Simplified Feature Set for Arabic Named Entity Recognition

Ahmed Abdul-Hamid, Kareem Darwish
Cairo Microsoft Innovation Center
Cairo, Egypt
{ahmedab,kareemd}@microsoft.com

Abstract

This paper introduces simplified yet effective features that can robustly identify named entities in Arabic text without the need for morphological or syntactic analysis or gazetteers. A CRF sequence labeling model is trained on features that primarily use character n-gram of leading and trailing letters in words and word n-grams. The proposed features help overcome some of the morphological and orthographic complexities of Arabic. In comparing to results in the literature using Arabic specific features such POS tags on the same dataset and same CRF implementation, the results in this paper are lower by 2 F-measure points for locations, but are better by 8 points for organizations and 9 points for persons.

1 Introduction

Named entity recognition (NER) continues to be an important part of many NLP applications such as information extraction, machine translation, and question answering (Benajiba et al., 2008). NER is concerned with identifying sequences of words referring to named entities (NE's) such as persons, locations, and organizations. For example, in the word sequence "Alan Mulally, CEO of Detroit based Ford Motor Company," Alan Mulally, Detroit, and Ford Motor Company would be identified as a person, a location, and an organization respectively.

Arabic is a Semitic language that present interesting morphological and orthographic challenges that may complicate NER. Some of these challenges include:

- Coordinating conjunctions, prepositions, possessive pronouns, and determiners are typically attached to words as prefixes or suffixes.
- Proper names are often common language words. For example, the proper name "Iman" also means faith.
- Lack capitalization of proper nouns.

The paper introduces a simplified set of features that can robustly identify NER for Arabic without the need for morphological or syntactic analysis. The proposed features include: word leading and trailing character n-gram features that help handle prefix and suffix attachment; word n-gram probability based features that attempt to capture the distribution of NE's in text; word sequence features; and word length.

The contributions of this paper are as follows:

1. Identifying simplified features that work well for Arabic without gazetteers and without morphological and syntactic features, leading to improvements over previously reported results.
2. Using leading and trailing character n-grams in words, which help capture valuable morphological and orthographic clues that would indicate or counter-indicate the presence of NE's.
3. Incorporating word language modeling based features to capture word associations and relative distribution of named entities in text.

Conditional Random Fields (CRF) sequence labeling was used in identifying NE's, and the experiments were performed on two standard Arabic NER datasets.

The rest of the paper is organized as follows: Section 2 surveys prior work on Arabic NER; Section 3 introduces the proposed features and motivates their use; Section 4 describes experimental setup and evaluation sets; Section 5 reports on experimental results; and Section 6 concludes the paper.

2 Background

Much work has been done on NER with multiple evaluation forums dedicated to information extraction in general and to NER in specific. Nadeau and Sekine (2009) surveyed lots of work on NER for a variety of languages and using a myriad of techniques. 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.,

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 110–115,
Uppsala, Sweden, 16 July 2010.

2007). Benajiba et al. (2007) used a maximum entropy based classification trained on a feature set that include the use of gazetteers and a stop-word 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. 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 improvement 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 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 morphological analyzer 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.

McNamee and Mayfield (2002) explored the training of an SVM classifier using many language independent binary features such as leading and trailing letters in a word, word length, presence of digits in a word, and capitalization. They reported promising results for Spanish and Dutch. In follow on work, Mayfield et al. (2003) used thousands of language independent features such character n-grams, capitalization, word length, and position in a sentence, along with language dependent features such as POS tags and BP chunking. For English, they reported 89%, 79%, and 91% F-measure for location, organization, and persons respectively.

The use of CRF sequence labeling has been increasing over the past few years (McCallum and Li, 2003; Nadeau and Sekine, 2009) with good success (Benajiba and Rosso, 2008). Though, CRF's are not guaranteed to be better than SVM's (Benajiba et al., 2008).

3 NER Features

For this work, a CRF sequence labeling was used. The advantage of using CRF is that they combine HMM-like generative power with classifier-like discrimination (Lafferty et al., 2001; Sha and Pereira, 2003). When a CRF makes a decision on the label to assign to a word, it also accounts for the previous and succeeding words. The CRF was trained on a large set of surface features to minimize the use of Arabic morphological and syntactic features. Apart from stemming two coordinating conjunctions, no other Arabic specific features were used.

The features used were as follows:

- Leading and trailing character bigrams (**6bi**). For a given word composed of the letter sequence l_0^n, where l_0 and l_n are a start and end word markers respectively, the first three bigrams (l_0^1, l_1^2, and l_2^3) and last three bigrams (l_{n-3}^{n-2}, l_{n-2}^{n-1}, and l_{n-1}^{n}) were used as features. Using leading and trailing character bigrams of a word was an attempt to account for morphological and orthographic complexities of Arabic and to capture surface clues that would indicate the presence of a NE or not. For example, plural forms of common words in Arabic are often obtained by attaching the suffixes *wn*[1] (ون) or *yn* (ين) for masculine nouns and *At* (ات) for feminine nouns. Presence of such plural form markers would generally indicate a plural noun, but would counter-indicate a NE. Also, verbs in present tense start with the letters *A* (ا), *t* (ت), *y* (ي), and *n* (ن). These would contribute to concluding that a word may not be a NE. Further, coordinate conjunctions, such as *f* (ف) and *w* (و), and prepositions, such as *b* (ب), *k* (ك), and *l* (ل), composed of single letters are often attached as prefixes to words. Accounting for them may help overcome some of the problems associated with not

[1] Arabic letters are presented using the Buckwalter transliteration scheme

stemming. Further, the determiner *Al* (ال) may be a good indicator for proper nouns particularly in the case of organizations. This would be captured by the second bigram from the head of the word. If the determiner is preceded by a coordinating conjunction, the third bigram from the head of the word would be able to capture this feature.

- Leading and trailing character trigrams (**6tri**). For a given word composed of the letter sequence l_0^n, where l_0 and l_n are a start and end word markers respectively, the first three trigrams (l_0^2, l_1^3, and l_2^4) and last three trigrams (l_{n-4}^{n-2}, l_{n-3}^{n-1}, and l_{n-2}^{n}) were used as features. The rationale for using these features is very similar to that of using character bigrams. The added value of using character trigrams, is that they would allow for the capture of combinations of prefixes and suffixes. For example, a word may begin with the prefixes *w+Al* (ال+و), which are a coordinating conjunction and determiner respectively.
- Leading and trailing character 4-grams (**6quad**). For a given word composed of the letter sequence l_0^n, where l_0 and l_n are a start and end word markers respectively, the first three 4 grams (l_0^3, l_1^4, and l_2^5) and last three 4 grams (l_{n-5}^{n-2}, l_{n-4}^{n-1}, and l_{n-3}^{n}) were used as features. Similar to leading and trailing trigrams, these features can capture combinations of prefixes and suffixes.
- Word position (**WP**). The feature captures the relative position of a word in a sentence as follows:

$$WP = \frac{Absolute\ position}{Sentence\ length}$$

Typically, Arabic is a VSO language. Thus, NE's in specific and nouns in general do not start sentences.

- Word length (**WL**). The feature captures the length of named entities, as some NE's, particularly transliterated NE's, may be longer than regular words.
- Word unigram probability (**1gP**). This is simply the unigram probability of word. Accounting for unigram probability would help exclude common words. Also, named entities are often out-of-vocabulary words.
- Word with previous and word with succeeding word-unigram ratio (**1gPr**). Given a word w_i, these two features are computed as:

$$1gPr_1 = \frac{p(w_i)}{p(w_{i-1})}$$

$$1gPr_2 = \frac{p(w_{i+1})}{p(w_i)}$$

This feature would potentially capture major shifts between word probabilities. For example, a named entity is likely to have much lower probability compared to the word before it and the word after it.

- Features that account for dependence between words in a named entity. Popular NE's are likely collocations, and words that make up named entities don't occur next to each other by chance. These features are as follows:
  - Word with previous and word with succeeding word bigram (**2gP**). For a given word w_i, the two bigram probabilities are $p(w_{i-1}w_i)$ and $p(w_iw_{i+1})$. Words composing named entities are likely conditionally dependent.
  - *t*-test between a word and the word that precedes and succeeds it (**T**). Given a word sequence w_i and w_{i+1}:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Wher$\bar{x} = p(w_i\ w_{i+1}), = p(w_i) * p(w_{i+1})$, $s^2 \approx \bar{x}$, and *N* is the number of words in the corpus (Manning and Schutze, 1999).

  - Mutual information between a word and the word that precedes and succeeds it (**MI**). Given a word sequence w_i and w_{i+1}:

$MI = log_2\left[\frac{\bar{x}}{\mu}\right]$, where $\bar{x}$ and μ are identical to those in the t-test.

- Character n-gram probability (**3gCLM**). Given character trigram language models for locations, persons, organizations, and non-NE's, the four features are just the character language model probabilities using the four different language models. The motivation for these features stem from the likelihood that NE's may have a different distribution of characters particularly for person names. This stems from the fact that many NE's are transliterated names.

4 Experimental Setup

4.1 Datasets

For this work, the NE's of interest were persons, locations, and organizations only. Two datasets were used for the work in this paper. The first

was a NE tagged dataset developed by Binajiba et al. (2007). The Binajiba dataset is composed of newswire articles totaling more than 150,000 words. The number of different NE's in the collection are:

Locations (LOC)	878
Organizations (ORG)	342
Persons (PER)	689

The second was the Arabic Automatic Content Extraction (ACE) 2005 dataset. The ACE dataset is composed of newswire, broadcast news, and weblogs. For experiments in this work, the weblogs portion of the ACE collection was excluded, because weblogs often include colloquial Arabic that does not conform to modern standard Arabic. Also, ACE tags contain many subcategories. For example, locations are tagged as regions, bodies of water, states, etc. All sub-tags were ignored and were conflated to the base tags (LOC, ORG, PER). Further, out of the 40 sub-entity types, entities belonging to the following 13 ACE sub-entity types were excluded because they require anaphora resolution or they refer to non-specific NE's: nominal, pronominal, kind of entity (as opposed to a specific entity), negatively quantified entity, underspecified entity, address, boundary (eg. border), celestial object (comet), entertainment venue (eg. movie theater), sport (eg. football), indeterminate (eg. human), vehicle, and weapon. The total number of words in the collection is 98,530 words (66,590 from newswire and 31,940 from broadcast news). The number of NE's is as follows:

Locations (LOC)	867
Organizations (ORG)	269
Persons (PER)	524

Since both collections do not follow the same tagging conventions, training and testing were conducted separately for each collection. Each collection was 80/20 split for training and testing.

4.2 Data Processing and Sequence Labeling

Training and testing were done using CRF++ which is a CRF sequence label toolkit. The following processing steps of Arabic were performed:

- The coordinating conjunctions *w* (و) and *f* (ف), which always appear as the first prefixes in a word, were optionally stemmed. *w* and *f* were stemmed using an in-house Arabic stemmer that is a reimplementation of the stemmer proposed by Lee et al. (2003). However, stemming *w* or *f* could have been done by stemming the *w* or *f* and searching for the stemmed word in a large Arabic corpus. If the stemmed word appears more than a certain count, then stemming was appropriate.
- The different forms of *alef* (*A* (ا), | (آ), > (أ), and < (إ)) were normalized to *A* (ا), *y* (ي) and *Y* (ى) were normalized to *y* (ي), and *p* (ة) was mapped to *h* (ه).

4.3 Evaluation

The figures of merit for evaluation were precision, recall, and F-measure ($\beta = 1$), with evaluation being conducted at the phrase level. Reporting experiments with all the different combinations of features would adversely affect the readability of the paper. Thus, to ascertain the contribution of the different features, a set of 15 experiments are being reported for both datasets. The experiments were conducted using raw Arabic words (**3w**) and stems (**3s**). Using the short names of features (bolded after feature names in section 3), the experiments were as follows:

- 3w
- 3w_6bi
- 3w_6bi_6tri
- 3w_6bi_6tri_6quad
- 3w_6bi_6tri_6quad_WL
- 3w_6bi_6tri_6quad_WP
- 3s
- 3s_6bi_6tri_6quad
- 3s_6bi_6tri_6quad_1gP
- 3s_6bi_6tri_6quad_1gPr_1gP
- 3s_6bi_6tri_6quad_2gP
- 3s_6bi_6tri_6quad_3gCLM
- 3s_6bi_6tri_6quad_MI
- 3s_6bi_6tri_6quad_T
- 3s_6bi_6tri_6quad_T_MI

5 Experimental Results

Table 1 lists the results for the Benajiba and ACE datasets respectively. Tables 2 and 3 report the best obtained results for both datasets. The results include precision (**P**), recall (**R**), and F-measure (**F**) for NE's of types location (LOC), organization (ORG), and person (PER). The best results for P, R, and F are bolded in the tables.

In comparing the base experiments **3w** and **3s** in which the only the surface forms and the stems were used respectively, both produced the highest precision. However, **3s** improved recall over **3w** by 7, 13, and 14 points for LOC, ORG, and PER respectively on the Benajiba dataset. Though using **3s** led to a drop in P for ORG

compared to **3w**, it actually led to improvement in P for PER. Similar results were observed for the ACE dataset, but the differences were less pronounced with 1% to 2% improvements in recall. However, when including the **6bi**, **6tri**, and **6quad** features the difference between using words or stems dropped to about 1 point in recall and nearly no difference in precision. This would indicate the effectiveness of using leading and trailing character n-grams in overcoming morphological and orthographic complexities.

		Benajiba			ACE		
Run Name	Type	P	R	F	P	R	F
3w	LOC	**96**	59	73	88	59	71
	ORG	**92**	36	51	**87**	50	63
	PER	90	32	48	**94**	47	63
3w_6bi	LOC	92	75	82	85	72	78
	ORG	83	57	67	76	54	63
	PER	87	68	76	89	70	78
3w_6bi_6tri	LOC	93	79	86	87	77	**82**
	ORG	82	61	70	77	56	65
	PER	89	72	80	89	**73**	80
3w_6bi_6tri_6quad	LOC	93	**83**	87	87	77	81
	ORG	84	64	72	77	55	65
	PER	90	73	81	92	71	80
3w_6bi_6tri_6quad_WL	LOC	93	82	87	87	**78**	**82**
	ORG	83	64	73	79	56	65
	PER	89	73	80	93	71	**81**
3w_6bi_6tri_6quad_WP	LOC	91	82	86	88	77	82
	ORG	83	62	71	77	**59**	**67**
	PER	89	74	81	91	70	79
3s	LOC	**96**	66	78	**89**	60	72
	ORG	88	49	63	86	52	65
	PER	**93**	46	61	92	49	64
3s_6bi_6tri_6quad	LOC	93	**83**	**88**	87	77	**82**
	ORG	84	63	72	78	58	**67**
	PER	90	74	81	91	70	80
3s_6bi_6tri_6quad_1gP	LOC	93	**83**	**88**	87	77	**82**
	ORG	84	64	73	79	57	66
	PER	90	**75**	**82**	93	70	80
3s_6bi_6tri_6quad_1gPr_1gP	LOC	93	81	87	87	77	81
	ORG	85	60	70	82	55	66
	PER	91	72	81	93	69	79
3s_6bi_6tri_6quad_2gP	LOC	93	81	87	88	77	**82**
	ORG	85	61	71	82	56	**67**
	PER	89	74	81	90	69	78
3s_6bi_6tri_6quad_3gCLM	LOC	93	82	87	87	76	81
	ORG	84	**65**	**74**	78	56	66
	PER	90	74	81	93	71	**81**
3s_6bi_6tri_6quad_MI	LOC	93	81	86	87	77	**82**
	ORG	84	59	69	82	56	66
	PER	90	72	80	93	70	80
3s_6bi_6tri_6quad_T	LOC	93	81	87	87	76	81
	ORG	85	61	71	82	55	66
	PER	90	72	80	93	69	79
3s_6bi_6tri_6quad_T_MI	LOC	93	80	86	87	76	81
	ORG	85	57	68	82	54	65
	PER	91	71	80	93	67	78

Table 1: NER results for the Benajiba and ACE datasets

	P	R	F
LOC	93	83	88
ORG	84	64	73
PERS	90	75	82
Avg.	89	74	81

Table 2: Best results on Benajiba dataset (Run name: 3s_6bi_6tri_6quad_1gP)

	P	R	F
LOC	87	77	82
ORG	79	56	65
PERS	93	71	81
Avg.	88	70	76

Table 3: Best results on ACE dataset (Run name: 3w_6bi_6tri_6quad_WL)

	P	R	F
LOC	93	87	90
ORG	84	54	66
PERS	80	67	73
Avg.	86	69	76

Table 4: The results in (Benajiba and Rosso, 2008) on Benajiba dataset

The **3s_6bi_6tri_6quad** run produced nearly the best F-measure for both datasets, with extra features improving overall F-measure by at most 1 point.

Using t-test **T** and mutual information **MI** did not yield any improvement in either recall or precision, and often hurt overall F-measure. As highlighted in the results, the 1gP, 2gP, WL, WP, and 3gCLM typically improved recall slightly, often leading to 1 point improvement in overall F-measure.

To compare to results in the literature, Table 4 reports the results obtained by Benajiba and Rosso (2008) on the Benajiba dataset using the CRF++ implementation of CRF sequence labeling trained on a variety of Arabic language specific features. The comparison was not done on their results on the ACE 2005 dataset due to potential difference in tags. The averages in Tables 2, 3, and 4 are macro-averages as opposed to micro-averages reported by Benajiba and Rosso (2008). In comparing Tables 2 and 4, the features suggested in this paper reduced F-measure for locations by 2 points, but improved F-measure for organizations and persons by 8 points and 9 points respectively, due to improvements in both precision and recall.

The notable part of this work is that using a simplified feature set outperforms linguistic features. As explained in Section 3, using leading and trailing character n-grams implicitly capture morphological and syntactic features that typically used for Arabic lemmatization and POS tagging (Diab, 2009). The improvement over using linguistic features could possibly be attributed to the following reasons: not all prefixes and suffixes types equally help in identifying named entities (ex. appearance of a definite article or not); not all prefixes and suffix surface forms equally help (ex. appearance of the coordinating conjunction *w* "و" vs. *f* "ف"); and mistakes in stemming and POS tagging. The lag in recall for locations behind the work of Benajiba and Rosso (2008) could be due to the absence of location gazetteers.

6 Conclusion and Future Work

This paper presented a set of simplified yet effective features for named entity recognition in Arabic. The features helped overcome some of the morphological and orthographic complexities of Arabic. The features included the leading and trailing character n-grams in words, word association features such as t-test, mutual information, and word n-grams, and surface features such word length and relative word position in a sentence. The most important features were leading and trailing character n-grams in words. The proposed feature set yielded improved results over those in the literature with as much as 9 point F-measure improvement for recognizing persons.

For future work, the authors would like to examine the effectiveness of the proposed feature set on other morphologically complex languages, particularly Semitic languages. Also, it is worth examining the combination of the proposed features with morphological features.

References

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 284–293, 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, LREC'08.

Y. Benajiba, P. Rosso and J. M. Benedí. 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.

M. Diab. 2009. *Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking*. Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2009.

B. Farber, D. Freitag, N. Habash, and O. Rambow. 2008. *Improving NER in Arabic Using a Morphological Tagger*. In Proc. of LREC'08.

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

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.

Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, Hany Hassan. 2003. *Language Model Based Arabic Word Segmentation*. ACL 2003: 399-406

C. Manning and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

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.

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.

F. Sha and F. Pereira. 2003. *Shallow parsing with conditional random fields*, In Proc. of HLT/NAACL-2003.

K. Shaalan and H. Raza. 2007. Person Name Entity Recognition for Arabic. Proceedings of the 5th Workshop on Important Unresolved Matters, pages 17–24, Prague, Czech Republic, June 2007.