RC25043 (W1009-001) September 1, 2010
Other

IBM Research Report

Enhancing Mention Detection Using Projection via Aligned Corpora

Yassine Benajiba
Center for Computational Learning Systems
Columbia University

Imed Zitouni
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Enhancing Mention Detection using Projection via Aligned Corpora

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

Imed Zitouni
IBM T.J. Watson Research Center
Yorktown Heights, NY
izitouni@us.ibm.com

Abstract

The research question treated in this paper is centered on the idea of exploiting rich resources of one language to enhance the performance of a mention detection system of another one. We successfully achieve this goal by projecting information from one language to another via a parallel corpus. We examine the potential improvement using various degrees of linguistic information in a statistical framework and we show that the proposed technique is effective even when the target language model has access to a significantly rich feature set. Experimental results show up to 2.4F improvement in performance when the system has access to information obtained by projecting mentions via a parallel corpus from a resource-rich-language mention detection system.

1 Introduction

The task of identifying and classifying entity textual references in open-domain texts, i.e. the *Mention Detection* (MD) task, has become one of the most important subtasks of Information Extraction (IE). It might intervene both as one step to structure natural language texts or as a text enrichment preprocessing step to help other Natural Language Processing (NLP) applications reach higher accuracy. Similarly to the Automatic Content Extraction (ACE) [1] nomenclature, we consider that a mention can be either named (e.g., John, Chicago), nominal (e.g., president, activist) or pronominal (e.g., he, she). It has also a specific class which describes the type of the entity it refers to. For instance, in the sentence:

Michael Bloomberg, the Mayor of NYC, declared his war on tobacco and sugary drinks in the city.

we find the mentions ‘Michael Bloomberg’, ‘Mayor’ and ‘his’ of the same person entity. Their types are named, nominal and pronominal, respectively. ‘NYC’ and ‘city’, on the other hand, are mentions of the same geopolitical (GPE) entity of type named and nominal, respectively. Consequently, MD is a more general and complex task than the well known Named Entity Recognition (NER) task which aims solely at the identification and classification of the named mentions.

The difficulty of the MD task is directly related to the nature of the language and the linguistic resources available, i.e. it is easier to build accurate MD systems for languages with a simple morphology and a high amount of linguistic resources. For this reason, we explore the idea of using a MD system which has been designed and built for a resource-rich language (RRL), to help enhance the performance of a MD system in a target language (TL). More specifically, the research work we present in this paper is that because English received significant attention in terms of creating and refining natural language resources, it would be highly relevant to use those same resource to raise the accuracy of MD systems in other languages. For instance, an English MD system might achieve a performance of $F_{\beta=1}$-measure=82.7 (Zitouni and Florian, 2009) when it resorts to a rich set of features extracted from diverse resources, namely: part-of-speech, chunk information, syntactic parse trees, word sense information, WordNet information and information from the output of other mention detection classifiers. In this paper, our research question revolves around how to use such a system to the benefit of other languages such as Arabic, Chinese, French or Spanish

[1] http://www.itl.nist.gov/iad/mig/tests/ace/2007/doc/ace07-evalplan.v1.3a.pdf

MD systems, which also have annotated resources but not at the same level of quantity and/or quality as English.

In this paper, we have targeted English and Arabic as the RRL and TL, respectively, because:

1. We have a very competitive English MD system;
2. The linguistic resources available for the Arabic language allow a simulation of different TL richness levels; and
3. The use of two languages of an utterly different nature makes the extrapolation of the results to other languages possible.

Our hypothesis might be expressed as follows: using a MD system resorting to a rich feature set (i.e. the RRL MD system) to boost a MD system performance in a TL can be very beneficial if the "donor" system surpasses its TL counterpart in terms of resources. To test this hypothesis, we have projected MD tags from RRL to TL via a parallel corpus, and then extracted several linguistic features about the automatically tagged words. Thereafter, we have conducted experiments adding these new features to the TL baseline MD system. In order to have a complete picture on the impact of these new features, we have used TL baseline systems resorting to a varied amount of features, starting with a case employing only lexical information to a case where we use all the resources we could gather for the TL. Experiments show that the gain is always statistically significant and it reaches its maximum when only very basic features are used in the baseline TL MD system.

2 Mention Detection

Similarly to classical NLP tasks, such as Base Phrase Chunking (Ramshaw and Marcus, 1999) (BPC) or NER (Tjong Kim Sang, 2002), we formulate the MD task as a sequence classification problem, i.e. the classifier assigns to each token in the text a label indicating whether it starts a specific mention, is inside a specific mention, or is outside any mentions. It also assigns to every non outside mention a class to specify its type: e.g., person, organization, location, etc. In this study, we chose the Maximum Entropy Markov Model (MEMM henceforth) approach because it can easily integrate arbitrary types of information in order to make a classification decision. To train our models, we have used the *Sequential Conditional Generalized Iterative Scaling* (SCGIS) technique (Goodman, 2002). This techniques uses a *Gaussian prior* for regularization (Chen and Rosenfeld, 2000). The features used by our MD systems can be divided into the following categories:

1- *Lexical*: these are token n-grams directly neighboring the current token on both sides, i.e. left and right. Empirical results have shown that the optimal span is $n = 3$.

2- *Syntactic*: they consist of the outcomes of several Part-Of-Speech (POS) taggers and BPCs trained on different corpora and different tag-sets in order to provide the MD system with a wider variety of information. Our model uses the POS and BPC information appearing in window of 5 (current, two previous, and two next) jointly with the tokens.

Both the English and the Arabic MD systems have access to lexical and syntactic features. The former one, however, also employs a set of features obtained from the output of other MD classifiers. In order to provide the MD system with complementary information, these classifiers are trained on different datasets annotated for different mention types, e.g. dates or occupation references (not used in our task).

3 Annotation, Projection and Feature Extraction

We remind the reader that our main goal is to use a RRL MD system to enhance the performance of a MD system in another language, i.e. the TL. In order to achieve this goal, we propose an approach that uses a RRL-to-TL parallel corpus to bridge between these two languages. This approach performs in three main steps, namely: annotation, projection and feature extraction. In this section, we describe in details each of these steps.

3.1 Annotation

This first step consists of MD tagging of the RRL side of the parallel corpus. Because in our case study we have chosen English as the RRL, we have used an accurate English MD system to perform the annotation step. Our English MD system achieves an F-measure of 82.7 (Zitouni and Florian, 2009) and has achieved significantly competitive results at the ACE evaluation campaign.

3.2 Projection

Once the RRL side of the parallel corpus is accurately augmented with MD tags, the projection step comes to transfer those tags to the TL side, Arabic in our case study, using the word alignment information. We illustrate the projection step with a relevant example. Let consider the following MD tagged English sentence:

*Bill/**B-PER-NAM** Clinton/**I-PER-NAM** is visiting North/**B-GPE-NAM** Korea/**I-GPE-NAM** today*

where "*Bill Clinton*" is a named person mention and "*North Korea*" is a named geopolitical entity (GPE) one. A potential Arabic translation of this sentence would be:

بيل كلينتون يزور كوريا الشمالية اليوم

which might be transliterated as:

byl klyntwn yzwr kwryA Al$mAlyA Alywm

After projecting the English mentions to the Arabic text, we obtain the following:

*byl/**B-PER-NAM** klyntwn/**I-PER-NAM** yzwr kwryA/**B-GPE-NAM** Al$mAlyp/**I-GPE-NAM** Alywm*

This tagged version of the Arabic text is provided to the third module of the process responsible on feature extraction (see Subsection 3.3). It is, however, pertinent to point out that the example we have used for illustration is relatively simple in the sense that almost all English and Arabic words have a 1-to-1 mapping. In real world translation (both human and automatic), one should expect to see 1-to-n, n-to-1 mappings as well as unmapped words on both sides of the parallel corpus rather frequently.

As stated by (Klementiev and Roth, 2006), the projection of NER tags is easier in comparison to projecting other types of annotations such as POS-tags and BPC[2], mainly because:

1. *Not all the words are mentions*: once we have projected the tags of the mentions from the RRL to TL side, the rest of tokens are simply considered as outside any mentions. This is different from the POS-tag and BPC where all the words are assigned a tag and thus when a word is unmapped, further processing is required (Yarowsky et al., 2001);

2. *In case of a 1-to-n mapping, the target n words are assigned the same class*: for instance, let consider the English GPE named mention "North-Korea". The segmented version of its Arabic translation would be "كوريا ال شمالية" (kwrya Al $mAlyp). The projection process consists in simply assigning the same class, i.e. GPE, to all Arabic tokens. The problem takes another dimension, however, in the case of propagating the POS-tags, because "North" is a NNP aligned with the determinant (DET) "Al" and the NNP "$mAlyp". Additional processing is needed to handle this difference of tags on the two sides.

3. *In case of n-to-1 mapping, the TL side word is simply assigned the class propagated from the RRL side*. For instance, if on the English side we have the named person multi-word mention "Ben Moussa", translated into the one-word mention بنموسى (bn-mwsY) on the Arabic side, then projection consists of simply assigning the person named tag to the Arabic word.

However, in our research study, new challenges arose because our RRL data are automatically annotated, which is different from what has been reported in the research works we have mentioned before, i.e. (Yarowsky et al., 2001) and (Klementiev and Roth, 2006), where gold annotated data were used. In order to relax the impact of the noise introduced by the English MD system, we :

1. *use mention "splits" to filter annotation errors:* We assume that when a sequence of tokens is tagged as a mention on the RRL side, its TL counterpart should be an uninterrupted sequence of tokens as well. When the RRL MD system captures incorrectly the span of a mention, e.g. in the sentence "*Dona Karan international reputation of ...*", the RRL MD system might mistakenly tag "Dona Karan international" as an organization mention instead of tagging "Dona Karan" as a person mention. It is possible to detect this type of errors on the TL side because "dwnA kArAn" (Dona Karan) is distant from "Al EAlmyp" (international), i.e. they do not form an uninterrupted token sequence. We use this "split" in the mentions as information in order to not use these mentions in the feature extraction step (see Subsection 3.3).

2. *do not use the projected mentions directly for training:* Instead, we use these tags as additional features to our TL baseline model and allow our MEMM classifier to weigh them according to their relevance to each mention type.

3.3 Feature Extraction

At this point, the parallel corpus should be annotated with mentions on both of its sides. Where the RRL side is tagged using the English MD system during the annotation step (c.f section 3.1) while the TL side is annotated by the propagation of these MD tags via the parallel corpus in the projection step (c.f. section 3.2). In this third step, the goal is to extract pertinent linguistic features of the automatically tagged TL corpus to enhance

[2] The claim is also valid for MD because it is the same type of annotation.

MD model in the TL. The explored features are as follows:

1. **Gazetteers:** we group mentions by class in different dictionaries. During both training and decoding, when we encounter a token or a sequence of tokens that is part of a dictionary, we fire its corresponding class; the feature is fired only when we find a complete match between sequence of tokens in the text and in the dictionary.

2. **Model-based features:** it consists of building a model on the automatically tagged TL side of the parallel corpus. The output of this model is used as a feature to enhance MD model in the target language. However, it is also possible to use this model to directly tag text in the TL. This would be useful in cases where we do not have any TL annotated data.

3. **n-gram context features:** it consists of using the annotated corpus in the TL to collect n-gram tokens surrounding a mention. We organize those contexts by mention type and we use them to tag tokens which appear in the same context in both the training and decoding sets. These tags will be used as additional feature in the MD model. For instance, if we consider that the person mention صدام حسين (SdAm Hsyn - Sadam Husein) appears in the following sentence:

صرح أمس أن صدام حسين يترأس نظاما فاشلا

which might be transliterated as: *SrH Ams An SdAm Hsyn ytrAs nZAmA fA$lA* and translated to English as: *declared yesterday that Sadam Husein governs a failed system*

the context n-grams that would be extracted are:

. **Left n-grams:** W_{-1}=أن (An - that), W_{-2}=أمس أن (Ams An - yesterday that), etc.

. **Right n-grams:** W_{+1}=يترأس (ystrAs - governs), W_{+2}=يترأس نظاما (ytrAs nZAmA - governs a system), etc.

. **Left and right n-grams:** joint feature of the two previous ones, W_{-i} and W_{+i}.

For both training and test data we create a new feature stream where we indicate that a token sequence is a mention if it appears in the same n-gram context.

4. **Head-word based features:** it considers that the lexical context in which the mention appeared is the sequence of the parent sub-trees head words in a parse-tree. For instance, if we consider the sentence which we have used in the previous example, the corresponding parse tree is shown in Figure 1.

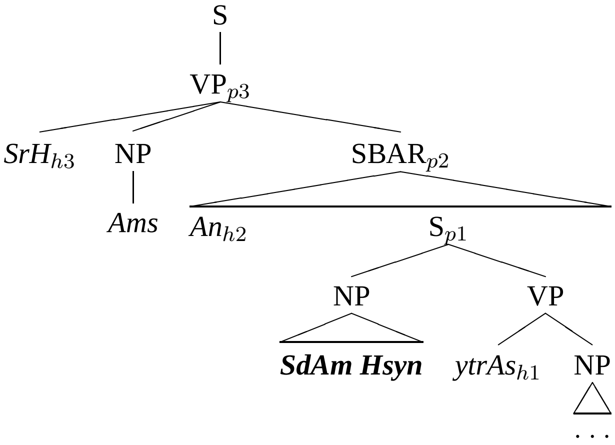

Figure 1: Parse tree

The parent sub-tree heads of ‘SdAm Hsyn’ are marked with h_i on the tree. Similarly to the other features, in both training and decoding sets, we create a new feature stream where we tag those token sequences which appear with the same n first parent sub-tree head words as a person mention in the annotated TL data.

5. **Parser-based features:** it attempts to use the syntactic environment in which a mention might appear. In order to do so, for each mention in the target language corpus we consider only labels of the parent non-terminals .We mark parent non-terminal labels of ‘SdAm Hsyn’ on the tree with p_i. Similarly to the features described above, we create during both training and test a new feature stream where we indicate the token sequences which appear in the same parent non-terminal labels.

Gazetteers and model-based features are the most natural and expected kind of features that one would extract from the automatically MD tagged version of the TL text. Our motivation of using n-gram context features, on one hand, and the head-word based and parse-based features on the other is to: (i) contrast the impact of local and global context features; and (ii) experiment the possibility of employing both of them jointly in order to test their complementarity.

4 The Target Language Mention Detection System

- **The Arabic language:** In our research study, we have intentionally chosen a TL which has a different strategy to form the words and the sentences than English. By doing so, we are seeking to avoid obtaining results which are biased by the similarity of the employed languages. For this reason, we have chosen the Arabic as a TL.

Due to its Semitic origins, the Arabic language is both *derivational*, i.e. it uses a templatic strategy to form a word, and *highly inflectional*, i.e. additional affixes might be added to a word in order to obtain further meaning. Whereas the former characteristic is common with most of the languages, the latter one, however, results in increasing *sparseness* in data and consequently forming an obstacle to achieve a high performance for most of the NLP tasks (Diab et al., 2004; Benajiba et al., 2008; Zitouni et al., 2005; Zitouni and Florian, 2008). From a NLP viewpoint, especially the supervised tasks such as the one we are dealing with in this paper, this implies that a huge amount of training data is necessary in order to build a robust model. In our study, to tackle the data sparseness problem, we have performed the *word segmentation*. This segmentation pre-processing step consists of separating the normal white-space delimited words into prefixes, stems, and suffixes. Thus, from a modeling viewpoint, the unit of analysis becomes the segments. We use a technique similar to the one introduced in (Lee et al., 2003) for segmentation with an accuracy of 98%.

- The Arabic MD system: Our Arabic MD system employs the same technique presented in Section 2. Compared to English MD model, Arabic MD system has access to morphological information (Stem) as we will explain next. Features used by the Arabic MD system are divided in three categories:

1. Lexical: Similar to the lexical features used by out English MD system (c.f. section 2);

2. Stem: This feature has been introduced in (Zitouni et al., 2005) as stem n-grams spanning the current stem; both preceding and following it. If the current token x_i is a stem, stem n-gram features contain the previous $n-1$ stems and the following $n-1$ stems. Stem n-gram features represent a lexical generalization that reduce data sparseness;

3. Syntactic: it consists of the output of POS taggers and the BPCs.

As we describe with more details in the experiments section (see Section 6), once we have extracted the new features from the parallel corpus, we contrast their impact with the level of richness in features of the TL MD system, i.e. we measure the impact of each feature f_i when the TL MD system uses: (i) only lexical features; (ii) both lexical and stem features; and (iii) lexical, stem and syntactic features.

5 Evaluation Data

Experiments are conducted on the Arabic ACE 2007 data. There are 379 Arabic documents and almost $98,000$ words. We find 7 classes of mentions: Person (PER), Organization (ORG), Geo-Political Entity (GPE), Location (LOC), Facility (FAC), Vehicle (VEH), Weapon (WEA). Since the evaluation test sets 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) for training and 56 documents ($18,000$ words) as a test set. This results in $17,634$ mentions ($7,816$ named, $8,831$ nominal and 987 pronominal) for training and $3,566$ for test ($1,673$ named, $1,682$ nominal and 211 pronominal). To facilitate future comparisons with work presented here, and to simulate a realistic scenario, the splits are created based on article dates: the test data is selected as the latest 15% of the data in chronological order, in each of the covered genres (newswire and webblog).

While performance on the ACE data is usually evaluated using a special-purpose measure - the ACE value metric, given that we are interested in the mention detection task only, we decided to use the more intuitive and popular (un-weighted) F-measure, the harmonic mean of precision and recall.

6 Experiments and Results

As we have stated earlier, our main goal is to investigate how a MD model of a TL might benefit from additional information about the mentions obtained by propagation from a RRL. In our research study we have chosen Arabic as the TL and English as the RRL. The English MD system we use has access to a large set of information (Zitouni and Florian, 2009) and has achieved a performance of 82.7F on ACE'07 data. In order to simulate different levels of resource-richness for the TL, we have employed four baseline systems which use different feature-sets. Following we present these feature-sets ranked from the resource-poorest to the resource-richest one: 1- $Lex.$: lexical features; 2- $Stem.$: $Lex.$ + stem features; and 3- $Syntac.$: $Stem.$ + syntactic features.

For each of these baseline systems, we study the impact of features extracted from the parallel corpus (c.f. Section 3) separately. We report the following results:

1- $Base.$: baseline system *without* the use of parallel-data extracted features;

2- $n-Lex.$: $Base. +$ n-gram context features;

3- $n-Head$: $Base. +$ head-word based features;

	$Lex.$	$Stem$	$Syntac$
$Base.$	74.14	74.47	75.53
$n-Lex.$	74.71	75.25	76.20
$n-Head$	74.63	75.29	75.93
$n-Pars.$	75.32	75.19	75.74
Gaz	74.90	74.79	75.66
$Model$	74.60	75.50	76.22
$Comb.$	**76.01**	**76.74**	**77.18**

Table 1: Obtained results when the features were extracted from a hand-aligned parallel corpus

4- $n-Pars.$: $Base.$ + parser-related features;
5- $Gaz.$: $Base.$ + automatically extracted gazetteers from the parallel corpus;
6- $Model$: $Base.$ + output of model trained on the Arabic part of the parallel corpus;
7- $Comb.$: combination of all the above.
In the rest of the paper, to measure whether the improvement in performance of a system using features from parallel data over baseline is statistically significant or not, we use the stratified bootstrap resampling significance test (Noreen, 1989) used in the NER shared task of CoNLL-2002[3]. We consider results as statistically significant when $p < 0.02$.

6.1 Hand-aligned Data

In our first experiment-set we use a hand-aligned English-to-Arabic parallel corpus of 1 million words approximately. After tagging the Arabic side by projection we obtain 86.5K mentions. As we have previously mentioned, in order to generate the model-based feature, $Model$, we have trained a model on the Arabic side of the parallel corpus. This model achieved an F-measure of 57.7F. This shows the performance that might be achieved when *no human annotated data* is available in the TL.

Results in Table 1 show that a significant improvement is obtained when the TL is poor in resources; for instance an improvement of ∼1.9 points was achieved when the TL uses only lexical features. The use of $n-Pars.$ features alone yielded 1.2 points of improvement. when the TL model uses a rich feature-set, we still can obtain ~ 1.7 points improvement. When the TL baseline model employs the $Syntac$ feature-set, the greatest improvement is obtained when we add the model-based feature. Improvement obtained by the system using $Comb.$ features is statistically significant compared to baseline model. This system also outperforms systems using the new feature set separately across the board. According to our error-analysis, the significant amount of Arabic mentions observed in the parallel corpus, where many of them do not appear in the training corpus, has significantly helped the $Lex.$, $Stem$ and $Syntac$ MD models to capture new mentions and/or correct the type assigned. Some of the relevant examples in out data are: (i) the facility mention مبنى بلفور (mbnY blfwr - Belvoir Building); (ii) the GPE mention كابول (kAbwl - Kabul); and (iii) the person mention البعثيّن (AlbEvyyn - the Baathists). These mentions have only been tagged correctly when we have added the new extracted features to our model.
In other words, the error-analysis points out clearly that one possible way to get further improvement is to increase the parallel data in order to increase the number of matches between (1) *the number of mentions which are wrongly tagged by the TL MD model* and (2) *the number of mentions in the TL side of the parallel corpus*. The second parameter can be, indirectly, increased by increasing the size of the parallel data. Getting 10 or 20 times more of parallel data that is hand-aligned is expensive and requires several months of human/hours work. For this reason we opted for using an unsupervised approach by selecting a parallel corpus that is automatically aligned as we discuss in the next section.

6.2 Automatically-aligned Data

We have used for this experiment-set an Arabic-to-English parallel data of 22 million words. The data in this corpus is automatically aligned using a technique presented in (Ittycheriah and Roukos, 2005). The alignment is one-to-many with a performance around 87 F-measure.

Because we are dealing with large amount of data and also because the word alignment is done automatically, meaning more noise, we have used the English MD model confidence for additional filtering. Such filtering consists in keeping, from the parallel corpus, only sentences which have all tokens tagged with a confidence greater than α. In this paper, we use a value of $\alpha = 0.94$, which results in a corpus of 17 million words. We notice that a lower value of α results in a radical increase in noise. Because of space limitation, we will report results only with this value of α.

Table 2 shows the obtained results for parallel-data based features using the 17M subset. Dif-

[3] http://www.cnts.ua.ac.be/conll2002/ner/

	$Lex.$	$Stem$	$Syntac$
$Base.$	74.14	74.47	75.53
$n - Lex.$	74.27	74.74	75.24
$n - Head.$	74.07	74.95	75.33
$n - Pars.$	75.62	75.22	76.02
Gaz	73.96	74.11	74.94
$Model$	74.87	75.12	75.76
$Comb.$	**75.56**	**75.93**	**76.46**

Table 2: Obtained results when the features were extracted from a automatically-aligned parallel corpus

ferently from experiments using hand-aligned data, best results have been obtained when we have used the parser-based feature, i.e. $n - Pars.$ On one hand, the overall behavior is comparable to the one obtained when using the 1M hand-aligned parallel data (see Table 1), i.e. (i) the greatest improvement has been obtained when the TL uses a poor feature-set; and (ii) when the TL baseline model is rich in resources, we still obtain 0.45 points absolute improvement when using $n - Pars.$ On the other hand, features extracted from automatically-aligned data, in comparison with the ones extracted from the hand aligned data, have helped the MD model to correct many of the TL baseline model false negatives. This has been observed when the TL baseline system uses a rich feature set as well. A side effect of the noisy word alignment, however, was an increase in the number of false positives. For instance, the word مستحضرات (mstHDrAt - preparations) which appeared in the following sentence:

عدم السماح لمستحضرات أخرى

which might be transliterated as:

Edm AlsmAH lmstHDrAt AxrY

and translated to English as:

not to allow other preparations

has been tagged as an organization mention because it has been mistakenly aligned, in the parallel corpus, with the word كاو, *KO*, in the sentence:

شركة كاو الكبرى للّمستحضرات التجميلية

meaning:

The big cosmetics company KO.

Class	Num. of mentions
FAC	285
GPE	2,145
LOC	239
ORG	1,135
PER	2,474
VEH	65
WEA	138

Table 3: Distribution over the classes of the blind test mentions

	$Lex.$	$Stem$	$Syntac$
$Base.$	74.26	73.54	73.61
$n - Lex.$	74.04	73.72	73.83
$n - Head$	74.14	73.64	73.83
$n - Pars.$	74.32	74.18	74.32
Gaz	71.49	72.13	73.39
$Model$	**75.01**	**74.66**	**74.78**

Table 4: Obtained results on blind test

In order to validate our results, we run our experiments on a blind test-set. We have selected the latest 5% of each genre of the hand-aligned data and they have been manually annotated by a human. The blind test-set consists of 51,781 tokens of which 6,481 are mentions. Table 3 shows the distribution of these mentions over the different classes. The results are shown in Table 4. These results confirm the conclusions we have deduced from the ones previously presented in Table 2, i.e.: (i) the highest improvement is obtained when the TL is resource-scarce.

6.3 Combining Hand-aligned and Automatically-aligned Data

Table 5 shows that combining both features extracted from hand-aligned and automatically-aligned corpora has led to better results. The improvement of using $Comb.$ compared to baseline is statistically significant. We again notice that when

	$Lex.$	$Stem$	$Syntac$
$Base.$	74.14	74.47	75.53
$n - Lex.$	74.60	75.08	75.58
$n - Head$	74.51	75.32	75.56
$n - Pars.$	75.46	75.90	76.22
Gaz	74.85	74.83	75.92
$Model$	74.83	75.59	75.40
$Comb.$	**76.39**	**76.85**	**77.23**

Table 5: Obtained results when the features were extracted from both hand-aligned and automatically-aligned parallel corpora

the TL baseline MD model uses a richer feature set, the obtained improvement from using RRL becomes smaller. We also observed that automatically aligned data helped capture most of the unseen mentions whereas the hand-aligned features helped decrease the number of false-alarms. It is important to notice that when features $Comb.$ is used with $Stem$ baseline model, the obtained F-measure (76.85) is 1.3 higher than the baseline model which uses lexical, stem and syntactic features – $Syntac$ (75.53). The type of errors which mostly occur and has not been fixed neither by using hand-aligned data, automatically aligned data nor the combination of both are the nominal mentions whose class depends fully on the context. For instance, the word موظف (mwZf - employee) which was considered as **O** by the MD model because it has not been seen in any of the parallel data in a context such as the following:

تعريف شكل الموظف المصري كان ...

transliterated as:

tEryf $kl AlmwZf AlmSry ...

and translated as: *"defining the life of the Egyptian employee ..."*

7 Previous Works

Several research works, in different NLP tasks, have shown that the use of a RRL to achieve a better performance in a resource-challenged language yields to successful results. In (Rogati et al., 2003), authors used a statistical machine translation (MT) system to build an Arabic stemmer. The obtained stemmer has a performance of 87.5%. In (Ide et al., 2002), authors use the aligned versions of George Orwell's *Nineteen Eighty-Four* in seven languages in order to determine sense distinctions which can be used in the Word Sense Disambiguation (WSD) task. They report that the automatically obtained tags are at least as reliable as the one made by human annotators. Similarly, (Ng et al., 2003) report a research study which uses an English-Chinese parallel corpus in order to extract sense-tagged training data. In (Hwa et al., 2002), authors report promising results of inducing Chinese dependency trees from English. The obtained model outperformed the baseline.

One of the significant differences between these works and the one we present in this paper is that instead of using the propagated annotation directly as training data we use it as an additional feature and thus allow the MEMM model to weigh each one of them. By doing so, the model is able to distinguish between the relevant and the irrelevant information propagated from the RRL.

Authors in (Zitouni and Florian, 2008) attempt to enhance a MD model of a foreign language by using an English MD system. They have used an MT system to (i) translate the text to English; (ii) run the English model on the translated text; (iii) and propagate outcome to the original text. The approach in (Zitouni and Florian, 2008) requires a MT system that needs more effort and resources to build when compared to a parallel corpus (used in our experiments); not all institutions may have access to MT and MD systems in plenty of language pairs.

8 Conclusions and Future Works

In this paper, we have presented our investigation of exploiting the richness, in terms of resources, of one language (English) to the benefit of a target language (Arabic). We have achieved successful results by adopting a novel approach performing in three main steps, namely: (i) Annotate the English side of an English-to-Arabic parallel corpus automatically; (ii) Project the obtained annotation from English to Arabic via the parallel corpus; and (iii) Extract features of different linguistic motivations of the automatically tagged Arabic tokens. Thereafter, each of the extracted features is used to bootstrap Arabic MD system. We use different Arabic baseline MD models which employ different feature sets representing different levels of richness in resources. We also use both a 1 million word hand-aligned parallel corpus and a 22 million word automatically aligned one in order to study size vs. noise trade-off.

Results show that a statistically significant improvement is always observed even when the Arabic baseline MD model uses all the available resources. When we use the hand-aligned parallel corpus, we obtain up to 2.2 points improvement when the Arabic MD model has access to very limited resources. It decreases to 1.7 points when we use all the resources we could gather for the Arabic language. When no human-annotated data is available in the TL, we show that we can obtain a performance of 57.6 using only mention propagation from RRL. The results also show that a greater improvement is achieved when using a small hand-aligned corpus than using a 20 times bigger automatically aligned data. However, in case both of them are available, combining them leads to even higher results.

References

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

Stanley Chen and Ronald Rosenfeld. 2000. A survey of smoothing techniques for ME models. *IEEE Transaction on Speech and Audio Processing*.

Mona Diab, Kadri Hacioglu, and Dan Jurafsky. 2004. Automatic tagging of arabic text: from raw text to base phrase chunks. In *Proc. of HLT/NAACL'04*.

Joshua Goodman. 2002. Sequential conditional generalized iterative scaling. In *Proceedings of ACL'02*.

Rebecca Hwa, Philip Resnik, and Amy Weinberg. 2002. Breaking the resource bottleneck for multilingual parsing. In *Proceedings of LREC*.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*, pages 54–60.

Abe Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of HLT/EMNLP'05*, pages 89–96.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of ACL'06*, pages 817–824, Sydney, Australia. Association for Computational Linguistics.

Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. 2003. Language model based Arabic word segmentation. In *Proc. of the ACL'03*, pages 399–406.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of ACL'03*, pages 455–462.

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

Lance Ramshaw and Mitchell Marcus. 1999. Text chunking using transformation-based learning. In S. Armstrong, K.W. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 157–176. Kluwer.

Monica Rogati, Scott McCarley, and Yiming Yang. 2003. Unsupervised learning of arabic stemming using a parallel corpus. In *Proceedings of ACL'03*, pages 391–398.

Eric. F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT'01*, pages 1–8.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, Hawaii, October.

Imed Zitouni and Radu Florian. 2009. Cross-language information propagation for arabic mention detection. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–21.

Imed Zitouni, Jeff Sorensen, Xiaoqiang Luo, and Radu 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.