L’annotation sémantique

Mohamed CHABCHOUB

IFT6010

Plan

1) Annotation Sémantique
2) Dbpedia Spotlight
3) AIDA
4) Comparaison
5) Conclusion

1. Annotation sémantique

L'Annotation sémantique est une tâche de fouille de texte proche des méthodes de traitement automatique des langues qui consiste à étiqueter dans un document les mots avec des liens qui pointent vers une description sémantique.

Défis (désambiguisation) :

Un des plus grands défis de l'annotation est l'ambiguïté.
“They performed **Kashmir**, written by **Page** and **Plant**. **Page** played unusual chords on his **Gibson**.”

Chanson (Led Zepline) Région de l’Himalaya	guitariste Jimmy Page fondateur de google Larry Page	Modèle de guitar acteur Mel Gibson

Pour les humains, la désambiguïsation qui est le choix entre plusieurs candidats pour un mot ambigu est basée sur les connaissances du lecteur et le contexte actuel. Cependant la désambiguïsation est un problème difficile.

2) Dbpedia Spotlight :

C’est un système permettant de connecter les documents présents sur le web avec le Linked Open Data. Et cela en effectuant des annotations de textes en entrés par des URI pointant sur des ressources en DBpedia.

Figure 1 dbpedia Spotlight Demo

Dataset:

Construction d'un graphe de label, en commençant par les noms de surfaces qui sont les titres des pages de wikipedia, les reliants avec les Redirects qui sont les URIs indiquant des synonymes ou des noms de surfaces alternatives qui deviennent eux-mêmes des formes de surfaces. Les désambiguïsations fournissent les noms de surfaces ambigus qui peuvent être confus avec les ressources auxquelles elles sont liées, leurs labels deviennent eux aussi des surfaces forms.

Approche:

1-Repérage (Spotting)

L'ensemble des étiquettes a été utilisé comme lexique pour la phase de spotting. L'implémentation utilisée est le *LingPipe Exact Dictionary-Based Chunker* basée sur l'algorithme de correspondance de chaîne de caractères de Aho-Corasick.

Le système ignore les spots qui se composent seulement verbes, adjectives, adverbes ou prépositions. Le part of speech Tagger utilisé est l'implémentation de LingPipe basé sur le modèle de Hidden Markov**.**

2-Candidats

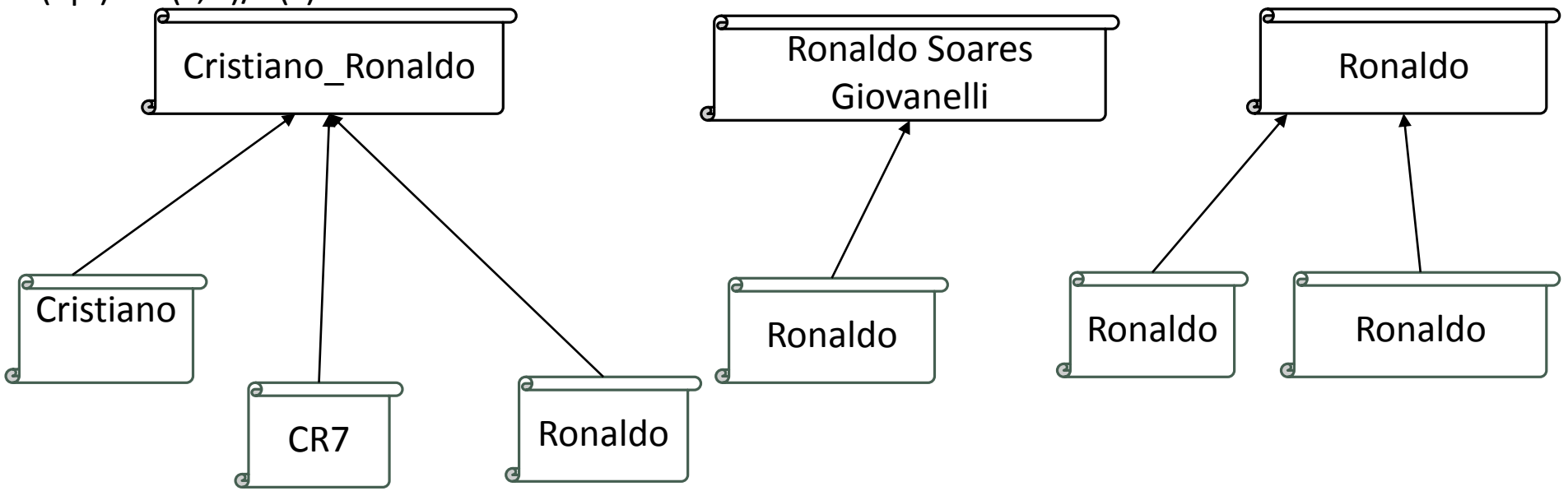
En utilisant la lexicalisation de dbpedia. Il relie chaque nom de surface à des ressources candidates.
(comme Washington à : la ville, la personne et l'état)

3-Désambiguïsation

-Métrique

- 1er métrique utilisé : probabilité à priori

P(r|s) = n(s, r)/n(s)



L'utilisation de la probabilité à priori serait comme une première phase pour ordonner les candidats.
Mais cette métrique seule ne suffit pas pour permettre la désambiguïsation, il faut prendre en considération le contexte.

❑ 2ème métrique tf :icf

-Le système utilise le paragraphe comme contexte pour les noms de surface

La modélisation des occurrences des ressources de dbpedia dans un vecteur (VSM) ou chaque ressource pointe à un espace multidimensionnel de mots

TF : Dans DBPedia Spotlight, il représente l'importance d'un mot pour une ressource donnée.

IDF : (Inverse Document Frequency) représente l'importance générale d'un mot dans une collection de ressource.

Cette métrique, bien qu'efficace dans le domaine de la recherche d'information, ne l'est pas pour la désambiguïsation parce qu'elle capture très bien l'importance global d'un mot (pour toutes les ressources) mais ne parvient pas à saisir l'importance d'un mot pour un ensemble spécifique de ressources candidats.

IDF ⟹ ICF

Soit Rs l'ensemble des ressources candidates pour un nom de surface s et n(wj) le nombre de ressources en Rs associé au mot wj.

$$ICF(w_j) = \log \frac{|R_s|}{n(w_j)} = \log |R_s| - \log n(w_j)$$

Avec la représentation VSM des ressources DBPedia et les poids associés TF*ICF, la désambiguïsation peut être vu comme un problème de classement ou l'on peut trouver la ressource la plus adéquate à la première position.
L'approche de DBPedia spotlight est de classer les ressources candidates à partir de leur score de similarité entre leur contexte et le contexte autour du nom de surface. Cosine Similarity
Le paramètre « Support » : le nombre de inlinks
Le paramètre « confidence » : si deux ressources pour une même forme de surface avec des scores élevés alors ils calculent la différence relative entre les deux. Pour P = 0,7 alors la différence relative entre les deux entités doit être inférieur à 1-P.

$$Mixed(r, s, C) = 1234.3989 * P(r|s) + 0.9968 * contextualScore(r, s, C) - 0.0275$$

Test

Test sur 155,000 wikilink

Disambiguation Approach	*Accuracy*
Baseline Random	17.77%
Baseline Default Sense	55.12%
Baseline TF*IDF	55.91%
DBpedia Spotlight TF*ICF	73.39%
DBpedia Spotlight Mixed	80.52%

Tableau 1 Précision des approches

Limites:

DBpedia Spotlight fait la désambiguïsation de chaque mot tout seul ce qui ne donne pas de résultats préférable si le texte en input est court .

They performed **Kashmir**, written by **Page** and Plant. Page played unusual chords on his **Gibson**.

3) AIDA

-Système pour annotation sémantique

-Des liens vers Wikipedia

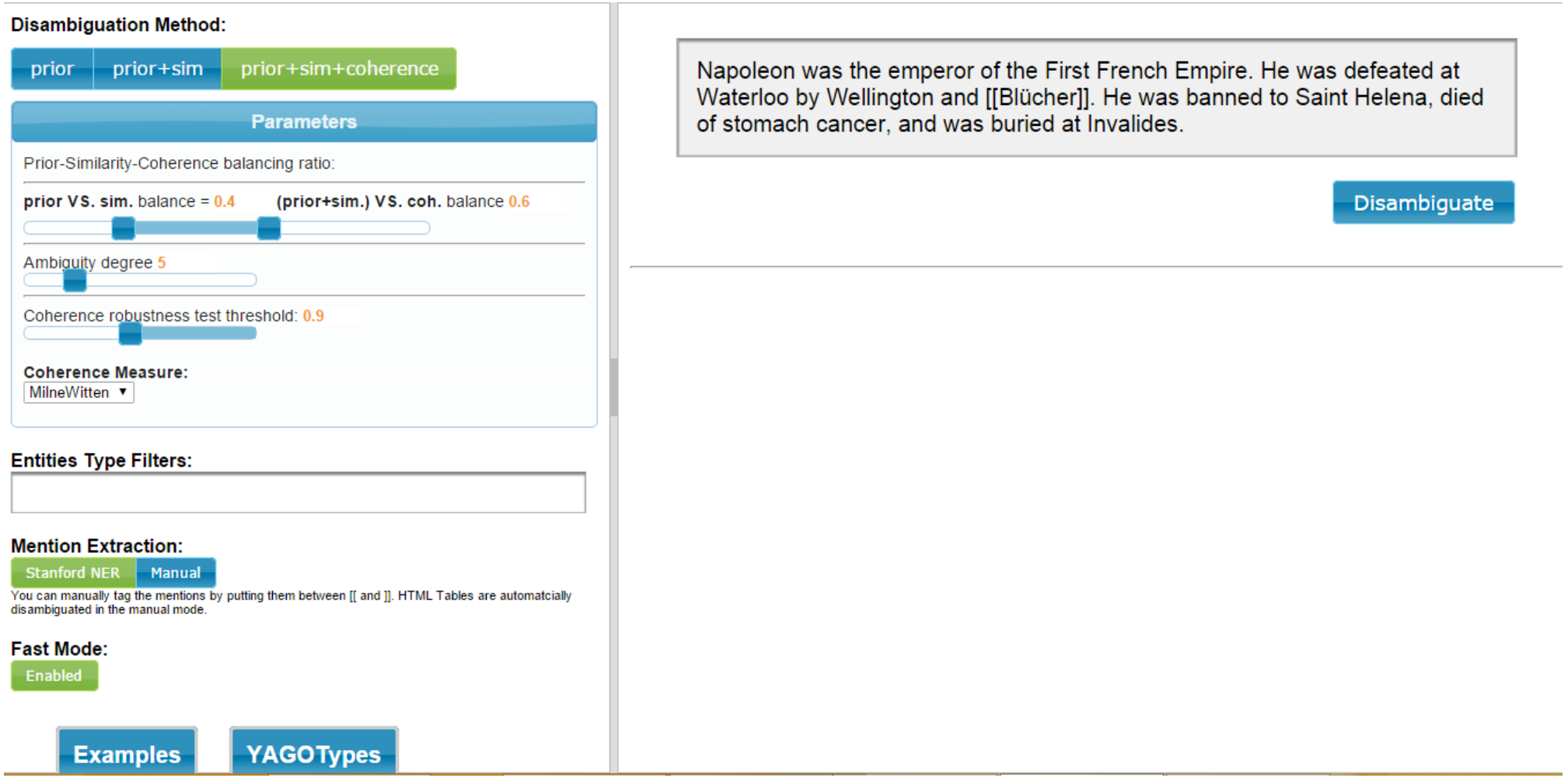


Figure 3 AIDA

Approche

1-Le spotting

Il utilise Stanford NER tagger pour identifier les formes de surfaces dans le textes en input

2-Candidates

La base de connaissances YAGO2

3-Désambiguïsation

-Métriques

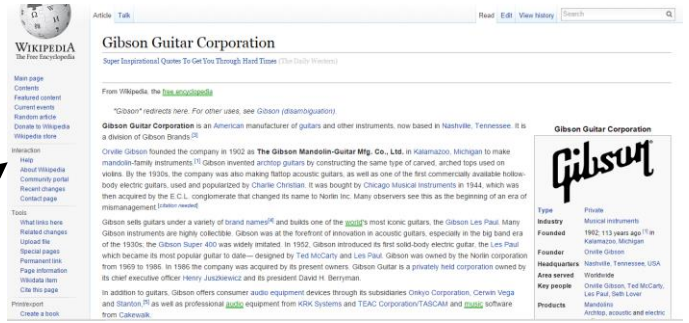
- La probabilité à priori (comme Dbpedia Spotlight)
- Similarité
  - Similarité basée sur la syntaxe:

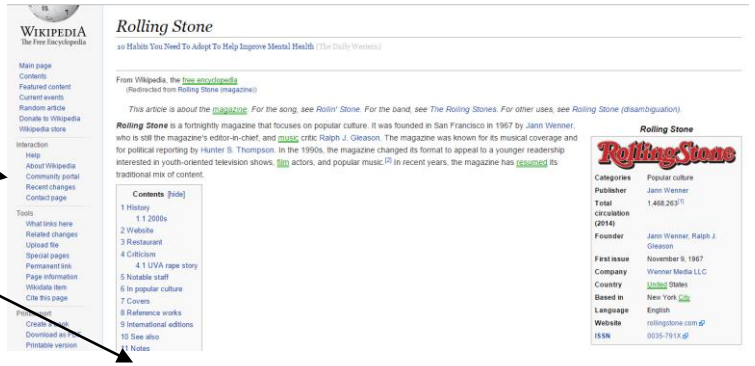
    Par exemple si une forme de surface S est le sujet d'un verbe B, en entrainant sur un large corpus de textes et extraire l'ensemble de mots qui peuvent être le sujet du verbe B. et après ordonner les entités candidats en se basant sur leur scores de compatibilités.

- Similarité
  - Similarité basée sur des mots clés :

-Du côté des noms de surface : prendre tous les mots en entrée en enlevant les Stopwords et la forme de surface elle-même (le contexte).

-Construction de mots-clés





❑ Cohérence

Le calcule se base sur le nombre de inlink que les articles des 2 entités se partagent

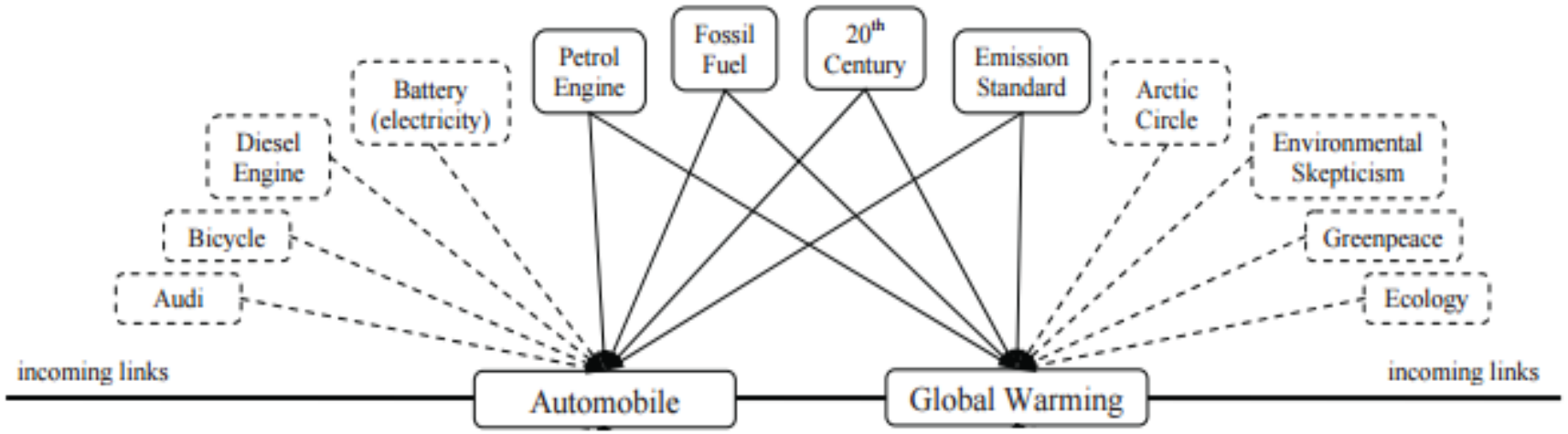
$$mw\_coh(e_1, e_2) = \quad 1 - \frac{\log\left(\max(|IN_{e_1}|, |IN_{e_2}|)\right) - \log(|IN_{e_1} \cap IN_{e_2}|)}{\log(|N|) - \log\left(\min(|IN_{e_1}|, |IN_{e_2}|)\right)}$$

$$\text{if} > 0 \text{ and else set to } 0.$$

IN : représente l'ensemble des inlinks pour chaque entité

N: le nombre totale dans la collection (Wikipedia)

(David Milne, University de Waikato, Nouvelle Zelande)



Formule Principale

L'objectif : choisir une seule entité pour chaque forme de surface repérée tel que

$$\alpha \cdot \sum_{i=1..k} \text{prior}(m_i, e_{j_i}) + \quad \beta \cdot \sum_{i=1..k} \text{sim}(\text{cxt}(m_i), \text{cxt}(e_{j_i})) + \quad \gamma \cdot \text{coh}(e_{j_1} \in \text{cnd}(m_1) \ldots e_{j_k} \in \text{cnd}(m_k)) = \max!$$

$$\alpha + \beta + \gamma = 1$$

$$\alpha = 0.43, \beta = 0.47, \gamma = 0.10$$

Modèles de graphe et algorithme

-Graphe

Construction d'un graphe pesé dans lequel les nœuds sont les formes de surfaces détectées ainsi que les ressources candidats.

Les arcs :

Mention-entité : une combinaison entre la probabilité à priori et la similitude :

Entité-entité : la cohérence

Le graphe est dense et le nombre d'arcs qui relient les formes de surfaces aux entités est très grand vu que dans une base de connaissance le nombre de résultats pour un mots donné est très grand.

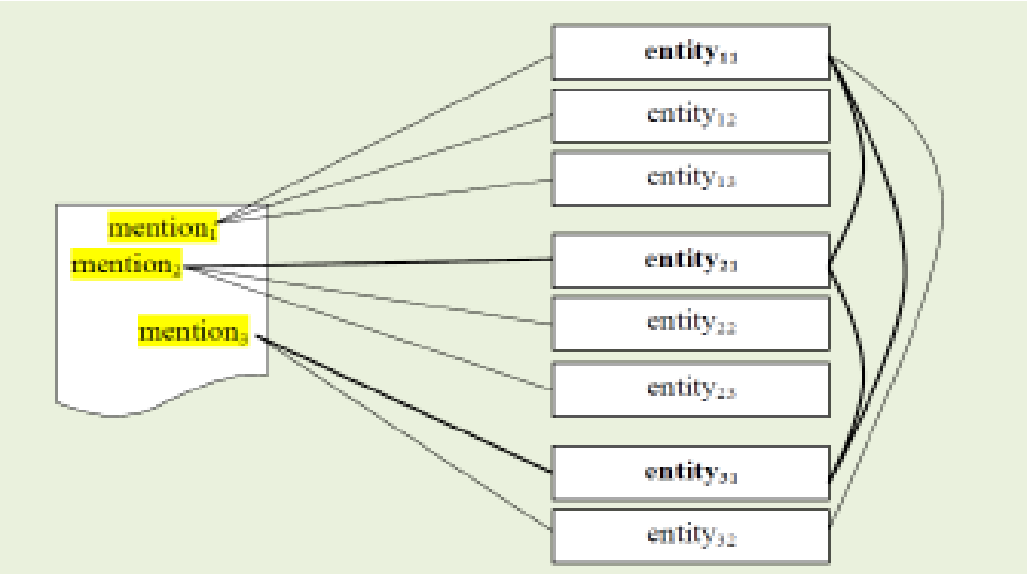


Figure 4 Graphe de désambiguïsation

A partir du graphe construit, l'objectif est de trouver un sous graphe dense qui relie exactement chaque forme de surface à un seul nœud ''Entité ''.

* Densité = Weighted degree

✓ Prétraitement :

Calcul des distances de chaque entité à l'ensemble des formes de surfaces et cela par le carré des distances des plus courts chemins. Ne garder que l'ensemble des entités qui sont les plus proches des mentions. Après des expériences effectuées ils ont trouvés que la taille de cet ensemble doit être 5 fois le nombre de mention.

✓ Traitement : Algorithme principale

✓ Post-traitement :

Le résultat peut amener à avoir des formes de surfaces qui sont reliées à plus qu'une entité. Mais le sous-graphe obtenu est assez petit pour appliquer un algorithme probabiliste proportionnel aux WD qui choisissent de manière aléatoire les entités. Cet algorithme est répétée N nombre de fois et le graphe obtenu avec un total de poids le plus grands est le résultat retourné.

Les premiers test:

- La définition d'un seuil S=0,9. Si la probabilité à priori est supérieur à S alors pour le calcul du poids de l'arc et la combinaison entre la similarité et la probabilité à priori sinon c'est seulement la similarité qui est utilisé.

- Le résultat est toujours entre 0 et 2 si il est supérieur à un seuil A alors les candidats sont inclus dans le graphe de désambiguïsation. Si ce n'est pas le cas alors l'application d'une combinaison entre la probabilité à priori et la similarité et seulement l'entité qui gagne est entré dans le graphe et tous les autres ne seront plus utilisé.

$$\sum_{i=1..k} |\text{prior}(m, e_i) - \text{simscore}(m, e_i)|$$

4) Comparaison

AIDA/CONLL : Construit par l'équipe de AIDA, ensemble de documents extraits d'actualités de Reuters Corpus V1.
(longueur moyenne 1039)
AQUAINT : sous ensemble du corpus AQUAINT, ensemble de textes de communiqués en anglais
(longueur moyenne 1415)
MSNBC : ensemble de communiqués extraits du réseau de nouvelles MSNBC
(longueur moyenne 3316)
IITB : textes extraits de pages WEB populaires (Sport, divertissement, science, technologie, santé)
(longueur moyenne 3879)
Mesure utilisé: F1-Score

	Pertinent	**Non pertinent**
Sélectionné	a	b
Non séléctionné	c	d

R = a / (a + c) P=a/(a+b)
F-1 Score = 2 (S[R]*S[P]) / (S[R]+S[P])

	Dbpedia Spotlight	AIDA
AIDA/CONLL	35,2	46,7
AQUAINT	27,6	21,2
MSNBC	33,5	47,4
IITB	48	7,7

Tableau 2 Comparaison AIDA Spotlight

IITB : textes longs (phrases ne sont pas fortement reliées)
Pour AIDA: faiblesse de rappel de Stanford NER tagger

Conclusion

-La désambiguïsation est le plus grand problème pour la tâche d'annotation sémantique,
-Le contexte est important
-Les problèmes que peut résoudre la désambiguïsation collective
-Se méfier de la désambiguïsation collective lorsque le texte parle de beaucoup de sujet qui sont faiblement connectés.

Références

- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen F¨urstenau,Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, Gerhard Weikum.(2011) '' Robust Disambiguation of Named Entities in Text '', Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 782–792.

- Pablo N. Mendes, Max Jakob, Andrés García-Silva, Christian Bizer(2011). '' DBpedia Spotlight: Shedding Light on the Web of Documents '', I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria.

- Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, Gerhard Weikum, ''AIDA-light:High-Throughput Named-Entity Disambiguation'', LDOW2014, April 8, 2014, Seoul, Korea.
- Marco Cornolti, Paolo Ferragina, Massimiliano Ciaramita, '' A Framework for Benchmarking Entity-Annotation Systems '', IW3C2 May 13–17, 2013, Rio de Janeiro, Brazil.