# Construction automatique d'une ressource de synonymes désambiguïsés pour l'aide à la lecture

Mokhtar Boumedyen BILLAMI

LIF-CNRS UMR 7279, Aix-Marseille Université

mokhtar.billami@lif.univ-mrs.fr

27 Juin 2016

# Sommaire

# BabelNet et Jeux De Mots (JDM)

## BabelNet

**Source** [Navigli et Ponzetto, 2012]

**Data** sens provenant de différentes sources (Wikipédia, Wiktionnaire, Wikidata, Omega wiki, Open Multilingual WordNet et WordNet [Fellbaum, 1998])

**Avantages** (1) très bonne couverture des mots polysémiques; (2) différenciation entre concepts et entités nommées

## JDM

**Source** [Lafourcade, 2007]

**Data** prise en compte de la relation de raffinement sémantique et de synonymie

**Avantage** représentation des sens d'un mot donné sous la forme d'un arbre [Lafourcade et Joubert, 2009]

# Difficultés de désambiguïsation lors de l'utilisation de BabelNet et JDM

## BabelNet

- Granularité de sens trop fine – *Souris* a 9 sens, parmi lesquels: *'espèce de petit rongeur'*, *'genre de rongeur'* et *'rongeur'*
- Existence de mots techniques et mots provenant de langues étrangères

## JDM

- Faible couverture de raffinements sémantiques et de liens entre sens et synonymes
- L'impossibilité d'utiliser des méthodes de désambiguïsation à base de traits sémantiques ([Billami et Gala, 2016, Billami, 2015])

# NASARI : a Novel Approach to a Semantically-Aware Representation of Items

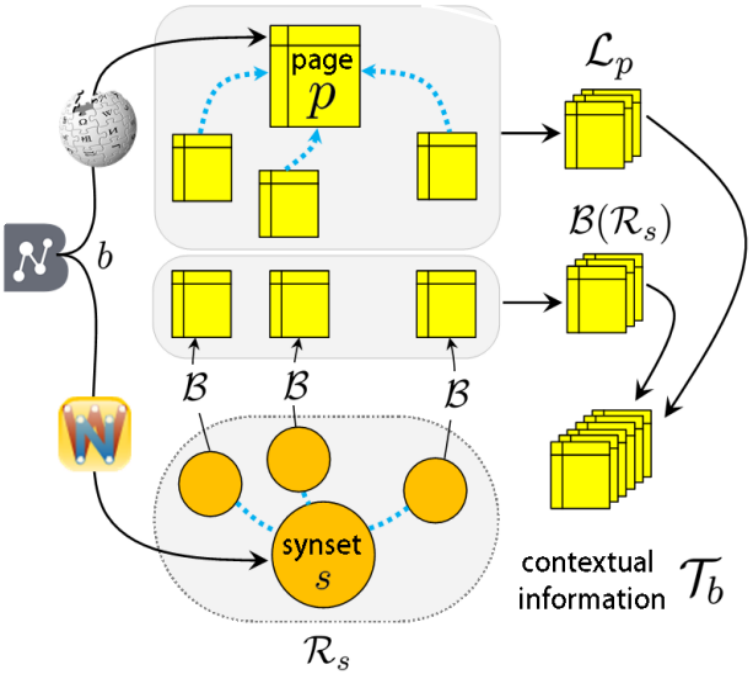


Figure 1: Processus d'obtention de l'information contextuelle provenant des sens de WordNet et des articles de Wikipédia [Camacho-Collados et al., 2015]

- $b$ sens $(p, s)$ décrit par un synset et/ou un article de Wikipédia
- $L_p$ page décrivant le sens $b$ et les pages ayant un lien sortant vers $b$
- $\mathbf{R_s}$ ensemble de sens de WordNet : le sens $b$ et tous les sens représentant un voisinage direct de $b$ y compris les sens désambiguïsés de la glose de $s$
- $\beta$ fonction de mapping entre un synset $s\prime$ de WordNet et sa page $p$
- $\beta(\mathbf{R_s})$ $\bigcup_{s\prime \in R_s} \beta(s\prime)$
- $\mathbf{T}_b$ information contextuelle (pages de Wikipédia) $\mathbf{T}_b = L_p \cup \beta(\mathbf{R_s})$

# Algorithme de filtrage

– Utilisation de BabelNet comme base de connaissances :

1. Tri décroissant des sens selon leur nombre de connexions sémantiques dans le réseau
2. Mesurer la similarité (forte *vs* faible) entre chaque paire de sens par utilisation de la fonction *Weighted Overlap (WO)* [Pilehvar et al., 2013]
   – $Seuil = 0.5$, sens le plus fort est gardé et le plus faible est supprimé

$$WO(Sens_1, Sens_2) = \frac{\sum_{q \in O}(r_q^1 + r_q^2)^{-1}}{\sum_{i=1}^{|O|}(2i)^{-1}} \quad (1)$$

– Un vocabulaire commun avec JDM y compris Lexique3 [New et al., 2007].

# L'exemple de « toile »

| Numéro | ID BabelNet | CS[1] | Vecteur de synonymes |
|---|---|---|---|
| 1 | bn:00019980n | 1665 | [tenue, tissu, toile, linge, matière, étoffe, chiffon, matériau, textile] |
| 2 | bn:00080772n | 1413 | [réseau, toile, web, internaute, toile d'araignée mondiale, world wide web, www] |
| 3 | bn:00073342n | 325 | [toile, toile d'araignée] |
| 4 | bn:00051362n | 323 | [toile, lin, toile de lin, lin textile] |
| 5 | bn:00015381n | 298 | [toile (peinture), toile] |
| 6 | bn:00062749n | 114 | [toile] |
| 7 | bn:00070989n | 111 | [toile] |
| 8 | bn:00015386n | 11 | [toile] |
| 9 | bn:00007787n | 7 | [bureau, milieu, fond, toile, origine, arrière-plan] |
| 10 | bn:04976972n | 6 | [toile (armure)] |
| 11 | bn:00051363n | 4 | [toile, lin] |
| 12 | bn:00015382n | 3 | [toile] |

Table 1: Sens de "toile" selon BabelNet + précisions sur le filtrage avec NASARI, JDM et Lexique3 (En préparation : [Gala et al., 2017])

[1] Nombre de connexions sémantiques (avec les voisins directs)

# ReSyf – ressource lexicale de synonymes désambiguïsés pour le français et triés selon leur niveau de difficulté


| POS | Nombre de lemmes | Proportion (%) |
|---|---|---|
| Noms | 15 494 | 79,88 % |
| Adjectifs | 1 741 | 8,97 % |
| Verbes | 1 516 | 7,82 % |
| Adverbes | 646 | 3,33 % |
| Total | 19 397 | 100 % |

Table 2: Distribution des entrées de ReSyf selon leur catégorie grammaticale

– *Nombre de sens* = 11 166

En vous remerciant pour votre attention

# Références


BILLAMI M. B. et Gala N. (2016)

Approches d'analyse distributionnelle pour améliorer la désambiguïsation sémantique.

*Actes des Journées internationales d'Analyse statistique des Données Textuelles (JADT), Nice*, pages 707 – 717.

BILLAMI M. B. (2015)

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

*Actes de la 22e conférence en Traitement Automatique des Langues Naturelles, session RECITAL, Caen, France*, pages 13 – 24.

Camacho-Collados J., Pilehvar M. T. et Navigli R. (2015)

NASARI: a Novel Approach to a Semantically-Aware Representation of Items.

*Proceedings of the 2015 NAACL:HLT Conference, Denver, Colorado*, pages 567 – 577. http://www.aclweb.org/anthology/N15-1059

Fellbaum C. (1998)

WordNet: an Electronic Lexical Database.

*MIT Press.*

# Références


Gala N., François T., Billami M. B. et Bernhard D. (2017)

ReSyf : a lexical resource with difficulty levels for reading assistance and automated text simplification (en préparation).

*Behavior Research Methods, Instruments, and Computers, Springer Journals, xx.*

Lafourcade M. et Joubert A. (2009)

Similitude entre les sens d'usage d'un terme dans un réseau lexical.

*Journal du Traitement Automatique des Langues,* **50***(1)*, pages 179 – 200.

Lafourcade M. (2007)

Making people play for Lexical Acquisition with the JeuxDeMots prototype.

*SNLP'07: 7th International Symposium on NLP, Pattaya, Chonburi, Thailande. http://hal-lirmm.ccsd.cnrs.fr/lirmm-00200883/file/MLF-snlp2007-v5.pdf.*

Navigli R. et Ponzetto S. (2012)

BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network.

*Artificial Intelligence, 193, Elsevier*, pages 217 – 250.

# Références


New B., Brysbaert M., Veronis J. et Pallier C. (2007)

The use of film subtitles to estimate word frequencies.

*Applied Psycholinguistics*, **28***(04)*, pages 661 – 677.

Pilehvar, M. T., Jurgens, D. et Navigli, R. (2013)

Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity.

*Proceedings of the 51st ACL Conference, Sofia, Bulgarie*, pages 1341 – 1351.