Désambiguïsation lexicale à base de connaissances par sélection distributionnelle et traits sémantiques

Mardi 23 Juin 2015

Mokhtar Boumedyen BILLAMI
Doctorant, équipe TALEP, Laboratoire LIF (Marseille)
mokhtar.billami@lif.univ-mrs.fr

Introduction

- WSD est nécessaire pour plusieurs applications : recherche d'information, extraction d'information, traduction automatique, fouille de textes, simplification lexicale de textes, etc.

- Systèmes de désambiguïsation lexicale existants s'appuient sur deux étapes (Navigli, 2009) :
  1) Représentation de l'ensemble des sens d'un mot (ressources lexicales : dictionnaires ou réseaux sémantiques)
  2) Choix du sens le plus probable d'un mot par rapport à son contexte

La meilleure possibilité d'identifier le sens d'un mot est de se référer à son contexte (Ide, Véronis, 1998)

Ambiguïté lexicale et sa résolution

- Quel contexte ?
  - Définir la taille de la fenêtre textuelle, nature des informations retenues
    - ✓ *n* mots à gauche/droite du mot cible, une phrase, un paragraphe, le document entier, …
    - ✓ Sacs de mots, dépendances syntaxiques, parties du discours spécifiques, …

- Méthode de désambiguïsation la plus directe
  - Estimer la proximité sémantique de chaque sens candidat par rapport à chaque sens de chaque mot appartenant au contexte du mot à désambiguïser
  - Complexité exponentielle

Complexité exponentielle (approche exhaustive)

- Temps d'exécution : critère important pour une désambiguïsation plus rapide

- Si $|W_i|$ est le nombre de sens du mot $|W_i|$ alors $\prod_{i=1}^{N} |W_i|$ est le nombre de combinaisons possibles avec *N* nombre de mots dans le contexte

- Exemple :

✓ Une phrase de 10 mots avec 10 sens en moyenne ⇨ 10^{10} combinaisons possibles (10^{10} séquences de 10 sens possibles)

Plan de la présentation

- Approches de désambiguïsation lexicale
  - Méthodes dirigées par les données
  - Méthodes basées sur des connaissances, utilisation de ressources lexicales
- Méthodologie
  - Corpus de travail et de test
  - Réseau sémantique BabelNet
- Évaluation
  - Données des expériences menées
- Conclusion et perspectives

Approches de désambiguïsation lexicale (1/2)

- Quelle méthode de désambiguïsation utiliser ?

- Dirigées par les données (supervisées/non supervisées)

  - ✓ Méthodes supervisées : s'appuient sur un corpus d'apprentissage, utilisation de données annotées sémantiquement
    - Exemple: Corpus SemCor (Miller *et al.*, 1993) avec 352 textes annotés et près de 234 000 sens au total
    - Elles ont les meilleures performances et donnent les meilleurs résultats

  - ✓ Méthodes non supervisées : exploitation des résultats de méthodes d'acquisition automatique de sens

Approches de désambiguïsation lexicale (2/2)

- À base de connaissances : exploitation de ressources prédéfinies

  - ✓ Elles reposent sur des connaissances issues de ressources lexicales et sémantiques
  - ✓ Les ressources fournissent les descriptions sémantiques nécessaires pour faire la désambiguïsation
  - ✓ Mesures de similarité sémantique locales donnant une valeur de proximité entre deux sens de mots
  - ✓ Algorithmes globaux utilisant ces mesures pour trouver le sens selon le contexte (phrase, paragraphe ou document)

Méthodologie

- Désambiguïsation lexicale traitant l'hypothèse de « **One sens Per Discourse** »
- Clé :
  - Observation des voisins de chaque mot polysémique dans le texte
- Comparaison du sens candidat uniquement avec les sens des voisins sélectionnés au moyen d'une similarité distributionnelle (Lin,1998)

- Avantages d'utilisation des voisins les plus proches (*k*-PPV)
  - Avoir des indices sur le sens le plus probable d'un mot polysémique dans le texte
  - Réduire la complexité exponentielle (temps d'exécution)
  - Ne pas perdre une cohérence au niveau de la désambiguïsation de tous les mots du texte
  - Plus de chance de retourner le sens le plus spécifique au lieu du sens le plus général

Corpus de travail

- Ensemble de trois corpus de différents genres
  - ✓ AFP, EST REP et Wikipédia

- Au total : 6 631 442 phrases correspondant à 147 648 986 tokens
- Corpus analysés par la chaîne de traitement Macaon (Nasr *et al.*, 2011)
- Textes lemmatisés et annotés en partie de discours
- Un ensemble de triplets de dépendances syntaxiques est stocké et indexé
  - ❖ Un triplet de dépendance se compose d'une tête, d'un type de dépendance et d'un modificateur

- Exemple d'un triplet de dépendance :

  « leurs regards recouvraient les eaux du fleuve »

  - ✓ Triplet : (recouvrir, sujet, regard)

Corpus de test

- Deux corpus différents
  - ✓ 10 textes du corpus IREST, textes standards pour des tests de vitesses de lecture
  - ✓ 20 textes de lectures pour enfants en école primaire
- 6 235 occurrences de mots (4 139 occurrences de mots pleins)
- Une moyenne de 208 occurrences (138 occurrences de mots pleins) par texte
- Textes lemmatisés et annotés en partie de discours par Macaon

Réseau sémantique BabelNet
(Navigli, Ponzetto, 2012)

- Il décrit des sens et des entités nommées

- Il intègre différentes ressources avec WordNet : Wikipédia, Wiktionary, Wikidata, OmegaWiki et Open Multilingual WordNet (Collections de WordNets disponibles dans différentes langues)

- Il regroupe des mots en différentes langues par groupes de synonymes (Babel Synsets). Pour chaque Babel synset, plusieurs définitions en différentes langues sont disponibles

Langue Française	Mots	Synsets	Mots monosémiques	Mots polysémiques
	1 088 170	708 802	959 906	128 264

- **BabelNet 2.5.1 (proportion de 95,4% pour les noms)**

Similarité distributionnelle (1/3)

- Mesure indiquant le degré de cooccurrence entre un mot cible et son voisin apparaissant dans des contextes similaires

- Cas de « *souris* »
  - ✓ Voisins les plus proches possible pour le sens « genre d'animaux » : rongeur, muridæ, espèce, animal, queue, oreille …
  - ✓ Voisins les plus proches possibles pour le sens « dispositif de pointage en informatique » : informatique, pointage, ordinateur, périphérique, surface, interface, …

- Plus la similarité distributionnelle entre le mot à désambiguïser et ses voisins les plus proches est forte plus la probabilité d'avoir le sens le plus probable est grande

Similarité distributionnelle (2/3)

- Utilisation de cette mesure pour des mots partageant la même partie du discours; cas d'utilisation sur les noms et les verbes

- Pour chaque mot du corpus de travail, des relations grammaticales (de cooccurrences) existent : objet-de, sujet-de, a-pour-objet, etc.

« leurs regards recouvraient les eaux du fleuve »

- Triplets de dépendances syntaxiques

(regard, **det**, leur), (recouvrir, **suj**, regard), (recouvrir, **obj**, eau), (eau, **det**, le), (eau, **dep**, de) et (de, **obj**, fleuve)

Similarité distributionnelle (3/3)

- Triplet de dépendance syntaxique ⇨ trait syntaxique
  - (recouvrer, suj, regard) ⇨ regard a pour trait syntaxique suj(recouvrer)

- Similarité distributionnelle entre deux mots w_1 et w_2

$$sim(w_1, w_2) = \frac{2\ x\ I(F(w_1) \cap F(w_2))}{I\big(F(w_1)\big) + I\big(F(w_2)\big)}$$

  - $F(w)$: ensemble des traits syntaxique possédés par w
  - $I(S)$: quantité d'information contenue dans les traits de S
  - $I(S) = -\sum_{f \in S} \log P(f)$ avec $P(f)$: probabilité d'avoir le trait f

Algorithme de Lesk (1/2)

- Similarités à base de traits sémantiques

- Principe de base (Lesk, 1986)
  - ✓ Déterminer le nombre de mots en communs entre deux définitions de mots pour avoir le sens partagé
  - ✓ Comparaison de mots pleins (traits sémantiques)

- Baseline, variante de Lesk
  - ✓ Comparaison des mots pleins du contexte avec les traits sémantiques des définitions de chaque sens candidat
  - ✓ Prise en compte des synonymes dans le cas d'absence des définitions

$$Lesk_{Variante} = \left| contexte(w) \cap D(S_i(\mathrm{w})) \right|$$

Exemple

Texte

Dans une **petite** ville, un marchand de fruits possédait un magasin situé juste au-dessus d'une cave profonde. Chaque nuit les souris venaient en foule de cette cave dans le magasin. Elles mangeaient les pommes et les poires, les raisins et les noix, et n'épargnaient pas non plus les légumes et les pommes de terre. Entre minuit et le lever de soleil, aucune marchandise se trouvant dans le magasin n'était épargnée des **petits rongeurs**. La nuit, tant qu'il y avait du bruit dans les rues et des chariots qui roulaient, les souris restaient silencieuses dans la cave. Mais dès minuit, elles venaient en foule, s'amusaient avec les fruits doux et célébraient de vrais banquets. Leurs traces de passage désespéraient le magasin. Il essayait en vain de se protéger contre les souris.

Sens 1

« Dispositif de pointage en informatique »

(nombre de connexions sémantiques : 1102)

- Dispositif de pointage en informatique
- Une souris est un dispositif de pointage pour ordinateur
- Un périphérique d'entrée qui, lorsqu'il est déplacé sur une surface plane, fait bouger un curseur sur les axes X et Y d'une interface utilisateur graphique, et possède un ou plusieurs boutons pour sélectionner des éléments de l'interface graphique comme des boutons

Sens 2

« espèce de souris »

(nombre de connexions sémantiques : 839)

- Espèce de **souris**
- La **souris** grise est une espèce parmi les **petits rongeurs** de la famille des muridés

Algorithme de Lesk (2/2)

- Similarités à base de traits sémantique
- Algorithme de base de Lesk
  - ✓ Donner un score à une paire de sens de deux mots différents
  - ✓ Comparaison des traits sémantiques entre les définitions des sens

$$Sim_{Lesk}(S_1, S_2) = \left| D(S_1) \cap D(S_2) \right|$$

- Extension de Lesk (Lesk étendu)
  - ✓ Calcul du recouvrement entre les définitions de deux sens de mots
  - ✓ Ajouter du recouvrement entre les définitions provenant de différentes relations : *hypernyms, hyponyms, meronyms, holonyms, attribute, similar-to et also-see* (7 relations ⇨ 49 paires de relations possibles)

$$Relpairs = \{(R_1, R_2) \mid R_1, R_2 \in Relations; \\ si(R_1, R_2) \in Relpairs \; alors (R_2, R_1) \in Relpairs\}$$

$$Sim_{LeskEtendu}(S_1, S_2) = \sum_{\forall (R_1, R_2) \in Relations^2} \left| D(R_1(S_1)) \cap D(R_2(S_2)) \right|$$

Désambiguïsation lexicale par sélection distributionnelle et traits sémantiques

- Méthode structurelle fondée sur une distance sémantique entre un sens candidat et chaque sens de chaque voisin plus proche du mot à désambiguïser

$$\boldsymbol{S^*} = argmax_{S \in \text{Sens(w)}} \sum_{N_i \in N_w:\, N_i \neq \text{w}} \boldsymbol{max} \quad \text{Score}\left(\text{S}, \text{S}'\right)$$

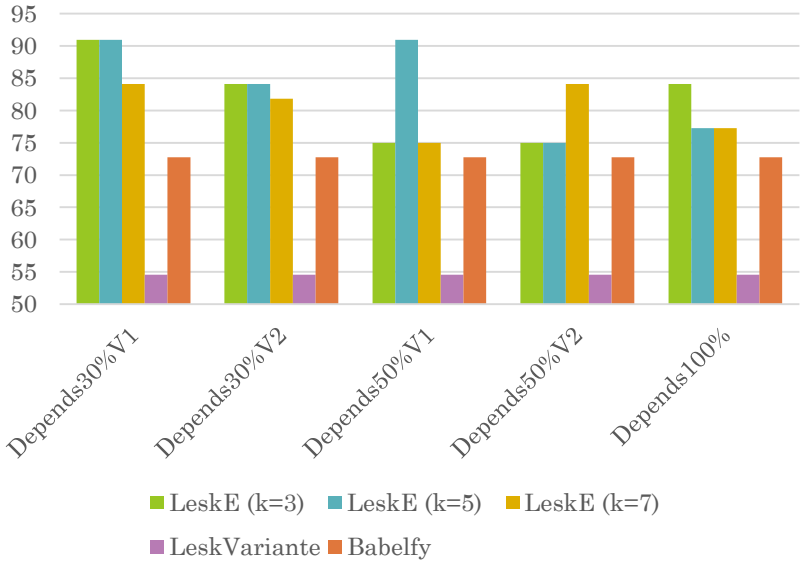
- $S' \in$ Sens (N_i) $avec$ $i = 1 \ldots k$
- $N_w = \{N_1, N_2, \ldots, N_k\}$: ensemble ordonné des k voisins les plus proches du mot cible w
- Sens (N_i) : ensemble des sens du voisin N_i
- Sens (w) : ensemble des sens du mot cible w
- Score $\left(\text{S}, \text{S}'\right)$: fonction pour mesurer la similarité entre deux sens S et S′

Données des expériences menées

Jeu de test	Fréquence	Nbre Sens	LeskBase	LeskÉtendu	LeskVariante	Babelfy
fleuve	3	**3**	100	**100**	0	100
fée	8	**3**	0	**100**	100	87,5
pêcheur	4	**4**	0	0	0	0
plante	15	**5**	86,67	**100**	80	100
castor	4	**9**	100	**100**	100	100
souris	10	**9**	30	**100**	0	30
planter	2	**3**	100	**100**	100	100
naître	7	**3**	0	85,71	85,71	100
obliger	2	**5**	50	**100**	50	100
taire	9	**5**	erreur POS	erreur POS	erreur POS	-
troubler	2	**7**	0	0	0	0
parler	6	**10**	16,67	**100**	16,67	50

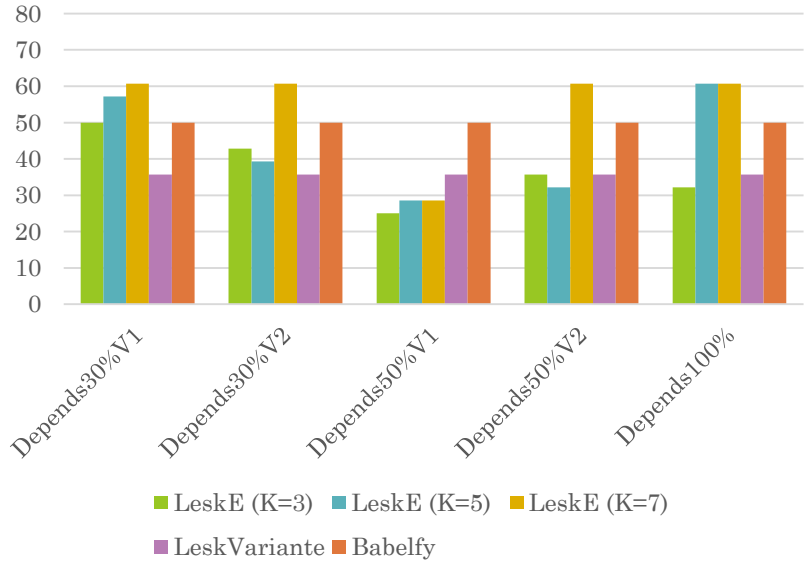
Taux d'exactitude (%) obtenus par méthode de Lesk de base et Lesk étendu par sélection aléatoire de 30% (V_1) sur l'ensemble des triplets pour les 5 plus proches voisins et comparaison avec la méthode de Lesk variante et Babelfy

Évaluation (1/2)



Écart-type = 3,76

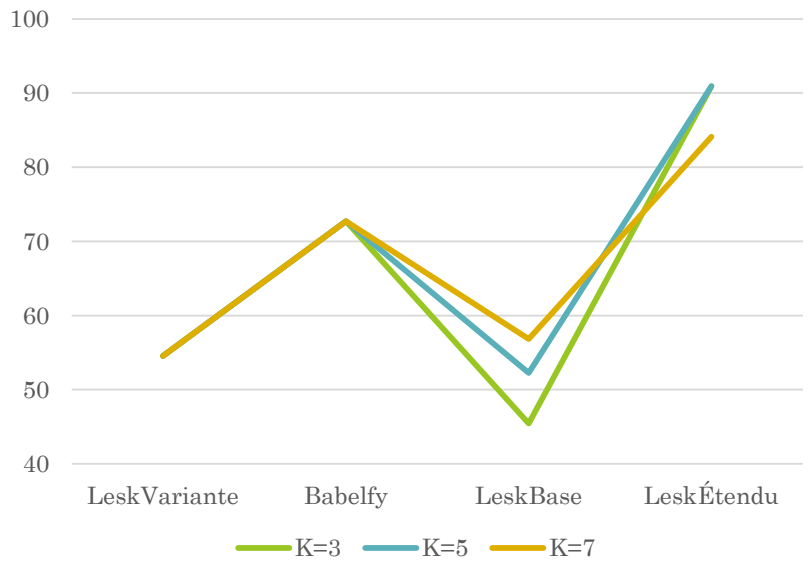
Taux d'exactitude sur les **noms** du jeu de test par utilisation de l'algorithme de Lesk étendu



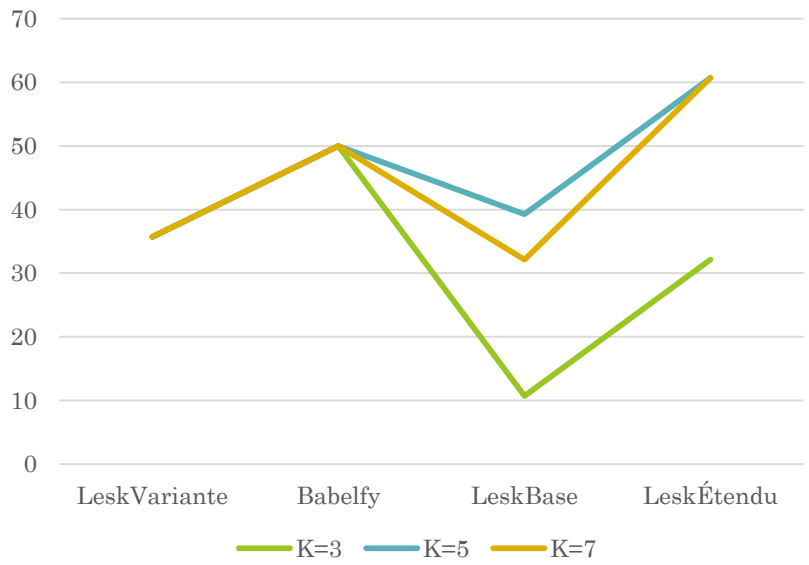
Écart-type = 9,80

Taux d'exactitude sur les **verbes** du jeu de test par utilisation de l'algorithme de Lesk étendu

Évaluation (2/2)



Taux d'exactitude sur les **noms** du jeu de test par sélection aléatoire de 30% sur les dépendances syntaxiques



Taux d'exactitude sur les **verbes** du jeu de test sur l'ensemble des dépendances syntaxiques

Conclusion

- Méthode de désambiguïsation lexicale traitant l'hypothèse du « **One sense Per Discourse** » en utilisant *k*-PPV par sélection distributionnelle et traits sémantiques selon le contexte du mot à désambiguïser.

- Utilisation de l'algorithme de désambiguïsation lexicale de Lesk, de sa variante et de son extension.

- Application de la méthode de désambiguïsation sur un domaine général (pas de domaine de spécialité) et utilisation du paragraphe comme contexte.

- *Taux d'exactitude* = 78% sur les données du jeu de test (meilleure combinaison est retournée par *k*=5 et 30%V1 comme sélection de triplets de dépendances syntaxiques).

Perspectives

- Comparaison des résultats avec l'utilisation des *k* mots pleins les plus proches linéairement de l'occurrence du mot à désambiguïser.

- Évaluation qualitative sur les différents sens d'un mot ambigu (vérification de l'existence d'une relation d'hyperonymie entre les sens candidats).

- Utilisation d'autres algorithmes locaux pour mesurer la similarité sémantique : algorithmes à bases de distance taxonomique (Wu, Palmer, 1994 ; Hirst, St-Onge, 1998) ou algorithms à base de contenu informationnel (Resnik, 1995 ; Seco *et al.*, 2004).

En vous remerciant pour votre attention

RÉCITAL : 17ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

Mes coordonnées

- **Mokhtar Boumedyen BILLAMI**

  Émail : mokhtar.billami@lif.univ-mrs.fr

  Tel : 04 91 82 96 37

  Fax : 04 91 82 92 75

  Aix-Marseille Université

  LIF, UMR 7279

  163 Avenue de Luminy

  Case 901

  13288 Marseille cedex 9, France

Références

- Fellbaum, C. Ed. (1998). WordNet: An Electronic Database. *MIT Press, Cambridge, MA.*

- Hirst, G., St-Onge, D. D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet : An electronic Lexical Database. C. Fellbaum. Ed. MIT Press,* 305–332.

- Ide, N., Véronis, J. (1998). Word sense disambiguation: The state of the art. *Computat. Ling. 24, 1,* 1–40.

- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation, SIGDOC '86, New York, NY, USA : ACM,* 24–26.

Références

- Lin, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning (ICML, Madison, WI),* 296–304.

- Miller, G. A., Leacock, C., Tengi, R., Bunker, R. T. (1993). A semantic concordance. *In Proceedings of the ARPA Workshop on Human Language Technology.* 303–308.

- Moro, A., Raganato, A., Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL), 2,* 231-244.

- Nasr, A., Béchet, F., Rey, J. F., Favre, B., Le Roux, J. (2011). MACAON: A linguistic tool suite for processing word lattices. *The 49th Annual Meeting of the Association for Computational Linguistics. ACTI,* 86-91.

Références

- Navigli, R., Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence, 193, Elsevier,* 217-250.

- Navigli, R. (2009). Word Sense Disambiguation : a Survey. *ACM Computing Surveys 41(2), ACM Press,* 1-69.

- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *In IJCAI'95, San Francisco, CA, USA,* 448–453.

- Seco, N., Veale, T., Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. *In Proceedings of ECAI'2004, Valencia, Spain,* 1089–1090.

Références

- Wu, Z., Palmer, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on ACL, volume 2 de ACL '94, Stroudsburg, PA, USA. Association for Computational Linguistics,* 133–138.