GermEval-2014: Nested Named Entity Recognition with Neural Networks

Nils Reimers† **Judith Eckle-Kohler**†‡ **Carsten Schnober**†‡ **Jungi Kim**† **Iryna Gurevych**†‡
†Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
‡Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
http://www.ukp.tu-darmstadt.de

Abstract

Collobert et al. (2011) showed that deep neural network architectures achieve state-of-the-art performance in many fundamental NLP tasks, including Named Entity Recognition (NER). However, results were only reported for English. This paper reports on experiments for German Named Entity Recognition, using the data from the GermEval 2014 shared task on NER. Our system achieves an F_1-measure of 75.09% according to the official metric.

1 Introduction

Neural network architectures using low-dimensional vector representations of words (word embeddings) as the (almost) only features have been shown to achieve state-of-the-art performance in many fundamental NLP tasks, such as POS tagging, parsing and Named Entity Recognition (NER) (Collobert et al., 2011). Word embeddings are distributed word representations that are learned in an unsupervised fashion. A distinguishing feature of word embeddings is their ability to capture properties of words at various levels, in particular semantic and morphosyntactic regularities: words with similar embeddings are semantically (or morphosyntactically) similar, i.e. they are close to each other in the low-dimensional embedding space (Mikolov et al., 2013).

Most previous NER shared tasks annotated named entities flatly (e.g. CoNLL (Tjong Kim Sang and De Meulder, 2003)) and ignored entities that are nested within each other, e.g., the top-level named entity "*Real Madrid*", an organization containing the nested location "*Madrid*". In contrast, the GermEval 2014 NER dataset also contains annotations of nested named entities (Benikova et al., 2014b). Besides the four main classes PERson, LOCation, ORGanization and OTHer, it also introduces for each main class the subtypes *-deriv* for adjectives referring to named entities (e.g. *euklidisch - Euclidean*) and *-part* for words only partly containing names (e.g. *deutschlandweit - Germany-wide*). The dataset is divided into a training set consisting of 24,000 sentences, a development set of 2,200 sentences and a test set of 5,100 sentences.

2 Named Entity Recognition using Neural Networks

Collobert et al. (2011) propose a unified neural network architecture that can be applied to various natural language processing tasks. The presented deep neural network architecture uses only features based on minimal preprocessing: lowercased words, capitalization of the words, part-of-speech and a small gazetteer of known named entities. The input sentence is fed into the architecture and several layers of abstractions are learned.

The first layer is a *lookup operation* which maps each word and its associated features (POS etc.) to a d-dimensional vector. The second layer

makes the assumption that the named entity tag of a word can be predicted from its neighboring words. The vectors from the lookup operation for the target word and the neighboring words are concatenated and fed through an affine transformation followed by a non-linear activation function like the hyperbolic tangent function.

There are two different approaches for the last layer of the network, depending on whether the *isolated tag criterion* or the *sentence tag criterion* is used. For the *isolated tag criterion*, each word in the sentence is considered independently. The probabilities of the different tags for each word are computed by the softmax-function.

The *sentence tag criterion* optimizes the label sequence over the entire sentence. Tag probabilities from each window are concatenated and the dependencies between tags are factored into the model by learning initial probabilities and transition probabilities between tags. The Viterbi algorithm is used during inference. Collobert et al. (2011) use the more expressive IOBES-tagging scheme in their experiments. It uses an S-tag to mark single word named entities and B-, I- and E-tags to mark the first, the intermediate and the last word of a multi-word named entity.

We address nested named entities by training two independent neural networks. The first one detects top-level named entities and the second one detects nested named entities. The neural network for the nested named entities is trained only on top-level named entities that span over two or more words. At inference time, the top-level model is applied first, and its classification result is fed as an additional feature into the model for nested named entities.

3 Word-Embeddings

Word embeddings are a representation of words in a dense vector space (Bengio et al., 2003). They serve as the main feature for our models and can be learned from unannotated text data.

We used the following six corpora with a total of 116 million sentences to pre-train the word embeddings: German Wikipedia, the Leipzig Corpora Collection (Biemann et al., 2007), the SDeWac corpus (Faaß and Eckart, 2013), the print archive of *Spiegel*[1], the print archive of *ZEIT*[2], and the articles of *ZEIT Online*[3]. We used the Word2Vec tool presented by Mikolov et al. (2013) to compute the word embeddings from our training corpus.

Apart from tokenization, we performed the following pre-processing steps: Numbers are substituted by the special token `0`, diacritics are removed, except for German umlauts. All tokens are lowercased; the semantics of capitalization in the German orthography is captured by the capitalization feature (cf. section 4) instead.

Decompounding could significantly increase the performance of named entity recognition, especially for *-part* named entites. Our system uses only a naïve decompounding strategy for out-of-vocabulary words. In case a word cannot be found in the vocabulary, we split it along non-alphabetic characters (e.g. hyphens or slashes). We then replace the word by the first part which can be found in the vocabulary.

4 Additional Features

We designed several features which we assume to be helpful for the task of tagging named entities.

Capitalization: A feature to cover the information whether the word is all uppercase, the initial character is uppercase or if any succeeding character is uppercase.

Hyphen-Decompound: This feature splits words with a hyphen and adds the word embedding for the first part of the splitted word.

POS: The POS-tags as assigned by TreeTagger (Schmid, 1995).

Gazetteer: A feature to cover the information if the word appears in various gazetteers with named entities which can freely be found on the internet. Most notably the provided gazetteers from (Tjong Kim Sang and De Meulder, 2003) and a city and country list by GeoBytes[4]. Additionally, we compiled a gazetteer for person names and locations based on the corresponding Wikipedia categories. Our gazetteers contains around 311,000 person names, 90,000 locations,

[1] `http://www.spiegel.de/spiegel/print/`
[2] `http://www.zeit.de/2014/index`
[3] `http://www.zeit.de/index`
[4] `http://www.geobytes.com/freeservices.htm`

	Pr	Re	F_1
STC	78.5%	69.1%	73.5%
STC+Hyphen	79.8%	71.4%	75.4%
STC+POS	78.8%	71.2%	74.8%
STC+POS+Hyphen	80.1%	72.1%	75.9%
STC+Gazetteer	79.0%	71.2%	74.9%
STC+Wikipedia	78.8%	71.6%	75.0%
STC+All Features	80.4%	74.1%	77.1%

Table 1: Performance for the *sentence tag criterion* (STC) and different hand-crafted features. Scores are computed for the top-level named entities on the GermEval 2014 test set.

3,800 organizations and 3,600 other named entities.

Wikipedia-Definition: A feature that uses the German Wikipedia as an external knowledge base. In contrast to (Kazama and Torisawa, 2007), we used the Mate dependency parser[5] to process the first sentence and from all nouns that are positioned after the root verb, we selected the one with the shortest path to the root.

5 Evaluation

The GermEval 2014 shared task is evaluated using precision, recall and F_1-measure. We have a true positive if we have an exact match on the span and an exact match on the assigned label. The offical metric for the shared task (Benikova et al., 2014a) also takes the level for an assigned label into account. This leads to some counter-intuitive behavior. For example, for the nested named entity *[[Fraunhofer]$_{ORG}$ FIT]$_{ORG}$*, a model that does not return any named entity is scored better than a model that returns only the nested named entity *Fraunhofer*. The latter model would place the tag for *Fraunhofer* on the first level and thus it would be considered a misclassification, resulting in a lower precision for this model. We provide results for a level-independent evaluation in section 5.2.

5.1 Separate Evaluation of top- and nested-level

Optimizing globally the label sequence over the entire sentence for the top-level named entities has a major impact on the performance of our system. Using no other features than the word-embeddings and the capitalization of the word, our system achieves an F_1-measure of F_1=69.9% for the *isolated tag criterion* and F_1=73.5% for the *sentence tag criterion*. We experimented with the IOB2- as well as with the IOBES-tagging scheme, but the difference was below 0.1% in F_1-measure. The nested named entities were covered by training a second, independent neural network. Our networks use a window size of 5, a decreasing learning rate between 0.1 and 0.01 and 150 hidden units.

Top-Level NE				
	#	Pr	Re	F_1
PER	1639	89.0%	84.7%	86.8%
PERderiv	11	–	0%	0%
PERpart	44	35.3%	13.6%	19.7%
LOC	1706	84.8%	83.8%	84.3%
LOCderiv	561	81.1%	88.8%	84.8%
LOCpart	109	77.8%	38.5%	51.5%
ORG	1150	71.8%	68.8%	70.3%
ORGderiv	8	–	0%	0%
ORGpart	172	70.6%	55.8%	62.3%
OTH	697	61.6%	43.3%	50.8%
OTHderiv	39	82.6%	48.7%	61.3%
OTHpart	42	63.6%	16.7%	26.4%
Nested NE				
PER	82	44.8%	31.7%	37.1%
LOC	210	58.0%	51.9%	54.8%
LOCderiv	159	68.1%	48.4%	56.6%
ORG	41	42.9%	7.3%	12.5%

Table 2: Number of named entities (#), Recall (Re), Precision (Pr) and F_1-measure for the differend named entity classes. Scores are for the test dataset using all features. Our model found none of the nested named entities with the classes PERderiv (#4), PERpart (#4), LOCpart (#5), ORGderiv (#1), ORGpart (#1), OTH (#7) or OTHpart (#1).

Table 1 gives an overview of the impact of the different features. By using POS-tags and the Hyphen-feature, we can increase the F_1-measure for the top-level named entities by 2.4% to F_1=75.9%. Adding external knowledge resources increases the score further by 1.2% to F_1=77.1% for the top-level named entities.

We can observe a large difference in F_1-measure for the different named entity classes. While for PER, our model achieves an F_1-measure of around 87%, we only achieve an F_1-

[5] http://code.google.com/p/mate-tools/

measure of 51% for OTH. Analyzing the data shows that OTH-named entities are often especially hard, for example titles of books or songs, and appear much less coherent than other classes.

5.2 Level-Independent Evaluation

Combining the scores for the top-level and the nested-level, our model achieves an F_1-measure of 75.1%. However, as noted above, the separate evaluation of top- and nested-level leads to some counter-intuitive behavior. When neglecting the level and only validating the span and the correct label, the F_1-measure for the same model is F_1=78.0%. This shows that in several cases our model finds only the nested named entity and not the corresponding top-level named entity.

Neglecting the level also allows to use an approach that learns the short named entities first, followed by the longer ones. With the proposed level-dependent evaluation, such an approach would be evaluated much worse because several named entities would probably be placed on the wrong level and would be considered as a misclassification. We therefore argue that future named entities evaluations should be level-independent.

6 Conclusion

We adapted the approach of Collobert et al. (2011) to German using the GermEval 2014 dataset. Without external resources, we achieve an F_1-measure of 75.9% on the test set for the top-level named entities. Adding gazetteers and knowledge extracted from the German Wikipedia increases the performance to 77.1% for the top-level named entities. Combined with the performance for the nested-level, we achieve an overall F_1-measure of 75.1% in the offical metric. When neglecting the two levels, and solely evaluating the correct span and the correct label, the performance of our model is 78.0%.

Acknowledgement

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1110D (DARIAH-DE), and by the Leibniz Association as part of the SAW project "Children and Their World".

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014a. GermEval 2014 Named Entity Recognition: Companion Paper. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, Hildesheim, Germany.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014b. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig corpora collection - monolingual corpora of standard size. In *Proceedings of Corpus Linguistics*, Birmingham, UK.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC - A Corpus of Parsable Sentences from the Web. In *Language Processing and Knowledge in the Web*, pages 61–68. Springer.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, pages 142–147, Edmonton, Canada.