Comparaison de deux critères de rupture thématique pour l'extraction d'indices de segmentation et de liage dans une grande collection de textes

Yves Bestgen*

*Chercheur qualifié du FNRS, UCL, 10 Place Mercier, 1348 Louvain-la-Neuve Belgique
yves.bestgen@uclouvain.be

Résumé. Une technique automatique est proposée et évaluée pour extraire de grands corpus de textes des indices de rupture et de continuité thématique, tels que les expressions adverbiales, les connecteurs et les anaphores, qui retiennent l'attention de chercheurs tant en linguistique qu'en psychologie cognitive et en TAL, mais qui sont le plus souvent étudiés par l'analyse qualitative d'un très petit nombre de textes.

1 Introduction et approche proposée

Depuis de nombreuses années, la linguistique textuelle s'intéresse aux indices de rupture et de continuité thématique, tels que les expressions adverbiales, les connecteurs et les anaphores (Adam, 2005 ; Charolles, 1995 ; Bestgen & Piérard, 2014). Ces indices retiennent également l'attention de chercheurs en psychologie cognitive et en traitement automatique du langage parce qu'ils sont susceptibles de faciliter la compréhension d'un texte et d'améliorer les performances d'algorithmes de segmentation automatique. En linguistique textuelle, ces marques sont étudiées par l'analyse qualitative approfondie d'un très petit nombre de textes, ce qui rend difficile toute comparaison entre des genres de textes. De plus, nombre d'indices linguistiques de la structure ont un taux d'occurrence très faible, imposant l'analyse de grands corpus si on cherche à en dresser un tableau représentatif.

L'objectif de la présente étude est d'évaluer une approche basée sur des techniques du TAL qui permet l'extraction de ces indices dans de grandes quantités de textes. L'approche, qui s'inspire des travaux sur la détection de paragraphes (Sporleder & Lapata, 2006), repose sur la construction, par d'une procédure d'apprentissage supervisé, de modèles prédictifs capables de discriminer dans des textes des paires de phrases contigües qui sont ou non séparées par une rupture thématique. Nous avons employé une machine à vecteurs de support (SVM, version linéaire) bien connue pour son efficacité en catégorisation de textes. Les indices potentiels analysés sont composés des unigrammes, bigrammes et trigrammes de lemmes et d'étiquettes morphosyntaxiques présents dans les phrases. Lors de leur extraction, les n-grammes en début de phrase sont distingués des autres parce que les indices de rupture occupent fréquemment cette position (Bestgen & Piérard, 2014). L'hypothèse sous-jacente à cette approche est que les indices les plus efficaces pour discriminer les phrases en situation de rupture de celles en situation de continuité seront des candidats à la fonction d'indices de segmentation et de liage.

Pour identifier les phrases en situation de continuité ou de discontinuité thématiques, deux critères ont été comparés (Piérard & Bestgen, 2006). Le premier est dérivé d'une mesure de cohésion lexicale issue de l'analyse sémantique latente (ASL, Landauer et al., 2007). Il s'agit d'une technique d'analyse statistique de corpus dont l'objectif est d'inférer les similarités sémantiques entre des mots sur la base de leurs occurrences dans des documents. Le sens de chaque mot ou de chaque phrase est représenté par un vecteur dans l'espace sémantique. Pour calculer la similarité sémantique entre deux phrases, on calcule le cosinus entre les vecteurs qui les représentent. Plusieurs autres approches pour extraire des espaces sémantiques ont été proposées (Levy et al., 2015). L'intérêt de l'ASL est que son efficacité pour estimer la cohérence dans des textes est considérée comme bien établie (Landauer et al., 2007, p. 10). Le second critère est plus simple puisqu'il s'appuie directement sur la structure du texte telle qu'elle est rendue visible par les sections (le niveau 1 dans Wikipédia pour la présente étude).

2 Évaluation

L'objectif de cette étude est d'évaluer l'intérêt de l'approche proposée en répondant aux deux questions suivantes : quelle est son efficacité pour discriminer les phrases en situation de rupture de celles en situation de continuité et les indices les plus utiles pour la SVM sont-ils compatibles avec les conclusions des analyses linguistiques plus qualitatives ?

2.1 Méthode

Les analyses ont porté sur des phrases extraites de l'encyclopédie Wikipédia francophone au moyen de *wikiextractor* de Attardi. Les textes ont été lemmatisés et étiquetés morphosyntaxiquement par le TreeTagger (Schmid, 1994). Les phrases utilisées ont été sélectionnées après suppression des documents très courts et très longs par rapport à la longueur moyenne ainsi que de ceux contenant un grand nombre de phrases très brèves ou un nombre disproportionné d'étiquettes morphosyntaxiques des catégories *abréviation*, *nom propre*, *nombre* et *signe de ponctuation*. L'espace sémantique employé pour calculer les cosinus a été extrait de cette collection de textes sur la base d'une segmentation arbitraire en unité de 250 mots.

Pour l'ASL, on a calculé le cosinus entre les deux phrases de toutes les paires de phrases contigües d'au moins 7 mots et d'au plus 43 mots. Ont été considérés en situation de rupture les 2,5% de paires de phrases dont le cosinus était le plus petit. Les paires de phrases en situation de continuité sont celles dont le cosinus fait partie des 2,5% les plus élevés. Les n-grammes servant de traits pour l'apprentissage supervisé ont été extraits de la deuxième phrase de chaque paire. Au total, il y a 35 629 phrases de chaque type.

Pour le critère basé sur les sections, on a analysé les quadruplets de phrases contigües répondant aux mêmes critères de longueur. Pour être considérée en situation de rupture, la troisième phrase de ces quadruplets devait être précédée par une rupture de section principale et il ne pouvait y avoir d'alinéas entre les autres phrases du quadruplet. La troisième phrase d'un quadruplet était considérée en situation de continuité s'il n'y a avait aucune rupture de section, ni aucun alinéa entre les quatre phrases. On a extrait le maximum possible de phrases en situation de rupture (N = 19 216) et un échantillon aléatoire d'une même taille de phrases en situation de continuité. Les n-grammes pour la SVM ont été extraits comme pour l'ASL.

Pour chaque analyse, les phrases cibles ont été divisées en 80% pour l'apprentissage et 20% pour l'évaluation. Deux paramètres ont été optimisés sur 25% des données d'apprentissage : le seuil de fréquence minimal des traits dans le matériel analysé et le paramètre C de régularisation. Les 200 indices les plus utiles pour la SVM (Chang & Lin, 2008) ont été sélectionnés pour être analysés qualitativement.

2.2 Résultats

La première analyse indique que les deux critères de continuité thématique sont liés puisque l'indice de continuité dérivé de l'ASL est plus faible lorsque deux phrases contigües sont séparées par un changement de section (moy. = 0.15) que lorsqu'elles ne le sont pas (moy. = 0.24) (test *t* pour comparaison de moyennes, $p < 0.0001$). Comme le montre le tableau 1, les performances [1] obtenues par la procédure d'apprentissage supervisée sur la base de chacun des deux critères sont relativement élevées. Discriminer les ruptures obtenues par l'ASL est plus simple que discriminer celles basées sur les sections. Le tableau 1 indique aussi les performances obtenues lorsque la SVM ne peut employer que les n-grammes en début de phrase ou que les autres n-grammes. La SVM basée sur l'ASL bénéficie nettement moins des n-grammes en début de phrase que la SVM basée sur les sections alors que les études qualitatives en linguistique soulignent l'importance de la position initiale (Adam, 2005).

Le tableau 2 présente le pourcentage d'indices parmi les 200 les plus utiles qui sont en position initiale, des lemmes et des unigrammes. On observe que les indices sélectionnés sur la base des deux critères sont très différents : l'ASL privilégie nettement des unigrammes de lemmes qui ne commencent pas les phrases. Il s'agit de lemmes comme *album, cheval, langue, Disney, mètre, vin, espèce* et *film* pour signaler une rupture et comme *jeune, peu, politique, siècle* et *aller* pour signaler une continuité. Ces indices ne semblent pas interprétables dans le cadre des travaux linguistiques sur les marques de segmentation et de liage (Adam, 2008), contrairement à ceux mis en évidence lorsqu'il s'agit de prédire la présence ou non d'un changement de section dont les plus utiles sont les pronoms personnels, les déterminants démonstratifs et des connecteurs comme *mais*, *ainsi* et *cependant* pour la continuité et une suite de deux noms propres, *naître*, *après_le* et *depuis_le* pour la discontinuité, à chaque fois en début de phrase.

	Nbr. d'exemplaires	Tout	Début	Suite
ASL	71258	78%	62%	77%
Section	38432	74%	67%	70%

TAB. 1 – *Efficacité (exactitude) pour distinguer les ruptures des continuités*

	Début	Lemme	Unigramme
ASL	10%	93%	86%
Section	40%	38%	31%

TAB. 2 – *Indices les plus utiles selon le type*

1. L'exactitude est employée comme mesure de performance parce que, dans toutes les analyses effectuées, elle est égale au rappel et à la précision et donc aussi à la mesure F1.

3 Conclusion

Les résultats obtenus sur la base des sections sont relativement élevés et les indices sont interprétables. Cette approche semble être une voie prometteuse pour étudier les indices de continuité et de liage dans de grands corpus et il serait intéressant d'appliquer la technique aux sous-sections et aux paragraphes, ainsi qu'à des corpus journalistiques et littéraires. Si le critère structurel dérivé de l'ASL donne lieu à une meilleure performance, les indices les plus utiles sont nettement moins interprétables. Ce résultat questionne l'emploi de l'ASL pour déterminer la cohérence de textes, une procédure largement popularisée par des outils libres d'accès comme Coh-Metrix. Tirer une conclusion définitive de la présente étude serait toutefois prématuré puisqu'on n'a pas procédé à une analyse approfondie des phrases considérées par l'ASL comme en situation de rupture ou de continuité. Il serait aussi intéressant d'évaluer des procédures d'extraction d'espaces sémantiques proposées récemment (Levy et al., 2015).

Références

Adam, J.-M. (2008). *La linguistique textuelle.* Paris : A. Colin.

Bestgen, Y. et S. Piérard (2014). Sentence-initial adverbials and text comprehension. In L. Sarda, S. Carter Thomas, B. Fagard, et M. Charolles (Eds.), *Adverbials in Use : From Predicative to Discourse Functions*, pp. 151–168. PUL.

Chang, Y.-W. et C.-J. Lin (2008). Feature ranking using linear SVM. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, pp. 53–64.

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique : Revue Internationale de Linguistique Française 29*, 125–151.

Landauer, T., D. McNamara, D. Simon, et W. Kintsch (2007). *Handbook of Latent Semantic Analysis*. Mahwah: Erlbaum.

Levy, O., Y. Goldberg, et I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *TACL 3*, 211–225.

Piérard, S. et Y. Bestgen (2006). Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL 47*(2), 89–110.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sporleder, C. et M. Lapata (2006). Broad coverage paragraph segmentation across languages and domains. *ACM Transactions on Speech and Language Processing 3*, 1–35.

Summary

An automatic technique is proposed and evaluated to extract from large corpora markers of thematic continuity and discontinuity, such as adverbials, connectives and anaphoras, which attract the attention of researchers in linguistics, cognitive psychology and natural language processing, but are often studied by the qualitative analysis of a very small number of texts.