TAL Synthèse
Constitution de lexiques: extraction de termes et de collocations, variations de termes

REGUIG Ghiles

April 27, 2017

1 Introduction

Dans le domaine du traitement automatique de la langue, la compréhension sémantique d'un document constitue un enjeu majeur. En effet, le fait de pleinement saisir le sens d'un texte permet, par exemple, de le traduire, d'en extraire des informations, de le résumer, de le classer,... Dans cette optique, on désire souvent traiter un document en le morcelant en termes. On qualifie de terme simple un mot et de terme complexe un ensemble de mots représentant un intérêt sémantique en tant que groupe. Il devient alors impératif de pouvoir correctement identifier et extraire les termes complexes d'un texte afin de pouvoir les traiter à part.
Dans cette synthèse nous aborderons dans un premier temps la problèmatique d'extraction de termes complexes au sein d'un document en présentant 3 approches : l'approche syntaxique, l'approche statistique ainsi que l'approche hybride. Dans un second temps, nous nous intéresserons à l'analyse de la compositionnalité de ces derniers, à savoir s'il est possible de traiter un terme complexe comme un composé de ses parties. Pour cela nous verrons la représentation de termes complexes au sein d'un espace de représentation numérique ainsi que les caractéristiques déduisibles par cette projection.

2 Extraction de termes

Les termes complexes (ou *Multiword Units*, en anglais) peuvent représenter jusqu'à un cinquième de la totalité d'un texte. De plus ces derniers peuvent prendre diverses formes syntaxiques qui peuvent dépendre aussi des langues.
On peut discerner 3 grandes familles de termes complexes :

1. les mots composés tels que *pomme de terre, arc-en-ciel*. Ces derniers sont toujours contigus, ce qui facilite leur extraction, cependant leur signification n'est pas forcèment compositionnelle.

2. les expressions idiomatiques telles que *poser un lapin, casser sa pipe, tomber dans les pommes*. Ces dernières sont délicates à extraire car leur syntaxe est très changeante. De plus, leur signification n'est pas compositionnelle.

3. les collocations telles que *caresser l'espoir, gros fumeur, exercer une profession*. Les composantes de ces dernières n'étant pas nécessairement contigues, il est souvent difficile de les repérer au sein d'un texte. En ce qui concerne l'aspect sémantique, il dépend fortement de la collocation considérée, certaines significations peuvent facilement être déduites de manière compositionnelles tandis que dans certains cas (comme dans une métaphore) il est nécessaire d'étudier le sens figuré des termes. De plus, il est aussi possible de retrouver plusieurs variations de la même collocation se reposant sur l'utilisation de synonymes ou les tournures de phrases (forme passive, forme interrogative,...).

Dans cette partie, nous nous intéressons à 3 approches de reconnaissance de termes complexes en explicitant leurs caractéristiques.

2.1 Approche syntaxique

Une première façon d'aborder le problème est l'approche syntaxique. Cette dernière se base sur une caractérisation purement syntaxique des termes complexes à extraire. On considère, par exemple, un corpus annoté dans lequel on souhaite repérer les termes complexes. On utilise alors un patron morpho-syntaxique, c'est-à-dire une configuration syntaxique spécifique, telle que *Nom + adjectif*, afin d'extraire tous les groupes de mots qui correspondent.
Il est éventuellement possible de rapprocher les variations de termes lorsqu'elles sont basées sur l'utilisation de synonymes en ayant recours à un dictionnaire de synonymes.
Cette approche nécessite obligatoirement l'intervention d'un expert qui devra définir à la main tous les patrons à utiliser. De plus, cette méthode est considérée comme monolinguale car les différents patrons peuvent sensiblement varier d'une langue à l'autre. Enfin, à moins de lister toutes les possibilités qui soient, les patrons pré-définis ne couvrent en général pas toutes les variations, ce qui donne une efficacité assez faible sur les collocations.

2.2 Approche statistique

Une seconde manière d'approcher le problème est la méthode statistique. Dans ce cas, on fait appel à différentes mesures statistiques permettant de repérer les termes complexes. Une architecture typique de ce type d'algorithme se fait en 2 étapes :

1. extraction des candidats : en analysant l'ensemble du corpus, certains termes sont sélectionnés via une mesure statistique, typiquement une mesure de co-occurence (mots apparaissant ensemble). Un raffinement est même possible en prenant en compte les co-occurences au sein de n-grams afin de pouvoir pallier au problème des variations de collocation.

2. sélection des candidats : dans un second temps, un score est attribué à chaque candidat en se basant sur diverses caractéristiques : fréquence, mesure de vraisemblance,... L'algorithme ne retient alors que les candidats ayant un score supérieur à un seuil pré-défini.

Le principal avantage d'une méthode statistique est de s'affranchir aussi bien de l'intervention d'un expert que de la barrière de la langue. En effet, les seuls paramètrages d'un algorithme d'extraction de termes complexes purement statistique sont les mesures statistiques à utiliser.

2.3 Approche hybride [Dia03]

Enfin, la troisième approche, et certainement la plus utilisée en pratique, est dite hybride car elle se base aussi bien sur des mesures statistiques que des aspects linguistiques. L'algorithme se déroule toujours en au moins 2 étapes, à savoir une phase d'extraction ainsi qu'une phase de sélection.
Dans les systèmes les plus récents, l'extraction de termes complexes se fait en général en comptant les co-occurences tout en prenant en compte une fenêtre de mots. Comme cette méthode nécessite une grande puissance de calcul en raison de l'exploration combinatoire, on se limite en général à une fenêtre réduite (jusqu'à 5 mots). Dans ce paradigme, il est aussi possible de faire 2 extractions (une basée sur la syntaxe et l'autre sur l'aspect statistique) afin de les comparer par la suite.
Par la suite, il est encore une fois nécessaire de pouvoir évaluer les différents candidats extraits, cela peut se faire via une mesure de cohésion lexicale, qui peut se baser sur un patron syntaxique, auquel un poids est attribué, ainsi qu'une mesure statistique telle que la vraisemblance. Tout cela donne un score final permettant d'allier aussi bien une cohésion lexicale du terme ainsi qu'une vraisemblance statistique basée soit sur les observations, soit sur un a priori spécifié auparavant.

3 Représentation numérique et compositionnalité

Les récents travaux de recherche ont permis l'émergence d'une représentation des mots sous forme de vecteurs numériques. L'algorithme précurseur est sans doute l'algorithme *word2vec* qui permet, à partir d'un corpus donné, de construire un espace de représentation assignant à chaque mot rencontré un vecteur numérique.
A partir de cet algorithme, plusieurs variantes ont été utilisées, notamment les représentations *phrase2vec* et *doc2vec* permettant, respectivement, de fixer dans l'espace les termes complexes et les documents/paragraphes.
L'espace produit par *word2vec* permet de situer les mots en fonction de leur sémantique. Pour ce faire, on prend en compte des n-grams du corpus et l'on rapproche les mots contenus dans ce dernier, tout en éloignant les autres afin d'éviter une convergence de tous les mots en un même point.
Dans cette partie, nous nous intéressons à la représentation numérique de termes complexes au sein de ces espaces ainsi que l'intérêt sémantique de ces projections.

3.1 Représentation de termes complexes

Etant donné que l'on dispose d'un algorithme permettant de générer un espace où tous les termes simples d'un corpus sont projetés, on peut se demander de quelle manière il est possible de représenter un terme complexe.

1. La manière la plus intuitive est l'approche compositionnelle, c'est-à-dire en faisant la moyenne des représentations des parties du terme complexe. Cette approche a l'avantage d'être simple à mettre en oeuvre. Cependant, la représentation obtenue peut être biaisée par l'utilisation

des parties du terme complexe en dehors de ce dernier. Par exemple, représenter un terme tel que *droit juridique* par la moyenne des représentations de *droit* et *juridique* ne donnera une représentation adéquate que si chaque mot co-occure avec l'autre.

2. La seconde manière de faire consiste à pré-traiter le corpus en entrée de *word2vec* en faisant des termes complexes une seule entité. Ainsi, on transformera *Los Angeles* en *Los_Angeles* afin d'en avoir une représentation précise.

3. Une troisième manière de procéder consiste à prendre en compte non seulement la représentation du terme complexe, mais aussi celle de son contexte (mots qui l'entourent). De cette manière, on affine la connaissance sémantique qui accompagne le terme.
   Ainsi, on ne représente plus le terme par un seul vecteur par rapport à l'espace initial *word2vec*, mais plutôt par une projection dans un sous-espace représentant son contexte. Pour cela, on commence par représenter notre ensemble de mots en concaténant les représentations de chaque partie. On obtient alors un sous-espace correspondant au contexte.On effectue ensuite une réduction de dimensions via une analyse en composantes principales (ACP), afin de garder un maximum d'informations. Enfin, on obtient la représentation de notre terme en en faisant une projection orthogonale par rapport au sous-espace du contexte.

3.2 Compositionnalité[GBV16]

Une fois la représentation des différents termes fixée, on peut alors s'intéresser à l'utilisation de cet espace d'un point de vue sémantique.
Les deux premières projections sont assez similaires, bien que théoriquement la seconde permette d'affranchir les termes complexes de leur compositionnalité et ainsi de pouvoir les situer indépendamment de leurs parties.
Dans notre cas nous nous intéressons à la troisième représentation, particulièrement pertinente car elle capture les caractéristiques les plus représentatives du contexte du terme. La projection orthogonale qui en suit permet, via une similarité cosine, de savoir à quel point le terme est sémantiquement proche de son contexte.
A partir de cette représentation, il a été possible d'approcher voire de dépasser l'état de l'art sur les tâches de détection de compositionnalité de termes complexes dans un certain contexte (savoir si la signification d'un terme complexe correspond à celles de ses parties), détection de métaphores ainsi que détection de sarcasme sur des tweets.

4 Conclusion

En définitive, l'extraction et le traitement de termes complexes, notamment dans le cas d'une représentation numérique, a permis de grandes avancées dans l'utilisation des caractéristiques sémantiques d'un texte.
De plus, avec l'utilisation d'une représentation sous forme vectorielle, on s'affranchit de toutes les ressources linguistique et des barrières de langues. Une piste de recherche intéressante serait un raffinement des représentations des termes afin de traiter d'autres tâches.

References

[Dia03] Gaël Dias. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 41–48. Association for Computational Linguistics, 2003.

[GBV16] Hongyu Gong, Suma Bhat, and Pramod Viswanath. Geometry of compositionality. *arXiv preprint arXiv:1611.09799*, 2016.