Mention Detection Crossing the Language Barrier

Imed Zitouni and Radu Florian
IBM T.J. Watson Research Center
1101 Kitchawan Rd, Yorktown Heights, NY 10598
{izitouni, raduf}@us.ibm.com

Abstract

While significant effort has been put into annotating linguistic resources for several languages, there are still many left that have only small amounts of such resources. This paper investigates a method of propagating information (specifically mention detection information) into such low resource languages from richer ones. Experiments run on three language pairs (Arabic-English, Chinese-English, and Spanish-English) show that one can achieve relatively decent performance by propagating information from a language with richer resources such as English into a foreign language alone (no resources or models in the foreign language). Furthermore, while examining the performance using various degrees of linguistic information in a statistical framework, results show that propagated features from English help improve the source-language system performance even when used in conjunction with all feature types built from the source language. The experiments also show that using propagated features in conjunction with lexically-derived features only (as can be obtained directly from a mention annotated corpus) yields similar performance to using feature types derived from many linguistic resources.

1 Introduction

Information extraction is a crucial step toward understanding a text, as it identifies the important conceptual objects and relations between them in a discourse. It includes classification, filtering, and selection based on the language content of the source data, i.e., based on the meaning conveyed by the data. It is a crucial step for several applications, such as summarization, information retrieval, data mining, question answering, language understanding, etc. This paper addresses an important and basic task of information extraction: *mention detection*[1]: the identification and classification of textual references to objects/abstractions *mentions*, which can be either named (e.g. John Smith), nominal (the president) or pronominal (e.g. he, she). For instance, in the sentence

President *John Smith* said *he* has no comments.

there are three mentions: *President, John Smith* and *he*. This is similar to the named entity recognition (NER) task with the additional twist of also identifying nominal and pronominal mentions.

A few languages have received a lot of attention in terms of natural language resources that were created – for instance, in English one has access to labeled part-of-speech data, word sense information, parse tree structure, discourse, semantic role labeles, named entity data, to name just a few (our apologies if we missed your favorite resource). There are a few other languages that also have annotated resources (such as Arabic, Chinese, German, French, Spanish, etc), but also a very large number of languages with few resources. It would be very useful if one could make use of the resources in the former languages to help bootstrapping (or just the projection) of resource in any resource-challenged language.

Information transfer from a language to another can be very useful when the "donor" language has more resources than the receiving one. As resources grow in quantity and quality in the receiving language, it becomes less and less likely that there will be a gain in performance by transfering information, as there are several sources of noise involved in the

[1] We adopt here the ACE (NIST, 2007) nomenclature

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 600–609,
Honolulu, October 2008.

process - such as the translation (machine generated or not) and the inherent imperfection of the mention detection in the donor language. To test this hypothesis, we conducted experiments on systems build with a varied amount of resources in the receiving language, starting with the case where there are none[2] (all information is transferred through translation alignment), and ending with the case where we used all the resources we could gather for that language. The experiments will show that the gain in performance decreases with the amount of resources used in the source language, but, still, even when all resources were used, a statistically significant gain was still observed.

Similarly to classical NLP tasks such as text chunking (Ramshaw and Marcus, 1995) and named entity recognition (Tjong Kim Sang, 2002), we formulate mention detection as a sequence classification problem, by assigning a label to each token in the text, indicating whether it starts a specific mention, is inside a specific mention, or is outside any mentions. The classification is performed with a statistical approach, built around the maximum entropy (MaxEnt) principle (Berger et al., 1996), that has the advantage of combining arbitrary types of information in making a classification decision.

2 Previous Work

There are several investigations in literature that explore using parallel corpora to transfer information content from one language (most of the time English) to another. The earliest investigations of the subject have been performed, on word sense disambiguation (Dagan et al., 1991; P.F.Brown et al., 1991; Gale et al., 1992) (perhaps unsurprisingly given its close connection to machine translation) – all propose and (lightly) evaluate methods to use word sense information extracted from the target language to help the sense resolution in the source language and machine translation. (Dagan and Itai, 1994) explicitly suggests performing word sense disambiguation in the target language (English in the article) with the goal of resolving ambiguity in the source language (Hebrew), and show moderate improvement on a small data set[3]. More recently, (Diab and Resnik, 2001) presents a method for performing word sense tagging in both the source and target texts of parallel bilingual corpora with the English WordNet sense inventory, by using translation correspondences.

On more general cross-language information transfer, (Yarowsky et al., 2001) proposed and evaluated a method of propagating POS tagging, named mention, base noun phrase, and morphological information from English into a foreign language, which is very similar to the one presented in this article (experiments were run on French, Chinese, Czech, and Spanish – on human-generated translations). Their results show a significant improvement in performance while building an automatic classifier on the projected annotations over the same automatic classifier trained on a small amount of annotated data in the source language. (Riloff et al., 2002) extends the ideas in (Yarowsky et al., 2001), by showing how it can be used, in conjunction with an automatically trained information extraction system on the source language, to bootstrap the annotation of resources in the target language. They show that they can obtain 48 F-measure on a information extraction task identifying locations, vehicles and victims in plane crashes. (Hwa et al., 2002) proposes a framework that enables the acquisition of syntactic dependency trees for low-resource languages by importing linguistic annotation from rich-resource languages (English). The authors run a large-scale experiment in which Chinese dependency parses were induced from English, and show that a parser trained on the resulting trees outperformed simple baselines. (Cabezas et al., 2001) investigates a similar method of propagating syntactic treebank-like annotations from English to Spanish.

Finally, a large body of research has been done on *cross-language information retrieval*, where the goal is to find information in one language (e.g. Chinese newswire) corresponding to a query in a different language (e.g. English) – although the list of relevant papers is too long to be mentioned here (see, for instance, (Grefenstette, 1998)).

The work presented here differs from the information extraction investigations presented above in two aspects:

- it handles unrestricted text and a full set of

[2]While applying this method in the case where the source language has absolutely no resources might be an interesting test case, we don't see it as being realistic. Resources are build nowadays in a large variety of languages, and not making use of them is rather foolish (a certain big bird and sand comes to mind).

[3]Very small by "modern" standards - 137 examples. Probably because at the time the article was written, there were no large publicly annotated databases, such as Semcor.

mention types (the ACE entity types) during the information transfer

- it investigates whether using a resource-rich language (English) can improve on the performance obtained by using various degrees of existent resources in the source language (Arabic, Chinese, Spanish)
- the information transfer is performed over machine generated translations and alignments.

3 Mention Detection

As mentioned in the introduction, the mention detection problem is formulated 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 any mentions.

Good performance in many natural language processing tasks has been shown to depend heavily on integrating many sources of information (Florian et al., 2004).[4] Given this observation, we are interested in algorithms that can easily integrate and make effective use of diverse input types. We select a exponential classifier, the Maximum Entropy (MaxEnt henceforth) classifier that integrates arbitrary types of information and makes a classification decision by aggregating all information available for a given classification. But the reader can replace it with her favorite feature-based classifier throughout the paper.

To help with the presentation, we introduce some notations: let $\mathcal{Y} = \{y_1, \ldots, y_n\}$ be the set of predicted classes, $\mathcal{X}$ be the example space and $\mathcal{F} = \{0,1\}^m$ be a feature space. Each example $x \in \mathcal{X}$ has associated a vector of m binary features $f(x) = (f_1(x), \ldots, f_m(x))$. The goal of the training process is to associate examples $x \in \mathcal{X}$ with either a probability distribution over the labels from $\mathcal{Y}$, $P(\cdot|x)$(if we are interested in *soft* classification) or associate one label $y \in \mathcal{Y}$ (if we are interested in *hard* classification).

The MaxEnt algorithm associates a set of weights $\{\alpha_{ij}\}_{j=1\ldots m}^{i=1\ldots n}$ with the features $(f_j)_i$, and computes the probability distribution as

$$P(y_i|x) = \frac{1}{Z(x)} \prod_{j=1}^{m} \alpha_{ij}^{f_j(x,y_i)}, \quad (1)$$

$$Z(x) = \sum_i \prod_j \alpha_{ij}^{f_j(x,y_i)}$$

where $Z(x)$ is a normalization factor. The $\{\alpha_{ij}\}_{j=1\ldots m}$ weights are estimated during the training phase to maximize the likelihood of the data (Berger et al., 1996). In this paper, the MaxEnt model is trained using the *sequential conditional generalized iterative scaling* (SCGIS) technique (Goodman, 2002), and it uses a *Gaussian prior* for regularization (Chen and Rosenfeld, 2000).

Now take $x_1^N = (x_1, x_2, \ldots x_N)$, a sequence of contiguous tokens (i.e., a sentence or a document) in the source language. The goal of mention detection system is to find the most likely sequence of labels $y_1^N = (y_1, y_2 \ldots y_N)$ that best matches the input x_1^N. In the mention detection case, each token x_i in x_1^N is tagged with a label y_i as follows:[5]

- if it's not part of any entity, $y_i = O$ (O for "outside any mentions")
- if it is part of an entity, it is composed of a subtag specifying whether it starts a mention (*B-*) or is inside a mention (*I-*), and a sub-type corresponding to mention type (e.g. *B-PERSON*). In ACE, there are seven possible types: person, organization, location, facility, geopolitical entity (GPE), weapon, and vehicle.

To compute the best sequence y_1^N, we use

$$\begin{aligned} y_1^N &= \arg\max_{\hat{y}_1^N} P\left(\hat{y}_1^N | x_1^N\right) \\ &= \arg\max_{\hat{y}} \prod P\left(\hat{y}_j | x_1^N, \hat{y}_1^{j-1}\right) \\ &= \arg\max_{\hat{y}} \prod_j P\left(\hat{y}_j | x_1^N, y_{j-k}^{j-1}\right) \end{aligned}$$

where $P\left(\hat{y}_j | x_1^N, y_{j-k}^{j-1}\right)$ has an exponential form of the type (2). We also used the standard Markov assumption that the probability $P\left(\hat{y}_j | x_1^N, \hat{y}_1^{j-1}\right)$ only depends on the previous k classifications. This model is similar to the MEMM model (McCallum et al., 2000), but it does not separate the probability into generation probabilities and transition probabilities, and, crucially, has access to "future" observed features (i.e. it can examine the entire x_1^N sequence, though in practice it will only examine some small part of it) – which is one way of eliminating label

[4] In fact, the feature set used for classification has a much larger impact on the performance of the resulting system than the classifier method itself.

[5] The mention encoding is the IOB2 encoding presented in (Tjong Kim Sang and Veenstra, 1999) and introduced by (Ramshaw and Marcus, 1994) for base noun phrase chunking.

bias observed by (Lafferty et al., 2001).[6]

The experiments are run on four languages, part of the ACE-2007 evaluation (NIST, 2007): Arabic, Chinese, English and Spanish.[7] Systems across the languages use a large range of features, including lexical (words and morphs in a 3-word window, prefixes and suffixes of length up to 4 characters, WordNet (Miller, 1995) for English), syntactic (POS tags, text chunks), and the output of other information extraction models. These features were described in (Florian et al., 2004), and are not discussed here. In this paper we focus on the examining the benefit of cross-language mention propagation information in improving mention detection systems.

Besides generic types of features, we also have implemented language-specific features:

- In Arabic, blank-delimited words are composed of zero or more prefixes, followed by a stem and zero or more suffixes. Each prefix, stem or suffix is a token; any contiguous sequence of tokens can represent a mention. Similar to the approaches described in (Florian et al., 2004) and (Zitouni et al., 2005), we decided to "condition" the output of the system on the segmented data: the text is segmented first into tokens and classification is then performed on tokens. The segmentation model is similar to the one presented by (Lee et al., 2003) and obtains an accuracy of 98%.
- In Chinese text, unlike in Indo-European languages, words neither are white-space delimited nor do they have capitalization markers. Instead of a word-based model, we build a character-based one, since word segmentation errors can lead to irrecoverable mention detection errors; Jing et al. (2003) also observes that character-based models are better performing than word-based ones. Word segmentation information is still useful and is integrated as an additional feature stream.
- In English and in Spanish mention detection systems are similar to those described in (Florian et al., 2004) where words are the tokens to classify.

[6] In fact their example of label bias can be trivially solved by allowing the classifier to examine features for subsequent words.

[7] The ACE data has the nice property of being consistent in annotations across these languages.

4 Cross-Language Mention Propagation

The approach proposed in this article requires a mention detection system build in a resource-rich language, and a *translation* from the source language to the resource-rich language, together with *word alignment*. This assumption is realistic: while truly parallel data (humanly created) might be in short supply or harder to acquire, adapting statistical machine translation (SMT) systems from one language-pair to another is not as challenging as it used to be (Al-Onaizan and Papineni, 2006). We also find that there is a large number of parallel corpora available these days which cover many language pairs. For example, for the European Union's 23 official languages we find 253 language pairs; each document in one language might have to be translated in all other 22 languages. This is in addition to parallel corpora one could get from books, including religious texts such as the Bible, that are translated to a large number of languages. On the other hand, even though mention detection system is important for many natural language processing applications, we still find lack of mention-annotated corpora in many languages. In the approach we propose below, the annotated corpus used to train the mention detection classifier does not have to be part of a parallel corpus.

To start the process, we first use a SMT system to translate the source unit (document or sentence) x_1^N into the resource-rich language, yielding the sequence $\xi_1^M = (\xi_1, \xi_2, \ldots \xi_M)$. Taking the sequence of tokens ξ_1^M as input, the MaxEnt classifier assigns a mention label to each token, building the label sequence $\psi_1^M = (\psi_1, \psi_2 \ldots \psi_M)$. Using the SMT-produced word alignment between source text x_1^N and translated text ξ_1^M (Koehn, 2004),we propagate the target labels ψ_1^M to the source language building the label sequence $\tilde{y}_1^N = (\tilde{y}_1, \tilde{y}_2 \ldots \tilde{y}_N)$.[8] As an example, if a sequence of tokens in the resource-rich language $\xi_i\xi_{i+1}\xi_{i+2}$ is aligned to $x_j x_{j+1}$ in the source language and if $\xi_i\xi_{i+1}\xi_{i+2}$ is tagged as a location mention, then the sequence $x_j x_{j+1}$ can be labeled as a location mention: B-LOC, I-LOC. Hence, each token x_i in x_1^N is tagged with a corresponding propagated label $\tilde{y}_i$ in $\tilde{y}_1^N$, $\tilde{y}_i = \phi\left(i, A, \psi_1^M\right)$, where A is the alignment between the source and resource-rich languages. In cases when the alignment is 1-to-1 the function becomes the identity, but one can imagine different scenarios which can be used in

[8] Or by using Giza++ if your favorite engine does not give you word alignment.

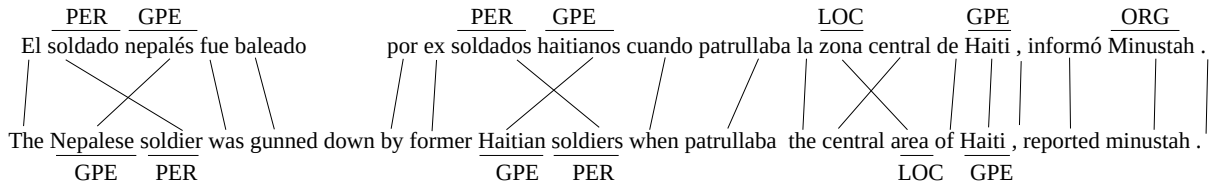

Figure 1: Word alignment for a Spanish sentence and its English machine-translation. The mention labels shown are the gold-standard ones for Spanish and the automatically detected ones for English. If mentions were to be propagated from English to Spanish, the last mention would be a miss, due to the fact that the English mention detection failed to identify 'minustah' as an organization.

many-to-many alignment cases. The alignement we use in this paper is 1-to-many ($\{1...n\}$) from the source language (eg., Arabic) to the resource-rich language (e.g., English). Once we use SMT word alignment to propagate label sequence ψ_1^M of ξ_1^M to the corresponding text x_1^N in the target language, we end up with a sequence of labels $\tilde{y}_1^N$ where for each token x_i in x_1^N we attach its label $\tilde{y}_i$ in $\tilde{y}_1^N$. Hence, we label te entire span and if the strategy results in two mentions where one contains the other, we eliminate the inner one.

Figure 1 displays the alignment between a Spanish sentence and its English automatic translation. It also shows a good match between the gold-standard tags in Spanish and the automatically extracted tags in English.

There are three ways in which we propose using these propagated labels:

1. Consider $\tilde{y}_1^N$ as the result of propagating the detected mentions in the original text x_1^N, basically selecting $y_1^N = \tilde{y}_1^N$. This situation corresponds to a case where no resources (annotated data) are available/needed on the source side, where the propagated labels are the output of the system.

2. Use the label sequence $\tilde{y}_1^N$ as an additional feature in the MaxEnt framework when predicting $P\left(y_j|x_1^N, y_{j-k}^{j-1}\right)$, together with other features built from resources available on the source language. We will call this model *CDP* (Context Dependent Propagation).

3. Starting with a large corpus (possibly including the training data), translate it into the resource-rich language and run mention detection. Then select the word sequences in the source language associated with the found mentions in the translation and add them to a machine-generated gazetteer $\mathcal{G}$[9]. This gazetteer $\mathcal{G}$ is then used to construct features for classification. We will call this model *CIP* (Context Independent Propagation).

From a runtime point of view, the CIP method has the advantage that there is no need to perform machine translation, and it can incorporate data from a very large amount of text. The CDP method, on the other hand, has the advantage that features are computed in context, and will not fire unless the corresponding mentions were found in the translated version (hence the name). Of course, the CDP method can incorporate features generated in the dictionary $\mathcal{G}$. The experimental section analyzes the impact of each of these techniques on mention detection task performance.

5 Resources

Experiments are conducted on the ACE 2007 data sets[10], in four languages: Arabic, Chinese, English, and Spanish. This data is selected from a variety of sources (broadcast news, broadcast conversations, newswire, web log, newswire, conversational telephony) and is labeled with 7 types: person, organization, location, facility, GPE (geo-political entity), vehicle and weapon. Besides mention level information, also labeled are coreference between the mentions, relations, events, and time resolution.

Since the evaluation tests set are not publicly available, we have split the publicly available *training* corpus into an 85%/15% data split. 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. This way, the documents in

[9]This is in fact a way to automatically construct a source-side mention dictionary.

[10]Same data as for ACE 2008.

Language	Training	Test
Arabic	323	56
Chinese	538	95
English	499	100
Spanish	467	52

Table 1: Datasets size (number of documents)

Language Pair	BLEU Score
Arabic-English	0.55
Chinese-English	0.32
Spanish-English	0.55

Table 3: BLEU performance of the SMT systems on the 3 language pairs

the training and test data sets do not overlap in time, and the content of the test data is more recent than the training data. Table 1 presents the number of documents in the training/test datasets for each of the four languages.

While performance on the ACE data is usually evaluated using a special-purpose measure - the ACE value metric (NIST, 2007), given that we are interested in the mention detection task only, we decided to use the more intuitive and popular (unweighted) F-measure, the harmonic mean of precision and recall.

6 Resource-Rich Languages

From the set of four languages in ACE 2007, we will unsurprisingly select English as the resource-rich language. Table 2 shows the performance of mention detection systems in all 4 languages one can obtain by using all available resources in that language, including lexical (words and morphs in a 3-word window, prefixes and suffixes of length up to 4, WordNet (Miller, 1995) for English), syntactic (POS tags, text chunks), and the output of other information extraction models.

	N	P	R	F
Arabic	3566	83.6	76.8	**80.0**
Chinese	4791	81.1	71.3	**75.8**
English	8170	84.6	80.8	**82.7**
Spanish	2487	79.1	73.5	**76.2**

Table 2: Performance of Arabic, Chinese, English and Spanish mention detection systems. Performance is presented in terms of Precision (P), Recall (R), and F-measure (F). The column (N) displays the number of mentions in the test set.

Results show that the English mention detection system has a better performance when compared to systems dealing with other languages such as Arabic, Chinese and Spanish. These results are not unexpected since the English model has access to a larger training data and uses richer set of information such as WordNet (Miller, 1995) and the output of a larger set of information extraction models.

7 Experiments

To show the effectiveness of cross-language mention propagation information in improving mention detection system performance in Arabic, Chinese and Spanish, we use three SMT systems with very competitive performance in terms of BLEU[11] (Papineni et al., 2002).

To give an idea of the SMT performance, Table 3 shows the performance of the translation systems on the three language pairs, computed on standard test sets. The Arabic to English SMT system is similar to the one described in (Huang and Papineni, 2007); it has 0.55 BLEU score on NIST 2003 Arabic-English machine translation evaluation test set. The Chinese to English SMT system has similar architecture to the one described in (Al-Onaizan and Papineni, 2006). This system obtains a score of 0.32 cased BLUE on NIST 2003 Arabic-English machine translation evaluation test set. The Spanish to English SMT system is similar to the one described in (Lee et al., 2006); it has a 0.55 BLEU score on the final text edition of the European Parliament Plenary Speech corpus in TC-STAR 2006 evaluation. As mentioned earlier, these three SMT systems have very competitive performance and are ranked among top 2 systems participating to NIST or TC-STAR evaluations. Also, the English mention detection system used for experiments has an F-measure of 82.7 and that has very competitive results among systems participating in the ACE 2007 evaluation.

Experiments are conducted under several conditions in order to investigate the effectiveness of our approach in improving mention detection system performance on languages with different levels of resource availability (from simple to more complex):

1. the system does not have access to any training data in the source language (no resources

[11] BLEU is an automatic measure for the translation quality which makes good use of multiple reference translations.

needed besides the MT system);

2. the system has access to only lexical information (information that can be directly derived exclusively from mention-labeled text);

3. the system has access to lexical and syntactic (e.g., POS tags, text chunks) information (requires mention-labeled text, and models to predict POS tags, etc);

4. the system that has access to lexical, syntactic, and semantic information (requires even more models and labeled data).

The rest of this section examines in detail these four cases.

To measure whether the improvement in performance of a particular system over another one is statistically significant or not, we use the stratified bootstrap re-sampling significance test (Noreen, 1989). This approach was used in the named entity recognition shared task of CoNNL-2002 (http://www.cnts.ua.ac.be/conll2002/ner/, 2002). In the following tables, we add a dagger sign † to results that are *not* statistically significant when compared to the baseline results.

7.1 No Source Language Training Data

In this first case, as described in Section 4, the mention labels in the source language are obtained directly through the alignment from the mentions in the translated text. This is a very simple scenario, which can be implemented with ease, and, as we will see, yields reasonable performance out-of-the-box.

	N	P	R	F
Arabic	3566	52.7	49.6	**51.1**
Chinese	4791	66.4	52.2	**58.5**
Spanish	2487	63.4	63.6	**63.5**

Table 4: Performance of the cross-language propagation from English mention detection system onto Arabic, Chinese and Spanish texts. Performance is presented in terms of Precision (P), Recall (R), and F-measure (F). The column (N) shows the number of mentions in the test set.

Experimental results presented in Table 4 show the performance of applying this information transfer approach. For each source language (Arabic, Chinese, or Arabic), we show the performance of propagating mentions from the English text. Even though no training data to build a source language mention classifier is available, we still can detect mentions with reasonably high accuracy. We consider the obtained accuracy as reasonably good because, as an example, the performance of a system that attaches to every word its most frequent label (unigram) is around 25% F-measure on Arabic. Results in Table 4 also show that even though the Chinese-to-English SMT system is lower in term of BLEU than the Arbic-to-English SMT system (0.32 vs. 0.55), performance of the cross-language propagation from English mention detection system onto Chinese is better than the performance of the propagation from English mention detection system onto Arabic. One reason for this is that we notice that Chinese-to-English SMT system translates and aligns ACE categories better than Arabic-to-English SMT system.

7.2 Lexical Resources

In this section, we consider the case when we have available training data in the source language to be able to train a statistical classifier. We also consider that the classifier has access to lexical information *only*. Our goal here is to study the effectiveness of *adding* cross-language mention propagation information to improve mention detection performance on languages with limited resources.

Table 5 shows the performance of the 3 languages with and without cross-language mention propagation information from English, with the 3 propagation methods described in Section 4. One can see that propagating mention propagation information results in system performance increase[12]. When systems use the CIP method, no improvement can be observed on Arabic and Chinese, while a small improvement of 0.5F point is obtained on Spanish (74.5 vs. 75.0). In contrast, when systems use the CDP method an improvement is obtained in recall – which is to be expected, given the method – leading to systems with better performance in terms of F-measure: 1.6F points improvement for Arabic, 1.5F points improvement for Chinese and almost 3F points improvement for Spanish. The results for all the CDP transfers and the CIP for Spanish are statistically significant.

7.3 Lexical and Syntactic Resources

We represent in Table 6 mention detection system performance when syntactic resources are available in the source language, in addition to lexical re-

[12]Only systems' performance marked with † is not statistically significantly better.

		Baseline			CIP			CDP		
	N	P	R	F	P	R	F	P	R	F
Arabic:	3566	81.8	71.7	**76.4**	82.2	71.3	**76.4**†	82.6	73.9	**78.0**
Chinese:	4791	79.3	70.2	**74.5**	79.4	70.5	**74.7**†	79.8	72.5	**76.0**
Spanish:	2478	79.1	70.4	**74.5**	79.7	70.8	**75.0**	80.4	74.6	**77.4**

Table 5: Performance of Arabic, Chinese and Spanish mention detection using lexical features ("Baseline" column). Columns "CIP" stands for systems that add cross-language context independent mention propagation information and column "CDP" is for systems that add cross-language context dependent mention propagation information.

		Baseline			CIP			CDP		
	N	P	R	F	P	R	F	P	R	F
Arabic:	3566	82.2	72.6	**77.1**	82.7	72.9	**77.5**	83.2	74.5	**78.6**
Chinese:	4791	80.0	71.3	**75.5**	79.9	71.5	**75.5**†	81.0	72.4	**76.5**
Spanish:	2487	79.1	71.2	**74.9**	79.9	71.9	**75.7**	80.7	74.6	**77.5**

Table 6: Performance of Arabic, Chinese and Spanish mention detection using lexical and syntactic features (POS tags, chunk information, etc).

sources available in the previous Subsection. This experiment is important because it tests the effectiveness of the propagation approach in improving performance on languages with a typical level of resources.

Results show that even in this situation, the use of cross language mention propagation information still lead to considerable improvement: using the CDP transfer method yields improvements from $1.1F$ in Chinese to $2.6F$ in Spanish. Similar to the previous section, the use of CIP information did not improve performance significantly on Arabic (77.5 vs. 77.1) and Chinese (75.5 vs. 75.5) systems, but we notice an improvement in Spanish[13].

7.4 Lexical, Syntactic and Semantic Resources

This final section investigates whether the access to cross-language mention propagation information can still improve the performance of existing competitive mention detection systems trained on languages with large resources. In this case, systems have access to a full array of lexical, syntax, semantic information, including the output from other information extraction models. Table 7 presents the performance of mention detection systems on the three languages, in the familiar 3 propagation methods: again, results show that better performance is obtained when cross language mention information is used. Under CIP, almost no change in terms of performance is obtained for Arabic and Spanish, though a slight improvement can be observed for Chinese (76.9F vs. 75.8F). When CDP is used the performance of mention detection systems is improved by 0.9F for Arabic (80.9 vs. 80.0), 2.3F for Chinese (78.1F vs. 75.8F) and 1.9F for Spanish (78.1 vs. 76.2F). Once again, the results prove that the use of cross language mention propagation information, especially through CDP, is effective in improving the performance even in this case.

By comparing results across tables, one can note that systems having access to only lexical and cross language mention propagation information are as effective as systems having access to large set of information. For Chinese, we obtain a performance of 75.8F when the system has access to lexical, syntactic and output of other information extraction models. On the other hand, the same system has a slightly better performance of 76.0 when it has access to lexical and cross language mention propagation information. The same behavior is observed for Spanish, we obtain a performance of 76.2F when the system has access to lexical, syntactic and output of other information extraction models; compared to 77.4F when lexical and cross language mention information are used. This is not true for Arabic where having access to larger set of information led to better performance when compared to systems having access to lexical information and CDP information (80.0F vs. 78.0). We attribute this difference to the fact that in Arabic we use the output of larger number of information extraction models, and consequently a richer set of information.

[13]The dagger sign † marks the systems that are not statistically significantly better.

		Baseline			CIP			CDP		
	N	P	R	F	P	R	F	P	R	F
Arabic:	3566	83.6	76.8	**80.0**	83.9	77.0	**80.2**†	84.2	77.8	**80.9**
Chinese:	4791	81.1	71.3	**75.8**	81.4	73.0	**76.9**	81.7	74.8	**78.1**
Spanish:	2487	79.1	73.5	**76.2**	79.3	73.4	**76.2**†	80.1	76.2	**78.1**

Table 7: Performance of Arabic, Chinese and Spanish mention detection using lexical, syntactic and output of other information extraction models: full-blown systems.

The other observation that is worth making is that the improvement in performance has a decreasing tendency as more resources are available. The performance gain for CDP in Arabic goes from 1.6 to 1.5 to 0.9, and the one on Spanish goes from 2.9 to 2.6 to 1.9. The one on Chinese follows part of this trend, as it goes from 1.4 to 1.1 to 2.3. While the evidence here is not definitive, one can indeed note the reduced effectiveness of the method as more resources are available, which was indeed what we expected.

Results obtained by all these experiments help answer an important question: when trying to improve mention detection systems in a resource-poor language, should we invest in building resources or should we use propagation from a resource-rich language to (at least) bootstrap the process? The answer seems to be the latter.

8 Conclusion

This paper presents a new approach to mention detection in low, medium or high-resource languages, which benefits from projecting the output from a resource-rich language such as English. We show that even when no training data is available in one source language, we can still build a decently performing baseline mention detection system by only using resources from English. This approach requires a mention detection system on a resource-rich language and an SMT system that translate text from the source to the resource-rich language, both of which can be attained.

In cases when large resources are available in the source language, our cross language mention propagation technique is still able to further improve mention detection system performance. Experiments performed on the four languages of ACE 2007, with English chosen as the *resource-rich* language, show consistent and significant improvements across conditions and levels of linguistic sophistication. The experiments are conducted on clearly specified partitions of the ACE 2007 data set, so future comparisons against the presented work can be correctly and accurately made. We also note that systems that have access to lexical and cross language mention propagation information are as accurate as those that have access to lexical, syntactic and output of other information extraction models in the source language (but no cross-language resources). As future work, we plan to extend this work to use semi-supervised and unsupervised approaches that can make use of cross-language information propagation.

We believe that it is important for the research community to continue to invest in building better resources in "source" languages, as it looks the most promising approach. However, using a propagation approach can definitely help bootstrap the process.

Acknowledgments

This work was supported by DARPA/IPTO Contract No. HR0011-06-2-0001 under the GALE program.

References

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 529–536, Sydney, Australia, July. Association for Computational Linguistics.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

C. Cabezas, B. Dorr, and P. Resnik. 2001. Spanish language processing at university of maryland: Building infrastructure for multilingual applications. In *Proceedings of the 2nd International Workshop on Spanish Language Processing and Language Technologies*.

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

I. Dagan and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Meeting of the Association for Computational Linguistics*, pages 130–137.

Mona Diab and Philip Resnik. 2001. An unsupervised method for word sense tagging using parallel corpora. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262, Morristown, NJ, USA. Association for Computational Linguistics.

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 *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 1–8.

W. Gale, K. Church, and D. Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

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

Gregory Grefenstette. 1998. *Cross-Language Information Retrieval*, volume 079238122X. Kluwer Academic Publishers.

http://www.cnts.ua.ac.be/conll2002/ner/. 2002.

Fei Huang and Kishore Papineni. 2007. Hierarchical system combination for machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 277–286.

Rebecca Hwa, Philip Resnik, and Amy Weinberg. 2002. Breaking the resource bottleneck for multilingual parsing. In *Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*.

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.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA'04*, Washington DC, September-October.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Y.-S. Lee, K. Papineni, S. Roukos, O. Emam, and H. Hassan. 2003. Language model based Arabic word segmentation. In *Proceedings of the ACL'03*, pages 399–406.

Young-Suk Lee, Yaser Al-Onaizan, Kishore Papineni, and Salim Roukos. 2006. Ibm spoken language translation system. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 13–18, Barcelona, Spain, June.

Andrew McCallum, Dayne Freitag, and Fernando Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *ICML*.

G. A. Miller. 1995. WordNet: A lexical database. *Communications of the ACM*, 38(11).

NIST. 2007. The ACE evaluation plan. www.nist.gov/speech/tests/ace/index.htm.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

P.F.Brown, S.A.Della Pietra, V.J. Della Pietra, and R.L.Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of ACL'91*.

L. Ramshaw and M. Marcus. 1994. Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging. In *Proceedings of the ACL Workshop on Combining Symbolic and Statistical Approaches to Language*, pages 128–135.

L. Ramshaw and M. Marcus. 1995. Text chunking using transformation-based learning. In David Yarowsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey. Association for Computational Linguistics.

E. Riloff, C. Schafer, and D. Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of Coling 2002*, Taipei, Taiwan.

E. F. Tjong Kim Sang and J. Veenstra. 1999. Representing text chunks. In *Proceedings of EACL'99*.

E. 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.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001*, San Diego, California, USA.

Imed Zitouni, Jeff Sorensen, Xiaoqiang Luo, and Radu Florian. 2005. The impact of morphological stemming on Arabic mention detection and coreference resolution. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 63–70, Ann Arbor, June.