Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances

Loïc Vial, Benjamin Lecouteux, Didier Schwab

Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances

Loïc Vial, Benjamin Lecouteux, Didier Schwab
LIG - GETALP, Univ. Grenoble Alpes, France

Désambiguïsation lexicale

La **désambiguïsation lexicale** consiste à attribuer le sens le plus probable à un mot donné dans un document, à partir d'un inventaire prédéfini de sens.
La désambiguïsation à base de similarités sémantiques utilise :

1. Un algorithme **local**, ou **mesure de similarité**, qui calcule un score de similarité $sim(S_1, S_2)$ entre deux sens.
2. Un algorithme **global**, qui va chercher la meilleure combinaison de sens à l'échelle du document, en utilisant l'algorithme local.

Mesures de similarité usuelles

Les **mesures de similarité** basées sur les définitions contenues dans un dictionnaire les plus utilisées sont :

— L'algorithme de **Lesk** [1], qui calcule le nombre de mots en commun entre les deux définitions : $Lesk(S_1, S_2) = |D(S_1) \cap D(S_2)|$,
avec $D(S) = \{w_0, w_1, \ldots, w_n\}$ la définition du sens S dans le dictionnaire

— L'algorithme de **Lesk étendu** [2], qui prend aussi en compte tous les sens proches dans un réseau lexical :
$ExtLesk(S_1, S_2) = |(D(S_1) \cup D(rel(S_1))) \cap (D(S_2) \cup D(rel(S_2)))|$,
avec $rel(S)$ l'ensemble des sens reliés à S à travers un lien explicite dans WordNet[3].

Notre mesure de similarité

Notre nouvelle mesure de similarité $VecLesk(S_1, S_2)$ prend en compte les sens proches dans notre modèle de vecteurs de sens, filtrés en fonction de leur similarité cosinus avec le **vecteur du lemme de S** au dessus d'un **seuil δ_1** et avec le **vecteur de S** au dessus d'un **seuil δ_2**.
Elle est définie formellement de la façon suivante :

$$VecLesk(S_1, S_2, \delta_1, \delta_2) = |(D(S_1) \cup D(rel(S_1, \delta_1, \delta_2))) \cap (D(S_2) \cup D(rel(S_2, \delta_1, \delta_2))|$$

$$rel(S, \delta_1, \delta_2) = \{S' \mid cosine(\phi(lemma(S)), \phi(S')) > \delta_1, cosine(\phi(S), \phi(S')) > \delta_2\}$$

Nous nous passons ainsi de tout réseau lexical manuellement créé pour étendre la mesure de Lesk.

Vecteurs de sens

La définition du vecteur du sens S, notée $\phi(S)$ correspond à :

$$\phi(S) = \sum_{i=0}^{n} (\phi(w_n) \times poids(pos(w_n)) \times idf(w_n))$$

— $D(S) = \{w_0, w_1, \ldots, w_n\}$ la définition du sens S dans le dictionnaire

— $\phi(w_n)$ le vecteur du mot w_n

— $pos(w_n) = \{n, v, a, r\}$ la partie du discours du terme w_n (nom, verbe, adjectif, ou adverbe)

— $poids(pos)$ le poids associé à une partie du discours

— $idf(w_n)$ la valeur IDF de w_n

$\phi(S)$ est ensuite normé afin d'avoir la même norme que les vecteurs de mots.

Quelques exemples de résultats

Nos vecteurs de sens peuvent être manipulés comme des vecteurs de mots.
Par exemple, on retrouve proches de **bank (institution financière)** les sens **account**, **deposit** et **money**, alors que l'on retrouve proche de **bank (rive, berge)** les sens **coast**, **sandbank** et **dip**.

Tous les vecteurs de sens sont disponibles à l'adresse suivante :

`https://github.com/getalp/WSD-TALN2017-Vialetal`

Évaluation

Système	SemEval 2007	SemEval 2015
S2C [4]	*75.80%**	
Lesk	68.70%	50.65%
Lesk étendu	78.01%	61.42%
VecLesk (Baroni C [5])	**75.29%**	58.02%
VecLesk (Baroni P [5])	73.52%	53.46%
VecLesk (Deps [6])	73.02%	56.40%
VecLesk (GloVe [7])	73.00%	**59.01%**
VecLesk (Word2Vec [8])	73.30%	57.00%

Table 1: Comparaison de nos résultats sur SemEval 2007 et SemEval 2015 pour chacun des modèles de vecteurs de mots par rapport aux étalons Lesk et Lesk étendu.
*Ce système est comparable au nôtre en termes de ressources utilisées, mais il comporte un biais : un paramètre δ est appris sur la tâche d'évaluation. Nous obtenons **76.50%** avec apprentissage de nos paramètres sur la tâche d'évaluation.

Modèle	Baroni C [5]		Baroni P [5]		Deps [6]		GloVe [7]		Word2Vec [8]	
Paramètres	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
SemEval 2007	0.6	0.6	0.5	0.5	0.6	0.8	0.5	0.6	0.5	0.6
SemEval 2015	0.5	0.8	0.5	0.6	0.6	0.8	0.5	0.7	0.5	0.6

Table 2: Estimation des paramètres δ_1 et δ_2 sur SemEval 2007 et SemEval 2015.

- La mesure de **Lesk** est **nettement améliorée** avec notre extension.
- Les résultats atteignent **presque** ceux du **Lesk étendu**,
sans la nécessité d'avoir un réseau lexical comme celui de **WordNet**.
- Notre système permet de **désambiguïser** des **langues peu dotées** avec de bons résultats.

Références

[1] Michael Lesk. Automatic sense disambiguation using mrd: how to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC '86*, pages 24–26, 1986.
[2] Satanjee Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing 2002*, Mexico City, February 2002.
[3] George A. Miller. Wordnet: A lexical database. *ACM*, Vol. 38(No. 11):p. 1–41, 1995.
[4] Xinxiong Chen *et al.* A unified model for word sense representation and disambiguation. In *EMNLP 2014*, pages 1025–1035.
[5] Marco Baroni *et al.* Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL 2014*, pages 238–247.
[6] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *ACL 2014*, pages 302–308.
[7] Jeffrey Pennington *et al.* Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
[8] Tomas Mikolov *et al.* Distributed representations of words and phrases and their compositionality. In *NIPS 2013*, pages 3111–3119.