Utiliser des sens de mots pour la segmentation thématique ?

Olivier Ferret
CEA, LIST
18 route du Panorama, BP6, Fontenay-aux-Roses, F-92265 France
olivier.ferret@cea.fr

Résumé. La segmentation thématique est un domaine de l'analyse discursive ayant donné lieu à de nombreux travaux s'appuyant sur la notion de cohésion lexicale. La plupart d'entre eux n'exploitent que la simple récurrence lexicale mais quelques uns ont néanmoins exploré l'usage de connaissances rendant compte de cette cohésion lexicale. Celles-ci prennent généralement la forme de réseaux lexicaux, soit construits automatiquement à partir de corpus, soit issus de dictionnaires élaborés manuellement. Dans cet article, nous examinons dans quelle mesure une ressource d'une nature un peu différente peut être utilisée pour caractériser la cohésion lexicale des textes. Il s'agit en l'occurrence de sens de mots induits automatiquement à partir de corpus, à l'instar de ceux produits par la tâche « Word Sense Induction and Discrimination » de l'évaluation SemEval 2007. Ce type de ressources apporte une structuration des réseaux lexicaux au niveau sémantique dont nous évaluons l'apport pour la segmentation thématique.

Abstract. Many topic segmenters rely on lexical cohesion. Most of them only exploit lexical recurrence but some of them makes use of knowledge sources about lexical cohesion. These sources are generally lexical networks built either by hand or automatically from corpora. In this article, we study to what extent a new source of knowledge about lexical cohesion can be used for topic segmentation. This source is a set of word senses that were automatically discriminated from corpora, as the word senses resulting from the Word Sense Induction and Discrimination task of the SemEval 2007 evaluation. Such a resource is a way to structurate lexical networks at a semantic level. The impact of this structuring on topic segmentation is evaluated in this article.

Mots-clés : Segmentation thématique, désambiguïsation sémantique.

Keywords: Topic segmentation, word sense disambiguation.

1 Introduction

Le travail que nous présentons dans cet article peut être appréhendé selon un double éclairage. Le plus évident est celui de la segmentation thématique, problème désormais classique consistant à découper des textes en une suite de segments thématiquement homogènes. De ce point de vue, ce travail explore l'intérêt de l'utilisation pour cette tâche d'une nouvelle source de connaissances sémantiques, en l'occurrence des sens de mots induits à partir de corpus[1]. Le second point de vue est celui des connaissances : les travaux sur la désambiguïsation sémantique

[1]Nous parlerons ici de « sens de mot » afin de reprendre un vocable largement reconnu mais celui de « contexte d'usage » utilisé dans (Véronis, 2003) nous semble plus juste.

ont fait émerger la problématique de la construction de répertoires de sens à partir de corpus afin de pallier les insuffisances des dictionnaires traditionnels (Kilgarriff, 1997), au point de donner lieu à une évaluation spécifique au sein de SemEval 2007 (Agirre & Soroa, 2007). L'utilisation des sens de mots ainsi définis, soit directement, soit par le biais de la désambiguïsation sémantique, reste néanmoins un champ à peu près vierge que nous nous proposons d'explorer ici dans le cadre de la segmentation thématique.

La plupart des travaux dans le domaine de la segmentation thématique s'appuient sur les seules caractéristiques intrinsèques des documents : la récurrence lexicale dans le cas de (Hearst, 1994), (Choi, 2000), (Utiyama & Isahara, 2001), (Galley *et al.*, 2003) ou plus récemment (Eisenstein & Barzilay, 2008) ; la présence de marques linguistiques pour (Passonneau & Litman, 1997) ou (Galley *et al.*, 2003). L'absence de recours à des connaissances externes donne à ces méthodes un champ d'application en apparence large mais la récurrence lexicale n'est un indice thématique fiable que si les concepts du document considéré ne sont pas exprimés sous des formes trop diverses (synonymes, etc.) et les marques linguistiques sont souvent peu nombreuses.

Pour surmonter ces limitations, un certain nombre de systèmes exploitent des connaissances sur les relations de cohésion lexicale, connaissances qui présentent elles aussi l'avantage d'une certaine généralité. Elles prennent la forme d'un réseau lexical construit à partir d'un dictionnaire dans (Kozima, 1993), d'un thésaurus dans (Morris & Hirst, 1991), de relations issues de WordNet dans (Stokes, 2003) ou encore d'un large ensemble de cooccurrences lexicales dans (Choi *et al.*, 2001). D'une certaine façon, ces connaissances permettent aux systèmes de segmentation thématique de détecter les récurrences à un niveau plus conceptuel en leur donnant accès à des relations d'équivalence lexicale. Elles sont néanmoins dépourvues de structuration thématique.

Ce dernier point peut être résolu en exploitant des connaissances sur les thèmes susceptibles d'être rencontrés dans les documents analysés. Ces connaissances sont généralement construites automatiquement à partir d'un ensemble de documents représentatifs des thèmes considérés, comme dans (Yamron *et al.*, 1998) ou (Beeferman *et al.*, 1999). L'amélioration de la précision ainsi obtenue se fait néanmoins au détriment de la couverture des systèmes considérés. Enfin, des systèmes hybrides combinant différentes approches parmi celles présentées ci-dessus ont également été développés : (Jobbins & Evett, 1998) associe ainsi la récurrence lexicale, l'utilisation de cooccurrences et celle d'un thésaurus ; (Beeferman *et al.*, 1999) s'appuie à la fois sur une modélisation statistique des thèmes et sur l'utilisation de marques discursives ; (Galley *et al.*, 2003) se fonde conjointement sur la récurrence lexicale et sur des marques discursives.

Dans ce contexte, le travail que nous présentons dans cet article se range parmi les approches reposant sur des connaissances relatives à la cohésion lexicale des textes. Plus précisément, il intègre l'utilisation de sens de mots induits à partir de corpus au sein de F06 (Ferret, 2006), un environnement dédié à la segmentation thématique fondé sur la cohésion lexicale, et cherche ainsi à situer l'intérêt de ce type de connaissances par rapport à l'exploitation de la récurrence lexicale ou de cooccurrences lexicales. Nous débuterons l'exposé de ce travail par un aperçu de la source de connaissances ainsi utilisée.

2 Des sens de mots construits à partir de textes

Du point de vue de la caractérisation des sens de mots obtenus, il est possible de distinguer deux grandes méthodes de construction à partir d'un corpus :

– les méthodes rassemblant les mots en classes d'équivalence selon un principe de distributionnalité, à l'instar de (Pantel & Lin, 2002). Dans ce cas de figure, les différents sens d'un mot correspondent aux différents classes auxquelles il appartient. Un sens de mot est ici l'équivalent d'un synset de WordNet ;
– les méthodes discriminant les sens d'un mot en opérant une classification non supervisée de ses cooccurrents. Chaque sens de mot est ainsi défini par un sous-ensemble des cooccurrents du mot considéré. C'est l'approche retenue dans (Véronis, 2003) mais également pour les sens de mots utilisés ici (Ferret, 2004).

Sans entrer dans les détails, explicités dans (Ferret, 2004), la construction des sens de mots utilisés ici s'effectue en deux grandes étapes. La première consiste à construire à partir d'un corpus un réseau de cooccurrences lexicales en utilisant une fenêtre graphique d'une taille assez large. 24 mois du journal *Le Monde* ont ici été utilisés avec une fenêtre de 20 mots pleins. La

Sens	**Définition (cooccurrents représentatifs du sens)**
barrage de protestation	conducteur, trafic, routier, route, camion, chauffeur, voiture, bloquer, poids_lourd
barrage hydraulique	eau, mètre, lac, pluie, rivière, bassin, fleuve, site, poisson, affluent, montagne, crue, vallée
barrage frontière	casque_bleu, soldat, tir, convoi, milicien, blindé, milice, aéroport, blessé, incident, croate

TAB. 1 – Sens discriminés pour le mot *barrage* à partir du journal *Le Monde*

seconde étape est une classification non supervisée des cooccurrents de chaque mot dont on souhaite discriminer les sens. Cette classification est plus précisément appliquée au sous-graphe du réseau de cooccurrences délimité par les cooccurrents du mot cible. La Figure 1 montre ce sous-graphe pour le mot *barrage*. La discrimination des sens se fait donc par l'identification des composantes de forte densité dans ce sous-graphe.

Cette identification est réalisée en dans le cas présent par l'algorithme Shared Nearest Neighbors (Ertöz *et al.*, 2001). Cet algorithme se décompose en deux grandes phases. La première vise à identifier les germes des futures classes (ici des sens) en commençant par « éclaircir » le graphe en ne conservant pour chaque nœud que ses k plus proches voisins[2], puis en le transposant en un graphe dans lequel la pondération d'un arc correspond au nombre des voisins partagés par les nœuds qu'il relie. Un seuil sur la distribution de ces valeurs permet d'identifier des liens dits forts. Chaque nœud se voit associer son nombre de liens forts et cette valeur est utilisée pour simultanément sélectionner les germes des futures classes et éliminer les nœuds les moins représentatifs. La seconde grande phase de l'algorithme consiste à rattacher les nœuds restants aux germes des classes. Un premier rattachement au germe le plus proche est réalisé sous condition d'une similarité suffisamment importante avec lui. Lors de cette étape, des germes peuvent ainsi s'associer, impliquant ainsi la fusion des classes qu'ils représentent. Une seconde étape de rattachement permet d'associer les nœuds restants aux classes en prenant en compte leur proximité par rapport à la globalité de chaque classe formée.

Le Tableau 1 donne les sens discriminés selon cette méthode pour le mot *barrage* à partir du graphe de coocurrents de la Figure 1. Plus globalement, le répertoire de sens obtenu pour le Français se compose de 7 373 lemmes avec une moyenne de 2,8 sens par lemme et un nombre

[2]Les valeurs de similarité dans le sous-graphe des cooccurrents correspondent à une estimation de l'information mutuelle dans le réseau de cooccurrences.

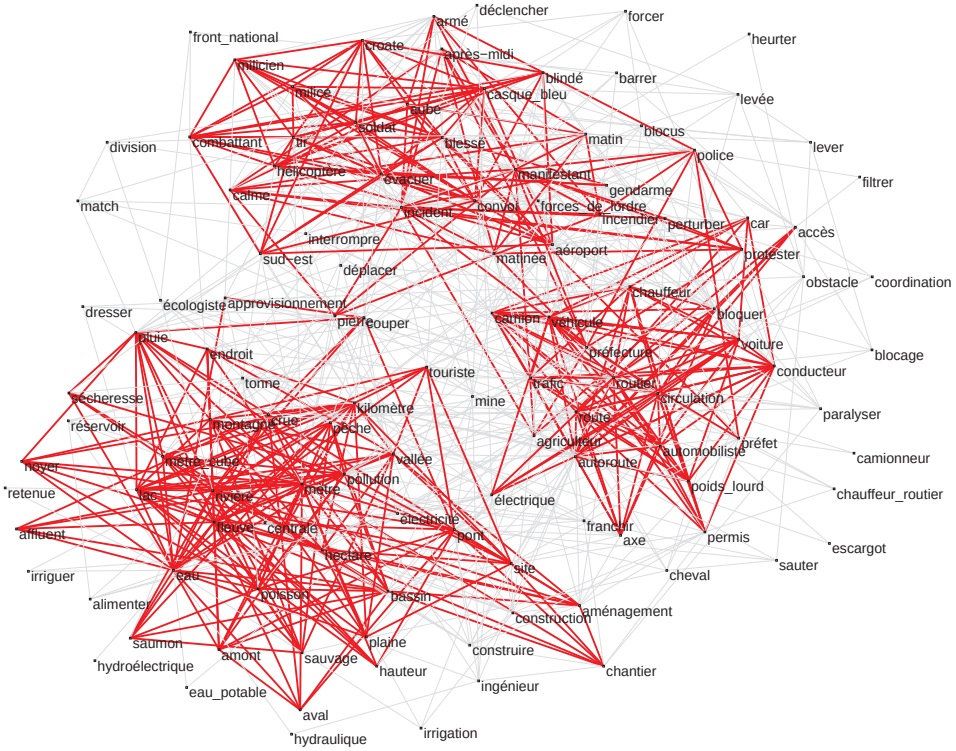

FIG. 1 – Cooccurrents du mot *barrage* (ceux impliqués dans ses 3 sens apparaissant en rouge)

moyen de mots formant la définition d'un sens égal à 16,1. Le vocabulaire initial était constitué de 17 261 lemmes mais pour un peu plus de la moitié des lemmes, la densité des liens dans le sous-graphe des cooccurrents n'était pas assez forte pour former des classes significatives.

3 F06 : un cadre pour la segmentation thématique

3.1 Principes

Après la présentation rapide de la source de connaissances à tester, nous allons maintenant présenter F06 (Ferret, 2006), le cadre retenu pour son application à la segmentation thématique. F06 reprend les principes proposés par Hearst dans *TextTiling* (Hearst, 1994) avec un processus de segmentation articulé en trois grandes parties :

– le prétraitement linguistique des documents ;
– l'évaluation de la cohésion lexicale au sein du document ;
– l'identification des changements de thème.

Le prétraitement linguistique, qui s'appuie sur l'outil *TreeTagger*, découpe les documents en phrases et représente chacune d'elles comme la séquence de ses mots pleins normalisés, c'est-à-dire ses noms (communs et propres), ses verbes et ses adjectifs. L'évaluation de la cohésion lexicale s'appuie comme dans *TextTiling* sur l'utilisation d'une fenêtre glissante de taille fixe. Cette fenêtre se déplace sur le texte de phrase en phrase. À chaque station de cette fenêtre, la cohésion lexicale est évaluée en son sein et affectée à la fin de phrase sur laquelle elle est centrée. Le résultat final est une courbe de cohésion couvrant l'ensemble du document.

La troisième partie de l'algorithme s'inspire quant à elle de son homologue dans le système *LCseg* (Galley *et al.*, 2003). Elle comprend elle-même trois étapes :

– le calcul d'un score évaluant la probabilité pour chaque minimum de la courbe de cohésion de correspondre à un changement de thème ;
– la suppression des candidats segments de trop petite taille ;
– la sélection des bornes de segments thématiques.

Le calcul du score initial d'un minimum commence par la recherche de la paire de maxima g et d qui l'entourent. En notant, $CL(x)$ la valeur de la cohésion lexicale à la position x, le score d'un minimum m est donné par :

$$score(m) = \frac{CL(g) + CL(d) - 2 \cdot CL(m)}{2} \quad (1)$$

Ce score, compris entre 0 et 1, est d'autant plus élevé que la différence entre le minimum considéré et les maxima qui l'entourent est plus importante. Il favorise ainsi comme changements de thème potentiels les minima caractérisés par une chute très nette de la cohésion lexicale. La suppression des candidats segments trop petits s'effectue quant à elle par une simple comparaison par rapport à un seuil de référence : les minima se trouvant à 2 phrases au plus du minimum qui les précède sont éliminés en tant que possibles changements de thème. Finalement, la sélection des bornes de segments thématiques est réalisée par l'utilisation d'un seuil s'adaptant à la distribution des scores des minima. Un minimum m est ainsi retenu comme borne de segment si $score(m) > \mu - \alpha \cdot \sigma$, où μ correspond à la moyenne des scores de minima, σ, à l'écart-type de ces scores et α, à un coefficient de modulation.

3.2 Évaluation de la cohésion lexicale

L'évaluation de la cohésion lexicale est l'étape la plus importante du processus de segmentation dans F06 et c'est à son niveau que sont introduites les différentes sources de cohésion lexicale à tester. Globalement, cette évaluation est réalisée suivant le principe proposé dans (Jobbins & Evett, 1998) : la cohésion est mesurée par l'application du coefficient de *Dice* entre les vecteurs représentant les deux moitiés de la fenêtre glissante. Dans le cas de la récurrence lexicale, ce principe est repris strictement : si F_g désigne le vocabulaire de la moitié gauche de la fenêtre et F_d, celui de sa moitié droite, la cohésion au sein de la fenêtre est donnée par :

$$cohésion\_réc(x) = \frac{2 \cdot card(F_g \cap F_d)}{card(F_g) + card(F_d)} \quad (2)$$

Lorsque les relations de cohésion ne sont pas des relations de récurrence[3], la mesure utilisée est une extension du coefficient de *Dice*. Dans chaque volet de la fenêtre, le nombre de mots liés par ces relations de cohésion avec les mots de l'autre volet de la fenêtre est déterminé en excluant les mots déjà impliqués dans des relations de récurrence. Les contributions des deux volets de la fenêtre sont ensuite sommées et ramenées au nombre total de mots dans la fenêtre[4] :

$$cohésion\_rel(x) = \frac{card(M_{rel}(g) - M_{réc}) + card(M_{rel}(d) - M_{réc})}{card(F_g) + card(F_d)} \quad (3)$$

[3]Plus généralement, on parlera de relations d'égalité ou d'équivalence entre mots, se limitant ici à la récurrence.
[4]Dans le cas du coefficient de *Dice*, la contribution de chaque volet correspond à leur intersection.

où $M_{rel}(x)$ représente les mots du volet x de la fenêtre (*g*auche ou *d*roit) sélectionnés sur la base des relations de cohésion lexicale et $M_{réc}$, les mots impliqués dans une relation de récurrence. La cohésion globale au sein de la fenêtre est finalement donnée par la somme de $cohésion\_réc(x)$ et des valeurs de $cohésion\_rel(x)$ correspondant aux différents types de relations de cohésion distingués.

4 Utiliser des sens de mots pour la segmentation thématique

Comme nous l'avons vu à la section précédente, l'intégration au sein de F06 d'une nouvelle source de connaissances sur la cohésion lexicale se fait au niveau de l'évaluation de la cohésion de la fenêtre glissante d'analyse et plus précisément lors de la détermination des mots de chacun de ses volets liés aux mots de l'autre volet. Dans le cas des sens de mots présentés à la Section 2, la méthode d'intégration en apparence la plus directe serait une forme d'extension de la récurrence lexicale : au lieu de se focaliser sur la répétition des mots, elle se focaliserait sur la répétition des sens de mots. Cette méthode impose de réaliser une désambiguïsation sémantique des textes. Sachant que dans un texte, il est peu fréquent qu'un même mot soit utilisé avec deux sens différents pour faire référence à deux thèmes différents, il apparaît probable que le gain de précision espéré du fait de l'utilisation de sens de mots soit en pratique effacé par le taux d'erreur du processus de désambiguïsation sémantique.

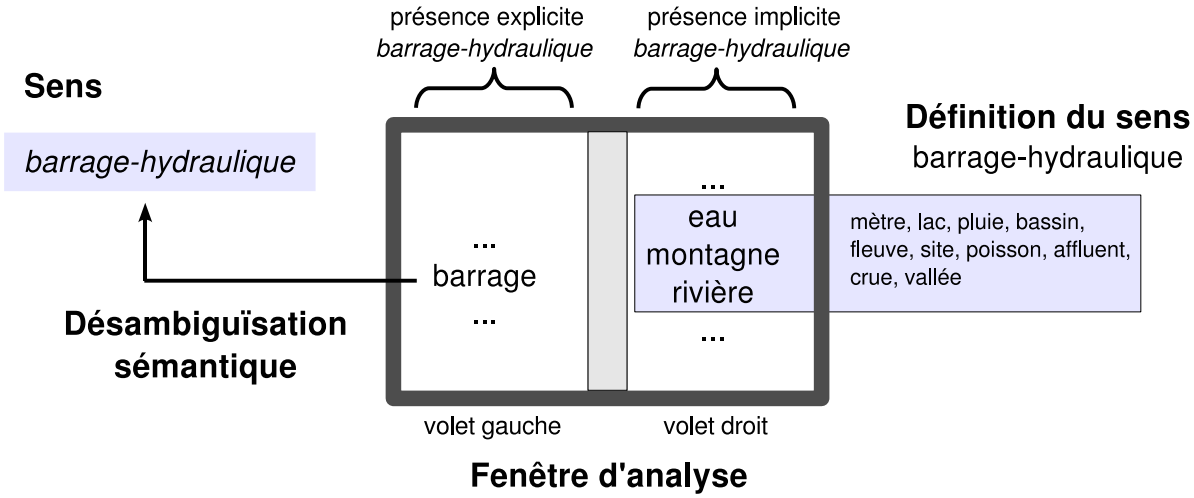

FIG. 2 – Utilisation des sens de mots pour l'évaluation de la cohésion lexicale

Nous avons donc opté pour une solution plus souple dans laquelle l'identification des sens de mots entre les deux volets de la fenêtre d'analyse ne passe pas par la présence explicite du même sens pour un mot donné dans ces deux volets mais par le fait qu'un lien peut être établi entre les deux volets si l'un d'eux contient un sens d'un mot et l'autre un nombre significatif d'éléments de définition de ce sens. Ainsi que l'illustre la Figure 2, si l'un des volets de la fenêtre contient le mot *barrage* avec le sens *barrage-hydraulique* et que l'autre volet contient les mots *eau*, *montagne* et *rivière*, tous trois faisant partie de la définition du sens *barrage-hydraulique* (cf. Tableau 1), nous faisons l'hypothèse de la présence d'un lien de cohésion entre les deux volets de la fenêtre, lien sous-tendu par la co-présence, l'une explicite, l'autre implicite, du sens *barrage hydraulique*.

Dans le cadre de F06, et plus précisément de l'équation 3, l'identification d'un tel lien se traduit par l'ajout des mots de la définition présents dans un volet de la fenêtre à l'ensemble $M_{rel}(x)$ associé à ce volet et l'ajout du mot défini à l'ensemble $M_{rel}(x)$ de l'autre volet. En pratique, ce mécanisme est déclenché dès lors qu'au moins 2 mots de la définition d'un sens de mot sont présents dans un volet de la fenêtre. Lorsque la discrimination de sens ne produit pas de résultat pour un mot, une procédure de rattrapage est appliquée pour maximiser la détection des relations de cohésion : le mot est supposé n'avoir qu'un seul sens, défini par ses cooccurrents dans le réseau de cooccurrences initialement utilisé pour discriminer les sens de mots.

Le mécanisme d'intégration décrit ci-dessus repose sur la possibilité d'identifier le sens des mots au sein de la fenêtre d'analyse. Pour ce faire, les mots de chacun de ses deux volets sont désambiguïsés en prenant comme contexte les mots du volet dans lequel ils se trouvent. Cette désambiguïsation est réalisée selon une approche de type Lesk : le recouvrement entre le contenu du volet de la fenêtre dans lequel un mot se trouve et la définition de chacun de ses sens est évalué et le sens retenu est celui pour lequel ce recouvrement est le plus important. Plus précisément, ce recouvrement est évalué par la mesure *Cosinus*.

5 Évaluation

5.1 Méthodologie

L'objectif de l'évaluation menée dans ce travail étant de déterminer l'intérêt de l'utilisation de sens de mots comme source de cohésion lexicale dans F06, nous reprendrons le même cadre d'évaluation que celui développé pour F06. Celui-ci s'inspire de la désormais classique méthode d'évaluation proposée par Choi dans (Choi, 2000). Cette méthode est fondée sur la constitution d'un corpus d'évaluation composé de morceaux de documents de taille variable collés les uns à la suite des autres. L'objectif pour le segmenteur évalué est de retrouver les frontières des morceaux de documents. Dans (Ferret, 2006), nous avons proposé une adaptation de cette procédure visant à contrôler plus précisément sa dimension thématique, ce qui permet d'ailleurs de se rapprocher de conditions plus réalistes d'après les résultats obtenus. Au lieu de tirer chaque extrait d'un document différent, nous n'utilisons que deux documents. Chacun d'entre eux est divisé comme dans le cas de Choi en segments de 3 à 11 phrases et le document d'évaluation est constitué en prenant, à partir du début des documents, alternativement un segment dans un des deux documents et le suivant dans l'autre et ce, jusqu'à obtenir 10 segments ou jusqu'à ce que le processus de construction atteigne la fin d'un des deux documents.

Pour nous assurer que deux segments consécutifs font référence à deux thèmes différents et que le changement de thème entre les deux est effectif, les deux documents sont sélectionnés de manière à appartenir à des thématiques différentes. Pour ce faire, nous nous sommes appuyés sur les données constituées pour l'évaluation CLEF sur la recherche d'information multilingue. Les documents source utilisés pour réaliser notre corpus étaient ainsi des documents issus du corpus CLEF pour lesquels nous disposions d'un jugement de pertinence par rapport à un des topics d'interrogation définis pour l'évaluation. Chaque document de notre corpus a ainsi été construit à partir de deux documents du corpus CLEF jugés pertinents pour deux topics différents. Nous avons ainsi constitué un corpus d'évaluation de 100 documents construits à partir de 11 topics et de documents issus de deux ans du journal *Le Monde*.

Classiquement, nous avons utilisé la mesure d'erreur P_k (Beeferman *et al.*, 1999) pour évaluer

la qualité des résultats des segmenteurs. P_k évalue la probabilité que deux mots choisis aléatoirement dans un document et séparés par k mots soient jugés comme appartenant au même segment alors qu'ils sont dans des segments différents (faux négatif) ou qu'ils soient jugés comme appartenant à des segments différents alors qu'ils sont dans le même (fausse alarme). k est égal à la moitié de la taille moyenne en mots des segments au niveau du corpus de référence. L'objectif est de minimiser P_k. *WindowDiff* (Pevzner & Hearst, 2002), dont nous donnons aussi les résultats, est une variante de P_k prenant en compte le nombre de frontières de segments séparant deux mots situés dans des segments différents.

5.2 Résultats et discussion

Le Tableau 2 donne les résultats obtenus par les segmenteurs testés dans le cadre F06 sur le corpus décrit à la Section 5.1 mais également les résultats sur ce même corpus de méthodes de référence : U00 est ainsi la méthode décrite dans (Utiyama & Isahara, 2001), C99, celle proposée dans (Choi, 2000) et LCseg est présentée dans (Galley *et al.*, 2003). TextTiling* est une variante de TextTiling dans laquelle la troisième étape d'identification des changements de thème est reprise de (Galley *et al.*, 2003).

Au sein de F06, F06R s'appuie sur la seule récurrence lexicale. F06C (Ferret, 2006) met en œuvre l'utilisation de relations de cooccurrence lexicale au sein de F06 en s'appuyant sur l'équation 3 : si un mot m d'un volet de la fenêtre d'analyse est lié à un nombre minimal (en l'occurrence 2) de mots de son autre volet par des relations de cooccurrence d'un niveau de cohésion suffisamment haut, ces cooccurrents de m sont ajoutés à l'ensemble $M_{rel}(x)$ associé au volet de la fenêtre où ils apparaissent. Pour les segmenteurs exploitant des sens de mots, 2 systèmes ont été testés. F06WS correspond strictement au descriptif de la Section 4 tandis que F06WSr est une variante sans désambiguïsation sémantique : les mots dans un des volets de la fenêtre peuvent correspondre à l'un quelconque des sens de l'un des mots de l'autre volet.

Les sens d'un mot pouvant être vus comme une forme de structuration de ses cooccurrents, F06WS et F06Wr se comparent naturellement à F06C, F06R servant de référence basse. Pour chaque résultat de ces méthodes, nous donnons donc le degré de signification p de sa différence avec F06R et F06C, niveau évalué grâce à un test de Student unilatéral dont les valeurs inférieures à 0,05 sont considérées comme significatives (en gras).

systèmes	**P_k (%)**			**WindowDiff (%)**		
	erreur	**p(F06R)**	**p(F06C)**	**erreur**	**p(F06R)**	**p(F06C)**
U00	25,91	**0,003**	**1,3e-07**	27,42	0,799	**0,032**
C99	27,57	**4,2e-05**	**3,6e-10**	35,42	**8,6e-07**	**6,5e-13**
TextTiling*	21,08	0,699	**0,037**	27,43	0,803	**0,032**
LCseg	20,55	0,439	0,111	28,31	0,767	**0.007**
F06R	21,58	/	**6,5e-05**	27,83	/	**4,8e-06**
F06C	16,48	**6,5e-05**	/	20,94	**4,8e-06**	/
F06WSr	*18,50*	**0,015**	0,10	*23,14*	**0,002**	0,11
F06WS	*18,17*	**0,006**	0,16	*23,20*	**0,002**	0,12

TAB. 2 – Résultats des différents segmenteurs testés sur le corpus de la Section 5.1

Le premier constat que l'on peut tirer de l'analyse de ce tableau est que globalement, l'apport de connaissances externes sur la cohésion lexicale, que ce soit sous la forme de cooccurrences

lexicales ou de sens de mots, représente un atout indéniable puisque les valeurs de P_k et de *WindowDiff* sont nettement meilleures pour les méthodes utilisant ces connaissances (F06C, F06WS et F06WSr) que pour celles n'exploitant que la récurrence lexicale et ce, de façon statistiquement significative dans la plupart des cas.

Concernant plus spécifiquement les sens de mots, cette évaluation fait apparaître que si leur utilisation permet d'améliorer les résultats par rapport à F06R, elle ne permet pas en revanche pas d'obtenir de progrès vis-à-vis de l'utilisation de cooccurrences lexicales. Les résultats avec les sens de mots sont même moins bons mais la différence n'est pas statistiquement significative. Les deux variantes testées, F06WS et F06WSr, sont à cet égard comparables, sans différence significative sur le plan statistique.

Ce résultat est décevant dans la mesure où les sens de mots utilisés ici, bien que proches des relations de cooccurrence de par leur définition, représentent un degré de structuration supérieur des connaissances, degré dont on pourrait attendre un impact positif sur la précision de la segmentation. Même s'il est assez difficile d'analyser dans le détail les causes de ces résultats, il est probable que le processus de structuration des cooccurrents intervenant lors de la discrimination des sens de mots provoque une perte d'informations de cohésion lexicale. Cette perte peut avoir pour origine à la fois l'élimination de certains cooccurrents jugés sans doute à tort non significatifs et le fait que la méthode de clustering utilisée est de type « hard clustering », conduisant à construire des sens sans aucun partage des mots constituant leur définition. Cette perte d'informations masque ainsi les gains potentiellement apportés par une plus grande structuration des connaissances.

6 Conclusion et perspectives

Dans cet article, nous avons étudié l'intérêt de l'utilisation de sens de mots discriminés à partir de corpus pour la segmentation thématique de textes. Cette étude a été réalisée en s'appuyant sur F06, un cadre d'étude de la segmentation thématique fondée sur la cohésion lexicale. Les résultats obtenus montrent que les sens de mots présentent un intérêt par rapport à l'exploitation de la simple récurrence lexicale mais qu'en revanche, ils ne se révèlent pas meilleurs de ce point de vue que de simples relations de cooccurrence lexicale.

Les perspectives de ce travail sont directement liées à l'analyse des résultats et poussent à porter les efforts davantage sur la construction des sens de mots que sur leur exploitation. En particulier, il serait intéressant de modifier la méthode de clustering utilisée afin de permettre un certain recouvrement entre les définitions des sens de mots. Le test d'autres méthodes de clustering serait également pertinent. Enfin, le recours à des sens de mots définis sur la base de classes d'équivalence distributionnelles constitue une autre piste à explorer, complémentaire de celle déjà empruntée. Au final, on notera qu'en dépit de résultats un peu décevants, l'utilisation de sens de mots dans le cadre de la segmentation thématique apparaît comme un moyen possible pour évaluer ce type de ressource et contribuer ainsi à cette tâche intrinsèquement difficile.

Références

AGIRRE E. & SOROA A. (2007). Semeval-2007 task 02 : Evaluating word sense induction and discrimination systems. In *Fourth International Workshop on Semantic Evaluations (SemEval-*

2007), p. 7–12.

BEEFERMAN D., BERGER A. & LAFFERTY J. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1), 177–210.

CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *NAACL'00*, p. 26–33.

CHOI F. Y. Y., WIEMER-HASTINGS P. & MOORE J. (2001). Latent semantic analysis for text segmentation. In *EMNLP'01*, p. 109–117.

EISENSTEIN J. & BARZILAY R. (2008). Bayesian unsupervised topic segmentation. In *2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*.

ERTÖZ L., STEINBACH M. & KUMA V. (2001). Finding topics in collections of documents : A shared nearest neighbor approach. In *Text Mine'01, Workshop of the 1st SIAM International Conference on Data Mining*.

FERRET O. (2004). Discovering word senses from a network of lexical cooccurrences. In *20th International Conference on Computational Linguistics (COLING 2004)*, p. 1326–1332.

FERRET O. (2006). Approches endogène et exogène pour améliorer la segmentation thématique de documents. *Traitement Automatique des Langues (TAL), numéro spécial Discours et Document*, **47**(2), 111–135.

GALLEY M., MCKEOWN K., FOSLER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, p. 562–569.

HEARST M. A. (1994). Multi-paragraph segmentation of expository text. In *32th Annual Meeting of the Association for Computational Linguistics*, p. 9–16.

JOBBINS A. C. & EVETT L. J. (1998). Text segmentation using reiteration and collocation. In *ACL-COLING'98*, p. 614–618.

KILGARRIFF A. (1997). I don't believe in word senses. *Computers and the Humanities*, **31**(2), 91–113.

KOZIMA H. (1993). Text segmentation based on similarity between words. In *31th Annual Meeting of the Association for Computational Linguistics (Student Session)*, p. 286–288.

MORRIS J. & HIRST G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, **17**(1), 21–48.

PANTEL P. & LIN D. (2002). Discovering word senses from text. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002*, p. 613–619.

PASSONNEAU R. J. & LITMAN D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, **23**(1), 103–139.

PEVZNER L. & HEARST M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, **28**(1), 19–36.

STOKES N. (2003). Spoken and written news story segmentation using lexical chains. In *HLT-NAACL 2003, student session*, p. 49–54.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *ACL 2001*, p. 491–498.

VÉRONIS J. (2003). Cartographie lexicale pour la recherche d'information. In *10ème Conférence sur le Traitement automatique des langues naturelles (TALN 2003)*, p. 265–274.

YAMRON J., CARP I., GILLICK L., LOWE S. & VAN MULBREGT P. (1998). A hidden markov model approach to text segmentation and event tracking. In *IEEE Conference on Acoustics, Speech and Signal Processing*, p. 333–336.