TALN 2008, Avignon, 9–13 juin 2008

Exploitation de treillis de Galois en désambiguïsation non supervisée d'entités nommées

Thomas Girault
France Télécom R&D
2, avenue Pierre Marzin 22307 Lannion Cedex
thomas.girault@orange-ftgroup.com

Résumé. Nous présentons une méthode non supervisée de désambiguïsation d'entités nommées, basée sur l'exploitation des treillis de Galois. Nous réalisons une analyse de concepts formels à partir de relations entre des entités nommées et leurs contextes syntaxiques extraits d'un corpus d'apprentissage. Le treillis de Galois résultant fournit des concepts qui sont utilisés comme des étiquettes pour annoter les entités nommées et leurs contextes dans un corpus de test. Une évaluation en cascade montre qu'un système d'apprentissage supervisé améliore la classification des entités nommées lorsqu'il s'appuie sur l'annotation réalisée par notre système de désambiguïsation non supervisée.

Abstract. We present an unsupervised method for named entities disambiguation, based on concept lattice mining. We perform a formal concept analysis from relations between named entities and their syntactic contexts observed in a training corpora. The resulting lattice produces concepts which are considered as labels for named entities and context annotation. Our approach is validated through a cascade evaluation which shows that supervised named entity classification is improved by using the annotation produced by our unsupervised disambiguation system.

Mots-clés : Désambiguïsation non supervisée, treillis de Galois, entités nommées.

Keywords: Unsupervised word sense disambiguation, concept lattice, named entities.

1 Introduction

Le traitement de la polysémie lexicale est un problème linguistique général qui est central dans de nombreuses tâches impliquant la manipulation de données en langue naturelle (*e.g.* recherche d'information, extraction d'information). Nous nous intéressons ici à un type d'unités lexicales (UL) particulier, les entités nommées (EN), une appellation générique pour les noms propres désignant entre autres des personnes, des lieux ou des organisations. Comme la plupart des UL considérées en dehors du contexte d'un énoncé, les EN sont polysémiques, c'est-à-dire qu'elles peuvent faire référence à des sens (et des référents) différents. Notre approche s'inscrit dans le cadre d'un apprentissage non supervisé pour résoudre en partie un problème de désambiguïsation du sens d'UL en contexte.

Plusieurs familles d'approches de désambiguïsation sémantique ont été développées. Une première tente de déterminer le sens de mots en utilisant entre autres la connaissance lexicale ré-

pertoriée dans des dictionnaires ou des thésaurus (*e.g.* WordNet). Les approches qui exploitent des corpus examinent, elles, les occurrences d'UL et de leurs contextes en utilisant des outils d'apprentissage artificiel. Lorsque les UL à désambiguïser sont étiquetées par des sens prédéfinis, les techniques d'apprentissage utilisées sont supervisées. L'apprentissage non supervisé intervient lorsque l'on ne dispose d'aucune annotation pour désambiguïser les UL du corpus.

L'étiquetage de corpus étant une tâche fastidieuse et coûteuse, nous nous intéressons aux approches non supervisées. Nous avons choisi d'utiliser l'*analyse de concepts formels* (ACF) (Ganter & Wille, 1999), une technique de classification non supervisée symbolique qui permet d'inférer des concepts formels pour mettre en relation des regroupements d'EN et des regroupements de contextes syntaxiques. Les concepts formels sont considérées comme des unités de sens qui sont organisées au sein d'une structure hiérarchique de treillis de Galois (ou treillis de concepts). Elle peut être interprétée comme une base de connaissance lexicale modélisant le recouvrement des unités de sens sur plusieurs niveaux de granularité. À notre connaissance, ces propriétés attachées aux treillis de Galois n'ont pas encore été exploitées dans une application de désambiguïsation lexicale. Dans cette perspective, nous proposons d'utiliser les concepts du treillis comme des étiquettes pour désambiguïser des UL de manière non supervisée.

Le plan de cet article est le suivant. Nous partons des approches existantes en désambiguïsation basée sur corpus (Section 2) pour introduire notre démarche. Nous décrivons ensuite notre approche basée sur l'*analyse de concepts formels* (Section 3) nous permettant de construire, à partir, d'un corpus la structure de connaissance des treillis de concepts. Nous détaillons comment exploiter les concepts du treillis pour réaliser une désambiguïsation non supervisée des entités nommées d'un corpus. Notre approche est validée par une évaluation en cascade (Section 4) qui consiste à montrer qu'un système de reconnaissance des entités nommées supervisé s'améliore lorsqu'il exploite l'annotation réalisée par notre système.

2 Désambiguïsation basée sur corpus

Dans cette section, nous introduisons notre problématique de désambiguïsation en donnant une limitation des systèmes d'apprentissage supervisé pour s'adapter à des corpus nécessitant une granularité d'étiquetage spécifique. Nous exposons les principes de l'analyse distributionnelle (Bouaud *et al.*, 1997) dirigée par l'étude de relations syntaxiques entre des UL et qui nécessite une extraction des contextes syntaxiques des EN pour notre problème. Nous expliquons enfin comment représenter l'ensemble de ces relations syntaxiques pour envisager une désambiguïsation des UL selon plusieurs points de vues.

2.1 Corpus et granularité d'étiquetage

De nombreux algorithmes d'apprentissage supervisé sont capables d'exploiter l'annotation sémantique de corpus pour désambiguïser des UL. Les systèmes de reconnaissance des EN les plus efficaces adoptent cette approche pour apprendre à détecter et classer des EN d'un corpus selon un inventaire de sens prédéfini (*e.g.* {`personne`, `lieu`, `organisation`, `autre`}). Il nous semble délicat de répertorier les EN selon ce type classification : cette granularité des sens n'est pas toujours appropriée pour désambiguïser les EN de corpus spécialisés.

À titre d'exemple, nous avons constitué un corpus d'étude à partir duquel nous avons sélec-

tionné quelques énoncés pour illustrer notre problème. Ce corpus a été élaboré en collectant les résultats renvoyés par un moteur de recherche interrogé avec des noms propres de personnes comme requêtes.

(1) filmographie de l'acteur Arnold Schwarzenegger, star de Terminator
(2) victoire d'Arnold Schwarzenegger en Californie
(3) élection d'Arnold Schwarzenegger au poste de gouverneur de la Californie
(4) élection d'Angela Merkel en Allemagne
(5) Angela Merkel dirige la politique gouvernementale
(6) concert de Yannick Noah au Zenith
(7) victoire de Yannick Noah à Roland Garros
(8) concert de Michael Jackson avec les Jackson Five

Dans ce corpus, on peut souhaiter différencier les acteurs, les musiciens, les politiciens et les sportifs : il s'agit d'une désambiguïsation fine des fonctions des EN (distinguer entre Schwarzenegger gouverneur et Schwarzenegger acteur). Cette tâche s'apparente au *template element filling* des conférences MUC (Grishman & Sundheim, 1996) qui visait à remplir une fiche (prédéfinie) pour affiner la description des EN déjà étiquetées dans un document. Notons que nous ne traitons pas le problème de l'ambiguïté des référents d'une EN (des personnes différentes peuvent porter le même nom).

Quelle que soit la tâche de désambiguïsation, il n'est pas envisageable d'annoter manuellement des nouveaux corpus pour s'adapter à un nouveau domaine ou un nouveau jeu d'étiquetage. De plus, les EN font référence à des objets du monde en constante évolution qui semblent délicates à répertorier selon une jeu d'étiquetage stable. La constitution d'un dictionnaire d'EN exhaustif serait coûteuse à mettre en place et à adapter à de nouveaux types de corpus. Comment déterminer automatiquement les différentes fonctions d'une EN dans un corpus sans utilisation d'un dictionnaire ou de corpus étiquetés ?

2.2 Contextes distributionnels des entités nommées

Plutôt que d'assigner des étiquettes connues à des UL, une stratégie alternative est de discriminer leurs sens en analysant leurs interactions avec d'autres UL présentes dans les énoncés d'un corpus. Cette approche non supervisée est développée à partir de l'hypothèse de Harris selon laquelle des UL (des EN en l'occurrence) qui apparaissent dans des contextes similaires tendent à avoir des sens proches. Les méthodes basées sur l'analyse distributionnelle (Bouaud *et al.*, 1997) s'appuient sur cette hypothèse et considèrent que le partage de contextes disposés selon une même configuration syntaxique (e.g. sujet-verbe, modifieur-nom) constitue un indice de proximité sémantique.

Nous adoptons ces dernières considérations pour notre problème : nous avons donc besoin d'identifier les UL pouvant être rattachées syntaxiquement aux EN des énoncés de notre corpus. Pour le moment, nous privilégions l'extraction des noms, verbes et prépositions qui entretiennent des liens syntaxiques avec des EN. Les patrons d'extraction que nous avons défini manuellement s'appuient sur un étiquetage morpho-syntaxique. Ils se traduisent par des constructions du type :
– nom commun + EN dans un chunk nominal (*e.g.* [acteur, Arnold Schwarzenegger]).

– nom commun + préposition + EN (*e.g.* [élection de, Angela Merkel], [victoire de, Arnold Schwarzenegger], [victoire de, Yannick Noah]).
– EN + chunk verbal (*e.g.* [Angela Merkel, dirige])

Ce travail d'extraction fournit un ensemble de couples (entité nommée, contexte syntaxique) dans lesquels chaque élément est potentiellement polysémique : l'entité nommée "Yannick Noah" peut se référer à la fois à un musicien et un sportif et le contexte syntaxique "victoire de" peut désigner une victoire politique ou sportive. Quelles représentations et quelles méthodes sont appropriées pour prendre en compte deux unités polysémiques en interaction ?

2.3 Bipartition des unités lexicales

Les approches distributionnelles classiques sont bien entendu applicables pour traiter la polysémie de toutes les UL. Cependant, elles les considèrent d'un seul point de vue alors que pour notre problème les données semblent plus naturellement représentées par deux vues interconnectées comme le montre la figure (1) :

– une vue sur les entités nommées qui est associée à un ensemble d'objets $O = \{o1, o2, \cdots, om\}$.
– une vue sur leurs contextes syntaxiques représentée par un ensemble d'attributs $A = \{a1, a2, \cdots, an\}$

Ces deux vues sont connectées par une relation $R \subseteq O \times A$, où $R(o, a)$ signifie que l'objet o possède l'attribut a (*i.e.* l'EN o possède le contexte syntaxique a).

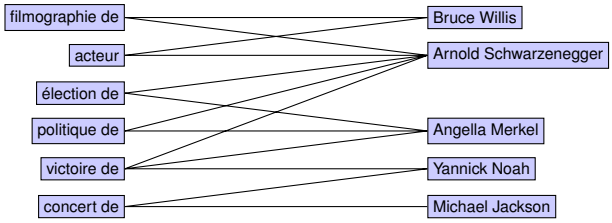

FIG. 1 – Relations entre les EN et leur contexte syntaxique local

Cette structure est un graphe biparti et ce mode de représentation est utilisé en *co-training* (Blum & Mitchell, 1998), une famille d'algorithmes d'apprentissage supervisé qui exploite plusieurs vues indépendantes et redondantes sur les données. Selon cette approche, il serait avantageux de combiner les prédictions réalisées par des apprentissages indépendants sur chaque vue. Dans le cadre de la classification supervisée d'EN, on cherche à déterminer l'étiquette $tag(x) \in Tagset$ d'un échantillon x représenté par deux vues : l'EN o et son contexte syntaxique a. Pour (Blum & Mitchell, 1998), la combinaison des prédictions peut être formulée par un produit de probabilités estimées avec des classifieurs bayesiens naïfs entraînés sur chaque vue :

$$tag(x) = \underset{c \in Tagset}{\text{argmax}}\ P(c|o)P(c|a) \tag{1}$$

Le *co-training* et ses versions dérivées ont été appliquées avec succès en désambiguïsation lexicale semi-supervisée (Yarowsky, 1995) et en extraction d'information (Jones, 2005). Ces travaux mériteraient d'être étendus au cadre de l'apprentissage non supervisé et c'est dans cette optique que nous avons emprunté une méthode appelée *analyse de concepts formels* (AFC) pour notre problème. Elle nous permet de réaliser en parallèle des regroupements d'EN et des regroupements de contextes syntaxiques. Ces deux types de regroupements sont attachés à des vues qui sont combinées pour désambiguïser à la fois les EN et leurs contextes syntaxiques.

3 Utilisation des treillis en désambiguïsation lexicale

Dans cette section, nous décrivons les principes de notre méthodologie basée sur l'ACF pour désambiguïser des EN selon une approche non supervisée. Nous donnons d'abord plusieurs définitions formelles que nous illustrons par un exemple d'application de l'ACF sur un petit corpus avant de détailler les principes de notre méthode de désambiguïsation.

3.1 Analyse de concepts formels

En analyse de concepts formels, le triplet (O, A, R) est appelé contexte formel. (O, A, R) correspond au graphe biparti (figure (1)) des objets (entités nommées) en relation avec des attributs (contextes lexico-syntaxique).

Pour $E \subseteq O$ et $I \subseteq A$, nous définissons deux ensembles $E' \subseteq A$ et $I' \subseteq O$ les étendant : $E' = \{a \in A | \forall o \in E : (o, a) \in R\}$ comme l'ensemble maximal des attributs communs aux objets de E et $I' = \{o \in O | \forall a \in C : (o, a) \in R\}$, l'ensemble maximal des objets partageant les attributs de I. Par exemple, si $I = \{$politique de, élection de$\}$ alors $I' = \{$Angela Merkel, Arnold Schwarzenegger$\}$. Pour $E = \{$Michael Jackson$\}$, on a, $E' = \{$album de, concert de, interview de, fan de$\}$.

Un concept formel d'un contexte formel (O, A, R) est une paire (E, I) telle que $E \subseteq O, I \subseteq A$, $E' = I$ et $I' = E$. On appelle E, l'extension du concept et I l'intention du concept. Par exemple la paire ($\{$Angela Merkel, Arnold Schwarzenegger$\}$,$\{$politique de, victoire de, élection de$\}$) est un concept formel. Pour $o \in O$ et $a \in A$, $\{o\}'$ est dénoté par o' et appelé l'intention d'un objet, et $\{a\}'$ est dénoté par a' et est appelé l'extension d'un attribut.

Les concepts formels sont partiellement ordonnés les uns par rapport aux autres selon une relation d'inclusion entre les extensions des concepts. Un concept $C_1 = (E_1, I_1)$ est plus spécifique ou égal qu'un concept $C_2 = (E_2, I_2)$ (ou $C_1 \leq C_2$) si $E_1 \subseteq E_2$ ou de manière équivalente $I_2 \subseteq I_1$. Par exemple, on a $C1 \leq C2$ pour les concepts $C1 = (\{$Michael Jackson, Yannick Noah$\}, \{$album de, concert de, interview de, fan de$\})$ et $C2 = (\{$Michael Jackson, Yannick Noah, Arnold Schwarzenegger$\}, \{$interview de, fan de$\})$.

Soit deux concepts X et Y : si $X \leq Y$ et qu'il n'y a pas de concept Z tels que $Z \neq X, Z \neq Y, X \leq Z \leq Y$, X est un enfant de Y et Y est un parent de X. Cette relation de parenté entre les concepts formels permet d'établir un graphe orienté des concepts formels qui est appelé treillis de Galois (ou treillis de concepts).

À quoi ressemble un treillis de concept construit à partir de relations entre des unités lexicale ? Cette structure est elle appropriée pour traiter une partie du problème de la polysémie lexicale ?

3.2 Constitution du treillis : structure de connaissance discriminante

Pour illustrer les définitions formelles énoncées dans la partie précédente, nous avons repris le contexte formel de la figure (1) pour constituer le treillis représenté graphiquement par le diagramme de Hasse de la figure (2).

La méthodologie employée pour bâtir le treillis de concepts à partir de notre corpus peut s'apparenter aux travaux de (Bendaoud *et al.*, 2007; Cimiano *et al.*, 2005). Nous avons utilisé l'al-

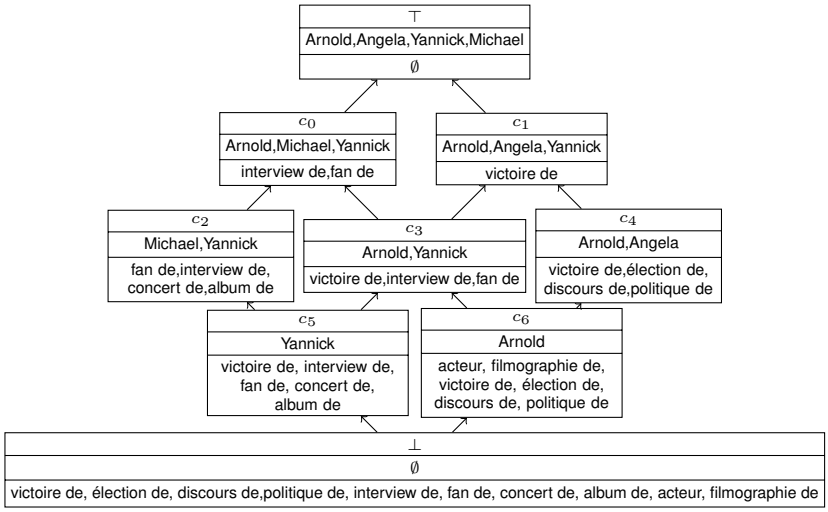

FIG. 2 – Treillis de Galois associé au contexte formel de la figure (1)

gorithme *AddIntent* (van der Merwe *et al.*, 2004) qui adopte une procédure de construction incrémentale : le treillis est construit dynamiquement avec les nouveaux objets et attributs découverts lors de l'extraction des UL du corpus. Cette construction permet d'utiliser un treillis comme une base de connaissances déjà structurée que l'on souhaiterait adapter à un nouveau corpus, propriété très intéressante au regard de la faible évolutivité de nombreuses ressources lexicales structurées. C'est dans cette perspective que (Priss & Old, 2004) à élaboré le modèle d'*analyse de concepts relationnels* pour une représentation FrameNet. Ces derniers travaux révèlent d'excellentes propriétés associées aux treillis pour traiter l'ambiguïté lexicale.

Comme nous pouvons l'observer en figure (2), différents niveaux de granularité apparaissent au sein de la structure hiérarchique d'un treillis. La partie supérieure du treillis englobe des concepts généraux qui regroupent plusieurs objets partageant des attributs ambigus. À l'opposé, la partie inférieure présente des concepts très spécifiques qui contiennent des objets ambigus. Les concepts de la zone supérieure se différencient dans une zone intermédiaire qui nous semble la plus informative. Les regroupements d'objets et d'attributs correspondent à des unités de sens qui semblent plus cohérentes pour désambiguïser à la fois les objets et les attributs. De manière générale, on observe une tendance au recouvrement des concepts qui traduit une certaine continuité entre les unités de sens alors que le modèle des treillis est normalement considéré comme symbolique et discret. À notre connaissance, ces propriétés attachées aux treillis n'ont pas encore été exploitées pour résoudre le problème de la désambiguïsation non supervisée et c'est dans cette perspective que nous cherchons à les exploiter.

3.3 Annotation conceptuelle pour désambiguïser les UL en interaction

Nous présentons ici notre modèle d'exploitation des treillis de Galois en désambiguïsation lexicale. Nous considérons désormais chaque concept structuré dans un treillis comme une étiquette sémantique potentiellement utile pour désambiguïser une UL en contexte.

Dans un énoncé, on suppose qu'une EN $o \in O$ et son contexte lexico-syntaxique $a \in A$ ont

été détectés avec les patrons d'extraction que nous avons élaborés. Le couple (o, a) est ajouté au treillis à l'aide de l'algorithme *AddIntent*. Nous cherchons à étiqueter le couple (o, a) par un couple de concepts (X, Y) du treillis. Notre proposition est la suivante. L'extension $E_x = Ext(X)$ du concept X qui annote o possède des objets similaires aux objets de l'extension de a (*i.e.* E_x doit être proche de $E_a = \{a\}'$). L'intention $I_y = Int(Y)$ du concept Y qui annote a possède des attributs similaires aux attributs de l'intention de o (*i.e.* I_y doit être proche de $I_o = \{o\}'$). Les deux concepts (X, Y) à déterminer se situent sur des chemins entre le concept $C_o = (\{o\}, \{o\}')$ associé à o et le concept $C_a = (\{a\}', \{a\})$ associé à a. Plus formellement, on a $C_a \leq X \leq C_o$ et $C_a \leq Y \leq C_o$: l'ensemble des concepts associés à cet intervalle se substitue au jeu d'étiquettes prédéfinies $Tagset$ dans la formule (1).

$$X = \underset{C_a \leq X \leq C_o}{\text{argmax}} \quad P(E_x|E_o)P(E_x|E_a) \tag{2}$$

$$Y = \underset{C_a \leq Y \leq C_o}{\text{argmax}} \quad P(I_y|I_o)P(I_y|I_a) \tag{3}$$

L'inférence bayesienne s'applique sur chaque vue du treillis (Knuth, 2004) et le terme $P(c|o)P(c|a)$ de la formule (1) est remplacé par $P(E_x|E_o)P(E_x|E_a)$ pour déterminer l'extension E_x la plus probable par rapport à E_o ou E_a pour annoter l'EN o. De manière similaire, l'estimation de $P(I_y|I_o)P(I_y|I_a)$ permet de choisir l'intention I_y pour le contexte syntaxique a.

À titre d'exemple, on peut vouloir désambiguïser l'expression "la victoire d'Arnold Schwarzenegger" dans lequel on extrait le couple $(o, a) =$ (Arnold Schwarzenegger,victoire de). L'objet o est représenté dans le treillis par le concept $C_6 = (\{$Arnold Schwarzenegger$\}, \{$politique de, discours de, victoire de, élection de, acteur, filmographie de, film avec, fan de$\})$ qui décrit des rôles d'homme politique tout autant que d'acteur de cinéma. L'attribut "victoire de" est aussi polysémique car une victoire peut être politique ou sportive. Cet attribut est représenté dans le treillis par le concept $C_1 = (\{$Arnold Schwarzenegger, Angela Merkel, Yannick Noah$\}$,$\{$victoire de$\})$. Les chemins possibles entre C_1 et C_6 sont $C_1C_4C_6$ et $C_1C_4C_6$. L'application des formules (2) et (3) permet de selectionner (C_4, C_4) pour annoter (o, a).

Notre proposition peut s'apparenter à plusieurs modèles qui ont été appliqués au calcul de similarité entre des ensembles de mots dans des ressources lexicales structurées. Les travaux de (Jacquet *et al.*, 2005) sur les graphes de synonymes ont permis de définir des unités de sens comme des cliques du graphes. Les auteurs utilisent des mesures basées sur le χ^2 pour estimer le degré d'affinité entre deux cliques et choisir le sens d'une UL. Nous pouvons noter que la notion de concept formel s'apparente à celle de clique : un concept formel pouvant être défini comme une bi-clique dans un graphe biparti. Dans le cadre de l'utilisation de WordNet en désambiguïsation lexicale, (Resnik, 1999) propose une mesure de similarité conceptuelle hybride : elle combine des contraintes structurelles (plus petit subsumant commun à deux UL) avec des critères probabilistes estimés sur corpus.

4 Évaluation

Il serait délicat de demander à un expert de juger de la qualité des concepts qui ont été produits par ACF : cette activité est subjective par nature car différents concepts pourraient être appropriés pour désambiguïser une EN. Il nous semble préférable de valider notre approche par rapport à une tâche existante. Nous proposons d'évaluer la pertinence de l'annotation non supervisée en prouvant que celle-ci apporte une connaissance nouvelle et intéressante pour améliorer une tâche de classification des EN supervisée.

Dans cette section, nous présentons d'abord le corpus utilisé pour cette tâche. Nous décrivons ensuite le protocole d'évaluation en cascade (Candillier *et al.*, 2006) que nous avons utilisé pour transférer les connaissances apprises avec notre système non supervisé vers le système supervisé. Cette section s'achève avec une étude des résultats fournis par notre système.

4.1 La campagne d'évaluation CoNLL 2003

Pour nos expérimentations, nous avons utilisé un corpus de la campagne d'évaluation CoNLL 2003 (Sang & Meulder, 2003) qui a été largement repris pour évaluer les technologies de reconnaissance des EN. Il est constitué d'un ensemble d'articles de presse anglophone (collection Reuters) découpés en phrases. Celles-ci sont fournies avec un étiquetage morpho-syntaxique et une segmentation en chunks qui ont été produits automatiquement et sont donc bruités. Les EN des phrases sont délimitées et annotées manuellement selon le jeu d'étiquettes {personne, lieu, organisation, "autre"}. Bien qu'il puisse paraître limité, c'est par rapport à ce jeu d'étiquettes que nous allons nous évaluer. La figure (3) donne un aperçu des caractéristiques du corpus.

	Articles	Phrases	Tokens	Lieux	Autres	Organisations	Personnes
Corpus d'apprentissage (train)	946	14987	203621	7140	3438	6321	6600
Corpus de développement (testa)	216	3466	51362	1837	922	1341	1842
Corpus de test (testb)	231	3684	46435	1668	702	1661	1617

FIG. 3 – Caractéristiques du corpus CoNLL 2003

4.2 Évaluation en cascade

En se plaçant dans le cadre d'une évaluation en cascade (Candillier *et al.*, 2006), on considère qu'un apprentissage non supervisé est un prétraitement d'une tâche supervisée que l'on sait évaluer. Cet assemblage en cascade permet de vérifier si notre système de désambiguïsation non supervisée apporte une connaissance nouvelle et intéressante pour améliorer une tâche de classification supervisée des EN sur le corpus CoNLL. Il s'agit de comparer les erreurs produites par deux classifieurs A et B, lorsqu'ils sont évalués sur le corpus de test (testb), après entraînement sur le même corpus d'apprentissage (train + testa).

Le système A est élaboré à partir d'un apprentissage supervisé sur le corpus d'apprentissage étiqueté. À la manière de (Ehrmann & Jacquet, 2006), le système B produit une double annotation des EN. La première couche d'étiquetage est appliquée avec une annotation conceptuelle suite à une AFC effectuée à partir du corpus d'apprentissage non étiqueté. Cette étape de prétraitement permet d'enrichir la description des corpus. Le système B pourra bénéficier de cet enrichissement durant la phase apprentissage supervisé pour produire la deuxième couche d'étiquetage.

4.3 Premiers résultats avec l'algorithme de Brill

Nous avons adapté l'algorithme *transformation-based learning* (Brill, 1995) pour réaliser un système de reconnaissance des EN supervisé. L'algorithme initialise la classification des EN grâce à un modèle de langue de type unigramme, entraîné sur le corpus d'apprentissage. La suite de la procédure est itérative : l'algorithme corrige progressivement la classification initiale

erronée en inférant une séquence de règles correctrices. Elles sont appliquées successivement sur les échantillons d'un corpus pour améliorer progressivement la classification des EN. Les règles produites par le système sont instanciées à partir d'une liste de patrons d'extraction définis manuellement. Ces patrons sont capables d'explorer les mots du contexte dans une fenêtre de +/- 3 mots. Différents types de propriétés sont pris en compte par les patrons : le mot, son étiquette morpho-syntaxique et, dans le cas du classifieur B, son étiquette conceptuelle produite par notre système non supervisé.

Le figure (4) présente les résultats de l'évaluation en cascade. La colonne de gauche présente les scores pour l'étiqueteur A appliqué sur les données de test dotées d'un étiquetage morpho-syntaxique. La colonne de droite correspond aux résultats obtenus avec l'étiqueteur B qui a été utilisé sur le jeu de test enrichi par l'annotation conceptuelle issue du treillis de Galois.

	A : Brill simple			B : annotation conceptuelle + Brill		
	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$
Lieu	66.56%	66.19%	66.38	75.09%	65.65%	70.06
Organisation	52.22%	55.18%	53.66	61.55%	46.91%	53.24
Personne	59.68%	68.62%	63.84	75.32%	57.82%	65.42
Autre	83.58%	60.74%	70.35	85.21%	67.46%	75.30
Total	62.67%	63.61%	63.14	73.81%	59.27%	65.75

FIG. 4 – Résultats de l'évaluation en cascade

Notre système de désambiguïsation non supervisé apporte donc une amélioration de 11.14% en précision, une régression de 4.4% en rappel et une progression de 2.61 pour $F_{\beta=1}$.

5 Conclusion, discussion et perspectives

Nous avons présenté une méthode non supervisée de désambiguïsation d'entités nommées, basée sur l'analyse de concepts formels. Cette technique produit un treillis de Galois pour hiérarchiser des relations entre des EN et leurs contextes syntaxiques qui ont été extraits de corpus. Le treillis résultant réalise une double structuration des UL facilitant l'interprétation linguistique, le lien avec des ressources lexicales existantes et l'adaptation à des corpus. Nous tirons profit de cette double structuration pour sélectionner les concepts désambiguïsant à la fois les entités nommées et leurs contextes syntaxiques. L'évaluation en cascade montre qu'un système de reconnaissance des EN supervisé améliore la précision d'étiquetage des entités nommées lorsqu'il s'appuie sur l'annotation réalisée par notre système de désambiguïsation non supervisée.

Les premiers résultats sont encourageants même si notre système n'atteint pas encore les performances obtenues par d'autres auteurs. Notre procédure d'extraction syntaxique réalisée en amont pourrait être améliorée par l'utilisation d'un analyseur en dépendances permettant de couvrir d'avantage de configurations syntaxiques et d'identifier différents types de dépendances (*e.g.* sujet-objet, verbe-objet). Ces différents types de relations pourraient permettre de constituer plusieurs treillis à combiner : cette stratégie pourrait réduire significativement le nombre de concepts assurant ainsi une meilleure généralisation. D'autre part nous devrons nous comparer avec d'autres modèles comme l'analyse distributionnelle classique et les techniques d'apprentissage non supervisé dérivées du *co-training*. Enfin, nous devrons vérifier que le système est capable de tenir la charge imposée par les volumes de données importants.

Remerciements : nous remercions vivement Pascale Sébillot, Laurent Candillier et Maxime Amblard pour leurs conseils et leurs remarques à la lecture de cet article.

Références

BENDAOUD R., HACENE M. R., TOUSSAINT Y., DELECROIX B. & NAPOLI A. (2007). Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. In *18e Journées Francophones d'Ingénierie des Connaissances (IC'2007)*.

BLUM A. & MITCHELL T. (1998). Combining labeled and unlabeled data with co-training. In *COLT : Proceedings of the Workshop on Computational Learning Theory*, p. 92–100 : Morgan Kaufmann Publishers.

BOUAUD J., HABERT B., NAZARENKO A. & ZWEIGENBAUM P. (1997). Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. In *Ingéniérie de la connaissance (IC'97)*, p. 207–223, Roskoff.

BRILL E. (1995). Transformation-based error-driven learning and natural language processing : A case study in part-of-speech tagging. *Computational Linguistics*, **21**(4), 543–565.

CANDILLIER L., TELLIER I., TORRE F. & BOUSQUET O. (2006). évaluation en cascade d'algorithmes de clustering. In L. MICLET, Ed., *8ième Conférence francophone sur l'Apprentissage automatique (CAp'2006)*, p. 109–124.

CIMIANO P., HOTHO A. & STAAB S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, **24**, 305–339.

EHRMANN M. & JACQUET G. (2006). Vers une double annotation des entités nommées. *Traitement automatique des langues*, **47**, 63–88.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis*. Springer-Verlag.

GRISHMAN R. & SUNDHEIM B. (1996). Message understanding conference- 6 : A brief history. In *COLING*, p. 466–471.

JACQUET G., VENANT F. & VICTORRI B. (2005). *Polysémie lexicale*, In P. ENJALBERT, Ed., *Sémantique et traitement automatique du langage naturel*. Hermès Sciences.

JONES R. (2005). *Learning to Extract Entities from Labeled and Unlabeled Text*. PhD thesis, Carnegie Mellon University.

KNUTH K. (2004). Lattice duality : the origin of probability and entropy. *Neurocomputing*, **67C**, 245–274.

PRISS U. & OLD L. J. (2004). Modelling lexical databases with formal concept analysis. *J. UCS*, **10**(8), 967–984.

RESNIK P. (1999). Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**, 95–130.

SANG E. F. T. K. & MEULDER F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *CoRR*, **cs.CL/0306050**.

VAN DER MERWE D., OBIEDKOV S. A. & KOURIE D. G. (2004). Addintent : A new incremental algorithm for constructing concept lattices. In P. W. EKLUND, Ed., *ICFCA*, volume 2961 of *Lecture Notes in Computer Science*, p. 372–385 : Springer.

YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, p. 189–196.