GermaNER:
Free Open German Named Entity Recognition Tool

Darina Benikova[1] Seid Muhie Yimam[1] Prabhakaran Santhanam[2] Chris Biemann[1]
(1) FG Language Technology, CS Dept., TU Darmstadt, Germany
benikova@aiphes.tu-darmstadt.de, {yimam,biem}@cs.tu-darmstadt.de
(2) IIT Patna, Dept. of CS and Eng., India
prabhakaran.cs11@iitp.ac.in

Abstract

With this paper, we release a freely available statistical German Named Entity Tagger based on conditional random fields (CRF). The tagger is trained and evaluated on the GermEval 2014 dataset for named entity recognition and comes close to the performance of the best (proprietary) system in the competition with 76% F-measure test set performance on the four standard NER classes. We describe a range of features and their influence on German NER classification and provide a comparative evaluation and some analysis of the results. The software components, the training data and all data used for feature generation are distributed under permissive licenses, thus this tagger can be used in academic and commercial settings without restrictions or fees. The tagger is available as a command-line tool and as an Apache UIMA component.

1 Introduction

Named Entity Recognition (NER) is the detection and classification task of proper names in continuous text. NER is used in information extraction, question answering, automatic translation, data mining, speech processing and biomedical science (Jurafsky and Martin, 2000). Moreover, it is a pre-processing step for deeper linguistic processing such as syntactic or semantic parsing, and co-reference resolution.

Despite German being a wide-spread and comparatively well-resourced language, German NER has not received a lot of attention. To the present day only three notable datasets exist, namely CoNLL-data (Tjong Kim Sang and De Meulder, 2003), an extension of this dataset to user-generated content by Faruqui and Padó (2010) and the NoSta-D NE dataset (Benikova et al., 2014b). So far, there has been no freely available German NE tagger. NER for German is especially challenging, as not only proper names, but all nouns are capitalized, which renders the capitalization feature less useful than in other Western-script languages such as English or Spanish. A baseline established on capitalized words therefore fails to show even moderate accuracy levels for German. This is reflected in previous results, e.g. from the CoNLL-2003 challenge, where German NER systems scored in the range of 70%-75% F-measure, as opposed to a recognition rate of 90% for English (Tjong Kim Sang and De Meulder, 2003).

We present GermaNER, a generic German NE tagger that can be easily executed from a command line or integrated into an NLP application. This paper presents the mechanism of the tagger, including the creation and experimental evaluation of the utilized features. The evaluation of the feature performance is accomplished using the F-measure, precision, and recall.

The tagger identifies the four default coarse named entity classes LOCation, PERson, ORGanisation, and OTHer. We have pragmatically excluded other NER subclasses and nested NERs from the GermEval 2014 task.

1.1 Free permissive licensing

Our most important contribution is the availability of GermaNER under a permissive license that allows academic and commercial use without licensing fees. The software components are mixed-licensed under modified BSD and ASL 2.0 licenses, the training and feature data is licensed under CC-BY. Unfortunately, these strict conditions on the permissiveness of licenses are not easy to meet. While it would have been possible to use more and better preprocessing components, more and better word lists for feature generation and possibly a better classifier, we had to exclude

Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology,
pages 31–38, University of Duisburg-Essen, Germany, Sep 30–Oct 2 2015.

the most part of them since many components are only free for academic use. We believe that placing these restrictions on software and data from publicly funded projects is hampering the development of language technologies as a whole, and German language processing in particular. The challenge of not being able to use standard preprocessing like part-of-speech tagging, however, led us to incorporate the output of several unsupervised methods that model required structural characterization in alternative ways.

2 Related Work

So far, two datasets were used for German NER in the academic community. The CoNLL-data by Tjong Kim Sang and De Meulder (2003) had flaws due to its inconsistencies in the training data, which were probably due to the circumstance that the annotators were non-native speakers (Leveling and Hartrumpf, 2008). Systems participating in the CoNLL 2003 contest achieved F-measure between 70%-75%. Faruqui and Padó (2010) have extended this data for evaluation purposes, and made available a German NER module for the Stanford NER tagger (Finkel et al., 2005) which is however, only free for academic use.

The NoSta-D NE data set by Benikova et al. (2014b) was used for the GermEval 2014 NER shared task (Benikova et al., 2014a). But the setting of the task is different to the one used for this project. In the GermEval 2014, the NE annotation were performed on nested-layers, so that entities like 'Madrid' in 'Real Madrid', were also detected. Moreover, in the shared task, derivations like 'German' and parts of NEs, such as 'Germany' in 'Germany-wide' are annotated.

The three best systems at GermEval 2014, ExB (Hänig et al., 2014), UKP (Reimers et al., 2014) and MoSTNER (Schüller, 2014) perform in a range of 73%-79% F-measure on the default set of four NER types (Metric 3 first-level spans) (Benikova et al., 2014a). All these systems implemented machine learning methods that make use of interdependencies among data points, such as Conditional Random Fields (CRFs) and Neural Networks.

While most participants used POS-level, character-level and gazetteer-based features, each of the three best performing systems (Reimers et al., 2014; Schüller, 2014; Hänig et al., 2014) operated with high-level semantic features, such as similarity clusters or word embeddings. These features were created using unsupervised learning methods on large corpora and successfully address the vocabulary problem and sparsity issues through vocabulary clusters (Schüller, 2014; Hänig et al., 2014) or dense vector representations (Reimers et al., 2014).

As previously shown, the use of such simple semantic generalization features improves the recall for NER (Biemann et al., 2007; Finkel and Manning, 2009; Faruqui and Padó, 2010). Moreover, the ExB system applied well-curated NE-specific suffix lists, containing entries such as '-stadt', '-hausen', or '-ingen' for locations.

3 Machine Learning Approach

There are different approaches to NER, including handcrafted rule-based algorithms, supervised machine learning, unsupervised machine learning and semi-supervised algorithms (Nadeau and Sekine, 2007). While a rule-based NER approach usually produces a high precision, it covers a single domain and fails to perform well when new entity types appear in a document (Petasis et al., 2001). Machine learning approaches perform more robustly and are more accurate if sufficient training data and adequate features are incorporated. Supervised NER approaches mainly depend on large collections of texts that are syntactically annotated to systematically recognize NEs based on syntactic patterns (Nadeau et al., 2006).

In our work, we focus on the development of a supervised machine learning NER system that can 1) be readily used from command line or integrated into any NLP applications to automatically tag NEs, and 2) be used as reasonable baseline system to further expand training data sets using active learning and adaptive annotation approaches, and 3) is not subject to license restrictions, thus can be freely downloaded and used by anyone.

3.1 Conditional Random Field (CRF)

While there are plenty of machine learning algorithms for sequence tagging, we choose to integrate a CRF (Lafferty et al., 2001) as it is highly accurate, scalable and easy to use as the training data can be prepared without the need of machine learning experts (Hoefel and Elkan, 2008). We have specifically integrated CRFsuite (Okazaki, 2007), a fast implementation of Conditional Random Fields, into a clearTK UIMA framework

(Bethard et al., 2014) to make training, feature annotation, classification and entity extraction more convenient.

The NER system is highly configurable, which allows users to either use the built-in model that is already optimized with our feature set, or train it with new training data and features sets. In order to make the NER system usable for both low-end and high-end machines, it provides a technique of data chunking, where users with high-end machines can use larger data chunks while users with low-end machines can still run the system on their laptop computer with smaller data chunks.

4 The NER system pipeline

The NER tagger pipeline consists of different components integrated into an UIMA (Ferrucci and Lally, 2004) pipeline written in the Java programming language. We have designed the NER tagger in such a way that each of the components can be replaced or modified easily. The first component of the system obtains the training and testing data and applies segmentation and tokenization that is stored in a UIMA CAS for further processing. The next component is the feature extraction process that internally annotates the documents accordingly. Feature extractors obtain different features either from the token and surrounding tokens, such as word and character n-grams, or the features are supplied from external sources, such as gazetteer lists or lists induced by unsupervised methods. The training component produces a CRFsuite classifier model based on the annotated features. The final component is a classifier component where unseen documents, which get feature-annotated in a similar way as the training file, are subject to prediction of NEs. Figure 1 shows a diagram of the GermaNER tagger system pipeline.

5 Data

5.1 Training Data

For training the NE-Tagger, we use the NoSta-D NE dataset. It consists of 31,300 sentences and more than 37,000 named entity span annotations that are used for training. The original dataset contains more annotations such as partial NEs and NE derivates and nested annotations, which were used in GermEval 2014, but have been excluded in GermaNER. The classes that are used in the training are LOCation, PERson, ORGanisation, and OTHer. All derivation and part classes that are contained in the original dataset were treated like unannotated tokens, because the task of GermaNER is the tagging of the four default coarse NE classes. However, these classes have been used for training and testing for the purpose of comparing the results of GermaNER to the systems that participated in the GermEval 2015 as shown in Table 2.

The training, development and test set were divided just as in the GermEval setting: 24,000 sentences for training, 2,200 for the development and 5,100 sentences for the test set. We optimized feature combinations on the development set and report evaluation scores for the same settings as in the GermEval 2014 challenge.

The final model of GermaNER as included in the GermaNER distribution was trained on the concatenation of training, development and test set. While we cannot assess the quality of the model, the test set performance as reported here can serve as a lower-bound estimate.

5.2 Data Input and Output Format

The input of GermaNER is a file, similar to the CoNLL format, which contains one token per line. Sentences should be separated by a blank line. The output of the tagger is a tab-separated file. The first column is the same as in the input file. The second column holds the predicted NE-tag. The NE-tags are similar to those employed in the training dataset, which made use of the BIO-scheme[1]. An exemplary output sentence of the tagger is presented below.

[1] "The **BIO** scheme suggests to learn classifiers that identify the **B**eginning, the **I**nside and the **O**utside of the text segments."(Ratinov and Roth, 2009)

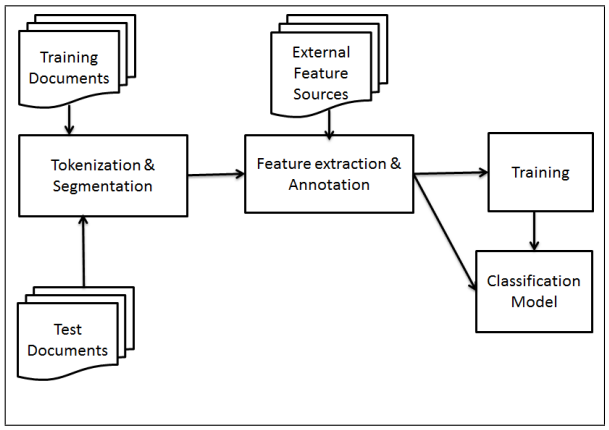

Figure 1: The German NER tagger pipeline

Nehmen	O
Sie	O
die	O
berühmte	O
Rede	O
von	O
Richard	B-PER
Feynman	I-PER
,	O
There	B-OTH
's	I-OTH
Plenty	I-OTH
of	I-OTH
Room	I-OTH
at	I-OTH
the	I-OTH
Bottom	I-OTH
,	O
von	O
1959	O
,	O
die	O
ist	O
damals	O
in	O
der	O
Zeitschrift	O
des	O
California	B-ORG
Institute	I-ORG
of	I-ORG
Technology	I-ORG
abgedruckt	O
worden	O
.	O

Table 1: Exemplary output of GermNER

6 Feature Representation

The creation and selection of features is a crucial part in the development of NER systems. The creation of all included features will be presented in feature groups, as the discussion of every single feature would be redundant. Some of the feature-groups are: 1) n-gram features such as character n-grams, unsupervised POS tag n-grams, and topic cluster n-grams, 2) time-shifted features, i.e. token-based features from surrounding tokens in relative position {-2,-1,0,+1,+2} to the current token 3) combinations of 1 and 2 such as as character n-grams features for one token to the left and right. We now provide details for all features.

6.1 Character and Word Features

This feature group consists of the first and last character uni-, bi- and trigrams of the current token, i.e.. prefixes and suffixes, time-shifted from -2 to +2. Similarly, character category pattern features, which are extracted from the current token based on unicode categories[2] from clearTK are used, and were found to be an influential feature for the system. Further, we use the words themselves as features in a window between -2 and 2.

6.2 NE Gazetteer

This gazetteer feature was created through the assembling of several lists containing NEs. Gazetteers may help to identify NEs that are known to be proper nouns in other contexts. For the *FreebaseList*, several Freebase lists containing proper nouns were merged. Freebase (Bollacker et al., 2008) is an English community-curated data-base containing well-known places, people and things under CC-BY-license. It contains 47 million lists, so-called topics, and 2 billion entities. The entities are ordered into different topics (e.g. Music Album, Family Name, or Continent) which are part of domains (e.g. People, Music, or Location). The largest task relevant lists as well as lists with frequent NEs such as Country or Currency were chosen for the final list. The following lists were incorporated in the gazetteer: Album, Mountain, Book, Musical Group, Book Edition, Organization, Citytown, Person, Country, River, Currency, Stock Exchange, Film Track, Human Language, TV-series-season, Lake, Work of Fiction, and Location.

Only the first column of the lists provided by Freebase was used for this task. The lists were stripped of all entries containing special characters or spaces only. Moreover, double entries of proper names in the same list were removed.

The final *FreebaseList* is a tab-separated file consisting of two columns. The first column contains the proper name and the second column contains the name of the list file it was extracted from. It was used as a look-up table to extract features for every token that was in the table for the corresponding class.

Several other gazetteers were incorporated as features, such as personal name lists extracted with the NameRec tool from ASV Toolbox (Biemann et al., 2008) from large, publicly available corpora. This merged feature group *Gazetteer features* is shown seperatly from the *FreebaseList* in Table 2.

[2] http://www.unicode.org/notes/tn36/

6.3 Parts of Speech

There are several approaches of machine learning for retrieving parts of speech (POS) of words in context automatically. We have incorporated automatically induced POS tags as POS features.

This POS induction is based on the system by Clark (2003), which clusters words into different classes in an unsupervised fashion, based on distributional and morphological information. For this setup, we have used 10 million sentences, which are part of the Leipzig Corpora Collection[3] Richter et al. (2006), and induced 256 different classes.

Additionally, we experimented with classical POS features using the Mate POS tagger (Bohnet, 2010). Our tool will not include this feature as the licenses of the POS tagger and its training data would render our tool unusable for commercial purposes. However, we will provide the possibility to add this feature so that it can be used in an academic setting.

6.4 Word Similarity

This feature group consists of the four most similar words of the current token, obtained from the JoBimText[4] (Biemann and Riedl, 2013) distributional thesaurus database, made available in a window of size 2.

6.5 Topic Clusters

Inspired by the semantic clusters of the ExB system, we have applied LDA topic modelling[5] to above-mentioned JoBimText German distributional thesaurus, using the thesaurus entries as 'documents' for LDA. This results in a fixed number of topic clusters, most of which are quite pure in terms of syntactic and semantic class. We have generated different sets of such clusters, each for all words and for uppercase words only, and use the number of its most probable topic as a token's feature – again, time-shifted in a range of -2 to 2. We experimented with sets of 50, 100, 200 and 500 clusters. In the final version we solely use the set of 200 clusters.

6.6 Other

There are two further features that were implemented: token *Position* and *Case*. *Position* feature is the position of the token in the sentence, while *Case* feature is the case of the token, distinguishing between uppercase and lowercase, the beginning of a sentence, camelCase and all uppercase, time shifted between -2 and 2.

7 Evaluation

In this section, the evaluation metric and other factors influencing the choice of features of the final GermaNER tagger will be presented.

7.1 Methods

For the evaluation of the feature performance, we report scores from the M3.1 metric described in the GermEval 2014 task. It calculates the precision, recall, accuracy and f-measure of the outer layer, which was the only layer of interest for the work described in this paper. To further investigate the issues of the current version of the tagger, the performance on individual classes will be discussed.

In order to determine the optimal feature set, different feature combinations have been tested in order to arrive at a final default feature set. The default feature set contains all features. To determine features that potentially reduce the performance of the NER by overfitting, we performed ablation tests.

7.2 Results

Table 2 shows the results of the previously described evaluation on the development and test data sets. The first line shows results with all features. The scores in boldface indicate the three most influential features, as leaving them out results in the most dramatic drops in tagging quality. The last line displays the performance of the full feature set including supervised POS features from the Mate POS tagger, which is however that is not part of our final system.

Table 2 shows that all features are relevant to the NER tagger, thus the final tagger makes use of all the described features. The best performing feature group are character n-grams. As not only n-grams of the current word, but also n-grams of preceding and following words are used, this features play a role that is on the one hand similar to the detection of prefixes and suffixes that indicate NEs, but are on the other hand similar the detection of words typically preceding or following NEs e.g. prepositions that precede NEs. The

[3] corpora.uni-leipzig.de

[4] http://www.jobimtext.org

[5] http://gibbslda.sourceforge.net/

Model	Precision (%)	Recall (%)	F-measure (%)
All features	83.16	74.32	78.49
no character n-grams	82.18	69.81	**75.49**
no case information	81.93	73.29	**77.37**
no gazetteers	82.75	73.92	78.08
no positions	82.69	74.23	78.23
no Freebase	82.56	73.56	77.80
no char cat. pattern	82.93	72.89	77.58
no similar words	82.29	73.25	**77.50**
no topic clusters	82.48	73.60	77.79
no clark POS induction	82.64	73.34	77.71
with Mate POS tagger	82.65	75.12	**78.71**

Table 2: Results of feature performance evaluation on the development set. Lower F-measure means a high impact of the corresponding feature

second best performing feature group is case information, meaning the classification of words in e.g. uppercase and lowercase. Although, as already mentioned earlier, this feature is not as distinguishing for proper nouns in German as it is in other languages, our experiments show that it still is an important feature in German NER. These two best performing features are standard features in NE detection, the advantage of which is confirmed through our experiment. The third best performing feature is a *similar words*, which is a semantic feature. This not only shows the importance of the semantic layer in this task, but also goes in line with the three best systems participating in the GermEval 2014, that also made use of high-level semantic features. Interestingly, the supervised POS tagger reduces precision and increases recall, resulting in a very modest increase of 0.22% F-score, thus is mostly subsumed by the unsupervised feature groups.

Table 3 reports overall P/R/F-results when training GermaNER on the concatenation of the training and development set and testing it on the official test set and also provides scores of the best GermEval 2014 participants. While the first two results are not directly comparable since the challenge participants were also asked to tag NE derivates and partial NEs, they indicate that GermaNER shows a competitive score to the UKP system and is outperformed by the proprietary ExB system. Interestingly, GermaNER has a comparatively high precision but a lower recall compared to ExB and UKP. To provide a better comparison to the other systems, the GermaNER system was trained on the concatinated training and development set including parts and derivations. The result, which is shown in the last line in Table 3, shows that although the tagging of parts and derivs was not the focus of this tagger, GermaNER is only outperformed by ExB and UKP.

	ExB	UKP	MoSTNER	GermaNER
PER	84.05	85.48	82.54	85.33
LOC	84.05	84.62	80.47	81.39
ORG	76.29	69.60	62.24	68.23
OTH	59.46	49.81	48.38	52.72

Table 4: Test set peformance in % F-measure by NE type for top GermEval 2014 systems

Both Table 2 and Table 3 show that the adjoining of a license resticted supervised POS tagger noticably improves the performance of GermaNER. These examples demonstrate the impact of permissive licensing on the performance of freely available tools.

Finally in Table 4, we provide an F-measure comparison broken down into the four coarse NER classes. Here, it becomes apparent that GermaNER is very strong on PERsons and that there is still some headroom for the other three classes, probably due to the lack of gazetteers for these other classes.

8 Conclusion and Future Work

We have developed GermaNER, a statistical German NER tagger which can be readily used from command line or can be integrated to an NLP application. While the architecture and the features of GermaNER are following common practice in sequence tagging and do not provide much

System	Precision (%)	Recall (%)	F-measure (%)
ExB	80.67	77.55	79.08
UKP	79.90	74.13	76.91
MoSTNER	79.71	67.74	73.24
GermaNER	82.72	71.19	76.52
GermaNER with POS	82.16	72.21	76.86
GermaNER including deriv and part	81.98	69.88	75.45

Table 3: Results of best GermEval 2014 systems and GermaNER on the test set, for different sets of classes from the NER dataset.

methodological novelty, we would like to stress the fact that the tagger is freely available in open source for download[6] under a permissive mixed license, allowing its use as a standalone or as a component in academic and commercial contexts without license restrictions or fees.

In its best configuration, GermaNER performs at an F-measure of 78.49% on the GermEval 2014 dev set and at 76.52% on the test set. The three features with the largest impact are the character n-grams, case information and similar words.

In summary: we provide a freely available German NER tagger for standard categories that comes close to the state of the art. Our largest challenges in creating this tagger were rooted in the fact that many resources and tools for language preprocessing are only available under restrictions. We hope to have advanced German language technology, both in academia and industry, by overcoming these limitations for named entity recognition and would like to see more free components in the future. Of course, everyone is welcome to add features and make high-quality German NER a community effort.

Acknowledgements

The authors would like to thank Martin Riedl for the idea and the implementation of the Topic Cluster feature in Section 6.5.

References

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. GermEval 2014 Named Entity Recognition: Companion Paper. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany*, pages 104–112, 2014a.

Darina Benikova, Chris Biemann, and Marc Reznicek. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of LREC*, pages 2524–2531, Reykjavik, Iceland, 2014b.

Steven Bethard, Philip Ogren, and Lee Becker. ClearTK 2.0: Design Patterns for Machine Learning in UIMA. In *Proceedings of LREC*, pages 3289–3293, Reykjavik, Iceland, 2014.

Chris Biemann and Martin Riedl. Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95, 2013.

Chris Biemann, Claudio Giuliano, and Alfio Gliozzo. Unsupervised Part of Speech Tagging Supporting Supervised Methods. In *Proceedings of RANLP-07*, Borovets, Bulgaria, 2007.

Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proc. International Conference on Computational Linguistics (COLING 2010)*, pages 89–97, Beijing, China, 2010.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, Vancouver,Canada, 2008. ACM.

Alexander Clark. Combining distributional and morphological information for part of speech induction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Budapest, Hungary, 2003.

Manaal Faruqui and Sebastian Padó. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS*, pages 129–133, 2010.

[6] https://github.com/tudarmstadt-lt/GermaNER

David Ferrucci and Adam Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. In *Journal of Natural Language Engineering 2004*, pages 327–348, 2004.

Jenny R. Finkel and Christopher D Manning. Joint Parsing and Named Entity Recognition. In *Proceedings of HLT-NAACL 2009*, pages 326–334, Boulder, CO, USA, 2009.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370, Ann Arbor, MI, USA, 2005.

Christian Hänig, Stefan Bordag, and Stefan Thomas. Modular Classifier Ensemble Architecture for Named Entity Recognition on Low Resource Systems. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 113–116, Hildesheim, Germany, 2014.

Guilherme Hoefel and Charles Elkan. Learning a Two-stage SVM/CRF Sequence Classifier. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 271–278, Napa Valley, California, USA, 2008.

Dan Jurafsky and James H Martin. *Speech & Language Processing*. Pearson Education India, 2000.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289, Williamstown, MA, USA, 2001.

Johannes Leveling and Sven Hartrumpf. On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science*, 22(3):289–299, 2008.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae investigationes: Revue internationale de linguistique française et de linguistique générale*, 30(1):3–26, 2007.

David Nadeau, Peter D. Turney, and Stan Matwin. Unsupervised Named-entity Recognition: Generating Gazetteers and Resolving Ambiguity. In *Proceedings of the 19th International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence*, AI'06, pages 266–277, Québec City, Québec, Canada, 2006.

Naoaki Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007. URL `http://www.chokkan.org/software/crfsuite/`.

Georgios Petasis, Frantz Vichot, Francis Wolinski, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D. Spyropoulos. Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Proceedings of ACL*, pages 426–433, Toulouse, France, 2001.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL*, pages 147–155. Association for Computational Linguistics, 2009.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, and Iryna Gurevych. GermEval-2014: Nested Named Entity Recognition with Neural Networks. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 117–120, Hildesheim, Germany, 2014.

Matthias Richter, Uwe Quasthoff, Erla Hallsteinsdóttir, and Chris Biemann. Exploiting the leipzig corpora collection. In *Proceedings of IS-LTC*, pages 68–73, Ljubljana, Slovenia, 2006.

Peter Schüller. MoSTNER: Morphology-aware split-tag German NER with Factorie. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 121–124, Hildesheim, Germany, 2014.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, pages 142–147, Edmonton, Canada, 2003.