Segmentation Multilingue des Mots Composés

Elizaveta Loginova Clouet, Béatrice Daille

(Juin 2013)

Cécile MARTIN

M1 LFC

2018/2019

Introduction

- **Qu'est-ce que la composition**:

  Mécanisme de formation: combination 2 ou plusieurs éléments lexicaux autonomes pour former une unité de sens.

- **Pourquoi le traiter en TAL?**

➢ Sujet difficile: la plupart des composés ne sont pas indexés dans les ressources lexcicales.

➢ Les usages: la Traduction automatique, la Recherche d'information, la Recherche d'informations multilingue, etc.

- **Selon les langues, les mécanismes de compositions sont plus ou moins complexes.** Cette analyse se porte surtout sur l'allemand et le russe.

Compositions

- **En français et anglais:** les composants sont concaténés

  → Fr: kilowatt-heure; En: parrotfish "poisson perroquet"

- **Les langues avec une morphologie riche, les frontières ne sont pas toujours bien définies.** (Terminaison de mot omise et/ou des morphèmes rajoutés)

  → De: Staatsfeind ("ennemie d'état") = Staat ("état") + Feind ("ennemie")

- **Cas particuliers: les "composés néoclassiques"** = composés ayant une origine latine ou grecque car pas autonomes. Souvent absents des bases de données.

  → Fr: multimédia; De: Turbomaschine

- **Certains systèmes TAL = stockage des composants connus dans le lexique.** Mais peu pratique pour du multilingue: trop grande couverture du dictionnaire.

Méthode

- **2 types de segmentations: manuelle et statistique.**

➢ **1. Manuelle:**

- définit des règles de segmentations (ex: transformations des composants aux frontières en allemand).

- composants identifiés soit dans un dictionnaire (Banana Split), soit dans un corpus monolingue (IMS Splitter).

- probabilité à chaque segmentation basée sur la fréquence du corpus.

Méthode

- **2. Statistiques:**

  - pas de règles spécifiques pour chaque langue

  - extraction automatique des opérations morphologiques sur les frontières de composants

  - Pour une nouvelle langue: besoin d'un corpus parallèle avec une partie anglaise.

  - algorithme basé sur la probabilité des séquences de caractères dans une langue

  - **modèle pas aussi précis** que le manuel. Avantage: réutilisation pour des langues variées.

Algorithme de segmentation (1/3)

- **Objectif:** créer un outil de segmentation des mots composés génériques et multilingue qui pourrait être appliqué à toutes les langues sans avoir besoin de connaissances au préalable. Un outil capable d'intégrer les règles préexistantes dans une langue.

- **Caractéristiques indépendantes de la langue exploitée:**
  - Fréquence des mots dans un corpus monolingue
  - Similarité entre une sous-chaine du mot et les lemmes candidats

Algorithme de segmentation (2/3)

- **Méthode de segmentation:**

  - Pour segmenter un composé: génération de toutes les segmentations possibles en deux parties de taille supérieure ou égale à la longueur acceptée par un composant.

Ex: De: Traktionsbatterie (“batterie de traction”)

Traktionsbatterie → tr + aktionsbatterie

Traktionsbatterie → tra + ktionsbatterie

…

Traktionsbatterie → traktionsbatter + ie

Algorithme de segmentation (3/3)

- Si des règles de transformations existent déjà, alors elles sont appliquées. Règles du type “s” → “ ” (“Staastsfeind”), etc.

- Pour chaque segmentation candidate: les parties recherchées dans un ditionnaire ou corpus monolingue.

- Lorsque plusieurs variantes possibles: le corpus calcule les fréquences de mots de ceux qui sont les plus probables.

Calculs

- **Calculs de fréquences pour voir quels composés sont les plus probables:**

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

→ nbEditOper: nombre minimal d'opérations d'éditions (substitution, suppression, insertion) nécessaires pour transformer un composant X en un lemme Y.

→ La segmentation est réitérée jusqu'à ce que les parties (à gauche et à droite) trouvent un composant attesté.

Calculs

- **Quand les lemmes sont acceptés = calcul du score de segmentations à chaque niveau de décomposition:**

$$S(segm) = \begin{cases} \frac{S(compA)+S(compB)}{2} \\ \frac{S(compA)+S(compB)}{nbComp} \end{cases}$$

→ nbComp: est le nombre de composant dans le mot

→ Besoin de trouver une coorespondance exact

- **Calcul du score du composant:**

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

- **L'algorithme donne le Top N des meilleurs segments (ordre décroissant).** Ex: traktion + batterie 1.50 / trakt + ion + batterie 1.25

Expérience

- **Application des algorithmes en allemand et russe (domaine: énergie éolienne)**

→ Allemand: 445 composés / Russe: 348 composés

→ Variations des paramètres: impact utilisation corpus/règle de transformation sur la qualité de la segmentation.

→ Segmentation avec le dictionnaire enrichi (prise en compte des règles de transformation + filtrage dans le corpus)

→ Explorations de corpus thématiques (énergie éolienne) venant du web et lemmatisé par TreeTagger

Règles

- En Allemand: basées sur les travaux de Langer (1998)
- En Russe:

    1. Connaissance basique: morphèmes “o” et “e” servent de morphèmes frontières pour des composés

    2. Jeu de règles élargi : connaissances morphologiques approfondies

- **Seuil de similarité =** valeur minimale acceptable entre un composant et un lemme du disctionnaire/corpus
- **Évaluation des résultats:** Calcul de l’exactitude de décomposition du Top 1 et Top 5 dans la liste de segmentation.
- Le bruit des faux positifs(mots non-composés segmentés par l’algorithme): pas pris en compte.

Résultats

- **Allemand:**

→Règle de transformation + mesure de similarité = meilleurs résultats que expérience avec juste dictionnaire

→Permet segmentation correcte de plus de mots qui ne sont pas dans le dictionnaire

→Le corpus améliore le classement dans certains cas.

→Dans d'autres cas, favorise composants plus courts et plus fréquents. Résolution problème: remplacer fréquence simple du corpus par spécificité (caractère terminologique composé)

→ Comparaison avec 2 outils libres: Banana Split et IMS Splitter → sur 445 composés = exactitude de plus de 85%

Résultats

- **Russe:**

→Différence significative entre l'expérience de base (dico) et ceux avec les règles et la mesure de similarité

→Corpus = bénéfique → pour certains composés: le corpus compense l'absence de règles

→ Ex: электромагнитный (*elektromagnitnyi)* « électromagnétique »

- magnitnyi *(« magnétique »)* n'existe pas dans dictionnaire: technique de base pas possible

- corpus = segmenté / règle = retrouvé mot associé magnit *(« aimant »)*.

Conclusion

- Présentation d'un algorithme de segmentation des mots composés
  - ✓ Combine caractéristiques indépendantes de la langue (mesure de similarité, fréquence des mots) avec caractéristiques dépendendantes de la langue (règles de transformations des composants)
  - ✓ Méthode plus performante
  - ✓ Résultats comparables aux méthodes de segmentations monolingues
  - ✓ Corpus = globalement bénéfique (surtout corpus spécialisé)
  - ✓ Code source accessible en ligne: application à d'autres langues en changeant les sources lexicales.