TALN 2010, Montréal, 19–23 juillet 2010

Segmentation Automatique de Lettres Historiques

Michel Généreux, Rita Marquilhas, Iris Hendrickx
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

Résumé. Cet article présente une approche basée sur la comparaison fréquentielle de modèles lexicaux pour la segmentation automatique de textes historiques Portugais. Cette approche traite d'abord le problème de la segmentation comme un problème de classification, en attribuant à chaque élément lexical présent dans la phase d'apprentissage une valeur de saillance pour chaque type de segment. Ces modèles lexicaux permettent à la fois de produire une segmentation et de faire une analyse qualitative de textes historiques. Notre évaluation montre que l'approche adoptée permet de tirer de l'information sémantique que des approches se concentrant sur la détection des frontières séparant les segments ne peuvent acquérir.

Abstract. This article presents an approach based on the frequency comparison of lexical models for the automatic segmentation of historical texts. This approach first addresses the problem of segmentation as a classification problem by assigning each token present in the learning phase a value of salience for each type of segment. These lexical patterns can both produce a segmentation and make possible a qualitative analysis of historical texts. Our evaluation shows that the approach can extract semantic information that approaches focusing on the detection of boundaries between segments cannot capture.

Mots-clés : Corpus comparables, Saillance, Segmentation, Textes historiques.

Keywords: Comparable corpora, Salience, Segmentation, Historical Texts.

1 Introduction

Dans le projet CARDS[1], des lettres privées allant du 16ième jusqu'au 19ième siècle au Portugal sont transcrites manuellement. Le corpus CARDS est étiqueté textuellement pour identifier les formules d'ouverture (*opening*) et de fermeture (*closing*), d'exorde (*harengue*) et de conclusion ou péroraison (*peroration*). Dans l'étude présentée ici, le but est de réduire la charge de travail manuel par le traitement automatique des corpus en ce qui concerne la segmentation afin de produire une édition critique électronique et une interprétation historique et linguistique des lettres. Ce papier présente donc un travail dont le but est de segmenter automatiquement des lettres faisant partie d'un corpus historique, en les séparant en cinq parties, le corps de la lettre et quatre parties formelles identifiées par les historiens (ouverture, exorde, conclusion et clôture). Le modèle choisi est purement lexical, chaque mot étant classifié en une de ces cinq classes pour indiquer qu'il appartient à une partie ou à une autre. Des scores d'associations de n-grammes (pour n=1,2 et 3) sont calculés pour choisir la classe d'un mot donné.

[1] `http://alfclul.clul.ul.pt/cards-fly`

2 Segmentation des Textes

La segmentation de lettres historiques est une tâche difficile parce que les outils qui nous permettent normalement d'extraire des informations significatives à partir du lexique des lettres (catégories grammaticales, lemmes) n'existent tout simplement pas. Par conséquent, il faut s'appuyer sur les formes dites de «surface» du mot (le *lexis*), ce qui représente une sérieuse limitation sur la capacité des outils automatiques de faire des généralisations utiles. Par exemple, les noms propres sont plus informatifs comme une catégorie que comme un mot, puisque le nom propre lui-même n'a pas tendance à réapparaître dans les textes. Nous avons tout de même fait une tentative pour étiqueter les noms propres sur la base d'une liste *ad hoc* de 6923 noms portugais provenant de diverses sources locales et en ligne. Les lemmes sont aussi utiles parce que la même information sémantique peut être capturée dans un lemme unique, qui peut être instancié sous plusieurs formes au travers des mots dans les textes. Un sous-ensemble de 402 textes pour la phase d'apprentissage et 100 textes pour la phase de test ont été choisis au hasard dans le corpus CARDS.

2.1 Création des modèles lexicaux

La tâche de segmentation consiste à attribuer à chaque mot (1-gramme) des lettres historiques un seul des quatre étiquettes/segments (*opener*, *harengue*, *peroration* ou *closer*) disponibles. Il est également possible qu'aucune étiquette ne soit attribuée à un mot (*free*). Contrairement à d'autres approches (Sporleder & Lapata, 2006) qui utilisent une variété de bases de connaissances (indicateurs textuels, information reliée à la syntaxe et au discours) ou cherchent à identifier les patrons lexicaux permettant d'identifier les changements de thèmes (Hearst, 1997; Ferret, 2002), nous nous appuyons sur des modèles lexicaux pour chacune des classes que nous cherchons à identifier, ce qui a l'avantage de produire du même coup un vocabulaire pour chaque segment. Notre approche est de récolter tous les n-grammes ($n \leq 3$) dans les données du corpus d'apprentissage et de calculer un score représentant leur saillance dans le segment dans lequel le n-gramme apparaît, ce qui la rattache à la composante *bottom-up* de l'analyse du discours présentée dans (Biber *et al.*, 2007). Nous utilisons le *log odds ratio* (Everitt, 1992) comme mesure statistique de la saillance d'un n-gramme. Le *log odds ratio* compare la fréquence d'occurrence de chaque n-gramme dans un corpus spécialisé à sa fréquence d'occurrence dans un corpus de référence :

$$log\ odds\ ratio = ln(ad/cb) = ln(a) + ln(d) - ln(c) - ln(b)$$

où a est la fréquence du mot dans le corpus spécialisé, b est la taille du corpus spécialisé moins a, c est la fréquence du mot dans le corpus général et d est la taille du corpus général moins c. Une grande valeur de saillance positive indique une saillance forte, alors qu'une grande valeur négative indique un mot sans importance pour le segment. Nous avons construit quatre corpus spécialisés, un pour chacun des quatre segments (*opener*, *closer*, *harengue* or *peroration*) et un pour les mots qui n'appartiennent à aucun segment (*free*). Nous avons adopté le corpus *Tycho Brahé*[2] comme corpus de référence. La comparaison avec un corpus de reference permet non seulement de comparer les classes entre elles, mais aussi de les situer par rapport à un discours *neutre*. Notons qu'il existe d'autres mesures de comparaison de fréquences (Frantzi *et al.*, 2000). Le corpus Tycho Brahé constitue un bon étalon de référence pour trois raisons principales :

1. Il est assez varié selon les genres tandis que CARDS ne dispose que de lettres privées.

[2] `http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html`

Corpus	Nb textes	1-gramme	2-gramme	3-gramme	Nb segments	Nb mots/segment
référence	44	1508386	1268345	1052225	-	-
opener	275	1366	1077	815	275	5.3
closer	343	4070	3585	3141	343	13.6
harengue	121	2788	2597	2408	124	25.6
peroration	231	3223	2890	2582	266	15.4
free	402	111745	105921	100242	-	-

TAB. 1 – Corpus de Référence et Spécialisés

Segment	2-gramme	3-gramme
opener	*Exmo Snr* 'Très Excellent Mr' 12.7	*Illmo e Exmo* 'Très Illustre et Excellent' 12.9
closer	*De V* 'De Vous' 11.2	*De V Exa* 'De Votre Excellence' 10.8
harengue	*saude q* 'santé que' 10.1	*da q me* 'de la que me' 8.6
peroration	*gde a* 'garde à' 13.0	*Ds gde a* 'Dieu garde à' 11.6

TAB. 2 – 2-grammes et 3-grammes les plus saillants

2. Il est presque entièrement constitué par des échantillons du portugais littéraire formel de l'ère moderne ; en revanche, le corpus CARDS est assez varié tant en termes de registre (formel et informel) que de représentativité sociale.
3. Il a été en partie normalisé en fonction de l'orthographe tandis que CARDS maintient systématiquement l'orthographe des manuscrits originaux[3].

Nos 402 lettres servant à l'apprentissage ont été utilisées pour créer les corpus spécialisés en concaténant, pour chaque lettre, tous les mots appartenant à un segment particulier, en préservant la ponctuation et les limites des segments de telle sorte qu'aucun *n* mots consécutifs ne peut appartenir à différentes phrases ou segments. Ces corpus d'apprentissage sont décrits dans le tableau 1. Les saillances pour chaque n-gramme ont ensuite été calculées et triées du plus grand au plus petit. Dans le tableau 2, nous affichons pour chaque segment le 2-gramme et le 3-gramme le plus saillant ainsi que sa valeur de saillance.

2.2 Classer chaque mot

Les listes de n-grammes avec des valeurs de saillance pour chaque segment constituent nos modèles lexicaux pour notre classificateur. Pour savoir à quel segment un mot appartient, le classificateur adopte la stratégie en deux étapes suivante :

1. On attribue à chaque 1-gramme les valeurs de saillance pour chaque segment que l'on peut trouver dans les modèles, zéro sinon ;
2. Chaque mot d'un n-gramme (n $\in$ (2,3)) voit sa valeur de saillance augmentée par la valeur de saillance correspondant aux n-grammes dans les modèles, si elles existent.

La stratégie ci-dessus peut être limitée à un sous-ensemble d'un, deux ou trois modèles. Par conséquent, chaque mot a une valeur de saillance pour chaque segment, et peut prendre en compte des informa-

[3]Nous avons inclu des textes de référence non normalisés pour éviter de se retrouver avec un corpus de référence trop petit.

n-gramme utilisé(s)	F-scores %				Exactitude générale %
	opener	harengue	peroration	closer	
{**1**}	4	15	6	25	53
{**2**}	9	20	8	33	62
{**3**}	30	16	14	24	88
{**1,2**}	5	14	8	25	47
{**1,3**}	4	18	8	29	53
{**2,3**}	9	21	8	35	63
{**1,2,3**}	5	15	9	26	47

TAB. 3 – F-scores et exactitudes pour la classification des mots

tions contextuelles (si les modèles supérieurs à 1-gramme ont été inclus dans le procédé de calcul décrit précédemment). Le segment final est celui correspondant à la saillance la plus élevée, diminuée de la valeur de la saillance des autres classes. L'évaluation sur 100 lettres sont présentés dans le tableau 3.

2.3 Production de segments

L'approche précédente pour classer les mots d'une lettre sur une base individuelle ne peut pas toujours produire des regroupements d'étiquettes continus semblables à de vrais segments, il y a donc inévitablement des discontinuités entre des groupes de mots distants ayant la même étiquette. Regardons l'exemple suivant[4], où l'indice[5] indique une étiquette proposée par le classifieur et où les balises indiquent la vraie classe telle qu'annotée par les humains.

> <**opener**> Meo$_o$ amo$_o$ e$_c$ Sr$_o$ <**/opener**> <**harengue**>Ainda$_c$ q$_h$ VM$_f$ me$_h$ não$_h$ quer$_h$ dar$_h$ o$_c$ alivio$_h$ de$_h$ suas$_h$ novas$_h$ a$_h$ minha$_h$ amizade$_h$ não$_h$ pide$_h$ tal$_h$ discuido$_c$ e$_h$ assi$_h$ lembresse$_h$ VM$_h$ de$_o$ mim$_f$ q$_h$ com$_h$ novas$_h$ suas$_h$ q$_h$ bem$_h$ sabe$_h$ q$_f$ não$_h$ tem$_h$ qm$_p$ lhas$_h$ dezeje$_h$ com$_h$ mais$_h$ veras$_p$. <**/harengue**> Sabado$_f$ nove$_f$ deste$_f$ mes$_f$ ... por$_f$ não$_f$ ficar$_f$ com$_f$ escrupello$_f$ <**peroration**> aqui$_p$ fico$_p$ ás$_h$ ordens$_p$ de$_p$ VM$_p$ pa$_p$ o$_f$ q$_p$ me$_p$ quizer$_p$ mandar$_p$ com$_f$ gde$_p$ vontade$_p$ Ds$_p$ gde$_p$ a$_p$ VM$_p$ <**/peroration**> <**closer**> Prada$_f$ 10$_f$ de$_f$ Julho$_c$ de$_c$ 1712$_f$ Mayor$_f$ Amo$_c$ e$_c$ Servidor$_c$ de$_f$ VM$_c$ Frando$_f$ de$_c$ Sá$_f$ Menezes$_f$ <**/closer**>

Bien que les patrons d'étiquettes calculés suivent à peu près l'annotation «vraie», une technique de lissage pourrait être appliquée pour tenter de rattacher les ilôts d'étiquettes disparates et de créer des segments proches de ceux créés par des annotateurs humains. Pour obtenir un lissage des patrons «segmentaires» obtenus par l'étiquetage automatique de chaque mot individuellement, nous choisissons un intervalle pour la longueur de chaque segment de telle sorte que 95% des valeurs moyennes (pour la longueur) se trouvent dans cette intervalle. Ceci est donné dans la distribution normale par le calcul (moyenne $\mp$ 2 * écart-type)[6].

[4]Traduction française : <**opener**> Mon ami et Seigneur <**/opener**> <**harengue**>Bien que votre Grâce ne veut pas me soulager avec des nouvelles de votre Grâce, mon amitié ne demande pas un tel manque d'attention, alors, accordez votre Grâce de moi avec vos nouvelles, parce que vous savez bien qu'il n'y a personne à les désirer plus que moi, vraiment <**/harengue**> Le Samedi 9 de ce mois ... de ne pas avoir des scrupules <**peroration**>ici je reste aux ordres de votre Grâce pour ce que votre Grâce le veuille ordonner, de toute ma volonté, Dieu garde à votre Grâce <**/peroration**> <**closer**> Prada, le 10 juillet de 1712 Le plus grand ami et serviteur de votre Grâce Fernando de Sá Menezes <**/closer**>

[5]o=opener, c=closer, h=harengue, p=peroration and f=free

[6]Nous avons calculé la moyenne et l'écart-type à partir des données d'apprentissage.

n-gramme utilisé(s)	F-scores %				Exactitude générale %
	opener	harengue	peroration	closer	
{**1**}	2	10	4	35	26
{**2**}	6	26	6	41	33
{**3**}	31	23	12	28	79
{**1,2**}	2	14	6	31	25
{**1,3**}	2	21	7	38	27
{**2,3**}	7	31	6	41	33
{**1,2,3**}	3	15	7	32	26

TAB. 4 – F-scores et exactitudes pour chaque mot après la production de segments

On obtient donc les valeurs d'intervalles suivantes : [1,15] pour *opener*, [1,60] pour *harengue*, [1,43] pour *peroration* et [1,28] pour *closer*. Cela signifie que nous allons examiner seulement les *opener* dont la taille varie entre 1 et 15 mots, etc.

À partir du premier mot de chaque lettre historique, on calcule un score pour chaque segment de chacune des quatre classes, compte tenu de la longueur des intervalles définis précédemment pour chaque segment. Les scores de chaque segment sont obtenus en retenant la classe pour laquelle la somme S des scores de chaque mot dans le segment est la plus élevée. En d'autres termes, un segment de N mots consécutifs est susceptible d'être étiqueté avec la classe C si elle a beaucoup de mots avec des valeurs de saillance élevées pour C. Nous conservons les segments au-dessus d'un certain seuil pour S et qui ne se chevauchent pas. L'évaluation de ce classificateur utilisant la même métrique que dans la section 2.2 sont indiqués dans le tableau 4.

2.4 Remarques

Bien que l'exactitude du classifieur soit nettement supérieure à celle d'une base aléatoire (cinq classes $\Rightarrow$ 20%), les valeurs pour les F-scores indiquées dans les tableaux 3 et 4 concernant les mots appartenant à l'un des quatre segments d'intérêt sont décevantes mais pas surprenantes : les lettres historiques présentent en effet un grand nombre de variantes orthographiques. Nous avions également à notre disposition un corpus d'apprentissage plutôt petit (402 textes). Néanmoins, nous pensons que notre approche basée sur la fréquence de comparaison des n-grammes et de lissage pour la création de véritables segments est un bon point de départ. Les résultats présentés dans les tableaux 3 et 4 nous permettent également de formuler trois observations intéressantes :

– Les 3-grammes et 2-grammes permettent d'établir une meilleure discrimination entre les quatre classes.
– *Harengue* et *closer* sont les classes qui peuvent être le plus facilement discriminées.
– Le lissage pour produire des segments plus réalistes permet d'améliorer la classification de chaque mot dans le cas de *harengue* et *closer*.

D'autre part, l'analyse des n-grammes les plus saillants nous a permis de faire les constatations suivantes :

– *opener* : la sémantique du respect social exprimé par des formes de courtoisie nominales (ce qui équivaut à *Très Excellent Monsieur*)
– *harengue* : la sémantique de la santé, combinée avec des verbes psychologiques et des expressions phatiques, typique des formules de souhaits dans les débuts de dialogue (équivalent à *J'espère que vous*

êtes en bonne santé)
– *peroration* : la sémantique de la religion, aussi combinée avec des expressions phatiques, typique de l'invocation de Dieu dans les fins de dialogue (l'équivalent de *Que Dieu soit avec vous*)
– *closer* : de nouveau la sémantique du respect social, exprimée ici par des formes adjectivales et nominales d'autodérision (équivalent à *Je suis votre humble serviteur*).

Notons que pour évaluer spécifiquement notre modèle lexical, nous avons préféré ne pas exploiter l'information relative au positionnement normal des segments dans les textes. Finalement, nous avons l'intention d'évaluer notre système avec des mesures classiques en segmentation (Sitbon & Bellot, 2006).

3 Conclusion

Nous avons présenté une étude visant à segmenter automatiquement des lettres historiques selon quatre classes. Étant donné l'absence d'outils permettant d'extraire des informations linguistiques sur lequel s'appuyer, nous avons adopté une approche essentiellement statistique, sur la base de modèles lexicaux et d'une comparaison fréquentielle avec un corpus de référence, ce qui nous a permis de voir les limites d'une telle approche, mais aussi de faire une analyse résolument objective de certaines caractéristiques des textes anciens. Nous pensons qu'avec l'assistance d'outils permettant l'acquisition d'information linguistique, les performances d'une telle approche peuvent être grandement améliorées.

De façon plus générale, le traitement informatisé des données historiques, tels que les lettres des sociétés du passé, permet d'établir une base de comparaison avec des échantillons comparables contemporains. Au cours du XXe siècle, des millions de lettres ont été écrites dans les sociétés occidentales, et certaines de ces lettres ont survécu. En termes d'étude du changement linguistique et social, des éléments de preuve comparables provenant du passé et du présent sont nécessaires, et la technologie informatique semble être un outil indispensable pour la réalisation de cet objectif.

Références

BIBER D., CONNER U. & UPTON T. (2007). *Discourse on the move : Using corpus analysis to describe discourse structure.* Amsterdam : John Benjamins.

EVERITT B. (1992). *The Analysis of Contingency Tables*. Chapman and Hall, 2nd edition.

FERRET O. (2002). Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. In *TALN 2002*, Nancy.

FRANTZI K., ANANIADOU S. & MIMA H. (2000). Automatic recognition of multi-word terms : the C-value/NC-value Method. *International Journal on Digital Libraries*, **3**(2), 115–130.

HEARST M. A. (1997). TextTiling : Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, **23**(1), 33–64.

SITBON L. & BELLOT P. (2006). Tools and methods for objective or contextual evaluation of topic segmentation. In *Proceedings of Language Resources and Evaluation (LREC) 2006*.

SPORLEDER C. & LAPATA M. (2006). Broad coverage paragraph segmentation across languages and domains. *ACM Trans. Speech Lang. Process.*, **3**(2), 1–35.