Training and Evaluating a German Named Entity Recognizer with Semantic Generalization

Manaal Faruqui
Dept. of Computer Science and Engineering
Indian Institute of Technology
Kharagpur, India 721302

Sebastian Padó
Maschinelle Sprachverarbeitung
Universität Stuttgart
70174 Stuttgart, Germany

Abstract

We present a freely available optimized Named Entity Recognizer (NER) for German. It alleviates the small size of available NER training corpora for German with distributional generalization features trained on large unlabelled corpora. We vary the size and source of the generalization corpus and find improvements of 6% F_1 score (in-domain) and 9% (out-of-domain) over simple supervised training.

1 Introduction

Named Entity Recognition is an important preprocessing step for many NLP tasks. It finds usage in applications like Textual Entailment, Question Answering, and Information Extraction. As is often the case for NLP tasks, most of the work has been done for English. To our knowledge, at this time there is no single "off-the-shelf" NER system for German freely available for academic purposes.

A major reason for this situation is the (un-) availability of labelled development data in the respective languages. For English, many large corpora annotated with named entities are available from a number of shared tasks and bakeoffs, including CoNLL 2003, MUC 2006/2007 and ACE 2008. For German, the only available dataset for NER seems to be the data from the CoNLL 2003 shared task on "Language-Independent Named Entity Recognition" (Tjong Kim Sang and De Meulder, 2003).

The German training part of the CoNLL 2003 data consists only of a total of 220,000 tokens. This is fairly small, but there must be a language-specific aspect at play as well: Even though the amount of training data for English is roughly comparable, the recall of the best system on English data, at 89%, is 25% higher than when trained on German data with 64% (Florian et al., 2003). We hypothesize that this difference is primarily due to the higher morphological complexity of German. Generally, this puts a higher strain on the lemmatization, and where lemmatization fails, tokens in the test set may simply be unknown. Also, morphological features, which can be learned from comparatively little data, are presumably less predictive for German than they are for English. For example, capitalization is a good predictor of NERs in English, where common nouns are not capitalized. In German, on the other hand, all nouns are capitalized, but most of them are not NEs.

While feature engineering for German is clearly one way out of this situation, the scarcity of labelled data remains a problem since it can lead to overfitting. In this paper, we therefore investigate an alternative strategy, namely *semantic generalization*. We acquire semantic similarities from large, unlabelled corpora that can support the generalization of predictions to new, unseen words in the test set while avoiding overfitting. Our contribution is primarily in evaluation and system building. We train the Stanford NER system (Finkel and Manning, 2009) on different German generalization corpora. We evaluate on both in-domain and out-of-domain data, assessing the impact of generalization corpus size and quality. We make the system with optimal parameters freely available for academic purposes. It is, to our knowledge, among the best available German NERs.

2 Named Entity Recognition with Semantic Generalization

We use Stanford's Named Entity Recognition system[1] which uses a linear-chain Conditional Random Field to predict the most likely sequence of NE labels (Finkel and Manning, 2009). It uses a variety of features, including the word, lemma, and POS tag of the current word and its context, n-gram features, and "word shape" (capitalization, numbers, etc.).

Importantly, the system supports the inclusion of distributional similarity features that are trained on an unrelated large corpus. These features measure

[1] `http://nlp.stanford.edu/software/`

how similar a token is to another in terms of its occurrences in the document and can help in classifying previously unseen words, under the assumption that strong semantic similarity corresponds to the same named entity classification. Specifically, the Stanford NER system is designed to work with the clustering scheme proposed by Clark (2003) which combines standard distributional similarity with morphological similarity to cover infrequent words for which distributional information alone is unreliable.[2] As is generally the case with clustering approaches, the number of clusters is a free parameter. The time complexity of the clustering is linear in the corpus size, but quadratic in the number of clusters.

To illustrate the benefit, imagine that the word "Deutschland" is tagged as location in the training set, and that the test set contains the previously unseen words "Ostdeutschland" and "Westdeutschland". During clustering, we expect that "Ostdeutschland" and "Westdeutschland" are distributed similarly to "Deutschland", or are at least morphologically very similar, and will therefore end up in the same cluster. In consequence, these two words will be treated as similar terms to "Deutschland" and therefore also tagged as LOC.

3 Datasets

German corpus with NER annotation. To our knowledge, the only large German corpus with NER annotation was created for the shared task "Language-Independent Named Entity Recognition" at CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003). The German data is a collection of articles from the Frankfurter Rundschau newspaper annotated with four entity types: *person* (PER), *location* (LOC), *organisation* (ORG) & *miscellaneous* (MISC). MISC includes, for example, NE-derived adjectives, events, and nationalities.[3] The data is divided into a training set, a development set, and a test set. The training set contains 553 documents and approximately 220,000 tokens. The development set (TestA) and test set (TestB) comprise 155 and 201 documents, respectively, with 55,000 tokens each.

Large unlabelled German corpora. For the semantic generalization step, we contrast two corpora that are representative of two widely available classes of corpora. The first corpus, the Huge German Corpus (HGC), consists of approximately 175M tokens of German newspaper text. The HGC is a relatively clean corpus and close in genre to the CoNLL data, which are also newswire. The second corpus is deWac (Baroni et al., 2009), a web-crawled corpus containing about 1.9M documents from 11,000 different domains totalling 1.71B tokens. deWac is very large, but may contain ungrammatical language, and is less similar to the CoNLL data than HGC.

4 Exp. 1: Testing on In-Domain Data

In this experiment, we replicate the CoNLL 2003 setup: We train the NER system on the training set, experiment with different generalization settings while evaluating on the the TestA development set, and validate the best models on the TestB test set. We tag and lemmatize both with TreeTagger (Schmid, 1994). We report precision, recall, and F_1 as provided by the CoNLL scorer.

Without any semantic generalization, on TestA we obtain a precision of 80.9%, a recall of 58.8%, and an F-Score of 68.1%. The poor recall corresponds to our expectations for the small size of the training set, and the experiences from CoNLL 2003. It also results in a low overall F_1 score.

For generalization, we apply Clark's (2003) clustering method to HGC and deWac. For each corpus, we vary two parameters: (a), the amount of generalization data; and (b), the number of clusters created. Following Clark (p.c.), we expect good performance for k clusters when $k^3 \approx n$ where n is the size of the generalization corpus. This leads us to consider at most 600 clusters, and between 10M and 175M tokens, which corresponds to the full size of the HGC and about 10% of deWac.[4]

Table 1 shows the results for using the HGC as generalization corpus. Already the use of 10M tokens for generalization leads to a drastic improvement in performance of around 5% in precision and 10% in recall. We attribute this to the fact that the semantic similarities allow better generalization to previously unknown words in the test set. This leads primarily to a reduction of recall errors, but to more robust regularities in the model, which improves precision. The beneficial effect of the generalization corpus increases from 10M tokens to 50M tokens, leading to a total improvement of 6-7% in precision

[2] Clark's system is available from `http://www.cs.rhul.ac.uk/home/alexc/pos2.tar.gz`

[3] See `http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt` for annotation guidelines.

[4] The deWac corpus supports the training of larger models. However, recall that the runtime is quadratic in the number of clusters, and the optimal number of clusters grows with the corpus size. This leads to long clustering times.

Tokens	Clusters	Precision	Recall	F_1
Baseline (0/0)		80.9	58.8	68.1
10M	100	85.2	68.1	75.7
10M	200	85.2	66.8	74.9
20M	100	83.0	64.9	72.9
20M	200	86.4	70.1	77.4
50M	200	86.7	69.3	77.0
50M	400	87.3	71.5	78.6
100M	200	85.4	69.4	76.6
100M	400	86.7	76.0	77.8
175M	200	86.2	71.3	78.0
175M	400	87.2	71.0	78.3
175M	600	**88.0**	**72.9**	**79.8**

Table 1: Performance on CoNLL German TestA development set, using HGC as generalization corpus

Tokens	Clusters	Precision	Recall	F_1
Baseline (0/0)		80.9	58.8	68.1
10M	100	83.5	65.5	73.4
10M	200	84.1	66.0	73.9
20M	100	84.2	66.2	74.1
20M	200	84.1	66.8	74.5
50M	200	85.4	68.9	76.3
50M	400	85.1	68.9	76.1
100M	200	84.9	68.6	75.9
100M	400	84.8	69.1	76.1
175M	200	85.0	69.4	76.4
175M	400	**86.0**	**70.0**	**77.2**
175M	600	85.4	69.3	76.5

Table 2: Performance on CoNLL German TestA development set, using deWac as generalization corpus

and 12-13% in recall, but levels off afterwards, indicating that no more information can be drawn from the HGC. For all but the smallest generalization corpus size, more clusters improve performance.

The situation is similar, but somewhat different, when we use the deWac corpus (Table 2). For 10M tokens, the improvement is considerably smaller, only 2.5% in precision and 6.5% in recall. However, the performance keeps improving when more data is added. At the size of the HGC (175M tokens), the performance is only about 1% worse in all statistics than for the HGC. As can be seen in Figure 1, the performances for HGC and deWac seem largely to converge. This is a promising result, given that we did not do any cleaning of deWac, since web corpora are cheaper than newswire corpora and can be obtained for a larger range of languages.

Finally, Table 3 validates the results for the best HGC and deWac models on the test set (TestB) and

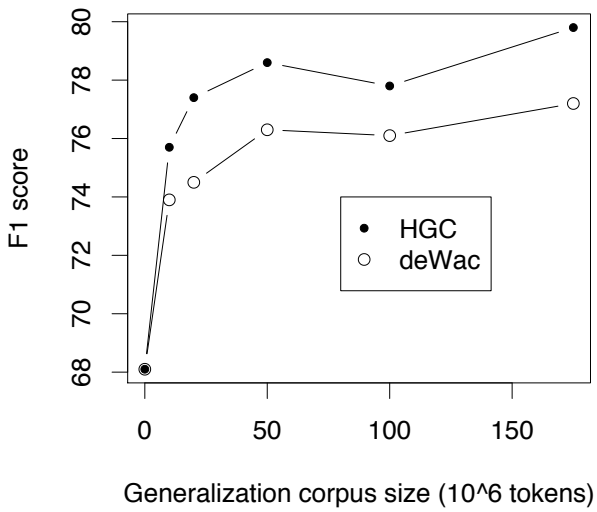

Figure 1: F_1 as function of generalization corpus

Model	Precision	Recall	F_1
Florian et al. (2003)	83.9	63.7	72.4
Baseline (0/0)	84.5	63.1	72.3
HGC (175M/600)	**86.6**	**71.2**	**78.2**
deWac (175M/400)	86.4	68.5	76.4

Table 3: Comparison to best CoNLL 2003 results for German on the CoNLL TestB test dataset

compares them to the best CoNLL 2003 shared task system for German (Florian et al., 2003). We see a small decrease of the performance of both systems by about 1% F-Score. Both models substantially outperform the baseline without generalization and Florian et al., a classifier combination system, by 4% and 5% F-Score, respectively. The improvement is mainly due to an 8% increase in recall.

5 Exp. 2: Testing on Out-of-Domain Data

This experiment assesses the performance of our CoNLL-trained German NER system on a different domain, namely the German part of the EUROPARL corpus (Koehn, 2005). EUROPARL consists of the Proceedings of the European Parliament, i.e., corrected transcriptions of spoken language, with frequent references to EU-related NEs. It thus differs from CoNLL both in genre and in domain. We annotated the first two EUROPARL files[5] with NEs according to the CoNLL guidelines, resulting in an out-of-domain test set of roughly 110,000 tokens.

Results. We tagged the test set with the baseline model and the best HGC and deWac models. The

[5] `ep-96-04-{15,16}`; tagging speed ≈2000 tokens/h.

Model	Precision	Recall	F_1
Baseline (0/0)	67.8	47.4	56.0
HGC (175M/600)	**78.0**	**56.7**	**65.6**
deWac (175M/400)	77.0	56.7	65.3

Table 4: Performance on EUROPARL

results are shown in Table 4. The performance of the baseline model without generalization is considerably worse than on the in-domain test set, with a loss of about 10% in both precision and recall. We see particularly bad recall for the MISC and ORG classes (34.4% and 46.0%, respectively), which are dominated by terms infrequent in newswire (nationalities and EU organizations and programs).

With semantic generalization, both recall and precision increase by roughly 10% for both HGC and deWac, indicating that corpus quality matters less in out-of-domain settings. We find a particularly marked improvement for the LOC category (deWac: P: 85.5% $\rightarrow$ 93.5%; R: 53.4% $\rightarrow$ 71.7%). We attribute this to the fact that location names are relatively easy to cluster distributionally and thus profit most from the semantic generalization step. Unfortunately, the same is not true for the names of EU organizations and programs. Even though the final performance of the models on EUROPARL is still around 10% worse than on the in-domain test data, the comparatively high precision suggests that the systems may already be usable for term extraction or in some semi-automatic setup.

6 Related Work

Rössler (2004) follows a similar motivation to ours by compiling resources with lexical knowledge from large unlabelled corpora. The approach is implemented and evaluated only for the PER(son) category. Volk and Clematide (2001) present a set of category-specific strategies for German NER that combine precompiled lists with corpus evidence. In contrast, Neumann and Piskorski (2002) describe a finite-state based approach to NER based on contextual cues and that forms a component in the robust SMES-SPPC German text processing system. Didakowski et al. (2007) present a weighted transducer-based approach which integrates LexikoNet, a German semantic noun classification with 60,000 entries.

Table 5 compares the performance of these systems on the only category that is available in all systems, namely PER(son). System performance is between 88% and 93% F-Score, with the best results for Didakowski et al. and our system. This comparison must however be taken with a grain of salt. Only our system and Rössler's are evaluated on the same data (CoNLL 2003), while the three other systems use their own gold standards. Still, our HGC model performs competitively with the best systems for German, in particular with respect to precision.

System	Data	Prec	Rec	F_1
HGC 175M/600	C	**96.2**	88.0	92.0
Rössler (2004)	C	89.4	88.4	88.9
Didakowski et al. (2007)	O	93.5	**92.8**	**93.1**
Volk and Clematide (2001)	O	92	86	88.8
Neumann and Piskorski (2002)	O	95.9	81.3	88.0

Table 5: Different German NER systems on category PER (C: CoNLL 2003 test set, O: own test set)

7 Conclusions

We have presented a study on training and evaluating a Named Entity Recognizer for German. Our NER system alleviates the absence of large training corpora for German by applying semantic generalizations learned from a large, unlabelled German corpus. Corpora from the same genre yield a significant improvement already when relatively small. We obtain the same effect with larger web-crawled corpora, despite the higher potential noise. Applied across domains, there is no practical difference between the two corpus types.

The semantic generalization approach we use is not limited to the four-class CoNLL setup. Even though its benefit is to decrease the entropy of the NE classes distribution by conditioning on clusters, and a higher number of NE classes could reduce the size of this effect, in practice the number of clusters is much higher than the number of NER classes. Therefore, this should not be an issue. Generalization can also be combined with any other models of NER that can integrate the class features. The extent to which other systems (like Florian et al., 2003) will improve from the features depends on the extent to which such information was previously absent from the model.

We hope that our results can be helpful to the German NLP community. Our two best classifiers (HGC 175M/600 and deWac 175M/400) as well as the EUROPARL test set are freely available for research at `http://www.nlpado.de/~sebastian/ner_german.html`.

Acknowledgements Many thanks to Jenny Rose Finkel and Alexander Clark for their support.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 59–66, Budapest, Hungary.

Jörg Didakowski, Alexander Geyken, and Thomas Hanneforth. 2007. Eigennamenerkennung zwischen morphologischer Analyse und Part-of-Speech Tagging: ein automatentheoriebasierter Ansatz. *Zeitschrift für Sprachwissenschaft*, 26(2):157–186.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the Conference on Natural Language Learning*, pages 168–171. Edmonton, AL.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X*, Phuket, Thailand.

Günter Neumann and Jakub Piskorksi. 2002. A shallow text processing core engine. *Journal of Computational Intelligence*, 18(3):451–476.

Marc Rössler. 2004. Corpus-based learning of lexical resources for German named entity recognition. In *Proceedings of the Language Resources and Evaluation Conference*, Lisbon, Portugal.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Natural Language Proceedings*, pages 44–49, Manchester, UK.

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 the Conference on Natural Language Learning*, pages 142–147, Edmonton, AL.

Martin Volk and Simon Clematide. 2001. Learn-filter-apply-forget. Mixed approaches to named entity recognition. In *Proceedings of the 6th International Workshop on Applications of Natural Language for Information Systems*, Madrid, Spain.