Named Entity Recognition for Arabic Social Media

Ayah Zirikly
Department of Computer Science
The George Washington University
Washington DC, USA
ayaz@gwu.edu

Mona Diab
Department of Computer Science
The George Washington University
Washington DC, USA
mtdiab@gwu.edu

Abstract

The majority of research on Arabic Named Entity Recognition (NER) addresses the the task for newswire genre, where the language used is Modern Standard Arabic (MSA), however, the need to study this task in social media is becoming more vital. Social media is characterized by the use of both MSA and Dialectal Arabic (DA), with often code switching between the two language varieties. Despite some common characteristics between MSA and DA, there are significant differences between which result in poor performance when MSA targeting systems are applied for NER in DA. Additionally, most NER systems rely primarily on gazetteers, which can be more challenging in a social media processing context due to an inherent low coverage. In this paper, we present a gazetteers-free NER system for Dialectal data that yields an F1 score of 72.68% which is an absolute improvement of $\approx 2-3\%$ over a comparable state-of-the-art gazetteer based DA-NER system.

1 Introduction

Named Entity Recognition (NER) is the task of tagging names with a predefined set of named entity types (e.g. Location, Person) in open-domain text (Nadeau and Sekine, 2007). NER has been shown to improve Information Retrieval performance (Thompson and Dozier, 1997) and Question Answering (QA) performance where (Ferrndez et al., 2007) shows that, on average, questions contain $\approx 85\%$ Named Entities.

In the current world of ubiquitous social media presence, processing informal genre is becoming ever more crucial. One of the prevalent genre in social media in need for text mining is microblog data such as Twitter. Twitter data is characterized by being massive. Off the shelf NER systems trained on formal genre such as newswire fail to process such data, thereby current research in information extraction has been specifically targeting this genre (Ritter et al., 2011). This problem is quite significant in English and is ever more pronounced in lower resourced languages such as Arabic.

Arabic has gained more attention recently due to the increased availability of annotated datasets. Arabic NER systems, as other languages, are domain dependent and mainly trained on news corpora or other well structured data that uses the Modern Standard Arabic (MSA) variety of the language (Benajiba et al., 2007) and (Abdallah et al., 2012). Arabic, in general, poses additional challenges to Natural Language Processing (NLP) tasks, as opposed to other languages, due its rich morphology and highly inflected nature. Moreover, Arabic is also one of those languages that exists in a state of diglossia where multiple forms of the language exist in the same context, the standard formal form, MSA, used in formal settings (education, formal speeches, etc.) and the spoken vernaculars that differ significantly from MSA, known as Dialectal Arabic (DA) that are used pervasively in informal settings such as in social media. Since MSA and DA co-exist, we note that people very often code switch between the two varieties within the same utterance which is reflected in microblog data. Hence NLP systems targeting the Twitter genre needs to account for this phenomenon.

Compared to English NER, here are some example challenges posed for Arabic NER (Abdul-Hamid and Darwish, 2010):

- Lack of capitalization: Capitalization in Latin languages is a strong indicator of Named Entity (NE). However, in Arabic,

Proceedings of NAACL-HLT 2015, pages 176–185,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

proper nouns are not capitalized, which renders the identification of NEs more complicated;

- Nominal Confusability: Some words can be proper nouns, nouns, or adjectives. For instance, *jamiyolap*[1] which means 'beautiful' can be a proper noun or an adjective. Another example, *jamAl*, which means 'beauty', is a noun but can be a common noun or a proper name;
- Agglutination: Since Arabic exhibits concatenative morphology, we note the pervasive presence of affixes agglutinating to proper nouns as prefixes and suffixes (Shaalan, 2014). For instance, determiners appear as prefixes *Al* as in (*AlqAhrp*, 'Cairo'), likewise with affixival prepositions such as *l*, 'for' (*ldm$q* 'for/to/from Damascus'), as well as prefixed conjunctions such as *w*, 'and', as in (*wAlqds* 'and Jerusalem');
- Absence of Short Vowels (Diacritic Markers): Written MSA, even in newswire, is underspecified for short vowels, aka undiacritized, which results in higher ambiguity that can only be resolved using contextual information (Benajiba et al., 2009). Examples of ambiguity are: *mSr*, may be *miSor* as in 'Egypt' or *muSir* as in 'insistent'; *qTr* may be the name of the country 'Qatar' if vowelized/diacritized as *qaTar*, *qaTor* for 'sugar syrup', *quTor* for 'diameter'.

In addition to the afore mentioned challenges, in general, for Arabic NER in general compared to Latin-based languages, DA NER faces additional issues:

- Lack of annotated data for supervised DA NER;
- Lack of standard orthographies or language academics (Habash et al., 2013): Unlike MSA, the same word in DA can be rewritten in so many forms, e.g. *mAtEyT$, mtEyt$, mA tEyT$*, 'do not cry', are all acceptable variants since there is no one standard;
- Lack of comprehensive Gazetteers: this is a problem facing all NER systems for all languages addressing NER in social media text, since by definition such media has a ubiquitous presence of highly productive names exemplified by the usage of nick names, hence the PERSON class in social media NER will always have a coverage problem;
- Applying NLP tools designed for MSA to DA results in considerably lower performance, thus the need to build resources and tools that specifically target DA (Habash et al., 2012).

The majority of existing NER systems rely on the use of gazetteers to improve the system accuracy (Kazama and Torisawa, 2007), however, large external resources are correlated with higher performance cost. In this paper, we study the impact of word representation and embedding features on Arabic NER performance for Twitter and Dialectal Arabic, and demonstrate that our proposed features show comparable and superior results to other NER systems that use large gazetteers. Our contributions are as follows:

- Show the impact of using word representations and embedding on NER performance;
- Propose a set of features that does not include the use of external resources;
- Produce comparable NER performance to other systems that use large gazetteers.

2 Related Work

Significant amount of work in the area of NER has taken place. In (Nadeau and Sekine, 2007), the authors survey the literature of NER and report on the different sets of used features such as contextual and morphological features. Although more research has been employed in the area of English NER, Arabic NER has been gaining more attention recently. Similar to other languages, several approaches have been used for Arabic NER: Rule-based methods, Statistical Learning methods, and a hybrid of both.
In (Shaalan and Raza, 2009), the authors present rule-based NER system for MSA that comprises gazetteers, local grammars in the form of regular expressions, and a filtering mechanism that mainly focuses on rejecting incorrect NEs based on a blacklist. Their system yields a performance of 87.7% F1 measure for the Person label (PER), 85.9% for Location (LOC), and 83.15% for Organization (ORG) when evaluated against corpora

[1] We use the Buckwalter encoding system to render Arabic. For a reference listing please see http://www.qamus.org/transliteration.htm

developed by the authors. (Elsebai et al., 2009) proposed a rule-based system targeting PER in MSA. It uses the Buckwalter Arabic Morphological Analyser (BAMA) and a set of keywords. The proposed system yields an F-score of 89% when tested on an in-house annotated dataset of 700 news articles extracted from Aljazeera television website.

Although rule based approaches proved successful to some extent, most recent NER research focuses on Statistical Learning techniques due to the shortcomings of rule based approaches in terms of coverage and robustness (Nadeau and Sekine, 2007). For example, (Benajiba et al., 2007) proposes an MSA NER system (ANERsys) based on n-grams and maximum entropy. The authors also introduce ANERCorp corpora and ANERGazet gazetteers. (Benajiba and Rosso, 2008) further modify ANERsys in terms of the underlying machinery to use Conditional Random Fields (CRF) sequence labeling as the statistical learning framework. ANERsys uses the following features: part of speech (POS) tags, Base Phrase Chunks (BPC), gazetteers, and nationality information. The latter feature is included based on the observation that PER occur after mentioning the nationality, in particular in newswire data. In (Benajiba et al., 2008), a different classifier is built for each NE type. The authors study the effect of features on each NE type, then the overall NER system is a combination of the different classifiers that target each NE class label independently. The set of features used is a combination of general features as listed in (Benajiba and Rosso, 2008) and Arabic-dependent (morphological) features. Their system's best performance is 83.5% for ACE 2003, 76.7% for ACE 2004, and 81.31% for ACE 2005, respectively. (Benajiba et al., 2010) presents an Arabic NER system that incorporates lexical, syntactic, and morphological features and augmenting the model with syntactic features derived from noisy data as projected from Arabic-English parallel corpora. The system F-score performance is 81.73%, 75.67%, 58.11% on ACE2005 Broadcast News, Newswire, and Web blogs, respectively. The authors in (Abdul-Hamid and Darwish, 2010) suggest a number of features, some of which we incorporate in our current DA-NER system, namely, the head and trailing 2-grams (L2), 3-grams (L3), and 4-grams (L4) characters in a word. (Abdul-Hamid and Darwish, 2010) produce near state-of-the-art results with the use of generic and language independent features that we use to generate baseline results (BL).
(Shaalan and Oudah, 2014) presents a hybrid approach that targets MSA and produces state-of-the-art results. However, we could not get access to the exact rules employed, we were not able to replicate their results. The rule-based component is identical to their previous proposed rule-based system in (Shaalan and Raza, 2009). The features used in their study are a combination of the rule-based rules in addition to morphological, capitalization, POS tag, word length, and dot (i.e. if a word has an adjacent dot) features. All the previous work mentioned above focused on MSA, albeit with variations in genres to the extent exemplified by the ACE data and author generated data. However, unlike the work mentioned above, (Darwish and Gao, 2014) proposed an NER system that specifically targets microblogs as a genre, as opposed to newswire data. Their proposed language-independent system relies on a set of features that is similar to (Abdul-Hamid and Darwish, 2010), with the use of a simple yet effective domain adaptation approach (Daumé et al., 2010) based on a two-pass semi supervised method. Their NER system on Twitter data yields an overall F-score=65.2% (76.7% for LOC, 55.6% for ORG, and 55.8% for PER).
In our prior work, (Zirikly and Diab, 2014), we proposed a small set of annotated DA data and DA-NER system that yields an F-score=70.3%. We used n-gram, gazetteers and an extensive set of morphological features. In our current work, we explore the impact of using word embedding features and how can word representations and embedding replace the use of dictionaries and even generate better performance.

3 Approach

In this paper, we adopt a supervised machine learning approach. Supervised approaches have been shown to outperform unsupervised approaches for the NER task (Nadeau et al., 2006). We use Conditional Random Fields (CRF) sequence labeling as described in (Lafferty et al., 2001). Guided by previous work, for example (Benajiba and Rosso, 2008) demonstrates that CRF yields better results over other supervised machine learning techniques for the task of NER. One of the goals of our empirical investigation is to show the impact of using skip-gram word embedding and word representations on the NER performance, with the potential of these features substituting the use of extensive large gazetteers.

3.1 Baseline

For our baseline system (BL), we reimplemented the features proposed in (Abdul-Hamid and Darwish, 2010) that produced the best results: previous and next word, and leading and trailing +/- (2-4) character ngrams. These features are chosen as a preferred set for the baseline since they are directly applicable to DA, as none of the features rely on the availability of morphological or syntactic analyzers. This baseline is also adapted from (Darwish and Gao, 2014). We opted for this baseline as opposed to (Darwish and Gao, 2014)'s NER system since they use Wikipedia gazetteers (WikiGaz) for their NER system.[2] However, we report the results of applying their features using exact match against the gazetteers' entries in Section 4.3.

3.2 Our NER Features

In addition to the features employed in the baseline BL, we introduce the following additional features in our NER system:

- **Lexical Features**: character n-gram features, the leading and trailing character bigrams (L2), trigrams (L3), and quadrigrams (L4);
- **Contextual Features** (CTX): The surrounding undiacritized words of a context window $= \pm 1$;(W-1,W0,W1);
- **Gazetteers** (GAZ): Although our work mainly targets NER systems without the use of external sources, but we added GAZ features for comparison purposes. The gazetteer used is the union of: i) ANERGaz: proposed by (Benajiba and Rosso, 2008), which contains 2183 LOC, 402 ORG, and 2308 PER; and ii) WikiGaz: large Wikipedia gazetteer (Darwish and Gao, 2014), which contains 50141 LOC, 17092 ORG, and 65557 PER. We followed this strategy for text matching against gazetteer entries:
  - Exact match (EM-GAZ): For more efficient search, we use Aho-Corasick Algorithm (Aho and Corasick, 1975) that has linear running time in terms of the input length plus the number of matching entries in a gazetteer. When a word sequence matches an entry in the gazetteer;
  - Partial match(PM-GAZ): This feature is created to handle the case of compound gazetteer entries. If the token is part of the compound name then this feature is set to true. For example, if the gazetteer entry is *yAsr ErfAt* 'Yasser Arafat' and the input text is *yAsr BrkAt* then PM-GAZ for the token *yAsr* will be set to true. This is particularly useful in persons names;
  - Levenshtein match (LVM-GAZ): We use Levenshtein distance (Levenshtein, 1966) to compare the similarity between the input and a gazetteer entry. This is based on the observation that social media data might contain non-standard spelling of words since it contains the DA variety.
- **Morphological Features**: The morphological features that we employ in our feature set are generated by MADAMIRA (Pasha et al., 2014). The set comprises the following:
  - Part of Speech (POS) tags: We use POS tags generated from MADAMIRA within a window of ±1 (POS-1, POS0, POS1);
  - Capitalization (CAPS): In order to circumvent the lack of capitalization in Arabic, we check the capitalization of the translated NE which could indicate that a word is an NE (Benajiba et al., 2008). This feature is dependent on the English gloss generated by MADAMIRA. This feature is set to true when the gloss starts with a capital letter;
  - Aspect (ASP), person (PERS), proclitics0 (PROC0), proclitics1 (PROC1), proclitics2 (PROC2), proclitics3 (PROC3), enclitics0 (ENC0); detailed description for these features is provided in Table 1;

[2] Though the authors kindly gave us access to the actual gazetteer, we were unable to replicate their results since the gazetteer matching method is not detailed in their paper.

Feature	**Feature Values**
Aspect	Verb aspect: Command, Imperfective, Perfective, Not applicable
Person	Person Information: 1st, 2nd, 3rd, Not applicable
Proclitic3	Question proclitic: No proclitic, Not applicable, Interrogative particle
Proclitic2	Conjunction proclitic: No proclitic, Not applicable, Conjunction *f*, Connective particle *f*, Response conditional *f*, Subordinating conjunction *f*, Conjunction *w*, Particle *w*, Subordinating conjunction *w*
Proclitic1	Preposition proclitic: No proclitic, Not Applicable, Interrogative *i$*, Particle *b*, Preposition *b*, Progressive verb particle *b*, Preposition *E*, Preposition *ElY*, Preposition *fy*, Demonstrative *hA*, Future marker *H*, Preposition *k*, Emphatic particle *l*, Preposition *l*, Preposition *l* + preposition *b*, Emphatic *l* + future marker *H*, Response conditional *l* + future marker *H*, Jussive *l*, Preposition *l*, Preposition *mn*, Future marker *s*, Preposition *t*, Particle *w*, Preposition *w*, Vocative *w*, vocative *yA*
Proclitic	Article proclitic: No proclitic, Not Applicable, Demonstrative particle *A*, Determiner, Determiner *Al* + negative particle *mA*, Negative particle *lA*, Negative particle *mA*, Negative particle *mA*, Particle *mA*, relative pronoun *mA*
Enclitics	Pronominals: No enclitic, Not applicable, 1st person plural/singular, 2nd person dual/plural, 2nd person feminine plural/singular, 2nd person masculine plural/singular, 3rd person dual/plural, 3rd person feminine plural/singular, 3rd person masculine plural/singular, Vocative particle, Negative particle *lA*, Interrogative pronoun *mA*, Interrogative pronoun *mA*, Interrogative pronoun *mn*, Relative pronoun *mn, m, mA*, Subordinating conjunction *m, mA*.

Table 1: Morphological Features

  - isNum: Binary feature that is set to true if the token is a number;
  - isNoun: Binary feature that is set to true if the token is proper noun (i.e. POS=noun_prop).

- **Brown Clustering IDs** (BC): Brown clustering (Brown et al., 1992) is a hierarchical clustering approach that maximizes the mutual information of word bigrams. Word representations, especially Brown Clustering, have been shown to improve the performance of NER system when added as a feature (Turian et al., 2010). In this work, we use Brown Clustering IDs (BC) of variable prefix lengths (4,5,6,7,10,13 and the full length of the cluster ID) as features resulting in the following set of features BC4, BC5, BC6, BC7, BC10, BC13, and BC, respectively. For example if *AmrykA* 'America' has the brown cluster ID BC=11110010 then BC7=1111001, whereas BC10 and BC13 are empty strings. This feature is based on the observation that semantically similar words will be grouped together in the same cluster and will have a common prefix;

- **Word2vec Cluster IDs** : Word2vec is an algorithm for learning embeddings using a neural network model (Mikolov et al., 2013). Embeddings are represented by a set of latent variables, where each word is represented by a specific instantiation of these variables. In our system, we apply K-means clustering on the word vectors and use the clusters IDs as features.

3.3 Datasets

We use Microblogs and Dialectal weblogs datasets for our experiments:

- Twitter dataset: We use the training and test data split proposed in (Darwish, 2013), where the training dataset contains 3,646 tweets which were randomly selected from tweets that were authored in the period of May 3-12, 2012. The tweets were scraped from Twitter using the query lang:ar. The testing data contains 1,423 tweets that were randomly selected from tweets authored between November 23, 2011 and November 27, 2011. This dataset has also been used in (Darwish and Gao, 2014) for testing. Both datasets are annotated using the Linguistics Data Consortium ACE tagging guidelines;

- Dialectal Arabic dataset (DA-EGY): The annotated data was chosen from a set of web blogs that are manually identified by LDC as Egyptian dialect and contains nearly 40k tokens. The data was annotated by one native Arabic speaker annotator who followed the Linguistics Data Consortium guidelines for tagging. We use the same 80/20 train/test 5-fold cross validation split proposed in (Zirikly and Diab, 2014)

Table 2 shows dataset statistics, namely number of tokens, and the named entity types: PER, LOC, and ORG.

Brown Clustering and word2vec Data In our work, we run brown clustering and word2vec three times based on the data genre: i) Newswire

	#Tokens	#PER	#LOC	#ORG
Twitter-Train	55k	788	713	449
Twitter-Test	26k	464	587	316
DA-EGY	24k	311	155	19

Table 2: Twitter and DA-EGY Evaluation data statistics

(NW): Arabic Gigaword, ANERCorp, and NW data of ACE2005 and ACE2006; ii) Broadcast News (BN): BN data of ACE2005 and ACE2006; iii) Weblogs (WL): Twitter data (training and testing), WL data of ACE2005 and ACE2006, and Arabic Dialect[3]. The number of Brown and word2vec clusters is empirically chosen; the best results achieved are: 500, 200, 200 for brown clustering on NW, BN, and WL respectively, as opposed to: 300,150,150 for word2vec on NW, BN, and WL respectively.

Parametric features values We use the following values for the parametric features:

- CTX features: we set context window = ± 1 for words;
- LM-GAZ: Threshold of the number of deletion, insertion, or modification ≤ 2;
- BC: Length of the prefixes of brown clusters ID is set to 4,5,6,7,10,13, and full length of cluster ID.

3.4 Data Normalization and Preprocessing

Arabic normalization has proven to improve retrieval results (Darwish et al., 2012). We apply the following normalizations on training, testing, BC and word2vec input data: i) Number normalizations: $[0-9] \rightarrow 8$; ii) Hamza normalization: hamza numerous forms are used interchangeably depending on the role of a word in the sentence. For instance, the term "his sky" can be written smA&h, smA'h, or smA}h, where the hamza takes its form based on the term being subject, object, or idafa (construct state indicating possessive), respectively $', >, <, \&, \}, |, \{, `, Y \rightarrow A$; iii) Normalizing elongated words: We remove consecutive repeated letters that occur > 2.

Tools In this work, we use the following tools:

1. MADAMIRA (Pasha et al., 2014): For tokenization preprocessing and morphological features such as gender and POS tags;
2. CRFSuite implementation (Okazaki, 2007).

4 Experiments & Discussion

4.1 Features set

The list of feature sets used in our experiments are:

- Feature set1 (FS1): Baseline (BL) features, as proposed in (Abdul-Hamid and Darwish, 2010) with the use of exact match against the Wikipedia gazetteers (WikiGaz) for PER, LOC, and ORG named entity types;
- Feature set2 (FS2): BL features with the use of CAPS (English gloss capitalization) and the current, previous and next POS;
- Feature set3 (FS3): FS2 features in addition to ENC0, PROC0, PROC1, PROC2, PROC3, as demonstrated in Table 1;
- Feature set4 (FS4): FS3 in addition to isNum, isNoun binary features;
- Feature set5 (FS5): FS4 features with the use of word2vec cluster IDs;
- Feature set6 (FS6): FS4 features with the use of BC cluster IDs with different prefixes length;
- Feature set7 (FS7): FS6 features with the use of word2vec cluster IDs;
- Feature set8 (FS8): FS7 features with the use of exact (EM-GAZ) and partial (PM-GAZ) match against WikiGaz gazetteers' entries;
- Feature set9 (FS9): FS8 features in addition to the use of Levenshtein gazetteers' entires match with distance threshold set to 1 (LVM-GAZ1);
- Feature set10 (FS10): FS7 features in addition to the use of Levenshtein gazetteers' entires match with distance threshold set to 2 (LVM-GAZ2);
- Feature set11 (FS11): FS8 features in addition to LVM-GAZ1 and LVM-GAZ2;
- Feature set12 (FS12): FS11 with the use pf ASP and PERS morphological features;
- Feature set13 (FS13): FS7 in addition to ASP and PERS features;
- Feature set14 (FS14): FS6 in addition to ASP and PERS features;
- Feature set15 (FS15): FS5 in addition to ASP and PERS features;

[3] LDC2012T09

4.2 Evaluation Metrics

We choose precision (PREC), recall (REC), and harmonic F-measure (F1) metrics to evaluate the performance of our NER system over accuracy. This decision is based on the observation that the baseline accuracy is always high as the majority of the words in free text are not named entities.

4.3 Results & Discussion

Twitter Results: Table 3 illustrates results of our NER system performance. We use the weighted macro-average across the three NEs (PER, LOC, ORG) to calculate the overall performance. Although we were not able to replicate (Darwish and Gao, 2014) results with WikiGaz (F1=55% vs. 51.62%), but our proposed features coupled with BC and word2vec surpass their performance yielding an F1=57.84% without the use of any external resources vs. 59.59% with the use of gazetteers.[4]Although word2vec and BC increase F1 $\approx$ 10% over BL, we note that BC impact (+6%) is more significant in comparison to word2vec with only 3% improvement. It is worth mentioning that this is aligned with (Turian et al., 2010) observations that Brown Clustering yields better English NER performance as opposed to word embedding. This is due to Brown Clustering's ability to induce rare words compared to word embedding. We also note that our intuition for using Levenshtein Matching approach, LVM-GAZ, against gazetteers' entries to overcome non-standardization issue in DA shows 0.8% improvement over EX-GAZ and PAR-GAZ. We should note that LVM-GAZ very much depends on the percentage of present DA variety in the data. The results achieved are promising, especially in the area of social media since generating gazetteers that have high coverage is a challenging and expensive task.

When observing the MORPH feature set in more details, we notice that CAPS and POS yield the highest improvement over the baseline, especially in the PER class, this is mainly due to the correct assignment of the Proper Noun POS tag to this class confirming that POS tag is a strong indicator for NE.

We study the impact of applying BC and word2vec on different data genre. We take as an example BC, shown in Figure 1. We note that genre variations impose minimum impact on word representations, thus we can induce that word2vec and BC presents robust and domain-independent features.

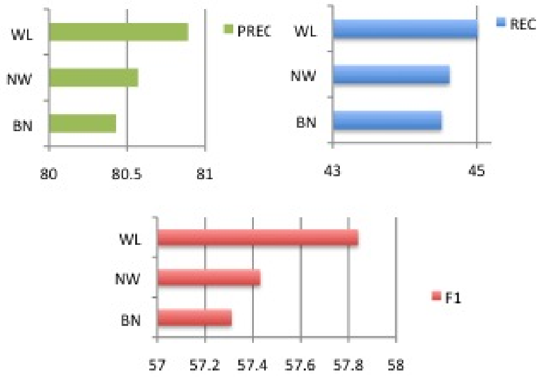

Figure 1: BC Data Genre and Performance correlation

DA-EGY Results: We apply the feature sets that yields the best result with and without the use of gazetteers in Table 3 to our second evaluation dataset DA-EGY. The reported result is the average of 5-fold cross validation. As proposed in (Zirikly and Diab, 2014), we omit ORG class because there is less than 0.05% instances of ORG in the annotated data, which does not represent a fair training data to the system. Our system outperforms the state-of-the-art results by $\approx$ 7% with the use of gazetteers, and $\approx$ 2% without the use of gazetteers. As shown in Table 4, we notice that FS15, which uses word2vec features and excludes BC features and gazetteers, generate very comparable results (72.61%) to the best gazetteers-free performance achieved 72.68%.

5 Conclusion & Future Work

In this paper we study the impact of word representations and embedding on Arabic NER system for social media data. We show that our proposed gazetteers-free features surpass other NER systems that use large gazetteers. This is a significant advantage since gazetteers are expensive to generate, especially in the area of social media due to the low coverage of dictionaries. We show that our proposed system improves NER performance and outperforms state-of-the-art results for Dialectal Arabic.

In future work, we would like to test the impact of cross-lingual word embedding and representation

[4] It should be noted that our use of the gazetteers is probably different from theirs thereby rendering our results with gazetteers incomparable to their results.

	LOC	ORG	PER	Overall		
	F1	F1	F1	PREC	REC	F1
BL	49.15	38.38	48.78	83.25	31.77	45.99
BL+WikiGaz (Darwish and Gao, 2014)	65.5	41.5	48.5	79.3	42.1	55
FS1=BL + EX-GAZ (LOC,ORG,PER)	52.37	41.72	57.09	83.07	37.44	51.62
FS2=BL+CAPS+POS[-1,0,1]	51.23	38.6	59.2	79.89	38.1	51.59
FS3=FS2 +PROC{3,2,1,0}+ENC0	51.73	39.55	60.78	79.01	39.53	52.7
FS4=FS3+isNum+isNoun	51.56	39.48	60.91	79.11	39.53	52.72
FS5=FS4+word2vec	53.18	38.37	62.74	79.54	40.57	53.74
FS6=FS4+BC	55.41	40.46	65.02	81.83	42.6	56.03
FS7=FS6+word2vec	56.78	39.74	65.58	82.01	43.12	56.52
FS8=FS7+EX-GAZ+PAR-GAZ	58.02	41.87	67.09	81.83	44.94	58.02
FS9=FS8+LVM-GAZ1	58.1	40.71	67.33	81.25	44.94	57.87
FS10=FS7+LVM-GAZ2	59.63	41.4	67.39	81.71	45.47	58.42
FS11=FS9+LVM-GAZ2	59.63	41.28	68.17	81.93	45.86	58.8
FS12=FS11+ASP+PERS	61.03	41.28	68.92	81.7	46.9	**59.59**
FS13=FS7+ASP+PERS	58.29	38.11	68.32	80.89	45.01	**57.84**

Table 3: Twitter NER Results

	LOC	PER	Overall		
	F1	F1	PREC	REC	F1
State-of-the-art	91.43	49.18	86.53	62.3	70.31
FS12	96.77	57.47	82.9	72.39	**77.12**
FS13	89.66	55.7	86.67	63.08	**72.68**
FS14	89.66	54.05	90	61.04	*71.86*
FS15	89.66	55.56	93.48	61.04	*72.61*

Table 4: DA-EGY NER Results

features on NER performance and test our system with numerous different domains.

6 Acknowledgment

This work was supported by the Defense Advanced Research Projects Agency (DARPA) Contract No. HR0011-12-C-0014, the BOLT program with subcontract from Raytheon BBN. We would like to thank three anonymous reviewers for their comprehensive feedback

References

Sherief Abdallah, Khaled Shaalan, and Muhammad Shoaib. 2012. Integrating rule-based system with classification for arabic named entity recognition. In *Computational Linguistics and Intelligent Text Processing*, pages 311–322. Springer.

Ahmed Abdul-Hamid and Kareem Darwish. 2010. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, NEWS '10, pages 110–115, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alfred V. Aho and Margaret J. Corasick. 1975. Efficient string matching: An aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153. Citeseer.

Yassine Benajiba, Paolo Rosso, and José-Miguel Benedí. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *CICLing*, pages 143–153.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2008. Arabic named entity recognition using optimized feature sets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 284–293. Association for Computational Linguistics.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2009. Arabic named entity recognition: A feature-driven study. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):926–934.

Yassine Benajiba, Imed Zitouni, Mona Diab, and Paolo Rosso. 2010. Arabic named entity recognition: Using features extracted from noisy data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 281–285, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Kareem Darwish and Wei Gao. 2014. Simple effective microblog named entity recognition: Arabic as an example. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 2513–2517.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2427–2430, New York, NY, USA. ACM.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hal Daumé, III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 53–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ali Elsebai, Farid Meziane, and Fatma Zohra Belkredim. 2009. A rule based persons names arabic extraction system. *Communications of the IBIMA*, 11(6):53–59.

Sergio Ferrndez, Antonio Toral, scar Ferrndez, Antonio Ferrndez, and Rafael Muoz. 2007. Applying wikipedias multilingual knowledge to cross-lingual question answering. In *In Zoubida Kedad, Nadira Lammari, Elisabeth Mtais, Farid Meziane, and Yacine Rezgui, editors, NLDB, volume 4592 of Lecture Notes in Computer Science*. Springer.

Nizar Habash, Mona T Diab, and Owen Rambow. 2012. Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal arabic. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 426–432.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

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

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

David Nadeau, Peter Turney, and Stan Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity.

Naoaki Okazaki. 2007. Crfsuite: A fast implementation of conditional random fields (crfs).

Arfath Pasha, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Khaled Shaalan and Mai Oudah. 2014. A hybrid approach to arabic named entity recognition. *Journal of Information Science*, 40(1):67–87.

Khaled Shaalan and Hafsa Raza. 2009. Nera: Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(8):1652–1663.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. *Comput. Linguist.*, 40(2):469–510, June.

Paul Thompson and Christopher C. Dozier. 1997. Name searching and information retrieval. In *In Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, pages 134–140.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Ayah Zirikly and Mona Diab. 2014. Named entity recognition system for dialectal arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 78–86, Doha, Qatar, October. Association for Computational Linguistics.