Construction d'une cascade de transducteurs pour la reconnaissance des dates à partir d'un corpus Wikipédia

Fatma Ben Mesmia*, Nathalie Friburger **, Kais Haddar* et Denis Maurel**

*Université de Sfax, Laboratoire MIRACL, Multimedia, InfoRmation Systems and Advanced Computing Laboratory

fatmabm@ymail.com, Kais.Haddar@fss.rnu.tn

** Université François-Rabelais de Tours, Laboratoire d'Informatique

{nathalie.friburger, denis.maurel}@univ-tours.fr

*RÉSUMÉ.*Les dates sont souvent des sources d'information et peuvent aussi être incluses dans des entités nommées représentant les lieux et les évènements. Leur reconnaissance doit être intégrée dans le processus de reconnaissance des entités nommées arabes. En conséquence, dans le présent article, nous élaborons une cascade de transducteurs reconnaissant les entités nommées arabes de type Date à partir d'un corpus extrait de Wikipédia. L'implémentation de cette cascade est établie en utilisant l'outil CasSysdisponible sous la plateforme linguistique libre Unitex.

*ABSTRACT.*The dates often are sources of informationand can beincluded in named entities representing the locations and events. Their recognition mustbe integrated in the process of recognition of Arab named entities.Consequently, in the presentpaper,we develop a cascade of transducers recognizing Arabic named entities withthe type Date from a corpus extracted of Wikipedia. The implementation of this cascade is established by using the tool CasSysavailable under the Unitex free linguistic platform.

*MOTS-CLÉS :*Cascade de transducteurs, Wikipédia, REN, Unitex, CasSys.

KEYWORDS: Cascade of transducers, Wikipedia, NER, Unitex, CasSys.

1. Introduction

La reconnaissance des entités nommées (REN) constitue une piste de recherche encore très innovante. Elle n'est pas une tâche facile car elle dépend en large partie d'un nombre important de ressources à exploiter. Autrement dit, la complexité de la REN peut être justifiée par l'incomplétude de ces ressources.Le critère d'exhaustivité est donc impossible. En contrepartie, le Web devient très exploité dans nos jours. Il fournit un nombre très intéressant des ressources libres sur lequel elles sont publiées. Parmi celles qui sont plus utilisées, citons Wikipédia.En ce sens, la Wikipédia arabe est considérée comme étant une ressource de connaissances pouvant illustrer des phénomènes linguistiques informatisés. Son exploitation offre l'opportunité pour la valorisation del'entité nommée arabe (ENA) de type Date. Les dates apparaissent dans différents textes (date de naissance, évènement…). Cependant, l'extraction des dates peut rencontrer plusieurs problèmes (en particulier l'existence de différentes écritures régionales).

C'est dans ce contexte que s'inscrit le présent article. Notre objectif est donc de proposer une démarche baséeune cascade de transducteurs reconnaissant les ENA de type Date. Pour ce faire, nous devons, d'une part, identifier un ensemble de mots déclencheurs permettant le repérage d'ENA et, d'autre part, construire un ensemble de transducteurs agissant sur un corpus avec un ordre prédéfini. La cascade proposée doit résoudre les problèmes d'ambiguïté.

Cet article s'articule autour de quatre sections. La première section permet de présenter les approches existantes pour la REN. La deuxième section est dédiée à la description de la catégorisation des dates à partir de Wikipédia. La troisième section est consacrée à détailler la démarche proposée qui va être expérimentée à l'aide du système CasSys de la plateforme

linguistique libre Unitex. Cette expérimentation est présentée et évaluée dans la section quatre.

2. Etat de l'art sur les systèmes de reconnaissance des entités nommées

Les approches de REN existantes sontde trois types : symbolique, statistique et hybrides. Les facteurs de distinction entre les trois approches citées sont leur acquisition et leur manipulation, ce n'est pas la nature des informations qui sera étudiées.L'approche symbolique s'appuie spécialement sur l'utilisation de grammaires formelles construites à la main par un linguiste (Friburger et Maurel, 2004 ; Maurel et al., 2011). Elle se fonde sur des règles exploitant des marqueurs lexicaux, des dictionnaires, etc. Parmi les travaux basés sur cette approche, citons :le système NERA développé par (Shaalan et Raza, 2009) reposant sur l'utilisation d'un ensemble de dictionnaires d'EN et sur une grammaire sous forme d'expressions régulières ; le module de repérage des EN à base de règles pour la langue arabe développé par (Zaghouani et al., 2010) en exploitant une première étape de prétraitement lexical qui prépare le texte pour son analyse linguistique ; le système de reconnaissance d'ENA pour le domaine de sportdéveloppé par (Fehri, 2012) à travers un ensemble de dictionnaires, des patrons syntaxiques etle formalisme de transducteurs sur la plateforme linguistique Nooj.L'approche statistiqueutilise des techniques statistiques sur de larges corpus de textes où les entités-cibles ont été étiquetées. Elle utilise aussi un algorithme d'apprentissage permettant d'élaborer automatiquement une base de connaissances.En se basant sur l'approche statistique, une technique d'apprentissage SVM a été conçue par (Benajiba et al., 2008) pour mettre en œuvre un système de reconnaissance d'entités nommées en exploitant les particularités de la langue arabe.L'approche hybride utiliseà la fois des règles écrites manuellement et des règles extraites grâce à des algorithmes d'apprentissage et à des arbres de décisions.Dans ce contexte, se situe le travail de (Shaalan et Oudah. 2014).

3. Catégorisation des dates à partir de Wikipédia

La catégorisation des ENA de type Date que nous proposons est basée sur l'étude effectuée sur le corpus Wikipédia d'étude constitué de 17 fichiers textes.De ce corpus, nous avons pu identifier trois formes de dates.

Première forme d'ENA de type Date. La première formeest composée par l'année uniquement. Cette formecontient un terme déclencheur qui peut la précéder et/ou la suivre.Par exemple, dans « 1434 عام هـ » (année 1934 hégirienne) le mot عام joue le rôle d'un mot déclencheur permettant d'identifier le nombre 1434 comme étant une année, tandis que le mot هـ ajoute un degré de certitude sur le nombre identifié. C'est un indice que l'année désignée est hégirienne.« 2004في » (en 2004)présente un deuxième exempled'apparition respectant la première forme déjà mentionnée. L'élément brillant est donc l'année. Le mot déclencheur في peut créer une ambiguïté sémantique dans la langue arabe. Il peut être suivi par un nombre désignant l'année (2004) ou suivie d'une suite de caractères indiquant une date. Comme par exemple 21في القرن (en 21ème siècle) ou في الربيع (auprintemps).

Deuxième forme de type Date. La deuxième forme décrit le contexte d'apparition d'une date dont le mois est un élément central. Cette date est incomplète car elle est composée à son tour de deux formes. Nous trouvons soit le nom et/ou le nombre du jour et le mois, soit le mois et l'année.D'après l'étude de corpus, nous constatons que certainesdates peuvent être détectées selon leur contexte d'apparition,Lorsqu'elles sont intégrées dans des événements (par ex., ثورة 14جانفي/ La révolution du 14 janvier) ou dans des noms de lieux (par ex,.ملعب 14 جانفي برادس / Stade 14 janvier de Rades). Quant aux mois hégiriens,ilsapparaissent généralement dans les évènements religieux (par ex,.يوم العيد 1 شوال / L'aïd 1er chawal). Les dates peuvent avoir des

écritures différentes dans les pays arabes. Par exemple, dans les pays orientaux, les mois syriaques et musulmans sont les plus utilisés. Par contre, les mois grégoriens sont utilisés d'une façon fréquente dans les pays magrébins. Au sein de cette union, il existe une différence aux niveaux des appellations des mois. En Tunisie, comme en Algérie,le mois d'août en arabe est « أوت », tandis qu'au Maroc, son appellation est « غشت ».

Troisième forme de type Date. La troisième formeà reconnaitre dans le corpus d'étude concerne une date complète, telle qu'elle composée par le nom et/ou le nombre du jour, le mois, l'année.« 2014 أكتوبر 26 يوم الأحد » (Dimanche 26 octobre 2014)illustre une forme possible d'une date complète figurant dans le corpus d'étude.

4. Démarche proposée pour la reconnaissance des entités nommée de type Date

La démarche que nous proposonsest composée par deux étapes : l'identification des ressources nécessaires pouvant cerner les entités nommées à reconnaitre et la création des transducteurs dont chacun possède son propre rôle.

1.1. Identification des ressources nécessaires

Les ressources nécessaires sont les dictionnaires, les mots déclencheurs, les règles d'extraction. Un dictionnaire doit être crééstockant les noms de la semaine et les noms du mois selon les différents calendriers.Les règles d'extraction des dates sont identifiées grâce aux mots déclencheurs. Par exemple, les mots déclencheurs ثورة,حتى et ليلةreconnaissent respectivement les formes suivantes : <NB><mois><NB>, <NB>عام <mois><NB> et <NB><mois><NB>.

1.2. Cascade de transducteurs proposée

La cascade de transducteurs proposée englobe trois transducteurs principaux. Ces transducteursdoivent être classés selon les trois formes identifiées. Cette décomposition est faite pour éviter les problèmes de chevauchement de certains chemins, d'une part, et les problèmes d'ambigüité, d'autre part.Donnons l'exemple du premier transducteur reconnaissant une date complète (figure 1).

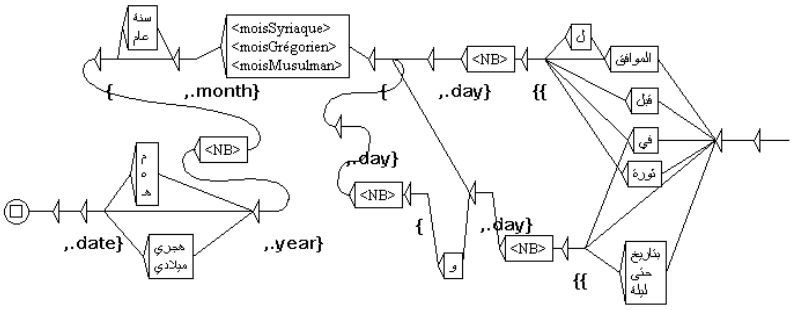

Figure 1. Exemple d'un transducteur reconnaissant une date complète

5. Expérimentation et évaluation

La cascade de transducteurs proposée est implémentéesous la plateforme linguistique Unitex. La figure 2 ci-dessous montre la forme de la cascade qui est générée grâce à l'outil CasSys. L'expérimentationeffectuée montre que chaque graphe ajoute ses propres annotationsà l'aide du mode « Merge ». Ce mode permet d'avoir, en sortie, une ENA reconnue entourée par une balise définie au sein des transducteurs.

#	Disabled	Name	Merge	Replace
1	☐	RecDateCompPrincipal.fst2	☑	☐
2	☐	RecFormemonthPrincipal.fst2	☑	☐
3	☐	RecFormeYearPrincipal.fst2	☑	☐

Figure 2. Cascade de transducteurs reconnaissant les dates

Dans le but d'effectuer une évaluation nous avons appliqué la cascade implémentée sur le corpus de test. Le corpus est composé de 50 fichiers textes dont sa construction rassemble à celle du corpus d'étude. Le résultat obtenu dépend en grande partie des mots déclencheurs établis précédemment.

Echantillons traités	**Entités de type Date trouvées**	**Entités détectées parerreur**
1260	1248	42

Tableau 1. Tableau récapitulatif des résultats obtenus

Nous avons évalué manuellement la qualité de notre travail sur le corpus de test. Les résultats sont satisfaisants (Tableau 1) car les transducteurs ont pu couvrir la majorité des ENA y figurant.Avec une précision de 0,96 et un rappel de 0,95.Nous constatons donc que la méthode proposée est efficace.

6. Conclusion et perspectives

Dans le présent article, nous avons construit un ensemble de transducteurset généré une cascade permettant la reconnaissance des ENA de type Date. La génération de cette cascade est réalisée à l'aide du système CasSys, intégré dans la plateforme linguistique Unitex. Le fonctionnement de la cascade de transducteurs a nécessité la construction d'un dictionnaire et une liste des mots déclencheurs. Dans un futur immédiat, nous tentons decouvrir les autres types (les noms de personnes, les évènements, les noms de lieux, etc.) afin de générer une cascade de transducteurs reconnaissant toutes les ENA. Nous continuons à travailler avec la ressource libre Wikipédia arabeenprofitant de sa richesse pour enrichir nos corpus.

7. Références

Benajiba Y. et Rosso P. 2008. Arabic Named Entity Recognition using Conditional Random Fields, In Proceedings of Workshop on HLT and NLP within the Arabic World, LREC.

Friburger N. et Maurel D. 2004, Finite-state transducer cascade to extract named entities in texts, Theoretical Computer Science, volume 313 : 94–104.

Fehri H. 2012.Reconnaissance automatique des entités nommées arabes et leur traduction vers le français, thèse de doctorat, Université de Sfax.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I. et Nouvel D. 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées, Traitement automatique des langues, 52(1) :69–961.

Shaalan K. et Raza H. 2009. NERA : Named entity recognition for Arabic, Journal of the American Society for Information Science and Technology, 60(9) :1652–1663.

Khaled Shaalan et Mai Oudah. 2014. A hybrid approach to Arabic named entity recognition. Journal of Information Science, 40(1) : 67–87

Zaghouani W., Pouliquen B., Ebrahim M. et Steinberger R. 2010. Adapting a resource-light highly multilingual named entity recognition system to arabic, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) 563–567.