La reconnaissance automatique de la fonction des pronoms démonstratifs en langue arabe

Yacine Ben Yahia, Souha Mezghani Hammami, Lamia Hadrich Belguith
ANLP Research Group – Laboratoire MIRACL/ FSEGS Sfax, Tunisie.
anlp-reserach-group@googlegroups.com
benyacine.sint@gmail.com, souha.mezghani@fsegs.rnu.tn, l.belguith@fsegs.rnu.tn

RESUME

La résolution d'anaphores est l'une des tâches les plus difficiles du Traitement Automatique du Langage Naturel (TALN). La capacité de classifier les pronoms avant de tenter une tâche de résolution d'anaphores serait importante, puisque pour traiter un pronom cataphorique le système doit chercher l'antécédent dans le segment qui suit le pronom. Alors que, pour le pronom anaphorique, le système doit chercher l'antécédent dans le segment qui précède le pronom. En outre, le nombre des pronoms a été jugée non-trivial dans la langue arabe. C'est dans ce cadre que se situe notre travail qui consiste à proposer une méthode pour la classification automatique des pronoms démonstratifs arabes, basée sur l'apprentissage. Nous avons évalué notre approche sur un corpus composé de 365585 mots contenant 14318 pronoms démonstratifs et nous avons obtenu des résultats encourageants : 99.3% comme F-Mesure.

ABSTRACT

Automatic recognition of demonstrative pronouns function in Arabic

Anaphora resolution is one of the most difficult tasks in NLP. Classifying pronouns before attempting a task of anaphora resolution is important because to handle the cataphoric pronoun, the system should determine the antecedent into the segment following the pronoun. Although, for the anaphoric pronoun, the system should look for the antecedent into the segment before the pronoun. In addition, the number of demonstrative pronouns is very important in Arabic. In this paper, we describe a machine learning method for classifying demonstrative pronouns in Arabic. We have evaluated our approach on a corpus of 365585 words which contain 14318 demonstrative pronouns and we have obtained encouraging results: 99.3% as F-Measure.

MOTS-CLES : Pronoms démonstratifs, résolution des anaphores, traitement de la langue arabe.
KEYWORDS: Demonstrative pronouns, anaphora resolution, ANLP.

1 Introduction

La résolution des anaphores pronominales est l'une des branches les plus actives du domaine de Traitement Automatique du langage Naturel (TALN). Elle consiste à identifier l'antécédent pour chaque pronom. La première étape de la tâche de résolution des pronoms anaphoriques consiste à distinguer les occurrences référentielles (anaphoriques et cataphoriques) de celles non-référentielles. Au niveau de cette étape, le système détecte toutes les occurrences non-référentielles afin d'éviter la recherche d'un antécédent qui n'existe pas. En outre, la classification des occurrences référentielles en anaphoriques et cataphoriques est importante pour un système de résolution. En effet,

pour les pronoms anaphoriques, le système doit chercher l'antécédent dans le segment localisé avant le pronom, alors que, pour les pronoms cataphoriques, le système doit chercher l'antécédent dans le segment qui suit le pronom. Par conséquent, cette classification peut améliorer la performance du système.

Considérons les exemples suivants :

(1) انه من الصعب إيجاد حل لهذا المشكل

/ Ain~ahu mina AlS~aEobi Ii?jaAdu HalK liha*aA Almu$okili/

Il est difficile de trouver une solution à ce problème.

(2) عندما دخلت أختي المستشفى العام الماضي، كنّا نحضر لها الكسكسي والعجين بأنواعه

/EinodamaA daxalato Oxotiy Almusota$ofaY AlEaAma AlmaADiy, kuna~A nuHaD~iru lahaA Alkusokusiy waAlEaji?na biOanowaAEihi/

Quand ma sœur entra à l'hôpital l'année dernière, nous lui avons apporté le couscous et les divers types de pâtes

(3) هذا الاختراع مهم جدا

/ha*aA AlIixotiraAEu muhimN jid~FA/

Cette invention est très importante.

Le pronom (هـ/hu/*il*), dans l'exemple (1) ne se réfère à aucun syntagme nominal. Il est donc non-référentiel. Cependant, dans l'exemple (2), le pronom (ها/hA/*lui*) possède comme antécédent le syntagme (أختي/>xty/*ma sœur*) ; de ce faite il est anaphorique. Le pronom (هذا/h*A/cette), dans l'exemple (3) est cataphorique puisqu'il se réfère au nom (الاختراع/AlAxtrAE/invention) situé après le pronom. Ainsi, un système de résolution des anaphores doit chercher les antécédents seulement pour les pronoms des exemples (2) et (3).

Vu l'importance de la classification automatique des pronoms, plusieurs chercheurs se sont intéressés à ce sujet. Certains travaux ont visé la distinction des pronoms personnels de ceux impersonnels ((Lappin et Leass, 1994), (Boyd et a, 2005), (Weissenbacher et Nazarenko, 2007), (Hammami et al, 2010)). D'autres travaux se sont intéressés aux pronoms démonstratifs tels que les travaux de (Muller, 2007), (Byron, 2002) pour l'anglais, le travail de (Navaretta, 2009) pour le danois et le travail de (Dutta et al, 2010) pour l'Hindo. A notre connaissance, il n'existe pas de travaux similaires pour la classification des pronoms démonstratifs en langue arabe.

Dans cet article, nous proposons une méthode d'apprentissage pour la classification automatique des pronoms démonstratifs en langue arabe. Cette méthode permet de classifier les pronoms démonstratifs en pronoms démonstratifs cataphoriques et pronoms démonstratifs anaphoriques. Elle se base, d'une part, sur un ensemble de critères contextuels et d'autre part sur des techniques d'apprentissage. La section 2 présente la spécificité des pronoms démonstratifs arabes. La section 3 donne un aperçu sur l'état de l'art pour la classification des pronoms. Dans la quatrième section, nous décrivons la méthode proposée pour la classification des pronoms démonstratifs. Enfin, nous présentons nos expérimentations et nous discutons les résultats obtenus.

2 Les pronoms démonstratifs en langue Arabe

Selon la littérature, les pronoms démonstratifs sont fréquemment utilisés en langue arabe. Comme pour l'anglais (this, these…) et le français (ceci, celui-ci, celle-là, celui-là,…), il existe dans la langue arabe des pronoms qui désignent le singulier (هذا/h*A/, هذه/h*h/) et le pluriel (هؤلاء/hWlA'/, أولئك/Awl}k/). Ce pendant, il existe des pronoms démonstratifs qui désignent le duel tels que : هذان/h*An/, هاتان/htAn/, ذلكما/*lkmA/, تلكما/tlkmA/, ذانك/*Ank/. Ce qui n'est pas le cas pour le français ou l'anglais. En outre, il existe des pronoms, qui sont considérés comme démonstratifs, et qui désignent le temps (حينذاك/Hyn*Ak/, أنذاك/On*Ak/) et le lieu (هنا/hnA/, هناك/hnAkA/, هنالك/hnAlkA/, ههنا/hhnA/, ثَمَّ/vm~/).

D'après notre étude statistique, il y a des pronoms qui sont utilisés beaucoup plus que d'autres tels que les pronoms هذا/h*A/، تلك/tlk/, هذه/h*h/، ذلك/*lk/. L'utilisation des pronoms démonstratifs (*lkmA/ذلكما, *lkm/ذلكم, tlkmA/تلكما, *lkn/ذلكن) est presque négligeable (ils apparaissent dans le saint coran ou dans les anciens livres arabes).

3 Travaux antérieurs

L'anaphore pronominale est le type d'anaphore le plus fréquent (Mitkov, 2002), c'est pourquoi la résolution automatique des pronoms est un domaine de recherche qui a suscité énormément d'attention depuis plusieurs années. Un système de résolution des anaphores pronominales doit être capable de distinguer les occurrences des pronoms non-référentielles de celles référentielles avant de s'attaquer à leur résolution. De nombreux travaux se sont intéressés particulièrement à cette étape vue son importance et sa difficulté.

La plus part des chercheurs se sont intéressés aux pronoms personnels. Nous pouvons distinguer trois types d'approches : une approche à base de règles telle que les travaux de (Paice et Husk, 1987), (Lappin et Leass, 1994) et (Denber, 1998) pour l'anglais, et (Hammami, 2009) pour l'arabe. Afin de remédier aux inconvénients rencontrés au niveau de l'approche précédente, d'autres auteurs ont adopté une approche numérique basée sur des méthodes d'apprentissages telle que les travaux d' (Evans, 2001) et (Bergsma, 2008).

D'autres, ont fait recours à la combinaison des deux approches précédentes telles que les travaux de (Boyd et al, 2005), (Weissenbacher et Nazarenko, 2007), (Hammami et al., 2010) et (AbdulMajeed, 2011).

Cependant, la classification des pronoms démonstratifs n'a pas encore reçu beaucoup d'attention. (Byron, 2002) décrit un système pour la résolution des pronoms it, this et that dans les dialogues dans un domaine spécifique. Ce système, appelé PHORA, est implémenté et basé sur des connaissances sémantiques. Le résultat d'évaluation était 67% et 62% respectivement pour la précision et le rappel.

(Müller, 2007) a proposé une approche basée sur l'apprentissage automatique pour la résolution des pronoms *it, this* et *that*. Cet algorithme a utilisé cinq corpus différents composés de dialogues pour l'apprentissage et le test, et il repose exclusivement sur l'annotation du corpus. Les résultats de cet algorithme sont moins performants que ceux

des algorithmes reposant sur des connaissances linguistiques et des structures de discoures complexes.

(Navaretta, 2009) décrit des expérimentations d'apprentissage supervisé (classification) et non supervisé (clustering) dans le but de reconnaitre la fonction du pronom neutre singulier dans la langue danoise. Le corpus utilisé est très hétérogène. Il est composé de quatre parties : des textes écrits, des transcriptions de monologue, des dialogues et une interview de TV. La classification de la fonction du pronom neutre singulier comprend neuf classes telles que la classe explétive (non-référentiel), cataphorique, déictique, anaphore individuelle, etc. Le meilleur algorithme pour le clustering est Expectation Maximisation de l'outil Weka. Les résultats obtenus pour les textes écrits sont plus intéressants que pour les autres corpus. Pour la classification, plusieurs algorithmes ont été examinés. Pour estimer la performance de son système, Navaretta utilise la méthode d'échantillonnage validation croisée (10 cross-validation). Elle a fixé l'algorithme ZeroR de Weka comme baseline. Les meilleurs algorithmes sont NBTree, SMO, SMO et KStar respectivement pour les corpus de textes, monologues, dialogue et l'interview.

(Dutta et al., 2010) proposent une application pour la classification des pronoms démonstratifs « yeh », « veh », « iss » et « uss » en langue Hindo. Cette classification est basée sur le formalisme du réseau de neurone probabiliste (PNN). Comme première étape, ils ont extrait des patrons et des caractéristiques pour l'identification des pronoms démonstratifs indirects. Ensuite, ils ont appliqué l'algorithme basé sur le modèle PNN en utilisant la validation croisée. Enfin, des expérimentations sont effectuées pour l'ensemble des données contenant les pronoms démonstratifs et aussi les occurrences des pronoms démonstratifs non référentielles. Les meilleurs résultats sont 94.90% pour tous les pronoms démonstratifs et 84.16% pour les pronoms non-référentiels comme taux de réussite.

4 Méthode proposée

La méthode que nous proposons pour la classification automatique des pronoms démonstratifs arabes est composée de deux phases à savoir la phase d'apprentissage et la phase de test.

La phase d'apprentissage permet d'apprendre à classifier les pronoms. Elle accepte, en entrée, un corpus annoté et elle est composée de trois étapes à savoir la segmentation, l'analyse morphologique et l'extraction des règles. L'étape de segmentation consiste à segmenter les textes de notre corpus. Les textes sont segmentés en phrases dans le but de connaitre les frontières des phrases contenant un pronom démonstratif. Le texte segmenté sera par la suite analysé morphologiquement afin d'identifier les caractéristiques morphologiques des mots de chaque texte de notre corpus. Cette étape va servir à déterminer les valeurs des critères de classification que nous utilisons dans l'étape d'extraction des règles. Pour établir cette dernière, nous avons dégagé huit critères de classification à savoir :

- POS+1 : (Part Of Speech) ce critère prend la catégorie du mot qui suit le pronom. Les valeurs possibles pour ce critère sont : Nom-propre, Nom, Particule, Délimiteur ou Inconnu (dans le cas où la catégorie du mot n'a pas pu être identifiée).

- Type : Dans le cas où le critère *POS+1* prend la valeur Particule, alors le critère *Type* prend le type de cette particule qui peut être : particule de coordination, particule de conjonction, particule d'exception, conjonction d'appel, conjonction de négation, conjonction de condition, etc.
- Determine : Dans le cas où le critère *POS+1* prend la valeur Nom, le critère *Determine* prend la valeur « match » si ce Nom est défini (c'est-à-dire agglutiné à ال). Sinon il prend la valeur « nomatch ».
- Enclitique : Si le mot qui suit le pronom est un *Nom*, et ce nom est agglutiné à un enclitique, alors, ce critère prend la valeur « match ».
- Proclitique+1 : Si le mot qui suit le pronom est agglutiné à un proclitique, alors, ce critère prend la valeur « match ».
- Bimafidhalika : Ce critère prend la valeur « match » si le pronom (*lk / ذلك) est précédé par une succession des deux mots (bmA /بما) et (fyفي/).
- MotSpec : Si le pronom est suivi d'un mot spécifique ce critère prend la valeur « match ». La liste des mots spécifiques est composée des mots suivants: OyDA/أيضا, gyr/غير, mmA/ممّا, kl/كل.
- Pronom : ce critère reçoit le pronom à apprendre.

L'étape d'extraction des règles exploite ces critères de classification pour produire des règles appelées règles de classification, en utilisant un algorithme d'apprentissage.

La phase de test permet de classifier un nouveau pronom démonstratif en pronom anaphorique ou pronom cataphorique. Elle accepte en entrée un texte brut qui sera par la suite segmenté en phrases et analysé morphologiquement. L'étape d'identification des pronoms démonstratifs consiste à identifier les pronoms démonstratifs figurant dans le texte afin de les classifier automatiquement. La détection des pronoms se fait d'une manière automatique en examinant le texte analysé morphologiquement et en faisant ressortir les mots qui ont comme catégorie pronom démonstratif (Asm I$Arp/ اسم إشارة). Enfin, les pronoms identifiés seront classifiés en pronoms démonstratifs anaphoriques ou pronoms démonstratifs cataphoriques en se basant sur les règles d'extraction générées par la phase d'apprentissage.

5 Expérimentations

5.1 Corpus

Le processus de classification à base d'apprentissage nécessite généralement un corpus annoté afin d'assurer la phase d'entraînement. Il est à signaler que la constitution d'un corpus de référence (corpus d'apprentissage) est coûteuse. Ainsi, et vu le manque de corpus étiquetés pour la langue arabe, nous avons procédé à une étape d'annotation binaire des documents constituant notre corpus. Il s'agit d'attribuer à chaque pronom démonstratif la classe anaphorique ou cataphorique. Pour accélérer notre travail, nous avons développé un système d'annotation manuelle AnnotAr. Ce système accepte en entrée un texte segmenté en mot sous le format XML et permet à l'utilisateur d'annoter les pronoms démonstratifs dans le texte en pronoms anaphoriques ou cataphoriques. Le texte annoté est enregistré sous format XML où chaque pronom démonstratifs est étiqueté par la balise qui lui correspond (c'est-à-dire <ANA> pour le pronom anaphorique et <CATA> pour le pronom cataphorique).

Le corpus d'apprentissage est composé d'un ensemble d'articles de presse d'ELMASRY ALYOUM 2010, de textes de livres scolaires, de l'enseignement Tunisien de différents niveaux (un texte contient en moyenne vingt cinq phrases), des manuels d'utilisation (la taille moyenne d'un manuel est de trente pages) et un extrait du Penn Arabic TreeBank (ATB). Nous avons choisi un corpus de nature variée parce que nous estimons que plus le corpus est diversifié plus il sera représentatif. Ce corpus contient un total de 365585 mots et nous a permis d'analyser 14318 pronoms démonstratifs (où 32.15% sont anaphoriques et 67.85% sont cataphoriques).

Corpus	Anaphorique		Cataphorique		Total
	Nombre	Pourcentage(%)	Nombre	Pourcentage(%)	
Livres	1072	33.29	2148	66.71	3220
Journaux	2562	31.88	5472	68.12	8034
Manuels	155	29.41	372	70.59	527
ATB	815	47.32	1722	52.68	2537
Total	4604	32.15	9714	67.85	14318

TABLE 1:Statistiques du corpus

5.2 Résultats et discussion

Nous avons effectué nos expérimentations de classification en utilisant le système Weka (Frank, Witten, 2005) qui permet de tester et de comparer plusieurs algorithmes. Nous avons choisi de tester les algorithmes suivants: IBk, JRip, NBTree et NaiveBayes.

Nous avons procédé à une validation croisée pour valider les résultats de nos expériences. Nous avons sélectionné aléatoirement le neuf dixième du corpus pour l'apprentissage. Nous avons ensuite appliqué notre système sur le un dixième restant. Nous avons réitéré dix fois ces opérations en changeant à chaque fois la partie de test, pour obtenir la moyenne des performances de chaque itération. Les attributs pertinents d'après Weka sont : POS + 1, Type, Determine, procltique + 1, Pronom et motSpec.

Afin de bien évaluer notre système, nous avons implémenté un système Baseline à base des règles. Le système Baseline repose sur six règles contextuelles qui se basent principalement sur les caractéristiques morphologiques du mot qui suit le pronom démonstratif (ex. pr_dem + Nom-défini --> pr-cataph, pr_dem + pr-relatif --> pr-cataph). Les résultats obtenus sont présentés dans le Tableau 2.

Algorithme	Résultats avec annotation morphologique automatique (MORPH2)			Résultats avec annotation morphologique manuelle		
	Précision	Rappel	F-Mesure	Précision	Rappel	F-Mesure
IBk	94.7%	94.7%	94.7%	99.3%	99.3%	99.3%

JRip	93.9%	93.9%	93.9%	99%	99%	99%
NBTree	94.7%	94.7%	94.7%	99.2%	99.2%	99.2%
NaiveBayes	92.8%	92.8%	92.8%	96.2%	96.2%	96.2%
Baseline	78.09%	75.39%	76.72%	97.87%	98.89%	98.38%

TABLE 2 : Résultats obtenus avec la validation croisée

En examinant les mesures de rappel, précision et F-mesure calculées sur le corpus d'évaluation, nous remarquons que les résultats sont très encourageants.

D'une part, l'utilisation de l'apprentissage a amélioré les résultats d'une manière significative (16.61% et 15.81% respectivement pour IBk et JRip) par rapport au Baseline (méthode à base de règles). Cela justifie notre choix d'une méthode d'apprentissage qui réduit l'erreur d'estimation en déterminant le poids des attributs discriminants pour le domaine du corpus.

D'autre part, nous remarquons que l'algorithme IBk (K Plus Proche Voisin) donne les meilleurs résultats par rapport aux autres algorithmes. Ensuite, nous avons mené deux évaluations. Au niveau de la première, nous avons utilisé l'analyseur morphologique MORPH2 (Chaaben et al, 2010) pour l'étiquetage morphologique de notre corpus. Au niveau de la deuxième évaluation, nous avons corrigé manuellement les résultats de MORPH2. En effet, l'utilisation d'un étiquetage morphologique manuel a amélioré les résultats d'une manière significative (environ 5% pour l'apprentissage et 21.66% pour le Baseline).

Les principales erreurs sont dues à des constructions spécifiques de phrases contenant un pronom démonstratif ainsi le manque de ponctuations en langue arabe (Belguith et al., 2005). Citons l'exemple suivant :

(4) ...قد يستهلك ذلك الكثير من الماء.
*/qado yasotaholiku *alika Alkaviyra mina AlmaA'i/*
Ça peut consommer beaucoup d'eau.

Dans cet exemple, le pronom démonstratif (ذلك, /*lk/, ça) est suivi par le nom défini (الكثير, /Alkvyr/, beaucoup). En appliquant l'apprentissage ou la méthode de Baseline, ce pronom démonstratif est classé cataphorique alors qu'il est anaphorique. Cette fausse classification est due à l'absence de la virgule après le pronom démonstratif.

6 Conclusion

La classification des pronoms démonstratifs est une étape très importante dans le processus de la résolution de l'anaphore pronominale. Dans cet article, nous avons proposé une méthode d'apprentissage pour la classification binaire des pronoms démonstratifs en langue arabe en pronom anaphoriques et cataphoriques. Cette méthode d'apprentissage proposée atteint des résultats meilleurs que celle à base des règles. Ainsi l'algorithme K-PPV a donné un meilleur résultat. En se basant sur ces résultats, nous envisageons de chercher les antécédents des pronoms et de terminer les étapes de la résolution d'anaphores.

Références

C. PAICE ET G. HUSK (1987). Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun it. *Computer Speech and Language.*

D. WEISSENBACHER ET A. NAZARENKO (2007). A bayesian classifier for the recognition of the impersonal occurrences of the it pronoun. *In Proceedings of DAARC'07, 2007.*

S. LAPPIN ET H. LEASS (1994). An algorithm for pronominal anaphora resolution, *Computational Linguistics, 20(4), 1994, p. 535–561.*

S. MEZGHANI HAMMAMI., R.SELLAMI, L. HADRICH BELGUITH (2010). A Bayesian Classifier for the Identification of Non-referential Pronouns in Arabic. *The 7th INFOS 2010.*

S. HAMMAMI, L. BELGUITH, A. BEN HAMADOU (2009). A Rule-Based Method for Detecting Arabic Anaphoric Pronouns. *Proceedings of the 7th DAARC'2009, Goa-India, 2009.*

R. EVANS (2001). Applying machine learning toward an automatic classifi cation of it. *Literary and linguistic computing, 16, 2001, p. 45–57.*

M. DENBER (1998). Automatic resolution of anaphora in English. *Eastman Kodak Co, 1998.*

ABDUL-MAGEED, M. 2011. Automatic detection of Arabic non-anaphoric pronouns for improving anaphora resolution. *Asian Lang. Inform. Process. (March 2011).*

S. BERGSMA, D. LIN ET R. GOEBEL (2008). Distributional Identification of Non-Referential Pronouns. *ACL, Columbus Ohio, 2008, p. 10-18.*

A. BOYD, W. GEGG-HARRISON ET D. BYRON (2005). Identifying non-referential it: a machine learning approach incorporating linguistically motivated features. *Workshop on Feature Engineering for Machine Learning in Natural Language Processing, 2005.*

C. MÜLLER (2007).Resolving It, This, and That in Unrestricted Multi-Party Dialog. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007.*

D. BYRON (2002). Resolving Pronominal Reference to Abstrac Entities. *In Proceedings of th 40th Annual Meeting of the Association of Computation Linguistics (ACL), Philadelphia 2002.*

C. NAVARETTA (2009). Automatic Recognition of the Function of Singular Neuter Pronouns in Texts and Spoken Data. *In DAARC 2009.*

K. DUTTA , N. PRAKASH, S. KAUSHIK (2010). Probabilistic neural network approach to the classification of demonstrative pronouns for indirect anaphora in Hindi. *International Journal Information Technology and Intelligent Computing, 2010.*

N. CHAÂBEN KAMMOUN, L. HADRICH BELGUITH ET A. BEN HAMADOU (2010). The MORPH2 new version: A robust morphological analyzer for Arabic texts. *JADT'2010.*

E. FRANK, IAN H. WITTEN (2005). Practical Machine Learning Tools and Techniques, Second Edition. *(Morgan Kaufmann series in data management systems).*

BELGUITH, L., BACCOUR, L. AND MOURAD, G. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules *(TALN'2005).*

MITKOV (2002): R. Mitkov, «Anaphora resolution». Longman. 2002.