Simple Effective Microblog Named Entity Recognition: Arabic as an Example

Kareem Darwish, Wei Gao

Qatar Computing Research Institute
Qatar Foundation, Doha, Qatar
{kdarwish,wgao}@qf.org.qa

Abstract

Despite many recent papers on Arabic Named Entity Recognition (NER) in the news domain, little work has been done on microblog NER. NER on microblogs presents many complications such as informality of language, shortened named entities, brevity of expressions, and inconsistent capitalization (for cased languages). We introduce simple effective language-independent approaches for improving NER on microblogs, based on using large gazetteers, domain adaptation, and a two-pass semi-supervised method. We use Arabic as an example language to compare the relative effectiveness of the approaches and when best to use them. We also present a new dataset for the task. Results of combining the proposed approaches show an improvement of 35.3 F-measure points over a baseline system trained on news data and an improvement of 19.9 F-measure points over the same system but trained on microblog data.

Keywords: Named Entity Recognition, Microblogs, Arabic

1. Introduction

Named Entity Recognition (NER) is essential for a variety of Natural Language Processing (NLP) and Information Retrieval (IR) applications such as information extraction. Though there has been a fair amount of work on NER, primarily for the news domain, little work has been done on microblogs (tweets) NER. NER on microblogs faces many challenges such as: (1) Microblogs are often characterized by informality of language, ubiquity of spelling mistakes, and the presence of Twitter name mentions (ex. @someone), hashtags, and URL's; (2) NE's are often abbreviated. For example, tweeps (tweet authors) may write "Real Madrid" as just "the Real"; (3) Tweeps often use brief and choppy expressions and incomplete sentences; (4) Word senses in tweets may differ than word senses in news. For example, "mary jane" in tweets likely refers to Marijuana as opposed to a person's name; (5) Tweeps may inconsistently use capitalization (for English), where capitalized words may not capitalized and ALL CAP words are used for emphasis; (6) We observed that NE's often appear in the beginning or the end of tweets and they are often abbreviated.
As for Arabic tweets, they exhibit more complications, namely:

- Tweets may contain transliterated words (ex. "LOL" → لول) and non-Arabic words, particularly hashtags (ex. #syria)

- Arabic lacks a capitalization feature

- The percentage of NE's that exist in our tweet test set that were seen in the training set was only 25%

- Tweets frequently use dialects, which may lack spelling standards (ex. معرفتش and ماعرفتش are varying spellings of "I did not know"), introduce a variety of new words (ex. محد means "no one"), or make different lexical choices for concepts (ex. باهي and صافي mean "good"). Dialects introduce morphological variations with different prefixes and suffixes. For example, Egyptian and Levantine tend to insert the letter ب before verbs in present tense.

In this paper, we employ simple effective methods for microblog NER, including building a large gazetteer, domain adaptation, and a two-pass semi-supervised method. We also present a new dataset. Our contributions are:

- We describe the peculiarities of tweets and how they affect microblog NER

- We introduce a new dataset for Arabic microblog NER with training and test tweets from different time periods. We plan to release it publicly

- We propose a novel semi-supervised two-pass approach that uses automatically tagged NE's as a gazetteer

- We compare the relative effectiveness of different methods for improving microblog NER.

The remainder of the paper is organized as follows: Section 2 surveys previous work on Arabic NER and NER on tweets; Section 3 describes our baseline system and discusses its shortcomings; Section 4 introduces the techniques that we employed to improve Arabic NER on tweets including building a large gazetteer, domain adaptation, and semi-supervised training; and Section 5 concludes the paper.

2. Background

Nadeau and Sekine (2009) surveyed NER research for a variety of languages and explored a variety of features, such as: (a) Contextual features, which are typically captured by the sequence labeling algorithms; (b) Character-level features that include the first or last few letters of words. For example, a word ending in "berg" is often a NE. Differences in morphology between formal Arabic and dialects lead to the attachment of different affixes; (c) Part-of-speech (POS) and morphological features: Linguistic tools for tweets, particularly for Arabic, lag behind tools for the news domain; and (d) Gazetteers: We extract a large gazetteer from Wikipedia category names and redirects. Due to the differences between tweets and new texts, in-domain training data would be required to capture contextual and character-level features. There has been limited work pertaining to NER on microblogs and hardly any for Arabic. The central problem most previous work attempted to solve is the lack of gazetteers or reliable NE candidates (Habib and Keulen, 2012; Li et al., 2012; Jung, 2012). We show in this work that building gazetteers is not the most effective approach. Li et al. (2012) developed an unsupervised NER system that exploits collocations, which may be NE's. Collocations were ranked using a random graph walk. Liu et al. (2011) used a k-nearest neighbor classifier to assign initial labels to words based on their contexts, and then they exposed the tags to a CRF labeler. Jung (2012) combined tweets in a single thread to add more context to facilitate microblog NER. Habib and Keulen (2012) used simple matching against a large knowledge base and then a disambiguation module that leveraged relationships in a knowledge base. Ritter et al. (2011) opted to retrain NLP tools specifically for tweets, including a POS tagger, a shallow parser, and a capitalization recovery classifier. They also used entities in Freebase[1], which is not available in Arabic. Liao and Veeramachaneni (2009) introduced a semi-supervised self-training, where automatically labeled segments were added to the training set if the labeler was confident. Unlike their work, we used a semi-supervised two-pass method, in which the maximum likelihood ratio that a supervised sequence labeler would label a word (or phrase) as a NE is used as a feature. For Arabic, Darwish (Darwish, 2013) tested a NER system that was trained on news data on Arabic tweets. He reported results that were far lower than those for news. We used simple yet reportedly effective domain adaptation based on instance weighting approach (Daumé III, 2007).

Significant work has been conducted by Benajiba and colleagues on Arabic NER (Benajiba and Rosso, 2008; Benajiba et al., 2008; Benajiba et al., 2007). They used a gazetteer, a stopword list, current, previous, and next words, cross-language capitalization, POS tagging, base-phrase chunking, and adjectives indicating nationality. For evaluation, they created a dataset called ANERCORP. Other significant work was done by others (Abdul-Hamid and Darwish, 2010; Darwish, 2013; Farber et al., 2008; Huang, 2005; Mohit et al., 2012; Shaalan and Raza, 2007) used an HMM-based NE recognizer for Arabic. Farber et al. (2008) used morphological features. Shaalan and Raza (2007) reported on a rule-based system that uses hand-crafted grammars and regular expressions in conjunction with gazetteers on a non-standard dataset. Abdul-Hamid and Darwish (2010) used a simplified feature set that relied primarily on character level features, namely leading and trailing letters in words. They experimented with a variety of other phrase level features with limited success. We used their simplified features in our baseline system. Mohit et al. (2012) attempted to improve the recall of NER on Arabic Wikipedia using self-training and recall-oriented classification. More recent work by Darwish (2013) used cross-lingual features from English to improve Arabic NER. Most previous work on Arabic NER focused on news.

3. Baseline System

For our baseline system, we used the CRF++[2] implementation of CRF sequence labeling. We opted to reimplement the most successful features that were reported in (Abdul-Hamid and Darwish, 2010), namely the leading and trailing 1, 2, 3, and 4 letters in a word as well as the current, previous, and next words in their raw forms (without linguistic processing). We tokenized tweets in the manner reported in (Darwish et al., 2012). For training, we used two training sets. The first was the full ANERCORP dataset (Benajiba and Rosso, 2007), which has approximately 150k tokens. For the second, we tagged a new training set of 3,646 tweets, which were randomly selected from tweets that were authored in the period of May 3-12, 2012. The tweets that were scraped from Twitter using the query "lang:ar". For testing, we tagged a set of 1,423 tweets containing nearly 26k tokens. The tweets were randomly selected from the period of Nov. 23-27, 2011. We picked the tweets from a time window that was shifted from the training set to ascertain how well the models generalize. For both of the tweet sets, the tweets were annotated by an annotator and reviewed again by another to ensure correctness. Both annotators were native Arabic speakers. They followed the Linguistics Data Consortium guidelines for tagging. Table 1 (a) reports on the results for the baseline system. Table 1 (b) reports on the results of using the tweets training set. The results show nearly 17.9 F-measure points improvement over using the ANERCORP news training data. This shows the critical importance of in-domain training data for the task.

4. Improving NER for Tweets

To improve NER for tweets, we used a large gazetteer and applied domain adaptation and two-pass training. We opted to use language independent methods that require no linguistic processing.

4.1. Using a Wikipedia Gazetteer

We built a large gazetteer from Wikipedia that hopefully has broad coverage in a way that is similar to that of Ratinov and Roth (2009). We heuristically classified Arabic Wikipedia titles into persons, organizations, locations, or others if they belong Wikipedia categories that contain trigger words in Figure 1. We added the Wikipedia redirects to grow the gazetteer. The gazetteer had 70,908 locations, 26,391 organizations, and 81,880 persons and covered 60% of the NE's that appear in the tweet-test set. When a word sequence matched an entry in the gazetteer, the feature associated with the first word was "B-" plus the entity type (PERS, LOC, or ORG), and "I-" plus the entity type for subsequent words. Table 1 (c) and (d) show the results using the large Wikipedia gazetteer with news and tweets training sets respectively. Adding the gazetteer improved

[1] http://www.freebase.com/

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

	(a)			(b)		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
LOC	82.0	24.6	37.9	87.1	46.1	60.3
ORG	42.6	7.4	12.7	77.1	20.7	32.7
PERS	39.0	24.9	30.4	63.2	21.2	31.7
Overall	53.7	20.7	29.9	78.8	31.7	45.3

	(c)			(d)		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
LOC	81.7	28.1	41.8	82.7	54.3	65.5
ORG	51.6	10.7	17.7	79.1	28.2	41.5
PERS	51.4	35.5	42.0	73.5	36.1	48.5
Overall	61.9	26.6	37.2	79.3	42.1	55.0

Table 1: Results of using: (a) news training data; (b) tweets training data; (c) news training data with gazetteer; (d) tweets training data with gazetteer

PERS: مواليد births, وفيات deaths, أشخاص علي قيد الحياة living people
LOC: دول countries, عواصم capitals, بلدان countries, محافظات provinces, ولايات states, مدن cities, مطارات airports, صحاري deserts, أنهار rivers, بحار seas, بحيرات lakes, محيطات oceans, مواقع locations, قرى villages, خلجان bays, مراكز محافظة provincial centers, أحياء suburbs, تضاريس geo-features, حدايق parks, أقاليم rural areas
ORG: منظمات organizations, شركات companies, مؤسّسات foundations, هيّئات institutes, نقابات syndicates, أندية clubs, اتحادات unions, أحزاب parties, جماعات groups, جمعيات organizations, مجالس councils, بنوك banks, وزارات ministries, مصانع factories, جامعات universities, قنوات channels, اذاعات stations, وكالات subsidiaries, كنائس churches

Figure 1: Trigger words for detecting NE types in Wikipedia

NER effectiveness when using either the news or tweets training sets with gains of 7.3 and 9.7 F-measure points respectively. However, though a large gazetteer improved NER effectiveness, using in-domain training data seems to be more important. Also, the large gazetteer was unable to raise the effectiveness when training on news data to that of using in-domain data even without a gazetteer.

4.2. Domain Adaptation

We used simple domain adaptation that combines ANERCORP news and tweets training sets. We used an instance-based weighting scheme to give different weights to training examples that entails replicating the tweets dataset multiple times during training. This method was shown to be effective by Daumé III (2007), with nearly state-of-the-art results for a variety of tasks. To determine the optimal number of times to replicate the tweets data, we split the data into training and validation sets using an 80/20 split and replicated the tweet set n times ($1 \leq n \leq 9$). A replication factor of 7 was best. Table 2 shows the results of domain adaptation. Performing domain adaptation (Table 2 (a)) is comparable to using in-domain training with a large gazetteer (Table 1 (d)). Using a large gazetteer generally improved precision while using domain adaptation generally improved recall. Combining both together yielded a 4.5 F-measure points improvement over using in-domain data with gazetteer and 4.9 using adaptation alone.

	(a)			(b)		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
LOC	85.0	57.2	68.4	84.3	61.6	71.2
ORG	75.4	33.7	46.5	78.0	35.6	48.9
PERS	60.9	30.2	40.4	67.3	39.9	50.1
Overall	76.0	42.6	54.6	77.7	48.2	59.5

	(c)			(d)		
	P	R	F	P	R	F
LOC	82.9	43.1	56.7	84.7	63.2	72.4
ORG	67.8	33.3	44.7	75.5	37.6	50.2
PERS	60.2	30.6	40.6	72.4	38.0	49.8
Overall	71.8	36.6	48.5	79.3	48.6	60.3

Table 2: Results of: (a) adaptation; (b) adaptation with large gazetteer; (c) semi-supervised two-pass training; (d) semi-supervised two-pass training with large gazetteer

4.3. Semi-supervised Two-pass Training

A previously unseen NE (in training data or in gazetteer) may appear in a context that was previously seen in training leading to recognition. Otherwise, a previously unseen NE may not be recognized. Thus, our intuition was that if we automatically tag a large set of tweets, then a NE may be tagged correctly multiple times. Then, automatically identified NE's can then be used as a "new gazetteer." The process proceeded as follows:

```
Require: Initial NER system $R$ and untagged set $S$
Ensure: Final recognizer $R'$
  tag $S$ using $R$ to produce gazetteer $G$
  remove items in $G$ that are recognized only once
  set $w(g_j \in G) = \frac{Count(R\ tags\ g_j\ as\ an\ entity\ type)}{Count(g_j\ in\ corpus)}$
  round weights $w$ to first significant figure
  use $G$ as weighted gazetteer to train $R'$
```

For R, we experimented with two of our recognizers, namely the baseline recognizer that is trained on tweets, and another that uses the large Wikipedia gazetteer. For the untagged set S, we used 65M random tweets that cover the time periods of training and test tweets. After running the baseline recognizer on S, we obtained a gazetteer containing 4,464, 3,974, and 13,057 locations, organizations, and person names respectively. When using the recognizer that uses the Wikipedia gazetteer, we obtained a gazetteer containing 5,188, 3,176, and 11,721 locations, organizations, and person names respectively. They contained 74% of NE's in test set. Due to the low precision of our best system (more than a quarter of the recognized items would be erroneous), we weighted each item in G using the maximum likelihood estimate that R would tag it as a NE type. We binned weights because CRF++ uses nominal features only. For example, if the sequence جامعة المنصورة (meaning "Mansoura University") was recognized as an organization 83% of the time, the first and second words in the sequence were "B-ORG-.8" and "I-ORG-.8" respectively. Out of gazetteer words were assigned the value "null". The results in Table 2 (c) and (d) show that the approach improved NER by 3.2 F-measure points over using a baseline system without the Wikipedia gazetteer and 5.3 F-measure points over using a baseline system with the Wikipedia gazetteer respectively. The proposed semi-supervised approach is more effective when starting with a better initial recognizer.

4.4. Putting it All Together

Table 3 reports on the results of using the Wikipedia gazetteer (**Wikigaz**), domain adaptation (**Adapt**), and the semi-supervised two-pass approach (**2Pass**). The results show that NER effectiveness improved upon training using the ANERCORP news and tweets baselines by 35.3 and 19.9 F-measure points respectively. The improvements surpass using any of the three approaches in isolation and improves upon using the Wikipedia gazetteer in combination with domain adaptation by 5.7 F-measure points. Figure 2 compares the results of all the runs. Some observations are: (1) Using in-domain training data has the most effect. This is the most laborious yet essential approach; (2) Building a large gazetteer and using domain adaptation have a comparable effect. Building a gazetteer and incorporating out-of-domain training data are both simple and effective; (3) Using our semi-supervised approach is more potent when using a better initial recognizer (i.e., **Adapt+Wikigaz**); (4) Combining different approaches yields improved overall effectiveness.

	Combining all features		
	P (%)	R (%)	F (%)
LOC	83.6	70.8	76.7
ORG	76.4	43.7	55.6
PERS	67.1	47.8	55.8
Overall	76.8	56.6	65.2

Table 3: Results of combining the Wikipedia gazetteer, domain adaptation, and semi-supervised learning

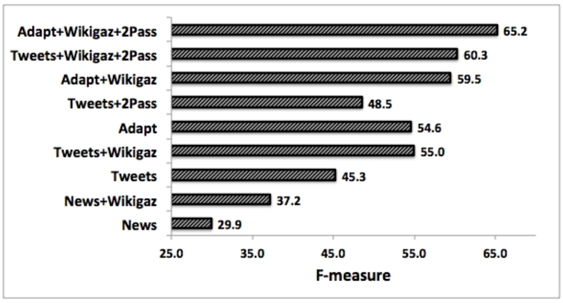

Figure 2: Results of All Runs

5. Conclusion

We presented simple, effective and language-independent approaches for improving NER in microblogs, with Arabic as an example. These approaches include using in-domain training data, building a large gazetteer from Wikipedia, domain adaption, and a semi-supervised two-pass approach. The approaches significantly improve Arabic microblog NER over our baseline. These approaches are language independent.

6. References

A. Abdul-Hamid and K. Darwish. 2010. Simplified Feature Set for Arabic Named Entity Recognition. In 2010 NEWS Workshop, ACL-2010, pages 110–115.

Y. Benajiba, M. Diab, and P. Rosso. 2008. Arabic Named Entity Recognition using Optimized Feature Sets. EMNLP-2008, pages 284–293.

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

Y. Benajiba, P. Rosso and J. M. Benedi Ruiz. 2007. ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. CICLing-2007, Springer-Verlag, LNCS(4394), pages 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. Workshop on Natural Language-Independent Engineering, IICAI 2007.

Kareem Darwish, Walid Magdy, Ahmed Mourad. 2012. Language Processing for Arabic Microblog Retrieval. CIKM-2012, pages 2427–2430

Kareem Darwish. 2013. Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. ACL-2013.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. ACL-2007, pages 256–263.

B. Farber, D. Freitag, N. Habash, and O. Rambow. 2008. Improving NER in Arabic Using a Morphological Tagger. LREC-2008.

Mena Habib, Maurice van Keulen. 2012. Unsupervised Improvement of Named Entity Extraction in Short Informal Context Using Disambiguation Clues. EKAW Workshop on Semantic Web and Information Extraction, pp. 1–10.

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

Jason J. Jung. 2012. Online Named Entity Recognition Method for Microtexts in Social Networking Services: A Case Study on Twitter. Journal of Expert Systems with Applications, 39(9):8066–8070.

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

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun and Bu-Sung Lee. 2012. TwiNER: Named Entity Recognition in Targeted Twitter Stream. SIGIR-2012, pages 721–730.

Wenhui Liao, Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In NAACL-HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pages 58–65.

Xiaohua Liu, Shaodian Zhang, Furu Wei, Ming Zhou. 2011. Recognizing Named Entities in Tweets. ACL-2011, pages 359–367.

A. McCallum and W. Li. 2003. Early Results for Named Entity Recognition with Conditional Random

Fields, Features Induction and Web-Enhanced Lexicons. CoNLL-2003, pages 188-191.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. EACL-2012, pages 162–173.

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

L. Ratinov and D. Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. CoNLL-2009, pages 147–155.

Alan Ritter, Sam Clark, Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In EMNLP-2011, pages 1524–1534.

K. Shaalan and H. Raza. 2007. Person Name Entity Recognition for Arabic. In 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources (Semitic 2007), pages 17–24.