Approche cognitive pour la désambiguïsation d'entités nommées dans les articles d'actualité

Nivo RANDRIAMBOLOLONA , Yvon ANDRIANAHARISON, Gerhard WEIKUM
RANDIMBINDRAINIBE Falimanana
Max-Planck-Institute for Informatics Germany(MPII)
Laboratoire de recherche en sciences cognitives et applications (LRSCA)

nivoran@gmail.comyvonkyo@gmail.com weikum@mpi-inf.mpg.defalimanana@mail.ru
yvonkyo@gmail.com falimanana@mail.ru

Résumé : -Cet article propose une approche cognitive basée sur l'utilisation d'une grammaire formelle pour la désambiguïsation d'entités nommées dans les articles d'actualité. Utilisée conjointement avec un modèle de régression logistique, son objectif principal est d'éviter l'exclusion d'entités émergentes qui sont difficilement reconnaissables par voie empirique sans pour autant désavantager les entités préétablies. Ce projet est né dans le contexte du projet originel AIDA[1] qui a été réalisé au laboratoire du MPII en Allemagne et pour cette raison on lui a attribué l'appellation AIDA-for-News.

*Mots clés :*approche cognitive -grammaire formelle –désambiguïsation d'entités nommées - régression logistique– AIDA-for-News

Abstract : - The main purpose of this paper is to present a cognitive approach based on the usage of a formal grammar to disambiguate named entities in news articles. Jointly used with a logistic model tree, its main goal is about avoiding the exclusion of emerging entities which are hardly recognizable by existing empiric methods, without disadvantaging the popular ones. This project was motivated by the original MPII-Germany project AIDA[1] and for this reason, we named it AIDA-for-News.

*Keywords :*cognitve approach-formal grammar - named entity disambiguation– logistic model tree– AIDA-for-News

1 Introduction

La désambiguïsation d'entités nommées est indissociable de tout traitement automatique du langage naturel contenant des noms propres [2]. Ces derniers peuvent effectivement désigner, de manière ambiguë, différentes entités à l'intérieur d'une base de connaissances. Plusieurs projets de recherche se sont attelés à cette problématique, reposant majoritairement sur des approches empiriques. Le projet AIDA qui en fait partie, se base d'une part sur la méthode CRF [3] et d'autre part sur d'autres mesures empiriques. Ces techniques lui permettent de discerner les entités mentionnées dans divers textes avec grande précision (95%), lesquels textes contiennent principalement des noms de célébrités. Cette précision baisse toutefois lorsqu'il est confronté à d'autres types de textes contenant des noms dont la notoriété est récente. Le but du présent travail est d'inclure les noms qui ne sont pas détectables par les méthodes empiriques dans le processus de désambiguïsation. Partant de l'idée qu'une entité émergente qui n'est pas détectable par voie statistique devrait être appréhendée par voie cognitive, une grande partie du travail est consacrée à la construction d'une grammaire formelle destinée à générer le langage des contextes sémantiques des entités nommées. La classification automatique des contextes se fait ensuite via une variante de l'arbre décisionnel. Une base de connaissances locale permet l'intégration des entités émergentes. Les articles d'actualité, de par leur qualité grammaticale d'une part et du fait de la présence massive d'entités émergentes d'autre part, représentent un champ d'expérimentation

parexcellence.

2 Problématique et approche générale

2.1 Problématique

Par définition, la désambiguïsation consiste à déterminer pour une mention donnée (m_i) parmi toutes les entités candidates (e_{ij}) celle qui est la plus apte à être l'entité canonique désignée par m_i[4]. La fonction de désambiguïsation (collective) de AIDA combine trois mesures empiriques dont la probabilité à priori, la similitude contextuelle et la cohérence entre les entités candidates. Pour chaque mention m_i, il s'agit de déterminer l'entité candidate e_{ij} ,qui maximise la valeur de la fonction pondérée de désambiguïsation [4]:

$\alpha . \sum_1^k prior\ (m_i, e_{ji}) + \beta . \sum_1^k sim\ (cxt(m_i), cxt(e_{ji})) + \gamma \cdot coh(e_{j1} \in cnd(m_1) \ldots e_{jk} \in cnd(m_k)) = max!$ (1)

où :

- α, β, γ poids avec $(\alpha + \beta + \gamma)=1$
- $cnd(m_i)$ désigne l'ensemble des entités candidates de m_i,
- cxt() dénote le contexte des mentions et des entités,
- coh() est la fonction de cohérence pour un ensemble d'entités donné.

La probabilité à priori *prior*(m_i , e_{ji}) est calculée en fonction de la fréquence avec laquelle la mention m_i est utilisée dans un lien hypertexte pour se référer à l'entité e_{ij}, par rapport au nombre total de fois où elle est utilisée dans un lien hypertexte en général. La similitude de contexte sim (($cxt(m_i)$, $cxt(e_{ji})$)) tient compte du contexte aussi bien de la mention que de l'entité candidate et se base entre autres sur la divergence de *Kullback-Leibler* pour mesurer la similarité de contexte entre les deux. La fonction de cohérence enfin utilise un graphe de cohérence pour déterminer la constellation générale et collective la plus probable des entités candidates.

Aussi bien la probabilité à priori que la similitude de contexte présuppose que les entités considérées existent et qu'elles aient un poids statistique quelconque. Cela exclut d'emblée toute entité émergente du processus de désambiguïsation. Par ailleurs, toute entité émergente ambiguë se ferait, à tort, systématiquement écraser par ses homonymes qui bénéficient d'une grande popularité.

2.2 Approche centrée sur le contexte

L'approche proposée pour éviter l'exclusion des entités émergentes est censée rattraper le déficit en données empiriques par une exploitation maximale et par voie cognitive de tout ce qui représente un contexte sémantique des entités nommées à l'intérieur d'un article d'actualité. Cela suppose :

a) L'élaboration d'un dispositif cognitif, en l'occurrence une grammaire formelle, permettant de détecter et d'appréhender systématiquement tous les contextes de manière à ce que toute entité nommée, pourvu qu'elle soit dotée d'un contexte, puisse être identifiée, quel que soit son degré de popularité.
b) L'attribution automatique d'une classe sémantique à une structure contextuelle donnée ainsi que l'interprétation automatique de ce contexte en connaissances.
c) La possibilité d'intégrer ces entités avec leurs contextes dans une base de connaissances (ontologie) afin qu'une désambiguïsation en bonne et due forme soit possible.

2.3 Méthodologie

Pour la construction de la grammaire formelle, les étapes suivantes sont nécessaires :

- Repérage des principales catégories de description d'entités nommées.
- Analyse morpho syntaxique des différentes catégories de description.
- Regroupement des schémas similaires et définition d'un ensemble représentatif des différentes catégories de schémas sémantiques.
- Création de différents types de canevas où chaque canevas est générateur systématique et exhaustif des schémas d'une catégorie, fonctionnant comme un arbre décisionnel avec ses propres règles de dérivation.
- Intégration de l'ensemble des canevas dans une grammaire formelle capable de

produire le langage des schémas sémantiques des entités nommées.

Les schémas dérivés et reconnaissables par la grammaire auraient encore besoin qu'on leur attribue une classe sémantique pour pouvoir les interpréter correctement. Cette classification se fera par apprentissage supervisé moyennant une variante de l'arbre décisionnel et avec pour ensemble d'apprentissage le langage des schémas qui seront augmentés chacun d'un attribut classe. L'avantage de cette approche est de permettre également la reconnaissance d'éventuels schémas irréguliers qui ne font pas partie du langage des schémas.

La connaissance de la classe sémantique d'un schéma nous permettra ensuite de l'interpréter et d'extraire toutes les informations explicatives qu'il contient. Dans un premier temps, par souci de rigueur, ces informations seront encapsulées dans un modèle orienté-objet sur lequel sera calqué plus tard le concept d'une ontologie. L'objectif final est de pérenniser les entités et leurs schémas contextuels. Il ne s'agit pas de les pérenniser en tant que données mais en tant que connaissances réutilisables. Dans la même foulée, l'intégration et la désambiguïsation – au moins locale - des entités émergentes deviendront ainsi possible.

3Résultats

Le repérage des catégories et l'analyse morphosyntaxique des catégories de description se sont faits via le décryptage manuel d'une centaine de descriptions extraites de vrais articles d'actualité. Nous allons intégrer les résultats obtenus directement dans la définition des canevas qui en découlent.

3.1 Les canevas de description

Un canevas est un vecteur de symboles avec une structure arborescente sous-jacente. Chaque position peut accueillir soit un symbole terminal soit un symbole non terminal qui est la racine d'une expression variable récursive. En tout, 6 catégories principales de description ont été répertoriées pour tous les types. Une description est dite préfixe lorsque le contexte précède la mention de l'entité et suffixe dans le cas contraire. Ces 6 catégories sont sources de 22 classes sémantiques qui sont représentées par des canevas. C'est pour le type *personne* que les descriptions sont les plus complexes et les plus variées. A titre d'illustration, nous allons considérer le canevas <det> X_{MOL} X_{KA} F P qui est à la base de descriptions préfixes de personne avec précision de son occupation (ou fonction). Dans ce canevas, il y a 3 symboles terminaux (<det>, F et P) et deux non terminaux (X_{MOL}et X_{KA}). Chaque symbole non terminal engendre des nœuds fils qui sont des options. Lorsque Ɛ fait partie des fils, alors le père est facultatif. Figure 1représentece canevas sous sa forme déployée avec ses sous arborescences et illustre en même temps comment le schéma <det>LFP est engendré par ce canevas. X_{MOL} peut générer au choix : le symbole vide(ε), une nationalité (M), une organisation (O), un type d'organisation (O') ou un lieu (L). Le même principe vaut pour X_{KA}. L'expression *the US president Barack Obama* est un exemple d'expression répondant au schéma <dét>LFP.

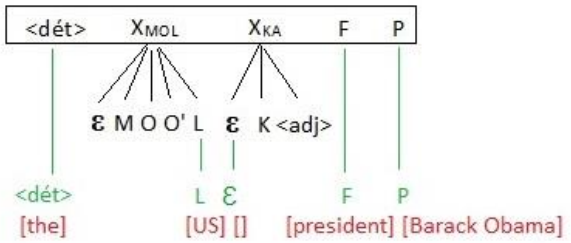

Figure 1 *:* Exemple de déploiementd'un canevas (Source : auteur)

3.2 Définition de la grammaire formelle

La grammaire G_{news} = (Σ_{news}, $\aleph_{news}$, R_{news}, S_{news}) est définie par l'alphabet des terminaux Σ_{news},,l'alphabet des non terminaux $\aleph_{news}$, l'ensemble des règles de productionR_{news} et enfin l'axiome S_{news}.

L'alphabet Σ_{news}des terminaux est composé de 3 sous-ensembles, $\Sigma_{new}=\Sigma_T \cup \Sigma_{\mp} \cup \Sigma_{\forall}$:

- Σ_T = {P, L, O}, contenant les symboles des 3 types d'entité*: person, location organization*

- Σ_T = {Ɛ , o, O', L', K, T, F, M, R}, contenant des méta-types et autres types supplémentaires : o (acronyme), O' (type d'organisation), L' (type de lieu), K (mot-clé), T (Titre), F (type de fonction), M (nationalité), R(lien familial). Il contient aussi le symbole vide Ɛ, qui est utilisé pour marquer l'omission d'un symbole facultatif.
- Σ_Y = {<dét>, <prép>, <adj>, <verb>, <ponct>, <verb>, <*mots_réservés*>},contenant principalement les fonctions grammaticales des mots.

L'alphabet N_{news} des symboles non-terminaux contient tous les symboles *germes* qui sont utilisés dans la partie variable d'un canevas ou dans les expressions variables dérivées.

Les règles de production appartenant à l'ensemble R_{news} sont définies par une fonction récursive non déterministe de dérivation r:

$r : (\Sigma \cup N)^* N (\Sigma \cup N)^* \rightarrow (\Sigma \cup N)^*$

$r(c_t) = cat(r(c_{it}))$, avec i allant de 0 à n_t.

C'est-à-dire que l'expression dérivée d'un canevas c_t via r est la concaténation des expressions dérivées des symboles individuels c_{it} qui composent c_t. Les conditions suivantes sont valables :

- Si c_{it} appartient à Σ_{new}, alors $r(c_{it}) = c_{it}$
- Sinon si c_{it} appartient à N_{news},,alors $r(c_{it})$ est défini par un tableau de dérivation.

Le langage des schémas généré par G_{news}est L_{news}. Il fait partie de la catégorie des *langages récursivement énumérables*, acceptable par une machine de Turing [5]. Cette grammaire nous a permis de dériver et de sélectionner 258 différents schémas pour décrire une personne, 45 pour décrire une organisation et 14 pour décrire un lieu. La prochaine étape est la définition d'un modèle d'apprentissage qui permettra de prédire la classe sémantique d'un schéma quelconque.

3.3 Apprentissage supervisé

3.3.1 Choix de l'ensemble d'apprentissage

Tous les schémas repérés dans un article d'actualité seront classifiés par apprentissage automatique. Certains de ces schémas sont des éléments du langage $L_{news,}$ tandis que d'autres sont des schémas en marge de ce langage, mais acceptés par la linguistique journalistique. La raison de cette marginalité n'est que d'origine structurelle. Les symboles qui les composent sont puisés dans le même alphabet que ceux des schémas du langage$L_{news.}$ Pour cette raison, nous estimons que L_{news}représente encore l'ensemble d'apprentissage le plus proche et le plus propice qu'on puisse mettre à la base de notre processus d'apprentissage. Pour rendre un apprentissage supervisé possible, il suffit d'attribuer manuellement sa classe à chaque schéma.

3.3.2 *Logistic model tree* (LMT)

Il s'agit d'un modèle de classification qui combine le modèle de régression logistique (type arbre modèle)[5]avec l'apprentissage par arbre décisionnel. La particularité est que chaque nœud est remplacé par un plan de régression au lieu d'une valeur constante.

L'objectif de la classification des schémas est d'associer chaque schéma à l'une des 12 classes de description. Testé sur le langage des schémas sémantiques avec le logiciel Weka, le LMT a livré de très bonnes performances de classification. Figure 2 nous montre les résultats du LMT et Figure 3 ceux d'un arbre décisionnel classique (C 4.5) [5].

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.27 seconds

=== Summary ===

Correctly Classified Instances         283               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0021
Root mean squared error                  0.0166
Relative absolute error                  2.7666 %
Root relative squared error              8.537  %
Total Number of Instances              283
```

Figure 2: Résultat de la classification avec LMT (Source : auteur)

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.23 seconds

=== Summary ===

Correctly Classified Instances         279               98.5866 %
Incorrectly Classified Instances         4                1.4134 %
Kappa statistic                          0.9831
Mean absolute error                      0.0021
Root mean squared error                  0.0322
Relative absolute error                  2.6981 %
Root relative squared error             16.5034 %
Total Number of Instances              283
```

Figure 3 : Résultat de la classification avec C4.5(Source : auteur)

3.4 L'ontologie des articles d'actualité

Notre ontologie dont l'utilité repose sur la nécessité de représenter les articles d'actualités et leurs entités nommées sous forme de connaissances est représentée sur la figure 4. Les relations entre les différents concepts qui la composent sont affichées dans le tableau 1.
L'article d'actualités qui y joue un rôle central est tout d'abord caractérisé par :

- ses métadonnées
- les entités nommées qu'il mentionne.

Une entité nommée sera à son tour définie par :

- les attributs qui la caractérisent (voir le modèle objet),
- les relations qui existent entre elle et les autres entités nommées de l'article,
- l'article qui la mentionne.

Des règles très simples ont été définies sur cette ontologie pour en faire une petite base de connaissances locale exploitable pour la désambiguïsation. Voici un exemple de règle :

P1 et P2 des individus de la classe P
SI P1 <estMariDe> P2
ALORS P2 <estFemmeDe> P1

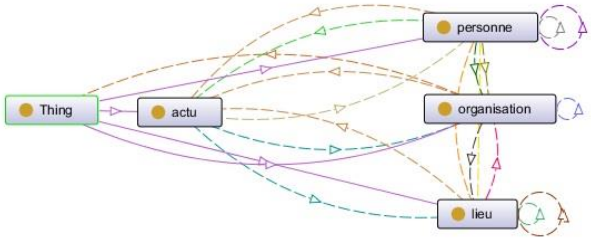

Figure 4 : Ontologie des articles d'actualité (Source : auteur)

	Actu	Personne	Organisation	Lieu
Actu		mentionne	mentionne	mentionne
Personne	estMentionnéeDans estAuteurDe	estEpouxDe	travailleChez	vitA, estNéA
Organisation	estMentionnéeDans	emploie	estFilialeDe	estSiseA
Lieu	estMentionnéDans		abrite	estSousLieu

Tableau 1 : Les relations entre les concepts (Source : auteur)

3.5 La désambiguïsation

Au lieu d'une seule fonction de désambiguïsation, il existe une fonction pour chaque type d'entité. Chaque fonction prend un objet complet comme argument et non une mention, comme c'est le cas avec AIDA.

3.5.1 Désambiguïsation collective des noms de lieux

Soient :

- $Réf_{GN}^{L}$ = répertoire des références de lieux dans le SIG GeoNames
- $Réf_{YG}^{L}$ = répertoire des références de lieux dans la base de connaissances YAGO
- $Réf_{ext}^{L} = Réf_{GN}^{L} \cup Réf_{YG}^{L} \cup \{\varepsilon\}$
- $Réf_{news}^{L}$= répertoire des références de lieux dans la base de connaissances locale $\cup \{\varepsilon\}$
- $\mathscr{L}$ = ensemble des objets de la classe LIEU pour un article donné

La fonction de désambiguïsation des lieux est définie comme suit:

$\mathscr{D}_L : \mathscr{L} \rightarrow Réf_{ext}^{L} \times Réf_{news}^{L}$

$\mathscr{D}_L(l) = (\alpha, \beta)$, avec :

- $\alpha \in Réf_{GN}^{L}$, si le nom de l est connu de GeoNames et qu'il existe un autre lieu $l' \in \mathscr{L}$ qui partage la même hiérarchie géographique que l[6]
- $\alpha \in Réf_{YG}^{L}$, si le nom de l n'est pas connu de GeoNames mais connu de YAGO.

Le contexte géographique $\mathscr{L}$ de l'article dans sa globalité est pris en compte. Si deux ou plusieurs noms de lieux partagent la même hiérarchie géographique, alors cette hiérarchie est certainement la bonne référence. Les lieux qui n'ont pas été reconnus par GeoNames sont recherchés dans YAGO moyennant, le cas échéant, un mot-clé. En cas d'échec répété, ils sont recherchés dansWikipedia. Si le nom n'existe pas non plus dans Wikipedia, on considère que le nom a été mal classifié et qu'il s'agit d'une erreur d'étiquetage. Les lieux doivent ainsi obtenir deux types d'identifiants après leur désambiguïsation :une référence GeoNames et une référence locale, ou alors une référence YAGO et une référence locale. Sinon c'est la paire (ε,ε).

3.5.2 Désambiguïsation de noms de personne et de noms d'organisation

Soient :

- $Réf_{YG}^{P}$ = répertoire des références de personnes dans la base de connaissances YAGO
- $Réf_{news}^{P}$= répertoire des références de personnes dans la base de connaissances locale U {ε}
- $\mathscr{P}$ = ensemble des objets de la classe PERSONNE pour un article donné

La fonction de désambiguïsation des noms de personne est définie comme suit :

$\mathscr{D}_P: \mathscr{P} \rightarrow Réf_{YG}^{P} \times Réf_{news}^{P}$

$\mathscr{D}_P(\wp) = (\alpha, \beta)$

- S'il existe une entité canonique dans YAGO dont les attributs correspondent à ceux de $\wp$, alors $\alpha \neq \varepsilon$

  Sinon $\alpha = \varepsilon$
- S'il existe une entité $\wp'$ dans la base de connaissances locale avec $\wp = \wp'$ et $\mathscr{D}_P(\wp)=(\alpha',\beta')$, alors $\beta = \beta'$. Dans ce cas, si $\alpha' \neq \varepsilon$, alors α et α' doivent être égaux.

  Sinon β obtient une nouvelle valeur.

Cette fonction vérifie dans YAGO s'il existe une entité dont les valeurs d'attributs correspondent complètement ou partiellement à celles des attributs de l'objet *personne* local. Si oui, l'identifiant YAGO est retenu comme référence externe, sinon il n'y aura pas de référence externe. La référence interne sera créée indépendamment de cela au moment de la matérialisation de l'objet dans l'ontologie.

La désambiguïsation de noms d'organisations se passe exactement de la même manière que pour les personnes.

3.6 Implémentation

Les tâches qui composent *AIDA for news* sont réparties dans 3 modules : le module de préparation des données, le module cognitif et le module de désambiguïsation. Figure 5 donne un schéma exprimant cette architecture avec les API utilisées à chaque niveau de traitement et figure 6 met en évidence les structures de données correspondantes aux tâches. Par rapport à l'AIDA classique, la phase de préparation de données complète l'étiquetage par Stanford-NER par une annotation fine de tous les mots qui ne représentent ni personne, ni organisation ni lieu.

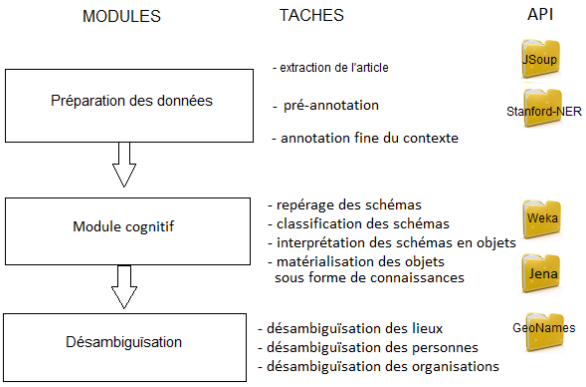

Figure 5: Architecture modulaire du système de désambiguïsation(Source : auteur)

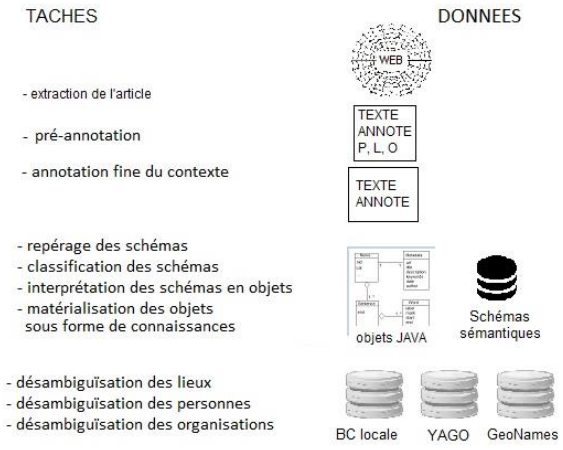

Figure 6: Correspondances tâches-structures de données(Source : auteur)

3.7 Evaluation

3.7.1 Evaluation interne

Une centaine d'articlesd'actualité de BBC-News, appartenant à 10 différentes rubriques ont été selectionnés pour l'évalation interne, c'est à dire 10 groupes de 10 articles.

Les rubriques suivantes ont été sélectionnées:

- *business* (10 articles)
- *entertainement and arts* (10 articles)
- *health* (10 articles)
- *science and environment (*10 articles)
- **world (us-canada, africa, asia, Europe, england)* (50 articles)
- *sport* (10 articles)

Les résultats des tests sur les différents groupes sont affichés dans le Tableau 2. La troisième, la quatrième et la cinquième colonne donnent respectivement les taux de réussite de la désambiguïsation pour les types *personne*, *lieu*, *organisation*.Le taux de réussite par type Xestcalculéselon(2):

$$TR(X) = \frac{N\_correct(X)}{N(X) - N\_ignoré(X)} \quad (2)$$

Le taux de réussitegénéral pour unecatégoriedonnéeestcalculéselon(3):

$$TR(cat) = \frac{\sum TR(X)}{3} \quad (3)$$

Catégories	*N*	*TR (P)*	*TR(L)*	*TR(O)*	*TR(cat)*
Business	10	100,00%	100,00%	100,00%	100,00%
World-Europe	10	100,00%	98,00%	100,00%	99,33%
World-Africa	10	92,50%	100,00%	90,00%	94,17%
World-Asia	10	100,00%	100,00%	96,67%	98,89%
World-US-Canada	10	80,00%	97,50%	100,00%	92,50%
World-Latin-America	10	90,00%	92,50%	100,00%	94,17%
Sport	10	90,83%	100,00%	91,67%	94,17%
Health	10	100,00%	96,67%	100,00%	98,89%
Science	10	70,00%	88,33%	95,00%	84,44%
Entertaiments	10	99,17%	75,00%	95,00%	89,72%

Tableau 2:Résultats de l'évaluation interne (Source : auteur)

Avec :

- *N_correct(X)*= Nombre d'entités de type X désambiguïsées correctement.
- *N(X)* = Nombre d'entités de type X
- *N_ignoré(X)*= Nombre d'entités ignorées de type X

3.7.2 Evaluation comparative

L'objectif est de comparer les performances des 2 systèmes *AIDA* et *AIDA for News* en matière de désambiguïsation, particulièrement vis-à-vis des entités émergentes. Pour cela, nous avons sélectionné sur BBC—News des articles contenant en tout approximativement une cinquantaine de noms propres dont certains sont célèbres et d'autres moins et lancé ensuite les 2 systèmes sur ces articles. Il s'agit des articles énumérés ci-dessous. Les deux derniers articles contiennent des entités émergentes:

- *Obama: 50 countries to take in 360,000 refugees this year*
- *Brexit: PM to trigger Article 50 by end of March*
- *Donald Trump rejects CIA Russia hacking report*
- *Cairobombing: Cairo Coptic Christian complex hit*
- *Hurricane Matthew: Category Four storm pounds Haiti*

Le tableau suivant montre les résultats obtenus.

	P	O	L
AIDA	66,66 %	74,64%	85,82%
A-F-N	100%	94,28%	94,28%

Tableau 3 : Récapitulation des résultats de l'évaluation comparative(Source : auteur)

Figure 7: Echec lors de l'identification des entités émergentes sur AIDA(Source : auteur)

Figure 8: Identification des mêmes entités émergentes sur AIDA-for-News (Source : auteur)

4 Discussions

4.1 Interprétation de l'évaluation interne

a) Tableau 2 nous montre que c'est dans les catégories *science* et *entertainement* que le système est le moins performant (taux

de réussite au-dessous de 90 %) et que les meilleurs résultats sont obtenus dans les catégories *business* et *world* (sociale, politique et économique) où ils présentent tous des taux supérieurs à 90%.

b) Les noms d' organisation sont moins exposés aux confusions dans les textes scientifiques par rapport aux deux autres types. Cela transparaît dans les résultats : 95% de taux de réussite pour le type *Organisation* contre 70% pour le type *Personne* et 88,33% pour le type *Lieu*.

Explications:

a) Les articles des catégories *business* et *world* rapportent toujours des faits réels vérifiables, impliquant des personnes, des organisations et des lieux qui existent réellement, avec des contextes tout aussi réels. Ils répondent ainsi très bien à la vocation du système, aussi bien au niveau de l'étiquetage qu'au niveau de la désambiguïsation. Les articles sur la cience et sur les loisirs par contres ont des terrains de *polysémie* par excellence, où non seulement des objets sont désignés par des noms de lieux ou de personnes, mais qui sont en plus riches en personnages fictifs qui n'ont aucune entité canonique correspondante.

b) Les noms d' organisation sont rarement ambigüs et (presque) jamais - candidats à la polysémie. Particulièrement dans les différents domaines scientifiques, les noms d'organisation sont très *objectifs* et loin d'être fantaisistes.

4.2 Evaluation comparative

Les résultats nous montrent que l'approche cognitive permet de désambiguïser plus de noms que l'approche empirique et ce « plus » qui fait la différence est composée exactement des entités émergentes. Les schémas descriptifs de personnes, qui sont les plus complexes, livrent en contrepartie de cette complexité plus d'informations permettant de les définir et de les identifier, ce qui n'est pas le cas des lieux et des organisations. Ceci explique la légère différence dans leurs résultats. Tant qu'un nom est livré avec un contexte, il est possible de l'identifier et de le désambiguïser.

Au-delà de l'identification des entités émergentes, le recours aux schémas sémantiques a un autre avantage. Il permet parfois de déjouer la polysémie. Dans le dernier article par exemple, AIDA-for-News a pu reconnaître que *Matthew* n'était pas une personne mais un cyclone grâce à l'analyse de contexte qui lui a permis de reconnaître le mot-clé *Hurricane* placé devant. AIDA par contre l'a interprété comme désignant la personne Matthew C. Perry, un contre-amiral de la marine américaine du 19ème siècle.

5.Conclusion

Par rapport à l'objectif défini, les résultats de l'évaluation comparative sont plutôt encourageants car ils révèlent la fiabilité de l'approche cognitive. Parallèlement à cela, les résultats de l'évaluation interne nous ont révélé l'existence d'un tout autre type de problème qui n'a strictement rien à voir avec la notoriété d'une entité.. Il provient tout simplement de la polysémie du langage naturel à laquelle les systèmes d'annotation existants ne sont pas encore préparés. Le traitement de ce problème est en tout cas un défi intéressant à relever et mérite d'être intégré dans les perspectives.

Références :

[1]J. Hoffart,M.A. Yosef,I. Bordino,M. Spaniol,G.Weikum: AIDA: An Online Tool for Accurate Disambiguation of Named Entities in

Text and Tables, Max Planck Institute for Informatics, Saarbrücken, Germany, 2011

[2]M. R. Vicente: La glose comme outil de désambiguïsation référentielle des noms propres purs, Corela – Université de poitiers, 2005

[3]J.R. Finkel, T. Grenager, C. Manning: Incorporating non-local information into information extraction system by Gibbs sampling, Stanford Univesity, 2005

[4]J. Hoffart,M.A. Yosef,I. Bordino,H. Fürstenau, M. Pinkal,M. Spaniol,B. Taneva,S. Thater,G.Weikum: Robust Disambiguation of Named Entities in Text, EMNLP, 2011, 784-785.

[5]J.-Y. Pollock, H.G. Obenhauer, Linguistique et cognition: Réponses à quelques critiques de la grammaire générative. Recherches Linguistiques de Vincennes, n. 19, 1990.

[6]I.H. Witten, E. Frank, M.A. Hall: Data mining Practical machine learning tools and techniques, Morgan Kaufmann, 2011

[7]J. L. Leidner: Toponym Resolution in Text Annotation, Evaluation and Applications of Spatial Grounding of Place Names, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, 2007

[8]I. Bensalem, M.-K.Kholladi: L'utilisation des Chemins hiérarchiques des lieux pour la Désambiguïsation des Toponymes, Département Informatique Université Mentouri Constantine, Algérie2009