Une approche mixte morpho-syntaxique et statistique pour la reconnaissance d'entités nommées en langue chinoise

Zhen Wang[1,2]
(1) GEOLSemantics, 32, rue Brancion, 75015, Paris
(2) INALCO, ERTIM, 2 rue de Lille, 75343, Paris
zhen.wang@geolsemantics.com

RÉSUMÉ

Cet article présente une approche mixte, morpho-syntaxique et statistique, pour la reconnaissance d'entités nommées en langue chinoise dans un système d'extraction automatique d'information. Le processus se divise principalement en trois étapes : la première génère des noms propres potentiels à l'aide de règles morphologiques ; la deuxième utilise un modèle de langue afin de sélectionner le meilleur résultat ; la troisième effectue la reconnaissance d'entités nommées grâce à une analyse syntaxique locale. Cette dernière permet une reconnaissance automatique d'entités nommées plus pertinente et plus complète.

ABSTRACT

A Mixed Morpho-Syntactic and Statistical Approach to Chinese Named Entity Recognition

This paper presents a morpho-syntactic and statistical approach for Chinese named entity recognition which is a part of an automatic system for information extraction. The process is divided into three steps : first, the generation of possible proper nouns is based on morphological rules; second a language model is used to select the best result, and last, a local syntactic parsing performs the named entity recognition. Syntactic parsing makes named entity recognition more relevant and more complete.

MOTS-CLÉS : Reconnaissance de noms propres, Reconnaissance d'entités nommées, Traitement automatique du chinois, Extraction d'information, Analyse syntaxique

KEYWORDS : Propre noun recogition, Named entity recognition (NER), Chinese Natural Language Processing, Information extraction, Syntactic parsing.

1 Introduction

La reconnaissance d'entités nommées (EN) joue un rôle crucial dans l'extraction de connaissances, les systèmes de question/réponse, la traduction et les résumés automatiques, ainsi que dans l'indexation inter-lingue. Cette tâche a pour objectif de déterminer les frontières d'une entité nommée, et de lui attribuer un type.

La définition d'une « entité nommée » varie selon les systèmes et l'utilisation qui en est faite. La campagne d'évaluation *Automatic Content Extraction* (ACE) en 2007 a étendu la définition traditionnelle des entités (ACE 2007) en ajoutant aux types habituels

(personne, organisme, lieu, expressions numériques) les types véhicule et arme. Dans le programme Quaero (Rosset et al., 2011), la définition des entités nommées a également été étendue en prenant en compte de nouveaux types (tels que les civilisations et les fonctions), et des expressions ne contenant pas de nom propre.

La définition des entités nommées que nous allons employer ici s'approche de celles d'ACE et de Quaero. Nous traitons donc les entités nommées de type : personne, organisme, lieu, expressions numériques (dates, heures, quantités, mesures, nombres), avec ou sans nom propre. Dans cet article, nous nous concentrons sur la présentation du traitement des entités nommées des type personne, organisme et lieu, en chinois, qui contiennent au moins un nom propre.

Cet article propose une approche mixte morpho-syntaxique et statistique pour la reconnaissance d'entités nommées en langue chinoise. Les méthodes existantes utilisées pour la reconnaissance d'EN en langue chinoise seront résumées dans la section 2. Les composants de notre approche seront exposés en section 3. Puis, nous présenterons dans la section 4 l'évaluation de notre système sur un corpus. Enfin, nous énoncerons des perspectives d'amélioration dans la section 5.

2 État de l'art

Les particularités de la langue chinoise rendent la reconnaissance d'EN plus difficile que pour les langues occidentales. En effet, l'un des premiers problèmes rencontrés est l'absence de séparateurs de mots. Par exemple, la phrase « Je vais à Paris. » se traduit ainsi en chinois « 我去巴黎。 ». Une étape de segmentation en mots est effectuée avant la reconnaissance d'EN. Cependant, la segmentation est très souvent ambiguë, ce qui multiplie les hypothèses de découpage. Par exemple, une suite de caractères comme « 的确切 » peut être découpée en « 的确/切 » (sûrement / couper) ou « 的/确切 ». (de / exactitude). Ceci rend difficile la détermination des frontières des mots, notamment dans le cas de mots inconnus, qui sont le plus souvent des noms propres (Sun et al., 2009). À cela, on peut ajouter le manque d'information typographique. Par exemple, il n'existe pas de différenciation entre majuscule et minuscule, qui constitue un critère de reconnaissance efficace pour les noms propres en français.

Les premiers travaux sur la reconnaissance d'EN en langue chinoise ont débuté au début des années 1990 (Sun et al., 2010). Les méthodes basées sur des règles (Wang et al., 1992) ont autant été employées que les méthodes statistiques (Sproat et al., 1990), telles que le modèle de Markov Caché (Liu et al., 2005, Wang et al., 2012) et les champs aléatoires conditionnels (*Conditional Random Fields*) (Chen et al., 2006, Mao et al., 2008). Les deux types de méthode ont chacun des avantages ainsi que des inconvénients. En effet, les méthodes basées sur des règles s'appuient sur des dictionnaires ou/et des règles linguistiques élaborés par des experts de la langue. Elles permettent d'obtenir une bonne précision pour certains cas, mais le processus de construction est long. De plus, il est difficile d'inclure tous les cas linguistiques existants. De ce fait, la portabilité de ces méthodes est faible. Les méthodes statistiques, quant à elles, s'appuient sur un corpus d'apprentissage. La construction d'un système est rapide mais la pertinence de ces méthodes dépend de la taille et du contenu du corpus d'apprentissage. D'autres recherches (Chen et al., 2000, Cao et al., 2002) se fondent sur une approche qui combine

les deux types de méthodes afin de profiter des avantages et de pallier les inconvénients de chaque approche. Nous nous plaçons dans cette dernière catégorie.

3 Notre approche

Notre module de reconnaissance d'entités nommées est basé sur un système d'automates à états finis pondérés. Ces automates effectuent l'analyse morphologique et l'analyse syntaxique des textes. Un automate particulier sert à la désambiguïsation. Il utilise un corpus étiqueté afin de créer un modèle de langue.

Nous utilisons les notions d'annonceur (McDonald , 1993) et de déclencheur , très importantes pour la reconnaissance des noms propres et des EN. En effet, le déclencheur peut être un mot ou une catégorie, qui permet de provoquer la détection d'un nom propre. Les déclencheurs doivent pouvoir être définis par une liste finie de mots ou catégories afin de pouvoir être intégrés dans les règles. Par exemple, nous avons désigné les noms de famille, qui représentent environ 131 mots pour les noms de personne chinois, comme mots déclencheurs pour la détection de prénoms potentiels. Quant aux annonceurs de nom propre, ce sont des mots qui sont suivis ou précédés par un nom propre faisant partie d'une entité nommée. Ces annonceurs de nom propre sont souvent utilisés comme annonceurs d'entité nommée. Il peut s'agir de mots qui désignent un métier, le titre d'une personne, un type de lieu, d'organisation ou de produit. Par exemple, le nom de famille «WANG» est le mot déclencheur permettant la reconnaissance du prénom «Zhen», et «Mademoiselle» peut être un annonceur de nom propre qui se situe au début d'une entité nommée de type personne comme «Mademoiselle Wang Zhen». Que ce soit le nom de famille «WANG» ou l'annonceur «Mademoiselle», l'un ou l'autre peut être utilisé comme mot déclencheur pour la reconnaissance de l'entité nommée «Mademoiselle Wang Zhen».

Dans notre corpus d'apprentissage, seuls les annonceurs immédiatement suivis ou précédés d'un nom propre ont été étiquetés avec la catégorie « annp », afin de faciliter la désambiguïsation de la phrase. Nous attribuons éventuellement une propriété à la catégorie « annp » afin de distinguer sa position par rapport au nom propre, il s'agit d' « antérieur » ou « postérieur ». Ceci permet aussi de mieux repérer les noms propres potentiels. Les autres potentiels annonceurs sont étiquetés comme tels juste avant l'analyse syntaxique, une fois que la désambiguïsation a été effectuée grâce aux catégories positionnelles. Ils vont permettre de détecter des structures syntaxiques particulières, de reconnaître des EN aux structures plus complexes et éventuellement de déterminer les types des EN. Par exemple, dans l'entité nommée « 胡主席 » (le président Hu), « 胡 » (hu) est étiqueté comme nom propre et « 主席 » (président) comme annonceur postérieur. En revanche, dans l'entité nommée « 中国石油大学 » (L'Université du pétrole de Chine) , puisque le nom propre « 中国 » (Chine) est suivi immédiatement par le nom « 石油 » (pétrole) mais pas par l'annonceur « 大学 » (Université), celui-ci n'est pas étiqueté comme annonceur dans la meilleure hypothèse de catégorisation après désambiguïsation (voir plus bas), mais le sera lors de l'analyse syntaxique.

3.1 Architecture

Notre procédure (figure 1) débute par une étape de tokenization qui permet de

reconnaître les caractères latins, et de les étiqueter de façon simple (seulement deux catégories sont utilisées). Ensuite, une segmentation en mots est effectuée à l'aide de dictionnaires. Elle est suivie par une étape de reconnaissance et de normalisation d'expressions numériques. Plusieurs hypothèses de segmentation du texte sont produites à l'issue de cette étape. Dans le texte, chaque mot est associé à une ou plusieurs catégories.

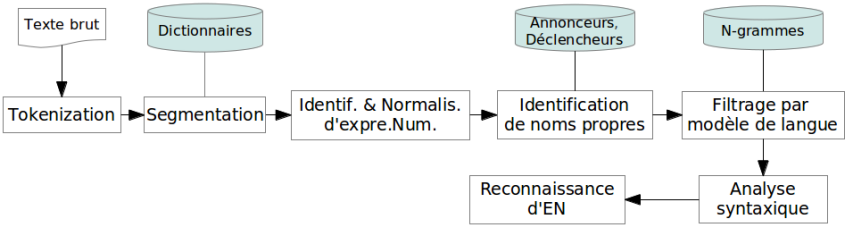

FIGURE 1: Procédure de reconnaissance d'entités nommées

Puis, l'identification de noms propres potentiels à l'aide de règles linguistiques et de ressources linguistiques (liste d'annonceurs, etc.) est réalisée. Celle-ci génère plusieurs hypothèses de segmentation et de catégorisation supplémentaires. Ensuite, un modèle de langue permet de désambiguïser et de sélectionner la meilleure segmentation. Enfin, la reconnaissance et le typage d'entités nommées est réalisée juste après l'analyse syntaxique.

3.2 Automates morphologiques

La construction morpho-syntaxique des noms propres en chinois est très variée. En effet, un nom propre peut être composé de n'importe quels caractères et a une longueur variable. Afin de tenir compte de ces phénomènes, lors du découpage et de la consultation des dictionnaires, nous ajoutons aux hypothèses d'interprétation des mots, des noms propres potentiels, obtenus grâce à des règles linguistiques.

La langue chinoise a peu de formes fléchies. Mais certaines catégories peuvent aider à exclure des mots qui ne peuvent pas entrer dans la composition des noms propres, tels que certaines dates, les interjections, etc. De ce fait, nous avons choisi de procéder à la reconnaissance des noms propres après l'étape de segmentation et de reconnaissance d'expressions numériques.

Les automates morphologiques (Eilenberg, 1974) sont des transducteurs qui contiennent des opérations telles que la composition et la concaténation. Ils permettent de traiter à la fois les caractéristiques communes à tous les noms propres, mais aussi leurs différences. Par exemple, en chinois, les noms propres de personnes chinois ont des particularités par rapport aux noms propres étrangers. C'est pour cette raison que nous avons décidé de traiter les noms propres de personnes chinois différemment. Par ailleurs, les noms propres transcrits de type personne, organisation ou produit ont des caractéristiques

communes. À ce stade, nous les traitons ensemble sans différencier leur type. De plus, nous avons pris en compte la relation entre un nom propre et son annonceur. C'est un autre critère sur lequel les automates morphologiques se basent pour identifier les noms propres.

Par conséquent, nous avons divisé la reconnaissance de noms propres en deux phases : l'identification des noms propres de personnes chinois, et l'identification des noms propres étrangers (personnes, lieux, organismes ou de type inconnu).

Notons que, dans cet article, nous n'avons pas décrit l'identification de noms propres en écriture latine, puisque les expressions en écriture latine sont traitées avec une méthode relativement simple, basée sur la reconnaissance typographique. Par exemple, une suite de caractères en écriture latine contenant une arobase « @ » est identifiée comme un nom propre, ainsi que celles commençant par un majuscule.

3.2.1 Noms de personne chinois

Les noms de personne chinois ont des particularités morphologiques. Tout d'abord, un nom de personne chinois ne possède qu'entre 2 et 4 caractères : il peut contenir un nom de famille de 1 à 2 caractères et un prénom de 1 à 2 caractères. Ensuite, le nom de famille précède toujours le prénom. Enfin, le nombre de noms de famille en chinois est limité, et les cent noms de famille les plus fréquents concernent 87% de la population chinoise (De La Robertie, 2005). Cependant, l'ambiguïté de segmentation entraîne souvent la confusion entre nom de famille et prénom.

Après avoir analysé les noms de personne chinois mal reconnus, nous pouvons lister les causes suivantes : 1) un caractère du prénom est inconnu ; 2) le deuxième caractère du prénom est ambigu avec un autre mot : « 薄/熙来/自 » (Bo / Xilai / à partir de) ou « 薄/熙/来自 » (Bo / Xi / vient de); 3) le premier caractère du prénom est un annonceur spécial : « 刘/庄 » (Liu / Zhuang) est le nom d'une personne ou « 刘/庄 » (Liu / Village) est le nom d'un village dont le nom est Liu, ce qui pose aussi une difficulté pour le typage des entités nommées; 4) le prénom ou le nom de famille font partie d'un nom commun, comme par exemple dans 2) ; 5) le prénom est un nom commun : « 李/建国 » (Li / Jiangguo ou Li / fonder le pays) ; 6) le nom de personne est un nom commun : « 汪/洋 » (Wan / Yang) ou « 汪 洋 » (Océan). Par conséquent, la reconnaissance de noms de personne chinois revient le plus souvent à une résolution des ambiguïtés de segmentation.

Notre automate utilise les noms de famille comme déclencheurs pour détecter des prénoms potentiels. Ces prénoms sont sélectionnés par leur catégorie et par leur nombre de caractères. Par exemple, nous pouvons exclure des prénoms potentiels les expressions numériques, notamment les dates ainsi que les expressions quantitatives, et les mots qui ont plus de 2 caractères.

La détection de noms propres potentiels s'effectue après la consultation du dictionnaire, lorsque la phrase a déjà plusieurs hypothèses de segmentation et d'étiquetage. Les noms propres potentiels sont identifiés à l'aide des règles de construction de noms propres. Par exemple, pour la phrase « 王 珍 去 北 京 » (Wang Zhen va à Beijing), nous obtenons plusieurs hypothèses de catégorisation après la consultation de dictionnaire(voir figure 2). En utilisant le nom de famille « 王 » (Wang) comme déclencheur, des possibilités ont

été ajoutées après la détection de noms propres (voir figure 3).

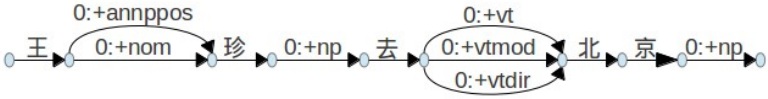

FIGURE 2: Transducteur de l'exemple "Wang Zhen va à Beijing" après la consultation de dictionnaire

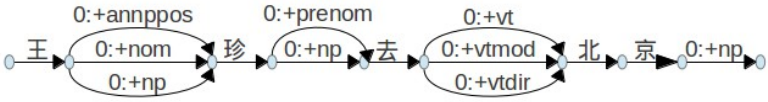

FIGURE 3:Transducteur de l'exemple "Wang Zhen va à Beijing" après la détection de nom de personne

Les exemples montrent que les automates morphologiques génèrent d'avantage d'hypothèses de segmentation et de catégorisation pour une phrase donnée. Ces hypothèses sont par ailleurs pondérées grâce aux règles linguistiques ce qui permettra d'effectuer une désambiguïsation à l'aide d'un modèle de langue, comme présenté dans la section 3.3.

3.2.2 Noms propres transcrits

À l'inverse des noms de personne chinois, les noms propres étrangers, notamment européens, sont transcrits en écriture chinoise d'après la phonétique du nom d'origine. Des caractères spécifiques sont utilisés pour la composition de ces noms propres. Les caractères utilisés sont peu présents dans la composition de mots communs par rapport aux autres caractères de même prononciation. Ceci réduit la possibilité de former un mot commun à l'intérieur d'un nom propre transcrit. Autrement dit, après notre segmentation, un nom propre étranger est séparé en plusieurs caractères individuels. Par exemple, la transcription du nom de personne « Nicolas Sarkozy » en écriture chinoise est « 尼古拉·萨科齐 », et le résultat de sa segmentation est « 尼/古/拉/·/萨/科/齐 », sans ambiguïté. La reconnaissance de ces noms propres transcrits consiste donc à regrouper les caractères qui les composent. De plus, les caractères utilisés dans la transcription appartiennent généralement à un ensemble limité.

Notre automate utilise une liste de caractères de transcription comme déclencheur afin de détecter des noms propres étrangers potentiels. Prenons l'exemple de « Nicolas Sarkozy », les caractères de sa translittération en écriture chinoise sont dans la liste. Notre automate les repère, puis les regroupe en deux noms propres potentiels, « 尼古拉 » et « 萨科齐 ». Grâce à la fonction conventionnelle du point de séparation « · », le prénom et le nom de famille seront identifiés.

Dans le cas où un nom propre transcrit ne contient que des caractères hors liste qui ont été étiquetés en +unk[1], nous avons mis en place un automate qui permet de prendre en compte ces caractères, le nombre de caractères, ainsi que le contexte de ce nom propre potentiel. Quand une chaîne de caractères n'est composée que de caractères transcrits, ceux-ci sont regroupés en un mot et nous considérons que c'est un nom propre. En observant dans un dictionnaire de noms de personnes du monde (Xinhua News Agency, 1993, 650 milles noms propres transcrits), nous avons constaté que le nombre de caractères du nom propre transcrit dépasse rarement 8. Cette caractéristique a été appliquée dans les règles. Le nom propre peut être précédé ou suivi par un annonceur. Le déclencheur de ce type de nom propre est souvent un caractère inconnu non latin du dictionnaire utilisé. Par exemple, l'un des résultats de l'analyse de « 巴拉圭 » (Paraguay) est « 巴+unk/ 拉+vt /圭+unk». Puisque « 拉 » est un caractère qui peut entrer dans la composition d'un nom propre transcrit, et que les caractères à sa gauche ainsi qu'à sa droite sont des mots inconnus, nous pouvons les regrouper pour former un nom propre potentiel.

3.3 Automates statistiques

Les automates statistiques se composent de dictionnaires pondérés, de règles linguistiques pondérées, et d'un modèle de langue. Ils permettent d'attribuer et de calculer la probabilité d'une suite de catégories pour une phrase entrée. L'attribution de la pondération est effectuée au moment de l'application des automates de segmentation, de reconnaissance de noms propres et de lissage. Quant aux noms propres, les pondérations sont attribuées à l'aide de la fréquence du nom de famille dans le corpus d'apprentissage, et à l'aide de règles linguistiques. Par exemple, le caractère « 王 » individuel a plus de chances d'être un nom de famille (Wang) qu'un nom commun (roi).

Le dictionnaire utilisé pour la segmentation (voir un extrait figure 4) contient 119 859 couples (mot, catégorie), dont 4 526 mots de catégorie « nom propre ». Un couple (mot, catégorie) peut être associé à un poids. Les couples non présents ou présents une seule fois dans le corpus sont sans pondération. Comme montré sur la figure 4, la première colonne est un mot en écriture chinoise, la deuxième colonne est la catégorie grammaticale du mot, la troisième colonne est le poids associé à ce couple (mot, catégorie). Ce poids représente le nombre d'occurrences du couple (mot, catégorie) dans le corpus d'apprentissage. Il est calculé par la formule suivante : $poids(w_i, cat_j) = -\log (\Sigma(w_i, cat_j))$, où $\Sigma(w_i, cat_j)$ désigne le nombre d'occurrences du mot w_i avec la catégorie cat_j dans le corpus d'apprentissage.

[1]+unk : caractère/mot inconnu

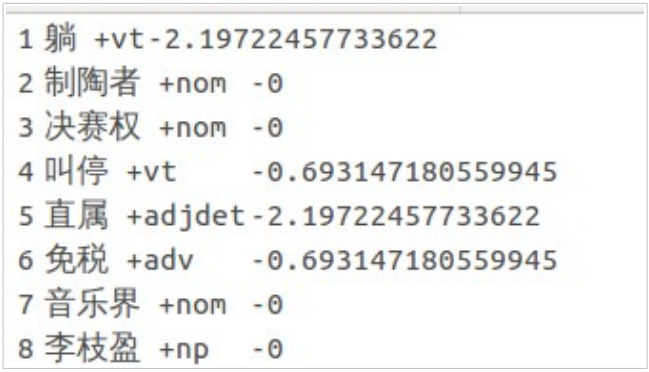

FIGURE 4: – Exemples de couples (mot, catégorie) pondérés.

Le modèle de langue utilise des trigrammes de catégories morphosyntaxiques. Par exemple, « annppos » est la catégorie d'annonceur de nom propre postérieur. Ils sont établis à partir du corpus étiqueté du LDC (Chinese Treebank 6.0, 2007), et permettent d'attribuer une pondération aux séquences de catégories (figure 5), afin de calculer la catégorie la plus probable d'un mot en contexte. Par exemple, un nom propre est plus probable après un verbe qu'après un adverbe, après un annonceur qu'après un adjectif. Nous avons modifié le corpus LDC avec un jeu de catégories morphosyntaxiques défini par notre équipe afin d'optimiser la désambiguïsation. Par exemple, la catégorie « +vtmod » qui désigne le verbe de modalité a été ajoutée afin de mieux désambiguïser le nom et le verbe.

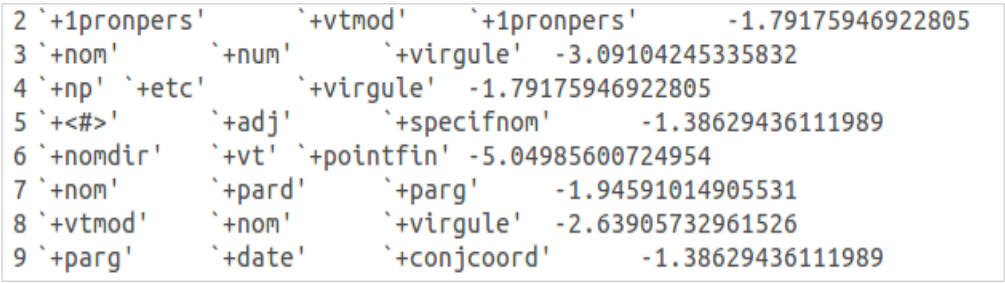

FIGURE 5 – Exemples des trigrammes de catégories.

Le modèle de langue est appliqué sur les textes étiquetés. Les probabilités des différents chemins possibles sont alors calculées afin de résoudre les ambiguïtés de segmentation et de catégorisation. Après l'application du modèle de langue, nous ne conservons que la solution la plus probable.

3.4 Automates syntaxiques et reconnaissance d'EN

Les automates syntaxiques établissent des relations de dépendance typées entre les mots, en s'appuyant surtout sur leur catégorie et sur leurs propriétés. Ils mettent en évidence les liens (attribut, appositif ou quantitatif) entre les mots au sein des groupes nominaux, permettant ensuite d'identifier une entité nommée, même lorsque son annonceur est éloigné du nom propre. Nous avons spécifié quelque types de relations afin de mieux repérer les entités nommées. Par exemple, ENpers désigne les relations entre un nom de

famille et un prénom, ENrel entre un annonceur et un nom propre, ATT désigne les relations de type attribut entre un annonceur et un nom ou un adjectif, QUN désigne les relations de type quantitatif entre un chiffre et un nom comme 3人(3 personnes).

Une fois que les relations syntaxiques au sein des groupes nominaux ont été repérées, la dernière phase de la reconnaissance des entités nommées commence. Elle consiste à regrouper les relations syntaxiques (ENrel, ENpers, ATT, etc) qui composent l'entité nommée, et à typer celle-ci. Un automate repère dans un premier temps, un nom propre ou un annonceur dans la phrase donnée. Un autre automate parcourt toutes les relations qui ont comme tête ce nom propre ou cet annonceur en prenant en compte les types des relations et la position du nom propre ou de l'annonceur dans celles-ci. Nous les récupérons jusqu'à rencontrer un nom propre, un annonceur, ou bien une frontière de groupe nominal. Certains types de relations ne peuvent pas faire partie d'une entité nommée, comme les relations entre un sujet et un verbe, mais aussi les relations entre un prénom et une date, ce qui est aussi une condition d'arrêt du parcours.

La figure 6 présente un exemple d'analyse syntaxique de la phrase « 国家主席习近平同志访问法国。 » (Le président de l'État, le camarade Xi Jinping, visite la France). L'entité nommée reconnue dans cette phrase est « le président de l'État, le camarade Xi Jinping » qui inclut tous les composants liés par les relations de type ENpers et ceux liés par la relation ATT, liée au mot « Président ».

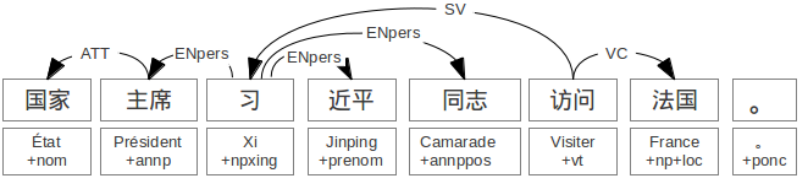

FIGURE 6 – Analyse syntaxique (exemple 1)

En revanche, dans la phrase « 摄影师拍摄阿拉伯海美景。 » (Le photographe photographie le beau paysage de la Mer d'Arabie)(voir la figure 7) , l'entité nommée reconnue est « Mer d'Arabie », mais n'inclut pas la relation entre l'annonceur postérieur « Mer » et le nom « beau paysage », car le nom « beau paysage » gouverne « Mer », mais n'est pas un annonceur.

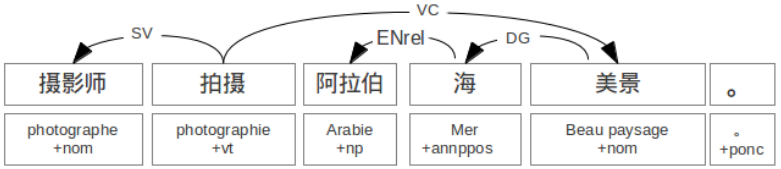

FIGURE 7 – Analyse syntaxique (exemple 2)

Le typage des entités nommées s'effectue après le regroupement de leurs composants. Le type de l'entité nommée est déterminé grâce au type de l'annonceur, qui est lui-même

déjà déterminé dans les dictionnaires. Par ailleurs, le type des entités nommées sans annonceur est déterminé grâce à la présence d'un nom de famille, et le type des entités nommées sans annonceur, ni nom de famille reçoit un type spécial « unk » (inconnu). Lorsque l'entité nommée contient plusieurs annonceurs, c'est le type du dernier qui détermine le type de cette entité nommée. Par exemple, l'entité nommée « 北京(Beijing +np+loc)市(Ville +annppos+loc[2])人民(peuple +nom)政府(gouvernement +annppos+org[3]) 网(net +annppos+prod[4]) » (le site web du gouvernement populaire de la ville de Beijing), est typée comme « produit », comme nous l'indique le type du dernier annonceur, 网(net +annppos +prod). Dans le cas d'entités nommées composées d'une seule unité comme « 北 京 »(Beijing), le type peut être indiqué directement dans le dictionnaire. Enfin, certaines entités nommées peuvent ne pas être typées à ce niveau, faute d'informations sémantiques. Nous utilisons alors le type spécial, « unk »(inconnu).

4 Évaluation

4.1 Corpus et calcul des scores

Pendant la deuxième campagne internationale SIGHAN[5] (*Second International Chinese Word Segmentation Bakeoff*), un corpus élaboré par l'Université de Pékin a été utilisé pour l'évaluation de la segmentation en chinois simplifié. Ce corpus est composé d'articles variés issus de journaux d'information, et est en chinois simplifié, la langue visée par notre traitement. Nous avons décidé d'annoter manuellement les entités nommées au sein d'un extrait[6] segmenté de ce corpus pour une première évaluation. 3 215 entités nommées ont été annotées, dont 1 055 de type personne, 1 565 de type lieu et 595 de type organisme.

Pour évaluer notre système, nous avons choisi d'utiliser le *slot error rate (SER)* (Makhoul et al., 1999) qui permet de mieux identifier les types d'erreurs. La formule est : SER = (I + D + TF + 0.5 * (T + F))/R ; un nombre plus petit indique une performance meilleure. La formule prend en compte :

- Insertion (I) : EN dans le test mais pas dans la référence ;
- Suppression (D) : EN dans la référence mais pas dans le test ;
- Type et frontière (TF) : EN dans le test avec erreurs de type et de frontière ;
- Type (T) : EN dans le test avec erreur de type ;
- Frontière (F): EN dans le test avec erreur de frontière ;
- Référence (R) : EN présentes dans la référence.

Par exemple, pour «国家主席习近平同志访问法国。» (Le président de l'État, le camarade Xi Jinping, visite la France), la référence est <pers>国家主席习近平同志</pers> 访问 <loc>法国</loc> où deux entités nommées ont été annotées. Supposons que nous ayons

[2]+loc : une propriété d'un mot qui signifie que le mot a un trait sémantique «lieu»
[3]+org : une propriété d'un mot qui signifie que le mot a un trait sémantique « organisme »
[4]+prod : une propriété d'un mot qui signifie que le mot a un trait sémantique « produit »
[5]Site web du SIGHAN : http://sighan.cs.uchicago.edu/
[6]Cet extrait du corpus annoté est disponible sous conditions d'utilisation.

les deux hypothèses suivantes :

1. 国家主席 <pers>习近平同志</pers> 访问 <loc>法国</loc>.
2. 国家主席 <pers>习近平同志</pers> 访问 <org>法国</org>.

L'hypothèse 1 contient une erreur de frontière pour l'EN de personne, donc le SER est 0.5*1/2=0,25. L'hypothèse 2 contient une erreur de frontière pour l'EN de personne et une erreur de type pour l'EN de lieu, donc le SER est (0.5*(1+1))/2=0,50.

4.2 Résultats et observations

Nous avons appliqué notre système sur cet extrait de corpus et obtenu un SER de 0,2589, 2 633 sur 3 215 entités nommées ont été correctement identifiées et typées. Le nombre d'erreurs par type est présenté dans le tableau ci-dessous (Table 1). ENpers désigne une entité nommée de type personne, ENloc est de type lieu, ENorg de type organisme et ENunk est le type inconnu.

	ENpers	ENloc	ENorg	ENunk	Total
Suppression	141	88	56	-----	285
Insertion	125	88	19	104	336
Type	5	35	3	0	43
Frontière	85	39	4	-----	128
Type + Frontière	25	39	62	0	126
Total	381	289	144	104	918
Total d'EN correctes	799	1 364	470	-----	2 633

TABLE 1 – Évaluation SER en chiffres

Certaines entités nommées de type personne (ENpers) ne sont par repérées à cause de l'absence de certains noms de famille qu'elles contiennent dans la liste des déclencheurs. Ces noms de famille ne sont pas intégrés dans la liste des déclencheurs à cause de leur faible présence dans le corpus, et de leur faible usage dans la vie courante. L'intégration de ces noms de famille doit être effectuée avec beaucoup de précautions. Elle nécessite plus de temps. Certaines autres entités nommées, telles que les surnoms, les noms de personne japonais, les abréviations d'organisme et des noms propres de lieux n'ont pas été identifiés à cause de leur structure morphosyntaxique très particulière.

Les entités nommées inconnues contiennent principalement des chiffres et des expressions en écriture latine. Ces chiffres sont écrits de manière inhabituelle. Par exemple : dans « 19·1% » (normalement écrit 19.1%), le point de séparation « · » est considéré dans nos règles comme un composant important pour identifier des entités

nommées. Dans notre système, une chaîne de caractères en écriture latine commençant par une majuscule est identifiée comme un nom propre. Certains de ces noms propres sont cependant des entités nommées de produit qui ne sont pas annotées dans le corpus de référence actuel.

Les entités nommées incorrectes détectées montrent des détections erronées de noms propres. Nous allons affiner l'utilisation de listes de déclencheurs en vue d'éviter cette détection erronée. Pour ce faire, la structure phonétique des noms propres transcrits ainsi que la structure interne des noms propres seront étudiées, afin d'intégrer nos nouvelles observations aux règles. Les erreurs de frontières et de type+frontière soulignent le problème de segmentation. Les entités nommées effacées et les erreurs de typage et de type+frontière révèlent des problèmes de catégorisation. Les entités nommées supprimées sont celles qui ne contiennent pas de nom propre ni d'annonceur, et qui ne sont pas repérables par un déclencheur, telles que des noms d'organisations ainsi que des produits.

5 Perspectives

Le résultat de l'évaluation est encourageant, 82 % des entités nommées ont été bien identifiées. Par la suite, nous allons également affiner la reconnaissance des entités nommées afin de traiter les entités nommées imbriquées les unes dans les autres. En effet, pour une exploitation d'entités nommées en vue de l'extraction d'information, nous souhaitons, pour certains cas, extraire plusieurs entités nommées à partir d'une même chaîne de caractères, afin de conserver des informations importantes. Par exemple, pour « 比利时(Belgique +np +loc)使馆(Ambassade +annppos +org) », l'entité nommée doit être l'Ambassade de Belgique, mais dans « 巴黎(Paris +np +loc)歌剧院(Opéra +annppos +org) », il est préférable d'extraire à la fois « Opéra de Paris » et « Paris », étant donné que « Paris » est le lieu où se trouve l'Opéra, contrairement à « Belgique » qui n'est pas le lieu où se trouve l'ambassade, et devrait être extrait et typé de manière différente, car « Belgique » dans « l'ambassade de Belgique » a davantage le sens d'organisation que de lieu.

Remerciements

Nous tenons à remercier l'Agence Nationale de la Recherche, projet portant la référence ANR-09-CSOSG-08-01, pour l'aide qu'elle nous a apportée pour mener à bien ce travail.

Références

ACE 2007. *The Automatic Content Extraction 2007 Evaluation Plan, Evaluation of the Detection and Recognition of ACE Entities, Value, Temporal Expressions, Relations, and Events*

Cao Wenjie, Zong Chengqing. *Chinese person name identification based on rules and statistics*. In : The 3th International Symposium on chinese spoken language processing, 2002

Chen Keh-Jiann and Chert Chao-jan. Knowledge Extraction for Identification of Chinese

Organization Names. In : *Second Chinese Language Processing Workshop, Association for Computational Linguistics,* 2000, pages 15-21

Chen W., Zhang Y., & Isahara H.. Chinese Named Entity Recognition with Conditional Random Fields. In : *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Association for Computational Linguistics,* 2006, pages 118-121

De La Robertie Pierre. *Le nom propre en chinois*. *Essai de morphosyntaxe*. In : CORELA – Numéros thématiques le traitement lexicographique des noms propres. *Publié en ligne le 02 décembre 2005*

Eilenberg Samuel. 1974 *Automata, Languages, and Machines*. Volume A. Academic Press, San Diego.

Liu, F., Zhao, J., Lv, B., Bo, X., & Yu, H. Product Named Entity Recognition Based on Hierarchical Hidden Morkov Model. In : *Proceedings of the Forth SIGHAN Workshop on Chinese Language Processing*, 2005

Makhoul J., Kubala F., Schwartz, R. et Weischedel R. (1999). Performance measures for information extraction. In *Darpa broadcast news workshop*.

Mao Xinnian, Dong Yuan, He Saike, Wang Haila, Bao Sencheng. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields. In *Sixth SIGHAN Workshop of Chinese Language Processing, 2008*

McDonald David D. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In *Corpus Processing for Lexical Acquisition, 1993, pages 61-67*

Rosset Sophie, Grouin Cyril, Zweigenbaum Pierre. *Entités nommées structurées : guide d'annotation Quaero*. Notes et Documents LIMSI N° : 2011-04, Septembre, 2011

Sproat, Richard. and Shih, Chilin. A statistic method for finding word boundaries in Chinese text. In : *Computer Processing of Chinese & Oriental Languages,* 1990*, Vol 4*, pages 336-351

Sun Tieli, Liu Yanji, 中文分词技术的研究现状与困难 [The State of the art and difficulties in Automatic Chinese word segmentation]. In 信息技术 *[Information Technology], 2009, Vol.7* [en chinois]

Sun Zhen, Wang Huilin, 命名实体识别研究进展综述 [Overview on the advance of the research on named entity recognition]. In 现代图书情报技术 *[New technology of library and information service]*, 2010, No.7, pages 42-47 [en chinois]

Propre Names And Translation Service, Xinhua News Agency. *Names of the world's peoples – A comprehensive dictionary of names in roman-chinese.* Published by Chian Translation & Publishing Corporation, 1993

Wang, L.-J.; Li, W.-C. & Chang, C.-H.. Recognizing Unregistered Names for Mandarin Word Identification. In : *Proceedings of 14th COLING,* 1992, pages 1239-1243

Wang Longyue, Li Shuoo, Wong Derek F., Chao Lidia, A Joint Chinese Named Entity Recognition and Disambiguation System. In *The second CIPS-SIGHAN joint Conference on Chinese Language Processing, 2012*

