A Named Entity Labeler for German: exploiting Wikipedia and distributional clusters

Grzegorz Chrupała, Dietrich Klakow

Saarland University
Saarbrücken, Germany
gchrupala@lsv.uni-saarland.de, dklakow@lsv.uni-saarland.de

Abstract

Named Entity Recognition is a relatively well-understood NLP task, with many publicly available training resources and software for English. Other languages tend to be underserved in this area. For German, CoNLL-2003 provides training data, but there are no publicly available, ready-to-use tools. We fill this gap and develop a German NER system with state-of-the-art performance. In addition to CoNLL 2003 labeled training data, we use two additional resources: (i) 32 million words of unlabeled text and (ii) infobox labels in German Wikipedia articles. We extract informative features of word-types from those resources and train a supervised model on the labeled training data. This approach allows us to deal better with word-types unseen in the training data and achieve state-of-the-art performance on German with little engineering effort.

1. Introduction

Named entity recognition (NER) as an NLP task has received a fair amount of attention over the last decade (for an overview see Nadeau and Sekine (2009)). The task is to identify and label certain kinds of proper names, such as people, organizations or locations in natural language text. This is useful in many applications including Question Answering and Information Extraction.

The predominant approach to Named Entity Recognition is to train supervised models on annotated text. For English, there are a number of datasets that can be used for this purpose, including MUC (Grishman and Sundheim, 1996), CoNLL-2003 Shared Task (Tjong Kim Sang and De Meulder, 2003), BBN (Weischedel and Brunstein, 2005) and ACE (Doddington et al., 2004). There are also publicly available, ready-to-use tools for English NER trained on those datasets: Lingpipe [1] or Stanford CRF NER [2] (Finkel et al., 2005).

There are far fewer resources for German: CoNLL-2003 provides training data, but there are no publicly available, ready-to-use tools to perform German NER.

The contribution of this paper is to fill this gap and develop a NER system with state-of-the-art performance on German data. In order to achieve this goal, in addition to CoNLL 2003 labeled training data, we use two additional resources:

- Large amount (32 million words) of unlabeled text from (ECI, `http://www.elsnet.org/eci.html`)
- Infobox labels in German Wikipedia articles

We extract informative features of word-types from the unlabeled text and Wikipedia articles, use them to augment the basic feature representation, and then train in supervised fashion on the labeled training data. The two extra resources allow us to deal better with word-types unseen in the training data and substantially improve the performance of the system.

SEQUENCEPERCEPTRON($\mathbf{x}_{1:N}, \mathbf{y}_{1:N}, I$):

1: $\mathbf{w} \leftarrow \mathbf{0}$; $\mathbf{w}_a \leftarrow \mathbf{0}$
2: $c \leftarrow 1$
3: **for** $i = 1...I$ **do**
4: **for** $n = 1...N$ **do**
5: $\hat{\mathbf{y}}_n \leftarrow \text{argmax}_{\mathbf{y}}\ \mathbf{w} \cdot \Phi(\mathbf{x}_n, \mathbf{y})$
6: **if** $\hat{\mathbf{y}}_n \neq \mathbf{y}_n$ **then**
7: $\mathbf{w} \leftarrow \mathbf{w} + \Phi(\mathbf{x}_n, \mathbf{y}_n) - \Phi(\mathbf{x}_n, \hat{\mathbf{y}}_n)$
8: $\mathbf{w}_a \leftarrow \mathbf{w} + c(\Phi(\mathbf{x}_n, \mathbf{y}_n) - \Phi(\mathbf{x}_n, \hat{\mathbf{y}}_n))$
9: $c \leftarrow c + 1$
10: **return** $\mathbf{w} - \mathbf{w}_a/c$

Figure 1: Sequence perceptron with weight averaging

2. Model Structure and Features

The approach we propose is based on augmenting representation of training examples with features extracted from extra resources and learning model parameters on the labeled training data using a standard supervised learning method. We chose the perceptron algorithm for sequence labeling, which is straightforward to implement, and efficient to train and decode with, and provides excellent performance.

2.1. Supervised Learning Model

Figure 1 shows the version of the perceptron algorithm we use. Detailed discussion of the sequence perceptron can be found in (Collins, 2002). In brief, given the set of input sequences $\{\mathbf{x}\}_{1:N}$ and the set of corresponding output label sequences $\{\mathbf{y}\}_{1:N}$ the algorithm updates the weight vector $\mathbf{w}$ in a online fashion, for I iterations, whenever the current weights predict an incorrect output label sequence for the current training instance.

The key points for sequence labeling are the Φ feature function and the argmax computation in line 5. Following (Collins, 2002) our global feature function Φ decomposes

[1] `http://alias-i.com/lingpipe/`

[2] `http://nlp.stanford.edu/software/CRF-NER.shtml`

as the sum of the local features at each position i as

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_i \phi(x_i, y_i).$$

The local features extracted by ϕ are described in Section 2.2.

The argmax can be computed using the Viterbi algorithm, or using beam search. In either case global score is computed incrementally as the sum of partial scores at each position i: $\mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}) = \sum_i \mathbf{w} \cdot \phi(x_i, y_i)$. In our implementation we use beam search rather than Viterbi. This has the advantage of allowing the easy use of non-local features, such as output label at position -2.

The CoNLL-2003 Shared Task (Tjong Kim Sang and De Meulder, 2003) provides annotated data for learning to recognize named entities. Newspaper text is manually annotated with four entity labels: PER, LOC, ORG and MISC (people, geographical locations, organizations and miscellaneous entities not fitting the other labels, respectively), using the IOB encoding for sequence labeling.

2.2. Features

We extract features of the current token to be labeled (content features) as well as features of a 5-token window centered around the current position (context features):

- Content features
  - Word form
  - Lowercase word form
  - Lemma
  - Suffixes of length 1..3
  - Word shape: encodes which character classes (upper-case, lower-case, digits, or punctuation) the word contains and in what order
  - POS label
  - Chunk label
- Context features
  - Word form for tokens at positions $\{-2, -1, 1, 2\}$
  - NE label at position -1
  - Concatenated NE labels at positions -2 and -1

The local features are encoded as binary vectors $\phi(x_i, y_i)$. On top of these standard content and context features extracted from the labeled training examples we use additional resources to enrich the feature representation. We describe this approach in sections 2.2.1. and 2.2.2.

2.2.1. Wikipedia infobox features

Wikipedia is an online encyclopedia with extensive coverage of entities. Articles on many types of entities contain so called *infoboxes*, which are informally specified records with a label and a number of attributes. For example the English Wikipedia article for *Malta* has an infobox with the label `Country` and attributes such as `capital, area km2, population estimate`. Similarly, in the German language article, the infobox has the label `Staat` and corresponding attributes named in German.

Clearly, this is very useful information that can be easily exploited by a NER system. Specifically, for each German Wikipedia article we extract its title and the list of labels of the infoboxes appearing in it. We discard any material inside parenthesis in the article titles, as its mostly used for disambiguation. We then generate the following features for token in training and test sentences.

- Current word appears at position 1 of article title associated with infobox label X.
- Current word appears at position > 1 of article title associated with infobox label X.

Those features allow the model to learn associations between infobox labels and entity labels, and thus provide generalization for unseen word types.

2.2.2. Distributional clustering features

The second external resource which we use to enhance generalization performance of our system is unlabeled text.

We make use of the fact that many features of language can be usefully approximated by how linguistic units behave in naturally occurring text.

Consider a simple example: we can use large quantities of raw text to compute the frequencies of the left and right context of the words in our vocabulary. If a particular word's most frequent immediate left contexts include the prepositions *in* and *to*, this indicates that the word in question is likely to be a named entity of type location.

Specifically we use the word clustering algorithm proposed by Brown et al. (1992) to partition words into classes based on their co-occurrence statistics in a large corpus.

These classes capture the essence of a words syntactic and some aspects of its semantic behavior and can be used in class-based n-gram language models where they provide some generalization over raw word forms. A bigram class model has the following form:

$$P(w_k|w_{k-1}) = P(w_k|\pi(w_k))P(\pi(w_k)|\pi(w_{k-1})),$$

where π is the partition function assigning words to classes. Brown et al. (1992) show that the partition π which maximizes the likelihood of the model with respect to the training data, also maximizes the average mutual information (MI) of classes of adjacent words.

They propose the following a greedy algorithm: Initially each word in the vocabulary is assigned to a unique class and the average MI of classes of adjacent words is computed from the training corpus. Then the pair of classes which minimizes the loss in mutual information is merged into a new class. In a second step words are moved to a different class if this improves MI. When no class reassignment further increases the MI, the algorithm terminates.

Like Miller et al. (2004) we use word-type membership in Brown classes as additional features of our training and test examples.

Table 1 shows a few examples from among the 500 clusters trained on the German unlabeled corpus, with some of their frequent members. It is obvious that membership in

Cluster	Example members
1	Groß, Rau, Müller, Zimmermann, Frei, Becker, Möllemann, Schmidt
2	Düsseldorf, Berlin, München, Köln, Stuttgart, Hannover, Hamburg
3	nahmen, macht, zeigt, gleichen, bringt, biete, machte, sorgt, enthält

Table 1: Example distributional clusters

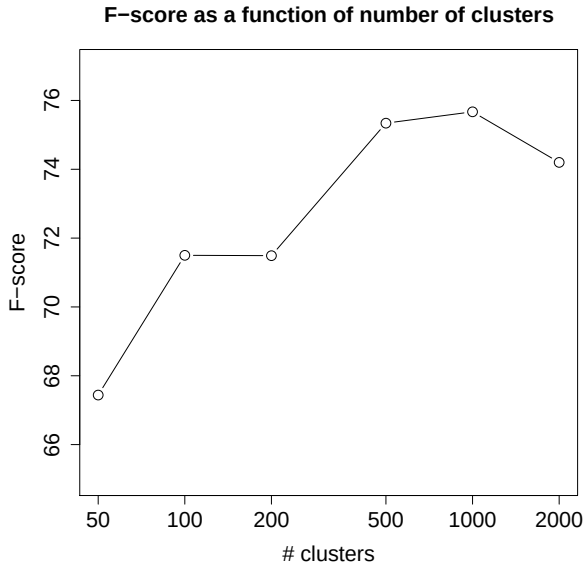

Figure 2: Varying number of clusters

clusters like 1 (person surnames) or 2 (cities) in this figure provide useful information for named entity recognition. Cluster 3 is also informative, grouping together many verb forms.

For both the Wikipedia and the cluster features, we create feature conjunctions between them and the following other features of the focus token: suffixes, word shape, POS tag, and chunk label. Experiments on the development set showed that these combinations improve accuracy without leading to an explosion in the number of features.

3. Evaluation

We evaluated our system on the German data from CoNLL 2003 Shared Task. We used all the available German text from ECI, i.e. approximately 32 million words, to train the distributional clusters. We experimented with different cluster granularities between 50 and 2000 clusters. Figure 2 shows the resulting performance: as can be seen it peaks for 1000 clusters. Based on this result, this is the cluster granularity we used in the subsequent experiments.

We processed the Wikipedia archive (downloaded on 28 October 2009) and found almost 218.000 articles which contained at least one infobox. We used all of them for infobox feature extraction.

We ran training for all the model variants for 20 iterations: this was sufficient for the performance curves to (nearly) converge. The beam size used for decoding in all the experiments was 9 – increasing it further did not appreciably

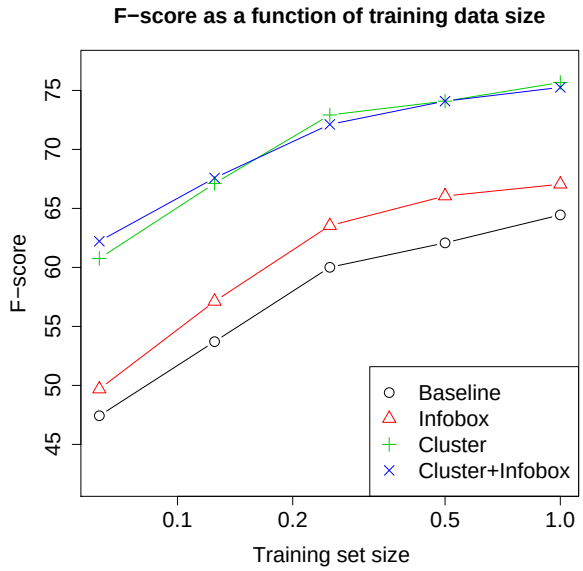

Figure 3: Performance of model versions with varying amounts of labeled training data

improve performance.

We tested four model versions:

- BASELINE: supervised model which only uses the features extracted from training data
- INFOBOX: Adds features extracted from Wikipedia articles
- CLUSTER: Adds cluster membership features
- INFOBOX+CLUSTER: Adds both infobox and cluster membership features

We scored NE-wise precision, recall and F-score as implemented in the CoNNL evaluation script.

Figure 3. shows the scores on the developments set of the four versions when trained on different amounts of training data.

For all sizes of labeled training data the models which use extra resources perform better than the baseline. For small training sets, the combination of infobox features and cluster features achieves the highest score. For larger sizes of the training set, using infobox features on top of distributional cluster features does not give any further improvements.

The issue seems to be that cluster features and infobox features contain similar type of information: they both provide generalization over unseen word types. For small amounts of training data, where relatively few features from each category fire, the overlap is minor, but with larger training sets, the cluster and infobox features will be increasingly redundant.

At all training data sizes, the cluster features on their gave a larger boost than the infobox features on their own. One reason for this performance gap may be the relatively low coverage of the German Wikipedia. It has around 218.000

Entity	Precision	Recall	F-score
All	82.19 (31.32)	71.72 (34.16)	76.60 (34.18)
LOC	76.85 (22.16)	79.85 (41.39)	78.32 (32.55)
MISC	77.06 (18.33)	55.54 (28.27)	64.56 (29.49)
ORG	81.20 (05.86)	61.97 (22.89)	70.29 (21.69)
PER	90.72 (63.46)	85.15 (51.85)	87.85 (56.95)

Table 2: NER results on German development set. Numbers in brackets show relative error reduction compared to the baseline

Entity	Precision	Recall	F-score
All	80.28 (18.21)	69.83 (22.46)	74.69 (21.67)
LOC	76.44 (12.71)	74.30 (23.56)	75.36 (19.13)
MISC	71.37 (11.61)	52.84 (17.28)	60.72 (17.22)
ORG	73.46 (03.39)	56.92 (11.67)	64.14 (10.24)
PER	91.59 (48.66)	83.85 (40.07)	87.55 (43.46)

Table 3: NER results on German test set. Numbers in brackets show relative error reduction compared to the baseline

article titles associated with an infobox, while the number of unique word forms in the ECI corpus is over 817.000. We also noticed that unlike in the English version, most German articles about people do not have an infobox. Since the English Wikipedia contains a large number of German entities, we are planning to try using features from both the German and English articles in future.

Combining cluster sets From the experiments with cluster granularities (cf. Figure 2) we saw that the two best performing cluster set sizes were 1000 and 500. We decided to see whether using those two cluster sets jointly would further improve the model. We tested several versions of this combination on development set. We obtained the best results using 500 clusters conjoined with content features, plus 1000 cluster without feature conjunctions. Thus cluster sets of different granularities do contain some complementary information.

Tables 2 and 3 show the performance of this best model on the development set and test set respectively. We report scores on all entity types, and broken down by entity type. The enriched model achieves noticeable error reduction rates for all entity types. The improvements are especially evident for person names

4. Related work

Our German NER system draws on ideas from several lines of previous research. We use the sequence perceptron learning algorithm introduced by (Collins, 2002). We adopt the idea of using distributional clustering in order to exploit unlabeled data in a semi-supervised scenario first proposed by (Miller et al., 2004), and more recently applied to dependency parsing by (Koo et al., 2008).

Wikipedia is routinely used for a wide range of NLP tasks. Wikipedia-derived resources have been used for NER in a number of papers. Kazama and Torisawa (2007) retrieve the English Wikipedia article for a candidate entity and extract the first NP after the verb *be* from the first sentence of the article, and use as a feature in a CRF NE labeler trained on CoNLL data. Mika et al. (2008) describe a method to automatically generate a domain-specific gazeteer from a seed list by exploiting the link structure of the English Wikipedia and evaluate on a corpus of archeology text. Zhang and Iria (2009). Wikipedia infoboxes are also being increasingly used in order to populate databases using heuristics (Suchanek et al., 2007), or as a resource to train information extraction models (Wu and Weld, 2007; Mika et al., 2008).

Our main contribution is taking some of these ideas and using them to develop a ready to use system for German with state-of-the-art performance. To our knowledge there is no publicly available data-driven NER system trained on German.

The open source Stanford CRF NER comes with models only for English. Evaluation scores for English on CoNLL 2003 data for this system are close to state-of-the-art (87.94% F-score). The Stanford CMM model which has been trained on German CoNLL data, with a reported test set F-score of 70.59%, does not seem to be available.

The Sprout platform (Drozdzynski et al., 2004) for finite-state grammar development has been used to develop NER and information extraction systems for multiple languages, including German. Our effort is complementary to the Sprout approach. Both data-driven and hand-crafted resources have their uses and we offer our semi-supervised system to fill the gap on the data-driven side.

5. Conclusion

We have created a German named entity labeler which will be released as a publicly available resource. Even though this initial version already has state-of-the-art performance on the CoNLL data, we would like to investigate how well it does on other domains and make sure that it is useful on a wide range of text types by employing domain adaptation methods.

We are also planning to further improve extracting informative features from Wikipedia, beyond infobox features.

Acknowledgments

Grzegorz Chrupała was funded by the BMBF project NL-Search under contract number 01IS08020B.

6. References

P. F. Brown, R. L. Mercer, V. J. Della Pietra, and J. C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

M. Collins. 2002. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of EMNLP 2002*.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of LREC*, volume 4, pages 837–840. Citeseer.

W. Drozdzynski, H. Krieger, J. Piskorski, U. Schäfer, and F. Xu. 2004. Shallow processing with unification and typed feature structures — foundations and applications. *Künstliche Intelligenz*, 1:17–23.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

R. Grishman and B. Sundheim. 1996. Message Understanding Conference-6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 466–471.

J. Kazama and K. Torisawa. 2007. Exploiting Wikipedia as external knowledge for Named Entity Recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707.

T. Koo, X. Carreras, and M. Collins. 2008. Simple semi-supervised dependency parsing. *Proc. ACL/HLT*.

P. Mika, M. Ciaramita, H. Zaragoza, and J. Atserias. 2008. Learning to tag and tagging to learn: A case study on wikipedia. *IEEE Intelligent Systems*, 23(5):26–33.

S. Miller, J. Guinness, and A. Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT-NAACL*, volume 4.

D. Nadeau and S. Sekine. 2009. A survey of named entity recognition and classification. In *Named entities: recognition, classification and use*, page 3. John Benjamins Pub Co.

F.M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, page 706. ACM.

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.

R. Weischedel and A. Brunstein. 2005. BBN pronoun coreference and entity type corpus. Linguistic Data Consortium, Philadelphia.

F. Wu and D.S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM.

Ziqi Zhang and José Iria. 2009. A novel approach to automatic gazetteer generation using wikipedia. In *Proceedings of the 2009 Workshop on the Peoples Web Meets NLP, ACL-IJCNLP 2009*.