Arabic Mention Detection: Toward Better Unit of Analysis

Yassine Benajiba
Center for Computational Learning Systems
Columbia University
ybenajiba@ccls.columbia.edu

Imed Zitouni
IBM T. J. Watson Research Center
izitouni@us.ibm.com

Abstract

We investigate in this paper the adequate unit of analysis for Arabic Mention Detection. We experiment different segmentation schemes with various feature-sets. Results show that when limited resources are available, models built on morphologically segmented data outperform other models by up to 4F points. On the other hand, when more resources extracted from morphologically segmented data become available, models built with Arabic TreeBank style segmentation yield to better results. We also show additional improvement by combining different segmentation schemes.

1 Introduction

This paper addresses an important and basic task of information extraction: *Mention Detection* (MD)[1]: the identification and classification of textual references to objects/abstractions (i.e., *mentions*). These mentions can be either named (e.g. Mohammed, John), nominal (city, president) or pronominal (e.g. he, she). For instance, in the sentence "President *Obama* said *he* will visit ..." there are three mentions: *President, Obama* and *he*. This is similar to the Named Entity Recognition (NER) task with the additional twist of also identifying nominal and pronominal mentions. We formulate the mention detection problem as a classification problem, by assigning to each token in the text a label, indicating whether it starts a specific mention, is inside a specific mention, or is outside all mentions. The selection of the unit of analysis is an important step toward a better classification. When processing languages, such as English, using the word itself as the unit of analysis (after separating punctuations) leads to a good performance (Florian et al., 2004). For other languages, such as Chinese, character is considered as the adequate unit of analysis (Jing et al., 2003). In this paper, we investigate different segmentation schemes in order to define the best unit of analysis for Arabic MD. Arabic adopts a very complex morphology, i.e. each word is composed of zero or more *prefixes*, one *stem* and zero or more *suffixes*. Consequently, the Arabic data is sparser than other languages, such as English, and it is necessary to "segment" the words into several units of analysis in order to achieve a good performance.

(Zitouni et al., 2005) used Arabic morphologically segmented data and claimed to have very competitive results in ACE 2003 and ACE 2004 data. On the other hand, (Benajiba et al., 2008) report good results for Arabic NER on ACE 2003, 2004 and 2005 data using Arabic TreeBank (ATB) segmentation. In all published works, authors do not mention a specific motivation for the segmentation scheme they have adopted. Only for the Machine Translation task, (Habash and Sadat, 2006) report several results using different Arabic segmentation schemes. They report that the best results were obtained when the ATB-like segmentation was used. We explore here the four known and linguistically-motivated sorts of segmentation: punctuation separation, ATB, morphological and character-level segmentations. To our knowledge, this is the first paper which investigates different segmentation schemes to define the unit of analysis which best fits Arabic MD.

2 Arabic Segmentation Schemes

Character-level Segmentation: considers that each character is a separate token.

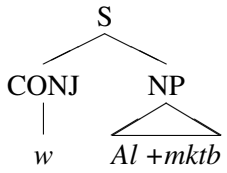
Morphological Segmentation : aims at segmenting

[1]We adopt here the ACE nomenclature: http://www.nist.gov/speech/tests/ace/index.html

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 709–712,
Los Angeles, California, June 2010.

all affixes of a word. The morphological segmentation for the word والمكتب (wAlmktb — and the office)[2] could be: "و +ال +مكتب" (w +Al +mktb).

Arabic TreeBank (ATB) segmentation : This segmentation considers splitting the word into affixes only if it projects an independent phrasal constituent in the parse tree. As an example, in the word shown above والمكتب, the phrasal independent constituents are: the conjunction و (*w* — and) and the noun المكتب (*Almktb* — the office). The morphological segmentation of this word would lead to the following parse tree:



Since the ال (*Al*, the definite article) is not an independent constituent, it is not considered for ATB segmentation. Hence, for والمكتب, the ATB segmentation would be و +المكتب (w +Almktb).

Punctuation separation : it consists of separating the punctuation marks from the word.

Both ATB and morphological segmentation systems are based on *weighted finite state transducers* (WFST). The decoder implements a general Bellman dynamic programming search for the best path on a lattice of segmentation hypotheses that match the input characters (Benajiba and Zitouni, 2009). ATB and morphological segmentation systems have a performance of 99.4 and 98.1 F-measure respectively on ATB data.

The unit of analysis when doing classification depends on the used segmentation. When using the punctuation separation or character-based segmentations, the unit of analysis is the word itself (without the punctuation marks attached) or the character, respectively. The ATB and morphological segmentations are language specific and are based on different linguistic viewpoint. When using one of these two segmentation schemes, the unit of analysis is the morph (i.e. prefix, stem or suffix). Our goal in this paper is to find the unit of analysis that fits best Arabic MD.

[2]Throughout the paper, for each Arabic example we show between parenthesis its transliteration and English translation separated by "—".

3 Mention Detection System

As explained earlier, we consider the MD task as a sequence classification problem where the class we predict for each unit of analysis (i.e., token) is the type of the entity which it refers to. We chose the maximum entropy (MaxEnt) classifier that can integrate arbitrary types of information and make a classification decision by aggregating all information available for a given classification. For more details about the system architecture, reader may refer to (Zitouni et al., 2009). The features used in our MD system can be divided into four categories:

Lexical Features: n-grams spanning the current token; both preceding and following it. A number of n equal to 3 turned out to be a good choice.

Stem n-gram Features: stem trigram spanning the current stem; both preceding and following it (Zitouni et al., 2005).

Syntactic Features: POS tags and shallow parsing information in a ± 2 window.

Features From Other Classifiers: outputs of MD and NER taggers trained on other data-sets different from the one we used here. They may identify types of mentions different from the mentions of interest in our task. For instance, such a tagger may identify dates or occupation references (not used in our task), among other types. Our hypothesis is that combining classifiers from diverse sources will boost performance by injecting complementary information into the mention detection models. We also use the two previously assigned classification tags as additional feature.

4 Data

Experiments are conducted on the Arabic ACE 2007 data. Since the evaluation tests set are not publicly available, we have split the publicly available *training* corpus into an 85%/15% data split. We use 323 documents (80,000 words, 17,634 mentions) for training and 56 documents (18,000 words, 3,566 mentions) as a test set. We are interested in 7 types of mentions: facility, Geo-Political Entity (GPE), location, organization, person, vehicle and weapon. We segmented the training and test set with four different styles building the following corpora:

Word$_s$: a corpus which is the result of running punctuation separation;

ATB$_s$: a corpus obtained by running punctuation separation and ATB segmentation;

Moph$_s$: a corpus where we conduct punctuation separation and morphological segmentation;

Char$_s$: a corpus where the original text is separated

into a sequence of characters.

When building MD systems on $Word_s$, ATB_s, $Morph_s$ and $Char_s$, the unit of analysis is the word, the ATB token, the morph and the character, respectively.

5 Experiments

We show in this section the experimental results when using Arabic MD system with different segmentation schemes and different feature sets. We explore in this paper four categories of features (c.f. Section 3):
Lex$_f$: lexical features;
Stem$_f$: Lex_f + morphological features;
Synt$_f$: $Stem_f$ + syntactic features;
Sem$_f$: $Synt_f$ + output of other MD classifiers. Lex_f and $Stem_f$ features are directly extracted from the appropriate corpus based on the used segmentation style. This is different for Sem_f: we first run classifiers on the morphologically segmented data. Thereafter, we project those labels to other corpora. This is because, we use classifiers initially trained on morphologically segmented data such as ACE 2003, 2004 and 2005 data. In such data, two morphs belonging to the same word or ATB token may have 2 different mentions. During transfer, a token will have the label of the corresponding stem in the morphologically segmented data. One motivation to not re-train classifiers on each corpus separately is to be able to extract Sem_f features from classifiers with similar performance.

Table 1: Results in terms of F-measure per feature-set and segmentation scheme

	Lex_f	$Stem_f$	$Synt_f$	Sem_f
$Word_s$	66.4	66.6	69.0	77.1
ATB_s	70.1	69.8	72.1	**79.0**
$Morph_s$	**74.1**	**74.5**	**75.5**	78.3
$Char_s$	22.3	22.4	22.5	22.6

Results in Table 1 show that classifiers built on ATB_s and $Morph_s$ have shown to perform better than classifiers trained on data with other segmentation styles. When the system uses character as the unit of analysis, performance is poor. This is because the token itself becomes insignificant information to the classifier. On the other hand, when only punctuation separation is performed ($Word_s$), the data is significantly sparse and the obtained results achieves high F-measure (77.1) only when outputs of other classifiers are used. As mentioned earlier, classifiers used to extract those features are trained on $Morph_s$ (less sparse), which explains their remarkable positive impact since they resolve part of the data sparseness problem in $Word_s$. When using full morphological segmentation, the data is less sparse, which leads to less Out-Of-Vocabulary tokens (OOVs): the number of OOVs in the $Morph_s$ data is 1,518 whereas it is 2,464 in the ATB_s. As an example, the word الرهينة (Alrhynp — the hostage), which is person mention in the training data. This word is kept unchanged after ATB segmentation and is segmented to "ال + رهين +ة" (Al+ rhyn +p) in $Morph_s$. In the development set the same word appears in its dual form without definite article, i.e. رهينتين. This word is unchanged in ATB_s and is segmented to "رهين +ت +ين" (rhyn +p +yn) in $Morph_s$. For the model built on ATB_s, this word is an OOV, whereas for the model built on $Morph_s$ the stem has been seen as part of a person mention and consequently has a better chance to tag it correctly. These phenomena are frequent, which make the classifier trained on $Morph_s$ more robust for such cases. Also, we observed that models trained on ATB_s perform better on long span mentions. We think this is because a model trained on ATB_s has access to larger context. One may argue that a similar behavior of the model built on the $Morph_s$ might be obtained if we use a wider context window than the one used for ATB_s in order to have similar contextual information. In order to confirm this statement, we have carried out a set of experiments using all features over $Morph_s$ data for a context window up to $-5/+5$, the obtained results show no improvement. Similar behavior is observed when looking to results on identified named (Nam.), nominal (Nom.) and pronominal (Pro.) mentions on ATB_s and $Morph_s$ (c.f. Table 2); we remind the reader that NER is about recognizing named mentions. When limited resources are available (e.g. Lex_f, $Stem_f$ or $Synt_f$), we believe that it is more effective to morphologically segment the text ($Morph_s$) as a pre-processing step. The use of morph as a unit of analysis reduces the data sparseness issue and at the same time allows better context handling when compared to character. On the other hand, when a larger set of resources are available (e.g., Sem_f), the use of the ATB token as a unit of analysis combined with morph-based features leads to better performance (79.0 vs. 78.3 on $Morph_s$). This is because (1) classifiers trained on ATB_s handle better the context and (2) the use of morph-based features (output of classi-

fiers trained on morphologically segmented data) removes some of the data sparseness from which classifiers trained on ATB_s suffer. The obtained improvement in performance is statistically significant when using the stratified bootstrap re-sampling significance test (Noreen, 1989). We consider results as statistically significant when $p < 0.02$, which is the case in this paper. For an accurate MD system, we think it is appropriate to benefit from ATB_s tokens and $Morph_s$. We investigate in the following the combination of these two segmentation styles.

Table 2: Performance in terms of F-measure per level on ATB_s and $Morph_s$

	$Seg.$	Lex_f	$Stem_f$	$Synt_f$	Sem_f
Nam.	ATB_s	68.2	69.0	72.8	79.1
	$Morph_s$	73.4	73.8	75.3	78.7
Nom.	ATB_s	65.6	64.6	66.9	75.8
	$Morph_s$	71.7	72.2	72.9	75.4
Pro.	ATB_s	60.7	60.1	59.9	66.3
	$Morph_s$	63.0	67.2	65.7	65.1

5.1 Combination of ATB and Morph

We trained a model on ATB_s that uses output of the model trained on $Morph_s$ as additional information ($M2A_f$ feature). We proceed similarly by training a model on $Morph_s$ using output of the model trained on ATB_s ($A2M_f$ feature). We have obtained the features by a 15-way round-robin. Table 3 shows the obtained results.

Table 3: Results in terms of F-measure of the combination experiments

	Lex_f	$Stem_f$	$Synt_f$	Sem_f
ATB_s	70.1	69.8	72.1	79.0
ATB_s+$M2A_f$	70.7	70.8	73.1	**79.1**
$Morph_s$	74.1	74.5	75.5	78.3
$Morph_s$+$A2M_f$	**74.9**	**75.2**	**75.4**	78.6

Results show a significant improvement for models that are trained on ATB_s using information from $Morph_s$ in addition to Lex_f, $Stem_f$ and $Synt_f$ features. This again confirms our claim that the use of features from morphologically segmented text reduces the data sparseness and consequently leads to better performance. For Sem_f features, only a 0.1 F-measure points have been gained. This is because we are *already* using output of classifiers trained on morphologically segmented data, which resolve some of the data sparseness issue. The $Morph_s$ side shows that the obtained performance when the ATB_s output is employed together with the $Stem_f$ (75.2) is only 0.3 points below the performance of the system using $Synt_f$ (75.5).

6 Conclusions

We have shown a comparative study aiming at defining the adequate unit of analysis for Arabic MD. We conducted our study using four segmentation schemes with four different feature-sets. Results show that when only limited resources are available, using morphological segmentation leads to the best results. On the other hand, model trained on ATB segmented data become more powerful and effective when data sparseness is reduced by the use of other classifier outputs trained on morphologically segmented data. More improvement is obtained when both segmentation styles are combined.

References

Y. Benajiba and I. Zitouni. 2009. Morphology-based segmentation combination for arabic mention detection. *Special Issue on Arabic Natural Language Processing of ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4).

Y. Benajiba, M. Diab, and P. Rosso. 2008. Arabic named entity recognition using optimized feature sets. In *Proc. of EMNLP'08*, pages 284–293.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, and S. Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proc.eedings of HLT-NAACL'04*, pages 1–8.

N. Habash and F. Sadat. 2006. Combination of arabic preprocessing schemes for statistical machine translation. In *Proceedings of ACL'06*, pages 1–8.

H. Jing, R. Florian, X. Luo, T. Zhang, and A. Ittycheriah. 2003. HowtogetaChineseName(Entity): Segmentation and combination issues. In *Proceedings of EMNLP'03*, pages 200–207.

E. W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley Sons.

I. Zitouni, J. Sorensen, X. Luo, and R. Florian. 2005. The impact of morphological stemming on arabic mention detection and coreference resolution. In *Proc. of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 63–70.

I. Zitouni, X. Luo, and R. Florian. 2009. A cascaded approach to mention detection and chaining in arabic. *IEEE Transactions on Audio, Speech and Language Processing*, 17:935–944.