Nessy: A Hybrid Approach to Named Entity Recognition for German

Martin Hermann, Michael Hochleitner, Sarah Kellner, Simon Preissner, Desislava Zhekova
CIS, Ludwig Maximilian University, Munich
{Martin.Hermann, M.Hochleitner, S.Kellner, Simon.Preissner, D.Zhekova}
@campus.lmu.de

Abstract

In this paper we present Nessy (Named Entity Searching System) and its application to German in the context of the GermEval 2014 Named Entity Recognition Shared Task (Benikova et al., 2014a). We tackle the challenge by using a combination of machine learning (Naive Bayes classification) and rule-based methods. Altogether, Nessy achieves an F-score of 58.78% on the final test set.

1 Introduction

Named Entity Recognition (NER) is a subtask of information extraction and is an important topic in natural language processing. It is useful for the identification of where information is located, how it may be connected and used for tasks such as text classification (Gui et al., 2012) and question answering (Mollá et al., 2006).

However, NER is not a simple task, especially for German, where capitalization is not as informative as in many other languages, such as English or Spanish. Following the NE annotation guidelines presented by Benikova et al. (2014b), the GermEval Shared Task on Named Entity Recognition (Benikova et al., 2014a) aims at detecting named entities (NEs) and assigning them to one of four classes: persons (-PER), locations (-LOC), organizations (-ORG), and the class of other (-OTH), where those NEs are assigned to which cannot be matched with the aforementioned classes. Furthermore, there are two subclasses (-part and -deriv) which are used for NEs that are subparts of bigger entities (-part, e.g. *deutschlandweit*) or derivatives (-deriv, e.g. *Bremer* Staatsanwaltschaft).

Named Entity Recognition and Classification (NERC) was introduced as a subtask of Information Extraction (IE) at the 6th Message Understanding Conference (MUC-6) in 1995 (Nadeau and Sekine, 2007). Since then, remarkable results have been reached for NER in English. Systems at the 7th Message Understanding Conference (MUC-7) reached scores of up to 93% (Mikheev et al., 1998), which is close to the inter-annotator agreement 96% for that task (Chinchor, 1998). So far, most work in NER for German was conducted in the context of the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition (Tjong Kim Sang and De Meulder, 2003). The systems reached F-scores of 72.41% on the German test set and 88.76% on the English test set. Among the machine learning techniques used for CoNLL-2003 Maximum Entropy (MaxEnt) and Hidden Markov Models (HMM) were most popular (Tjong Kim Sang and De Meulder, 2003).

Combining different classifiers also proved to be beneficial. Florian et al. (2003), for example, added robust linear classifier and transformation-based learning to MaxEnt and HMM. Additionally, to improve the performance of classification, it was common to make use of gazetteers.

Unfortunately, for German, there are not many freely available and simultaneously high-

fgroup	name	description
d_lex	*pos*	POS-tag of the token
	word	token itself
d_other	*prev_dec*	preceding IOB-tag
	all_caps	check if all characters are uppercased

Table 1: The feature groups (fgroup) used for NED.

performance NERs. One such system that applies semantic generalizations learned from unlabelled data was presented by Faruqui and Padó (2010).

In this paper, we describe the NER system Nessy developed for the GermEval 2014 Shared Task. We break NER down into two steps: named entity detection and named entity classification, both described in section 2 where all further details about the system pipeline are presented. In section 3, we provide a discussion on the results achieved by Nessy on the development set provided by the GermEval 2014 Shared Task and in section 4 we conclude our work.

2 The Nessy System

2.1 Preprocessing

Part-of-Speech (POS) tags and lemmas were acquired via the TreeTagger (Schmid, 1994; Schmid, 1995). Additionally large lists of known NEs (gazetteers) were prepared (containing 68922 entries). These NEs were taken directly from the already manually annotated data provided by the CoNLL-2003.

2.2 Named Entity Detection

For the task of named entity detection (NED), we use a Naive Bayes classifier and tag each of the words in an IOB-manner. The small set of features currently used in this classifier are presented in table 1. To make sure that the output contains only valid IOB-sequences any isolated I-tag is converted into a B-tag.

2.3 Named Entity Classification

For Named Entity Classification (NEC), we extract the presumable named entities found during NED. Again, these are passed to a naive Bayes classifier that uses the features given in table 2. In the case of one-word-entities, the features *ne*, *first_t* and *last_t* contain the same information. The feature *in_lookup* checks against the gazetteers prepared during preprocessing.

fgroup	name	description
c_lex	*ne*	the named entity itself
	lemmas	the sequence of lemmas in the NE
	first_t	the first word of the respective NE
	last_t	the last word of the respective NE
c_cont	*prev_t*	the word preceding the NE
	foll_t	the word following the NE
c_other	*num_t*	number of tokens in the NE
	all_caps	check if all characters are uppercased
	in_lookup	gazetteer lookup

Table 2: The feature groups (fgroup) used for NEC.

2.4 "part" and "deriv" Subclasses

Tags labeled with "part" and "deriv" are an individual characteristic of this data. Although many of them are already correctly found by the classifier, additional steps proved to be necessary.

2.4.1 The "part" Subclass

Tags ending in "part" are used to annotate tokens that are not NEs themselves, but contain a substring that does qualify as such. They make up about 5.5% of NEs in the training and 6.4% in the development data, most of which (96.4% in the training, 97.3% in the development data) occur in the outer layer. Hence, we neglect the inner layer completely in this step. Additionally, as we simply "overwrite" previously assigned tags, this may also correct mistakes in the detection step (e.g., if the phrase *EU-Kommissarin Viviane Reding* is (incorrectly) marked with "PER", detection of *EU-Kommissarin* as "ORGpart" would not only label this token appropriately, but also correct the span of *Viviane Reding*. Had we written the "ORGpart" label in the inner layer, we would end up with two wrong annotations.)

The detection of "part" tags is done with four lists of single-word NEs, one for every category, compiled from the training data and expanded with the list of stems described below. The list is revised, such that only entries are allowed that occur more often as a NE of the given category than not, in order to reduce ambiguity that may arise from either inaccuracies in the data, or, more likely, language itself (e.g. many surnames, such as *Gold*, are also common nouns).

By far, the biggest part (77.9% in the training, 77.7% in the development data) of partial NEs contains one or more hyphens ("-"), and in turn, a considerable amount of tokens (19.8% in the

NEs that are missing their "B-" tag are corrected.

training, 22.7% in the development data) containing hyphens are labeled with the "part" subclass, so it seems sensible to focus on these. Such tokens are separated at the hyphens and the first part is checked against the lists of single-word NEs. If a match is found, the token is labeled accordingly.

2.4.2 The "deriv" Subclass

Derivated forms of NEs are marked with tags ending in "deriv". As they account for about 11.9% of NE in the training and 10.5% in the development data, they should not be neglected. Especially LOCderiv, such as *deutschen* (German) or *Engländer* (Englishman) are very common in all datasets. Unlike the "part" labels, a considerable amount (16.5% in the training, 15.8% in the development data) of tags with "deriv" is found in the inner layer, so it is more reasonable here to check if the derivated form may already be part of a larger NE.

Similar to the "part" labels, we use four lists of single-word candidates, although this time, the entries are not simply taken from the training data, but suitable entries found there are stemmed, and then the stems are combined with a list of possible endings, e.g. *-lich*, *-istischer* or *-erin*. However, controlling this list with the test data is even more important than in the previous case, as from *deut*, which is generated as stem of *deutsch* (albeit linguistically not entirely correct) not only *deutsches*, *deutscher* or *deutsche* are derived, but also *deutlich* (clearly) or *deutung* (interpretation), which would cause many false-positives. A lot of nonsensical words are also generated, such as **deutistisch*, but as they seldom appear, they do not need to be considered.

2.5 Inner Layer

The data contains recursive NEs to the depth of one nested layer. This inner layer is filled with some of the "deriv" labeled tags and some NE found in the postprocessing step, but it is reasonable to further search for possible nested NEs. As they can only occur if the outer layer is not empty, the search is done only within previously found NEs. Here, we make further use of the list of NEs that has been compiled for finding "part" tags, as it proves to yield better results at this point than the gazetteers compiled from the CoNLL-2003 data

2.6 Additional Rules

Several rules have been written that account for special cases of NEs. These can be grouped into four different classes:

Hyperlinks: Hyperlinks are always annotated as NEs of the category OTH.

Hyphens: While hyphens usually are a sign for the "part" subclass (as described above), compounds that contain one or more hyphens and end in a NE usually obtain the class of that NE. This is so, since in German the last part of a word determines its class. So, for example, while both *Taiwan* and *Dollar* in *Taiwan-Dollar* are NEs, *Taiwan-Dollar* is a form of *Dollar*, and therefore should be categorized as OTH, just like *Dollar* itself.

Split-off parts: A hyphen at the end of a token (e.g. *Süd-*) and tokens such as *und* (and) or *oder* (or) following it may indicate split-off parts (e.g. *Süd- und Nordkorea*), both of which should have the class of the second token, in this case, LOC.

Tokens following nationalities: Nessy tends to mark any nationality and its following token as a two-word-NE. This, however, is hardly ever the case, unless the nationality starts with an uppercase letter (e.g. *Deutsches Theater*). Such subsequent tokens are discarded by using a list of nationalities during postprocessing.

3 Evaluation

The Nessy system was evaluated on the development set provided by GermEval 2014 (Benikova et al., 2014a). The results on the development and final test set are given in table 3. In order to see how informative the different feature types are (given in table 2), we evaluate separately a number of forward/backward inclusion/exclusion settings on the development data. First, we test each of the different feature groups separately, leading to settings +c_cont, +c_lex, +c_other in table 3 and then, we report results by excluding one of the groups, leading to settings -c_cont, -c_lex, -c_other. All three groups together are marked as +all in the table. Additionally, all seven variations are once tested on their own (-R) and once

Cases such as *EU-*, where the only hyphen in the word is at the end, are checked against.

setting	**Metric 1 (Strict)**				**Metric 2 (Loose)**				**Metric 3 - Outer Chunks**				**Metric 3 - Inner Chunks**			
	Acc.	**P**	**R**	**F1**	**Acc.**	**P**	**R**	**F1**	**Acc.**	**P**	**R**	**F1**	**Acc.**	**P**	**R**	**F1**
+c_cont-R	95.99	42.46	39.54	40.95	96.08	44.03	40.99	42.45	92.94	45.32	40.73	42.90	99.04	18.31	24.53	20.97
+c_other-R	96.16	47.79	44.14	45.89	96.24	49.89	46.08	47.91	93.20	50.33	45.21	47.64	99.13	24.62	30.66	27.31
+c_lex-R	96.74	55.37	47.89	51.36	96.83	57.25	49.51	53.10	94.02	55.44	49.89	52.52	99.45	53.33	22.64	31.79
-c_lex-R	96.59	55.26	50.42	52.73	96.64	56.29	51.35	53.71	93.96	57.81	51.91	54.70	99.22	28.88	31.60	30.18
-c_cont-R	96.77	59.02	52.81	55.74	96.81	59.91	53.60	56.58	94.23	60.82	54.67	57.58	99.31	34.83	29.25	31.79
-c_other-R	96.93	60.54	53.22	56.65	96.98	61.37	53.95	57.42	94.47	61.62	55.35	58.31	99.39	41.48	26.42	32.28
+all-R	96.90	61.40	55.06	58.06	96.94	62.13	55.72	58.75	94.49	63.56	57.07	60.14	99.30	33.69	29.72	31.58
+c_cont+R	96.13	44.00	40.68	42.28	96.21	45.58	42.13	43.79	93.12	46.19	41.92	43.95	99.15	21.99	25.00	23.40
+c_other+R	96.33	50.02	45.77	47.80	96.42	52.56	48.09	50.23	93.42	51.73	46.93	49.22	99.25	30.70	31.13	30.91
+c_lex+R	96.82	57.35	50.00	53.42	96.91	59.14	51.56	55.09	94.17	57.39	52.13	54.63	99.46	56.32	23.11	32.78
-c_lex+R	96.78	57.96	52.36	55.02	96.82	58.96	53.26	55.96	94.21	59.51	53.96	56.60	99.35	37.36	32.08	34.52
-c_cont+R	96.94	61.76	54.9	58.16	96.97	62.62	55.72	58.97	94.47	62.73	56.96	59.70	99.41	45.00	29.72	**35.80**
-c_other+R	97.02	62.43	55.27	58.63	97.07	63.25	55.99	59.40	94.62	63.40	57.52	60.31	99.41	44.19	26.89	33.43
+all+R	97.06	64.04	57.14	**60.39**	97.10	64.74	57.76	**61.05**	94.72	65.36	59.27	**62.17**	99.40	42.67	30.19	35.36
final test	97,07	63,57	54,65	58,78	97,11	64,34	55,31	59,48	94,77	64,83	56,93	60,62	99,38	42,86	27,38	33,41

Table 3: System results achieved on the GermEval 2014 development (upper part) and official test (last row) set.

with the supplementary use of the handcrafted rules presented in section 2.6, (+R).

As can be seen from the results of the strict evaluation setting (Metric 1), most informative to the learner on its own was the group of lexical features (c_lex), which reaches F-score of 51.36% when used alone during classification (setting *+c_lex-R*). This is a considerably big contribution regarding the fact that this feature group consists of four basic features representing the tokens and lemmas contained in one NE span. The other two groups (c_cont and c_other) also seem to carry very valuable information for the recognition process reaching scores of 40.95% and 45.89% respectively (settings *+c_cont-R* and *+c_other-R*), showing that both contextual and features carrying information about the number of tokens in a NE, their capitalization and presence in gazetteers should not be ignored for this task. The combination of all three groups (setting *+all-R*), reaches an improved F-score of 58.06%.

All these settings are then combined with the use of manually created rules leading to the *+R* settings in table 3. What can be seen is that the used rules do not interact with the separate feature group contribution, which leads to the same result tendencies as without the application of rules. However, the latter do increase the system performance for all tested variations, leading to an F-score of 60.39% (see setting *+all+R*), which is the highest score of our system based on the development set. Such a performance is competitive to the performance of systems applied to German on the CoNLL-2003 Shared Task ranging between F-scores of 47.74% to 72.41% (Tjong Kim Sang and De Meulder, 2003). We consider this to be a very good performance given the small feature set we employ.

The F-score of 60.39% is based mainly on the system performance for the outer layer of NE (62.17%), which seems to be weaker for the inner layer (achieving 35.36%). In fact, with respect to the inner layer, the system reaches best scores (35.80%) when context features are not used (setting *-c_cont+R*), which is surprising, since these features deliver information from the outer span, which should indicate the type of the outer NE in which the inner NE is included.

4 Future Work and Conclusion

In this paper, we presented the participation of Nessy, which is a hybrid approach to NER, at the GermEval 2014 Named Entity Recognition Shared Task for German. We evaluated the system (using Metric 1) on the development set provided by GermEval 2014, reaching an F-score of 60.39% on the development set and 58,78% on the final test set, which is considerably good for the small feature set that the system employs.

In the future, we would like to look deeper into the use of world knowledge for NER and explore the use of features carrying information about possible semantic relations between the tokens present in the NEs and tokens included in already known NEs present in available gazetteers.

References

[Benikova et al.2014a] 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.

[Benikova et al.2014b] Darina Benikova, Chris Biemann, and Marc Reznicek. 2014b. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

[Chinchor1998] Nancy A. Chinchor. 1998. Proceedings of the Seventh Message Understanding Conference (MUC-7) Named Entity Task Definition. page 21 pages, Fairfax, VA.

[Faruqui and Padó2010] Manaal Faruqui and Sebastian Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010*, page 129. Semantic Approaches in Natural Language Processing.

[Florian et al.2003] Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named Entity Recognition Through Classifier Combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 168–171, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Gui et al.2012] Yaocheng Gui, Zhiqiang Gao, Renyong Li, and Xin Yang. 2012. Hierarchical Text Classification for News Articles Based-on Named Entities. In Shuigeng Zhou, Songmao Zhang, and George Karypis, editors, *ADMA*, volume 7713 of *Lecture Notes in Computer Science*, pages 318–329. Springer.

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

[Mollá et al.2006] Diego Mollá, Menno van Zaanen, and Daniel Smith. 2006. Named Entity Recognition for Question Answering. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, pages 51–58.

[Nadeau and Sekine2007] David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1):3–26, January.

[Schmid1994] Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

[Schmid1995] Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

[Tjong Kim Sang and De Meulder2003] 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.