Named Entity Recognition for Dialectal Arabic

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

To date, majority of research for Arabic Named Entity Recognition (NER) addresses the task for Modern Standard Arabic (MSA) and mainly focuses on the newswire genre. Despite some common characteristics between MSA and Dialectal Arabic (DA), the significant differences between the two language varieties hinder such MSA specific systems from solving NER for Dialectal Arabic. In this paper, we present an NER system for DA specifically focusing on the Egyptian Dialect (EGY). Our system delivers $\approx$ 16% improvement in F1-score over state-of-the-art features.

1 Introduction

Named Entity Recognition (NER) aims to identify predefined set of named entities types (e.g. Location, Person) in open-domain text (Nadeau and Sekine, 2007). NER has proven to be an essential component in many Natural Language Processing (NLP) and Information Retrieval tasks. In (Thompson and Dozier, 1997), the authors show the significant impact NER imposes on the retrieval performance, given the fact that names occur with high frequency in text. Moreover, in Question Answering, (Ferrndez et al., 2007) report that Questions on average contain $\approx 85\%$ Named Entities.

Although NER has been well studied in the literature, but the majority of the work primarily focuses on English in the newswire genre, with near-human performance (f-score$\approx 93\%$ in MUC-7). Arabic NER has gained significant attention in the NLP community with the increased availability of annotated datasets. However, due to the rich morphological and highly inflected nature of Arabic language (Ryding, 2005), Arabic NER faces many challenges (Abdul-Hamid and Darwish, 2010), that manifest in:

- Lack of capitalization: Unlike English (and other Latin-based languages), proper nouns are not capitalized, which renders the identification of NER more complicated;
- Proper nouns can also represent regular words (e.g. *jamilah, gmylp* [1]" which means 'beautiful' and can be a proper noun or an adjective;
- Agglutination: Since Arabic exhibits concatenate morphology, we note the pervasive presence of affixes agglutinating to proper nouns as prefixes and suffixes (Shaalan, 2014). For instance: Determiners appear as prefixes as in *Al* (*AlqAhrp* 'Cairo'), likewise with affixival prepositions such as *l* meaning 'for' (*ldm$q* -'to/from Damascus'-), as well as prefixed conjunctions such as *w* meaning 'and' (*wAlqds* -'and Jerusalem'-);
- Absence of Short Vowels (Diacritics): Written MSA, even in newswire, is undiacritized; resulting in ambiguity that can only be resolved using contextual information (Benajiba et al., 2009). Instances of such phenomena: *mSr*, which is underspecified for short vowels, can refer to *miSor* 'Egypt' or *muSir* 'insistent'; *qTr* may be 'Qatar' if *qaTar*, 'sugar syrup' if *qaTor*, 'diameter' if *quTor*.

Previously proposed Arabic NER systems (Benajiba et al., 2007) and (Abdallah et al., 2012) were developed exclusively for MSA and primarily address the problem in the newswire genre. Nevertheless, with the extensive use of social networking and web blogs, DA NLP is gaining more

[1] The second form of the name is written in Buckwalter encoding http://www.qamus.org/transliteration.htm

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 78–86,
October 25, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

attention, yielding a more urgent need for DA NER systems. Furthermore, applying NLP tools, such as NER, that are designed for MSA on DA results in considerably low performance, thus the need to build resources and tools that specifically target DA (Habash et al., 2012).
In addition to the afore mentioned challenges for Arabic NER in general compared to Latin based languages, DA NER faces additional issues:

- Lack of annotated data for supervised 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 enough 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.

In this paper, we propose a DA NER system – using Egyptian Arabic (EGY) as an example dialect. Our contributions are as follows:

- Provide an annotated dataset for EGY NER;
- To the best of our knowledge, our system is one of the few systems that specifically targets DA.

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 set of used features such as contextual and morphological. 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 PER, 85.9% for LOC, and 83.15% for ORG when evaluated on corpora built by the authors. (Elsebai et al., 2009) proposed a rule-based system that is targeted for personal NEs in MSA and utilizes the Buckwalter Arabic Morphological Analyser (BAMA) and a set of keywords used to introduce a PER NE. The proposed system yields an F-score of 89% when tested on a dataset of 700 news articles extracted from Aljazeera television website. Although this approach proved to be successful, but most of the recent research focuses on Statistical Learning techniques for NER (Nadeau and Sekine, 2007).
In the area of Statistical Learning for NER, numerous research studies have been published. (Benajiba et al., 2007) proposes a system (ANERsys) based on n-grams and maximum entropy. The authors also introduce ANERCorp corpora and ANERGazet gazetteers. (Benajiba and Rosso, 2008) presents NER system (ANERsys) for MSA based on CRF sequence labeling, where the system uses language independent features: POS tags, Base Phrase Chunking (BPC), gazetteers, and nationality information. The latter feature is included based on the observation that personal NEs come 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 are a combination of general features as listed in (Benajiba and Rosso, 2008) and Arabic-dependent (morphological) features. Their system's best performance was 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, that we incorporate a subset of in our DA NER

system, namely, the head and trailing bigrams (L2), trigrams (L3), and 4-grams (L4) characters.
(Shaalan and Oudah, 2014) presents a hybrid approach that targets MSA and produces state-of-the-art results. However, due to the lack of availability of the used rules, it is hard 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 are a combination of the rule-based features in addition to morphological, capitalization, POS tag, word length, and dot (has an adjacent dot) features. We reimplement their Machine Learning component and present it as one of our baselines (BAS2). (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 (BAS1). The proposed system does not rely on any external resources and the system outperforms (Benajiba and Rosso, 2008) performance with an F-score of 81% on ANERCorp vs. the latter's performance of 72.68% F-score. All the work mentioned has 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 set of features that are similar to (Abdul-Hamid and Darwish, 2010). Their dataset contains dialectal data, since it is collected from Twitter. However, the dataset contains English and Arabic; in this work we only target Dialectal Arabic. Their overall performance, on their proposed data, is 65.2% (LOC 76.7%, 55.6% ORG, 55.8% PER).

3 Approach

In this paper, we use a supervised machine learning approach since it has been shown in the literature that supervised typically outperform unsupervised approaches for the NER task (Nadeau et al., 2006). We use Conditional Random Field (CRF) sequence labeling as described in (Lafferty et al., 2001). Moreover, (Benajiba and Rosso, 2008) demonstrates that CRF yields better results over other supervised machine learning techniques.

3.1 Baseline

In this paper, we introduce two baselines to compare our work against. The first baseline (BAS1) is based on work reported in (Abdul-Hamid and Darwish, 2010). We adopt their approach since it produces near state-of-the-art results. Additionally, the features proposed are applicable to DA as they do not rely on the availability of morphological or syntactical analyzers. We reimplement their listed features that yield the highest performance and report those results as our BAS1 system. The list of features used are: previous and next word, in addition to the leading and trailing character bigrams, trigrams, and 4-grams.
The second baseline (BAS2) adopted is the work proposed in (Shaalan and Oudah, 2014). The authors present state-of-the-art results when evaluated on ANERcorp (Benajiba and Rosso, 2008) using the following features: Rule-based features, Morphological features generated by MADAMIRA (Pasha et al., 2014) presented in Table 1, targeted word POS tag, word length flag which is a binary feature that is true if the word length is ≥ 3, a binary feature to represent whether the word has an adjacent dot, capitalization binary feature which is dependent on the English gloss generated by MADAMIRA, nominal binary feature that is set to true if the POS tag is noun or proper noun, and binary features to represent whether the current, previous, or next word belong to the gazetteers. We omit Rule-based features in our baseline since we do not have access to the exact rules used and their rules specifically targeted MSA, hence would not be directly applicable to DA.

3.2 NER Features

In our approach, we propose the following NER features:

- **Lexical Features**: Similar to BAS1 (Darwish and Gao, 2014) character n-gram features, the head and trailing bigrams (L2), trigrams (L3), and 4-grams (L4) characters;
- **Contextual Features** (CTX): The surrounding undiacritized lemmas and words of a context window = ± 1; (LEM-1, LEM0, LEM1) and (W-1,W0,W1)
- **Gazetteers** (GAZ): We use two sets of gazetteers. The first set (ANERGaz) proposed by (Benajiba and Rosso, 2008), which

Feature	**Feature Values**
Aspect	Verb aspect: Command, Imperfective, Perfective, Not applicable
Case	Grammatical case: Nominative, Accusative, Genitive, Not applicable, Undefined
Gender	Nominal Gender: Feminine, Masculine, Not applicable
Mood	Grammatical mood: Indicative, Jussive, Subjunctive, Not applicable, Undefined
Number	Grammatical number: Singular, Plural, Dual, Not applicable, Undefined
Person	Person Information: 1st, 2nd, 3rd, Not applicable
State	Grammatical state: Indefinite, Definite, Construct/Poss/Idafa, Not applicable, Undefined
Voice	Verb voice: Active, Passive, Not applicable, Undefined
Proclitic3	Question proclitic: No proclitic, Not applicable, Interrogative particle
Proclitic2	Conjunction proclitic: No proclitic, Not applicable, Conjunction *fa*, Connective particle *fa*, Response conditional *fa*, Subordinating conjunction *fa*, Conjunction *wa*, Particle *wa*, Subordinating conjunction *wa*
Proclitic1	Preposition proclitic: No proclitic, Not applicable, Interrogative *i$*, Particle *bi*, Preposition *bi*, Progressive verb particle *bi*, Preposition *Ea*, Preposition *EalaY*, Preposition *fy*, Demonstrative *hA*, Future marker *Ha*, Preposition *ka*, Emphatic particle *la*, Preposition *la*, Preposition *li* + preposition *bi*, Emphatic *la* + future marker *Ha*, Response conditional *la* + future marker *Ha*, Jussive *li*, Preposition *li*, Preposition *min*, Future marker *sa*, Preposition *ta*, Particle *wa*, Preposition *wa*, Vocative *wA*, vocative *yA*
Proclitic	Article proclitic: No proclitic, Not applicable, Demonstrative particle *Aa*, 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 *man*, Relative pronoun *man, ma, mA*, Subordinating conjunction *ma, mA*.

Table 1: Morphological Features

contains a total of 4893 names between Person (PER), Location (LOC), and Organization (ORG). The second gazetteer is a large Wikipedia gazetteer (WikiGaz) from (Darwish and Gao, 2014); 50141 locations, 17092 organizations, 65557 persons. which represents a significantly more extensive and comprehensive list. We introduce three methods for exploiting GAZ:

- Exact match (EM-GAZ): For more efficient search, we use Aho-Corasick Algorithm 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, EM-GAZ for the first word will take the value ”B-<NE class>” where <NE class>is one of the previously discussed classes (PER, LOC, ORG), whereas the following words will be assigned I-<NE class>, where <NE class>will be assigned the same value of the matched sequence's head;

- 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 we have in the gazetteer the compound name *yAsr ErfAt* 'Yasser Arafat' and the input text is *yAsr BarakAt* then PM-GAZ for the token *yAsr* will be set to true. This is particularly useful in the case of PER as it recovers a large list of first names in compounds;

- Levenshtein match (LVM-GAZ): Due to the non-standard spelling of words in dialectal Arabic, we use Levenshtein distance (Levenshtein, 1966) to compare the similarity between the input and a gazetteer entry;

• **Morphological Features**: The morphological features that we employ in our feature set are generated by MADAMIRA (Pasha et al., 2014):

  - Gender (GEN): Since Arabic nouns are either masculine or feminine, we believe that this information should help NER. Moreover, instances of the same name will share the same gender. MADAMIRA generates three values for this feature: Feminine, Masculine, or Not Applicable (such as the case for prepositions, for instance);

  - 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 that is generated by MADAMIRA;

  - Part of Speech (POS) tags: We use POS tags generated from MADAMIRA, where the POS tagger has a reported accuracy of 92.4% for DA;

- **Distance from specific keywords** within a window (KEY): This feature captures certain patterns in person names that are more commonly used in DA (e.g. using the nickname pattern of *Abw* + proper noun instead of an actual name). In this feature, if the distance is set to one, the feature will be true if the previous token equals an entry in a keywords list, otherwise false. Examples of keywords: *Abw* 'father of', *yA* invocation particle, typically used before names to call a person, terms of address, or honorifics, such as *dktwr/dktwrp* 'doctor -masculine and feminine-', and *AstA*/AstA*p* 'Mr/Mrs/Ms/teacher -masculine and feminine-';

- **Brown Clustering** (BC): Brown clustering as introduced in (Brown et al., 1992) is a hierarchical clustering approach that maximizes the mutual information of word bigrams. Word representations, especially Brown Clustering, have been demonstrated to improve the performance of NER system when added as a feature (Turian et al., 2010). In this work, we use Brown Clustering IDs of variable prefixes length (4,7,10,13) as features resulting in the following set of features BC4, BC7, BC10, BC13. For example if *AmrykA* 'America' has the brown cluster ID 11110010 then BC4 = 1111, 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.

4 Experiments & Discussion

4.1 Datasets and Tools

Evaluation Data Due to the very limited resources in DA for NER, we manually annotate a portion of the DA data collected and provided by the LDC from web blogs.[2] 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 (LDC) guidelines for NE tagging. Our dataset is relatively small and contains 285 PER, 153 LOC, and 10 ORG instances.

Brown Clustering Data In our work, we run Brown Clustering on BOLT Phase1 Egyptian Arabic Treebank (ARZ)[3], where the chosen number of clusters is 500.

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

- CTX features: we set context window = ±1 for lemmas and tokens;

- Keyword distance: we set the distance from the token to a keyword to 1 and 2, namely, KEY1 and KEY2, respectively;

- LM-GAZ: The threshold of the number of deletion, insertion, or modification ≤ 2;

- BC: the length of the prefixes of the Brown Clusters ID is set to 4,7,10,13;

Tools In this work, we used the following tools:

1. MADAMIRA (Pasha et al., 2014): For tokenization and other features such as lemmas, gender and Part of Speech (POS) tags, and other morphological features;

2. CRFSuite implementation (Okazaki, 2007).

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 on the token level in NER is not

[2] GALE Arabic-Dialect/English Parallel Text LDC2012T09

[3] LDC2012E98

a fair assessment, since NER accuracy is always high as the majority of the tokens in free text are not named entities.

4.3 Results & Discussion

In our NER system, we solely identify PER and LOC NE classes and omit the ORG class. This is due to the small frequency ($\leq 0.05\%$) of ORG instances in our annotated data, which does not represent a fair training data to the system. The reported results are the average of 5-fold cross validation on the blog post level. Also, it is worth mentioning that we use IOB tagging scheme; Inside *I* NE, Outside *O*, and Beginning *B* of NE. Table 2 depicts the two baselines discussed in 3.1. BAS1 yields a weighted macro-average F-score=54.762% using near state-of-the-art features on our annotated data. On the other hand, BAS2 F-score is 31%. Although BAS2 presents state-of-the-art results, it actually produces lower performance than BAS1. It should be noted that our implementation of BAS2 does not incorporate rule-based features (Shaalan and Oudah, 2014). However, by extrapolation using their performance improvement of $\approx 6\%$ attributed to rule-based features alone, such a relative gain in performance for BAS2 in our setting would still be outperformed by both BAS1 and our current system.

In Table 3, we show our NER system performance using different permutations of features proposed in Section 3.2. Additionally, in Table 3, we use the weighted macro-average (Overall) in order to assess the system's overall performance. We use the following abbreviation annotation:

- FEA1: includes n-gram characters and CTX on the word and lemma level features;
- FEA2: includes FEA1 in addition to KEY features with distance 1&2;
- FEA3: includes FEA2 in addition to the morphological features (MORPH) and it is subcategorized as follow: FEA3-GEN takes into account the gender feature only, FEA3-POS takes into account POS tag (FEA2+POS), whereas FEA3-CAPS takes into account the use of CAPS with FEA2;
- FEA4: shows the impact of adding EM-GAZ features (FEA3+EM-GAZ);
- FEA5: shows the impact of adding PM-GAZ features (FEA4+PM-GAZ);
- FEA6: shows the impact of adding LVM-GAZ features (FEA5+LM-GAZ);
- FEA7: shows the impact of adding Brown Clustering (BC) features on the performance;

The best results for precision, recall and F1-score are bolded in Table 3. FEA6 delivers the best NER performance of F1-score=70.305%

Baseline		PREC	REC	F1
BAS1	*LOC*	80	72.727	76.191
	PER	56.25	23.684	33.333
	AVG	68.125	48.201	**54.762**
BAS2	*LOC*	47.368	52.941	50
	PER	8.571	20	12
	AVG	27.97	36.471	31

Table 2: Baseline NER performance

In comparing FEA1, FEA2 results, we note that KEY features increase the F1-score by 2% absolute. This improvement mirrors the fact that *Abw*+name, for example, is very commonly used in dialects, where it represents $\approx 46\%$ of PER names. The morphological features (GEN, POS, CAPS), produce the most significant improvement $\approx +9\%$ absolute. Although the gazetteers help NER performance overall, the boost is not as significant as with using the MORPH features. Likewise, we note that LVM-GAZ using Levenshtein distance addresses the spelling variation challenge that DA pose and yields the best performance (F1-score=70.305%) when combining all features except the Brown clustering. Unlike the BC effect noted in English NER case studies, BC degrades the performance of our DA NER system. We further analyze this result by closely examining the clustering quality obtained on the dataset. For example, the following instances of the LOC class from our dataset: *mSr* 'Egypt', *AmrykA* 'America', and *qtr* 'Qatar'; the cluster IDs assigned by the Brown Clustering algorithm are 111101110, 11110010, 00111000, respectively. The common prefix among the three instances is very short (1111 in case of Egypt and America and none with Qatar), thus leading to poorer performance.

Overall, we note more stable performance for LOC class in comparison to PER. This is mainly due to the high PER singleton instances frequencies which results in high unseen vocabulary in

Features	***LOC***			***PER***			***Overall***		
	PREC	*REC*	*F1*	*PREC*	*REC*	*F1*	*PREC*	*REC*	*F1*
FEA1={L2,L3,L4,W-1,W0,W1,LEM-1,LEM0,LEM1}	93.333	77.778	84.849	54.546	14.286	22.642	73.94	46.032	53.746
FEA2={FEA1, KEY1, KEY2}	93.75	83.333	88.235	60	14.286	23.077	76.875	48.81	55.656
FEA3-GEN={FEA2, GEN}	93.75	83.333	88.235	63.636	16.667	26.415	78.693	50	57.325
FEA3-POS={FEA2, POS}	93.333	77.778	84.849	78.571	26.191	39.286	85.952	51.985	62.068
FEA3-CAPS={FEA2, CAPS}	93.333	77.778	84.849	78.571	26.191	39.286	85.952	51.985	62.068
FEA3={FEA2, MORPH}	94.118	88.889	91.429	83.333	23.81	37.037	88.7255	56.3495	64.233
FEA4={FEA3, EM-GAZ}	94.118	88.889	91.429	72.222	30.952	43.333	83.17	59.9205	67.381
FEA5={FEA4, PM-GAZ}	94.118	88.889	91.429	73.684	33.333	45.902	83.901	61.111	68.666
FEA6={FEA5, LVM-GAZ}	94.118	88.889	**91.429**	78.947	35.714	**49.18**	86.533	62.302	**70.305**
FEA7={FEA6, BC}	93.333	77.778	84.849	77.778	33.333	46.667	85.556	55.556	65.758

Table 3: Dialectal Arabic NER

the test data. In addition, LOC members, unlike PER, convey tag consistency, where most of the time it will be tagged as NE. For instance, *mSr* ‘Egypt’ occurred in the data 35 times and in all of which it was assigned a LOC tag, unlike *EAdl* that appears as an adjective ’fair/rightful’ and proper name ’Adel’ in the same dataset. The former reason explains why the GAZ helps PER class performance but does not affect LOC performance.

If we discuss in more detail the MORPH feature set, we notice that CAPS and POS produce identical results in terms of PREC, REC, and F-1 score on each of the NE classes. However, CAPS and POS help in PER class, whereas GEN helps in the LOC class. For example in LOC class, the number of false negatives, when POS is employed, is higher as opposed to GEN.

As mentioned earlier, LVM-GAZ produces the best F-score. However, LVM main contribution is on the PER class which is caused by the nature of Arabic names’ different spelling variations, especially the last name (e.g. with or without Al).

5 Conclusion & Future Work

In this paper we present Dialectal Arabic NER system using state-of-the-art features in addition to proposing new features that improve the performance. We show that our proposed system improves over state-of-the-art features performance. Our contribution is not solely limited to the NER system, but further includes, our manually annotated data.[4] In future work, we would like to annotate more data in more variable genre and with more dialects including code switched data.

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.

[4] Please contact the authors for access to the annotated data.

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.

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.

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.

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.

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*.

Karin C Ryding. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

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.