Named Entity Recognition on Noisy Data using Images and Text (1-page abstract)

Diego Esteves
SDA Research, University of Bonn, Germany
esteves@cs.uni-bonn.de

Motivation The importance of real-world knowledge (a.k.a. *common sense*) for NLP was first discussed as early as 1960 (Bar-Hillel, 1960). However, up to now a substantial fraction of problems involving NLP could only be fully resolved if a rich understanding of the world is available **Current Advances** Recently, deep neural network architectures have achieved one step further in NER on noisy data - successfully overcoming the dependency of *gazetteers* and encoded rules (Limsopatham and Collier, 2016) - but are still far from performing as good as in formal language domains[1]. For instance, SOTA NER have (AVG) F1-measure ranging from 0.30 to 0.50 depending on the number of classes and the dataset, which confirms the very challenging characteristic of the task in noisy data. In short, this occurs due to the lack of linguistic formalism. Findings of a recent and comprehensive qualitative study of this gap are presented by (Augenstein et al., 2017). **Embedding world-knowledge** The main insight underlying the proposed work[2] is that we can enrich NER models by adding global features extracted from images and text in a word level perspective. To this end, we train a set of computer vision classifiers to recognize a set of pre-defined objects (each set associated to each named entity) as well as a set of text classification classifiers to label documents. A sentence example of the produced features extracted by the computer vision module is shown in Figure 1. Results of the performance (5-fold) of a 3-classes CRF model in Ritter is shown in Figure 2 ([*] with features), confirming the potential of this approach. **Work in progress** The preliminary version of this framework uses TF-IDF based text classification and

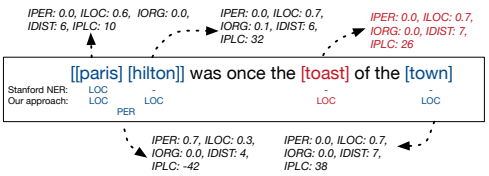

Figure 1: Example of extracted features.

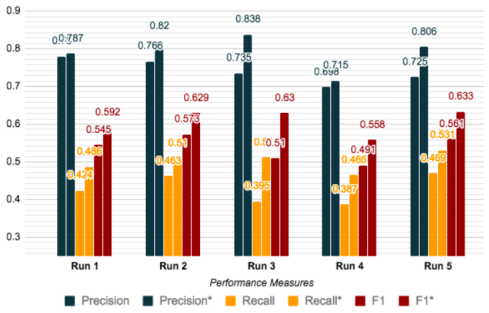

Figure 2: Performance measures in Ritter dataset (3-MUC).

SIFT features for computer vision. These components are being extended with more robust algorithms such as Topic Modeling and CNNs. Furthermore, due to its comprehensiveness, we are retraining the model using a recently released corpus: The Broad Twitter Corpus as well as extending the coverage of NE classes.

References

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*.

Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. In *Advances in computers*, volume 1, pages 91–163. Elsevier.

Nut Limsopatham and Nigel Collier. 2016. Bidirectional lstm for named entity recognition in twitter messages. *WNUT 2016*, page 145.

[1] NER over newswire often perform (F1) above 0.90 0.95.

[2] Esteves et al. 2017. *Named Entity Recognition in Twitter Using Images and Text*. Current Trends in Web Engineering - ICWE 2017 International Workshops.