Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées

Frédéric Béchet, Benoît Sagot, Rosa Stern

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées

Frédéric Béchet[1], Benoît Sagot[2], Rosa Stern[2,3]
(1) Aix Marseille Université, LIF-CNRS, route de Luminy, Marseille
(2) Alpage, INRIA & Univ. Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(3) Agence France-Presse – Medialab, 2 place de la Bourse, 75002 Paris, France
frederic.bechet@lif.univ-mrs.fr, benoit.sagot@inria.fr, rosa.stern@afp.com

Résumé. La détection et le typage des entités nommées sont des tâches pour lesquelles ont été développés à la fois des systèmes symboliques et probabilistes. Nous présentons les résultats d'une expérience visant à faire interagir le système à base de règles NP, développé sur des corpus provenant de l'AFP, intégrant la base d'entités Aleda et qui a une bonne précision, et le système LIANE, entraîné sur des transcriptions de l'oral provenant du corpus ESTER et qui a un bon rappel. Nous montrons qu'on peut adapter à un nouveau type de corpus, de manière non supervisée, un système probabiliste tel que LIANE grâce à des corpus volumineux annotés automatiquement par NP. Cette adaptation ne nécessite aucune annotation manuelle supplémentaire et illustre la complémentarité des méthodes numériques et symboliques pour la résolution de tâches linguistiques.

Abstract. Named entity recognition and typing is achieved both by symbolic and probabilistic systems. We report on an experiment for making the rule-based system NP, a high-precision system developed on AFP news corpora and relies on the Aleda named entity database, interact with LIANE, a high-recall probabilistic system trained on oral transcriptions from the ESTER corpus. We show that a probabilistic system such as LIANE can be adapted to a new type of corpus in a non-supervized way thanks to large-scale corpora automatically annotated by NP. This adaptation does not require any additional manual anotation and illustrates the complementarity between numeric and symbolic techniques for tackling linguistic tasks.

Mots-clés : Détection d'entités nommées, adaptation à un nouveau domaine, coopération entre approches probabilistes et symboliques.

Keywords: Named entity recognition, domain adaptation, cooperation between probabilistic and symbolic approaches.

1 Introduction

La reconnaissance d'entités nommées est une des tâches les plus étudiées du domaine du traitement automatique des langues. Reconnaître dans du texte ou de la transcription de parole les mentions d'entités nommées reste un préalable nécessaire avant toute tâche plus complexe telle que l'analyse syntaxique, l'analyse sémantique ou l'extraction d'informations. Ainsi, la tâche de reconnaissance d'entités nommées fait l'objet de nombreuses campagnes d'évaluation depuis plus d'une vingtaine d'années, dont les premières ont été les campagnes MUC (Message Understanding Conference). Ces campagnes ont donné lieu à la construction de corpus de référence, notamment pour l'anglais, le chinois, l'espagnol ou le japonais. Pour le français, on peut citer notamment l'évaluation menée dans le cadre de la campagne ESTER sur des transcriptions de nouvelles, qui a donné naissance à un corpus français annoté en entités nommées selon des directives assez différentes de celles des campagnes MUC, notamment pour les emplois polysémiques et métaphoriques (Galliano *et al.*, 2009). Pour une discussion plus précise de ces questions et des différents types d'ambiguïtés rencontrées en reconnaissance des entités nommées, on pourra se reporter à (Béchet, 2011).

Toutes les campagnes d'évaluation ont montré que la tâche de reconnaissance d'entités nommées peut être traitée efficacement aussi bien avec des systèmes symboliques qu'avec des systèmes probabilistes. Elles ont également montré que les systèmes symboliques couplés à de très grands lexiques donnent de meilleurs résultats que les

systèmes probabilistes sur du texte canonique, en particulier en termes de précision. En revanche lorsque le texte à traiter contient du bruit (absence de capitalisation, de formatage, de ponctuation, sortie d'un système de reconnaissance de la parole...), les systèmes probabilistes sont plus robustes et obtiennent de meilleurs scores en rappel, du moins tant que les données à annoter sont d'un genre similaire à celui du corpus de leur corpus apprentissage, voire que les entités nommées y sont globalement les mêmes.

C'est à partir de ce constat que nous avons décidé de tirer le meilleur parti des avantages des deux types de systèmes, en effectuant des expériences d'adaptation non supervisée du système probabiliste LIANE au domaine des dépêches d'agence, qui n'est pas celui du corpus ESTER sur lequel il est entraîné. Nous avons utilisé pour cela les annotations produites par le système symbolique NP couplé à la base d'entités Aleda, NP ayant été développé plus particulièrement pour le traitement de dépêches d'agence. Après avoir décrit successivement NP (section 2) et LIANE (section 3), nous décrivons le détail de ce processus d'adaptation (section 4) puis en montrons les résultats sur le corpus d'évaluation formé de dépêches AFP développé par (Stern & Sagot, 2010a) (section 5).

2 NP, système à base de règles reposant sur la ressource lexicale Aleda

Le système NP, dont une version précédente a été décrite par Stern & Sagot (2010a), est un système de détection, typage et résolution d'entités nommées développé initialement pour le français et en cours d'adaptation à l'anglais. Bien que générique, il a été conçu en priorité pour le traitement de dépêches de l'Agence France-Presse. Il est donc adapté aux données traitées dans ce travail. NP est constitué de deux modules, l'un pour la détection ambiguë et l'autre pour la désambiguïsation et la résolution des mentions.

Les deux modules qui constituent NP reposent sur **Aleda**, une base de données d'entités et de mentions possibles de ces entités construite automatiquement à partir de ressources librement disponibles. Aleda est distribuée librement en tant que composante du projet lexical libre Alexina [1]. Une version préliminaire d'Aleda (intégrée à l'époque à la distribution de SXPipe) est décrite par Stern & Sagot (2010b). Aleda contient 855 403 entités et 2,32 million de variantes dénotationnelles qui dénotent des lieux (LOC), des organisations et entreprises (ORG), des personnes (PERS), des produits, des noms d'œuvres, des noms de produits et de marques, des animaux remarquables (*Dolly*) et des personnages de fiction (*Arsène Lupin*). Ces entités ont été extraites principalement à partir de deux sources d'informations librement disponibles, à savoir `geonames` pour les noms de lieu [2] et la Wikipedia française pour les autres types d'entités [3]. Chaque entité a un identifiant unique qui contient un identifiant de la ressource d'où elle a été extraite, un identifiant interne à chaque ressource (un identifiant `geonames` ou un identifiant d'article de la Wikipedia française) et des informations supplémentaires dont un poids heuristique (le nombre d'habitants du lieu pour `geonames`, le nombre de lignes de l'article correspondant dans Wikipedia).

Pour les noms de lieux, nous avons tout d'abord extrait de la volumineuse base `geonames` les entités dont le type a été jugé pertinent pour le traitement de dépêches d'agence (villes, pays, etc., mais pas les noms de montagnes par exemple). Même ainsi filtrée, la base `geonames` contient beaucoup trop d'entités pour que toutes puissent être utilisées. Nous avons donc sélectionné heuristiquement un certain nombre d'entités jugées pertinentes au vu de la nature des données à traiter, à savoir des dépêches AFP émanant du bureau français de l'Agence (474 509 entités de lieu). Pour les autres types d'entités, nous nous sommes appuyés sur la Wikipedia française, dans la lignée de travaux antérieurs (Balasuriya *et al.*, 2009; Charton & Torres-Moreno, 2009). Pour optimiser la couverture de la ressource extraite, nous avons procédé à une extraction en deux étapes, qui repose sur les *catégories wikipedia* [4]. Ce mécanisme, bien qu'initié manuellement par un petit volume de données annotées, permet de récupérer un nombre très important d'entités nommées de la Wikipedia (400 169 entités). Pour chacune de ces entités, nous extrayons un nom normalisé qui est le titre de son article, un poids correspondant simplement au nombre de lignes de l'article, ainsi que différentes variantes dénotationnelles à partir des liens de redirection. Pour les noms de personnes, des variantes supplémentaires sont calculées avec prénoms abrégés et sans prénoms.

1. `http://gforge.inria.fr/projects/alexina/`
2. `geonames` est librement téléchargeable sur `http://www.geonames.org`.
3. `http://download.wikimedia.org/frwiki/latest/frwiki-latest-pages-articles.xml.bz2`
4. Tout d'abord, nous avons associé manuellement à quelques dizaines de catégories wikipedia un type associé (par exemple *Né en 1983* indique presque certainement une entité de type *personne*). Ceci nous a permis de typer un certain nombre d'articles wikipedia, dont nous avons extrait *toutes* les catégories. Nous avons alors identifié les catégories associées à au moins 2 articles et permettant de prédire le type des entités déjà typées avec une précision acceptable (75% sauf pour les catégories rares). Nous avons alors repris l'ensemble des articles de Wikipedia, et pour chaque article nous avons compté pour chaque type possible le nombre de fois qu'il est déclenché par l'une des catégories associées à l'article. Le type le mieux classé est alors attribué à l'article.

Le résultat de ces deux processus d'extraction est enrichi et corrigé par de courtes listes d'entités à éliminer et à ajouter, développées à la main. Les 855 403 entités qui forment ainsi Aleda se répartissent de la façon suivante : 573 074 lieux et monuments, 215 179 personnes, 29 181 titres d'œuvres, 25 247 organisations, 8 862 entreprises, 3 846 produits et marques, 997 personnages de fiction et 17 animaux remarquables.

Une grammaire non-contextuelle de 137 règles a été développée pour **détecter et typer les mentions** d'entités nommées à partir des mentions présentes dans Aleda ; des motifs contextuels (p.ex. *ville/localité/village de* ou *Dr/M./Mme...*) sont également utilisés. La reconnaissance est faite de façon ambiguë par l'architecture `dag2dag` de SXPipe (Sagot & Boullier, 2008). Des heuristiques de désambiguïsation sont appliquées pour réduire en partie cette ambiguïté. La sortie de ce module est donc un graphe (DAG) dans lequel chaque entité nommée candidate (empan et type) est représentée par une transition différente.

Un second module effectue alors deux tâches de façon conjointe : il résout les ambiguïtés d'empan et de type (PERS, LOC, ORG...) et assigne à chaque mention une entrée dans la base (résolution). Comme pour la détection, ce module de typage et de résolution repose sur des heuristiques utilisant des informations qualitatives et quantitatives. Il fait par exemple usage du poids attribué aux entités par Aleda ainsi que de la notion de *saillance* pour favoriser les analyses impliquant des entités déjà rencontrées dans le même document (ici, la même dépêche) ou qui sont cohérentes avec le contexte (pays, ville) identifié dans le document.

3 LIANE, système hybride génératif/discriminant

Le processus de détection d'entités nommées dans un texte est composé de deux sous-tâches : une tâche de segmentation consistant à trouver les bornes de début et de fin des entités ; une tâche de classification consistant à attribuer la bonne catégorie à l'entité détectée (PERS, LOC, ORG...). Ces deux tâches peuvent être effectuées de manière jointe si le processus de classification est appliqué au niveau des mots. Dans ce cas chaque mot appartenant à l'expression d'une entité nommée reçoit une étiquette correspondant à la fois au type de l'entité mais aussi à sa position à l'intérieur de celle-ci. Tous les mots ne participant pas à l'expression d'une entité reçoivent une étiquette vide. Trois étiquettes de position sont utilisés : *B* indique le début d'un chunk ; *I* indique que le mot est situé à l'intérieur d'un chunk mais ne le débute pas ; et *O* correspond à tous les mots hors chunks.

En considérant le processus de détection des entités nommées comme un processus d'étiquetage, grâce aux étiquettes de positionnement et de type d'entités, toutes les méthodes développées dans le cadre de l'étiquetage en parties du discours (POS) peuvent être utilisées, notamment les méthodes numériques. Parmi celles-ci, deux principales approches ont été suivies : les modèles génératifs tels que les Modèles de Markov Cachés (Hidden Markov Models - HMM) (Bikel *et al.*, 1999) et les modèles discriminants tels que MaxEnt (Brothwick *et al.*, 1998) ou les Champs de Markov Aléatoires (Conditional Random Field - CRF) (McCallum & Li, 2003).

Le système LIANE utilisé dans cette étude est décrit dans Bechet & Charton (2010). Il est basé sur une approche mixte utilisant tout d'abord un processus génératif à base de HMM pour prédire une étiquette syntaxique (POS) et sémantique à chaque mot d'un texte ; ensuite un processus discriminant à base de CRF est utilisé pour trouver les bornes et le type complet de chaque entité en utilisant le modèle BIO présenté précédemment. Il y a deux raisons pour justifier l'emploi d'un HMM en amont d'un étiqueteur à base de CRF :

– Tout d'abord les modèles d'étiquetage par HMM permettent de rajouter très facilement plusieurs sources d'information dans les estimations des probabilités des mots sachant les classes. Ils sont également assez tolérants au bruit dans les données d'apprentissage, à partir du moment où les fréquences relatives des différents événements modélisés sont respectées.
– Ensuite cette première phase d'étiquetage va permettre d'enlever un certain nombre d'ambiguïtés en attribuant des étiquettes syntaxiques et sémantiques aux mots d'un texte. Cela permettra de simplifier l'étiquetage par le CRF qui pourra se concentrer sur les aspects de segmentation et d'attribution d'une étiquette finale.

Les modèles HMM et CRF sont appris sur un corpus annoté selon le format suivant (sur la phrase d'exemple "*Investiture aujourd'hui à Bamako Mali du président Amadou Toumani Touré*") :

```
investiture     NFS      O          du             PREPDU  O
aujourd'hui     ADV      B-TIME     président      NMS     O
à               PREPADE  O          Amadou         PERS    B-PERS
Bamako          LOC      B-LOC      Toumani        PERS    I-PERS
Mali            LOC      B-LOC      Touré          PERS    I-PERS
```

4 Adaptation du système statistique à un nouveau domaine

Comme décrit dans Bechet & Charton (2010), le système LIANE a été développé dans le cadre de la campagne d'évaluation ESTER portant sur la transcription et l'annotation en entités nommées d'émissions de radio, principalement des journaux d'information. Les corpus d'apprentissage utilisés sont ceux de la campagne ESTER, constitué d'une centaine d'heures de transcriptions annotées en entités nommées. Le corpus cible de cette étude est un corpus de dépêches AFP. Bien que le contenu thématique de ces deux cadres applicatifs soit proche (des données d'actualité), le type de langue utilisé est différent : transcriptions de l'oral pour le corpus ESTER provenant de sources multiples (France Inter, RFI, Radio Africa, Radio Tele Maroc, etc.). L'étude des résultats obtenus après application du système LIANE au corpus AFP nous a permis de distinguer deux types de problèmes :

– manque de couverture lexicale : les données ESTER contiennent des émissions diffusées en 2007 et début 2008 alors que les dépêches AFP analysées couvrent des périodes plus récentes ;
– mauvaise modélisation des phénomènes spécifiques à l'écrit : le corpus ESTER étant composé de transcription de l'oral, il ne contient aucun « raccourci » spécifique à l'écrit tels que l'emploi d'abréviations (la graphie *M.* par exemple dans la séquence *M. Dupond*) ou l'ajout d'informations complémentaires par l'emploi de caractères de formatage (p.ex. les incises avec des parenthèses telles que : *Le syndicat Force Ouvrière (FO) a annoncé...*).

L'adaptation d'un système statistique à un nouveau domaine nécessite généralement l'annotation manuelle d'un corpus d'adaptation couvrant les nouveaux phénomènes à modéliser. Ce corpus d'adaptation est souvent de taille modeste en raison du coût élevé des annotations en entités nommées. Comme indiqué dans l'introduction, nous avons exploré une voie alternative qui consiste à annoter automatiquement un corpus de très grande taille. Nous avons utilisé pour cela le système NP, spécialisé dans notre domaine cible, celui des dépêches d'agence. L'objectif est d'obtenir sans coût supplémentaire (aucune supervision n'est nécessaire) un système statistique adapté à un domaine particulier à partir d'un système symbolique existant. Ces deux types de systèmes ayant des particularités complémentaires (précision élevée pour le système symbolique, rappel important et robustesse au bruit pour le système statistique), il est intéressant de disposer des deux selon les applications visées.

La procédure d'adaptation non supervisée utilisée ici a consisté tout d'abord à collecter un corpus d'adaptation brut, non annoté, appelé C_A. Nous avons utilisé ici un corpus de dépêches de l'AFP des années 2009 et 2010 constitué de 83M de mots (3M de phrases). Nous avons annoté en entités nommées le corpus C_A avec le système NP. Nous avons ainsi détecté 3,2M d'occurrences d'entités nommées, représentant 168K entités distinctes.

Le corpus C_A a aussi été étiqueté avec l'étiqueteur morphosyntaxique HMM du système LIANE. Nous avons enfin « fusionné » les étiquettes de parties de discours et les étiquettes d'entités nommées comme suit : tout nom propre ou mot inconnu contenu dans une entité nommée de type t détecté par NP reçoit l'étiquette t comme partie de discours. Nous avons obtenu ainsi le corpus C'_A étiqueté à la fois en partie de discours et en entités nommées.

Par exemple, pour la phrase *le commissaire du Portugal Jorge Palmeirim*, le système NP a produit :

```
le commissaire du <EN type="Location" name="Portuguese Republic">Portugal</EN>
<EN type="Person" name="Jorge Palmeirim">Jorge Palmeirim</EN>.
```

De son coté, l'étiqueteur HMM de LIANE a produit l'annotation suivante :

```
(le DET)(commissaire N)(du PREP)(Portugal ORG)(Jorge PERS)(Palmeirim UNK).
```

À partir de ces deux annotations, nous produisons automatiquement le corpus :

```
le              DET     O       Portugal    LOC     B-LOC
commissaire     N       O       Jorge       PERS    B-PERS
du              PREP    O       Palmeirim   PERS    I-PERS
```

Les distributions du modèle HMM ont été directement réestimées sur C'_A, produisant ainsi la version adaptée LIANE 1. Pour l'étiqueteur CRF, le corpus d'apprentissage constitué des corpus ESTER annotés manuellement a été enrichi avec des phrases sélectionnées du corpus C'_A, avec un critère qui repose sur la fréquence des entités détectées : en sélectionnant les phrases contenant les entités les plus fréquentes on espère diminuer le risque d'erreur d'étiquetage dans les données d'apprentissage. Pour éviter que quelques entités ne représentent à elles seules la majorité des exemples retenus, nous avons également limité le nombre maximum de phrases contenant chaque entité. Ainsi nous gardons n phrases parmi toutes celles contenant les entités détectées plus de x fois dans le corpus. En faisant varier n et x on peut mesurer l'impact de l'ajout de données annotées automatiquement dans les corpus d'apprentissage de LIANE. L'utilisation complète du corpus ($n = 400$, $x = 10$) permet de produire le second système adapté, nommé LIANE 2.

Systèmes	**NP**		**LIANE** (sans adapt.)		**LIANE 1** (adapt. lex.)		**LIANE 2** (adapt. lex.+segm.)	
Entité	Precision	Rappel	Precision	Rappel	Precision	Rappel	Precision	Rappel
LOC	**86.4**	**86.3**	81.6	80.7	82.2	81.5	69.8	85.1
ORG	**94.6**	48.1	51.4	55.5	52.6	58.0	53.0	**59.1**
PERS	**92.1**	82.1	81.3	88.4	81.1	91.0	78.3	**93.6**

TABLE 1 – Résultats comparatifs en précision et rappel de NP, LIANE, et des deux adaptations de LIANE réalisées grâce aux données annotées automatiquement par NP

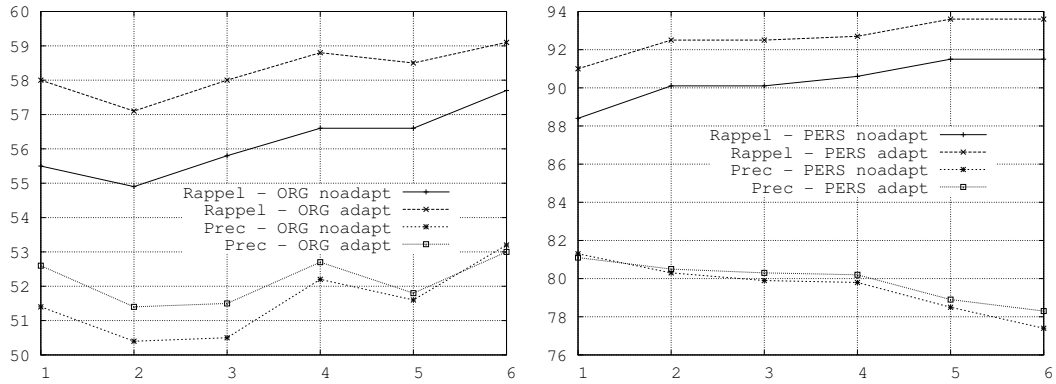

FIGURE 1 – Courbes d'apprentissage de LIANE sur les catégories *PERS* et *ORG* en fonction du volume des données d'apprentissage issues du corpus étiqueté par NP. Les résultats sont donnés avec et sans adaptation lexicale.

5 Résultats

L'évaluation de NP, de la version initiale de LIANE et des versions adaptées décrites à la section précédente a été réalisée sur une nouvelle version du corpus de référence développé et utilisé par Stern & Sagot (2010b), corpus qui est disponible librement dans le cadre de la distribution de SXPipe. Ce corpus, formé de 100 dépêches AFP contenant chacune une moyenne de 300 mots, a été annoté manuellement en entités nommées sous forme de marqueurs XML balisant le texte. Ces balises, outre l'empan de chaque mention, contiennent des informations sur son type ainsi que sur son référent représenté par un numéro d'entrée dans la base Aleda. Dans son état actuel, ce corpus de référence ne contient des annotations que pour des entités de type PERS, LOC et ORG (y compris les noms d'entreprises) [5]. Il contient un total de 1456 mentions d'entités.

Les résultats sont donnés en fonction des mesures de rappel et précision strictes : une hypothèse est considérée comme correcte uniquement si sa segmentation et son typage son corrects [6]. Comme le montre le tableau 1, le système NP donne de bons résultats en terme de précision pour les trois types d'entités. Le rappel sur les LOC est également élevé (86.3%), illustrant ainsi l'apport de la base d'entités Aleda. Comme prévu, le système LIANE obtient une précision plus faible, mais un rappel intéressant pour les entités ORG et PERS. Les résultats de LIANE s'améliorent significativement en adaptant ce système, d'une part au niveau lexical en réestimant les distributions de l'étiqueteur HMM, mais également au niveau segmentation en ajoutant au corpus d'apprentissage des CRF des phrases annotés par NP. Le gain en terme de rappel est de l'ordre de 4% en absolu pour chaque catégorie. Rappelons que cette adaptation est non supervisée et qu'elle a pour but d'obtenir un système statistique complémentaire au système symbolique NP. La figure 1 illustre cette complémentarité en montrant les courbes d'apprentissage du système LIANE sur les catégories ORG et PERS lorsque l'on fait croître le volume de données AFP annotées par NP ajouté au corpus d'apprentissage des CRF [7]. On constate que le rappel augmente en fonction de l'ajout de données, mais au prix d'une certaine diminution en précision. L'augmentation du rappel est encourageante dans la perspective d'intégrer plus encore les deux types de systèmes, par rapport à un cadre applicatif précis, pour exploiter au mieux la bonne précision des modèles symboliques et le bon rappel des modèles numériques.

5. Les annotations de mentions ne comprennent pas les tokens non constitutifs du nom de l'entité lui-même : les titres précédant les noms de personnes notamment en sont exclus (*Mme* ou *Dr*).

6. Les mesures utilisées dans ESTER étaient plus graduelles, donnant un poids différents aux erreurs de frontière, de typage ainsi qu'aux erreurs de reconnaissance dues à la transcription automatique de la parole

7. les 6 points d'abscisse de ces courbes correspondent respectivement à des volumes de 0, 127K, 209K, 498K, 1,2M et 1,9M mots ajoutés aux 1,3M mots du corpus ESTER

6 Conclusion et perspectives

La traditionnelle opposition entre méthodes symboliques et méthodes numériques a nourri quantité de débats à l'occasion de campagnes d'évaluation sur des tâches telles que l'étiquetage en parties de discours ou l'annotation en entités nommées. Si aucune méthode ne s'est avérée gagnante dans toutes les conditions de tests, ces deux familles de méthodes ont montré des comportements différents, et fortement complémentaires, que l'on peut résumer de façon simplificatrice comme suit : bonne précision pour les approches symboliques, bon rappel pour les approches numériques. Nous avons montré dans cette étude comment cette complémentarité pouvait être exploitée pour adapter un système numérique d'annotation en entités nommées à une nouvelle tâche à l'aide d'un système symbolique. Cette adaptation nous permet de disposer de deux systèmes complémentaires, l'un privilégiant la précision, l'autre le rappel, sans nécessiter aucune annotation supplémentaire, illustrant la complémentarité des méthodes numériques et symboliques pour la résolution de tâches linguistiques.

L'interaction NP et LIANE, qui est donc ici une instance simplifiée du paradigme du *co-training*, pourrait toutefois être poussée plus avant. Des expériences en marge de ce travail visant à intégrer un lexique issu d'Aleda aux versions adaptées de LIANE se sont avérées inopérantes, probablement parce que les corpus annotés par NP utilisés pour l'entraînement intégraient déjà une partie importante de ces informations lexicales. Mais un couplage plus fin avec Aleda, notamment en prenant en compte les poids associés aux entités, devrait permettre d'améliorer les résultats de LIANE. À l'inverse, la sortie de LIANE pourrait servir de source d'information au module de NP chargé de la désambiguïsation. Ces pistes, et d'autres, devraient confirmer les résultats présentés ici et valider la pertinence d'approches mixtes ou hybrides pour des tâches telles que la reconnaissance d'entités nommées.

Remerciements

Ce travail a été effectué dans le cadre du projet ANR EDyLex (ANR-09-CORD-008).

Références

BALASURIYA D., RINGLAND N., NOTHMAN J., MURPHY T. & CURRAN J. R. (2009). Named entity recognition in wikipedia. In *People's Web '09 : Proceedings of the 2009 Workshop on The People's Web Meets NLP*, p. 10–18, Suntec, Singapour.

BECHET F. & CHARTON E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

BIKEL D. M., SCHWARTZ R. L. & WEISCHEDEL R. M. (1999). An algorithm that learns what's in a name. *Machine Learning*, **24**(1-3), 211–231.

BROTHWICK A., STERLING J., AGICHTEIN E. & GRISHMAN R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *6th Workshop on Very Large Corpora (ACL '98)*, Montréal.

BÉCHET F. (2011). Named Entity Recognition. In G. TUR & R. DE MORI, Eds., *Spoken Language Understanding : systems for extracting semantic information from speech*, p. 257–290. Wiley.

CHARTON E. & TORRES-MORENO J.-M. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Actes de TALN 2009*, Senlis, France.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The Ester 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *Interspeech 2009*.

MCCALLUM A. & LI W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Seventh Conference on Natural Language Learning (CoNLL)*.

SAGOT B. & BOULLIER P. (2008). SXPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, **49**(2), 155–188.

STERN R. & SAGOT B. (2010a). Détection et résolution d'entités nommées dans des dépêches d'agence. In *Actes de la Conférence TALN 2010*, Montréal, Canada.

STERN R. & SAGOT B. (2010b). Resources for named entity recognition and resolution in news wires. In *Proceedings of LREC 2010 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, La Valette, Malte.