Named Entity Recognition using Cross-lingual Resources: Arabic as an Example

Kareem Darwish
Qatar Computing Research Institute
Doha, Qatar
kdarwish@qf.org.qa

Abstract

Some languages lack large knowledge bases and good discriminative features for Name Entity Recognition (NER) that can generalize to previously unseen named entities. One such language is Arabic, which: a) lacks a capitalization feature; and b) has relatively small knowledge bases, such as Wikipedia. In this work we address both problems by incorporating cross-lingual features and knowledge bases from English using cross-lingual links. We show that such features have a dramatic positive effect on recall. We show the effectiveness of cross-lingual features and resources on a standard dataset as well as on two new test sets that cover both news and microblogs. On the standard dataset, we achieved a 4.1% relative improvement in F-measure over the best reported result in the literature. The features led to improvements of 17.1% and 20.5% on the new news and microblogs test sets respectively.

1 Introduction

Named Entity Recognition (NER) is essential for a variety of Natural Language Processing (NLP) applications such as information extraction. There has been a fair amount of work on NER for a variety of languages including Arabic. To train an NER system, some of the following feature types are typically used (Benajiba and Rosso, 2008; Nadeau and Sekine, 2009):

- Orthographic features: These features include capitalization, punctuation, existence of digits, etc. One of the most effective orthographic features is capitalization in English, which helps NER to generalize to new text of different genres. However, capitalization is not very useful in some languages such as German, and nonexistent in other languages such as Arabic. Further, even in English social media, capitalization may be inconsistent.

- Contextual features: Certain words are indicative of the existence of named entities. For example, the word "said" is often preceded by a named entity of type "person" or "organization". Sequence labeling algorithms (ex. Conditional Random Fields (CRF)) can often identify such indicative words.

- Character-level features: These features typically include the leading and trailing letters of words. In some languages, these letters could prefixes and suffixes. Such features can be indicative or counter-indicative of the existence of named entities. For example, a word ending with "ing" is typically not a named entity, while a word ending in "berg" is often a named entity.

- Part-of-speech (POS) tags and morphological features: POS tags indicate (or counter-indicate) the possible presence of a named entity at word level or at word sequence level. Morphological features can mostly indicate the absence of named entities. For example, Arabic allows the attachment of pronouns to nouns and verbs. However, pronouns are rarely ever attached to named entities.

- Gazetteers: This feature checks the presence of a word or a sequence of words in large lists of named entities. If gazetteers are small, then they would have low coverage, and if they are very large then their entries may be ambiguous. For example, "syntax" may refer to sentence construction or the music band "Syntax".

Typically, a subset of these features are available for different languages. For example, morphological, contextual, and character-level features have been shown to be effective for Arabic NER (Benajiba and Rosso, 2008). However, Arabic lacks indicative orthographic features that generalize to previously unseen named entities. Also, although some

of the Arabic gazetteers that were used for NER were small (Benajiba and Rosso, 2008), there has been efforts to build larger Arabic gazetteers (Attia et al., 2010). Since training and test parts of standard datasets for Arabic NER are drawn from the same genre in relatively close temporal proximity, a named entity recognizer that simply memorizes named entities in the training set generally performs well on such test sets. Thus, the results that are reported in the literature are generally high (Abdul-Hamid and Darwish, 2010; Benajiba et al., 2008). We illustrate the limited capacity of existing recognizers to generalize to previously unseen named entities using two new test sets that include microblogs as well as news texts that cover local and international politics, economics, health, sports, entertainment, and science. As we will show later, recall is well below 50% for all named entity types on the new test sets.

To address this problem, we introduce the use of cross-lingual links between a disadvantaged language, Arabic, and a language with good discriminative features and large resources, English, to improve Arabic NER. We exploit English's orthographic features, particularly capitalization, as well as Arabic and English Wikipedias, including existing annotations from large knowledge sources such as DBpedia. We also show how to use transliteration mining to improve NER, even when neither language has a capitalization (or similar) feature. The intuition is that if the translation of a word is in fact a transliteration, then the word is likely a named entity. Cross-lingual links are obtained using Wikipedia cross-language links and a large Machine Translation (MT) phrase table that is true cased, where word casing is preserved during training. We show the effectiveness of these new features on a standard dataset as well as two new test sets. The contributions of this paper are as follows:

- Using cross-lingual links to exploit orthographic features in other languages.
- Employing transliteration mining to improve NER.
- Using cross-lingual links to exploit a large knowledge base, namely English DBpedia, to benefit NER.
- Introducing two new NER test sets for Arabic that include recent news as well as microblogs. We plan to release these test sets.
- Improving over the best reported results in the literature by 4.1% (Abdul-Hamid and Darwish, 2010) by strictly adding cross-lingual features. We also show improvements of 17.1% and 20.5% on the new test sets.

The remainder of the paper is organized as follows: Section 2 provides related work; Section 3 describes the baseline system; Section 4 introduces the cross-lingual features and reports on their effectiveness; and Section 5 concludes the paper.

2 Related Work

2.1 Using cross-lingual Features

For many NLP tasks, some languages may have significantly more training data, better knowledge resources, or more discriminating features than other languages. If cross-lingual resources are available, such as parallel data, increased training data, better resources, or superior features can be used to improve the processing (ex. tagging) for other languages (Ganchev et al., 2009; Shi et al., 2010; Yarowsky and Ngai, 2001). Some work has attempted to use bilingual features in NER. Burkett et al. (2010) used bilingual text to improve monolingual models including NER models for German, which lacks a good capitalization feature. They did so by training a bilingual model and then generating more training data from unlabeled parallel data. They showed significant improvement in German NER effectiveness, particularly for recall. In our work, there is no need for tagged text that has a parallel equivalent in another language. Benajiba et al. (2008) used an Arabic English dictionary from MADA, an Arabic analyzer, to indicate if a word is capitalized in English or not. They reported that it was the second most discriminating feature that they used. However, there seems to be room for improvement because: (1) MADA's dictionary is relatively small and would have low coverage; and (2) the use of such a binary feature is problematic, because Arabic names are often common Arabic words and hence a word may be translated as a named entity and as a common word. To overcome these two problems, we use cross-lingual features to improve NER using large bilingual resources, and we incorporate confidences to avoid having a binary feature. Richman and Schone (2008) used English linguis-

tic tools and cross language links in Wikipedia to automatically annotate text in different languages. Transliteration Mining (TM) has been used to enrich MT phrase tables or to improve cross language search (Udupa et al., 2009). Conversely, people have used NER to determine if a word is to be transliterated or not (Hermjakob et al., 2008). However, we are not aware of any prior work on using TM to determine if a sequence is a NE. Further, we are not aware of prior work on using TM (or transliteration in general) as a cross lingual feature in any annotation task. In our work, we use state-of-the-art TM as described by El-Kahki et al. (2011)

2.2 Arabic NER

Much work has been done on NER with multiple public evaluation forums. Nadeau and Sekine (Nadeau and Sekine, 2009) surveyed lots of work on NER for a variety of languages. Significant work has been conducted by Benajiba and colleagues on Arabic NER (Benajiba and Rosso, 2008; Benajiba et al., 2008; Benajiba and Rosso, 2007; Benajiba et al., 2007). Benajiba et al. (2007) used a maximum entropy classifier trained on a feature set that includes the use of gazetteers and a stopword list, appearance of a NE in the training set, leading and trailing word bigrams, and the tag of the previous word. They reported 80%, 37%, and 47% F-measure for locations, organizations, and persons respectively on the ANERCORP dataset that they created and publicly released. Benajiba and Rosso (2007) improved their system by incorporating POS tags to improve NE boundary detection. They reported 87%, 46%, and 52% F-measure for locations, organizations, and persons respectively. Benajiba and Rosso (2008) used CRF sequence labeling and incorporated many language specific features, namely POS tagging, base-phrase chunking, Arabic tokenization, and adjectives indicating nationality. They reported that tokenization generally improved recall. Using POS tagging generally improved recall at the expense of precision, leading to overall improvements in F-measure. Using all their suggested features, they reported 90%, 66%, and 73% F-measure for location, organization, and persons respectively. In Benajiba et al. (2008), they examined the same feature set on the Automatic Content Extraction (ACE) datasets using CRF sequence labeling and a Support Vector Machine (SVM) classifier. They did not report per category F-measure, but they reported overall 81%, 75%, and 78% macro-average F-measure for broadcast news and newswire on the ACE 2003, 2004, and 2005 datasets respectively. Huang (2005) used an HMM-based NE recognizer for Arabic and reported 77% F-measure on the ACE 2003 dataset. Farber et al. (2008) used POS tags obtained from an Arabic tagger to enhance NER. They reported 70% F-measure on the ACE 2005 dataset. Shaalan and Raza (2007) reported on a rule-based system that uses hand crafted grammars and regular expressions in conjunction with gazetteers. They reported upwards of 93% F-measure, but they conducted their experiments on non-standard datasets, making comparison difficult. Abdul-Hamid and Darwish (2010) used a simplified feature set that relied primarily on character level features, namely leading and trailing letters in a word. They also experimented with a variety of phrase level features with little success. They reported an F-measure of 76% and 81% for the ACE2005 and the ANERCorp datasets datasets respectively. We used their simplified features in our baseline system. The different experiments reported in the literature may not have been done on the same training/test splits. Thus, the results may not be completely comparable. Mohit et al. (2012) performed NER on a different genre from news, namely Arabic Wikipedia articles, and reported recall values as low as 35.6%. They used self training and recall oriented classification to improve recall, typically at the expense of precision. McNamee and Mayfield (2002) and Mayfield et al. (2003) used thousands of language independent features such as character n-grams, capitalization, word length, and position in a sentence, along with language dependent features such as POS tags and BP chunking. The use of CRF sequence labeling for NER has shown success (McCallum and Li, 2003; Nadeau and Sekine, 2009; Benajiba and Rosso, 2008).

3 Baseline Arabic NER System

For the baseline system, we used the CRF++[1] implementation of CRF sequence labeling with default parameters. We opted to reimplement the most suc-

[1] http://code.google.com/p/crfpp/

cessful features that were reported by Benajiba et al. (2008) and Abdul-Hamid and Darwish (2010), namely the leading and trailing 1, 2, 3, and 4 letters in a word; whether a word appears in the gazetteer that was created by Benajiba et al. (2008), which is publicly available, but is rather small (less than 5,000 entries); and the stemmed form of words (after removing coordinating conjunctions, prepositions, and determiners using a rule-based stemmer akin to (Larkey et al., 2002)). As mentioned earlier, the leading and trailing letters in a word may indicate or counter-indicate the presence of named entities. Stemming is important due to the morphological complexity of Arabic. We used the previous and the next words in their raw and stemmed forms as features. For training and testing, we used the ANERCORP dataset (Benajiba and Rosso, 2007). The dataset has approximately 150k tokens and we used the 80/20 training/test splits of Abdul-Hamid and Darwish (2010), who graciously provided us with their splits of the collection and they achieved the best reported results on the dataset. We will refer to their results, which are provided in Table 1, as "baseline-lit". Table 2 (a) shows our results on the ANERCORP dataset. Our results were slightly lower than their results (Abdul-Hamid and Darwish, 2010). It is noteworthy that 69% of the named entities in the test part were seen during training.

We also created two new test sets. The first test set is composed of news snippets from the RSS feed of the Arabic (Egypt) version of news.google.com from Oct. 6, 2012. The RSS feed contains the headline and the first 50-100 words in the news articles. The set has news from over a dozen different news sources and covers international and local news, politics, financial news, health, sports, entertainment, and technology. This set contains roughly 15k tokens. The second set contains a set of 1,423 tweets that were randomly selected from tweets authored between November 23, 2011 and November 27, 2011. We scraped tweets from Twitter using the query "lang:ar" (language=Arabic). This set contains approximately 26k tokens. The test sets will be henceforth be referred to as the NEWS and TWEETS sets respectively. They were annotated by one person, a native Arabic speaker, using the Linguistics Data Consortium tagging guidelines. Table 2 (b) and (c) report on the results for the baseline system on both test sets. The results on the NEWS test are substantially lower than those for ANERCORP. It is worth noting that only 27% of the named entities in the NEWS test set were observed in the training set (compared to 69% for ANERCORP). As Table 3 shows for the ANERCORP dataset, using only the tokens as features, where the labeler mainly memorizes previously seen named entities, yields higher results than the baseline results for the NEWS dataset (Table 2 (b)). The results on the TWEETS test are very poor, with 24% of the named entities in the test set appearing in the training set.

ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	93	83	88
ORG	84	64	73
PERS	90	75	82
Overall	89	74	81

Table 1: "Baseline-lit" Results from (Abdul-Hamid and Darwish, 2010)

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	93.6	83.3	88.1
ORG	83.8	61.2	70.8
PERS	84.3	64.4	73.0
Overall	88.9	72.5	79.9
(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	84.1	53.2	65.1
ORG	73.2	23.2	35.2
PERS	74.8	47.1	57.8
Overall	78.0	41.9	54.6
(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	79.9	27.1	40.4
ORG	44.4	9.1	15.1
PERS	45.7	27.8	34.5
Overall	58.0	23.1	33.1

Table 2: Baseline Results for the three test sets

ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	95.3	62.7	75.6
ORG	86.3	44.7	58.9
PERS	85.4	36.4	51.0
Overall	91.0	50.0	64.5

Table 3: Results of using only tokens as features on ANERCORP

4 Cross-lingual Features

We experimented with three different cross-lingual features that used Arabic and English Wikipedia cross-language links and a true-cased phrase table that was generated using Moses (Koehn et al., 2007). True-casing preserves case information during training. We used the Arabic Wikipedia snapshot from September 28, 2012. The snapshot has 348,873 titles including redirects, which are alternative names to articles. Of these articles, 254,145 have cross-lingual links to English Wikipedia. We used DBpedia 3.8 which includes 6,157,591 entries of Wikipedia titles and their "types", such as "person", "plant", or "device", where a title can have multiple types. The phrase table was trained on a set of 3.69 million parallel sentences containing 123.4 million English tokens. The sentences were drawn from the UN parallel data along with a variety of parallel news data from LDC and the GALE project. The Arabic side was stemmed (by removing just prefixes) using the Stanford word segmenter (Green and DeNero, 2012).

4.1 Cross-lingual Capitalization

As we mentioned earlier, Arabic lacks capitalization and Arabic names are often common Arabic words. For example, the Arabic name "Hasan" means good. To capture cross-lingual capitalization, we used the aforementioned true-cased phrase table at word and phrase levels as follows:

Input: True-cased phrase table PT, sentence S containing n words $w_{0..n}$, max sequence length l, translations $T_{1..k..m}$ of $w_{i..j}$
for $i = 0 \rightarrow n$ **do**
  $j = min(i + l - 1, n)$
  if PT contains $w_{i..j}$ & $\exists\, T_k\ isCaps$ **then**

$$weight(w_{i..j}) = \frac{\sum_{T_k\text{-}isCaps} P(T_k)}{\sum_{T_k\text{-}isCaps} P(T_k) + \sum_{T_k\text{-}notCaps} P(T_k)}$$

    round $weight(w_{i..j})$ to first significant figure
    set tag of $w_i = B\text{-}CAPS\text{-}weight$
    set tag for words $w_{i+1..j} = I\text{-}CAPS\text{-}weight$
  else
    if $j > i$ **then**
      j- -
    else
      tag of $w_i = null$
    end if
  end if
end for

Where: PT was the aforementioned phrase table; $l = 4$; $P(T_k)$ equaled to the product of $p(source|target)$ and $p(target|source)$ for a word sequence; isCaps and notCaps were whether the translation was capitalized or not respectively; and the weights were binned because CRF++ only takes nominal features. In essence we tried every subsequence of S of length l or less to see if the translation was capitalized. A subsequence can be 1 word long. We tried longer sequences first. To determine if the corresponding phrase was capitalized ($isCaps$), all non-function words on the English side needed to be capitalized. As an example, the phrase المحيط الهادي (meaning "Pacific Ocean") was translated to a capitalized phrase 36.7% of the time. Thus, the word المحيط was assigned B-CAPS-0.4 and الهادي was assigned I-CAPS-0.4. Using weights avoids using capitalization as a binary feature.

Table 4 reports on the results of the baseline system with the capitalization feature on the three datasets. In comparing baseline results in Table 2 and cross-lingual capitalization results in Table 4, recall consistently increased for all datasets, particularly for "persons" and "locations". For the different test sets, recall increased by 3.1 to 6.1 points (absolute) or by 8.4% to 13.6% (relative). This led to an overall improvement in F-measure of 1.8 to 3.4 points (absolute) or 4.2% to 5.7% (relative). Precision dropped overall on the ANERCORP dataset and dropped substantially for the NEWS and TWEETS test sets.

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.0/-1.6/-1.7	86.8/3.5/4.2	89.3/1.2/1.4
ORG	82.8/-1.1/-1.3	63.9/2.7/4.4	72.1/1.4/1.9
PERS	86.0/1.7/2.0	75.4/11.0/17.1	80.3/7.3/10.1
Overall	88.4/-0.4/-0.5	78.6/6.1/8.4	83.2/3.4/4.2
(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	82.1/-2.0/-2.4	59.0/5.8/11.0	68.7/3.5/5.4
ORG	68.4/-4.9/-6.6	23.2/0.0/0.0	34.6/-0.6/-1.7
PERS	70.7/-4.0/-5.4	55.6/8.4/17.9	62.2/4.4/7.6
Overall	74.5/-3.5/-4.5	47.0/5.1/12.2	57.7/3.1/5.7
(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	76.9/-3.0/-3.7	27.9/0.9/3.2	41.0/0.5/1.4
ORG	44.4/0.0/0.0	10.4/1.3/14.3	16.8/1.8/11.6
PERS	40.0/-5.7/-12.5	35.0/7.3/26.2	37.3/2.8/8.1
Overall	51.8/-6.2/-10.7	26.3/3.1/13.6	34.9/1.8/5.4

Table 4: Results with cross-lingual capitalization with /absolute/relative differences compared to baseline

4.2 Transliteration Mining

An alternative to capitalization can be transliteration mining. The intuition is that named entities are often transliterated, particularly the names of locations and persons. This feature is helpful if cross-lingual resources do not have capitalization information, or if the "helper" language to be consulted does not have a useful capitalization feature. We performed transliteration mining (aka cognate matching) at word level for each Arabic word against all its possible translations in the phrase table. We used a transliteration miner akin to that of El-Kahki et al. (2011) that was trained using 3,452 parallel Arabic-English transliteration pairs. We aligned the word-pairs at character level using GIZA++ and the phrase extractor and scorer from the Moses machine translation package (Koehn et al., 2007). The alignment produced mappings between English letters sequences and Arabic letter sequences with associated mapping probabilities. Given an Arabic word, we produced all its possible segmentations along with their associated mappings into English letters. We retained valid target sequences that produced translations in the phrase table.

Again we used a weight similar to the one for cross-lingual capitalization and we rounded the values of the ratio the significant figure. The weights were computed as:

$$\frac{\sum_{T_k\_isTransliteration} P(T_k)}{\sum_{T_k\_isTransliteration} P(T_k) + \sum_{T_k\_notTransliteration} P(T_k)} \quad (1)$$

where $P(T_k)$ is probability of the k^{th} translation of a word in the phrase table.

If a word was not found in the phrase table, the feature value was assigned null. For example, if the translations of the word حسن are "Hasan", "Hassan", and "good", where the first two are transliterations and the last not, then the weight of the word would be:

$$\frac{P(Hasan|\text{حسن}) + P(Hassan|\text{حسن})}{P(Hasan|\text{حسن}) + P(Hassan|\text{حسن}) + P(good|\text{حسن})} \quad (2)$$

In our experiments, the weight of حسن was equal to 0.5 (after rounding). Table 5 reports on the results using the baseline system with the transliteration mining feature. Like the capitalization feature, transliteration mining slightly lowered precision – except for the TWEETS test set where the drop in precision was significant – and positively increased recall, leading to an overall improvement in F-measure for all test sets. Overall, F-measure improved by 1.9%, 3.7%, and 3.9% (relative) compared to the baseline for the ANERCORP, NEWS, and TWEETS test sets respectively. The similarity of results between using transliteration mining and word-level cross-lingual capitalization suggests that perhaps they can serve as surrogates.

4.3 Using DBpedia

DBpedia[2] is a large collaboratively-built knowledge base in which structured information is extracted from Wikipedia (Bizer et al., 2009). DBpedia 3.8, the release we used in this paper, contains 6,157,591 Wikipedia titles belonging to 296 types. Types vary in granularity with each Wikipedia title having one or more type. For example, NASA is assigned the following types: Agent, Organization, and GovernmentAgency. In all, DBpedia includes the names of 764k persons, 573k locations, and 192k organizations. Of the Arabic Wikipedia titles, 254,145 have Wikipedia cross-lingual links to English Wikipedia, and of those English Wikipedia titles, 185,531 have entries in DBpedia. Since Wikipedia titles may have multiple DBpedia types, we opted to keep the most popular type (by count of how many Wikipedia titles are assigned a particular type) for each title, and we disregarded the rest. We also chose not to use the "Agent" and "Work" types because they were highly ambiguous. We found word sequences in the manner described in the pseudocode for cross-lingual capitalization. For translation, we generated two features using two translation resources, namely the aforementioned phrase table and Arabic-English Wikipedia cross-lingual links. When using the phrase table, we used the most likely translation into English that matches an entry in DBpedia provided that the product of $p(source|target)$ and $p(target|source)$ of translation was above 10^{-5}. We chose the threshold qualitatively using offline experiments. When using Arabic-English Wikipedia cross-lingual links, if an entry was found in the Arabic Wikipedia, we performed a look up in DB-

[2] http://dbpedia.org

pedia using the English Wikipedia title that corresponds to the Arabic Wikipedia title. We used Arabic Wikipedia page-redirects to improve coverage. For both features (using the two translation methods), for an Arabic word sequence corresponding to the DBpedia entry, the first word in the sequence was assigned the feature "B-" plus the DBpedia type and subsequent words were assigned the feature "I-" plus the DBpedia type. For example, for حزب الله (meaning "Hezbollah"), the words حزب and الله were assigned "B-Organization" and "I-Organization" respectively. For all other words, the feature was assigned "null". Using the phrase table for translation likely yielded improved coverage over using Wikipedia cross-lingual links. However, Wikipedia cross-lingual links likely led to higher quality translations, because they were manually curated. Table 6 reports on the results of using the baseline system with the two DBpedia features. Using DBpedia consistently improved precision and recall for named entity types on all test sets, except for a small drop in precision for locations on the ANERCORP dataset and for locations and persons on the TWEETS test set. For the different test sets, improvements in recall ranged between 4.4 and 7.5 points (absolute) or 6.5% and 19.1% (relative). Precision improved by 0.9 and 5.5 points (absolute) or 1.0% and 7.1% (relative) for the ANERCORP and NEWS test sets respectively. Overall improvement in F-measure ranged between 3.2 and 7.6 points (absolute) or 4.1% and 13.9% (relative).

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.9/-0.7/-0.7	83.5/0.2/0.3	88.0/-0.2/-0.2
ORG	82.9/-0.9/-1.0	61.8/0.6/1.0	70.9/0.1/0.1
PERS	84.5/0.3/0.3	71.9/7.5/11.7	77.7/4.7/6.5
Overall	88.3/-0.5/-0.6	75.5/2.9/4.1	81.4/1.5/1.9
(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	84.9/0.7/0.9	53.6/0.5/0.9	65.7/0.6/0.9
ORG	67.2/-6.1/-8.3	22.9/-0.3/-1.1	34.2/-1.0/-2.9
PERS	72.8/-1.9/-2.6	55.0/7.8/16.7	62.7/4.8/8.4
Overall	75.9/-2.1/-2.6	45.0/3.1/7.4	56.6/2.0/3.7
(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	79.1/-0.8/-1.0	27.1/0.0/0.0	40.3/-0.1/-0.3
ORG	41.8/-2.7/-6.0	9.1/0.0/0.0	14.9/-0.2/-1.1
PERS	40.0/-5.7/-12.5	35.5/7.7/27.8	37.6/3.1/8.8
Overall	51.7/-6.3/-10.9	25.8/2.6/11.3	34.4/1.3/3.9

Table 5: Results with transliteration mining with /absolute/relative differences compared to baseline

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.7/-0.9/-0.9	87.1/3.9/4.6	89.9/1.7/1.9
ORG	84.6/0.8/0.9	66.6/5.3/8.7	74.5/3.7/5.3
PERS	87.8/3.6/4.2	69.9/5.5/8.6	77.8/4.8/6.6
Overall	89.8/0.9/1.0	77.2/4.7/6.5	83.0/3.2/4.0
(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	87.8/3.6/4.3	61.8/8.6/16.2	72.5/7.4/11.3
ORG	76.1/2.9/3.9	30.2/7.0/30.1	43.2/8.0/22.7
PERS	83.2/8.5/11.3	54.2/7.1/15.0	65.7/7.8/13.6
Overall	83.5/5.5/7.1	49.5/7.5/18.0	62.2/7.6/13.9
(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	77.4/-2.5/-3.1	30.5/3.5/12.9	43.8/3.4/8.4
ORG	57.0/12.5/28.2	15.9/6.8/75.1	24.8/9.8/64.9
PERS	40.8/-4.9/-10.6	31.7/4.0/14.3	35.7/1.2/3.4
Overall	55.3/-2.6/-4.5	27.5/4.4/19.1	36.8/3.7/11.2

Table 6: Results using DBpedia with /absolute/relative differences compared to baseline

4.4 Putting it All Together

Table 7 reports on the results of using all aforementioned cross-lingual features together. Figures 1, 2, and 3 compare the results of the different setups. As the results show, the impact of cross-lingual features on recall were much more pronounced on the NEWS and TWEETS test sets – compared to the ANERCORP dataset. Further, the recall values for the ANERCORP dataset in the baseline experiments were much higher than those for the two other test sets. This confirms our suspicion that the reported values in the literature on the standard datasets are unrealistically high due to the similarity between the training and test sets. Hence, these high effectiveness results may not generalize to other test sets. Of all the cross-lingual features that we experimented with, the use of DBpedia led to improvements in both precision and recall (except for precision on the TWEETS test set). Other cross-lingual features yielded overall improvements in F-measure, mostly due to gains in recall, typically at the expense of precision. Overall, F-measure improved by 5.5%, 17.1%, and 20.5% (relative) compared to the baseline for the ANERCORP, NEWS, and TWEETS test sets respectively. For the ANERCORP test set, our results improved over the baseline-lit results (Abdul-Hamid and Darwish, 2010) by 4.1% (relative).

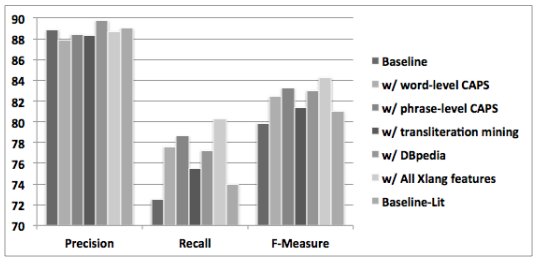

Figure 1: ANERCORP Dataset Results

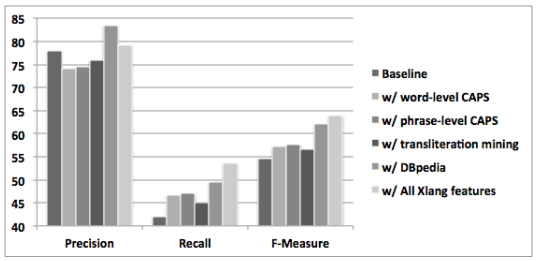

Figure 2: NEWS Test Set Results

When using all the features together, one notable result is that precision dropped significantly for the TWEETS test sets. We examined the output for the TWEETS test set and here are some of the factors that affected precision:

- the presence of words that would typically be named entities in news but would generally be regular words in tweets. For example, the Arabic word "Mubarak" is most likely the name of the former Egyptian president in the context of news, but it would most likely mean "blessed", which is common in expressions of congratulations, in tweets.

- the use of dialectic words that may have transliterations or a named entity as the most likely translation into English. For example, the word شي is typically the dialectic version of the Arabic word شيء, meaning something. However, since the MT system that we used was trained on modern standard Arabic, the dialectic word would not appear in training and would typically be translated/transliterated to the name "Che" (as in Che Guevara).

- Since tweets are restricted in length, authors frequently use shortened versions of named entities. For example, tweets would mostly have "Morsi" instead of "Mohamed Morsi" and without trigger words such as "Dr." or "president". The full version of a name and trigger words are more common in news. This same problem was present in the NEWS test set, because it was constructed from an RSS feed, and headlines, which are typically compact, had a higher representation in the test collection. We observed the same phenomenon for organization names. For example, "the Real" refers to "Real Madrid". Nicknames are also prevalent. For example, "the Land of the two Sanctuaries" refers to "Saudi Arabia".

We believe that this problem can be overcome by introducing new training data that include tweets (or other social text) and performing domain adaptation. New training data would help: identify words and expressions that are common in conversations, account for common dialectic words, and learn a better word transition model. Further, gazetteers that cover shortened versions of names could be helpful as well.

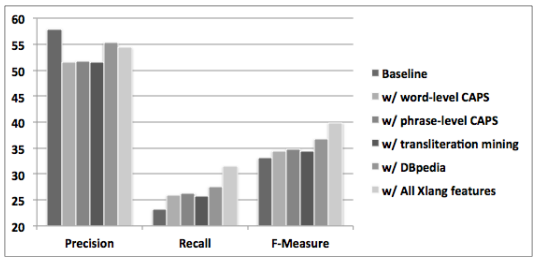

Figure 3: TWEETS Test Set Results

(a) ANERCORP Dataset			
	Precision	Recall	$F_{\beta=1}$
LOC	92.3/-1.3/-1.4	87.8/4.6/5.5	90.0/1.9/2.1
ORG	81.4/-2.4/-2.9	66.0/4.7/7.7	72.9/2.1/3.0
PERS	87.0/2.8/3.3	77.7/13.3/20.7	82.1/9.1/12.5
Overall	88.7/-0.2/-0.2	80.3/7.8/10.7	84.3/4.4/5.5
(b) NEWS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	85.1/1.0/1.2	64.1/11.0/20.6	73.1/8.0/12.3
ORG	73.8/0.5/0.7	29.4/6.2/26.9	42.1/6.8/19.4
PERS	76.8/2.0/2.7	63.4/16.3/34.5	69.5/11.7/20.2
Overall	79.2/1.2/1.6	53.6/11.6/27.7	63.9/9.4/17.1
(c) TWEETS Test Set			
	Precision	Recall	$F_{\beta=1}$
LOC	81.4/1.5/1.8	33.5/6.5/23.9	47.5/7.1/17.4
ORG	52.1/7.6/17.2	16.2/7.1/78.6	24.7/9.6/64.1
PERS	40.5/-5.2/-11.4	39.2/11.5/41.3	39.8/5.3/15.4
Overall	54.4/-3.6/-6.2	31.4/8.3/35.9	39.9/6.8/20.5

Table 7: Results using all the cross-lingual features with /absolute/relative differences compared to baseline

5 Conclusion

In this paper, we presented different cross-lingual features that can make use of linguistic properties and knowledge bases of other languages for NER. For translation, we used an MT phrase table and Wikipedia cross-lingual links. We used English as the "helper" language and we exploited the English capitalization feature and an English knowledge base, DBpedia. If the helper language did not have capitalization, then transliteration mining could provide some of the benefit of capitalization. Transliteration mining requires limited amounts of training examples. We believe that the proposed cross-lingual features can be used to help NER for other languages, particularly languages that lack good features that generalize well. For Arabic NER, the new features yielded an improvement of 5.5% over a strong baseline system on a standard dataset, with 10.7% gain in recall and negligible change in precision. We tested on a new news test set, NEWS, which has recent news articles (the same genre as the standard dataset), and indeed NER effectiveness was much lower. For the new NEWS test set, cross-lingual features led to a small increase in precision (1.6%) and a very large improvement in recall (27.7%). This led to a 17.1% improvement in overall F-measure. We also tested NER on the TWEETS test set, where we observed substantial improvements in recall (35.9%). However, precision dropped by 6.2% for the reasons we mentioned earlier. For future work, it would be interesting to apply cross-lingual features to other language pairs and to make use of joint cross-lingual models. Further, we also plan to investigate Arabic NER on social media, particularly microblogs.

References

A. Abdul-Hamid and K. Darwish. 2010. Simplified Feature Set for Arabic Named Entity Recognition. Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 110115.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini, and Josef van Genabith. 2010. An automatically built named entity lexicon for Arabic. In: LREC 2010 - 7th conference on International Language Resources and Evaluation, 17-23 May 2010, Valletta, Malta.

Y. Benajiba, M. Diab, and P. Rosso. 2008. Arabic Named Entity Recognition using Optimized Feature Sets. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 284293, Honolulu, October 2008.

Y. Benajiba and P. Rosso. 2008. Arabic Named Entity Recognition using Conditional Random Fields. In Proc. of Workshop on HLT & NLP within the Arabic World, LREC08.

Y. Benajiba, P. Rosso and J. M. Benedi. 2007. ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. In Proc. of CICLing-2007, Springer-Verlag, LNCS(4394), pp.143-153

Y. Benajiba and P. Rosso. 2007. ANERsys 2.0: Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information. In Proc. of Workshop on Natural Language-Independent Engineering, IICAI-2007.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. 2009. DBpedia A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7, Pages 154165, 2009.

D. Burkett, S. Petrov, J. Blitzer, D. Klein. 2010. Learning Better Monolingual Models with Unannotated Bilingual Text. Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 46–54.

A. El Kahki, K. Darwish, A. Saad El Din, M. Abd El-Wahab and A. Hefny. 2011. Improved Transliteration Mining Using Graph Reinforcement. EMNLP-2011.

B. Farber, D. Freitag, N. Habash, and O. Rambow. 2008. Improving NER in Arabic Using a Morphological Tagger. In Proc. of LREC08.

K. Ganchev, J. Gillenwater, and B. Taskar. 2009. Dependency grammar induction via bitext projection constraints. In ACL-2009.

Spence Green and John DeNero. 2012. A Class-Based Agreement Model for Generating Accurately Inflected Translations. In ACL-2012.

Ulf Hermjakob, Kevin Knight, and Hal Daum III. 2008. Name translation in statistical machine translation: Learning when to transliterate. ACL-08: HLT, Pages 389-397.

F. Huang. 2005. Multilingual Named Entity Extraction and Translation from Text and Speech. Ph.D. Thesis. Pittsburgh: Carnegie Mellon University.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit

for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML, pp.282-289, 2001.

Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell. 2002. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. SIGIR-2002.

J. Mayfield, P. McNamee, and C. Piatko. 2003.Named Entity Recognition using Hundreds of Thousands of Features. HLT-NAACL 2003-Volume 4, 2003.

A. McCallum and W. Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons. In Proc. Conference on Computational Natural Language Learning.

P. McNamee and J. Mayfield. 2002. Entity extraction without language-specific. Proceedings of CoNLL, .2002

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), pp. 162-173. 2012.

D. Nadeau and S. Sekine. 2009. A Survey of Named Entity Recognition and Classification. Named Entities: Recognition, Classification and Use, ed. S. Sekine and E. Ranchhod, John Benjamins Publishing Company.

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2008.

K. Shaalan and H. Raza. 2007. Person Name Entity Recognition for Arabic. Proceedings of the 5th Workshop on Important Unresolved Matters, pages 1724, Prague, Czech Republic, June 2007.

L. Shi, R. Mihalcea, M. Tian. 2010. Cross Language Text Classification by Model Translation and Semi-supervised Learning. Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2010.

Raghavendra Udupa, Anton Bakalov, and Abhijit Bhole. 2009. They Are Out There, If You Know Where to Look: Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval. Advances in Information Retrieval. Pages: 437-448.

D. Yarowsky and G. Ngai. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In NAACL-2001.