Segmentation Multilingue des Mots Composés

Elizaveta Loginova-Clouet[1] Béatrice Daille[1]
(1) LINA, 2, rue de la Houssinière 44322 Nantes Cedex 03
elizaveta.loginova@univ-nantes.fr, beatrice.daille@univ-nantes.fr

RÉSUMÉ

La composition est un phénomène fréquent dans plusieurs langues, surtout dans des langues ayant une morphologie riche. Le traitement des mots composés est un défi pour les systèmes de TAL car pour la plupart, ils ne sont pas présents dans les lexiques. Dans cet article, nous présentons une méthode de segmentation des composés qui combine des caractéristiques indépendantes de la langue (mesure de similarité, données du corpus) avec des règles de transformation sur les frontières des composants spécifiques à une langue. Nos expériences de segmentation de termes composés allemands et russes montrent une exactitude jusqu'à 95 % pour l'allemand et jusqu'à 91 % pour le russe. Nous constatons que l'utilisation de corpus spécialisés relevant du même domaine que les composés améliore la qualité de segmentation.

ABSTRACT

Multilingual Compound Splitting

Compounding is a common phenomenon for many languages, especially those with a rich morphology. Dealing with compounds is a challenge for natural language processing systems since all compounds can not be included in lexicons. In this paper, we present a compound splitting method combining language independent features (similarity measure, corpus data) and language dependent features (component transformation rules). We report on our experiments in splitting of German and Russian compound terms giving accuracy up to 95% for German and up to 91% for Russian language. We observe that the usage of a corpus of the same domain as compounds improves splitting quality.

MOTS-CLÉS : segmentation des mots composés, outil multilingue, mesure de similarité, règles de transformation des composants, corpus spécialisés.

KEYWORDS: compound splitting, multilingual tool, similarity measure, component transformation rules, specialized corpora.

1 Introduction

La composition est un mécanisme de formation des mots qui consiste à combiner deux (ou plusieurs) éléments lexicaux autonomes pour former une unité de sens. Ce phénomène est notamment présent dans les langues allemande, néerlandaise, grecque, suédoise, danoise, finlandaise et russe. Le traitement des mots composés est une difficulté pour les systèmes de traitement automatique des langues parce que la plupart des composés ne sont pas recensés dans les ressources lexicales. Ainsi leur reconnaissance et leur segmentation seraient bénéfiques pour des tâches variées du TAL : traduction automatique (Macherey *et al.* (2011), Weller et Heid (2012)), recherche d'information (Braschler et Ripplinger, 2004), recherche d'information multilingue (Chen et Gey, 2001), etc.

Les mécanismes de composition sont plus ou moins complexes en fonction des langues. Dans les langues très analytiques comme les langues française et anglaise les composants sont simplement concaténés : FR *kilowatt-heure*, EN *parrotfish*[1], « poisson perroquet ».

Dans les langues ayant une morphologie riche, des transformations sont possibles aux frontières des parties composantes. La terminaison du mot peut être omise, et/ou des morphèmes « frontières » rajoutés, par exemple en allemand :

Staat<u>s</u>feind (« ennemie d'état ») = Staat (« état ») + Feind (« ennemie ») ;

Pour certaines langues, les règles sont peu nombreuses et exhaustives. Pour d'autres, des phénomènes plus complexes interviennent comme la modification du radical en russe :

ветрогенератор (« générateur éolien »)
vet<u>ro</u>generator[2] = vet<u>er</u> (« vent ») + generator (« générateur ») ;

Les « composés néoclassiques », c'est-à-dire des composés ayant un ou plusieurs éléments d'origine latine ou grecque (Namer, 2009), sont un cas particulier de composition où les éléments lexicaux ne sont pas autonomes : FR *multimédia*, DE *Turbomaschine* (« turbomachine »), etc. Ces éléments néoclassiques sont généralement absents des dictionnaires ou des bases de données lexicales.

Certains systèmes de TAL optent pour le stockage de tous les composants connus dans le lexique (à notre connaissance, c'est généralement le cas des systèmes pour le russe). Cette solution nous semble insatisfaisante pour des tâches multilingues car ceci augmente considérablement la couverture du dictionnaire.

Dans cet article, nous faisons le tour d'horizon des méthodes de segmentation automatique des mots composés. Ensuite, nous proposons une méthode combinant des traits dépendants et indépendants de la langue. Enfin, nous présentons nos expériences de segmentation des composés allemands et russes.

2 Méthodes de segmentation des mots composés

Parmi les méthodes de segmentation des composés, on peut distinguer les méthodes utilisant des règles formulées manuellement et des méthodes complètement statistiques.

1. EN - langue anglaise, DE - langue allemande, RU - langue russe
2. Les exemples russes sont translittérés.

Le premier type de méthodes définit des règles de segmentation telles que celles de transformations aux frontières des composants en allemand. Généralement celles-ci utilisent des règles de formation des composés décrites par Langer (1998).

Pour choisir parmi plusieurs segmentations, les composants ainsi identifiés sont ensuite recherchés soit dans un dictionnaire (segmenteur Banana Split [3]), soit dans un corpus monolingue (Koehn et Knight (2003), IMS Splitter [4]). Les approches basées sur le corpus affectent également une probabilité à chaque segmentation, estimée sur la base de la fréquence des composants dans le corpus. Un corpus parallèle allemand-anglais peut être exploité afin d'y vérifier les correspondances des parties décomposées (Koehn et Knight, 2003).

Les approches du deuxième groupe ne requièrent pas de règles spécifiques pour chaque langue donnée. Macherey *et al.* (2011) proposent d'extraire automatiquement des opérations morphologiques sur les frontières de composants. L'entraînement du modèle pour une nouvelle langue nécessite un corpus parallèle contenant une partie anglaise. Hewlett et Cohen (2011) détectent automatiquement la place des frontières de composants. L'algorithme est basé sur la probabilité des séquences de caractères dans une langue.

Actuellement, les modèles purement statistiques ne sont pas aussi précis que des modèles utilisant des règles, leur avantage réside toutefois dans la possibilité de réutilisation pour des langues variées.

3 Algorithme de segmentation

Notre objectif est de créer un outil de segmentation des mots composés générique et multilingue qui pourrait être appliqué à des différentes langues grâce aux traits indépendants de la langue sans nécessiter de connaissances préalables. Néanmoins si des règles existent, cet outil doit être capable de les intégrer. Les caractéristiques indépendantes de la langue exploitées sont la fréquence des mots dans un corpus monolingue, et la similarité entre une sous-chaîne du mot et les lemmes candidats.

Pour segmenter un composé, nous commençons par générer toutes ses segmentations possibles en deux parties, de taille supérieure ou égale à la longueur minimale acceptée pour un composant. Par exemple DE *Traktionsbatterie* (« batterie de traction ») :

traktionsbatterie → tr + aktionsbatterie
traktionsbatterie → tra + ktionsbatterie
...
traktionsbatterie → traktionsbatter + ie

Si des règles de transformation des composants en lexèmes indépendants sont disponibles pour la langue donnée, elles sont appliquées aux composants candidats. Ce sont des règles de type : *« s » → « »*, (cf. DE exemple *Staatsfeind*), *« en » → « um »*, etc.

Pour chaque segmentation candidate, les deux parties sont recherchées dans un dictionnaire monolingue, et optionnellement dans un corpus monolingue. Le corpus permet de calculer les fréquences des mots, ce qui aide à choisir les composants candidats les plus plausibles lorsque plusieurs variantes sont possibles.

3. http://niels.drni.de/s9y/pages/bananasplit.html
4. http://www.ims.uni-stuttgart.de/~wellermn/tools.html

Nous calculons ensuite la similarité entre chacune des deux parties de segmentation et les lemmes du dictionnaire/corpus afin de choisir les lemmes « les plus proches ». Nous utilisons « la distance d'édition normalisée » basée sur la distance de Levenshtein comme mesure de similarité (pour la description détaillée des mesures existantes cf. (Frunza et Inkpen, 2009)) :

$$sim(X, Y) = 1 - \frac{nbEditOper}{max(length(X), length(Y))}$$

où nbEditOper est le nombre minimal d'opérations d'édition (substitution, suppression, insertion) nécessaires pour transformer un composant X en un lemme Y.

Si certains lemmes sont acceptables (i.e. avec une similarité supérieure ou égale à un seuil défini) pour la partie gauche de la segmentation courante, mais non pour la partie droite, nous réitérons la segmentation jusqu'à trouver des composants attestés ou jusqu'à un nombre maximal de composants.

RU килоэлектронвольт (« kiloélectronvolt ») :
kiloelektronvolt → kilo + elektronvolt
elektronvolt → elektron + volt

Dans le cas où des lemmes candidats sont acceptables pour chaque composant, nous calculons le score de cette segmentation à chaque niveau de décomposition :

$$S(segm) = \begin{cases} \frac{S(compA)+S(compB)}{2} & \text{si correspondance exacte} \\ \frac{S(compA)+S(compB)}{nbComp} & \text{sinon} \end{cases}$$

où nbComp est le nombre de composants dans le mot, et « correspondance exacte » signifie que tous les composants ont été trouvés en l'état dans le dictionnaire/corpus. Le score d'un composant est calculé de la manière suivante :

$$S(comp) = sim(comp, lemma)^{nbComp} \times (inDico + inCorpus + freqCorpus)$$

où inDico et inCorpus sont des valeurs attestant la présence ou l'absence du lemme dans le dictionnaire et le corpus, et freqCorpus est égale à la fréquence relative du lemme dans le corpus. La mesure de similarité est élevée à la puissance nbComp pour augmenter son impact lorsque le niveau de décomposition croît : plus il y a de composants dans la segmentation candidate, plus il est accordé d'importance au fait que les composants soient proches des lemmes trouvés (le cas le plus favorable étant celui d'une mesure de similarité égale à 1).

Enfin, l'algorithme retourne le Top N des meilleures segmentations classées par score décroissant. Par exemple, pour DE *Traktionsbatterie* (« batterie de traction ») le résultat affiché est le suivant :

traktion + batterie 1.50
trakt + ion + batterie 1.25

La segmentation correcte est *Traktion + Batterie*, et celle-ci obtient le meilleur score d'après le programme.

4 Expériences et données

Dans cette section, nous décrivons nos expériences en utilisant le précédent algorithme. Jusqu'à présent il a été appliqué à deux langues : l'allemande et le russe. La composition en

allemand est très productive et bien décrite. La composition en russe l'est moins, même si elle est plus fréquente dans les domaines de spécialité que dans la langue générale.

Pour les deux langues, nous avons analysé des mots composés appartenant au domaine de l'énergie éolienne. Pour la langue allemande, nous avons pris comme jeu de tests 445 composés extraits des expériences de Weller et Heid (2012) [5]. Pour la langue russe, nous avons compilé le jeu de tests à partir d'un corpus de l'énergie éolienne [6]. Parmi les 7 000 lexèmes les plus fréquents du corpus, 348 sont des composés.

Nous avons fait varier les paramètres pour observer l'impact de l'utilisation du corpus et des règles de transformation sur la qualité de segmentation. Comme la segmentation de base, nous avons retenu la segmentation avec le dictionnaire, ce qui correspond à la technique utilisée dans les systèmes n'ayant pas de module élaboré de segmentation. Nous avons enrichi cette segmentation de base premièrement avec la prise en compte de règles de transformation et l'utilisation de la mesure de similarité, et deuxièmement avec le filtrage dans le corpus.

Pour l'allemand, nous avons utilisé la partie allemande du dictionnaire libre allemand-anglais Dict.cc [7]. Pour le russe, nous avons exploité la version électronique du dictionnaire de Ozhegov [8], complétée par une liste d'éléments néoclassiques extraits du travail de Béchade (1992) et traduits en russe. Les éléments néoclassiques sont très fréquents dans les composés russes et leur repérage s'avère nécessaire pour une segmentation correcte. Comme nous travaillons avec des composés spécialisés, nous avons exploité des corpus thématiques du domaine de l'énergie éolienne compilés à partir du web [9] (environ 300 000 mots pour le russe et 1.7 million mots pour l'allemand) et lemmatisés par TreeTagger [10].

Les règles pour l'allemand sont basées sur (Langer, 1998). Pour la langue russe nous avons testé deux jeux de règles. Le premier jeu contient deux règles exprimant une connaissance basique du russe selon laquelle les morphèmes « o » and « e » servent de morphèmes « frontières » pour des composés. Le jeu de règles élargi (13 règles) intègre des connaissances morphologiques approfondies extraites de (Zaliznjak, 1977).

Un paramètre important pour notre algorithme est le seuil de similarité qui désigne la valeur minimale acceptable de similarité entre un composant candidat et un lemme du dictionnaire/corpus. Pour trouver la valeur optimale, nous avons testé l'algorithme avec des seuils différents sur le même corpus de l'énergie éolienne (cf. Figure 1). Sur nos données la valeur de 0.7 s'avère la plus satisfaisante pour les deux langues.

Pour évaluer les résultats, nous avons calculé l'exactitude (EN « accuracy ») de décomposition en position 1 (« Top 1 ») et en position 5 (« Top 5 ») dans la liste de segmentations candidates classées par l'algorithme. L'exactitude est obtenue en divisant le nombre de composés qui ont une segmentation correcte dans Top N produit par l'algorithme par le nombre total de composés. Jusqu'à présent, nous avons effectué l'évaluation seulement sur des mots composés, et nous n'avons pas évalué le bruit introduit par les faux positifs (non-composés qui sont segmentés par l'algorithme par erreur). L'identification des composés potentiels d'une langue relève d'une autre problématique.

5. http://www.ims.uni-stuttgart.de/~wellermn/tools.html
6. http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html
7. http://www.dict.cc
8. http://speakrus.ru/dict/ozhegovw.zip
9. http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html
10. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

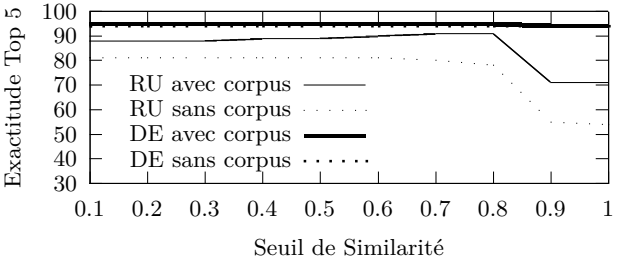

FIGURE 1 – Exactitude de la segmentation des mots composés (Top 5) en fonction du seuil de similarité

	Dictionnaire	Dictionnaire + Règles + Similarité	Dictionnaire + Règles + Similarité + Corpus	Banana Split	IMS Splitter
Top 1	57 %	91 %	91 %	86 %	87 %
Top 5	57 %	94 %	95 %	-	92 %

TABLE 1 – Exactitude de segmentation pour l'allemand

5 Résultats

Les résultats de la segmentation pour les langues allemande et russe sont présentés dans les tableaux 1 et 2.

5.1 Composés allemands

Les résultats en ajoutant les règles de transformation et la mesure de similarité sont nettement meilleurs que ceux obtenus dans l'expérience de base (seulement avec le dictionnaire).

L'utilisation du corpus améliore légèrement l'exactitude pour le Top 5. Cela permet la segmentation correcte d'un nombre supérieur de mots dont les composants ne sont pas présents dans le dictionnaire (*Netzanschluß*, « connexion réseau »). Dans certains cas, cela améliore aussi le classement : *Traktionsbatterie* sans corpus retourne deux segmentations classées à égalité *traktion+batterie 1.0* et *trakt+ion+batterie 1.0*. L'utilisation de corpus fait apparaître la segmentation correcte avant celle incorrecte : *traktion+batterie 1.50*, *trakt+ion+batterie 1.25*.

Dans d'autres cas le corpus nuit au classement parce qu'il favorise les segmentations constituées de composants plus courts et plus fréquents : *Aussichtsplattform*, « observation deck », est correctement segmenté sans corpus en *aussicht+plattform*, alors qu'avec le corpus la meilleure segmentation est *aus+sicht+plattform*. Ce problème peut être résolu en remplaçant la fréquence simple du corpus par la spécificité qui rend compte du caractère terminologique des composés. La spécificité est obtenue en divisant la fréquence dans le corpus spécialisé par la fréquence dans un corpus général (Ahmad *et al.*, 1992).

	Dictionnaire	Dictionnaire + Règles + Similarité		Dictionnaire + Règles + Similarité + Corpus	
		Règles restreintes	Règles élargies	Règles restreintes	Règles élargies
Top 1	35 %	62 %	78 %	72 %	82 %
Top 5	35 %	68 %	80 %	81 %	91 %

TABLE 2 – Exactitude de segmentation pour le russe

Nous avons comparé notre outil à deux outils libres disponibles pour l'allemand : Banana Split [11] et IMS Splitter [12]. Sur les mêmes 445 composés, le segmenteur Banana Split donne une exactitude de 86 % pour le Top 1 ; IMS Splitter aboutit à une exactitude de 87 % pour le Top 1 et 92 % pour le Top 5.

5.2 Composés russes

Nous avons observé une différence significative entre les résultats de l'expérience de base et ceux avec les règles et la mesure de similarité (cf. tableau. 2). L'utilisation de corpus a été également bénéfique. Notons que les résultats avec l'utilisation des règles élargies sans corpus sont proches de ceux avec les règles restreintes mais avec corpus. En fait pour certains composés le corpus compense l'absence de règles. Ainsi l'adjectif « électromagnétique » электромагнитный (*elektromagnitnyi*) ne pouvait pas être segmenté correctement avec la méthode de base, parce que son composant de droite *magnitnyi* (« magnétique ») n'est pas présent dans le dictionnaire. Il peut être segmenté soit en utilisant le corpus (où « magnétique » est présent), soit grâce à une règle qui permet de retrouver le nom associé *magnit* (« aimant »).

6 Conclusion

Nous avons présenté un algorithme de segmentation des mots composés combinant des caractéristiques indépendantes de la langue (mesure de similarité, fréquence des mots) avec des caractéristiques dépendantes de la langue (règles de transformation des composants). Cette méthode est beaucoup plus performante que celle de base consistant à vérifier la présence des composants dans un dictionnaire. Elle donne des résultats comparables aux méthodes de segmentation monolingues : pour le Top 5, exactitude jusqu'à 95 % pour l'allemand et jusqu'à 91 % pour le russe.

L'utilisation d'un corpus est globalement bénéfique. Un corpus spécialisé permet de segmenter correctement plus de mots dont les composants sont inconnus du dictionnaire et de filtrer des mauvaises segmentations. Le corpus permet dans une certaine mesure de compenser des règles morphologiques. Il peut cependant dégrader le classement des candidats dans certains cas.

Le code source avec une description détaillée de l'algorithme sont accessibles en ligne [13]. Le programme peut être appliqué à des langues différentes en changeant les sources lexicales

11. http://niels.drni.de/s9y/pages/bananasplit.html
12. http://www.ims.uni-stuttgart.de/~wellermn/tools.html
13. http://www.lina.univ-nantes.fr/?Compound-Splitting-Tool.html

et en ajoutant éventuellement des règles de transformation. Néanmoins un paramétrage préliminaire est préférable pour obtenir de meilleurs résultats pour une nouvelle langue. Nous prévoyons de tester l'algorithme pour d'autres langues et domaines ainsi que d'évaluer l'impact de la segmentation sur la qualité de la traduction automatique.

Remerciements

Les travaux ayant mené à ces résultats ont reçu le financement du programme European Community's Seventh Framework (FP7/2007-2013), sous l'agrément de bourse no. 248005.

Références

Ahmad, K., Davies, A., Fulford, H. et Rogers, M. (1992). What is a term ? the semi-automatic extraction of terms from text. *In Translation Studies : An Interdiscipline*, pages 267–278, Amsterdam/Philadelphia. John Benjamins.

Braschler, M. et Ripplinger, B. (2004). How effective is stemming and decompounding for german text retrieval. *In Information Retrieval*, volume 7, pages 291–316.

Béchade, H.-D. (1992). *Phonétique et morphologie du français moderne et contemporain*. Presses Universitaires de France, Paris.

Chen, A. et Gey, F. (2001). Translation term weighting and combining translation resources in cross-language retrieval. *In Proceedings of TREC Conference*.

Frunza, O. et Inkpen, D. (2009). Identification and disambiguation of cognates, false friends, and partial cognates using machine learning techniques. *In International Journal of Linguistics*, volume 1.

Hewlett, D. et Cohen, P. (2011). Fully unsupervised word segmentation with bve and mdl. *In Proceedings of ACL 2011*, pages 540–545, Portland, Oregon.

Koehn, P. et Knight, K. (2003). Empirical methods for compound splitting. *In Proceedings of EAC 2003*, Budapest, Hungary.

Langer, S. (1998). Zur Morphologie und Semantik von Nominalkomposita. *In Proceedings of KONVENS 1998*, pages 83–97, Bonn.

Macherey, K., Dai, A., Talbot, D., Popat, A. et Och, F. (2011). Language-independent compound splitting with morphological operations. *In Proceedings of ACL 2011*, pages 1395–1404, Portland, Oregon.

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues*. Lavoisier, Paris.

Ott, N. (2005). Measuring semantic relatedness of german compounds using germanet. `http://niels.drni.de/n3files/bananasplit/Compound-GermaNet-Slides.pdf`. [consulté le 20/03/2013].

Weller, M. et Heid, U. (2012). Analyzing and aligning german compound nouns. *In Proceedings of LREC 2012*, Istanbul.

Zaliznjak, A. A. (1977). *Grammaticheskij Slovar' Russkogo Jazyka [Grammatical Dictionary of the Russian Language]*. Russkij jazyk, Moscow.