aux relations étudiées = texte produit à partir d'un autre

- Qu'est-ce que la production d'un texte à partir d'un autre ?
- Différentes méthodes de production ? Différents produits ? Base commune ?
- Comment détecter une telle relation entre des textes ?
- Comment évaluer un système chargé de détecter ces relations ?

Problématique

Point commun aux relations étudiées = texte produit à partir d'un autre

- Qu'est-ce que la production d'un texte à partir d'un autre ?
  $\Rightarrow$ introduction de la notion de dérivation
- Différentes méthodes de production ? Différents produits ? Base commune ?
- Comment détecter une telle relation entre des textes ?
- Comment évaluer un système chargé de détecter ces relations ?

Problématique

Point commun aux relations étudiées = texte produit à partir d'un autre

- Qu'est-ce que la production d'un texte à partir d'un autre ?
  $\Rightarrow$ introduction de la notion de dérivation
- Différentes méthodes de production ? Différents produits ? Base commune ?
  $\Rightarrow$ proposition d'un modèle pour caractériser les dérivations
- Comment détecter une telle relation entre des textes ?
- Comment évaluer un système chargé de détecter ces relations ?

Problématique

Point commun aux relations étudiées = texte produit à partir d'un autre

- Qu'est-ce que la production d'un texte à partir d'un autre ?
  $\Rightarrow$ introduction de la notion de dérivation
- Différentes méthodes de production ? Différents produits ? Base commune ?
  $\Rightarrow$ proposition d'un modèle pour caractériser les dérivations
- Comment détecter une telle relation entre des textes ?
  $\Rightarrow$ couverture d'éléments singuliers et invariants
- Comment évaluer un système chargé de détecter ces relations ?

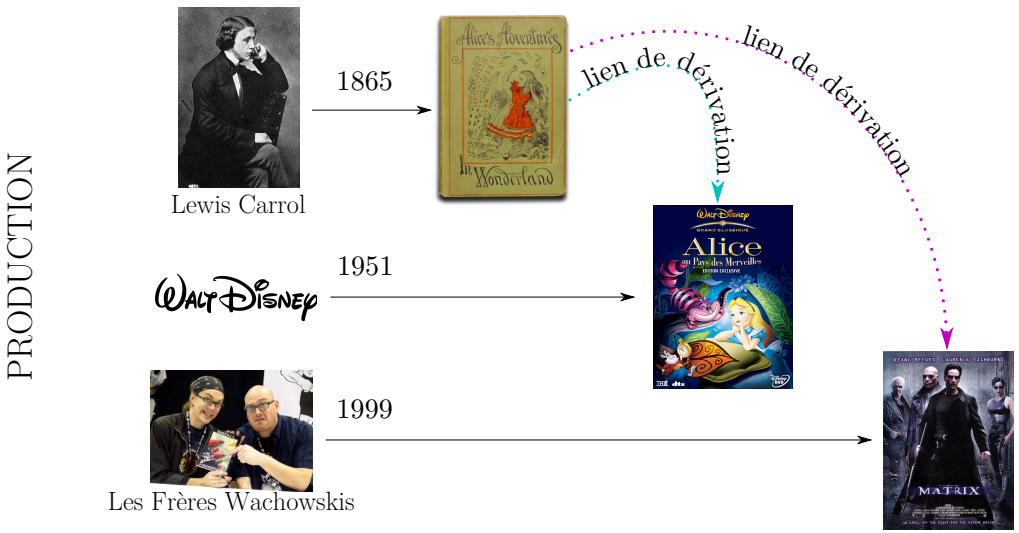
Problématique

Point commun aux relations étudiées = texte produit à partir d'un autre

- Qu'est-ce que la production d'un texte à partir d'un autre ?
  $\Rightarrow$ introduction de la notion de dérivation
- Différentes méthodes de production ? Différents produits ? Base commune ?
  $\Rightarrow$ proposition d'un modèle pour caractériser les dérivations
- Comment détecter une telle relation entre des textes ?
  $\Rightarrow$ couverture d'éléments singuliers et invariants
- Comment évaluer un système chargé de détecter ces relations ?
  $\Rightarrow$ évaluation comme un système d'aide à la décision

Cadre de nos travaux de thèse

- Détection des relations “texte A produit à partir de B”
- Recherche des textes A (dérivés) parmi une collection fermée de textes suspects
- Texte B (source) identifié
- Relations identifiées à l’échelle des documents



Sommaire

1. La dérivation de texte
2. Détecter les relations de dérivation
3. Évaluation comme des systèmes d'aide à la décision
4. Modéliser par des éléments singuliers et invariants
5. Discussion et perspectives

1 La dérivation de texte
- Cadre théorique global
- Modèle multidimensionnel du processus de dérivation

2 Détecter les relations de dérivation

3 Évaluation comme des systèmes d'aide à la décision

4 Modéliser par des éléments singuliers et invariants

5 Discussion et perspectives

Concept transversal à la littérature

Problématique abordée sous diverses appelations

- selon les degrés de modification :
  - Documents distincts "qui sont identiques ou presque" [Broder et al., 1997]
  - Duplications et presque-duplications [Shivakumar and Garcia-Molina, 1995, Yang, 2006, Bernstein et al., 2006]
- selon l'intention de l'auteur :
  - Copies de documents [Brin et al., 1995]
  - Plagiat, collusions [Lyon et al., 2001]
- caractérisant une forme particulière :
  - Réutilisation de texte (*Text reuse*) [Clough, 2003, Bendersky and Croft, 2009]
  - Versions, résumés, citations, références, transpositions de genre, traductions. . .

Concept transversal à la littérature

Problématique abordée sous diverses appelations

- selon les degrés de modification :
  - Documents distincts "qui sont identiques ou presque" [Broder et al., 1997]
  - Duplications et presque-duplications [Shivakumar and Garcia-Molina, 1995, Yang, 2006, Bernstein et al., 2006]
- selon l'intention de l'auteur :
  - Copies de documents [Brin et al., 1995]
  - Plagiat, collusions [Lyon et al., 2001]
- caractérisant une forme particulière :
  - Réutilisation de texte (*Text reuse*) [Clough, 2003, Bendersky and Croft, 2009]
  - Versions, résumés, citations, références, transpositions de genre, traductions. . .

Concept transversal à la littérature

Problématique abordée sous diverses appelations

- selon les degrés de modification :
  - Documents distincts "qui sont identiques ou presque" [Broder et al., 1997]
  - Duplications et presque-duplications [Shivakumar and Garcia-Molina, 1995, Yang, 2006, Bernstein et al., 2006]
- selon l'intention de l'auteur :
  - Copies de documents [Brin et al., 1995]
  - Plagiat, collusions [Lyon et al., 2001]
- caractérisant une forme particulière :
  - Réutilisation de texte (*Text reuse*) [Clough, 2003, Bendersky and Croft, 2009]
  - Versions, résumés, citations, références, transpositions de genre, traductions...

Problème

Problèmes voisins traités séparément $\Rightarrow$ pas de vision globale

- Réutilisabilité des méthodes ?
- Problèmes théoriques communs sous-jacents ?

Proposition

Introduction de la notion de **dérivation de texte** afin d'étudier le problème général

- Cadre théorique global cohérent avec l'existant
- Modèle multidimensionnel issue de ce cadre théorique permettant d'y repositionner les notions précédentes

Problème

Problèmes voisins traités séparément $\Rightarrow$ pas de vision globale

- Réutilisabilité des méthodes ?
- Problèmes théoriques communs sous-jacents ?

Proposition

Introduction de la notion de **dérivation de texte** afin d'étudier le problème général

- Cadre théorique global cohérent avec l'existant
- Modèle multidimensionnel issue de ce cadre théorique permettant d'y repositionner les notions précédentes

Un cadre théorique global cohérent avec l'existant

Idée générale

Production d'un nouveau texte à partir d'un ou plusieurs textes préexistant

Condition de dérivation de texte

Si le retrait de T_s lors du processus de production de T_d résulte en un texte $T_{d'}$ différent de T_d alors T_d résulte d'un processus de dérivation impliquant T_s.

Un cadre théorique global cohérent avec l'existant

Idée générale

Production d'un nouveau texte à partir d'un ou plusieurs textes préexistant

Condition de dérivation de texte

Si le retrait de T_s lors du processus de production de T_d résulte en un texte $T_{d'}$ différent de T_d alors T_d résulte d'un processus de dérivation impliquant T_s.

Un cadre théorique global cohérent avec l'existant

Relation de dérivation $\mathbf{R_D}$

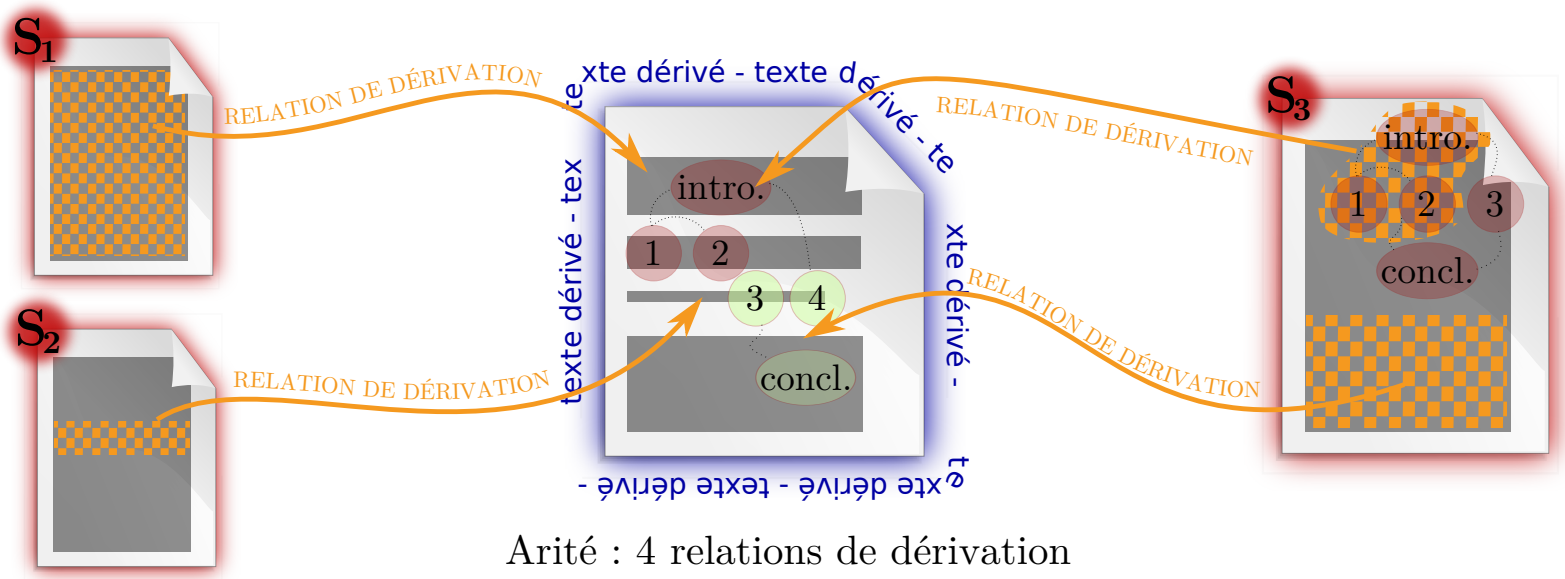
Soient T_d le *texte dérivé* et T_s un *texte source*
T_d dérive de $T_s \Leftrightarrow (T_s, T_d)$ appartient à la relation de dérivation $\mathbf{R_D}$ notée $T_s \, \mathbf{R_D} \, T_d$

Processus de dérivation de $T_d = \bigcup (T_s, T_d) \; \forall T_s$ tel que $T_s \, \mathbf{R_D} \, T_d$

Place des notions de la littérature dans ce cadre théorique ?

- Proposition d'un modèle multidimensionnel
- Caractérisation des relations de dérivation composant un processus
- Chaque notion de la littérature = combinaison particulière de caractéristiques
- 1 seule caractérisation par relation de dérivation

Dérivation = processus impliquant plusieurs relations de dérivation (arité)



Arité : 4 relations de dérivation

Caractérisation des relations de dérivation

- Nature des éléments dérivés depuis la source (textuelle, sémantique, discursive et stylistique)
- Granularité des éléments dérivés du texte source et ce qu'ils représentent dans le texte dérivé [Seo and Croft, 2008]
- Paternité des textes (même auteur ou auteurs différents)
- Intention de l'auteur du texte dérivé (tromper ou reconnaître)
- Similarité du contenu entre les éléments source et dérivé (identique, presque-identique, différent)
- Intégration des séquences textuelles (verbatim, syntaxique, paraphrase, contraction/dilatation, adaptation de genre et traduction)

Caractérisation des relations de dérivation

- Nature des éléments dérivés depuis la source (textuelle, sémantique, discursive et stylistique)
- Granularité des éléments dérivés du texte source et ce qu'ils représentent dans le texte dérivé [Seo and Croft, 2008]
- Paternité des textes (même auteur ou auteurs différents)
- Intention de l'auteur du texte dérivé (tromper ou reconnaître)
- Similarité du contenu entre les éléments source et dérivé (identique, presque-identique, différent)
- Intégration des séquences textuelles (verbatim, syntaxique, paraphrase, contraction/dilatation, adaptation de genre et traduction)

Caractérisation des relations de dérivation

- Nature des éléments dérivés depuis la source (textuelle, sémantique, discursive et stylistique)
- Granularité des éléments dérivés du texte source et ce qu'ils représentent dans le texte dérivé [Seo and Croft, 2008]
- Paternité des textes (même auteur ou auteurs différents)
- Intention de l'auteur du texte dérivé (tromper ou reconnaître)
- Similarité du contenu entre les éléments source et dérivé (identique, presque-identique, différent)
- Intégration des séquences textuelles (verbatim, syntaxique, paraphrase, contraction/dilatation, adaptation de genre et traduction)

Caractérisation des relations de dérivation

- Nature des éléments dérivés depuis la source (textuelle, sémantique, discursive et stylistique)
- Granularité des éléments dérivés du texte source et ce qu'ils représentent dans le texte dérivé [Seo and Croft, 2008]
- Paternité des textes (même auteur ou auteurs différents)
- Intention de l'auteur du texte dérivé (tromper ou reconnaître)
- Similarité du contenu entre les éléments source et dérivé (identique, presque-identique, différent)
- Intégration des séquences textuelles (verbatim, syntaxique, paraphrase, contraction/dilatation, adaptation de genre et traduction)

Caractérisation des relations de dérivation

- Nature des éléments dérivés depuis la source (textuelle, sémantique, discursive et stylistique)
- Granularité des éléments dérivés du texte source et ce qu'ils représentent dans le texte dérivé [Seo and Croft, 2008]
- Paternité des textes (même auteur ou auteurs différents)
- Intention de l'auteur du texte dérivé (tromper ou reconnaître)
- Similarité du contenu entre les éléments source et dérivé (identique, presque-identique, différent)
- Intégration des séquences textuelles (verbatim, syntaxique, paraphrase, contraction/dilatation, adaptation de genre et traduction)

Caractérisation des relations de dérivation

- Nature des éléments dérivés depuis la source (textuelle, sémantique, discursive et stylistique)
- Granularité des éléments dérivés du texte source et ce qu'ils représentent dans le texte dérivé [Seo and Croft, 2008]
- Paternité des textes (même auteur ou auteurs différents)
- Intention de l'auteur du texte dérivé (tromper ou reconnaître)
- Similarité du contenu entre les éléments source et dérivé (identique, presque-identique, différent)
- Intégration des séquences textuelles (verbatim, syntaxique, paraphrase, contraction/dilatation, adaptation de genre et traduction)

1 La dérivation de texte

2 Détecter les relations de dérivation
- Problématique
- Approches existantes
- Approche par couverture de texte

3 Évaluation comme des systèmes d'aide à la décision

4 Modéliser par des éléments singuliers et invariants

5 Discussion et perspectives

Problématique

Détection de dérivation = problème difficile

- Nombreuses variations dans le processus
- Implique des transformations variées à tous les niveaux du texte : paraphrases, lexicales, syntaxiques, typographiques . . .

Texte source (AFP)

Et elle souligne que la France va plus loin que le réglement européen, qui n'envisage pas le recueil de l'empreinte de huit doigts, mais de deux, ni la conservation des données en base centrale, qui comporte « des risques d'atteintes graves à la vie privée et aux libertés individuelles » et qui « ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient ».

Texte dérivé (Le Monde)

La CNIL note que la France va plus loin que la réglementation européenne, qui ne prévoit pas le recueil de l'empreinte de huit doigts mais de deux, ni la conservation des données en base centrale. Le dispositif français comporte " des risques d'atteintes graves à la vie privée et aux libertés individuelles " et il " ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient".

Problématique

Détection de dérivation = problème difficile

- Nombreuses variations dans le processus
- Implique des transformations variées à tous les niveaux du texte : paraphrases, lexicales, syntaxiques, typographiques . . .

Texte source (AFP)

Et elle souligne que la France va plus loin que le réglement européen, qui n'envisage pas le recueil de l'empreinte de huit doigts, mais de deux, ni la conservation des données en base centrale, qui comporte « des risques d'atteintes graves à la vie privée et aux libertés individuelles » et qui « ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient ».

Texte dérivé (Le Monde)

La CNIL note que la France va plus loin que la réglementation européenne, qui ne prévoit pas le recueil de l'empreinte de huit doigts mais de deux, ni la conservation des données en base centrale. Le dispositif français comporte " des risques d'atteintes graves à la vie privée et aux libertés individuelles " et il " ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient".

Problématique

Détection de dérivation = problème difficile

- Nombreuses variations dans le processus
- Implique des transformations variées à tous les niveaux du texte : paraphrases, lexicales, syntaxiques, typographiques . . .

Texte source (AFP)

Et elle souligne que la France va plus loin que le réglement européen, qui n'envisage pas le recueil de l'empreinte de huit doigts, mais de deux, ni la conservation des données en base centrale, qui comporte « des risques d'atteintes graves à la vie privée et aux libertés individuelles » et qui « ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient ».

Texte dérivé (Le Monde)

La CNIL note que la France va plus loin que la réglementation européenne, qui ne prévoit pas le recueil de l'empreinte de huit doigts mais de deux, ni la conservation des données en base centrale. Le dispositif français comporte " des risques d'atteintes graves à la vie privée et aux libertés individuelles " et il " ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient".

Problématique

Détection de dérivation = problème difficile

- Nombreuses variations dans le processus
- Implique des transformations variées à tous les niveaux du texte : paraphrases, lexicales, syntaxiques, typographiques . . .

Texte source (AFP)

Et elle souligne que la France va plus loin que le réglement européen, qui n'envisage pas le recueil de l'empreinte de huit doigts, mais de deux, ni la conservation des données en base centrale, qui comporte « des risques d'atteintes graves à la vie privée et aux libertés individuelles » et qui « ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient ».

Texte dérivé (Le Monde)

La CNIL note que la France va plus loin que la réglementation européenne, qui ne prévoit pas le recueil de l'empreinte de huit doigts mais de deux, ni la conservation des données en base centrale. Le dispositif français comporte " des risques d'atteintes graves à la vie privée et aux libertés individuelles " et il " ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient".

Problématique

Détection de dérivation = problème difficile

- Nombreuses variations dans le processus
- Implique des transformations variées à tous les niveaux du texte : paraphrases, lexicales, syntaxiques, typographiques . . .

Texte source (AFP)

Et elle souligne que la France va plus loin que le réglement européen, qui n'envisage pas le recueil de l'empreinte de huit doigts, mais de deux, ni la conservation des données en base centrale, qui comporte « des risques d'atteintes graves à la vie privée et aux libertés individuelles » et qui « ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient ».

Texte dérivé (Le Monde)

La CNIL note que la France va plus loin que la réglementation européenne, qui ne prévoit pas le recueil de l'empreinte de huit doigts mais de deux, ni la conservation des données en base centrale. Le dispositif français comporte " des risques d'atteintes graves à la vie privée et aux libertés individuelles " et il " ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient".

Problématique

Détection de dérivation = problème difficile

- Nombreuses variations dans le processus
- Implique des transformations variées à tous les niveaux du texte : paraphrases, lexicales, syntaxiques, typographiques . . .

Texte source (AFP)

Et elle souligne que la France va plus loin que le réglement européen, qui n'envisage pas le recueil de l'empreinte de huit doigts, mais de deux, ni la conservation des données en base centrale, qui comporte « des risques d'atteintes graves à la vie privée et aux libertés individuelles » et qui « ne peut être admis que dans la mesure où des exigences en matière de sécurité ou d'ordre public le justifient ».

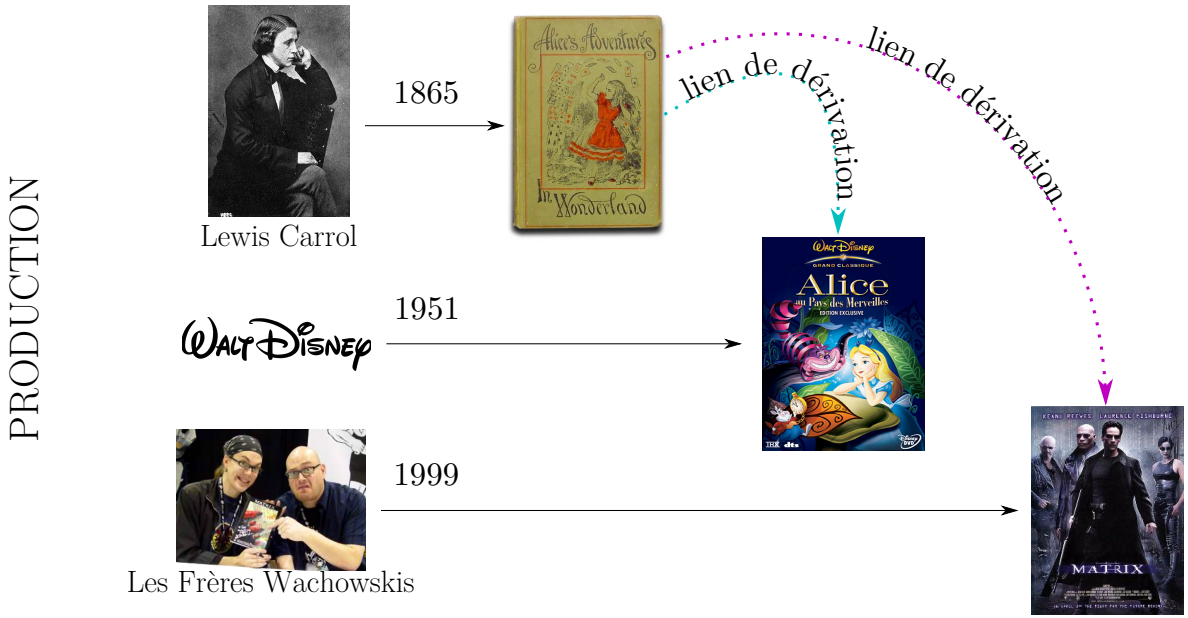
Texte dérivé traduit

And she stressed that France goes beyond the European regulation, which does not consider the collection of eight fingerprints, but only two, neither the recording of the data in a central database, which includes "serious risks of privacy and individual liberties damages" which "can only be assumed in the extent of security or public order requirements."

Détecter les relations de dérivation

Relations de dérivation entre sources et dérivés :

- Connus de l'auteur
- Inconnus des lecteurs



Détecter les relations de dérivation

Relations de dérivation entre sources et dérivés :

- Connus de l'auteur
- Inconnus des lecteurs

Détecter les relations de dérivation

Relations de dérivation entre sources et dérivés :

- Connus de l'auteur
- Inconnus des lecteurs

Deux types d'approche :

- Détection intrinsèque : rechercher des indices au sein du texte
- Détection extrinsèque : confronter un texte à des textes suspects

Deux types d'approche :

- Détection intrinsèque : rechercher des indices au sein du texte
- Détection extrinsèque : confronter un texte à des textes suspects

Détection extrinsèque

Détection extrinsèque : comparer le texte à d'autres

- Alignement de sous-chaînes
- Similarités de mots-clés
- Couverture de texte

Détection extrinsèque

Détection extrinsèque : comparer le texte à d'autres

- Alignement de sous-chaînes
  - distances d'édition [Yang, 2006]
  - recherche de sous-chaînes communes (*greedy string tiling*) [Wise, 1996]
  - alignement de passages [Monostori et al., 2001, Bourdaillet, 2007]
- Similarités de mots-clés
- Couverture de texte

Détection extrinsèque

Détection extrinsèque : comparer le texte à d'autres

- Alignement de sous-chaînes
  - distances d'édition [Yang, 2006]
  - recherche de sous-chaînes communes (*greedy string tiling*) [Wise, 1996]
  - alignement de passages [Monostori et al., 2001, Bourdaillet, 2007]

  Approches coûteuses et sensibles aux réécritures
- Similarités de mots-clés
- Couverture de texte

Détection extrinsèque

Détection extrinsèque : comparer le texte à d'autres

- Alignement de sous-chaînes
- Similarités de mots-clés
  - modèle vectoriel classique [Chowdhury et al., 2002, Özlem Uzuner and Davis, 2003, Clough, 2003, Hose, 2003]
  - modèle vectoriel à fréquences relatives [Shivakumar and Garcia-Molina, 1995, Hoad and Zobel, 2002, Metzler et al., 2005]
  - signature floue (*fuzzy-fingerprint*) [Stein, 2005]
- Couverture de texte

Détection extrinsèque

Détection extrinsèque : comparer le texte à d'autres

- Alignement de sous-chaînes
- Similarités de mots-clés
  - modèle vectoriel classique [Chowdhury et al., 2002, Özlem Uzuner and Davis, 2003, Clough, 2003, Hose, 2003]
  - modèle vectoriel à fréquences relatives [Shivakumar and Garcia-Molina, 1995, Hoad and Zobel, 2002, Metzler et al., 2005]
  - signature floue (*fuzzy-fingerprint*) [Stein, 2005]

  Mesure d'une similarité globale entre les textes
- Couverture de texte

Détection extrinsèque

Détection extrinsèque : comparer le texte à d'autres

- Alignement de sous-chaînes
- Similarités de mots-clés
- Couverture de texte
  - Calculer le taux de recouvrement entre deux textes [Brin et al., 1995, Heintze, 1996, Broder et al., 1997]

Détection extrinsèque

Détection extrinsèque : comparer le texte à d'autres

- Alignement de sous-chaînes
- Similarités de mots-clés
- Couverture de texte
  - Calculer le taux de recouvrement entre deux textes [Brin et al., 1995, Heintze, 1996, Broder et al., 1997]

Coût intermédiaire

Moins sensible aux réécritures

Formalisation de l'approche proposée par [Broder, 1997]

1. Modélisation des textes = signatures
   - Découpage du texte en sous-séquences contiguës de taille fixée (n-grammes mots)
   - Regroupement de toutes ces sous-séquences en une collection (multiensemble)
2. Calcul d'un score de similarité entre les signatures
   - Mesures de similarité ancrées dans la théorie des ensembles
   - *resemblance* : $r(\Pi(d_1), \Pi(d_2)) = \frac{|\Pi(d_1) \cap \Pi(d_2)|}{|\Pi(d_1) \cup \Pi(d_2)|}$
   - *containment* : $c(\Pi(d_1), \Pi(d_2)) = \frac{|\Pi(d_1) \cap \Pi(d_2)|}{|\Pi(d_1)|}$
3. Catégorisation selon le score de similarité

Exemple : *2-shingling* sur les mots

1) Choix d'un couple de textes à tester

Texte source (Ts)

Texte suspect (Td)

Source AFP

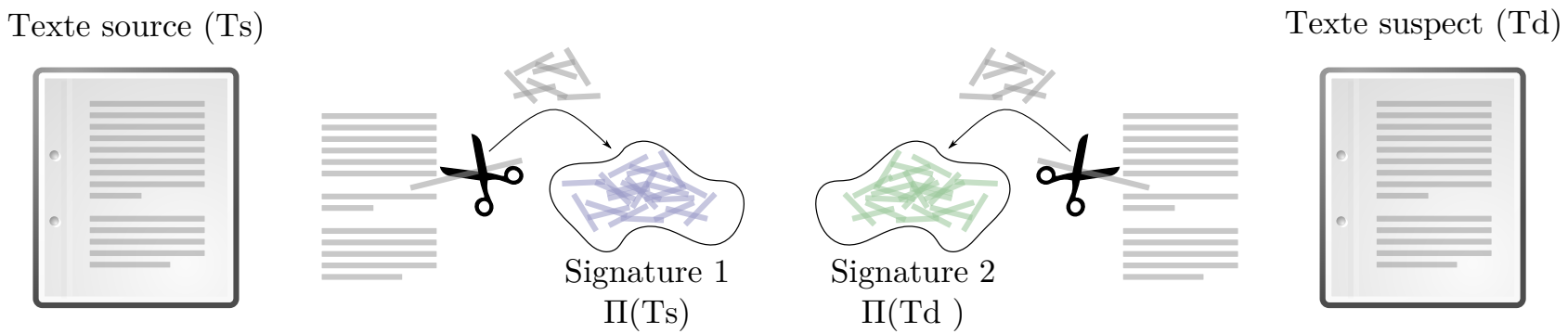
Et elle souligne que la France va plus loin que le réglement européen, qui n'envisage pas le recueil de l'empreinte de huit doigts, mais de deux, ni la conservation des données en base centrale,

Dérivé Le Monde

La CNIL note que la France va plus loin que la réglementation européenne, qui ne prévoit pas le recueil de l'empreinte de huit doigts mais de deux, ni la conservation des données en base centrale.

Exemple : *2-shingling* sur les mots

2) Calcule des signatures pour chacun des textes



Source AFP

$\Pi(T_s) = \{$ (Et, elle) (elle, souligne) (souligne, que) (que, la) (la, France) (France, va) (va, plus) (plus, loin) (loin, que) (que, le) (le, réglement) (réglement, européen) (européen, qui) (qui, n') (n', envisage) (envisage, pas) (pas, le) (le, recueil) (recueil, de) (de, l') (l', empreinte) (empreinte, de) (de, huit) (huit, doigts) (doigts, mais) (mais, de) (de, deux) (deux, ni) (ni, la) (la, conservation) (conservation, des) (des, données) (données, en) (en, base) (base,

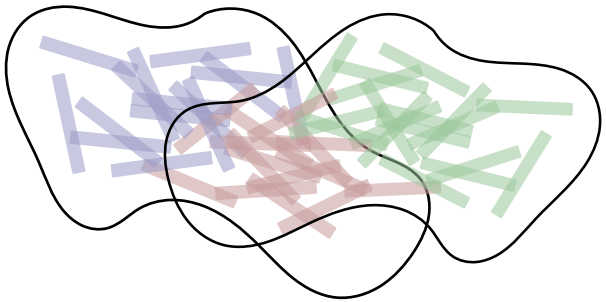
Dérivé Le Monde

$\Pi(T_d) = \{$ (La, CNIL) (CNIL, note) (note, que) (que, la) (la, France) (France, va) (va, plus) (plus, loin) (loin, que) (la, réglementation) (réglementation, européenne) (européenne, qui) (qui, ne) (ne, prévoit) (prévoit, pas) (pas, le) (le, recueil) (recueil, de) (de, l') (l', empreinte) (empreinte, de) (de, huit) (huit, doigts) (doigts, mais) (mais, de) (de, deux) (deux, ni) (ni, la) (la, conservation) (conservation, des) (des, données) (données, en) (en, base) (base,

Exemple : *2-shingling* sur les mots

3) Mesure de similarité entre les signatures (*containment*)

Signature 1



Signature 2

Source AFP

$\Pi(T_s) = \{$ (Et, elle) (elle, souligne) (souligne, que) (que, la) (la, France) (France, va) (va, plus) (plus, loin) (loin, que) (que, le) (le, réglement) (réglement, européen) (européen, qui) (qui, n') (n', envisage) (envisage, pas) (pas, le) (le, recueil) (recueil, de) (de, l') (l', empreinte) (empreinte, de) (de, huit) (huit, doigts) (doigts, mais) (mais, de) (de, deux) (deux, ni) (ni, la) (la, conservation) (conservation, des) (des, données) (données, en) (en, base) (base,

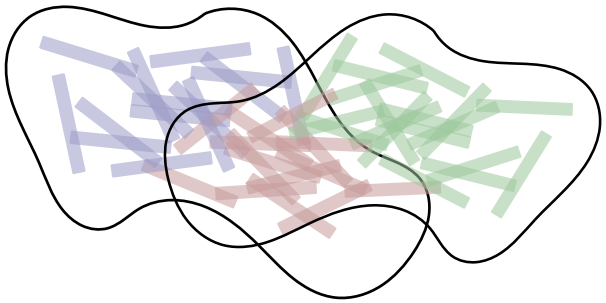
Dérivé Le Monde

$\Pi(T_d) = \{$ (La, CNIL) (CNIL, note) (note, que) (que, la) (la, France) (France, va) (va, plus) (plus, loin) (loin, que) (la, réglementation) (réglementation, européenne) (européenne, qui) (qui, ne) (ne, prévoit) (prévoit, pas) (pas, le) (le, recueil) (recueil, de) (de, l') (l', empreinte) (empreinte, de) (de, huit) (huit, doigts) (doigts, mais) (mais, de) (de, deux) (deux, ni) (ni, la) (la, conservation) (conservation, des) (des, données) (données, en) (en, base) (base,

Exemple : *2-shingling* sur les mots

3) Mesure de similarité entre les signatures (*containment*)

Signature 1



Signature 2

Source AFP vs. dérivé Le Monde

Taux d'inclusion du suspect dans la source

$$\frac{\text{\# éléments communs}}{\text{\# éléments dans la source}}$$

$$c = \frac{|\Pi(T_s) \cap \Pi(T_d)|}{|\Pi(T_s)|} = \frac{25}{35} \approx 71\%$$

Limitations par rapport à notre contexte

1. Méthode orientée classification alors que la complexité du problème nécessiterait une orientation "aide à la décision"
2. Le coût de la méthode est élevé et évolue avec le nombre de couples à comparer et la taille des documents

1 La dérivation de texte

2 Détecter les relations de dérivation

3 Évaluation comme des systèmes d'aide à la décision
- Évaluation : corpus et mesures
- Résultats de référence pour comparaison

4 Modéliser par des éléments singuliers et invariants

5 Discussion et perspectives

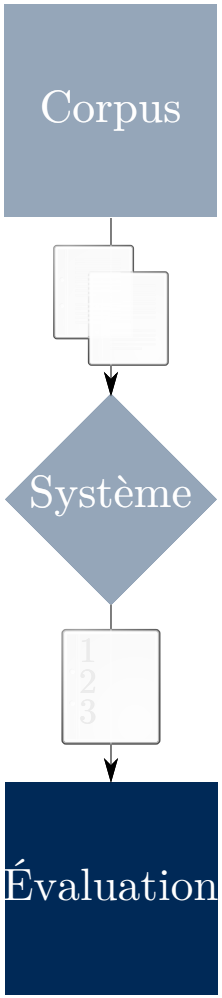
Données en entrée

- Constitution de trois corpus
- Trois formes de dérivation
  - Reprises de dépêches de presse (Piithie, projet ANR PIITHIE)
  - Révisions de documents de type articles de presse (Wikinews)
  - Plagiat artificiel de textes littéraires (PANini)
- Deux langues
  - Français (Piithie, Wikinews)
  - Anglais (PANini)
- Partition des textes de chaque corpus
  - P_1 = Textes sources
  - P_2 = Textes suspects composés de textes dérivés et de textes tiers

Système à évaluer

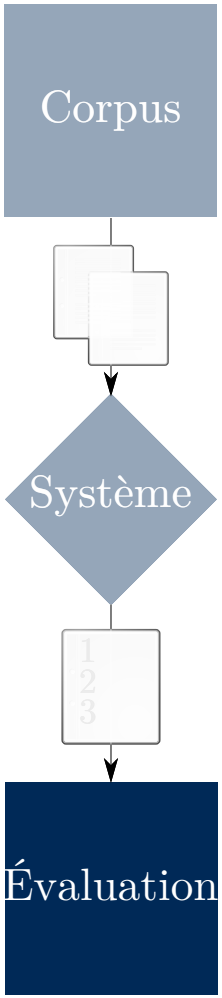
- Entrée : couples $P_1 \times P_2$
- Calcul des scores de similarité entre textes des couples
- Sortie : classement des couples par score de similarité décroissant
- Sortie idéale :
  - Haut du classement = couples correspondant à des liens de dérivation (positifs)
  - Bas du classement = couples ne correspondant pas à des liens de dérivation (négatifs)

Mesures d'évaluation selon trois critères



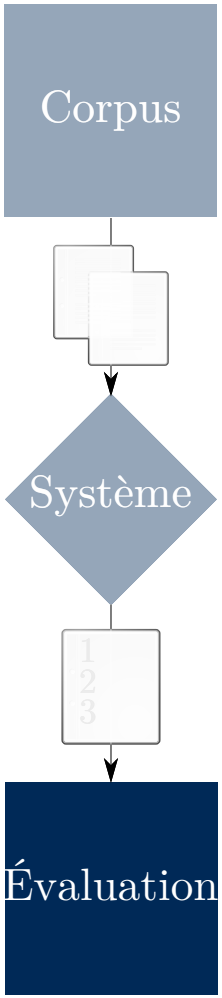
- Qualité de la classification ⇒ MAP
  - Positifs en haut et négatifs en bas
- Capacité de discrimination ⇒ SepQ
  - Écart important entre les scores des positifs et des négatifs
- Coût de mise en œuvre ⇒ Taille de la signature
  - Principalement porté par le coût de comparaison des signatures
  - Coût comparaison ⇔ Coût de la mesure de similarité
  - *containment* = Linéaire en la taille des signatures (nombre d'éléments)

Mesures d'évaluation selon trois critères



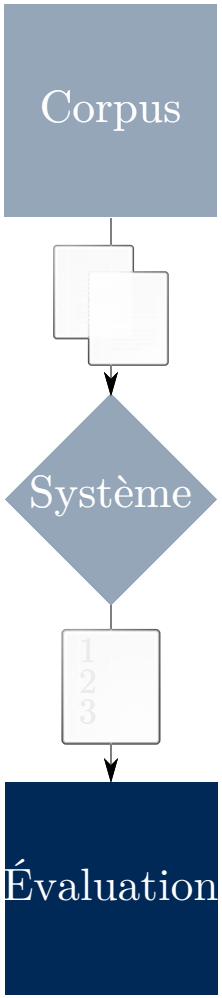
- Qualité de la classification ⇒ MAP
  - Positifs en haut et négatifs en bas
- Capacité de discrimination ⇒ SepQ
  - Écart important entre les scores des positifs et des négatifs
- Coût de mise en œuvre ⇒ Taille de la signature
  - Principalement porté par le coût de comparaison des signatures
  - Coût comparaison ⇔ Coût de la mesure de similarité
  - *containment* = Linéaire en la taille des signatures (nombre d'éléments)

Mesures d'évaluation selon trois critères



- Qualité de la classification ⇒ MAP
  - Positifs en haut et négatifs en bas
- Capacité de discrimination ⇒ SepQ
  - Écart important entre les scores des positifs et des négatifs
- Coût de mise en œuvre ⇒ Taille de la signature
  - Principalement porté par le coût de comparaison des signatures
  - Coût comparaison ⇔ Coût de la mesure de similarité
  - *containment* = Linéaire en la taille des signatures (nombre d'éléments)

Mesures d'évaluation selon trois critères

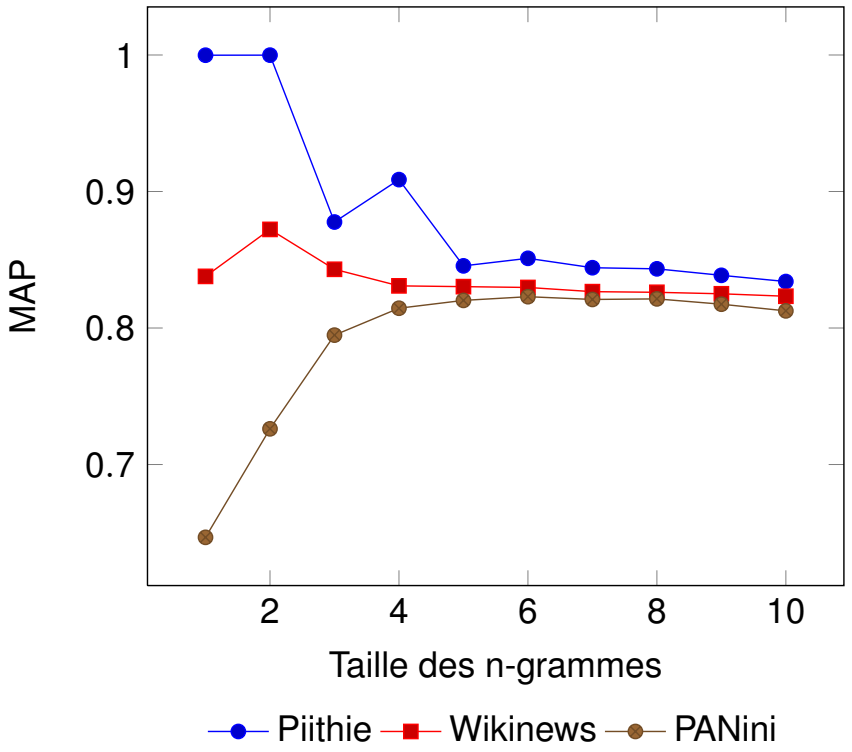


- Qualité de la classification ⇒ MAP
  - Positifs en haut et négatifs en bas
- Capacité de discrimination ⇒ SepQ
  - Écart important entre les scores des positifs et des négatifs
- Coût de mise en œuvre ⇒ Taille de la signature
  - Principalement porté par le coût de comparaison des signatures
  - Coût comparaison ⇔ Coût de la mesure de similarité
  - *containment* = Linéaire en la taille des signatures (nombre d'éléments)

Recherche de résultats de référence auxquels nous comparer

- Expérimentation de l'approche *w-shingling* sur nos corpus
- Meilleur paramètrage pour les différentes formes de dérivation
  - Modélisation ensembliste des signatures
  - Mesure de similarité $c_{max}(a, b) = \max(c(a, b), c(b, a))$
  - Normalisation des textes (filtrage mots outils + racinisation)
- Taille des n-grammes (variation de 1 à 10)

Recherche de résultats de référence



	Piithie	Wikinews	PANini
n	1	2	6
MAP	0,999	0,872	0,823

Causes supposées des variations

- Dérivés de Piithie plus similaires que Wikinews
- PANini : granularité partielle pour documents de grande taille

1 La dérivation de texte

2 Détecter les relations de dérivation

3 Évaluation comme des systèmes d'aide à la décision

4 Modéliser par des éléments singuliers et invariants
- Singularité et invariance
- Exploitation des n-grammes rares
- Exploitation des entités nommées et des composés nominaux
- Combinaison des approches

5 Discussion et perspectives

Limitation par rapport à notre contexte

Le coût de la méthode évolue avec le nombre de couples à comparer et la taille des documents

Comment réduire ce coût en maintenant les performances ?

- Filtrer le nombre de candidats en amont du système
  [Hose, 2003, Clough, 2003]
- Réduire la taille des signatures
  [Heintze, 1996, Schleimer, 2003, Stein, 2005]
- Réduire le nombre d'éléments dans la signature
  - Approche peu explorée (sélection aléatoire)
  - Conserver une approche exacte permettant un retour au texte
  - Profitable aux deux approches précédentes

Limitation par rapport à notre contexte

Le coût de la méthode évolue avec le nombre de couples à comparer et la taille des documents

Comment réduire ce coût en maintenant les performances ?

- Filtrer le nombre de candidats en amont du système [Hose, 2003, Clough, 2003]
- Réduire la taille des signatures [Heintze, 1996, Schleimer, 2003, Stein, 2005]
- Réduire le nombre d'éléments dans la signature
  - Approche peu explorée (sélection aléatoire)
  - Conserver une approche exacte permettant un retour au texte
  - Profitable aux deux approches précédentes

Limitation par rapport à notre contexte

Le coût de la méthode évolue avec le nombre de couples à comparer et la taille des documents

Comment réduire ce coût en maintenant les performances ?

- Filtrer le nombre de candidats en amont du système [Hose, 2003, Clough, 2003]
- Réduire la taille des signatures [Heintze, 1996, Schleimer, 2003, Stein, 2005]
- Réduire le nombre d'éléments dans la signature
  - Approche peu explorée (sélection aléatoire)
  - Conserver une approche exacte permettant un retour au texte
  - Profitable aux deux approches précédentes

Limitation par rapport à notre contexte

Le coût de la méthode évolue avec le nombre de couples à comparer et la taille des documents

Comment réduire ce coût en maintenant les performances ?

- Filtrer le nombre de candidats en amont du système
  [Hose, 2003, Clough, 2003]
- Réduire la taille des signatures
  [Heintze, 1996, Schleimer, 2003, Stein, 2005]
- Réduire le nombre d'éléments dans la signature
  - Approche peu explorée (sélection aléatoire)
  - Conserver une approche exacte permettant un retour au texte
  - Profitable aux deux approches précédentes

Intuition

Réduire le nombre d'éléments dans la signature

- Filtrer les éléments non conservés lors de la dérivation
  $\Rightarrow$ conserver les éléments invariants
- Filtrer les éléments non spécifiques au texte source
  $\Rightarrow$ conserver les éléments singuliers

Proposition

Exploiter uniquement des éléments aux propriétés de singularité et d'invariance pour différencier les dérivés des non-dérivés

Intuition

Réduire le nombre d'éléments dans la signature

- Filtrer les éléments non conservés lors de la dérivation
  $\Rightarrow$ conserver les éléments invariants
- Filtrer les éléments non spécifiques au texte source
  $\Rightarrow$ conserver les éléments singuliers

Proposition

Exploiter uniquement des éléments aux propriétés de singularité et d'invariance pour différencier les dérivés des non-dérivés

Intuition

Réduire le nombre d'éléments dans la signature

- Filtrer les éléments non conservés lors de la dérivation
  $\Rightarrow$ conserver les éléments invariants
- Filtrer les éléments non spécifiques au texte source
  $\Rightarrow$ conserver les éléments singuliers

Proposition

Exploiter uniquement des éléments aux propriétés de singularité et d'invariance pour différencier les dérivés des non-dérivés

Éléments que nous avons sélectionnés

- N-grammes de l'approche de référence
  - Filtrage pour ne retenir que les n-grammes singuliers
    $\Rightarrow$ Reflétant la rareté
    $\Rightarrow$ Reflétant le poids informatif
- Éléments linguistiquement ancrés
  - Taille variable des éléments mieux en accord avec la dérivation
  - Éléments porteurs de contenu ou de structure $\rightarrow$ invariance
    $\Rightarrow$ Entités nommées
    $\Rightarrow$ Composés nominaux
- Combinaison de ces éléments

Éléments que nous avons sélectionnés

- N-grammes de l'approche de référence
  - Filtrage pour ne retenir que les n-grammes singuliers
    $\Rightarrow$ Reflétant la rareté
    $\Rightarrow$ Reflétant le poids informatif
- Éléments linguistiquement ancrés
  - Taille variable des éléments mieux en accord avec la dérivation
  - Éléments porteurs de contenu ou de structure $\rightarrow$ invariance
    $\Rightarrow$ Entités nommées
    $\Rightarrow$ Composés nominaux
- Combinaison de ces éléments

Éléments que nous avons sélectionnés

- N-grammes de l'approche de référence
  - Filtrage pour ne retenir que les n-grammes singuliers
    $\Rightarrow$ Reflétant la rareté
    $\Rightarrow$ Reflétant le poids informatif
- Éléments linguistiquement ancrés
  - Taille variable des éléments mieux en accord avec la dérivation
  - Éléments porteurs de contenu ou de structure $\rightarrow$ invariance
    $\Rightarrow$ Entités nommées
    $\Rightarrow$ Composés nominaux
- Combinaison de ces éléments

Éléments que nous avons sélectionnés

- N-grammes de l'approche de référence
  - Filtrage pour ne retenir que les n-grammes singuliers
    $\Rightarrow$ Reflétant la rareté
    $\Rightarrow$ Reflétant le poids informatif
- Éléments linguistiquement ancrés
  - Taille variable des éléments mieux en accord avec la dérivation
  - Éléments porteurs de contenu ou de structure $\rightarrow$ invariance
    $\Rightarrow$ Entités nommées
    $\Rightarrow$ Composés nominaux
- Combinaison de ces éléments

Principe

- N-grammes qui n'apparaissent que dans un seul document
  $\Rightarrow$ n-grammes hapax
- Extrêmement singuliers
- Nécessite des corpus pour les distributions de référence
  - pour le français $\Rightarrow$ reliquats des articles Wikinews
  - pour l'anglais $\Rightarrow$ reliquats du corpus PAN [Potthast et al., 2010]

Résultats

		MAP		Coût	
Piithie	Référence (unigrammes)	0,999		100 %	
	Hapax (bigrammes)	0,999	=	90 %	↗
Wikinews	Référence (bigrammes)	0,872		100 %	
	Hapax (bigrammes)	0,856	↘	87 %	↗
PANini	Référence (6-grammes)	0,823		100 %	
	Hapax (bigrammes)	0,834	↗	18 %	↗

Synthèse : n-grammes hapax

- Amélioration plus nette sur le corpus PANini
  $\Rightarrow$ plus grande marge de progression (réduction du bruit dû à la faible granularité)
- Décalage de la meilleure taille des n-grammes
- N-grammes de petite taille (unigrammes et bigrammes) mieux appropriés
  $\Rightarrow$ meilleur compromis entre singularité et invariance ?

Principe : entités nommées

Intuition

- Entité nommée = parmi les entités les plus signifiantes (référent unique)
- Invariance du référent

Mise en œuvre

- Extraction automatique = technologie mature
- *Némesis* pour le français [Fourour, 2004]
- *Illinois Named Entity Tagger* pour l'anglais [Ratinov and Roth, 2009]

Principe : composés nominaux

Intuition

- Composés nominaux = privilégiés pour exprimer des idées et des concepts précis

Mise en œuvre

- Extraction automatique à l'aide de motifs syntaxiques
- Pour le français = N A, N (Prep (D)) N et N à Vinf
- Pour l'anglais = N N et A N

Résultats

		MAP		Coût	
Piithie	Référence	0,999		100 %	
	Entités nommées	0,839	↘	87 %	↗
	Comp. nominaux	0,889	↘	90 %	↗
Wikinews	Référence	0,872		100 %	
	Entités nommées	0,646	↘	61 %	↗
	Comp. nominaux	0,831	↘	67 %	↗
PANini	Référence	0,823		100 %	
	Entités nommées	0,774	↘	2 %	↗
	Comp. nominaux	0,793	↘	3 %	↗

Synthèse : entités nommées

- Baisse de la qualité de classification
- Plus forte diminution des coûts que n-grammes hapax
- Principales causes des erreurs
  - Approche peu fiable pour textes de tailles différentes ou dérivation partielle
  - Certaines entités nommées (toponymes communs)
    $\Rightarrow$ pas assez singulières et génèrent du bruit
  - Cas particulier des variations liées aux modifications des référentiels spatio-temporels
    $\Rightarrow$ *aujourd'hui* vs. *hier* vs. *mardi*
    $\Rightarrow$ *place de la Bastille* vs. *la capitale* vs. *Paris*

Synthèse : composés nominaux

- Meilleure préservation de la qualité de classification
- Diminution des coûts comparable aux entités nommées
- Globalement mêmes causes d'erreur que précédemment

Principe

Dérivation = processus complexe multidimensionnel

- Différentes méthodes proposées ciblent une dimension
- Considérer les différentes natures et transformations
- Hybridisation des méthodes

Combinaison des approches

- Fonction linéaire des scores de similarité
- Fusion des modélisations

Principe

Dérivation = processus complexe multidimensionnel

- Différentes méthodes proposées ciblent une dimension
- Considérer les différentes natures et transformations
- Hybridisation des méthodes

Combinaison des approches

- Fonction linéaire des scores de similarité
- Fusion des modélisations

Principe

Dérivation = processus complexe multidimensionnel

- Différentes méthodes proposées ciblent une dimension
- Considérer les différentes natures et transformations
- Hybridisation des méthodes

Combinaison des approches

- Fonction linéaire des scores de similarité
- Fusion des modélisations (meilleurs résultats)

Résultats

		MAP		Coût	
Piithie	Référence	0,999		100 %	
	$\Pi^{1}_{H \cdot E \cdot C}$	0,999	=	92 %	↗
Wikinews	Référence	0,872		100 %	
	$\Pi^{2}_{H \cdot E}$	0,874	↗	90 %	↗
	$\Pi^{1}_{H \cdot E \cdot C}$	0,855	↘	68 %	↗
PANini	Référence	0,823		100 %	
	$\Pi_{E \cdot C}$	0,803	↘	4 %	↗
	$\Pi^{1}_{H \cdot C}$	0,837	↗	4 %	↗

Conclusion

- Meilleurs résultats que les approches individuelles
  $\Rightarrow$ parfois mieux que l'approche de référence
- Complémentarité des approches linguistiques entre elles
- . . . et avec les approches statistiques

$\Rightarrow$ validation de l'hypothèse d'une prise en compte multidimensionnelle de la dérivation

Conclusion

- Meilleurs résultats que les approches individuelles
  $\Rightarrow$ parfois mieux que l'approche de référence
- Complémentarité des approches linguistiques entre elles
- . . . et avec les approches statistiques

$\Rightarrow$ validation de l'hypothèse d'une prise en compte multidimensionnelle de la dérivation

1 La dérivation de texte

2 Détecter les relations de dérivation

3 Évaluation comme des systèmes d'aide à la décision

4 Modéliser par des éléments singuliers et invariants

5 Discussion et perspectives
- Contributions
- Conclusion générale
- Perspectives

Contributions

1. Cadre théorique de la dérivation unifiant les propositions éparses de la littérature
2. Méthodologie d'évaluation orientée "aide à la décision", construction de corpus et résultats de référence pour le français
3. Amélioration des performances des méthodes de détection par couverture de texte
4. Boîte à outils logicielle pour expérimenter et évaluer les méthodes de détection de dérivation (TDDTS) [1]

1. http://www.fabienpoulard.info/download/Recherche/TDDTS/

Conclusion

- Cadre théorique
  - Structuration du domaine
  - Nécessite plus de confrontations expérimentales (modèle)
- Singularité et invariance
  - Validation expérimentale du principe
  - Mesurer l'invariance *a priori* ?
  - Nombre d'éléments communs parfois trop bas
- Combinaison
  - Résultats prometteurs
  - Toutes les combinaisons n'améliorent pas les résultats

Perspectives

- Maîtriser le nombre d'éléments communs dans les signatures
  - Capturer les variations des formes textuelles (réécritures)
  - Risque de réduction de la spécificité au texte source
  - Pondération des variations par des distances d'édition ?
- Capturer les granularités partielles
  - Projection d'éléments extrêmement singuliers
  - Densité d'éléments repris autour
- Mesurer la complémentarité des descripteurs
  - Quelles combinaisons améliorent les résultats ?

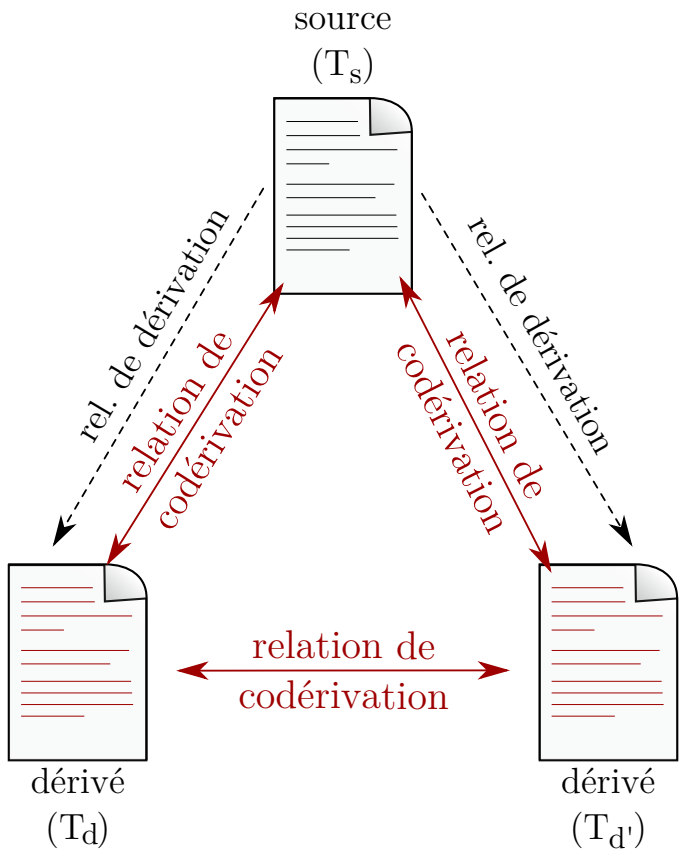
Merci !

Un cadre théorique global cohérent avec l'existant

Formalisation cohérente avec la notion de codérivation de [Bernstein and Zobel, 2004] :

Relation de codérivation $\mathbf{R_C}$

$$T_d \, \mathbf{R_C} \, T_{d'} \Leftrightarrow \begin{cases} T_d \, \mathbf{R_D} \, T_{d'} \vee T_{d'} \, \mathbf{R_D} \, T_d \\ \text{ou} \\ \exists T_s, \, T_s \, \mathbf{R_D} \, T_d \wedge T_s \, \mathbf{R_D} \, T_{d'} \end{cases}$$



Relation de codérivation

Codérivation [Bernstein and Zobel, 2004]

Documents are co-derivative if they share content : for two documents to be co-derived, some portion of one must be derived from the other or some portion of both must be derived from a third document.

- Principe de dérivation de [Bernstein and Zobel, 2004] = partage de contenu entre documents
- Notre principe de dérivation = rôle joué par un document source sur la production d'un document dérivé (pas nécessairement de contenu commun)

Détection intrinsèque de dérivation de texte

Détection intrinsèque : rechercher des indices au sein du texte

- Exploitation des marques de contextualisation
  - pour les citations [Giguet and Lucas, 2004, Mourad and Desclés, 2004, Poulard et al., 2008]
  - ou les références [Teufel et al., 2006]
- Exploitation des irrégularités stylistiques [Eissen and Stein, 2006, Stein et al., 2010]

Calcul de couverture de texte

Hypothèse

$\forall d_1, d_2 \in \mathcal{D}$ deux documents

$\mathcal{D}$ l'ensemble des documents,

$\mathcal{S}$ l'ensemble des séquences textuelles de $\mathcal{D}$,

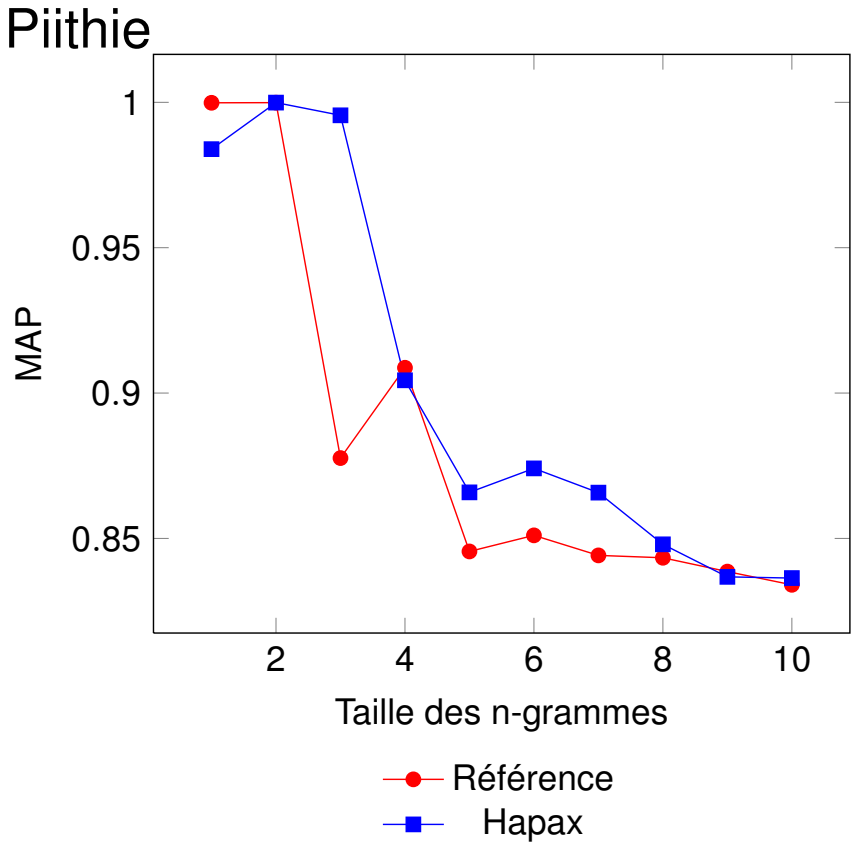
$$\begin{cases} p((d_1 \ \mathbf{R_D} \ d_2) \vee (d_2 \ \mathbf{R_D} \ d_1)) \propto \varphi(\Pi(d_1), \Pi(d_2)) \\ \quad \text{avec } \Pi : \mathcal{D} \mapsto \mathcal{P}(\mathcal{S}), \mathcal{P}(\mathcal{S}) \text{ ensemble des parties de } \mathcal{S} \\ \quad \text{et } \varphi : \mathcal{S}^n \times \mathcal{S}^m \mapsto [0..1], \text{ mesure de similarité} \end{cases}$$

Complexité des méthodes

Complexité répartit entre

- Complexité de la construction des signatures
  - Opération réalisée une seule fois par texte considéré
  - Variable selon le type d'éléments (n-grammes vs. entités nommées)
- Complexité de la comparaison des signatures
  - Dépend majoritairement de la mesure de similarité
  - containment $\Rightarrow O(|s_1| + |s_2|)$ (bornée par calcul d'intersection)

Résultats

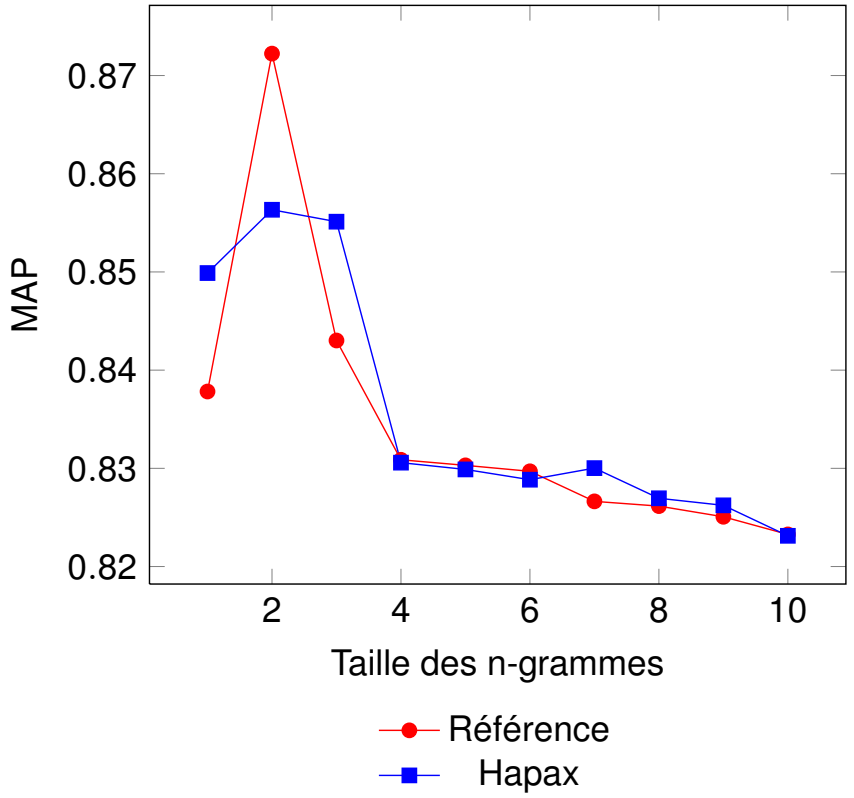


$\Rightarrow$ comportement similaire à l'approche de référence

	Référence (unigrammes)	Hapax (bigrammes)
MAP	0,999	0,999
Coût	100 %	90 %

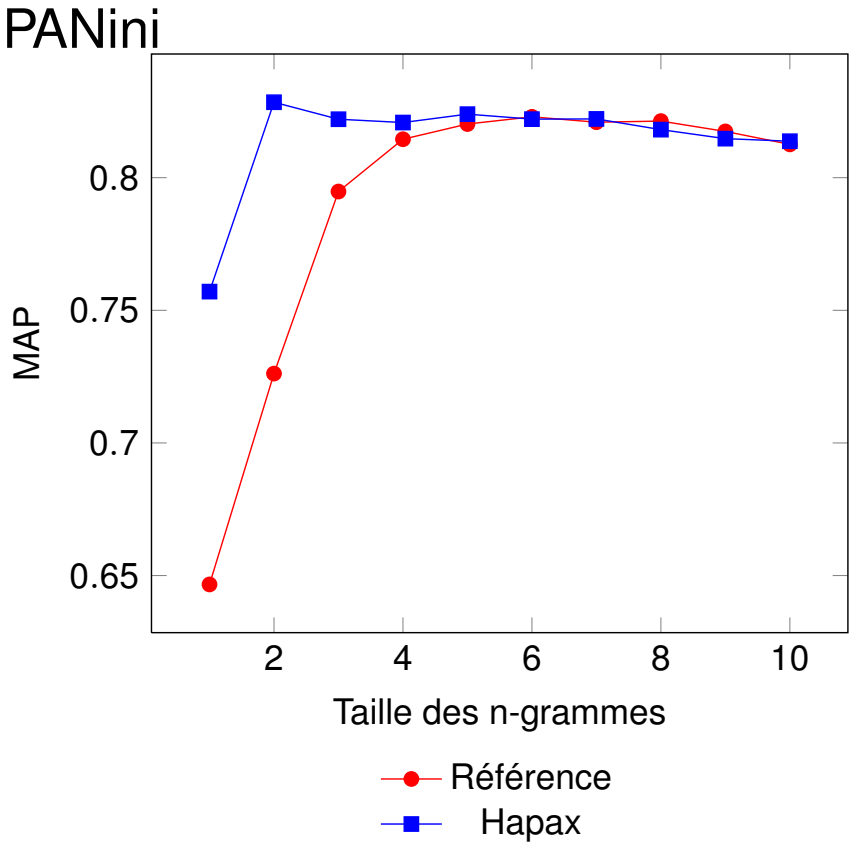
Résultats

Wikinews



	Référence (bigrammes)	Hapax (bigrammes)
MAP	0,872	0,856
Coût	100 %	87 %

Résultats



	Référence (6-grammes)	Hapax (bigrammes)
MAP	0,823	0,834
Coût	100 %	18 %

Bibliographie I


Bendersky, M. and Croft, B. (2009).
Finding text reuse on the web.
In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 262–271, Barcelona, Spain. ACM.


Bernstein, Y., Shokouhi, M., and Zobel, J. (2006).
Compact features for detection of near-duplicates in distributed retrieval.
In *In Proceedings of String Processing and Information Retrieval Symposium*.


Bernstein, Y. and Zobel, J. (2004).
A scalable system for identifying Co-Derivative documents.
In Proceedings of the Symposium on String Processing and Information Retrieval, pages 55–67.

Bibliographie II


Bourdaillet, J. (2007).
Alignement textuel monolingue avec recherche de délocationments : algorithmique pour la critique génétique.
PhD thesis, Pierre et Marie Curie.


Brin, S., Davis, J., and Garcia-molina, H. (1995).
Copy detection mechanisms for digital documents.
In Proceedings of the ACM SIGMOD Annual Conference, 24 :398–409.


Broder, A. Z. (1997).
On the resemblance and containment of documents.
In *In Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29.

Bibliographie III


Broder, A. Z., Glassman, S. C., Manasse, M. S., and Zweig, G. (1997).
Syntactic clustering of the web.
Computer Networks and ISDN Systems, 29(8-13) :1157–1166.


Chowdhury, A., Frieder, O., Grossman, D., and Mccabe, M. C. (2002).
Collection statistics for fast duplicate document detection.
ACM Transactions on Information Systems, 20 :2002.


Clough, P. (2003).
Measuring text reuse.
PhD thesis, University of Sheffield.


Eissen, S. M. Z. and Stein, B. (2006).
Intrinsic plagiarism detection.
Proceedings of the European Conference on Information Retrieval (ECIR-06).

Bibliographie IV


Fourour, N. (2004).
Identification et catégorisation des entités nommées dans les textes français.
PhD thesis, Université de Nantes.


Giguet, E. and Lucas, N. (2004).
La détection automatique des citations et des locuteurs dans les textes informatifs.
In noz, J.-M. L.-M., Marnette, S., and Rosier, L., editors, *Le discours rapporté dans tous ses états : question de frontiéres*, pages 410–418. L'Harmattan, Paris.


Heintze, N. (1996).
Scalable document fingerprinting (extended abstract).
http ://www.cs.cmu.edu/afs/cs/user/nch/www/koala/main.html.

Bibliographie V


Hoad, T. C. and Zobel, J. (2002).
Methods for identifying versioned and plagiarised documents.
Journal of the American Society for Information Science and Technology, 54 :203–215.


Hose, R. (2003).
Investigation of Sentence Level Text Reuse Algorithms.
Master's thesis, Cornell University.


Lyon, C., Barrett, R., and Malcolm, J. (2006).
Plagiarism is easy, but also easy to detect.
Plagiary, 1 :1–10.


Lyon, C., Malcolm, J., and Dickerson, B. (2001).
Detecting short passages of similar text in large document collections.
Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, pages 118–125.

Bibliographie VI

Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A., and Zobel, J. (2005).
Similarity measures for tracking information flow.
In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524. ACM New York, NY, USA.

Monostori, K., Alej, A. Z., and Bia, R. (2001).
Using the MatchDetectReveal system for comparative analysis of texts.
In *Proceedings of the 6th Australian Document Computing Symposium*, Coffs Harbour, Australia.

Bibliographie VII


Mourad, G. and Desclés, J.-P. (2004).
Identification et extraction automatique des informations citationnelles dans un texte.
In noz, J.-M. L.-M., Marnette, S., and Rosier, L., editors, *Le discours rapporté dans tous ses états : question de frontiéres*. L'Harmattan, Paris.


Potthast, M., Stein, B., and Rosso, P. (2010).
An evaluation framework for plagiarism detection.
In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010*.


Poulard, F., Waszak, T., Hernandez, N., and Bellot, P. (2008).
Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques.
In *Actes de la 15e Conférence sur le Traitement Automatique des Langues Naturelles Traitement Automatique des Langues Naturelles*, pages 450–459, Avignon France.


Ratinov, L. and Roth, D. (2009).
Design challenges and misconceptions in named entity recognition.
In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Bibliographie IX


Schleimer, S. (2003).
Winnowing : Local algorithms for document fingerprinting.
null, pages 76–85.


Seo, J. and Croft, W. B. (2008).
Local text reuse detection.
In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 571–578. ACM New York, NY, USA.


Shivakumar, N. and Garcia-Molina, H. (1995).
SCAM : A copy detection mechanism for digital documents.
In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*.


Stein, B. (2005).
Fuzzy-fingerprints for text-based information retrieval.
In *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz, Journal of Universal Computer Science*, pages 572–579.


Stein, B., Lipka, N., and Prettenhofer, P. (2010).
Intrinsic plagiarism analysis.
Language Resources and Evaluation.


Teufel, S., Siddharthan, A., and Tidhar, D. (2006).
Automatic classification of citation function.
In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110. Association for Computational Linguistics.


Wise, M. J. (1996).
YAP3 : improved detection of similarities in computer program and other texts.
ACM SIGCSE Bulletin, 28(1) :130–134.


Yang, H. (2006).
Near-duplicate detection by instance-level constrained clustering.
In *In Proceedings of the 29th ACM Conference on Research and Development in Information Retrieval (SIGIR-06). 2006*, pages 421–428.


Özlem Uzuner and Davis, A. (2003).
Content and Expression-Based copy recognition for intellectual property protection.
In the Proceedings of the 3rd ACM Workshop on Digital Rights Management (DRM'03), 2003 :103–110.