DRIM: Named Entity Recognition for German using Support Vector Machines

Roman Capsamun, Daria Palchik, Iryna Gontar, Marina Sedinkina, Desislava Zhekova
CIS, Ludwig Maximilian University, Munich
{R.Capsamun,D.Brykova,Iryna.Gontar,Marina.Sedinkina,D.Zhekova}
@campus.lmu.de

Abstract

This paper[1] describes the DRIM Named Entity Recognizer (DRIM), developed for the GermEval 2014 Named Entity (NE) Recognition Shared Task.[2] The shared task did not pose any restrictions regarding the type of named entity recognition (NER) system submissions and usage of external data, which still resulted in a very challenging task. We employ Linear Support Vector Classification (Linear SVC) in the implementation of SckiKit,[3] with variety of features, gazetteers and further contextual information of the target words. As there is only one level of embedding in the dataset, two separate classifiers are trained for the outer and inner spans. The system was developed and tested on the dataset provided by the GermEval 2014 NER Shared Task. The overall strict (fine-grained) score is 70.94% on the development set, and 69.33% on the final test set which is quite promising for the German language.

1 Introduction

Named Entity Recognition aims to detect and classify nominal phrases into predefined categories such as organization, person, location and other. So far, mostly flat NEs were the target of identification (Benikova et al., 2014), which has been changed for GermEval 2014. This task is very important for many NLP challenges, such as information retrieval, speech processing, data mining, question answering, automatic summarization etc.

Most of the research in this field has been carried out for English with systems achieving considerably high levels of recall (97%) and precision (95%) (Mikheev et al., 1998; Stevenson and Gaizauskas, 2000). Though those results are substantial, the situation for other languages, especially for German, seems to be different.

Rules that are applied to English are not always useful for German. For example, in German not only NEs, but all the nouns are capitalized. In distinction to English, German adjectives such as "deutsch" are not to be capitalized. In comparison to English, German has higher morphological complexity, most productive type of which are compounds that are not found in a dictionary, for example, *AXA-Kunde, ADAC-Mitglied, Victoria-Agentur*. Except compounds, there are also derivations containing NEs, for instance, *die Deuscthen, die Bremer Staatsanwaltschaft*. The GermEval 2014 Shared Task sets as a goal the identification of both levels. A big obstacle is that existing training datasets for German are hindered by license problems. Also, there are not many open source NER taggers for German that perform at high levels of accuracy.

Because of these facts, proper identification and classification of NEs in German are very crucial and set a big challenge to the NLP research.

In Section 2, we describe related NER research.

[2]https://sites.google.com/site/germeval2014ner

[3]http://scikit-learn.org/stable

In Section 3, the data sets and the tagset provided by the GermEval 2014 NER Shared Task are presented, while in Section 4, we give an overview of Linear SVC. Following, we focus on the features that were used (see Section 5). Finally, we present our results on the development set provided by GermEval 2014 in Section 6, and in Section 7 we summarize our work and give suggestions for future directions.

2 Related Work

Since the Sixth Message Understanding Conference (MUC-6)[4], NER has become a well-established task of information extraction systems. MUC was initiated and financed by the Defense Advanced Research Projects Agency[5] to encourage the development of new and better methods of information extraction. Such competitions aimed at establishing frameworks for the proper and objective evaluation of various systems performing the same task (providing datasets and scoring possibilities).

For NER different approaches have been developed so far. There is a freely available Java implementation of a Named Entity Recognizer for English, namely Stanford NER.[6] As for the other languages, in particular German, one of the most significant works were presented by Faruqui and Padó (2010). Their German NER tagger has been trained on the CoNLL 2003 Shared Task[7] (Tjong Kim Sang and De Meulder, 2003) train set and uses semantic generalization information from two large German corpora, namely the HGC (Stuttgart University Newspaper Corpus) and deWac (the .de top-level domain "web as corpus"). Since 2010, this system is among the best NER systems for German with precision of 88.0% and recall of 72.9% (Faruqui and Padó, 2010).

There are also other machine learning systems for German NER. For example, Rössler (2004) similar to Faruqui and Padó (2010) uses resources with lexical knowledge from untagged corpora, reaching 78% recall and 71% precision (Rössler, 2004).

Rule-based approaches are also used for NER. The manually created rule-based system elaborates a set of patterns to accurately recognize and tag NEs (Volk and Clematide, 2001). They have reached 86%(recall) and 92%(precision). Another well-known rule-based system is Syntactic Constraint Parser (SynCoP), that is based on TAGH-morphology and gazetteers (Geyken and Schrader, 2006). Using the largest annotated corpus in the molecular biology domain, namely GENIA, the NER from Shen et al., (2003) trained a Hidden Markov model over the inner named entities, and then used a rule-based approach to identify the named entities containing the inner entities (Shen et al., 2003).

In our work, we implement a machine-learning approach with two separate linear SVM classifiers which are trained for the outer and nested spans of the NEs present in the GermEval 2014 dataset.

3 Named Entity Data and Tagset

The GermEval 2014 NER Shared Task provides a new dataset. This data was sampled from the German Wikipedia and News Corpora as a collection of citations. The dataset covers over 31,000 sentences corresponding to over 590,000 tokens. It is publicly available for download[8] under the permissive CC-BY license. The data has been annotated by two native speakers according to the semantic-based guidelines (Benikova et al., 2014). The entities from the dataset are to be classified in four main categories (*PER* – person; *ORG* – organization; *LOC* – location; *OTH* – other) with three subclasses (*main*, a NE comprises the full span; *part*, a NE takes only part of the span and *deriv*, the span is a derivation of a NE).

As for the format, each sentence is encoded as one token per line, with information provided in tab-separated (TSV) columns. The first column contains the token number within the sentence. The second column is the token itself. Name spans are encoded in the BIO-scheme (begin-

[4] http://cs.nyu.edu/cs/faculty/grishman/muc6.html

[5] http://www.darpa.mil

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

[7] http://www.cnts.ua.ac.be/conll2003/ner/

[8] https://sites.google.com/site/germeval2014ner/data

inside-outside). An example of the data format used in this shared task can be seen in Table 1.

TokenId	Token	Outer	Inner
21	Troia	B-OTH	B-LOC
22	-	I-OTH	O
23	Traum	I-OTH	O
24	und	I-OTH	O
25	Wirklichkeit	I-OTH	O

Table 1: Example of the data format.

4 Linear Support Vector Machine

Support Vector Machines (SVMs) are set of supervised learning methods used for classification, regression and solving various pattern recognition problems. This state-of-the-art classification method was introduced in 1992 by Boser, Guyon and Vapnik (Boser et al., 1992). Even though it is a relatively new machine learning approach, SVMs are well known for their good generalization performance and efficiency in high dimensional spaces (Kudo and Matsumoto, 2001). In the field of NLP, SVMs are reported to have achieved high accuracy in text categorization without falling into over-fitting because of a large number of words taken as a feature (Kudo and Matsumoto, 2000). Linear SVC has also been used in DRIM. The model assumes that the data is linearly separable. Linear SVC implements "one-vs-the-rest" multi-class strategy, thus training class models.

5 Feature Description

The most significant role in Support Vector Machines (SVM) plays feature selection (Ekbal and Bandyopadhyay, 2008). As there is one level of embedded NEs, two different classifiers were trained for each layer of embedding (further called outer and inner span).

5.1 Outer Span

5.1.1 Morphological Features

This class of features includes the most informative characteristics such as the token itself, Part of Speech (POS) information, lemma, token suffix, prefix and root. Morphological features are very basic but at the same time significant features which we take as a baseline.

POS information and lemmas are obtained via the TreeTagger (Schmid, 1994; Schmid, 1995), developed by Helmut Schmid.[9] TreeTagger makes use of a decision tree to get more reliable estimates for contextual parameters. This method has resulted in a higher accuracy than a standard trigram tagger (Schmid, 1994).

Token suffix, prefix and root are also informative features for NER. Considering the variety of German morphological entities we use a fixed length (four characters) of token suffix or prefix in a respective suffix/prefix feature. This length is very useful in detecting German suffixes, like *-land, -burg*, English suffixes like *-town, -city* or Russian suffixes like *-grad*.

5.1.2 Word Context Features

Morphological information (POS and lemma) of three previous and one following words of the target word are used as features. The NE annotations of three previous tokens concatenated in a string is also considered as a feature of the Word Context Class. This feature has been seen as a dynamic one in the experiment. That means it depends on the previous decisions of the classifier. Another new informative 'in bracket'-feature looks whether the current token is in apostrophes.

5.1.3 Encoded Context (Word-Shapes)

These features carry information about the local context. The current token and its immediate context are encoded according to their orthographic pattern, which is derived equally for all tokens. In such a way, distinctive types of entities can be better detected, like web and email addresses (e.g. www.cip.ifi.lmu.com → xxx.xxx, email@gmx.de → xxx@x.xx), companies (e.g. GmbH → XxxX) and other organizations or proper names (e.g. EUROPARLAMENT → XXXXX).

5.1.4 Key-Words

Specific lists of key-words signal the belonging of a token to a particular NE category. For example, such words like *'denken', 'sagen'* may

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

	Strict			Loose			Outer			Inner		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Morphological	49.63	46.64	48.09	50.22	47.19	48.66	49.53	49.25	49.39	54.72	13.68	21.89
+ Context	75.18	61.09	67.41	75.78	61.57	67.94	76.06	63.20	69.04	59.35	34.43	43.58
+ Word-Shapes	76.15	65.59	70.48	76.83	66.18	71.11	77.22	67.69	72.14	58.45	39.15	46.89
+ Key-words	76.45	65.70	70.67	77.14	66.29	71.30	77.48	67.80	72.32	59.29	39.15	**47.16**
+ Gazetteers	76.76	65.94	**70.94**	77.45	66.53	**71.58**	77.84	68.06	**72.63**	58.87	39.15	47.03

Table 2: Results on the development set.

indicate PER NE; *'gründen', 'arbeiten'* are particular for ORG but also for PER; words *'Kino', 'Musik', 'Werk'* characterize the category other.

5.1.5 Gazetteers

Various gazetteers from different sources such as Wikipedia, DBpedia, the GeoNames geographical database etc. have been analysed. NEs were automatically extracted from these resources, categorized into different NE classes and written into lists. The size of the elaborated lists varies from 434 for category OTH to 339392 for category PER.

5.2 Inner Span

For the inner classifier a similar set of features has been used. However, the feature class *key-words* and the 'in bracket'-features are excluded as they lose their relevance for the sub-structure. The features from class *Word-Shapes* are also limited to two tokens.

Because the inner classifier is trained after the outer classifier, information about the NE tags the outer classifier assigns to the target, previous and following tokens is accessible. We use this information as additional features for the inner span.

Additionally, we include the NE tags of the three previous tokens for the inner span as a concatenated string.

6 Evaluation

DRIM has been evaluated on the development set provided by GermEval via the distributed scorer, which requires six tab-separated columns: index, token, first-level NEs (gold), second-level NEs (gold), first-level NEs (prediction), second-level NEs (prediction).

In our system, we define the baseline model where the NE tag probabilities depend on the morphological features with a current token, POS and lemma information, specifying token suffix, prefix and root. With these features, the system achieves an F-score of 48.09% (see first line of Table 2).

Including the features of the Word-Context-Class demonstrates that the performance of the NER system can be improved up to 19% (see second line of Table 2). Whereas, in other languages such morphological characteristics as capitalization are useful, for German it is almost impossible to find out the right definition of the word without a context. That is why using the information about POS, lemma and NE annotations of the surrounding words of the target token increases significantly the recognition of NEs in German.

Another important feature class is Word-Shapes. Using these features additionally to Morphological features and Word-Context features improved the F-score to 70.48% (see third line of Table 2).

Light improvements could be seen by adding Key-Words and Gazetteer features. With the Key-Words features the score is improved to 70.67% (see forth line of Table 2). We assume that Key-Word features would be better represented with the elaboration of the key words, particular to a certain category. Adding the Gazetteers features improves the final score to 70.94% (see fifth line of Table 2).

7 Conclusion and Future Work

The current work presented the SVM-based named entity recognition system DRIM and its participation at the GermEval 2014 NER Shared Task. The context of the current token has turned out to be the most informative feature class for NER for German. Experimental results on the strict (fine-grained) setting have shown a reasonably good system performance reaching 70.94% on the development set, and 69.33% on the final test set. In the future, we plan to explore variations of the current features, extending the Gazetteers and separating the common key words into groups particular to the different NE cate-

gories. Since context features have shown to be highly informative for this task, we plan on exploring further the optimal size of the context window that should be considered.

References

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *Proceedings of LREC-14*.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA. ACM.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Bengali named entity recognition using support vector machine. In *Proceedings of Workshop on NER for South and South East Asian Languages, 3rd International Joint Conference on Natural Languge Processing (IJCNLP)*, pages 51–58, India.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Alexander Geyken and Norbert Schrader. 2006. LexikoNet, a lexical database based on role and type hierarchies. In *Proceedings of LREC*.

T. Kudo and Y. Matsumoto. 2000. Use of Support Vector Learning for Chunk Identification. In *Proceedings of Sixth Conference on Computational Natural Language Learning (CoNLL-2000)*.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrei Mikheev, Claire Grover, and Marc Moens. 1998. Description of the ltg system used for muc-7. In *Proceedings of 7th Message Understanding Conference (MUC-7)*.

Marc Rössler. 2004. Corpus-based learning of lexical resources for german named entity recognition. In *LREC*. European Language Resources Association.

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

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 49–56, Sapporo, Japan, July. Association for Computational Linguistics.

Mark Stevenson and Robert Gaizauskas. 2000. Using corpus-derived name lists for named entity recognition. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 290–295, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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 to Information Systems*, NLDB'01, pages 153–163. GI.