Named Entity Recognition (NER) for Turkish

10 Nisan 2017

Named entity recognition is the task for detecting the names for people, locations, organizations; temporal or numeric expressions. It can be treated as a word or phrase-level text classification / sequence tagging task. For example, in the sentence “Achille-Claude Debussy, born in 1862 in St. Germain and known since the 1890s, was a talented French composer.”, ideally the named entities are Achille-Claude Debussy, 1862, St. Germain, 1890s and French.

It is one of the primary information extraction tasks and has many use cases in problems such as event detection, web search and translation. More pervasively, it is quite useful in text classification. In sentiment analysis, for example, it helps detecting the sentiment towards an entity. In message classification and generation, it is useful for capturing what / whom the message is talking about and for generating a meaningful answer. Generally, feature extraction mechanisms can make use of named entities; for example, the number of named entities in a text might indicate the formality level of it and can contribute to the overall performance when made use of.

We focus on Turkish NER in this report since for English a number tools are available such as [15] and [18] and the approach suggested here is easily applicable for other languages.

How it is done

Basically, there are three methods for detecting named entities: 1) rule-based 2) statistical and recently 3) neural.

- In the rule-based methods [9], a huge list of named entities such as person names, organizations, countries and so on is prepared and the input text is searched for those entities with the help of some text processing methods like regular expressions or more sophisticated tools like POS taggers (to search for only nouns, for example).
- In the statistical & neural methods, a model is built by determining features / using word vectors for how to represent the texts for named entities; a set of texts where the named entities are annotated is used as the training set and the words in it are input to a classifier as windows (by its neighbour words); the classifier learns the model with the

examples in the training set, then this learned model can be used as named entity recognizer ready to accept new texts whose entities are unknown.

Requirements

We went through a subset of studies on NER (details follow below), largely on Turkish NER. It appears that methods employing (neural) word embeddings (vector representations of words using neural networks) together with a sequence tagger or a neural network give very good results and are open to much more improvements compared to classical methods. Indeed, such an approach is applicable to other languages without difficulty.

Roughly, a method combined from those of [5], [13] and [16] should satisfy the needs and not hard to realize. The system will learn named entities from a set of texts annotated for named entities; using the embedding vectors of those words. The classifier will be either a CRF (Conditional Random Field) or a neural network, both available within open source packages.

1. Turkish word embeddings
    a. A set of embeddings is available from [16], produced out of unannotated news texts and tweets.
    b. Depending on our preprocessing requirements and possible use of these vectors in other tasks (like sentiment analysis), it is better we produce our own embeddings. For this we need:
        i. Lots of texts from news, social media, or any other domain to work on: at least 500M words
        ii. To test the performance of the produced embeddings
2. Annotations
    a. We need to decide on
        i. Labels: person, location, organization only or include numeric and temporal expressions or even domain-specific names (if they are so frequent).
    b. We need a set of texts whose words / word groups are to be annotated for named entities (the label above).
        i. Most of the existing annotations are not open to commercial use. The publicly available ones:
            1. A set of ~1300 entities from tweets [10] (no information about the licence).
            2. Some 300K entities automatically extracted from Wikipedia [14].
        ii. For a remarkable performance we need around 10K named entities annotated, which corresponds to a dataset of ~1000 texts.
        iii. An annotation tool for word-level labelling is available: http://brat.nlplab.org/
3. Type of the texts

    a. Domains (like business, email, social media) of the target texts should be clarified as much as possible so that sources of words embeddings and training texts will be better chosen and target domain can be better sampled and thus the classifier can give better results.
4. Learning
    a. We will hold a number of experiments (each expected to take some hours) to determine / test
        i. The feature set: we may need to add extra features
        ii. The type of the classifier and its parameters
        iii. The performance / quality of the embeddings, annotations and the classifier.

Studies on Turkish NER

The earliest NER studies on Turkish texts are [1], which is based on many hand-crafted features and statistical learning methods; and [2], which employs a language independent machine learning approach.

In [1], a purely rule-based system is presented. The authors extract lists of 1) person names 2) well-known people 3) well-known locations 4) well-known organizations from a large corpus, choosing only nouns from the output of a morphological analyser. And they build pattern bases for 1) location name (X Sokağı) 2) organization names (Y Üniversitesi) 3) temporal and numeric expressions. The tags they consider are ENAMEX, NUMEX and TIMEX. For the detection of named entities in a new text, their system searches for the match of name and pattern bases. As for performance, the system gives an f-score of 78% on news texts and 55% on historical texts. The authors say that pattern matches for non-entity phrases are common and lead to poor performance. And in [4], this rule-based system is improved with some additional lexical resources.

Recent studies on Turkish NER make use of rather statistical and lately neural approaches. The systems mostly rely on the features mostly found in formal text while social media data or texts from other domains may not carry such features or might be very different stylistically [5] and to overcome such noise or to extend to other domains 1) normalization 2) domain adaptation is suggested [6, 8]. Currently, using word embeddings beside named entity features is very popular and said to be overcoming such problems.

In [11], the authors extract various linguistic features for NER and for learning them use Conditional Random Fields, which is one of the top algorithms for sequence tagging [12]. In addition to morphological, lexical and symbolic features extracted, a list of names for entities (gazetteers) is also used. Their training set contains annotated news and tweets. The f-score performance of this system is 91%. Such a system is quite dependent on linguistic tools and

language-specific features, which may not be suitable for the cases having unstructured or informal texts. Indeed, neural approaches give comparable / slightly higher results [13].

In [5], the neural network based approach of [17], which in both studies called “NLP from Scratch” approach, is applied for Turkish texts. First, the embedding vectors are produced, using the Skip-Gram approach [7], from a huge amount of texts (a huge news corpus combined with a dump of Turkish Wikipedia). As for preprocessing prior to embedding production, the urls and numbers are converted to their respective symbols; all the words are lowercased; they indicate that dismissing capitalization is proven to be more effective on social media data. In the learning phase, the authors build a window-based neural network which makes the assumption that the tag of the target word depends on the neighbouring words [17]. This system performs much better than the similar Turkish systems on the social media dataset but not on the news texts. Among the named entities that the system can detect, the authors say there are misspelled, asciified (ü->u) or extended words so text normalization phase seems not very critical once high quality embeddings are used.

Another study that uses embeddings is [16] where in addition to word vectors, some contextual linguistic and stylistic features are also used and for learning a window-based perceptron is applied. The domain is tweets for which NER is notably more difficult. The f-score performance of the system is 56% when trained on annotated news texts and 48% when trained only on annotated tweets.

Paper	Dataset	Learning approach / algorithm	Labels	Performance	Availability
[15]	Unannotated wikipedia	Embeddings & NN	Person, location, organization	None for not having annotated data	Yes
[5]	News, tweets, forum, speech	Embeddings & NN	Person, location, organization	83% for news, 57% for twitter (f-score)	No
[16]	News, tweets	Embeddings & perceptron	Person, location, organization	56% (f-score)	Yes (both data and executables)
[11]	News, tweets	CRFs	Person, location, organization, date, time,	91% f-score	No, only online for one-time use.

			money, percentage		

Table1 - Sketch of the most recent Turkish NERs

Resources

[1] Tür, Gökhan, Dilek Hakkani-Tür, and Kemal Oflazer. "A statistical information extraction system for Turkish." Natural Language Engineering 9.02 (2003): 181-210.

[2] Cucerzan, Silviu, and David Yarowsky. "Language independent named entity recognition combining morphological and contextual evidence." Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC. 1999.

[3] Küçük, Dilek and Adnan Yazıcı. "Named entity recognition experiments on Turkish texts." International Conference on Flexible Query Answering Systems. Springer Berlin Heidelberg, 2009.

[4] Küçük, Dilek, and Adnan Yazıcı. "A hybrid named entity recognizer for Turkish." Expert Systems with Applications 39.3 (2012): 2733-2742.

[5] Önal, Kezban Dilek and Pınar Karagoz, “Named Entity Recognition from Scratch on Social Media”, ECML-PKDD, MUSE Workshop, pp2-17, September 2015.

[6] Eisenstein, Jacob. "What to do about bad language on the internet." *HLT-NAACL*. 2013.

[7] Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, (2013).

[8] Augenstein, Isabelle, Leon Derczynski, and Kalina Bontcheva. "Generalisation in named entity recognition: A quantitative analysis." *arXiv preprint arXiv:1701.02877* (2017).

[9] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30.1 (2007): 3-26.

[10] Küçük, Dilek, Guillaume Jacquet, and Ralf Steinberger. "Named Entity Recognition on Turkish Tweets." *LREC*. 2014.

[11] Şeker, Gökhan Akın, and Gülşen Eryiğit. "Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content1." *Semantic Web* Preprint: 1-18.

[12] Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." *Foundations and Trends® in Machine Learning* 4.4 (2012): 267-373.

[13] Demir, Hakan, and Arzucan Özgür. "Improving named entity recognition for morphologically rich languages using word embeddings." *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, 2014.

[14] Sahin, H. Bahadir, et al. "Automatically Annotated Turkish Corpus for Named Entity Recognition and Text Categorization using Large-Scale Gazetteers." *arXiv preprint arXiv:1702.02363* (2017).

[15] Al-Rfou, Rami, et al. "Polyglot-NER: Massive multilingual named entity recognition." *Proceedings of the 2015 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2015.

[16] Okur, Eda, Hakan Demir and Arzucan Özgür, "Named Entity Recognition on Twitter for Turkish using Semi-supervised Learning with Word Embeddings", LREC. 2016.
[17] Collobert, Ronan, et al. "Natural language processing (almost) from scratch." *Journal of Machine Learning Research* 12.Aug (2011): 2493-2537.
[18] Lample, Guillaume, et al. "Neural architectures for named entity recognition." *arXiv preprint arXiv:1603.01360* (2016).