Institut National des Langues et des Civilisations Orientales (INaLCO)

Département Textes, Informatique, Multilinguisme

La reconnaissance des entités nommées des personnes dans un corpus chinois

MASTER TRAITEMENT AUTOMATIQUE DES LANGUES

SPÉCIALISÉ INGÉNIEGIE MULTILINGUE

Li Yun YAN

Directeurs de recherche :

Cyril GROUIN

Pierre ZWEIGENBAUM

Encadrant d'entreprise :

Bianka BUSCHBECK

Présenté et soutenu en

novembre 2014

Remerciements

Je tiens à remercier dans un premier temps, toute l'équipe pédagogique de TAL tout au long de mon master pour leur professionnalité, expériences et patiences.

Je remercie tout particulièrement à mon directeur de mémoire, Cyril Grouin, pour tous ses conseils, sa disponibilité pendant la réalisation du mémoire.

Je remercie également à mon co-directeur de mémoire, Pierre Zweigenbaum, dont ses conseils au début m'ont beaucoup inspiré.

Je tiens également à remercie mon encadrant de stage, Bianka Buschbeck, pour m'avoir intégré rapidement au SYSTRAN et pour son expérience enrichissante et plein d'intérêts qu'elle m'a fait vivre durant ces six mois au sein de l'entreprise.

J'exprime mes profonds remerciements à mes collègues de stages pour toute l'aide qu'ils m'ont apporté tous les jours.

Table des matières

Liste des figures

Liste des tableaux

Introduction

1. Définition de REN

La reconnaissance des entités nommées (ci-après REN) est d'abord considérée comme une reconnaissance des noms propres [Coates-Stephens, 1992]. Pendant MUC-6 (New York University, 1995), pour la première fois, les entités nommées (ci-après ENs) sont catégorisées en 3 grandes classes : « Entity Name », « Temporal Expressions » et « Number Expressions » dont « Entity Name » se compose de personne, organisation et lieu. Ces 3 grandes catégories sont aussi nommées respectivement « ENAMEX », « TIMEX » et « NUMEX » selon l'évaluation 863NE en 2004. Par la suite, ces 3 catégories principales sont détaillées et élargies avec les ajouts des sous-catégories. Les villes font partie d'une sous-catégorie de « lieu» [Fleischman, 2001; Lee et al., 2005], et les hommes politiques font partie de « person » [Fleischman, 2001; Lee et al., 2005]. En outre, les ENs ne se contentent plus d'être un nom de personne, un lieu ou une suite de chiffres. Car aux besoins de recherches variées, les entités spécifiques sont proposées pour un domaine précis dans le but d'observer « les points de recouvrement » [Dutrey et al., 2012]. Par exemple, la reconnaissance des entités biologiques nommées aide davantage à relever les interactions entre gènes [Charnois et al., 2009]. L'extraction des entités spécifiques du domaine géographique est une étape essentielle pour extraire les informations géographiques [Gaio et al., 2012].

2. Difficultés générales et particulières en chinois

Le REN consiste à repérer des unités significatives qui appartiennent à des catégories prédéfinies [Sun et al., 2010]. Ce travail consiste à résoudre deux problèmes particuliers: premièrement le repérage des portions significatives, délimitées par des frontières, et deuxièmement l'attribution de la bonne catégorie à la portion précédemment repérée. Pour trouver les solutions à ces problèmes, les difficultés générales de REN se divisent en 3 types [Grouin et al., 2011]:

I. Choisir une bonne catégorie pour une EN en cas d'ambiguïté : « Washington » est à la fois un nom de personne et un lieu.

II. Détecter les frontières de l'EN : une entité est au milieu d'une autre entité plus longue, comme Ambassade de Chine en France dont « Chine » et « France » sont des lieux

III. Métonymies d'annotation : « Madrid » avait été considéré comme nom de ville. Aujourd'hui il est également utilisé pour désigner l'équipe de football « Real Madrid ».

Avec le développement de l'internet et de l'interaction multiculturelle, le phénomène du néologisme et de l'emprunt augmente le pourcentage du recouvrement des « mots inconnus ». Cela rend la REN plus difficile.

La REN en chinois comprend toutes les difficultés mentionnées ci-dessus. En outre, distingué de l'anglais ou d'autres langues latines, le chinois ne contient pas des marqueurs morphologiques évidents. Par exemple, il ne possède pas un espace comme un marqueur naturel de segmentation, et il n'est pas possible de reconnaître les noms propres par la lettre initiale en majuscule. Par conséquent, il faut d'abord définir les frontières des mots avant de définir si le(les) mot(s) est (sont) une EN. D'ailleurs, presque tous les idéogrammes sont polysémiques. La nuance de l'ordre ou de la combinaison des caractères change peut-être complètement le sens des mots. Cette caractéristique rend plus compliqué le choix de catégorie.

Toutes les particularités du chinois se manifestent plus évidemment à la catégorie « person ». Primo, tous les idéogrammes ayant un sens positif peuvent être employés comme un prénom. Sur 100 idéogrammes, [Ni et al., 2011] relèvent que 1,192 correspondent à un mot inconnu. Parmi ces mots inconnus, 48% sont des noms de personnes chinoises. Secundo, la longueur des noms n'est pas fixée. Bien que la longueur d'un nom chinois varie de 2 à 4, il n'existe pas de limite des noms pour les minorités ethniques (56 ethnies en Chine) ni pour des translittérations des noms des personnes étrangères. La longueur des noms de ce genre est de 2 à plus de 10. Par exemple, la translittération peut être longue comme卡尔科特·马塔斯凯莱凯莱 (Kalkot Mataskelekele) ou阿卜杜拉布·曼苏尔·哈迪 (Abdorabuh Mansour Hadi) et courte comme柳佑益 (Yu Woo Ik). D'une part, il manque de normalisation des noms étrangers, y compris la forme d'écriture et les idéogrammes utilisés. Le symbole raccordant le nom et les prénoms peut être un espace, « - », « _ », « · », « . », « / », « ・ » ou un mélange tel que « 塞茜尔·玛丽 - 安热・马努鲁汉塔 » . D'autre part, il peut exister plusieurs

translittérations pour un même nom. Ainsi, les translittérations suivantes se rapportent toutes au prénom Marie : « 马莉 », « 玛丽 », « 玛莉 », « 马丽 » ou « 马利 ». Enfin, en raison du caractère aléatoire et ambigu d'un nom, il est très probable que les frontières d'un nom font partie du contexte. Par exemple, dans la phrase « 张少刚刚下班 » (Zhang Shao vient de finir son travail.), « 刚 » qui est un prénom chinois fréquent n'est pas une partie du nom dans ce contexte, car « 刚刚 » ensemble a un sens de « venir de ».

3. Problématique de recherche

La reconnaissance correcte des ENs des personnes aide non seulement à un bon repérage de ce type d'ENs, mais aussi à éliminer le taux d'erreur de l'extraction d'information ou les autres traitements automatiques du chinois en évitant le plus possible la mauvaise interférence des ENs de personnes.

Au regard des problèmes présentés dans la section précédente, l'enjeu du mémoire est de trouver une approche pertinente de **la REN des personnes** dans un corpus chinois **général**, en prenant compte du problème de segmentation (configuration 1 et 3), la normalisation des noms étrangers (configuration 2), ainsi que la polysémie des idéogrammes chinois (configuration 4). Ces expériences se réalisent à l'aide d'outil UNITEX (section 3.2) et un post-traitement statistique (section 3.4.4). En fonction des observations et des analyses sur les erreurs et la tendance de progression des scores (chapitre 4), j'envisage de trouver une approche qui équilibre la particularité linguistique du chinois et l'universalité des méthodes statistiques afin d'économiser le temps consacré à la construction des règles ou dictionnaires, et en même temps garder la généralité de l'approche.

L'annotation des ENs s'effectue en combinant les règles, les dictionnaires ainsi que le post-traitement statistique concernant le contexte gauche et droite des ENs. Le travail de recherche expérimentera sur une variation horizontale des échantillonnages (section 4.1), une variation verticale des configurations (section 4.2) ainsi que les comparaisons plus en détail des doubles sous-catégories à l'intérieur de la REN des personnes (section 4.3). Dans la partie perspective (section 4.4), plusieurs approches potentielles concernant les domaines relatifs sont proposées afin d'insérer ce travail dans une discipline plus étendue et considérer ce travail comme une introduction des futures recherches.

I. Etat de l'art

Le chapitre de l'état de l'art présentera les recherches existantes par rapport à notre travail. Il se divisera en 3 sections ou 3 approches différentes. Ces approches seront illustrées par recherches précises, en particulier les recherches sur la REN chinoises.

Les études portant sur la REN en anglais ont commencé à partir de 1991. Pendant la 7ème série d'IEEE, [Rau, 1991] a publié, pour la première fois, un article sur l'extraction et la reconnaissance des noms des entreprises dont la méthodologie est basée sur l'algorithme heuristique et les règles [Sun et al., 2010]. Ensuite, l'évaluation des entités nommées a été introduite comme une sous-tâche de l'extraction d'information pendant MUC-6 en 1996. Dans la suite, MUC-7 et les conférences internationales comme IEER-99 , CoNLL-2002 , CoNLL-2003 , IREX , LREC ont tous pris en compte la REN.

1. Approche symbolique

Basée sur les motifs de filtrage et les lexiques, cette approche est souvent construite manuellement [Han et al., 2004]. Une partie de notre approche est orientée par cette approche symbolique: l'établissement des graphes avec Unitex qui contient la définition des motifs du contexte des ENs et la construction des lexiques. Ces traitements sont plus ou moins influencés par les recherches suivantes.

La plupart des recherches basées sur l'approche symbolique reposent sur les traits morphosyntaxiques du langage naturel. [Nouvel et al., 2013] présentait des règles d'annotation sur un corpus de transcription bruité. Au lieu d'établir un modèle venant du l'apprentissage sur un corpus d'entraînement, ils extrayaient les règles d'annotation depuis une fouille des données en prenant compte la morphosyntaxe du langage.

L'utilisation des dictionnaires (listes de lexique) n'est plus une méthode innovante. Par exemple, le système ANNIE[1] du projet GATE pendant l'évaluation de MUC-7, est un exemple représentatif [Sun et al., 2010]. Présenté sur sa page d'accueil, ANNIE est développé au contexte de l'extraction d'information multilingue. Elle se compose de plusieurs outils

[1] https://gate.ac.uk/sale/tao/splitch6.html#chap:annie

dont « Gazetteer » joue le rôle de REN basé sur une liste de lexique. Ces listes sont en texte brut avec une entrée par ligne. Chaque liste représente un ensemble des ENs, tels que les noms des villes, des organisations, des jours de la semaine, etc. Trois types d'index (nom, type majeur et type mineur de la liste) sont définis pour y accéder. Les entités sont annotées selon leur existence dans les listes. Cette idée principale est empruntée par notre approche au moment d'appeler un dictionnaire dans des graphes de l'Unitex.

L'approche symbolique est aussi largement utilisée sur la REN chinoise. Comme les noms des ethnies minoritaires sont un de nos objets, les méthodes existantes sur la REN des noms des minorités sont aussi utiles pour notre travail. [Altenbek, 2005] a effectué une recherche sur la reconnaissance des noms de la minorité Xinjiang basée sur les règles. Un modèle était construit pour vérifier les motifs comme étiquettes de segment, existence de flexion dans le dictionnaire, titre ou suffixe. La langue de Xinjiang est une autre langue complètement différente du chinois mandarin. Les idées de l'utilisation du dictionnaire et de la « match » des motifs pour reconnaître des personnes sont pourtant valables pour le mandarin. En revanche, au lieu d'avoir un suffixe descripteur, il n'existe pas de suffixe ni préfixe comme un déclencheur de personne.

Ce que nous allons expérimenter pour la partie « rule-based » est aussi similaire à l'approche de [Wang et Shi, 2005], sauf que notre but est la REN des noms de personnes, le leur est celle des organisations. Sur la base de connaissances du chinois, cette approche détectait potentielles limites gauche et droite des ENs dans un texte, puis déterminait si une paire frontière gauche-droite renfermait un nom de l'organisation en utilisant une contrainte de longueur et « POS-tag ». Similaire à ce qu'elles ont effectué, la détection des frontières potentilles d'une entité ainsi que sa limite de la longueur sont insérées dans les graphes de l'Unitex.

En ce qui concerne les caractéristiques d'une langue, l'approche symbolique est meilleure que celle de statistique, car elle permet de gérer plus facilement les caractéristiques de la langue. Cependant à la suite du changement des systèmes ou du corpus, les règles et les ressources linguistiques ont besoin d'être construites au moins rajustées manuellement [Sun et al., 2010]. En conséquence, cette approche est limitée par l'exigence de la compétence linguistique et par les contraintes du domaine.

2. Approche statistique

À partir de MUC, la REN utilise plus les approches statistiques. C'est parce qu'elles n'ont plus besoin d'effectuer manuellement le développement de ressources pour obtenir un bon résultat [Ezzat, 2010]. Les théories statistiques sont peu à peu adaptées à la reconnaissance des entités nommées. Les divers modèles sont établis pour la REN.

[Bikel et al., 1999] ont mentionné dans leur recherche que le problème de REN était considéré comme un problème de classification : soit le mot faisait une partie des entités, soit il ne l'était pas. Ils proposaient un modèle de Markov caché de bi-grams et observaient le mot précédent et sa classe des ENs. Ensuite ils généraient le mot initial à l'intérieur de cette classe en conditionnant le mot courant et le mot précédent. Enfin ils passaient toutes les séquences de mots suivants de cette classe à un prédécesseur.

Les théories statistiques sont aussi appliquées à la REN en chinois. Par exemple, le modèle de Markov [Hu et al., 2013] ou Markov caché [Zhang et al., 2006] ainsi que CRF [Li et al., 2010].

3. Approche hybride

Dans le domaine du traitement automatique des langues, il est difficile d'éviter l'aspect statistique et les caractéristiques des langues. Cependant la plupart des recherches existantes sont hybrides. Nous présentons cette approche scindée en 3 manières[Sun, 2010].

- La première manière consiste à varier les paramètres du modèle statistique, comme la recherche de [Lv et al., 2006] qui utilisait un modèle de Markov caché en cascade.
- La 2ème manière combine les règles, les dictionnaires et le modèle d'apprentissage. A partir des modèles présentés dans la dernière section (2.2), il ajoute un autre module ou un post-traitement pour ajuster les résultats préalables. Comme le contexte gauche et droite d'une entité de personne possède certaines caractéristiques similaires, la recherche de [Li et al., 2010] appliquait un module de frontière des personnes à la base d'un modèle CRF. Ce module de frontière tenait compte de la probabilité d'un segment apparu à gauche et à droite d'une personne. La valeur du Rappel et du F-mesure augmentaient respectivement 1.99% et 1.25% par rapport à l'application du CRF tout seul.

- La 3ème manière combine les différents modèles statistiques dont les résultats d'un premier modèle seront les corpus d'apprentissage pour le prochain modèle. [Chou et al., 2004] employaient d'abord le modèle de l'entropie maximale et obtenaient les résultats préalables sur la REN biomédicales. En face des erreurs des frontières et des catégories, ils fixaient les résultats en appliquant un modèle basé sur les règles et dictionnaires.

Notre travail de recherche utilisera plutôt l'approche hybride, plus précisément, la 2ème manière de cette approche, car nous envisagerons de combiner les graphes de l'Unitex, les listes de lexique et un post-traitement avec les règles d'annotation extraites à partir des données.

4. Mesure d'évaluation

[Bikel et al., 1999] L'évaluation de la qualité de REN se réalise par intermédiaire d'un programme de scores qui est développé pour les évaluations de MUC et de MET. Ce programme mesure la valeur de rappel (R), de précision (P) dans lesquels

$$R = \frac{\text{le nombre de réponses pertinentes données}}{\text{le nombre de réponses pertinentes existant}}$$

$$P = \frac{\text{le nombre de réponses pertinentes données}}{\text{le nombre de réponses données}}$$

Enfin, la F-mesure est la moyenne harmonique pondérée du rappel et de la précision:

$$F = \frac{R + P}{2\,(R + P)}$$

Les résultats peuvent être évalués respectivement par rapport aux corpus de référence et aux corpus de test [McNamee, et al., 2011]:

- vrais positifs = nombre de réponses pertinentes données

- faux positifs = nombre de réponses non pertinentes ramenées par le système

- faux négatifs = nombre de réponses pertinentes non ramenées par le système

Le rappel et la précision peuvent aussi se reformuler comme :

R = vrais positifs / (vrais positifs + faux négatifs)

P = vrais positifs / (vrais positifs + faux positifs)

II. Matériel et méthode

La partie suivante présentera le contexte du travail, les outils utilisés pour la REN, les corpus et les règles d'annotations les plus importantes dans Unitex et post-traitement. Avant de passer à l'analyse des données, cette partie donne une vue globale sur tout le travail apporté.

1. Projet NER au SYSTRAN

Pour l'entreprise SYSTRAN, la REN est une étape essentielle, avant toutes les étapes suivantes comme normalisation, enrichissement, et traduction des ENs.

Au total, 19 langues sont à traiter dont allemand, chinois, arabe et japonais posent plus de difficultés. Ce projet commence par les langues de l'UE et les langues aux besoins des clients. 9 langues se mettent au premier niveau par le degré d'urgence. C'est-à-dire qu'ils sont marqué « level one » dans l'**Annexe 1**.

La REN ne contient pas seulement une valeur scientifique, l'application en pratique est tout de même le besoin des clients de TAL. Les objectifs du NER sont multiples au SYSTRAN :

- Fournir un outil autonome pour l'extraction d'entités qui peuvent être exploitées dans des applications différentes, telles que l'extraction d'information, l'indexation des documents, etc. La suite présente une interface des ENs dans un corpus choisi par utilisateur :

• Améliorer la traduction automatique des noms propres. D'une part, les ENs reconnues peuvent être « protégées » pour ne pas être traduites ou être traduites différemment. Par exemple, les noms chinois ont besoin seulement d'une translittération en Pinyin[2] au lieu d'une traduction. D'autre part, le type des entités nommées détectées, même si les entités sont des mots inconnus, influence la manière de traduction des contextes de ces entités.

• Enrichir le catalogue de produits multilingues à répondre à des besoins de nouveaux domaines et de nouveaux clients. Comme mentionné dans la partie introduction, la reconnaissance des entités spécifiques devient plus demandée par les entreprises spécialisées à certains domaines.

2. Outil de REN : UNITEX et Brat

2.1 La présentation générale d'UNITEX

Développé par l'Université Paris Est-Marne-la Vallée[3], Unitex est un système de traitement de corpus, basé sur la technologie orientée-automates. Le concept de ce logiciel vient du

[2] http://fr.wikipedia.org/wiki/Hanyu_pinyin
[3] http://www-igm.univ-mlv.fr/~unitex/

LADL (Laboratoire d'Automatique Documentaire et Linguistique), sous la direction du directeur Maurice Gross. Avec cet outil, nous pouvons gérer les ressources électroniques telles que des dictionnaires électroniques et des grammaires. Nous pouvons aussi travailler au niveau de la morphologie, du lexique et de la syntaxe. Il est à la fois assez puissant et simple à manipuler, car premièrement Unitex est conforme à la norme Unicode 3.0, qui permet aux utilisateurs de gérer presque tous les caractères de toutes les langues, y compris les langues asiatiques, en dépit de leurs conventions d'espacements particuliers. Deuxièmement, l'établissement des grammaires peut être réalisé avec les expressions régulières et la construction du dictionnaire ne demande pas une conversion complexe. De plus, par rapport aux autres outils de REN comme NooJ, Unitex est « open source ». Il est plus économique pour une entreprise à but lucratif. Ses principales fonctions sont:

- Construction, vérification et application des dictionnaires électroniques
- « pattern matching » avec les expressions régulières et les réseaux de transition récursifs
- Application des tableaux lexique-grammatique
- Manipulation de l'ambiguïté
- Alignement des textes
- Construction d'un automate à partir d'un corpus certifié

2.2 L'interface visuelle du processus

Ayant des fonctions variées, UNITEX est un outil très puissant. Dans ce mémoire, seulement quelques fonctions utilisées seront présentées. Pour plus d'information, cf. le manuel de l'UNITEX[4] complet.

Etape 1 : choisir une langue de travail

Unitex offre un contexte multilingue, y compris les langues fréquentes comme anglais, français, allemand, italien, russe, et aussi les langues asiatiques comme chinois, japonais, thaï, coréen, etc. Il contient au total 23 langues différentes.

[4] http://www-igm.univ-mlv.fr/~unitex/ManuelUnitex3.1.pdf

Etape2 : charger un corpus depuis local

Après avoir choisi la langue de travail, il faut ouvrir un corpus depuis local par la rubrique « Text ». Ce corpus demande à l'avance une segmentation. Par ailleurs, les utilisateurs peuvent aussi définir les paramètres de prétraitement du corpus dans la fenêtre de « Preprocessing », présenté ci-dessous :

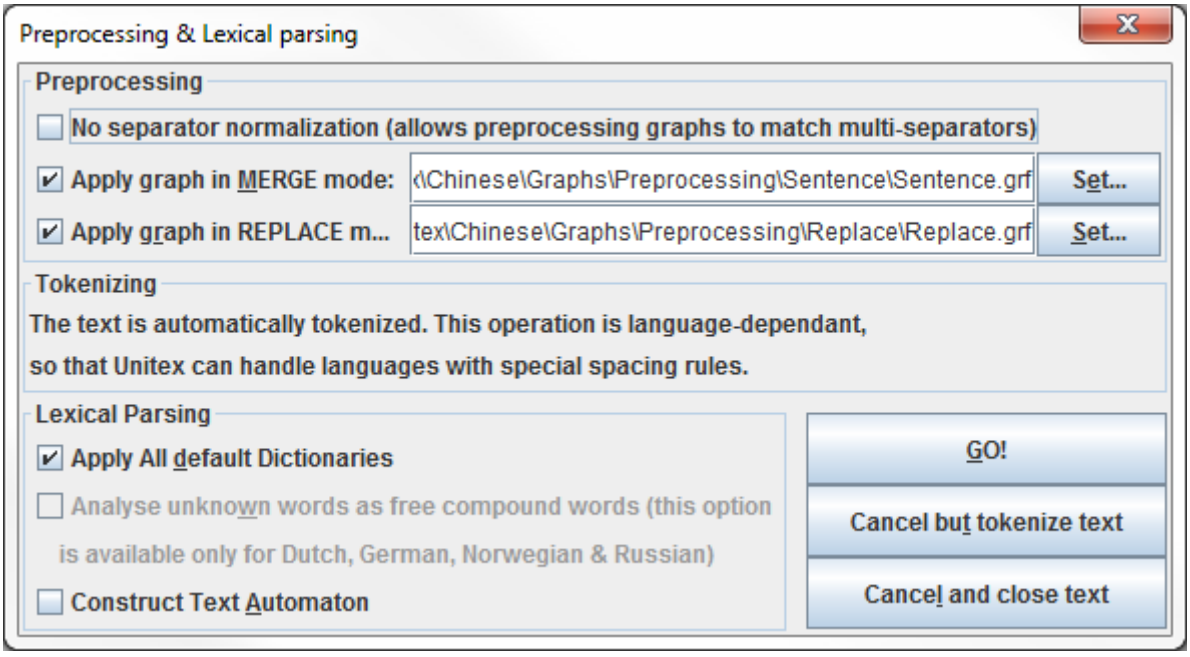


Figure 3 Interface de traitement du corpus dans Unitex

A la fin du chargement, une fenêtre affiche le texte et les informations associées.

Etape 3 : Appliquer les dictionnaires préparés:

Le dictionnaire est une liste de lexique au format :

flexion, Lemme.étiquette1+étiquette2+étiquette3

Les étiquettes sont définies par utilisateurs eux-mêmes. Ils aident à catégoriser certains groupes des chaînes en rappelant les étiquettes. En chinois, comme il n'existe pas de phénomène de flexion, les deux premières colonnes gardent une même forme. Il est possible que les utilisateurs paramètrent leurs propres étiquettes afin de l'appeler dans les automates. Tous les dictionnaires doivent être d'abord compressés au format FST avant de les appeler. La structure du dictionnaire est comme :

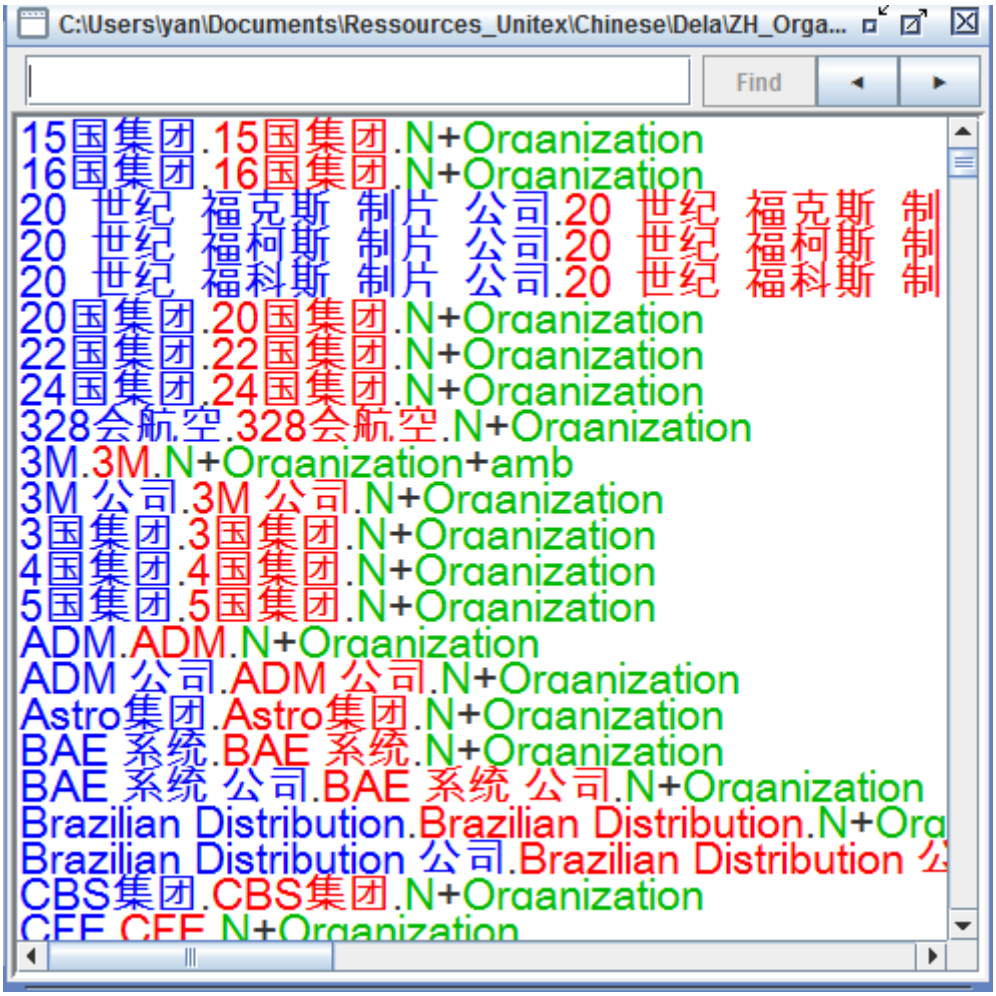


Figure 4 Structure du dictionnaire DELAF

Etape 4 : « dessiner » les graphes : formalisation des grammaires locales

Les grammaires locales sont un moyen puissant de représenter la plupart des phénomènes linguistiques. Les grammaires d'Unitex sont seulement une variante de grammaire algébrique, car elles intègrent la notion de transduction.

Un nouveau graphe contient deux boîtes : une de départ et une de fin. Les règles se réalisent avec l'ajout des boîtes au milieu et les connections des suites de boîtes avec les flèches en

deux sens. En adaptant des suites des boîtes, autrement dit des automates, les entités nommées sont reconnues. Un exemple de graphe se présente ci-dessous :

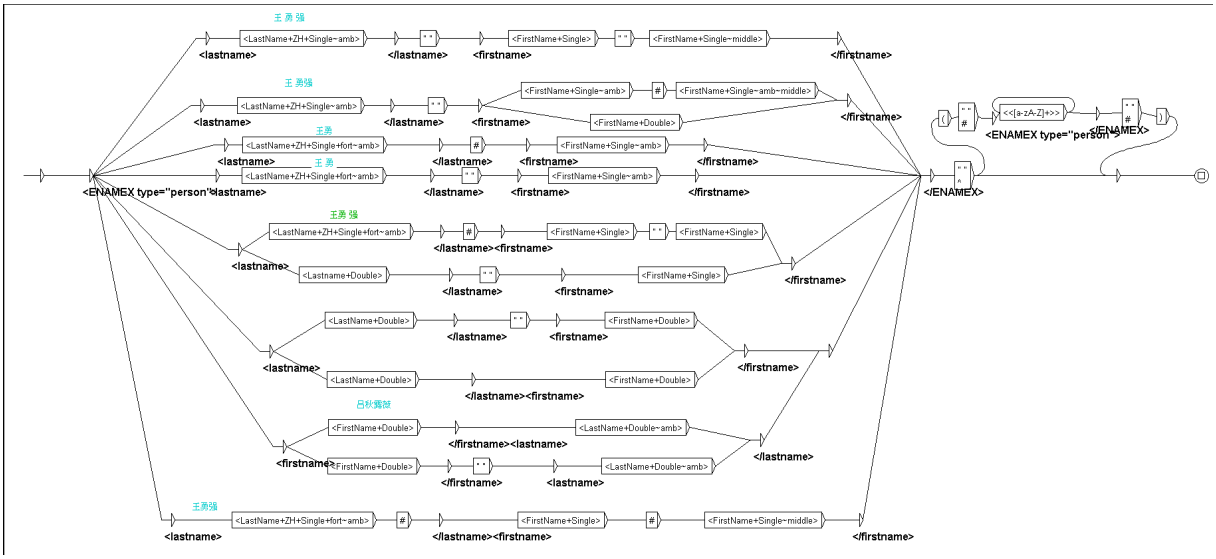


Figure 5 Aperçu des graphes Unitex

Etape 5 : Application des dictionnaires et des graphes sur un corpus avec les différentes options :

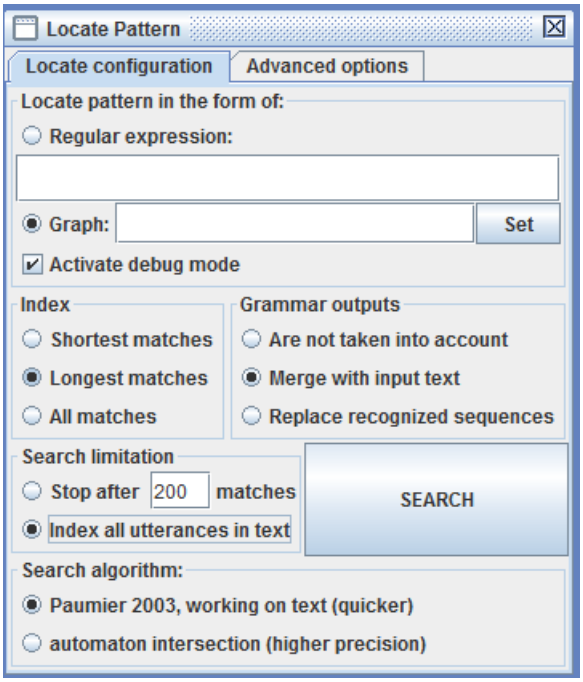


Figure 6 Appliquer un graphe Unitex

Dans cette fenêtre, les utilisateurs ont le droit de sélectionner le graphe, la façon de l'appairage (match la plus courte/longue chaîne), la sortie de grammaire (fusionner le fichier d'entrée ou remplacer les chaînes reconnues), la limite du nombre de l'appairage, etc.

2.3 Brat

BRAT est un outil d'annotation manuelle fonctionnant à partir d'un navigateur. Il permet deux sortes d'annotations :

- Annotation des entités : annoter les entités nommées du type « person », « organization » ou « location ».

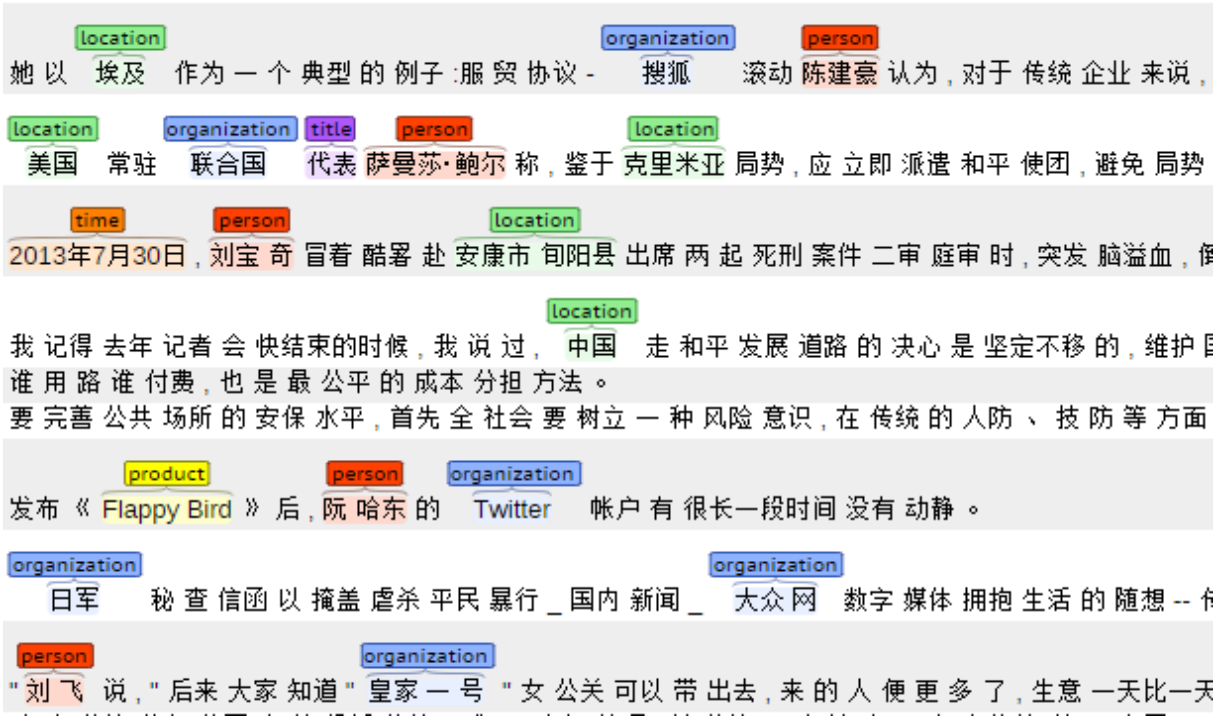
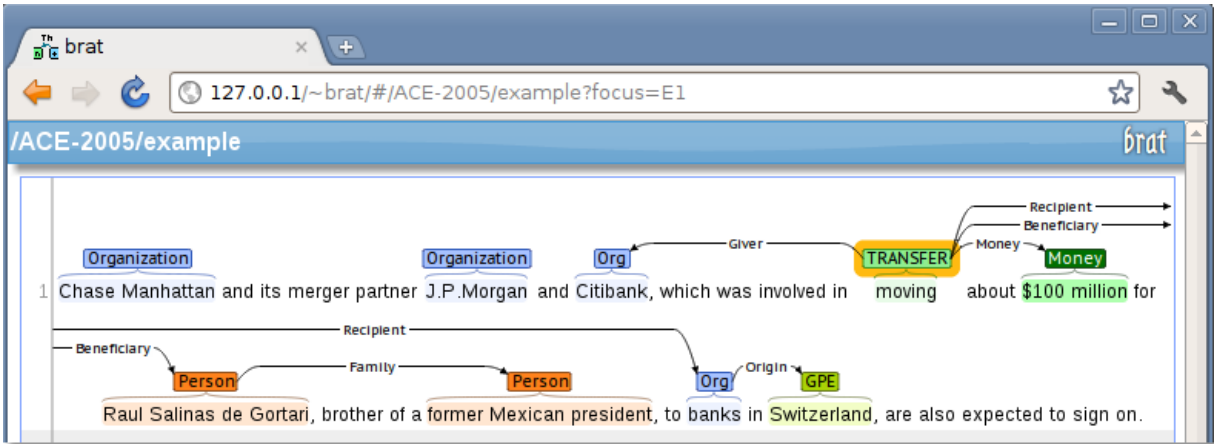


Figure 7 Interface Brat (ENs)

- Annotation des relations : Il peut aussi annoter la relation des entités dans le texte.

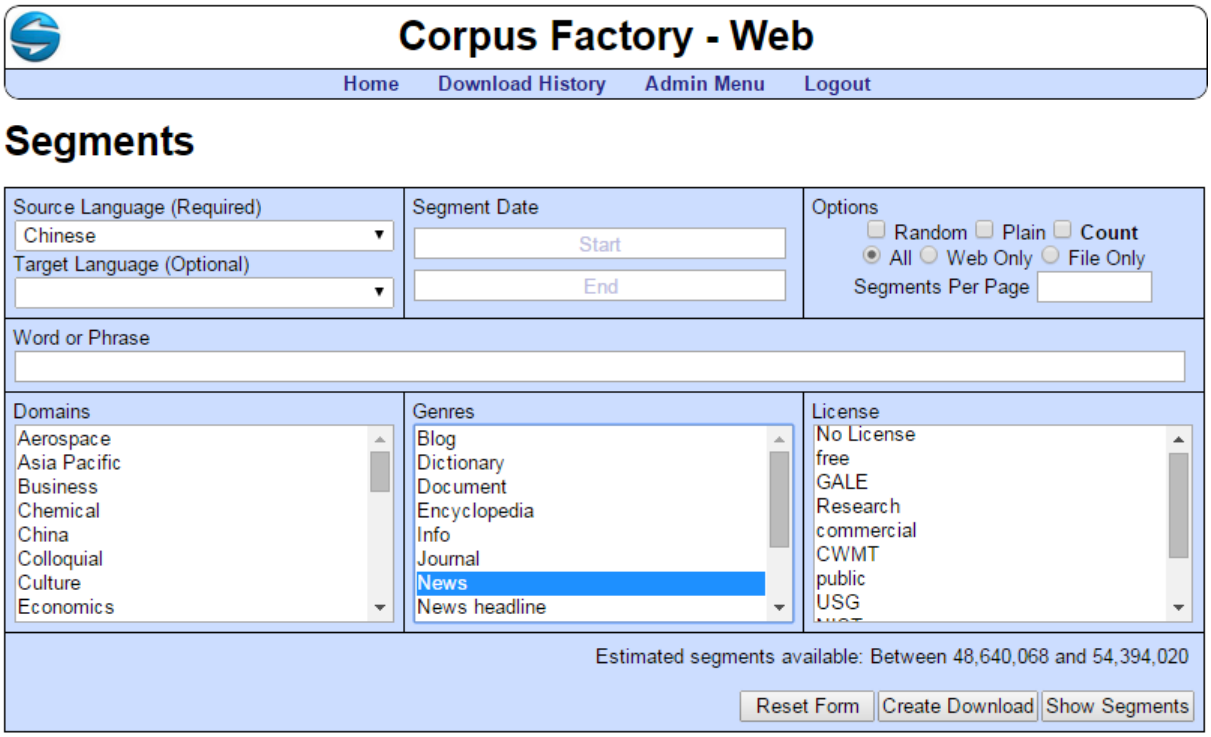


Brat nous aide à établir un référence de jeux de test. Nous avons converti le format de Brat à un texte annoté par des balises avec un script de Perl (ANNEXES 3.1).

3. Présentation du corpus

3.1 La généralité du corpus

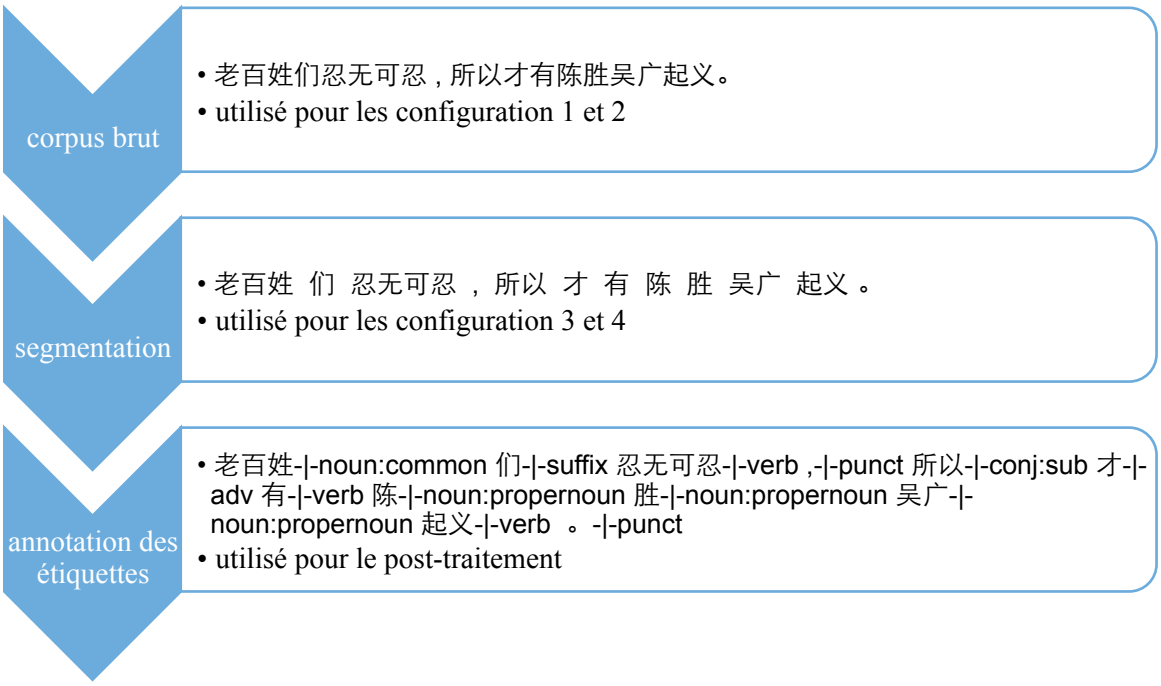
Le corpus choisi vient de la presse quotidienne en mélangeant les thèmes différentes : Une, économique, sports, politique, international, IT, etc. La préparation de corpus est facilitée par le « corpus factory » du SYSTRAN. Il existe 2 moyens de prendre un corpus de généralité : Soit le corpus est sélectionné par domaine et langue (Figure 9), soit est extrait automatiquement en définissant des portions de chaque corpus dans un fichier de CSV. Les deux moyens évitent la supervision du corpus à un certain domaine, ce qui est important pour la généralité de l'approche finale.



3.2 Le prétraitement du corpus

Segmentation et étiquette grammaticale du corpus

Les deux traitements se réalisent par « Systran Translation Engine ». Même s'il existe d'autres outil pour segmentation et étiquetage (ex. The Stanford NLP), encadrés dans le projet NER du SYSTRAN, il est plus cohérent d'utiliser son propre outil de traitement. Le processus de préparation et le format du corpus sont présentés ci-dessous :



Annotation manuelle pour le corpus de référence

Afin d'avoir deux corpus de référence, notre travail a besoin d'une annotation manuelle pour l'étape de développement et de test:

a) Référence du corpus d'entraînement

Les balises sont simplement définies comme « < » et « > » pour faciliter l'annotation manuelle comme il n'y a qu'une seule catégorie à traiter. L'annotation établie par les locuteurs natifs chinois est composée de 2 phases :

- Manuellement : chaque locuteur annote un corpus de 1000 lignes en ajoutant les balises « < », « > » pour les personnes.

- Semi-automatiquement : le corpus de 1000 lignes est déjà annoté par le modèle d'annotation préalable construit à partir deux tiers du corpus de développement. C'est-à-dire que les fautes restent à corriger par les locuteurs.

locuteur	manuellement	Demi-automatiquement	heures	Heures moyennes	Nombre de lignes
1[5]	F1[6]		2.4		1000

[5] Le premier locuteur natif chinois
[6] Le premier fichier

1	F2		2.3h	L1[7]: 2.35h	1000
2	F3		2.1h		1000
2		F6	0.5h	L2 : 1.3h	1000
3		F7	0.7h		1000
3		F8	0.5h	L3: 0.6h	1000

Tableau 1 Résumé des informations d'annotation manuelle

Le résultat de la référence est présenté comme :

- <卢-|-noun:propernoun 永仁>-|-noun:propernoun 曾-|-adv 就读-|-verb 香港-|-noun:propernoun 圣类斯中学-|-noun:propernoun
- 网站-|-noun:common 负责人-|-noun:common <波普肯>-|-noun:propernoun 说-|-verb
- 证券-|-noun:common 分析师-|-noun:common <严-|-noun:propernoun 小-|-noun:propernoun 飞>-|-noun:propernoun

b) Réréfence du corpus de test

1000 phases de jeux de test sont annotées par un outil d'annotation Brat. La sortie de Brat est sous le format :

<SYSTRAN sentence_id 24><entity_PHYSUB>京华时报</entity_PHYSUB> 讯 (记者 <entity_HUMANS>龚 棉</entity_HUMANS>) <entity_DATE>5月28日</entity_DATE> , <entity_GROUP><entity_GPE>北京市</entity_GPE> 红十字会</entity_GROUP>

Construction du dictionnaire

Unitex permet d'utiliser des dictionnaires DELAF qui contient la flexion, son lemme et les étiquettes associées. C'est une suite des informations liées par « , », « . » et « + » sans espace :

Flexion,lemme.étiqutte1+étiquette2+étiquette3

Précisé par un exemple en français :

Président,président.N+nounpropre+title

[7] Les heures utilisées en moyenne du premier locuteur

Comme le phénomène de flexion n'existe pas en chinois, la « flexion » et le « lemme » sont la même forme:

陆,陆.N+Hum+LastName+ZH+Single+fort

Les étiquettes définies personnellement, ne sont pas forcément grammaticales. Elles sont plutôt un nom de catégorie, ce qui est appelé dans les graphes.

Les dictionnaires chinois sont extraits de SYSTRAN. Ils sont aussi utilisés comme un filtrage dans le post-traitement. Les entités entre 2 à 5 caractères dans la liste de lieu et d'organisation constituent un filtrage de lexique. Cela garantit le fait que les ENs récupérés par le post-traitement ne sont pas dans ces listes. À l'inverse, le premier caractère des entités doit être dans la liste de noms et translittération.

4. Annotation du corpus

4.1 Guide d'annotation

Les entités nommées reconnues servent à améliorer la traduction de manière que sachant un mot inconnu annoté comme une personne, la traduction de ce mot et les mots du contexte ne seront plus les mêmes. Il demande ainsi à repérer nom, prénom et indicateur (titre et occupation) de personne.

La catégorie « person » doit être composée d'un prénom (ou plusieurs prénoms), suivie d'un nom de famille et, dans certains cas, une particule (Mme, M., Melle., Dr., etc). Les titres sont également extraits, mais non inclus dans la balise. Ils ne font pas partie de l'évaluation, mais ils servent à mieux repérer les ENs à côté d'eux. Un titre est la fonction d'une personne ou réfère à une personne (titre: président de Etats-Unis = personne: Barack Obama). Ils peuvent englober d'autres entités nommées telles que les lieux ou organisations, mais l'expression entière est considérée comme « title ».

Les titres contiennent:

- La fonction politique : 主席，国务卿，国防部长，纽约市长 (secrétaire d'État, ministre de la Défense, le maire de New York City, etc)
- Titre militaire : 队长，将军，上校 (capitaine, général, colonel, etc)
- État civil : 先生，女士，医生，女士 (Mr., Ms., Dr., Mrs., etc.)

- Fonction professionnelle : 首席执行官，人力资源总监，教授 (CEO, le directeur des ressources humaines, professeur, etc)

4.2 Annotation avec Unitex

Hiérarchie des graphes

Nous pouvons détailler la REN des personnes en plusieurs tâches :

A. Certaines personnes sont des noms et prénoms connus de dictionnaires. Ce genre de personnes sont les hommes politiques, les stars ou les cadres supérieurs des sociétés connues. Au niveau de la segmentation, il n'y aura pas d'espace à l'intérieur de ces entités. Ils sont entièrement reconnus par l'application de dictionnaires. La Figure 4 extrait toutes les personnes dans le dictionnaire. Un des étiquettes est « Person ».

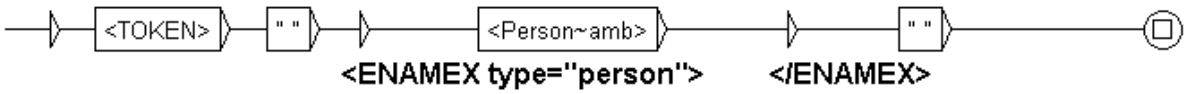


Figure 10 Graphe 1

B. La majorité de noms chinois apparus dans un corpus sont considérées comme des mots inconnus. Par conséquent, ils sont très souvent segmentés, plutôt mal segmentés. Dans ce cas-là, les graphes essaient d'épuiser toutes les possibilités de combinaisons à l'aide des dictionnaires. L'exemple suivant présente tous les cas possibles pour un nom avec un prénom au total de 3 idéogrammes. En réalité la situation est plus compliquée, car tout le monde ne procède pas un nom de 3 caractères. Les autres possibilités sont présentées dans les annexes 4.

   a. idéogramme1(espace)idéogramme2(espace)idéogramme3

Figure 11 Graphe 2

   b. idéogramme1idéogramme2(espace)idéogramme3

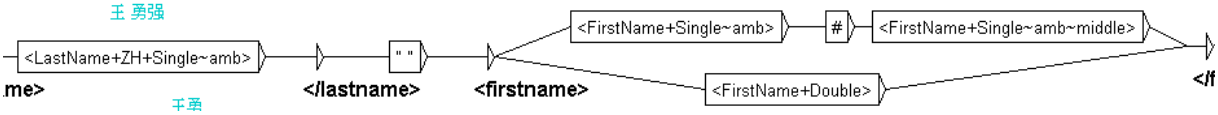


Figure 12 Graphe 3

c. idéogramme1(espace)idéogramme2idéogramme3

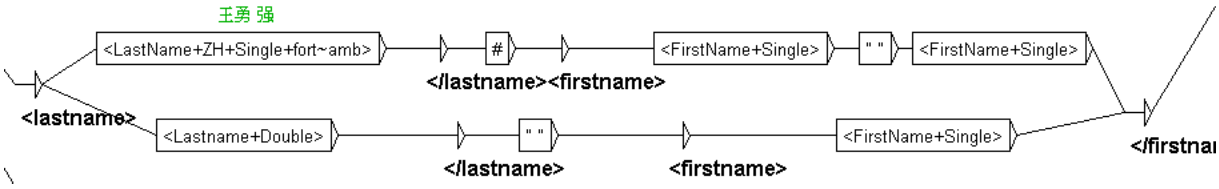


Figure 13 Graphe 4

En évitant le plus possible les bruits, la liste de noms de famille fréquents sert à limiter la frontière gauche de l'entité. C'est-à-dire qu'en appliquant un dictionnaire de noms chinois, il élimine une partie d'erreurs.

C. Les noms et prénoms étrangers de translittération phonétique occupent une grande partie parmi toutes les personnes reconnues, surtout dans un corpus de presse internationale. Le nombre d'idéogrammes varie selon la nationalité des personnes. Par exemple, souvent 2 à 3 idéogrammes pour des personnes coréennes, 4 à 6 des personnes japonaises, 2 à plus de 10 pour les personnes occidentales selon le nombre de syllabes. En face de cette variété, une normalisation des noms étrangers est effectuée en deux aspects :

   a. Une normalisation des idéogrammes réservés aux noms et prénoms étrangers : établissement d'un dictionnaire des idéogrammes courants pour la traduction des noms étrangers.
   b. Une normalisation des symboles de raccordement entre noms et prénoms.

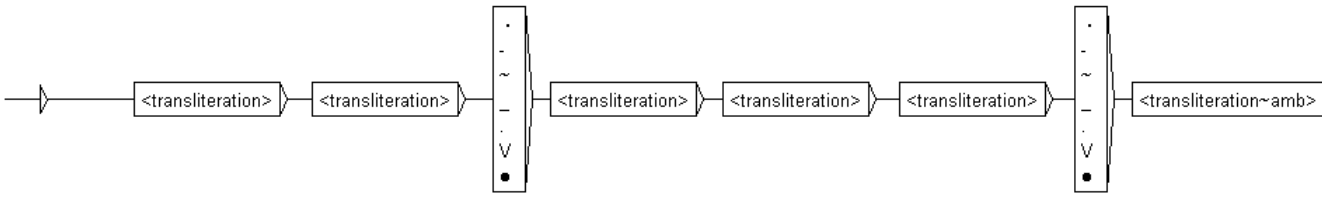
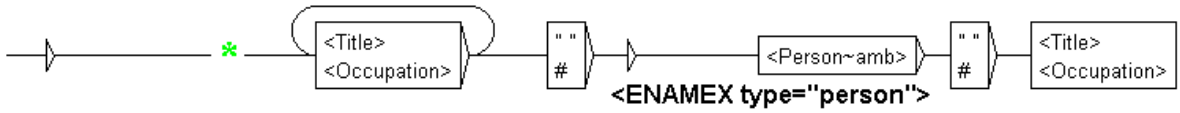


Figure 14 Graphe 5

D. Certaines personnes s'entourent des indicateurs onomastiques. Les indicateurs sont souvent les titres, les occupations ou les mots fréquents auprès d'une personne (comme verbe « parler »). Dans Unitex ils sont précisés par un contexte gauche ou droit :

   a. Contexte gauche :



b. Contexte droit :

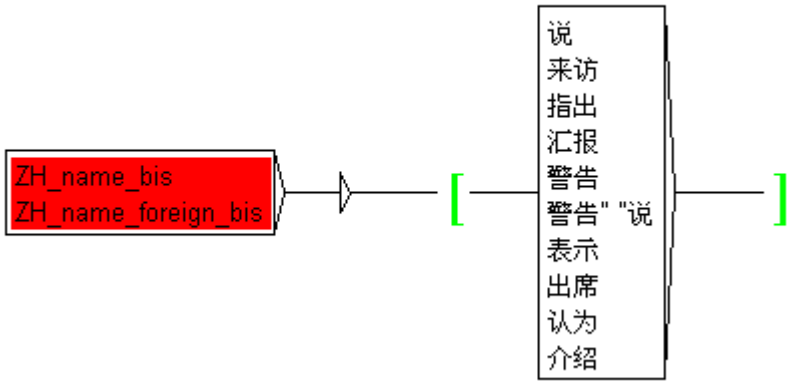


Figure 16 Graphe 7

4.3 Algorithme de post-traitement

L'approche basée sur les règles et les dictionnaires demande avant tout une connaissance linguistique. La construction des grammaires et dictionnaires prennent trop de temps. De plus, les règles ne peuvent pas prendre en considération toutes les possibilités, en particulier les phénomènes qui ne paraissent pas dans le corpus de développement. Un post-traitement est ajouté dans le but de récupérer les noms de famille et prénoms isolés en dehors des règles. Inspiré par la théorie de CRF et HMM présentés dans la partie de l'état de l'art, ce post-traitement s'appuie sur la relation des ENs des personnes avec les mots précédents et suivants.

4.3.1 Statistique préalable

La première étape calcule la plus haute fréquence des étiquettes précédentes, actuelles et suivantes (en ajoutant deux marqueurs BOS et EOS représentant une personne apparue au début ou à la fin de la ligne). Réalisé par deux scripts dont le premier sert à calculer la fréquence de toutes les étiquettes apparues et le deuxième regroupe ces étiquettes de plusieurs catégories en une seule de cette manière :

Selon le regroupement, voici les étiquettes (et leurs fréquence sur le corpus de développement) sélectionnées pour un filtrage :

catégories	Contexte gauche	actuel	Contexte droite
noun	458	1000	269
punct	330		310
BOS/EOS	265		23
verb	158	2	413
prep	100		109
conj	61		54
particle	36		78
adv	26		175
adj	18	2	24
aspect	17		22
number	2		31

Tableau 2 Etiquettes dans post-traitement

4.3.2 *Un filtrage par des étiquettes*

Selon le résultat de la statistique préalable, ce filtrage vise à sélectionner seulement des chaines de mots restreintes par les étiquettes du contexte gauche et droite. En outre, les étiquettes de ces mots sont aussi limitées. 97%[8] de personnes sont étiquetées comme « noun:propernoun », 1.5%[9] sont « noun:common ». Selon le corpus de développement, « noun:common » occupe la plus grande portion. En vue d'éviter les bruits, nous avons supposé que toutes les personnes sont étiquetées comme « noun:propernoun ».

[8] 1125 « noun:propernoun » parmi 1162 entités personnelles.
[9] 18 « noun:common » parmi 1162 entités personnelles.

4.3.3 *L'application des lexiques*

Il est possible que certains éléments erronés soient présents dans la liste de personnes. Les listes d'organisations et de lieux servent à tout d'abord éviter ces deux sortes d'entités nommées. Ensuite, une liste de noms de famille contrôle le premier caractère du segment. Par la suite, il faut collecter tous les segments étiquetés comme « noun:propernoun » mais qui ne sont pas une personne dans le corpus de développement. Cette liste filtre une quantité de lexique qui ne sont pas une organisation ni un lieu, comme 互联网 (Internet), 伊斯兰教 (Islam) et 红楼梦 (« dream of Red Mansions »), etc.

Pour conclure, le modèle peut être présenté comme ci-dessous :

<table>
<tr><th>Position</th><th colspan="2">Contexte gauche</th><th colspan="2">Token actuel</th><th>Contexte droite</th></tr>
<tr><td>BOS</td><td rowspan="3">(! org) &&
(! loc) &&
(! list_noisy)
&&
(list_lastname)
&&
(list_tranlitteration)</td><td></td><td rowspan="3">noms
de
famille</td><td>noun:propernoun</td><td rowspan="2">verb|punct|noun|
particle|
adv|prep|conj|particle|
relpron|aspect|
num|adj</td></tr>
<tr><td>milieu</td><td rowspan="2">verb|punct|noun|particle|conj|prep|conj|aspect|adv|adj|num |prefix</td><td>noun:propernoun</td></tr>
<tr><td>EOS</td><td>noun:propernoun</td><td></td></tr>
</table>

Tableau 3 Résumé du modèle post-traitement

4.4 Description des 4 configurations

4.4.1 *Application sur le texte tel quel sans prétraitement*

Cette première expérience annote un corpus avec Unitex sans aucun prétraitement. Les grammaires s'appliquent sur un texte sans segmentation. Nous supposons que la qualité de cette annotation sera plus modeste, le résultat sera pourtant utile pour prouver les difficultés intrinsèques du chinois.

File	TUs	Matche	G_part[10]	T_part[11]	G_total[12]	T_total[13]	P	R	F
Conf 1	6000	562	10	56	1332	905	0.621	0.422	0.502

Tableau 4 Configuration 1

La table ci-dessus montre que plus de 50% des entités de personne ne sont pas reconnues par Unitex. Quant aux entités trouvées, seulement 60% sont correctes. La plupart des entités correctes viennent du dictionnaire. Autrement dit, sans segmentation les grammaires ne font qu'un match des suites de caractères. Cela est limité par le complet du dictionnaire. Les entités correctes contiennent peu de personnes étrangères, surtout les noms de famille suivis par un ou plusieurs prénoms. Par exemple

欧锦赛冠军谢尔盖·卡扎科夫 VS. 欧锦赛冠军<谢尔盖·卡扎科夫>

En revanche, une partie des noms étrangers sont précédés ou suivis par un titre sont repérés par leurs indicateurs :

儿子<珀罗普斯> (fils<珀罗普斯>).

Le manque des noms de personnes explique le faible score du rappel (R= 0.422).

4.4.2 *Application sur le texte avec normalisation des noms étrangers*

Dans le but de reconnaître plus de noms étrangers ignorés, des graphes de normalisation s'ajoutent dans la 2ème expérience. À part des noms étrangers existant dans le dictionnaire, une liste d'idéogrammes de translittérations est passée à Unitex. En même temps, les symboles de raccordement sont aussi définis. Avec ces modifications, les résultats étaient légèrement progressés.

File	TUs	Matche	G_part	T_part	G_total	T_total	P	R	F
Conf 2	6000	575	12	55	1332	921	0.624	0.432	0.510

Tableau 5 Configuration 2

En ajoutant les graphes concernant les personnes étrangères, 16 entités de plus sont acquises. Un résumé de *diff* sur le corpus annoté pendant les 2 premières configurations, nous avons observé que, d'un part, les entités comme 欧锦赛冠军<谢尔盖·卡扎科夫> et 涉案嫌疑人<

[10] Un match partiel dans le corpus de référence
[11] Un match partiel dans le corpus de test
[12] ENs total dans le corpus de référence
[13] ENs total dans le corpus de test

汤姆·斯蒂芬斯> sont reconnues. D'autre part, quelques entités sont seulement partiellement reconnues 巴西老教练<吉尔森·塞凯拉>·努涅斯 par rapport à la référence 巴西老教练<吉尔森·塞凯拉·努涅斯>. Du point de vue global, la performance de l'Unitex est limitée sans segmentation.

4.4.3 *Application sur le texte avec normalisation des noms étrangers et ajout d'espaces autour des unités significatives*

D'après les deux expériences précédentes, nous pouvons déjà conclure que la reconnaissance des entités nommées sans segmentation ne suffit pas pour obtenir de bons résultats. C'est justement cette réflexion qui a inspiré la 3ème expérience où une segmentation préalable est apportée au corpus avant de le passer à Unitex. Nous avons utilisé la propre segmentation de SYSTRAN afin de garder l'homogénéité des traitements ultérieurs. Les résultats sont présentés ci-dessous :

File	TUs	Matche	G_part	T_part	G_total	T_total	P	R	F
Conf 3	6000	684	19	41	1332	920	0.743	0.514	0.607

Tableau 6 Configuration 3

Les résultats montrent que le nombre total des ENs trouvés n'a pas changé, alors que le nombre des entités nommées correctes augmente, et en même temps, celui des entités reconnues partiellement réduit. Ce changement améliore la valeur de la précision (de 0.621 à 0.743). Au contraire, la valeur du rappel n'est pas considérablement améliorée. Autrement dit, une grande quantité d'entités nommées restent toujours non reconnues.

4.4.4 *Idem + post-traitement*

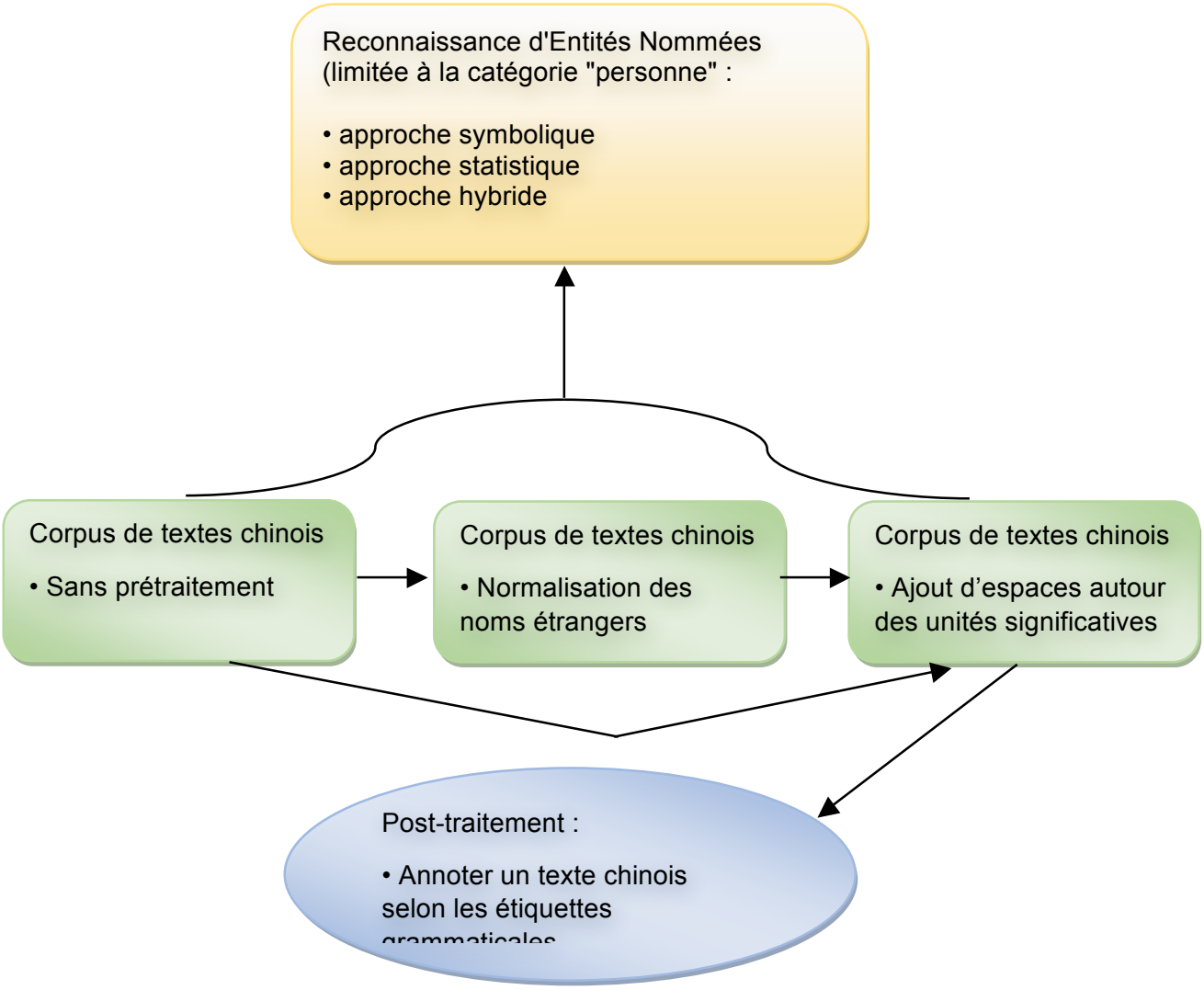
Le post-traitement de la dernière configuration essaie de corriger le défaut des 3 premières expériences : l'absence des ENs. Ce post-traitement est orienté par les relations des étiquettes entre ENs de personnes et leurs contextes. Inspiré par les principes de CRF et HMM, ceux qui analysent non seulement les entités elles-mêmes, mais aussi l'état des entités associées ainsi que la probabilité de leurs transitions. Au lieu de construire un modèle complet de CRF ou HMM, ce post-traitement n'utilise qu'une partie des algorithmes. En conséquence, cette approche calcule d'abord les étiquettes de deux côtés d'une entité de personne et sa propre étiquette grammaticale et met en ordre toutes les étiquettes de fréquence décroissante. Ensuite, selon des étiquettes choisies, nous avons construit un modèle d'annotation dans lequel chaque entité est associée d'une étiquette grammaticale. Le but de ce processus est de récupérer les

ENs oubliées par la sortie d'Unitex. Le résultat escompté est une perfection de la valeur du rappel. La suivante présente les informations sorties de cette étape :

File	TUs	Matche	G_part	T_part	G_total	T_total	P	R	F
Conf 4	6000	1081	27	70	1332	1478	0.731	0.812	0.769

Tableau 7 Configuration 4

Pour conclure la partie de « matériel et méthode », nous avons visualisé le processus avec un schéma suivant:



III. Evaluation et Discussion

1. Variation des échantillonnages

En employant BASIS, un outil basique de REN, nous avons varié le type des échantillonnages par domaine pour avoir d'abord une vue globale sur le recouvrement des personnes par rapport aux autres catégories des entités nommées. Les corpus et leurs domaines sont présentés dans le tableau 8.

Nom du fichier	Domaine	Nb de ligne
zhcna	CNA news	10000
zhcol	Transcription des paroles	10000
zhft	Presse financielle	10000
zhmil	Presse militaire	10000
zhst	Scientifique & technologie	10000

Tableau 8 Echantillonnages

Avant de vérifier les résultats, l'hypothèse est que les actualités générales contiennent plus d'ENs de personne alors que les articles technologiques ou scientifiques moins.

La tableau 9 donne un résumé du nombre des ENs de différentes catégories par domaine : en général, les « person » (18252) occupent 21% parmi toutes les entités nommées (87294). Le corpus de technologie a la moins grande portion (12%) des personnes. Cette observation conforme à l'hypothèse. La presse générale et économique contient le plus de personnes. Lorsque le nombre de titre est 2 fois moins que celui de personnes, une grande moitié des personnes ne peuvent pas être reconnues par les titres. C'est-à-dire que les motifs confirmés aux contextes ne suffiront pas à toutes les situations.

tags	zhcna	zhcol	zhft	zhmil	zhst
<IDENTIFIER: PHONE_NUMBER>	0	0	2	0	0
<IDENTIFIER: DISTANCE>	216	92	33	324	186
<IDENTIFIER:	2	0	0	5	2

LATITUDE_LONGITUDE>					
<IDENTIFIER:URL>	11	0	17	1	14
<IDENTIFIER:EMAIL>	2	0	1	1	0
<NATIONALITY>	54	41	131	60	7
<TEMPORAL:DATE>	1104	345	277	591	508
<TEMPORAL:TIME>	2511	920	464	658	324
<TITLE>	3238	1251	886	1285	715
<PERSON>	6883	2273	3739	2937	1497
<LOCATION>	8684	5709	5728	6472	3372
<RELIGION>	72	72	37	29	4
<ORGANIZATION>	6922	1909	4880	2763	5642
Total des ENs	29699	12612	16195	15126	12271
Pourcentage pour « PERSON »	23%	18%	23%	19%	12%

Tableau 9 Etiquettes dans échantillonnages

Le tableau 10 présente les noms de personnes en alphabet latin. Surtout dans la presse économique et générale, plus de 100 entités concernant une personne sont en anglais. Cela prouve que le traitement sur les personnes en anglais est aussi indispensable dans un corpus chinois.

Tags-anglais	zhcna	zhcol	zhft	zhmil	zhst
PERSON	83	2	810	9	46
TITLE	30	1	123	7	21

Tableau 10 "Person" et "title" dans échantillonnages

2. Variation des expériences

2.1 Résultats des prétraitements du corpus

	ENs correctes	P	R	F
Sans seg	562	0.621	0.422	0.502
seg	684	0.743	0.514	0.607

Tableau 11 Segmentation vs. Non segmentation

Nous avons obtenu une progression du score avec une segmentation du corpus, car au lieu de traiter chaque caractère comme un élément essentiel, la segmentation regroupe les caractères en mot. Ce traitement préalable analyse des lignes de caractères collés. C'est une sorte de désambigüisation qui transforme une simple correspondance des motifs à une correspondance sémantique et morphologique. Au lieu de prendre en compte un groupe de caractères indépendant qui joue un rôle grammatical, il extrait une partie du groupe si le dernier correspond à un motif ou un lexique du dictionnaire. Précisons par un exemple « 布宜诺斯艾利斯 » (Buenos Aires), une partie (布宜<诺斯艾利斯>) de cette entité complète est extraite comme un nom de personne. En outre, sans segmentation le concept des frontières des entités n'existe pas. C'est-à-dire que nous ne pouvons pas repérer les personnes par les traits morphologiques du chinois, comme les déclencheurs descriptifs ou des indicateurs des noms. Un indicateur verbal fort suivi une entité de personne « 认为 » n'a pas aidé à reconnaître le nom le précédant.

Après une comparaison entre les résultats avec et sans segmentation, nous pouvons déjà conclure que la REN sur le chinois, au moins pour les ENs des personnes, un corpus non segmenté empêche la performance des outils et des modèles.

2.2 Normalisation des noms étrangers

Les noms des étrangers, en translittération ou en lettres alphabétiques, apparaissent fréquemment dans les articles de presse. En face de la généralité du corpus, les noms des étrangers occupent presque une moitié des tous les noms.

	ENs correctes	P	R	F
Sans Norm	562	0.621	0.422	0.502
Norm	575	0.624	0.432	0.510

Tableau 12 Normalisation vs. Non normalisation

Le tableau 12 montre une comparaison entre les configurations avant et après la normalisation des noms étrangers. La qualité de REN est légèrement améliorée, mais les graphes de normalisation sont appliqués en repérant 13 noms de plus étrangers surtout des noms complets (nom avec prénom liés par des symboles) comme « 侯赛因 · 胡塞 », « 史蒂文 · 乔布斯 », « 黛丽 · 赫本 », etc. Cependant le nombre de reconnaissance partielle augmente aussi par manque des graphes assez complets (ex. <欧内斯特 · 约瑟夫> · 金). À travers les résultats

ci-dessus, cette manière de configuration dans laquelle seulement les caractères et l'écriture des symboles sont définis n'est pas suffisante pour une normalisation des noms étrangers.

2.3 Post-traitement

Les entités nommées reconnues par Unitex sont possibles d'être repérées par certains motifs, ou recouvertes par des dictionnaires. Dans ce cas-là, une grande partie des noms de personnes sont oubliés si elles ne satisfont pas à ces deux conditions. Autrement dit, les personnes sont en manque des indicateurs repérables ou contiennent un prénom rare. C'est pourquoi le Rappel est inférieur à 0.6. Un modèle de post-traitement vise à améliorer ce résultat en reconaissant les noms ou prénoms isolés dans le corpus. Nous avons obtenu des résultats du post-traitement par rapport à la sortie de l'Unitex.

	ENs correctes	P	R	F
Unitex	684	0.743	0.514	0.607
Post-traitement	1081	0.731	0.812	0.769

Tableau 13 Unitex vs. Post-traitement

Evidemment, le score R augmente de 30% avec un ajout de 400 entités.

2.4 Analyse des 4 configurations

D'une vue globale sur les 4 configurations, nous avons obtenu une progression significative des scores (la Figure 17).

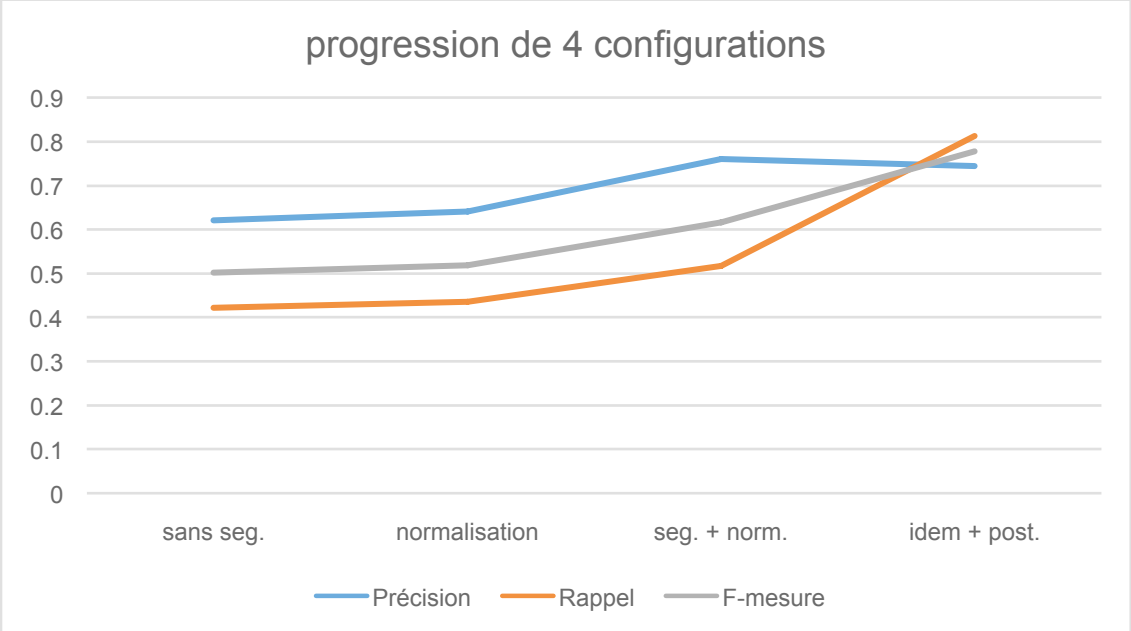


Figure 17 Progression de 4 configurations

La première configuration (*sans seg.*) prouve que l'application sur le texte tel quel sans aucun traitement ne suffit pas pour REN des personnes. La meilleure résolution est la configuration 4 (*idem+post.*). Elle combine les 4 premières étapes concernant les entités repérables par leurs contextes morphologiques en ajoutant les influences des étiquettes grammaticales du contexte. Même si une entité est isolée des indicateurs qui n'existent pas dans un dictionnaire, nous pouvons essayer de le repérer avec les contraintes des étiquettes contextuelles et la longueur de cette entité. Cette expérience a obtenue un meilleur résultat seulement comparé aux 3 premières expériences.

En comparant ce résultat à celui de SYSTRAN avec le même corpus, nous avons obtenu la figure suivante (Figure 18) :

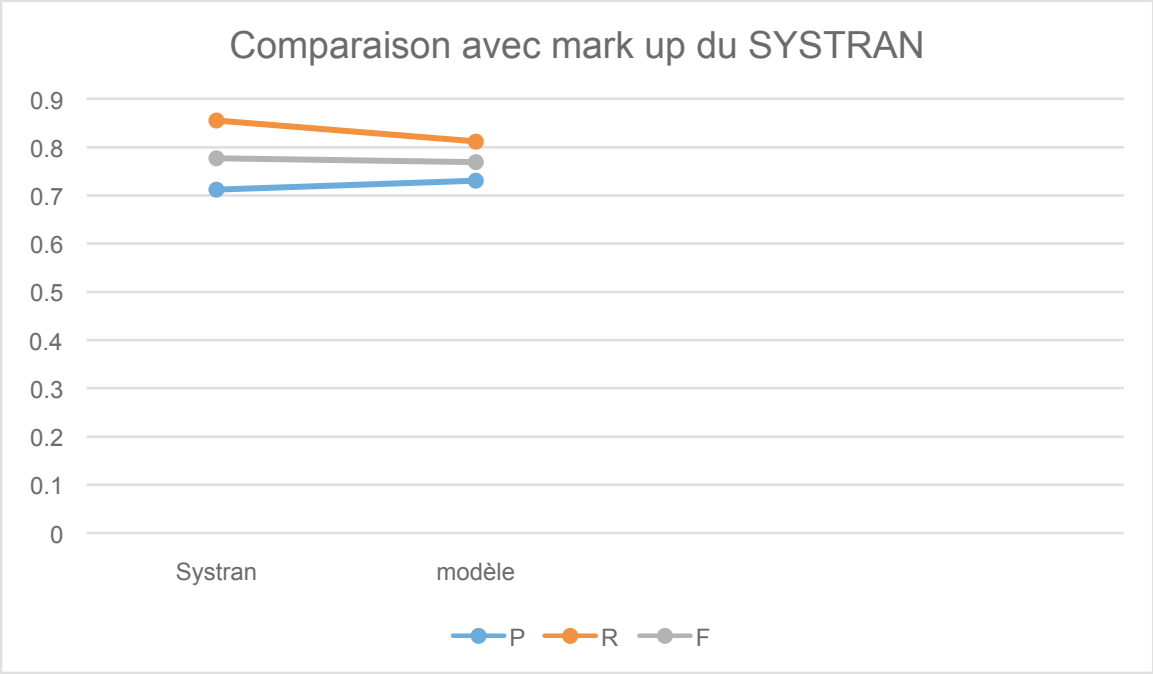


Figure 18 SYSTRAN vs. "Nous"

Nous avons atteint à peu près la même F-mesure (*la courbe grise*) avec celui du SYSTRAN (0.769 vs. 0.777). Quant au Rappel et Précision, chacun a son avantage. Le modèle établi a un meilleur score du Rappel (0.855 vs. 0.812) lorsque SYSTRAN a une meilleure Précision (0.731 vs. 0.712). Cela illustre que SYSTRAN définit plus correctement les frontières et les étiquettes des ENs tandis que le modèle repère plus d'ENs parmi les ENs de référence.

Même si la performance du modèle est passable, sa F-mesure de 0.769 est encore loin des résultats des autres approches présentées dans l'état de l'art. Les faux négatifs peuvent être divisés en 5:

a) Erreurs de frontières :

- Le titre avec l'entité suivant est considéré comme une personne.
  联合国 助理<秘书长 穆雷>, « 秘书长 » (le Secrétaire général) est un titre du prénom « 穆雷 ».

- La correspondance la plus longue garde plusieurs personnes comme une entité

  3 personnes successives <刘良伟> <董晓斌> <郭莹>

  sont étiquetées comme

  <刘良伟 董晓斌 郭莹>

b) Erreurs dues à ajout de segmentation non pertinent :
   En général, l'application de segmentation facilite la REN en chinois. Cependant, quelques fois un espace sépare des entités significatives, surtout pour un prénom non fréquent. 记者 <席 则> même précédé par une occupation « 记者 » (journaliste), le modèle n'a pas reconnu cette personne avec un nom et prénom séparés, car premièrement le nom chinois « 席 » n'est pas commun et ensuite le prénom « 则 » est une conjonction très fréquenté ayant un autre sens « alors ». Le segmenteur ne le considère surtout aux personnes chinoises ayant 3 caractères.

   Le même problème se trouve aussi aux noms étrangers. 埃及 外援 <依 哈比> dont « 依 » est utilisé au sens de « selon » par rapport à un caractère de translittération. Le résultat de modèle est donc 埃及 外援 依 <哈比>.

c) Erreur d'ambiguïté :
   En raison de la polysémie des caractères chinois, certaines combinaisons des caractères pour un prénom chinois sont utilisées aussi comme une organisation

ou un lieu. Quand ces combinaisons font partie des autres catégories des ENs, il est difficile de l'éliminer.

- <陆军> dont « 陆 » et « 军 » sont respectivement le nom et le prénom communs en chinois, mais les deux ensemble ont un autre sens complètement différent : « interarmes ». Il est pareil pour « <安卓> 系统 » qui devient un nom propre « Android » système et « <基尔福特> 大学 » qui est une université (Université Keele Ford), ainsi que <法 也 斯 法> qui ressemble à un prénom de translittération, est un nom de loi. 李鹏 会见 <斯里兰卡> 总统, <斯里兰卡> est le nom du pays.

d) Une partie des personnes ne sont pas identifiées :
   - 美国耶鲁大学医学院 副教授 <索尼娅·卡普里奥> dans lequel le nom de translittération respecte la graphe de la normalisation des noms étrangers, mais il n'a pas été identifié.

e) Les bruits ajoutés par le post-traitement : le post-traitement porte à la fois les entités oubliées et des bruits. Comme le modèle est établi par la catégorie du contexte et la longueur du mot, les entités satisfaisant à ces deux conditions sont toutes entrées dans la liste.
   - Par exemple, 那些 上百 年 的 <古 杉> 在 今天 肯定 寸 树 寸 金, « 古 » est bien un nom chinois, « 杉 » peut être utilisé comme un prénom. De plus, le contexte gauche « 的 » et droit « 在 » sont respectivement une particule et une préposition définie pour les catégories grammaticales du contexte. La longueur du mot est entre 2 à 5. Cependant « 古 杉 » est un « ancien cèdre » au lieu d'une personne.

3 Analyse des doubles sous-catégories

3.1 Noms chinois vs. Noms étrangers

À l'examen du corpus annoté selon des noms chinois ou étranger, et tout en nous appuyant sur la sortie de 4ème configuration, nous avons obtenu une comparaison :

File name	TUs	Matched	G_part	T_part	G_total	T_total	P	R	F
1000-NomChinois	1003	112	1	3	131	138	0.812	0.855	0.833
1000-NomEtranger	1003	87	1	11	123	131	0.664	0.707	0.685

Tableau 14 Noms chinois vs. Noms étrangers

Pour un corpus de 1000 lignes sélectionnées aléatoirement dans le corpus de développement, Ce mini corpus contient presque le même nombre de noms chinois (131) et étrangers (123). La F-mesure pour chacun les deux (0.833 vs. 0.685) implique que les noms étrangers sont un obstacle pour REN des personnes. Pour les noms étrangers apparus dans le corpus de test, le nombre de correspondance partielle est plus que celui des noms chinois. Cela décide une basse valeur de la précision. C'est majoritairement causé par la longue chaîne et la mauvaise influence de la polysémie des caractères de translittération au cours de la segmentation. En face de cette difficulté, nous analysons ensuite plus en détail les noms étrangers avec leurs longueurs et l'influence de la normalisation.

3.2 Noms étrangers avec 2 ou 3 caractères vs. Noms étrangers plus de 4 caractères

File name	TUs	Matched	G_part	T_part	G_total	T_total	P	R	F
1000-NomEtranger-2vs3	1003	58	1	3	75	87	0.667	0.773	0.716
1000-NomEtranger-4Plus	1003	30	0	8	51	50	0.600	0.588	0.594

Tableau 15 2 et 3 caractères vs. Plus de 4 caractères

En nous appuyant sur les noms étrangers dans le même corpus de 1000 lignes, nous avons constaté que, plus un nom est court, plus il est facile de l'identifier. Ayant une précision de 0.6 environ, le rappel de ces deux sous-catégories varie. Une partie des longues translittérations sont ignorées par le modèle. Ce qui cause cette différence c'est d'abord la difficulté intrinsèque des noms étrangers assez longs et ensuite que le post-traitement pour améliorer le rappel total n'applique pas aux entités plus de 5 caractères. Par conséquent, cette partie de noms devient un point aveugle.

3.3 Noms étrangers avec une normalisation vs. Noms étrangers sans normalisation

En dehors de distinguer la longueur limite des noms étrangers, nous avons analysé aussi le changement apporté par la normalisation des noms de translittération.

File name	TUs	Matched	G_part	T_part	G_total	T_total	P	R	F
1000-NomEtranger-Nomalise	1003	77	1	4	92	82	0.939	0.837	0.885
1000-NomEtranger-NonNomalise	1003	18	0	7	36	29	0.621	0.500	0.554

Tableau 16 Noms étrangers norm. vs. Noms étrangers non norm.

Dans le but de comparer seulement la performance des règles de normalisation, nous n'avons pris en considération que les noms étrangers en enlevant tous les faux positifs. En comparant

la différence de deux corpus annotés, nous avons constaté que plusieurs types de noms sont ajoutés :

3.3.1 Les prénoms isolés des indicateurs sont reconnus par la normalisation du dictionnaire de translittération

3.3.1.1 <希特勒> (<Hitler>)

3.3.1.2 <霍姆斯> (<Holmes>)

3.3.1.3 <弗朗西斯> 花 4万 美元 (<Francis> a depensé 40000 dollars.)

3.3.2 Noms complets liés avec un symbole de séparateur (· -._/)

3.3.2.1 前 中央 情报 局 局长 <迈克尔 · 海登> (Ancien directeur de la CIA <Michael Hayden>)

3.3.2.2 著名 裁判员 <阿兰 · 密尔斯> (Arbitre célèbre <Alan Mills>)

3.3.3 Noms étrangers en lettres

3.3.3.1 <michael hayden>

3.3.3.2 <kerry brinkert>

3.4 Noms ambigus vs. Noms non ambigus

Comme les idéogrammes sont généralement polysémiques et qu'il n'existe pas de lettre majuscule ni une limite pour le choix des idéogrammes, une personne est quelque fois ambigüe. Afin de prouver qu'une entité ambigüe va mal influencer la qualité de REN, nous avons effectué une comparaison de double sous-catégorie « Noms ambigus vs. Noms non ambigus ». Les critères de l' « ambiguïté » sont :

a) Les ENs précèdent ou suivent par un titre, une occupation ou un déclencheur, par exemple <习>主席 (président <XI>), <马>先生 (M. MA), 新华社 记者 <尹鸿祝> (le journaliste <YIN Hongzhu>) ou <奥拉拉 · 奥图诺> 发言 (<奥拉拉 · 奥图诺> prend la parole).
b) Les ENs ne font pas partie des autres catégories des entités nommées plus longues. Ex : <董存瑞>烈士陵园 (<Dong Cunrui> Cimetière des martyrs) ensemble serait une organisation.
c) Les idéogrammes des noms et surtout les prénoms possédant une forte ambiguïté sont utilisés plus souvent avec un autre sens. Par exemple,

a. la phrase <景 文辉> 现在 登封 的 一 家 工商银行 做 保安, le nom « 景 » est utilisé plus fréquemment au sens de « paysage ».
b. 特派 记者 <曾 向 荣> 摄影 报道, le nom « 曾 » est plutôt un marqueur du temps passé.

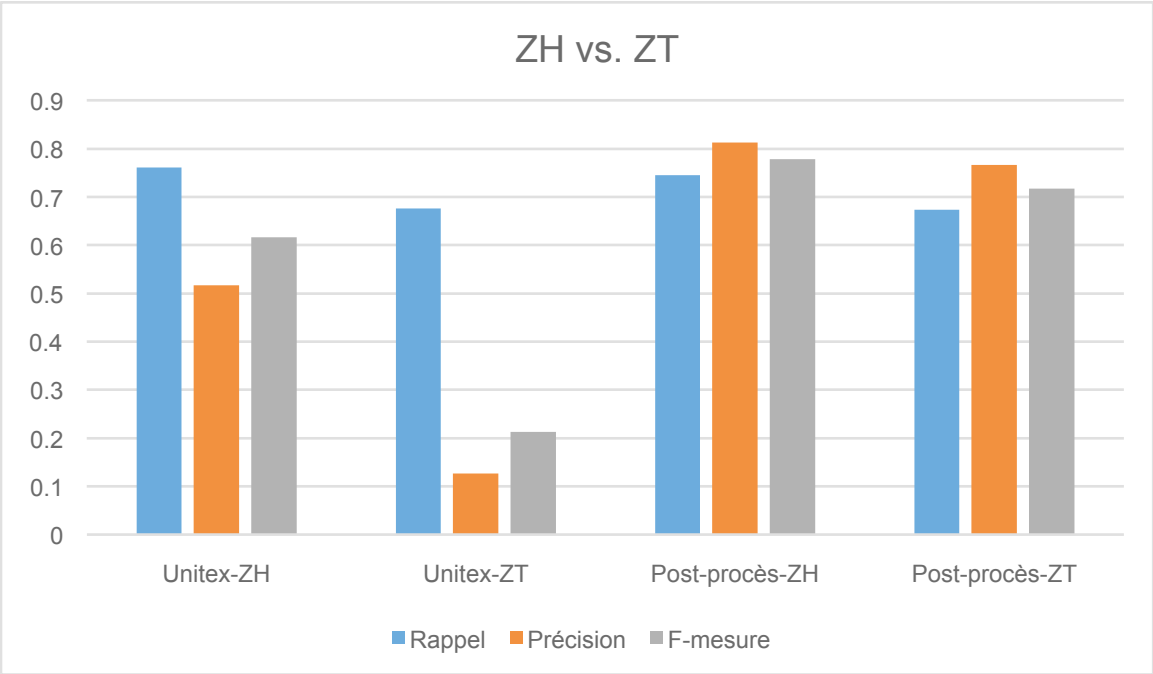
File name	TUs	Matched	G_part	T_part	G_total	T_total	P	R	F
1000-ambigu	1003	58	0	8	92	102	0.569	0.630	0.598
1000-NonAmbigu	1003	141	1	7	160	151	0.934	0.881	0.907

Tableau 17 Mots ambigus vs. Mots non ambigus

Visiblement, d'un part il y a moins d'ENs non ambiguës qui ne sont pas identifiées avec un rappel de 0.630 par rapport à 0.881. D'autre part les frontières et les étiquettes conforment à la configuration standard avec une précision de 0.934.

3.5 Chinois simplifié vs. Chinois traditionnel

Le chinois simplifié est utilisé en Chine continentale et à Singapour alors que le chinois traditionnel est utilisé à Hong Kong, Macao et Taiwan. Bien que les caractères du chinois simplifiés viennent de chinois traditionnel, la transformation entre les deux n'est pas simplement une traduction mot à mot [Xiaodong S. et al., 2013]. De même pour la REN, les règles, les dictionnaires ainsi que le modèle ne conviendront pas à ceux du chinois traditionnel. En base de cette hypothèse, nous avons comparé ces deux langues respectivement avec la sortie de l'Unitex et de post-traitement.



Les deux premiers groupes de la Figure 19 indiquent que les règles et les dictionnaires du chinois simplifié ne sont pas applicables pour le chinois traditionnel. Les 167/1345 (ENs trouvées correctes/ENs correctes) entités partagent les mêmes caractères du chinois simplifié. C'est en cette raison qu'elles sont correspondantes, comme par exemple <李宇春>, 記者 <徐江善>. Quand il rencontre un mot traditionnel, Unitex ne peut plus le repérer, comme <李登>輝.

Au cours du post-traitement, comme il ne prenait en compte que les étiquettes et la longueur des entités au lieu de chaque idéogramme, le modèle appliqué a amélioré la qualité de REN. En particulier, l'augmentation du rappel implique une amélioration de l'absence d'ENs de la dernière étape.

En face de cette comparaison, nous pouvons déjà conclure que le traitement du chinois simplifié est différent du chinois traditionnel, même s'ils partagent quelques points communs. Pendant la REN, les deux sortes de caractères distincts seront traitées différemment. La solution proposée peut être :

- Donner deux modèles pour chinois simplifié et traditionnel.

- Transformer d'abord le corpus traditionnel en chinois simplifié. Après le traitement, le retransformer en chinois traditionnel.

4 Perspectives

4.1 Limites de l'expérience

L'approche que nous avons utilisée satisfait plus au moins nos attentes de départ. Cependant il existe quand même des limites perceptibles.

Dans un premier temps, même si nous avons l'intention d'utiliser un modèle purement statistique, la construction du modèle n'a pas strictement respecté les algorithmes de CRF ni de HMM. Il est considéré plutôt comme un post-traitement à la base d'une approche de règles. Un autre obstacle nous semble être la quantité du corpus de développement contenant seulement 6000 lignes. L'établissement du corpus de référence demande un travail manuel. Il n'est pas possible d'éviter les erreurs d'annotation, ce qui conduit aux résultats imprécis. Il est peut-être plus pertinent d'emprunter les références existantes, mais dans ce cas la différence de définition des catégories doit être prise en considération. En outre, la conséquence du traitement monotone d'une catégorie « PERSON » est que la performance attendue de cette approche changera en fonction du changement des catégories ou de l'ajout des catégories. Nous nous interrogeons aussi sur le complet du dictionnaire, surtout pour les ENs non reconnues sans raison évidente.

4.2 Travaux futurs

À l'examen des expérimentations, nous trouvons plusieurs pistes potentielles pour la recherche future. Premièrement, étant donné que le travail s'applique seulement à la catégorie « person », il est nécessaire de compléter la tâche de REN pour d'autres catégories en tenant compte des caractéristiques de chaque catégorie. En appuyant sur le résultat de cette approche, la tâche complète de REN demande aussi une adaptation initiative. Par exemple, les étiquettes du contexte des noms et prénoms ne seront plus les mêmes pour celles des organisations. Idem pour les dictionnaires et les règles. En face des nouveaux problèmes, au lieu de se limiter à un support, nous nous ingénions à intégrer les connaissances des autres disciplines. Par exemple, la REN basée sur l'ontologie [Shi et al., 2009] ou sur un système multi-agents [Ye et al., 2002].

Premièrement, vu qu'une fausse segmentation du corpus enchaîne une suite d'erreurs, par exemple, des étiquettes grammaticales, de la segmentation des phrases, et que la mauvaise segmentation est partiellement causée par les mots inconnus dans un corpus, il est intéressant de développer des approches à identifier, extraire la néologie depuis les mini-blogs des

réseaux sociaux comme Facebook ou Twitter en chinois. Etant donné que la segmentation aide à la REN, nous pouvons, à l'inverse, bénéficier des sorties de REN pour segmenter un corpus. Dans ce cas-là, tenant compte des ENs, la séparation les longues ENs se réduira. Par ailleurs, les ENs reconnues aident à identifier les catégories grammaticales de leurs contextes.

Par ailleurs, la REN ne se contente pas d'avoir un score parfait. Un modèle de REN a besoin d'être appliquée à la pratique du TAL. La REN peut se combiner avec la traduction automatique, la reconnaissance des nouveaux mots sur l'internet et l'extraction d'information dans différents domaines, etc.

Conclusion

Au cours de ce travail de REN, nous avons essayé de trouver une approche d'annotation pertinente et assez générale sur la catégorie « PERSON ». Les résultats de chaque phase correspondent majoritairement à nos attentes au départ. D'après ces différentes expérimentations et résultats préliminaires, nous pouvons conclure que, dans un premier temps, la REN du chinois, au moins les ENs de personnes dans le corpus chinois, a besoin de segmentation avant tout traitement ultérieur. Ensuite, vue des caractéristiques intrinsèques du chinois, une meilleure approche serait plutôt hybride dans laquelle les règles, dictionnaires et modèles d'apprentissage a tous un rôle indispensable. L'approche hybride utilisée pendant le travail est de combiner une approche basée sur les caractéristiques linguistiques et une autre basée sur les données statistiques d'expérimentation. C'est en exploitant les avantages de chaque méthode que nous obtiendrons une résolution à la fois générale et efficace.

Nous estimons que l'approche utilisée répond essentiellement aux exigences de REN. Cependant il existe aussi des limites. Au un premier lieu, par rapport aux résultats des recherches présentées dans l'état de l'art, nos résultats sont modestes. Autrement dit cette approche ne convient pas à toutes les ENs personne. Par la suite, le modèle de post-traitement n'est pas strictement un modèle statistique, mais seulement inspiré par les principes des théories statistiques. Enfin, la quantité de corpus est assez petite. En face de limites et réflexions des travaux futurs, nous envisageons de commencer à améliorer cette approche par la segmentation en ajoutant les néologismes dans la base de données. Ensuite, un « vrai » modèle d'apprentissage sera établi en élargissant la taille du corpus de développement.

Dans les perspectives, cette approche de REN se porte aussi sur les autres catégories. Les règles et les modèles changeront avec un ajustement pour les cas particuliers de chaque catégorie. Sachant que la segmentation est la base de tout traitement automatique du chinois, c'est nécessaire d'évoluer la base de données en ajoutant les nouveaux lexiques apparus, surtout la néologie venant de l'Internet. Pour ce faire, il est aussi intéressant de chercher les méthodes d'identifier, extraire et implémenter ces mots à partir des réseaux sociaux comme RenRen (facebook en chinois) et Weibo (Twitter en chinois), etc. L'approche de REN fait une partie indispensable dans ce genre de recherches.

Bibliographies

[Altenbek, 2005] Gulila Altenbek. Rule-based Person Name Recognition for Xinjiang Minority Languages. (2005). In Journal of Chinese Language and Computing Vol.15, No.4, page 219-226.

[Bikel et al., 1999] Daniel M. Bikel, Richard Schwartz, Ralph M. Weischedel. (1999). An Algorithm that Learns What's in a Name. In Machine Learning Vol.34, pages 211-231.

[Charnois et al., 2009] Thierry Charnois, Marc Plantevit, Christophe Rigotti, et Bruno Crémilleux. (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. In TAL. Vol. 50, No.3, pages 59-87

[Chou et al., 2004] W. Chou,Y. Lin, T. Tsai, K. Wu, T. Sung and W. Hsu. A Maximum Entropy Approach to Biomedical Named Entity Recognition. (2004). In Proceeding of the 4th Workshop on Data Mining in Bioinformatics, pages 56-61.

[Coates-Stephens, 1992] Sam Coates-Stephens. (1992). The Analysis and Acquisition of Proper Names for the Understanding of Free Text. In Computers and the Humanities, Vol.26, pages 441-456.

[Dutrey et al., 2012] Camille Dutrey, Chloé Clavel, Sophie Rosset, Ioana Vasilescu, et Martine Adda-Decker. (2012). Quel est l'apport de la détection d'entités nommées pourl'extraction d'information en domaine restreint ? JEP-TALN-RECITAL, vol.2, pages 359–366.

[Ezzat, 2010] Mani EZZAT. (2010). Acquisition de grammaires locales pour l'extraction de relations entre entités nommées. In Conférence TALN 2010, Montréal, Canada, juillet 2010.

[Fleischman, 2001] Michael Fleischman. (2001). Automated subcategorization of named entities. In Proc. of the ACL 2001 Student Research Workshop, pages 25–30.

[Fleischman et Hovy, 2002] Michael Fleischman and Eduard Hovy. (2002). Fine grained classification of named entities. In Proc. Of COLING, vol.1, pages 1–7..

[Gaio et al., 2012] Mauro GAIO, Christian SALLABERRY, et Van Tien NGUYEN. (2012). Typage de noms toponymiques à des fins d'indexation géographique. In TAL Vol.53, No.2, pages 143-176.

[Grouin et al., 2011] Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. (2011). Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In Proceedings of the Fifth Law Workshop (LAW V), pages 92-100.

[Han et al., 2004] Tzong-Han Tsai, Shih-Hung Wu, Cheng-Wei Lee, Cheng-Wei Shih, and Wen-Lian Hsu. (2004). Mencius: A Chinese Named Entity Recognizer Using Maximum Entropy-based Hybrid Model. In Computational Linguistics and Chinese Language Processing, Vol. 9, No.1, page 65-82.

[Hu et al., 2013] Hu Yimin, Song Liangtu, Chen Pen, Wei Yuanyuan, Su Yaru. (2013). Chinese Geographic Entity Resolution Based on Markov Logic Network. In PR & AI, Vol.26, No.1, pages 114-122

[Lee et al., 2005] Seungwoo Lee and Gary Geunbae Lee. (2005). Heuristic methods for reducing errors of geographic named entities learned by bootstrapping. In IJCNLP, pages 658–669.

[Li et al., 2010] Lishuang Li, Zezhong Li, Zhuoye Ding and Degen Huang. (2010). A Hybrid Model Combining CRF with Boundary Templates for Chinese Person Name Recognition. In International Journal of Advanced Intelligence Vol.2, No.1, pages 73-80.

[Lv et al., 2006] LV Xueqiang,YU HongKui, ZHANG Huaping, LIU Qun, and SHI Shuicai. (2006). Chinese named entity identification using cascaded hidden Markov model. In Journal on Communications, Vol.27, No.2, p87-94.

[McNamee, et al., 2011] Paul McNamee, James C. Mayfield, and Christine D. Piatko. (2011). Processing Named Entities in Text. Johns Hopkins APL Technical Digest, Vol.30, No.1, pages 31-40.

[[Ni et al., 2011] Ni Ji, Kong Fang, Zhu Qiaoming, and Li Peifeng. (2011). Research on Chinese Name Recognition Base on Trustworthiness. In JOURNAL OF CHINESE

INFORMATION PROCESSING, Vol. 25, No. 3, pages 45-50.

[Nouvel et al., 2013] Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, Arnaud Soulet. (2013). Fouille de règles d'annotation partielles pour la reconnaissance des entités nommées. In TALN-RÉCITAL 2013, 17-21 Juin, Les Sables d'Olonne

[Rau, 1991] Lisa F. Rau. 1991. Extracting Company Names from Text. In Proc. (1991). IN Conference on Artificial Intelligence Applications of IEEE.

[Shi et al., 2009] Shi Shumin, Feng Chong, Huang Heyan, and Liu Dongsheng. (2009). Recognition of Chinese Domain Named Entities Based on Ontology. In Journal of the China Society and Technical Information, vol. 28, No.6, page 857-863.

[Sun et al., 2010] Sun Zhen and Wang Huilin. Overview on the Advanced if the Reasearch on Named Enetity Recognition. (2010). In New Technology of Library and Information Service vol.6, pages 42-47.

[Wang et Shi, 2005] Wang Houfeng and Shi Wuguang. (2005). A Simple Rule-based Approach to Organization Name Recognition in Chinese Text. In CICLing, LNCS 3406, pages 769-777.

[Xiaodong S. et al., 2013] Xiaodong Shi, Yidong Chen and Xiuping Huang. (2013). Key Problems in Conversion from Simplified to Traditional Chinese Characters. In Proceedings of the XIV Machine Translation Summit, pages 287–293.

[Ye et al., 2002] Shiren Ye, Tat-Seng Chua, Liu Jimin. (2002). An Agent-based Approach to Chinese Named Entity Recognition. School of Computing, National University of Singapore.

[Yu et al., 1998] Shihong Yu, Shuanhu Bai and Paul Wu. (1998). DESCRIPTION OF THE KENT RIDGE DIGITAL LABS SYSTEM USED FOR MUC-7. Proceedings of the Seventh Message Understanding Conference (MUC-7).

[Zhang et al., 2006] Zhang Xiaoyan, Wang Ting, Chen Huo-wang. (2006). A Mixed StatisticaI ModeI-based Method for Chinese Named Entity Recognition. In COMPUTER ENGINEERING & SCIENCE, Vol.28, No.6, pages 135-139

Annexes

1. Langues pour le projet NER de SYSTRAN

Language	Difficulty	Type
English	easy	level one
French	easy	level one
Spanish	easy	level one
Portuguese	easy	level one
German	hard	level one
Italian	easy	level one
Russian	medium	level one (USG)
Chinese	hard	level one (USG)
Somali	medium	level one (Mindef)
Korean	medium	level two
Japanese	hard	level two
Dutch	easy	level two
Arabic	hard	level two
Swedish	easy	level two
Polish	medium	level two
Greek	medium	level two
Romanian	medium	level two
Turkish	medium	level two
Slovenian	medium	level two

2. Corpus d’annotation

2.1 Corpus sans prétraitement (Annotation d’Unitex)

它包括摩尔法、<ENAMEX type="person">伏尔哈德</ENAMEX>法、法也斯法。{S}

(<title>记者</title><ENAMEX type="person">徐江善</ENAMEX>、<title>通讯员</title><ENAMEX type="person">常玉礼</ENAMEX>)科研可否“借船出海”?{S}

老百姓们忍无可忍,所以才有陈胜<ENAMEX type="person">吴广</ENAMEX>起义,才有<ENAMEX type="person">刘邦</ENAMEX>入关告父老书曰：{S}

卢现祥{S}

<ENAMEX type="person">利马索尔</ENAMEX>处{S}

当天,数十万朝鲜民众聚集在平壤<ENAMEX type="person">金日成</ENAMEX>广场,<TIMEX type="time"><time>中午12时</time></TIMEX>,朝鲜各地民众默哀三分钟,平壤和各道首府所在地以鸣炮、车船鸣笛等方式志哀。{S}

<ENAMEX type="organization">阿根廷航空</ENAMEX>(<ENAMEX type="organization">ar</ENAMEX>)(布宜<ENAMEX type="person">诺斯艾利斯</ENAMEX>、<ENAMEX type="location">马德里</ENAMEX>){S}

特别值得一提的是内地部分的提名艺人,<ENAMEX type="person">李宇春</ENAMEX>(听歌blog)、花儿乐队(听歌blog)、汪峰(听歌blog)和<ENAMEX type="person">杨坤</ENAMEX>(听歌blog),风格迥然不同。{S}

前中央情报局<title>局长</title><ENAMEX type="person">迈克尔·海登</ENAMEX>(<ENAMEX type="person">michaelhayden</ENAMEX>)表示,美国的网络制止通过网络窃取他人财产,但鼓励言论自由。{S}

<ENAMEX type="person">王立军</ENAMEX>收下了这珍贵的礼物,他双手端碗,将水一饮而尽。{S}

他曾向公司<title>副总</title><ENAMEX type="person">张光明</ENAMEX>问过此事,<ENAMEX type="person">张光明</ENAMEX>告诉他,“集团数目不多,老板自己就不好说了。{S}

龟兹石窟研究所研究室副主任台来提告诉记者,克孜尔千佛洞壁画的状况一年不如一年,需要尽快采取保护措施,把最好的美术、壁画保存下来,给子<ENAMEX type="person">孙</ENAMEX><title>后代</title>留下一些佛教艺术精品。{S}

盖茨与艾伦创办了企业。{S}

<ENAMEX type="person">温家宝</ENAMEX><title>总理</title>在部长级会议开幕式上提出了新形势下发展<ENAMEX type="location">中</ENAMEX><ENAMEX type="location">非</ENAMEX>关系的<TIMEX type="time"><time>四点</time></TIMEX>建议：{S}

毛<ENAMEX type="person">里塔尼亚</ENAMEX>军人政变后成立的民主与公正军事委员会四日宣布解散国民议会,但决定维持国家宪法并将军事委员会宪章添加到宪法中。{S}

曾经在温网比赛中执法了23年的着名裁判员阿兰·密尔斯对温网的雨有着特殊的情怀,对于这位有着“雨人”称号的裁判员来说,温网迫于电视转播等各方面的压力,必须在<TIMEX type="time"><date>2009年</date></TIMEX>安上活动顶棚的事实绝对是一种遗憾。{S}

专家评<ENAMEX type="person">陈水扁</ENAMEX>过境纽约{S}

到了<TIMEX type="time"><time>下午二点</time></TIMEX>以后,<ENAMEX type="location">南京城</ENAMEX>区的雨势逐渐变小,傍晚前后雨就基本上停了。{S}

对江津和<ENAMEX type="person">区楚良</ENAMEX>两人的精彩动作,<ENAMEX type="person">王建英</ENAMEX>则鼓掌以示嘉奖。{S}

高科技股<ENAMEX type="person">纳斯达克</ENAMEX>指数也上涨<TIMEX type="time"><time>六点五</time></TIMEX>五点,来到二千三百<TIMEX type="time"><time>六十三点八</time></TIMEX>四点,小升百分之<TIMEX type="time"><time>零点二</time></TIMEX>八。{S}

当然我们更不会忘了<ENAMEX type="person">雷锋</ENAMEX>和顺姬这两个有着不幸童年的苦孩子,没有他们,我们那次的作文就不会得优,姐姐也不会那么露脸,更重要的是,他们让我们觉得自己的童年真是幸福无比。{S}

2.2 Corpus avec segmentation (annotation d'Unitex)

它 包括 摩尔 法 、 <ENAMEX type="person">伏尔 哈德</ENAMEX> 法 、 法 也斯 法 。{S}

(<title>记者</title> <ENAMEX type="person">徐江善</ENAMEX> 、 <title>通讯员</title> <ENAMEX type="person">常玉礼</ENAMEX>) 科研 可否 " 借 船 出海 " ?{S}

老百姓 们 忍无可忍 , 所以 才 有 <ENAMEX type="person"><lastname>陈</lastname> <firstname>胜</firstname></ENAMEX> <ENAMEX type="person"><lastname>吴 </lastname><firstname>广</firstname></ENAMEX> 起义 , 才 有 <ENAMEX type="person"><lastname>刘</lastname> <firstname>邦</firstname></ENAMEX> 入关 告 父老书 曰 : {S}

卢现祥{S}

<ENAMEX type="location">利马索尔</ENAMEX> 处{S}

当天 , 数十万 <ENAMEX type="location">朝鲜</ENAMEX> 民众 聚集 在 <ENAMEX type="organization">平壤 金日成广场</ENAMEX> , <TIMEX type="time"><time>中午 12 时 </time></TIMEX> , <ENAMEX type="location">朝鲜</ENAMEX> 各地 民众 默哀 三 分钟 , <ENAMEX type="location">平壤</ENAMEX> 和 各 道 首府 所在地 以 鸣炮 、 车船 鸣笛 等 方式 志哀 。{S}

<ENAMEX type="organization">阿根廷航空</ENAMEX> (<ENAMEX type="organization">ar</ENAMEX>) (<ENAMEX type="location">布宜诺斯艾利斯 </ENAMEX> 、 <ENAMEX type="location">马德里</ENAMEX>){S}

特别 值得一提 的 是 内地 部分 的 提名 艺人 , <ENAMEX type="person"><lastname>李 </lastname><firstname>宇春</firstname></ENAMEX> (听歌 blog) 、 花儿乐队 (听歌 blog) 、 汪 峰 (听歌 blog) 和 <ENAMEX type="person"><lastname>杨</lastname><firstname>坤 </firstname></ENAMEX> (听歌 blog) , 风格 迥然不同 。{S}

前 <ENAMEX type="organization">中央 情报 局</ENAMEX> <title>局长</title> <ENAMEX type="person">迈克尔 · 海登</ENAMEX> (<ENAMEX type="person">michael hayden</ENAMEX>) 表示 , <ENAMEX type="location">美国</ENAMEX> 的 网络 制止 通过 网络 窃取 他人 财产 , 但 鼓励 言论 自由 。{S}

<ENAMEX type="person"><lastname>王</lastname><firstname>立军 </firstname></ENAMEX> 收下 了 这 珍贵 的 礼物 , 他 双 手端 碗 , 将 水 一饮而尽 。{S}

他 曾 向 公司 <title>副总</title> <ENAMEX type="person">张光明</ENAMEX> 问 过 此事 , <ENAMEX type="person"><lastname>张</lastname><firstname>光明 </firstname></ENAMEX> 告诉 他 , " 集团 数目 不 多 , 老板 自己 就 不 好说 了 。{S}

<ENAMEX type="location">龟兹</ENAMEX> 石窟 研究所 研究 室 副主任 台 来 提 告诉 记者 , <ENAMEX type="location">克孜尔 千佛洞</ENAMEX> 壁画 的 状况 一 年 不如 一 年 , 需要 尽快 采取 保护 措施 , 把 最 好 的 美术 、 壁画 保存 下来 , 给 子<ENAMEX type="person">孙 </ENAMEX> <title>后代</title> 留下 一些 佛教 艺术 精品 。{S}

盖茨 与 艾伦 创办 了 企业 。{S}

<ENAMEX type="person">温家宝</ENAMEX> <title>总理</title> 在 部长 级 会议 开幕式 上 提出 了 新 形势 下 发展 <ENAMEX type="location">中非</ENAMEX> 关系 的 <TIMEX type="time"><time>四 点</time></TIMEX> 建议 ： {S}

<ENAMEX type="location">毛里塔尼亚</ENAMEX> 军人 政变 后 成立 的 民主 与 公正 军事 委员 会 四 日 宣布 解散 国民 议会 , 但 决定 维持 国家 宪法 并 将 军事 委员 会 宪章 添加 到 宪法 中 。{S}

曾经 在 温网 比赛 中 执法 了 23 年 的 着名 裁判员 <ENAMEX type="person">阿兰 · 密尔斯 </ENAMEX> 对 温网 的 雨 有着 特殊 的 情怀 , 对于 这 位 有着 “ 雨 人 ” 称号 的 裁判员 来说 , 温网 迫于 电视 转播 等 各 方面 的 压力 , 必须 在 <TIMEX type="time"><date>2009 年 </date></TIMEX> 安上 活动 顶棚 的 事实 绝对 是 一 种 遗憾 。{S}

专家 评 <ENAMEX type="person">陈水扁</ENAMEX> 过境 纽约{S}

对 <ENAMEX type="location">江津</ENAMEX> 和 <ENAMEX type="person">区楚良 </ENAMEX> 两人 的 精彩 动作 , <ENAMEX type="person"><lastname>王 </lastname><firstname>建英</firstname></ENAMEX> 则 鼓掌 以示 嘉奖 。{S}

高 科技 股 <ENAMEX type="organization">纳斯达克</ENAMEX> 指数 也 上涨 <TIMEX type="time"><time>六点五</time></TIMEX>五 点 , 来到 二 千 三百<TIMEX type="time"><time>六十三点八</time></TIMEX>四 点 , 小 升 百分之<TIMEX type="time"><time>零点二</time></TIMEX>八 。{S}

当然 我们 更 不 会 忘 了 <ENAMEX type="person">雷锋</ENAMEX> 和 顺姬 这 两 个 有着 不幸 童年 的 苦 孩子 , 没有 他们 , 我们 那 次 的 作文 就 不 会 得 优 , 姐姐 也 不 会 那么 露脸 , 更 重要 的 是 , 他们 让 我们 觉得 自己 的 童年 真是 幸福 无比 。{S}

2.3 Post-traitement

它 包括 摩尔 法 、 伏尔 哈德 法 、 <法 也斯 法> 。

(记者 <徐江善> 、 通讯员 <常玉礼>) 科研 可否 “ 借 船 出海 ” ?

老百姓 们 忍无可忍 , 所以 才 有 <陈 胜> <吴广> 起义 , 才 有 <刘 邦> 入关 告 父 老书 曰 :

<卢现祥>

利马索尔 处

当天 , 数十万 朝鲜 民众 聚集 在 平壤 金日成广场 , 中午 12 时 , 朝鲜 各地 民众 默哀 三 分钟 , 平壤 和 各 道 首府 所在地 以 鸣炮 、 车船 鸣笛 等 方式 志哀 。

阿根廷航空 (ar) (布宜诺斯艾利斯 、 马德里)

特别 值得一提 的 是 内地 部分 的 提名 艺人 , <李宇春> (听歌 blog) 、 花儿乐队 (听歌 blog) 、 <汪 峰> (听歌 blog) 和 <杨坤> (听歌 blog) , 风格 迥然不同 。

前 中央 情报 局 局长 <迈克尔 · 海登> (<michael hayden>) 表示 , 美国 的 网络 制止 通过 网络 窃取 他人 财产 , 但 鼓励 言论 自由 。

<王立军> 收下 了 这 珍贵 的 礼物 , 他 双 手端 碗 , 将 水 一饮而尽 。

他 曾 向 公司 副总 <张光明> 问 过 此事 , <张光明> 告诉 他 , “ 集团 数目 不 多 , 老板 自己 就 不 好说 了 。

龟兹 石窟 研究所 研究 室 副主任 台 来 提 告诉 记者 , 克孜尔 千佛洞 壁画 的 状况 一 年 不如 一 年 , 需要 尽快 采取 保护 措施 , 把 最 好 的 美术 、 壁画 保存 下来 , 给 子孙 后代 留下 一些 佛教 艺术 精品 。

<盖茨> 与 <艾伦> 创办 了 企业 。

<温家宝> 总理 在 部长 级 会议 开幕式 上 提出 了 新 形势 下 发展 中非 关系 的 四 点 建议 :

毛里塔尼亚 军人 政变 后 成立 的 民主 与 公正 军事 委员 会 四 日 宣布 解散 国民 议会 , 但 决定 维持 国家 宪法 并 将 军事 委员 会 宪章 添加 到 宪法 中 。

曾经 在 温网 比赛 中 执法 了 23 年 的 着名 裁判员 <阿兰 · 密尔斯> 对 温网 的 雨 有着 特殊 的 情怀 , 对于 这 位 有着 “ 雨 人 ” 称号 的 裁判员 来说 , 温网 迫于 电视 转播 等 各 方面 的 压力 , 必须 在 2009 年 安上 活动 顶棚 的 事实 绝对 是 一 种 遗憾 。

专家 评 <陈水扁> 过境 纽约

对 江津 和 <区楚良> 两人 的 精彩 动作 , <王建英> 则 鼓掌 以示 嘉奖 。

高 科技 股 纳斯达克 指数 也 上涨 六点五五 点 , 来到 二 千 三百六十三点八四 点 , 小 升 百分 之零点二八 。

当然 我们 更 不 会 忘 了 <雷锋> 和 顺姬 这 两 个 有着 不幸 童年 的 苦 孩子 , 没有 他们 , 我们 那 次 的 作文 就 不 会 得 优 , 姐姐 也 不 会 那么 露脸 , 更 重要 的 是 , 他们 让 我们 觉得 自己 的 童年 真是 幸福 无比 。

2.4 Référence

它 包括 摩尔 法 、 伏尔 哈德 法 、 法 也斯 法 。

(记者 <徐江善> 、 通讯员 <常玉礼>) 科研 可否 “ 借 船 出海 ” ?

老百姓 们 忍无可忍 , 所以 才 有 <陈 胜> <吴广> 起义 , 才 有 <刘 邦> 入关 告 父 老书 曰 :

<卢现祥>

利马索尔 处

当天 , 数十万 朝鲜 民众 聚集 在 平壤 金日成广场 , 中午 12 时 , 朝鲜 各地 民众 默哀 三 分钟 , 平壤 和 各 道 首府 所在地 以 鸣炮 、 车船 鸣笛 等 方式 志哀 。

阿根廷航空 (ar) (布宜诺斯艾利斯 、 马德里)

特别 值得一提 的 是 内地 部分 的 提名 艺人 , <李宇春> (听歌 blog) 、 花儿乐队 (听歌 blog) 、 <汪 峰> (听歌 blog) 和 <杨坤> (听歌 blog) , 风格 迥然不同 。

前 中央 情报 局 局长 <迈克尔 · 海登> (<michael hayden>) 表示 , 美国 的 网络 制止 通过 网络 窃取 他人 财产 , 但 鼓励 言论 自由 。

<王立军> 收下 了 这 珍贵 的 礼物 , 他 双 手端 碗 , 将 水 一饮而尽 。

他 曾 向 公司 副总 <张光明> 问 过 此事 , <张光明> 告诉 他 , “ 集团 数目 不 多 , 老板 自己 就 不 好说 了 。

还是 他 不 符合 今天 日本 社会 对 " 王者 的 要求 " 。

龟兹 石窟 研究所 研究 室 副主任 <台 来 提> 告诉 记者 , 克孜尔 千佛洞 壁画 的 状况 一 年 不 如 一 年 , 需要 尽快 采取 保护 措施 , 把 最 好 的 美术 、 壁画 保存 下来 , 给 子孙 后代 留下 一些 佛教 艺术 精品 。

<盖茨> 与 <艾伦> 创办 了 企业 。

<温家宝> 总理 在 部长 级 会议 开幕式 上 提出 了 新 形势 下 发展 中非 关系 的 四 点 建议 :

曾经 在 温网 比赛 中 执法 了 23 年 的 着名 裁判员 <阿兰 · 密尔斯> 对 温网 的 雨 有着 特殊 的 情怀 , 对于 这 位 有着 “ 雨 人 ” 称号 的 裁判员 来说 , 温网 迫于 电视 转播 等 各 方面 的 压力 , 必须 在 2009 年 安上 活动 顶棚 的 事实 绝对 是 一 种 遗憾 。

专家 评 <陈水扁> 过境 纽约

对 <江津> 和 <区楚良> 两人 的 精彩 动作 , <王建英> 则 鼓掌 以示 嘉奖 。

高 科技 股 纳斯达克 指数 也 上涨 六点五五 点 , 来到 二 千 三百六十三点八四 点 , 小 升 百分之零点二八 。

当然 我们 更 不 会 忘 了 <雷锋> 和 <顺姬> 这 两 个 有着 不幸 童年 的 苦 孩子 , 没有 他们 , 我们 那 次 的 作文 就 不 会 得 优 , 姐姐 也 不 会 那么 露脸 , 更 重要 的 是 , 他们 让 我们 觉得 自己 的 童年 真是 幸福 无比 。

3. Scripts

3.1 renameTag.pl : Unifier les balises dans les sorties différentes

```
# ================================================================================================
# renameTags.pl  --  replaces the tag names to the desired ones (specified in the input_file)
#                    and removes all the other tags, not listed in the input file
#                    So, the replacement file serves two purposes:
#                    replacement of category names and declaration of categories to be validated
# ================================================================================================

#Note1. The script works fine with nested tags
#Note2. The script does not check for correspondence of open and close tag names, it assumes that the close tag corresponds to the latest open

#!/usr/bin/perl;

use Getopt::Std;
use strict;
use warnings;

my $USAGE = "
  $0 [-h] -i <input_filename.txt> -r <tags_to_replace.txt> -o <output_filename.txt>

  tags_to_replace.txt file has the following format:
  <tag_to_replace_from><tabulation><tag_to_replace_to>

  example: 1.renameTags.pl -i tagged_text.txt -r list_of_taggs_to_replace.txt -o filtered_text.txt
           1.renameTags.pl -h  for help

  ";

use vars qw ($opt_h $opt_i $opt_r $opt_o);
getopts ('i:o:r:h:');

die $USAGE if $opt_h;

die "Error in command line:  -i not defined$USAGE" unless defined $opt_i;
die "Error in command line:  -o not defined$USAGE" unless defined $opt_o;

my ($infile,$outfile,$replfile,$features,@tags,@items,@stack,$tg,$txt,$nt,$content,%tagMap,$key);

main();

sub main
{
  %tagMap = ();
  if (defined $opt_r){
    open(REPLACE, $opt_r) || die "Can't open file $opt_r: $!";
   while (<REPLACE>){
     chop;
     next if /^$/;
```

```
    @items = split(/\t/,$_);
    $tagMap{$items[0]} = $items[1];
  }
  close REPLACE;
}

open INPUT, $opt_i || die "Can't open file $opt_i: $!";
open OUTPUT, ">".$opt_o || die "Can't open file $opt_o: $!";

while (my $line = <INPUT>) {
  chomp $line;
    if ($line =~ /^\s*$/){
      print "Empty line\n";
      next;
    }
  $line =~ s/^\[.*?\] //;
  $line =~ s/{S}//g;
  @stack = ();
  while ($line =~ /^([^<]*)<(.*?)>(.*$)/){
    print OUTPUT $1;
    $tg = $2;
    $line = $3;
    if ($tg =~ /^\/.*$/){
      next if (scalar(@stack)<1);
      my $tg1 = pop(@stack);
      if ($tg1 ne " "){
        print OUTPUT ">";
      }
    }
    else{
      my $newtg = $tagMap{$tg};
      if (defined($newtg)){
        print OUTPUT $newtg;
        push @stack,$newtg;
      }
      else{
        push @stack," ";
      }
    }
  }
  print OUTPUT $line."\n";
}
close OUTPUT;
close INPUT;
}
```

3.2 countMatches.pl : évaluation et calcul des scores

```
# ================================================================================
# 2.countMatches.pl  -- calculates presicion, recall and F-score of NER
#                       using two input files: golden data and test file.
#                       Outputs all the matches, unmatches and partial matches
#                       of tags
# ================================================================================

#!/usr/bin/perl;
use strict;
use warnings;
use Getopt::Std;
use utf8;

my $USAGE = "
  $0 [-h] -g <goldendata_filename.txt> -t <testdata_filename.txt> [-o <output_filename.txt>]
  example: 2.countMatches.pl -g time.basis -t time.systran -o time.compare
           2.countMatches.pl -h  for help
  ";

use vars qw ($opt_h $opt_g $opt_t $opt_o);
getopts ('g:t:o:h:');

die $USAGE if $opt_h;

die "Error in command line:  -i not defined$USAGE" unless defined $opt_g;
die "Error in command line:  -o not defined$USAGE" unless defined $opt_t;

if ($opt_o) {
    open(OUT,">:utf8",$opt_o) || die "Can't open file $opt_o: $!";
} else {
   open(OUT, '>&', \*STDOUT) || die "Can't open STDOUT: $!";
   binmode( STDOUT, 'utf8:' );
}

open(G,"<:encoding(utf-8)",$opt_g) || die "Can't open file $opt_g: $!";
open(T,"<:encoding(utf-8)",$opt_t) || die "Can't open file $opt_t: $!";

my ($tot_tp,$tot_fn,$tot_fp,$tot_t,$tot_g,$gold,$test,$g_true_pos,$false_neg,$t_true_pos,$false_pos,$cond,$count,$pmatchgold,$pmatchtest,$tot_pmatchgold,$tot_pmatchtest);

$tot_tp = $tot_fn = $tot_fp = $tot_t = $tot_g = $count =
$pmatchgold = $pmatchtest = $tot_pmatchgold = $tot_pmatchtest = 0;
while ($gold = <G>) {
  $test = <T>;
  $count++;
  chomp $gold;
  chomp $test;
  print OUT "Gold:\t$gold\nTest:\t$test\n";
  ($g_true_pos,$false_neg,$t_true_pos,$false_pos,$pmatchgold,$pmatchtest) = matchUp ($gold, $test);
  print STDERR "True positives are not equal:\n$gold\n$test\nGold true pos:$g_true_pos\tTest true pos: $t_true_pos" if ($g_true_pos ne $t_true_pos);
```

```
  print OUT "True pos: $g_true_pos\tFalse pos: $false_pos\tFalse neg: $false_neg\n\n";

  $tot_tp += $g_true_pos;
  $tot_fn += $false_neg;
  $tot_fp += $false_pos;
  $tot_pmatchgold += $pmatchgold;
  $tot_pmatchtest += $pmatchtest;
  $tot_t += $t_true_pos+$false_pos;
  $tot_g += $g_true_pos+$false_neg;
}
close G;
close T;

my $precision = $tot_tp / $tot_t;
my $recall = $tot_tp / $tot_g;
my $f1 = (2 * $precision * $recall) / ($precision + $recall);

my $prec = sprintf '%.3f', $precision;
my $rec = sprintf '%.3f', $recall;
my $fscore = sprintf '%.3f', $f1;

print OUT "\nFile name\tTUs\tMatched_tags\tPart_matched_gold\tPart_matched_test\tTotal_gold_tags\tTotal_test_tags\tPresicion\tRecall\tF1\n";
print OUT "\n$opt_g\t$count\t$tot_tp\t$tot_pmatchgold\t$tot_pmatchtest\t$tot_g\t$tot_t\t$prec\t$rec\t$fscore\n";


sub matchUp
{
  my ($gold, $test) = @_;
  $gold =~ s/ //g;
  $test =~  s/ //g;
  my ($totalg,$totalt,$foundg,$foundt,$missedg,$missedt,$pmgold,$pmtest);
  $totalg =  $totalt = $foundg = $foundt = $missedg = $missedt = $pmgold = $pmtest = 0;
  my %gtags = ();
  my %ttags = ();
  while ($gold =~ s/(<.*?>.*?<\/.*?>)//) {
    $gtags{$1}++;
    $totalg++;
  }
  while ($test =~ s/(<.*?>.*?<\/.*?>)//) {
    $ttags{$1}++;
    $totalt++;
  }
  foreach my $key(sort keys %gtags){
    if (!exists $ttags{$key}){
      $missedg += $gtags{$key};
      print OUT "unmatched:\tgold: $key\ttest:\n";
      next;
    }
```

```
    if ($gtags{$key} > $ttags{$key}){
      $missedg += $gtags{$key} - $ttags{$key};
      $foundg += $ttags{$key};
      print OUT "unmatched:\tgold: $key\ttest:\n";
      next;
    }
    $foundg += $gtags{$key};
    print OUT "matched:\t$key\n";
  }
  foreach my $key(sort keys %ttags){
    if (!exists $gtags{$key}){
      $missedt += $ttags{$key};
      print OUT "unmatched:\tgold: \ttest: $key\n";
      next;
    }
    if ($ttags{$key} > $gtags{$key}){
      $missedt += $ttags{$key} - $gtags{$key};
      $foundt +=$gtags{$key};
      print OUT "unmatched:\tgold: \ttest: $key\n";
      next;
    }
    $foundt += $ttags{$key};
  }
  if ($missedg or $missedt){
    foreach my $key1(keys %gtags){
      $key1 =~ /(<.*?>)(.*?)(<\/.*?>)/;
      my $gtg = $1;
      my $gcnt = $2;
      foreach my $key2(keys %ttags){
        $key2 =~ /(<.*?>)(.*?)(<\/.*?>)/;
        my $ttg = $1;
        my $tcnt = $2;
        if (($gtg eq $ttg) and ($gcnt ne $tcnt)){
          if (index($gcnt, $tcnt) ne -1){
            print OUT "partial matched:\tgold: $key1\ttest: $key2\n";
            $pmtest++;
          }
          if (index($tcnt, $gcnt) ne -1){
            print OUT "partial matched:\tgold: $key1\ttest: $key2\n";
            $pmgold++;
          }
        }
      }
    }
  }
  return ($foundg,$missedg,$foundt,$missedt,$pmgold,$pmtest);
}
```

3.3 calcul_pos_frequence_categorise.pl : catégoriser les étiquettes du contexte

```perl
#author: YAN Liyun
#!/usr/bin/perl

use strict;
use warnings;
use utf8;
use Data::Dumper;
binmode STDIN, ':utf8';
binmode STDOUT, ':utf8';
binmode STDERR, ':utf8';

my (@parts, %hash_pre, %hash_post, %hash_actuel, %all_pos, $token_actuel, $pos_actuel);
my $line_nb=0;
open (FILE, "<:encoding(UTF8)", "$ARGV[0]") || die $!;
while (<FILE>){
    chomp;
    @parts = split(" ",$_);
    for(my $i=0; $i<scalar@parts; $i++){
        if ($parts[$i] =~ m/([^-]+)-\|-([^ ]+)/){
            $token_actuel = $1;
            $pos_actuel = $2;
            $all_pos{$pos_actuel}++;
            ### 1er cas: <拉尔松>-|-noun:propernoun ###
            if ($token_actuel =~ /^<[^>]+>$/){
                $hash_actuel{$pos_actuel}++;
                if ($i > 0 && $i < scalar@parts){
                    if ($parts[$i-1] =~ m/([^-]+)-\|-([^ ]+)/){
                        my ($surface, $tag) = ($1, $2);
                        get_pre_tag_freq($tag);
                    }
                    if ($parts[$i+1] =~ m/([^-]+)-\|-([^ ]+)/){
                        my ($surface, $tag) = ($1, $2);
                        get_post_tag_freq ($tag);
                    }
                ## si le token est le premier de cette ligne
                }elsif ($i == 0){
                    $hash_pre{"BOS"}++;
                    if ($parts[$i+1] =~ m/([^-]+)-\|-([^ ]+)/){
                        my ($surface, $tag) = ($1, $2);
                        get_post_tag_freq ($tag);
                    }
                ## si le token est le dernier de cette ligne
                }elsif ($i == scalar@parts-1){
                    $hash_post{"EOS"}++;
                    if ($parts[$i-1] =~ m/([^-]+)-\|-([^ ]+)/){
                        my ($surface, $tag) = ($1, $2);
                        get_pre_tag_freq($tag);
                    }
                }
            }
```

```
            }
            ### 2er cas :  <明-|-noun:propernoun   ###
            if ($token_actuel =~ /^<[^>]+$/){
                if ($i > 0 && $i < scalar@parts){
                    if ($parts[$i-1] =~ m/([^-]+)-\|-([^ ]+)/){
                        my ($surface, $tag) = ($1, $2);
                        get_pre_tag_freq($tag);
                    }
                }elsif ($i == 0){
                    $hash_pre{"BOS"}++;
                }elsif ($i == scalar@parts-1){
                    print SDTERR "WRONG PLACE!!!!\n";
                }
            }
            ### 3er cas : 卫华>-|-noun:propernoun
            if ($token_actuel =~ /^[^><]+>$/){
                if ($i > 0 && $i < scalar@parts){
                    if ($parts[$i+1] =~ m/([^-]+)-\|-([^ ]+)/){
                        my ($surface, $tag) = ($1, $2);
                        get_post_tag_freq ($tag);
                    }
                }elsif ($i == scalar@parts-1){
                    #####print "POS-1 pour le dernier mot de la ligne : ", $parts[$1-1], "\n";
                    $hash_post{"EOS"}++;
                }elsif ($i == 0){
                    print SDTERR "WRONG PLACE!!!!\n";
                }
            }

            ### 4er cas: <赵晓>博-|-noun:propernoun 士-|-noun:common

            if ($token_actuel =~ /^<[^>]+>[^-]$/){
                print $token_actuel, "\n";
                if ($i > 0 && $i < scalar@parts){
                    ####print "POS-1 : ", $parts[$i-1], "\n" ;
                    if ($parts[$i-1] =~ m/([^-]+)-\|-([^ ]+)/){
                        my ($surface, $tag) = ($1, $2);
                        get_pre_tag_freq($tag);
                    }
                }elsif ($i == 0){
                    ####print "POS+1 pour le 1er mot de la ligne : ", $parts[$i+1], "\n";
                    $hash_pre{"BOS"}++;
                }
            }
        }
    }
    $line_nb++;
}

```

```
#################################    RESULTATS FREQ    ##########################################
#All pos
print "ALL POS : \n";
foreach (sort {$all_pos{$b} <=> $all_pos{$a}} keys %all_pos){
    print "$_\t$all_pos{$_}\n";
}
#pre pos
print "\nPREP POS : \n";
foreach (sort {$hash_pre{$b} <=> $hash_pre{$a}} keys %hash_pre){
    print "$_\t$hash_pre{$_}\n";
}
#post pos
print "\nPOST POS: \n";
foreach (sort {$hash_post{$b} <=> $hash_post{$a}} keys %hash_post){
    print "$_\t$hash_post{$_}\n";
}
#actuel pos:
print "\nactuel pos : \n";
foreach (sort {$hash_actuel{$b} <=> $hash_actuel{$a}} keys %hash_actuel){
    print "$_\t$hash_actuel{$_}\n";
}
```

```
120 ######################################      SUB      ##########################################
121 # TODO : faut classifier
122 sub get_pre_tag_freq {
123     my $tag = shift;
124     if ($tag =~ m/noun/){
125         $hash_pre{"noun"}++;
126     }elsif ($tag =~ m/verb/){
127         $hash_pre{"verb"}++;
128     }elsif ($tag =~ m/classifier/) {
129         $hash_pre{"classifier"}++;
130     }elsif ($tag =~ m/conj/) {
131         $hash_pre{"conjonction"}++;
132     }elsif ($tag =~ m/prefix/){
133         $hash_pre{"prefix"}++;
134     }elsif ($tag =~ m/adv/){
135         $hash_pre{"adv"}++;
136     }elsif ($tag =~ m/adj/) {
137         $hash_pre{"adj"}++;
138     }elsif ($tag =~ m/pron/){
139         $hash_pre{"pron"}++;
140     }elsif ($tag =~ m/numeric/) {
141         $hash_pre{"number"}++;
142     }elsif ($tag =~ m/particle/){
143         $hash_pre{"particle"}++;
144     }elsif ($tag eq "punct") {
145         $hash_pre{"punct"}++;
146     }elsif ($tag =~ m/aspect/ ) {
147         $hash_pre{"aspect"}++;
148     }elsif ($tag =~ m/suffix/) {
149         $hash_pre{"suffix"}++;
150     }elsif ($tag eq "intj") {
151         $hash_pre{"intj"}++;
152     }elsif ($tag =~ m/prep/){
153         $hash_pre{"prep"}++;
154     }
155     return (%hash_pre);
156 
157 }
158 
159 sub get_post_tag_freq {
160     my $tag = shift;
161     if ($tag =~ m/noun/){
162         $hash_post{"noun"}++;
163     }elsif ($tag =~ m/verb/){
164         $hash_post{"verb"}++;
165     }elsif ($tag =~ m/classifier/) {
166         $hash_post{"classifier"}++;
167     }elsif ($tag =~ m/conj/) {
168         $hash_post{"conjonction"}++;
169     }elsif ($tag =~ m/prefix/){
170         $hash_post{"prefix"}++;
171     }elsif ($tag =~ m/adv/){
172         $hash_post{"adv"}++;
173     }elsif ($tag =~ m/adj/) {
174         $hash_post{"adj"}++;
175     }elsif ($tag =~ m/pron/){
176         $hash_post{"pron"}++;
177     }elsif ($tag =~ m/numeric/) {
178         $hash_post{"number"}++;
179     }elsif ($tag =~ m/particle/){
180         $hash_post{"particle"}++;
181     }elsif ($tag eq "punct") {
182         $hash_post{"punct"}++;
183     }elsif ($tag =~ m/aspect/ ) {
184         $hash_post{"aspect"}++;
185     }elsif ($tag =~ m/suffix/) {
186         $hash_post{"suffix"}++;
187     }elsif ($tag eq "intj") {
188         $hash_post{"intj"}++;
189     }elsif ($tag =~ m/prep/){
190         $hash_post{"prep"}++;
191     }
192     return (%hash_post);
193 
194 }
```

3.4 apply_pos.pl : Annoter les ENs avec le script

```
#Author: YAN Liyun
#!/usr/bin/perl

use strict;
use warnings;
use utf8;
use Data::Dumper;
binmode STDIN, ':utf8';
binmode STDOUT, ':utf8';
binmode STDERR, ':utf8';

my (@parts, $token_actuel, $pos_actuel, $temp_chaine, $premier );
my $line_nb=1;
my %list_loc = ();
my %list_org = ();
my %results = ();
my %list_noy = ();
my %list_lastname = ();

open (LIST, "<:encoding(UTF8)", "list_loc") or die $!;
while(<LIST>) {
    chomp;
    $list_loc{$_}++;
}
close LIST;

open (LIST, "<:encoding(UTF8)", "list_org") or die $!;
while(<LIST>) {
    chomp;
    $list_org{$_}++;
}
close LIST;

open (LIST, "<:encoding(UTF8)", "list_noisy") or die $!;
while(<LIST>) {
    chomp;
    $list_noy{$_}++;
}
close LIST;

open (LIST, "<:encoding(UTF8)", "list_lastname") or die $!;
while(<LIST>) {
    chomp;
    $list_lastname{$_}++;
}
close LIST;

open (LIST, "<:encoding(UTF8)", "list_translitteration") or die $!;
while(<LIST>) {
    chomp;
    $list_lastname{$_}++;
}
close LIST;

open (FILE, "<:encoding(UTF8)", "$ARGV[0]") || die $!;
while (<FILE>){
    chomp;
    @parts = split(" ",$_);
    for(my $i=0; $i<scalar@parts; $i++){
        if ($parts[$i] =~ m/([^-\| ]+)-\|-(.+)/){
            $token_actuel = $1;
            $pos_actuel = $2;
            $premier = substr($token_actuel, 0, 1);
            if (($pos_actuel eq "noun:propernoun") && (length($token_actuel) <= 5) && (length($token_actuel) >= 2)) {
                if ($list_loc{$token_actuel} || $list_org{$token_actuel} || $list_noy{$token_actuel} || !$list_lastname{$premier} ) {
                    next;
                }else{
                    ## détecter les tokens au début d'une ligne
                    if ($i == 0){
                        if ( $parts[$i+1] =~ m/([^-\| ]+)-\|-(.+)/) {
                            my ($surface, $tag) = ($1, $2);
                            if (($tag =~ /verb|punct|noun|particle|adv|prep|conj|particle|pron|aspect|num|adj|prefix/) ){
                                $parts[$i] = "<$token_actuel>-|-$pos_actuel";
                            }
                        }
                    ## détecter les tokens à la fin d'une ligne
                    }elsif ($i == scalar@parts-1 ){
                        if ($parts[$i-1] =~ m/([^-\| ]+)-\|-(.+)/) {
                            my ($s, $t) = ($1, $2);
                            if ($t =~ /verb|punct|noun|particle|conj|prep|conj|aspect|adv|adj|num|pron|classifier/){
                                $parts[$i] = "<$token_actuel>-|-$pos_actuel";
                            }
                        }
                    }
                    ## détecter les tokens au milieu d'une ligne
                    elsif ( ($i > 0) || ($i < scalar@parts) ){
                        if ( $parts[$i+1] =~ m/([^-\| ]+)-\|-(.+)/) {
                            my ($surface, $tag) = ($1, $2);
                            if ($parts[$i-1] =~ m/([^-\| ]+)-\|-(.+)/) {
                                my ($s, $t) = ($1, $2);
                                if  (($t =~ /verb|punct|noun|particle|conj|prep|adv|aspect|adv|adj|num|pron|classifier/) &&
                                ($tag =~ /verb|punct|noun|particle|adv|prep|conj|particle|pron|aspect|num|adj|prefix/)){
                                    $parts[$i] = "<$token_actuel>-|-$pos_actuel";
                                }
                            }
                        }
                    }
```

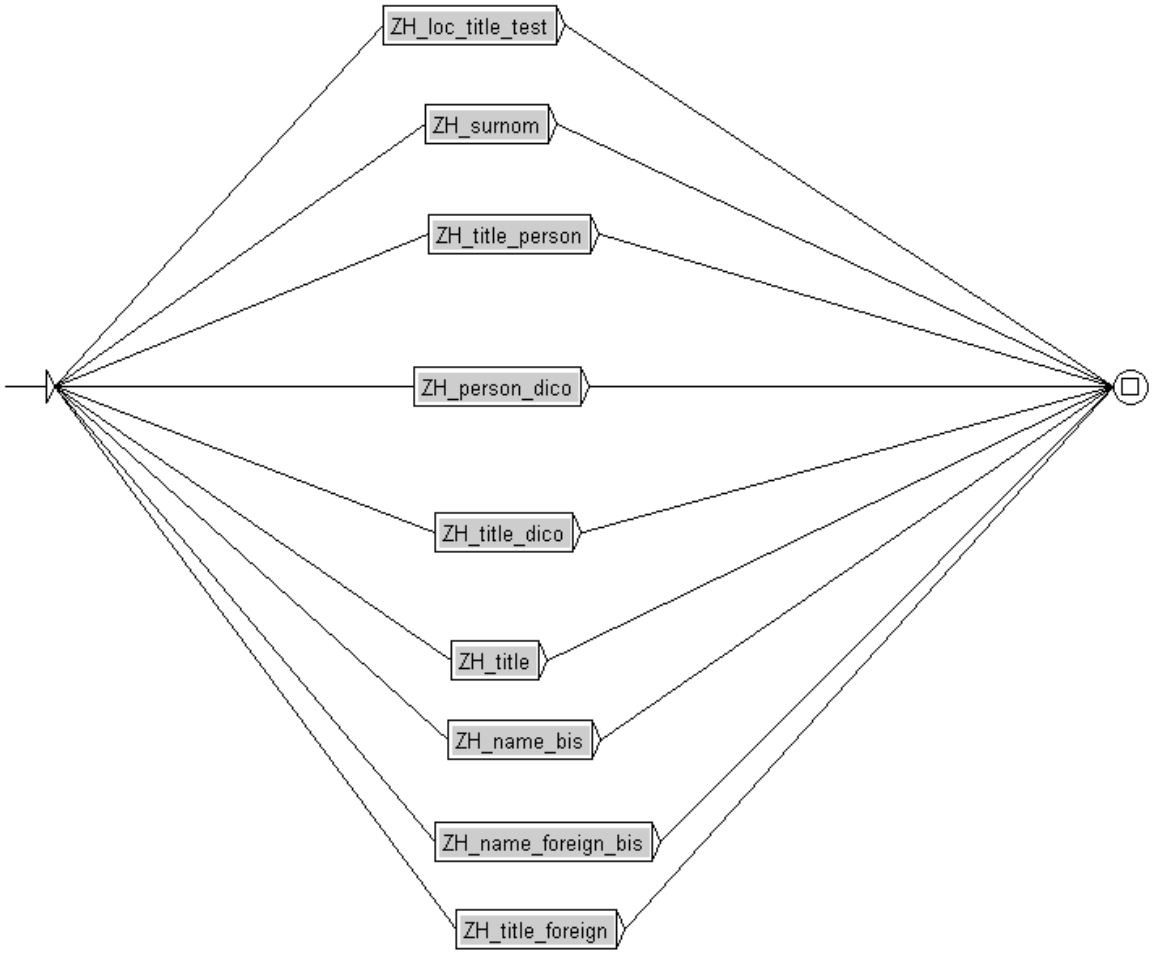
```
                    }
                }
            ##!!!## 2er cas : if la longeur de la chaîne est 1
            ## il faut regarder si son pre et post pos est "noun:propernoun"
            }elsif ($pos_actuel eq "noun:propernoun" && length($token_actuel) == 1 ){
                if ((my($surface, $tag) = $parts[$i+1] =~ m/([^-\| ]+)-\|-(.+)/) && (my($ss, $tt) = $parts[$i+2] =~ m/([^-\| ]+)-\|-(.+)/)){
                    if (($tag eq "noun:propernoun") && (($tt eq "noun:propernoun") || ($tt eq "noun:common")) && (length($ss) == 1 )){
                        $parts[$i] = "<$token_actuel-|-$pos_actuel";
                        $parts[$i+1] = "$surface-|-$tag";
                        $parts[$i+2] = "$ss>-|-$tt";
                        $i=$i+3;
                    }
                    elsif ($tag eq "noun:propernoun" && length($surface) <= 2) {
                        $parts[$i] = "<$token_actuel-|-$pos_actuel";
                        $parts[$i+1] = "$surface>-|-$tag";
                        $i=$i+2;
                    }
                }
            ## 3er cas: if la longeur de la chaîne est 2
            }elsif ($pos_actuel eq "noun:propernoun" && length($token_actuel) == 2 ){
                if ((my($surface, $tag) = $parts[$i+1] =~ m/([^-\| ]+)-\|-(.+)/) ){
                    if ($tag eq "noun:propernoun"  && length($surface) <= 2  ){
                        $parts[$i] = "<$token_actuel-|-$pos_actuel";
                        $parts[$i+1] = "$surface>-|-$tag";
                        $i = $i + 2;
                    }
                }
            }
        }
    }
    $line_nb++;
    print join (" ",@parts);
    print "\n";
}
close FILE;
```

3.5. Post-traitement.pl : Combiner les ENs d'Unitex avec celle de script

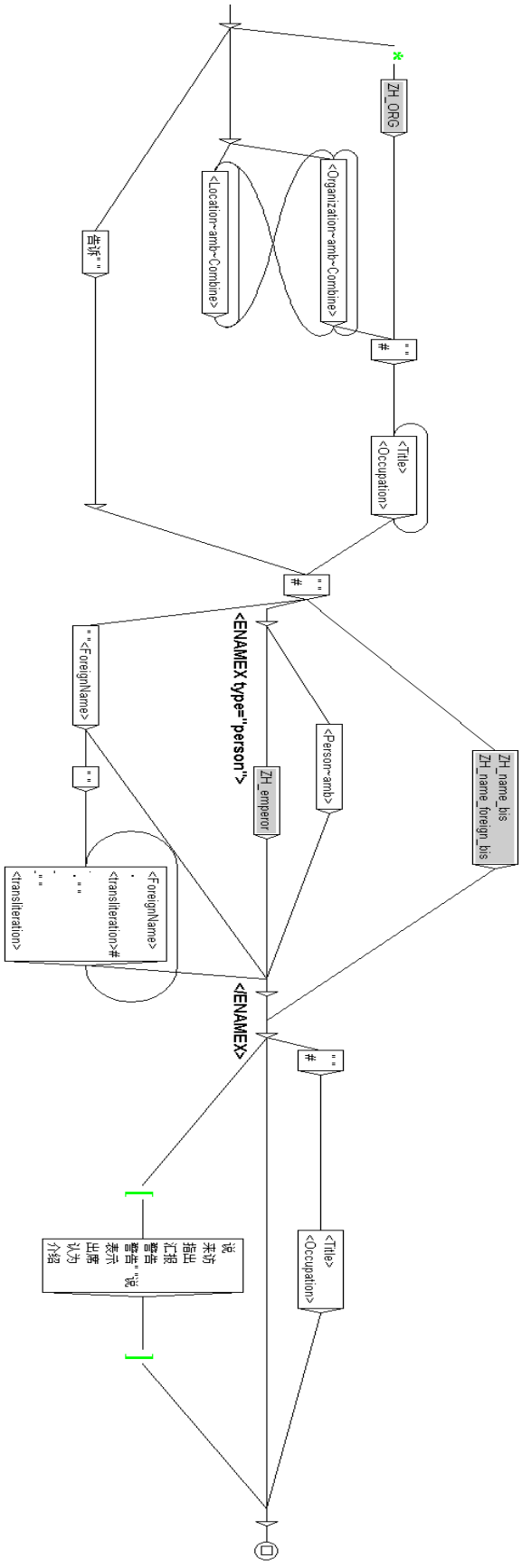
```perl
#author: YAN Liyun
#!/usr/bin/perl

use strict;
use warnings;
use utf8;
use Data::Dumper;
binmode STDIN, ':utf8';
binmode STDOUT, ':utf8';
binmode STDERR, ':utf8';

open (REF, "<:encoding(UTF8)", "$ARGV[0]") || die $!; # sortie unitex
open (TEST, "<:encoding(UTF8)", "$ARGV[1]") || die $!; # sortie script

my %match = ();
my $right = 0;
my $right_ref = 0;
my $right_test = 0;
my $line = 1;
my (@ref, @test);
while ((my $l1 = <REF>) && (my $l2 = <TEST>)) {
    chomp $l1;
    chomp $l2;
    @ref = split (" ",$l1);
    @test = split (" ",$l2);
    my $long_ref = scalar(@ref);
    my $long_test = scalar(@test);
    #calculer les tags de personne dans le fichier de reference
    foreach my $line (@ref){
        if($line =~ /</){
            $right_ref++;
        }
    }
    # calculer les tags de personne dans fichier de test
    foreach my $line (@test){
        if($line =~ /</){
            $right_test++;
        }
    }
    #calcul les bonnes reponses
    my $i = 0;
    while ($i >= 0 && $i <= scalar@ref){
        my $refToken = $ref[$i];
        my $testToken = $test[$i];
        my $post_refToken = $ref[$i+1];
        my $post_testToken = $test[$i+1];
        my $post2_refToken = $ref[$i+2];
        my $post2_testToken = $test[$i+2];
       # unitex abc, script <abc> cas : <习近平>
       if (($refToken =~ /^[^<>]+$/) && ($testToken =~ /^<[^<>]+>$/)){
           $ref[$i] = "<$ref[$i]>";
           # unitex <a b> c, script <a b c>
       }elsif (($refToken =~ m/^[^<>]+$/) && ($post_refToken =~ m/^[^<>]+>$/) && ($testToken =~ /^<[^<>]+$/)
       && ($post_testToken =~ /^[^<>]+$/)&& ($post_testToken =~ /^[^<>]+$/) && ($post2_testToken =~ /^[^<>]+>$/)){
           $ref[$i] = "$ref[$i]";
           $ref[$i+1] = "$test[$i+1]";
           $ref[$i+2] = "$test[$i+2]";
           $i = $i +3;
       # unitex a <b c>, script <a b c>
       }elsif (($refToken =~ /^[^<>]+$/) && ($post_refToken =~ /^<[^<>]+$/) && ($post2_refToken =~ /^[^<>]+>$/)
       && ($testToken =~ /^<[^<>]+$/) && ($post_testToken =~ /^[^<>]+$/) && ($post2_testToken =~ /^[^<>]+>$/)){
           $ref[$i] = "<$ref[$i]";
           $ref[$i+1] = "$test[$i+1]";
           $ref[$i+2] = "$ref[$i+2]";
           $i = $i +3;
       # unitex a b, script <a b>
       }elsif (($refToken =~ /^[^<>]+$/) && ($post_refToken =~ /^[^<>]+$/) && ($testToken =~ /^<[^<>]+$/)
       && ($post_testToken =~ /^[^<>]+>$/)){
           $ref[$i] = "<$ref[$i]";
           $ref[$i+1] = "$ref[$i+1]>";
           $i = $i +2;
       # unitex a b c, script <a b c>
       }elsif (($refToken =~ /^[^<>]+$/) && ($post_refToken =~ /^[^<>]+$/) && ($post2_refToken =~ /^[^<>]+$/)
       && ($testToken =~ /^<[^<>]+$/) && ($post_testToken =~ /^[^<>]+$/) && ($post2_testToken =~ /^[^<>]+>$/)){
           $ref[$i] = "<$ref[$i]";
           $ref[$i+1] = "$ref[$i+1]";
           $ref[$i+2] = "$ref[$i+2]>";
           $i = $i +3;
       }
       $i++;
    }
    $line++;
    print join (" ", @ref);
    print "\n";
}
```

4. Liste des patrons d'Unitex

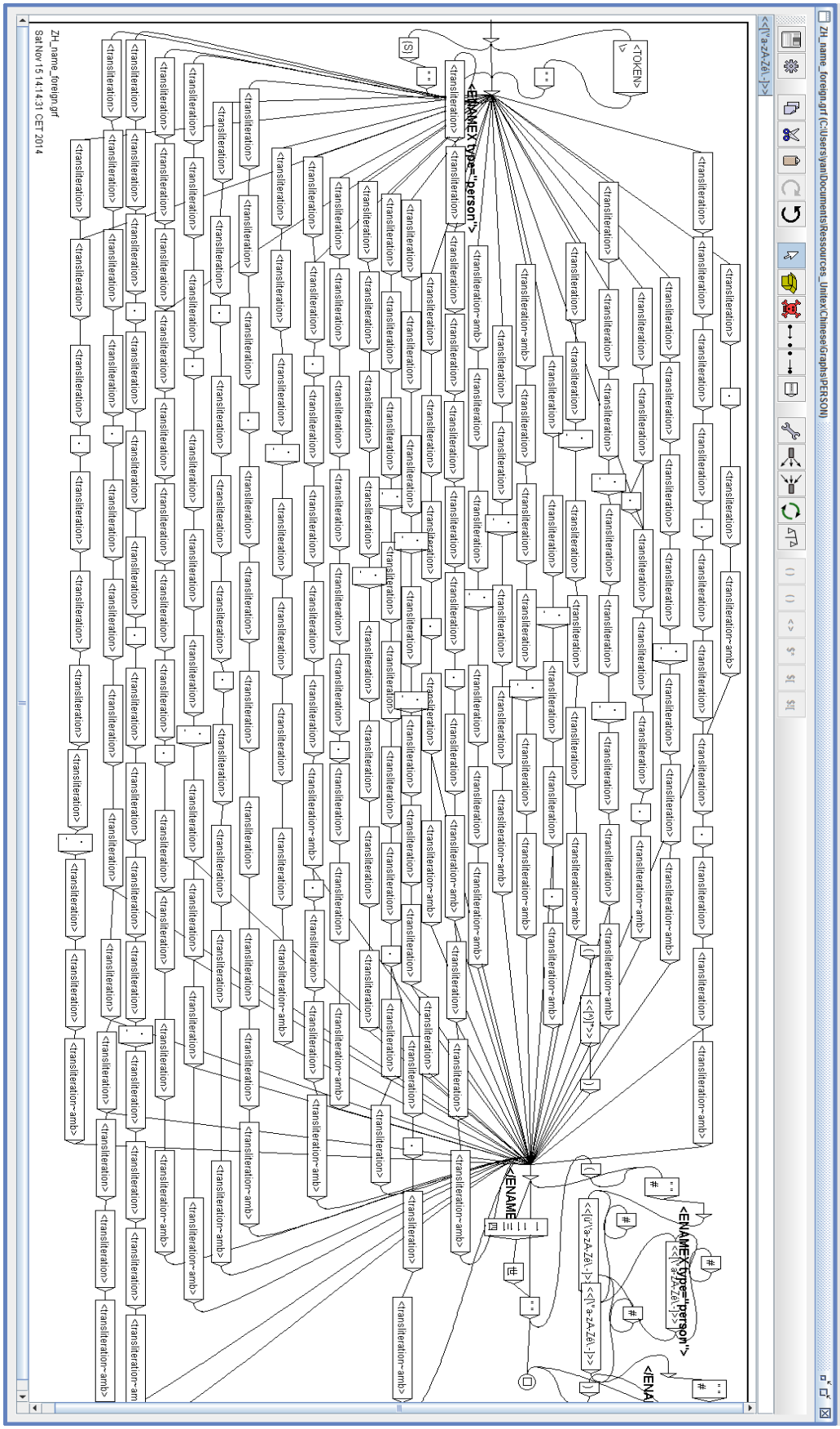
4.1 Le graphe PERSON : contient tous les sous-graphes



4.2 Exemple d’un graphe du « contexte »



4.3 Exemple d’un graphe de normalisation des noms étrangers



4.4 Exemple d’un graphe des personnes chinoises