German NER with a Multilingual Rule Based Information Extraction System: Analysis and Issues

Anna Druzhkina
National Research University
Higher School of Economics;
ABBYY
annarya@gmail.com

Alexey Leontyev
ABBYY
Aleksey_L@abbyy.com

Maria Stepanova
ABBYY
Maria_Ste@abbyy.com

Abstract

This paper presents a rule-based approach to Named Entity Recognition for the German language. The approach rests upon deep linguistic parsing and has already been applied to English and Russian. In this paper we present the first results of our system, ABBYY InfoExtractor, on GermEval 2014 Shared Task corpus. We focus on the main challenges of German NER that we have encountered when adapting our system to German and possible solutions for them.

1 Introduction

Named Entity Recognition (NER), which is a subtask of information extraction (Grishman, 2003), is a well-studied, yet challenging task. Various competitions have been held to evaluate quality of named entity recognition for different languages (MUC, CONLL-2002, IREX). German is no exception: there have already been two evaluation tracks for German, ConLL 2003 (Tjong Kim Sang and De Meulder, 2003) and GermEval 2014 (Benikova et al., 2014). The best results reported for the tracks are 72.41% and 76.38% respectively, which is still considerably below the results for English. One of the observed reasons is that noun capitalization in German differs considerably from that in English. Another reason is the smaller number of gazetteers and other linguistic resources available for German. In this paper we present an overview of our approach to Named Entity Recognition and discuss the issues that we have observed while adapting our information extraction system to German.

ABBYY InfoExtractor has already been applied to English and Russian. We evaluated our named entity recognition system for English on MUC-6 corpus and achieved the F-measure of approximately 83% with no prior adjustments. We performed this evaluation ourselves. As for the Russian language, we took part in FactRuEval 2016 competition[1] and showed the best results in the Russian Named Entity Recognition Track with the F-measure of 86.7% (Stepanova et al., 2016; Starostin et al., 2016).

The paper is structured as follows. In Section 2 we review some of the previous works in the field of German NER. In Section 3 an outline of the system architecture is given. In Section 4 we discuss the issues that we have faced in German NER and comment on them. In Section 5 we present performance of our system on GermEval 2014 corpus. Section 6 provides conclusion and discussion of our future work.

2 Related Work

There have been two main approaches to named entity recognition: rule-based and classifier-based. Most of the systems that work with the German language are classifier-based: only four of the systems that took part in either GermEval 2014 or CONLL 2003 tracks used handcrafted rules for named entity recognition (Bobkova et al., 2014; Hermann et al., 2014; Weber and Pötzl, 2014; Watrin et al., 2014), three of these systems used rules in combination with classifier-based approaches. Hatner (2014) reports that the combination of rules and a classifier performed actually worse than the classifier alone. However, Early-

[1] http://www.dialog-21.ru/en/evaluation/

Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, pages 28–33,
Berlin, Germany, August 12, 2016.

Tracks reported that the use of linguistic resources and rules improved the resulting F-measure considerably. NERU was the only system which relied mainly on handcrafted rules, the system was able to achieve the F-measure of 54.55% on the test set. Overall, the systems participating in GermEval track showed F-measures from 37.23% to 76.38%.

3 System Architecture

Our approach rests upon deep linguistic parsing performed by ABBYY Compreno parser. The input of information extraction module is semantic and syntactic structures of text produced by the parser.

3.1 Language model

The language model we use can be described as projective dependency trees. Alternatively, it can be viewed as flat constituent structures where every constituent is built around a word. Dependency arcs have two types of labels, syntactic and semantic ones. Syntactically, our structures are very similar to universal dependency model (Nivre, 2015). Semantic labels include traditional semantic roles (Agent, Instrument, Object, Time etc.) as well as several 'dummy' labels (such as Specification) devoid of semantic content.

Leaves of dependency trees are mostly word forms annotated with lexeme and lexical meaning (similar in spirit to WordNet (Fellbaum, 1998) meanings). Quoted expressions, e.g. "'War and Peace' is a masterpiece", have their own dummy nodes as well. The system is designed as multilingual, which is manifested as follows. Lexical meanings in a given language are leaves in a language-independent hierarchy of meanings. Semantic labels are universal for all languages as well. Syntactic structures across different languages are aligned as much as possible.

The pipeline of the system consists of the following steps. Input text undergoes lexical and morphological analysis, at which stage all possible lexical classes for each word form with all possible grammatical values are suggested. At the next step the syntactic module builds a graph of all possible syntactic and semantic dependencies between the lemmas. At the same time non-tree relations, if any, are checked. Incompatible meanings are gradually filtered, and a set of valid syntactic-semantic trees remains. Then the tree with the best score is selected. This final tree is passed to the ontological rules, which yield an RDF graph. More on the parser is given in (Anisimovich et al., 2012; Goncharova et al., 2015).

Statistics is gathered on parallel texts, and lexical ambiguity in one language is resolved with the help of other languages.

3.2 Rules

The information extraction system can be considered a rule-based one. The input accepted by the information extraction mechanism is a sequence of syntactic-semantic trees described above (one tree per sentence), the output of the mechanism is an RDF graph. A detailed description of our information extraction module can be found in (Starostin et al., 2014). A rule is itself a condition on parse tree or an object condition (i.e. that an object of a certain type must be "linked" to a certain node in a parse tree). Consider an example:

> Die AfD hat gesagt, dass sie über eine vollständige Alternative verhandeln will.
>
> *AfD said that it will carry on negotiations regarding a complete alternative.*

Figure 1 shows a fragment of a parse tree generated for the sentence.

```
$Verb, Predicate: ":TO_SAY_SPEAK_TELL_TALK"
  $Subject, Agent: "ACRONYM"
    $Article: ":ARTICLES"
```

Figure 1: Parse tree fragment

Figure 2 shows an example of a rule that would extract an organization on the fragment "AfD".

```
"COMMUNICATION_AND_SPEECH_ACTIVITY"
    [
      Agent: name "ACRONYM"
    ]
=>
Organization Org (name);
```

Figure 2: Rule

The main advantage of this approach is that the rules become language independent. Thus, we could reuse the set of rules that we use for both

English and Russian for German. However, German capitalization was a challenge: in English and Russian capitalization is a good marker to discriminate between abstract and named entities:

1. relationship between Church and State in the Middle Ages

2. He went to church on Sunday

In German this marker does not work because all nouns are capitalized. So far we have decided to create an extra rule, which checks if the noun has dependent adjectives that are capitalized. If such adjectives can be found, the entity is labelled as named, otherwise, the entity is labeled as abstract. The approach, however, works only for part of the cases and tends to miss out correct named entities (consider the example above).

As we evaluated our system on GermEval corpus, we wrote some additional rules to extract entities which are specific to this corpus, i.e. OTHER, ENTITYderiv and ENTITYpart (Benikova et al., 2014). The rules for extraction entities of the type OTHER included conditions on currency names, bracketed names, urls and so on. To extract ENTITYpart type we added a rule that marked the constituent that had the "Composite" grammeme.

Extracting of ENTITYderiv type of entities is challenging for us, because our system rests upon the hierarchy of semantic meanings and does not preserve information that some word was derived from another. For instance, the words "Deutscher" and "Deutsch" are not connected in the hierarchy. We created a rule that added a "deriv" tag to the names of nationalities and to named entities, which were extracted on adjectives.

We did not extract any gazetteers from the corpus, although we plan to do it in the future.

4 Issues

Extending our system to German, we have faced a number of challenges. In this section we will discuss them and offer solutions where possible.

4.1 Organization names

Organization-denoting expressions can have different syntactic structures.

Syntactic structures can be classified as follows. In the easiest case there is: generic word (*Inc., Ltd., GmbH* etc.) and the name of organization in quotes or italics/bold. See below a text example and our structure of its fragment:

> "Deutsche Post" AG ist eine große Firma.
>
> ist → AG → #BracketedProperName → AG.

2. In the second case, there is only the name of organization in quotes or italics/bold without a generic word. Ex.:

> "Deutsche Post" ist eine große Firma.
>
> *Deutsche Post* ist eine große Firma.
>
> ist → AG → #BracketedProperName → Post.

3. The third case is similar to the first one, but the name of organization does not have quotes or formatting. Ex.:

> Deutsche Post AG ist eine große Firma.
>
> Alpha Versicherung GmbH ist auch eine Firma.

Here we can easily identify only one of the two borders of the named entity, the one where the generic name is. The other border is harder to guess: in the examples above it can be either "Post AG" or "Deutsche Post AG" and "Versicherung GmbH" or "Alpha Versicherung GmbH".

In such cases elements of organization name in English and Russian are normally capitalized, and a dummy node #CapitalizedProperName, an analogue of #BracketedProperName, can be identified on the basis of this capitalization. Ex.:

> I shop at Healthy Soups.
>
> shop → #CapitalizedProperName → Soups → Healthy

In German this does not work. Thus we reworked our system and introduced direct links between the generic word and (the main word of) organization title as such.

Organization name itself can belong to one of the several categories (ranked from easier to more difficult):

1. Proper name present in the dictionary. Ex.: "Raiffeisen Bank".

2. Unknown word, i.e. not present in the dictionary. Ex.: "Gruffalo Bank".

3. Common name present in the dictionary in one or more meanings. Ex.: "Deutsche Versicherung Bank".

In the first case, the name helps building the correct structure because the model takes into account correlation statistics between the semantic class of the child node and semantically labelled link, and the statistics for a proper organization name and the link between it and the generic company name is normally good. In the third case the title is misleading because "Versicherung" (insurance) is a common verb noun, which is not a typical child of a *Name* link. Yet, as we cannot use a dummy node in this case in German, we have created a new semantically labelled link between generic company names and this type of organization titles that would allow for such connection. (This work is now in progress.) In this way German has forced us to face this problem. With only English and Russian, such cases were infrequent and could be ignored.

4. In the most difficult case there is neither generic word nor quotes/formatting. Ex.:

Deutsche Post ist eine große Firma.

In this case identification of both borders and thus the very presence of organization name is troublesome. Yet, from our experience, this type is not less frequent.

4.2 Composites

When analyzing a word form, the parser can have several hypotheses about its lemma. If the word can in theory be split into several known word forms, the parser may try to analyze it as a composite word. While it is generally reasonable for a language so rich in composites as German, it may cause problems when analyzing unknown proper names (not present in the dictionary). The inner structure of proper names, particularly person or location names, can be rather complicated: "Tempelhof", "Schimmelmann" etc., yet such words should be treated as single lemmas. But there are also common nouns that are actual composites and that should be split into parts during the analysis: "Tempelarchitektur" (temple architecture), "Schimmelkultur" (fungus culture). In both cases the parser will have two hypotheses for these words: 1) an unknown word (*Tempelhof*, *Tempelarchitektur*); 2) a composite consisting of several known parts: *Tempel* + *Architektur*; *Tempel* + *Hof*. Thus, choosing the correct hypothesis may be challenging.

4.3 Non-German fragments

We have also observed that sequences of foreign words incorporated in German text can create problems for named entity recognition. If none of the foreign words in a sequence looks like a German word, they will all be interpreted as unknown words, which is the best variant in this situation. But if any of the foreign words in the sequence is homonymous to any known word form in German, the parser is more likely to interpret this word as a known German lexeme rather than an unknown one and build a syntactic structure based on this interpretation. This results in incorrect parsing, impeding NE identification. Example: "Zimbabwe Conservation Task Force". The word "Force" is present in the German dictionary as a town name, and the parser can recognize this word form as a location name instead of an element of a sequence of unknown words that together form the name of an organization. A possible solution here is to detect borders of a foreign language fragment and treat all words within it as one unknown item. Yet, this will not a complete solution because there are cases when a German word is incorporated into a foreign language string, which makes border detection more difficult:

Nutzung von Business-zu-Customer-Beziehungen

Here "zu" is a German word, but "customer" and probably "business" are English fragments.

4.4 Abbreviations

Besides high lexical ambiguity resulting from capitalization of all nouns (ex., a noun "Kraft" can be either a common noun or a family name), German also features heavy use of ambiguous abbreviations. For example, "AG" can stand both for "Aktionsgesellschaft" (joint stock company) or "Auftraggeber" (customer). If an abbreviation has a limited number of interpretations or several most frequent interpretations, it can be placed in the dictionary in several meanings. And then disambiguation can be helped by the context: if "AG" is an object of "gründen", the statistical score for the combination of "gründen" + company will be better than for "gründen" + person. But if an abbreviation has an unlimited number of meanings that are relatively equally frequent, it is impossible to include all of them into the dictionary, and it makes little sense to include only some of meanings. In such cases the abbreviation is not included

it into the dictionary, and is interpreted as an unknown acronym at the parsing stage.

4.5 Quotation marks

If a fragment of text has quotation marks around it, there is a high probability that it denotes a named entity such as an organization name. However, quotation marks can also be used in some other contexts such as quoting: "'Excuse me', the boy said". For the languages we have previously dealt with we can look up for speech verbs in order to disambiguate between the two usages of quotation marks. However, for German the presence of a speech verb is not always obligatory since there exists a special verb form (Konjunktiv I) designating reported speech. Thus for a fragment in a long sequence of Konjunktiv I sentences quotation marks can signify either reported speech or a named entity, even if there is not any speech verb nearby.

5 Evaluation

Precision (%)	**Recall (%)**	**FB1 (%)**
45.22	44.87	45.04

Table 1: Overall results. Strict metric

We have tested our system on GermEval 2014 corpus using the evaluation script provided by the organizers. Overall results of strict evaluation are presented in Table 1, results of strict evaluation by categories are given in Tables 2, 3, 4, 5. Predictably, extraction of organizations has turned out to be the most challenging task for us, due to the parsing problems mentioned above. We hope that the implementation of changes to the parser suggested in Section 4 will improve the quality of parse trees and entity extraction. Enrichment of Organization gazetteer is also likely to help.

	Precision	**Recall**	**FB1**
PER (outer)	63.02	65.68	64.33
PER (inner)	30.43	20.59	24.56
PERpart (outer)	0	0	0
PERpart(inner)	0	0	0
PERderiv(outer)	0	0	0
PERderiv (inner)	0	0	0

Table 2: Results for PER (in %)

	Precision	**Recall**	**FB1**
LOC (outer)	78.5	49.28	60.55
LOC (inner)	15.38	9.43	11.7
LOCpart (outer)	50.94	51.92	51.43
LOCpart(inner)	8.33	100	15.38
LOCderiv(outer)	88.46	39.15	54.28
LOCderiv (inner)	28.57	17.02	21.33

Table 3: Results for LOC (in %)

	Precision	**Recall**	**FB1**
ORG (outer)	18.7	32.46	23.73
ORG (inner)	0	0	0
ORGpart (outer)	77.14	29.67	42.86
ORGpart(inner)	0	0	0
ORGderiv(outer)	0	0	0
ORGderiv (inner)	0	0	0

Table 4: Results for ORG (in %)

The results of "Person" type extraction are quite unexpected, first of all, because recall-better-than-precision pattern is not typical for us. These results require further analysis, which is beyond the scope of this paper.

The precision of "Location" type extraction is relatively high and we believe that further Location gazetteer enrichment will improve the recall considerably.

6 Conclusions

In this paper we presented our rule-based approach to named entity recognition for the German language. The approach has previously been applied to Russian and English languages and has shown good results. However, we have found out that several changes should be made to the parser to obtain better results on German. We have evaluated our system on GermEval 2014 corpus and presented the results as well as the analysis of problems we as a parser- and rule-based system have faced. In the nearest future we plan to implement several changes to the parser as well as enrich organizations and locations gazetteers in order to obtain better results for German.

References

K.V. Anisimovich, K.Ju. Druzhkin, F.R. Minlos, M.A. Petrova, V.P. Selegey, and K.A. Zuev. 2012. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In *Computational*

	Precision	Recall	FB1
OTH (outer)	39.34	39.78	39.56
OTH (inner)	0	0	0
OTHpart (outer)	12.9	22.22	16.33
OTHpart(inner)	0	0	0
OTHderiv(outer)	33.33	50	40
OTHderiv (inner)	0	0	0

Table 5: Results for OTH (in %)

Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2012), number 11, pages 91–103.

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

Yulia Bobkova, Andreas Scholz, Tetiana Teplynska, and Desislava Zhekova. 2014. HATNER: Nested Named Entitiy Recognition for German. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany*.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, speech, and communication. MIT Press.

M.B. Goncharova, E.A. Kozlova, A.V. Pasyukov, R.V. Garashchuk, and V.P. Selegey. 2015. Model-based WSA as means of new language integration into a multilingual lexical-semantic database with interlingua. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2015)*, volume 1, pages 169–182.

Ralph Grishman. 2003. Information Extraction. *The Handbook of Computational Linguistics and Natural Language Processing*, pages 515–530.

Martin Hermann, Michael Hochleitner, Sarah Kellner, Simon Preissner, and Desislava Zhekova. 2014. Nessy: A Hybrid Approach to Named Entity Recognition for German. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany*.

Joakim Nivre, 2015. *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I*, chapter Towards a Universal Grammar for Natural Language Processing, pages 3–16. Springer International Publishing, Cham.

A.S. Starostin, I.M. Smurov, and M.E. Stepanova. 2014. A Production System for Information Extraction Based on Complete Syntactic-semantic Analysis. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2014)*, pages 659–667.

A.S. Starostin, V.V. Bocharov, S.V. Alexeeva, A.A. Bodrova, A.S. Chuchunkov, S.S. Dzhumaev, I.V. Efimenko, D.V. Granovsky, V.F. Khoroshevsky, I.V. Krylova, M.A. Nikolaeva, I.M. Smurov, and S.Y. Toldova. 2016. FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian. In *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference «Dialogue» (2016)*, number 15, pages 702–720.

M.E. Stepanova, E.A Budnikov, A.N. Chelombeeva, P.V. Matavina, and D.A. Skorinkin. 2016. Information Extraction Based on Deep Syntactic-Semantic Analysis. In *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference «Dialogue» (2016)*, number 15, pages 721–732.

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 seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Patrick Watrin, Louis de Viron, Denis Lebailly, Matthieu Constant, and Stephanie Weiser. 2014. Named Entity Recognition for German Using Conditional Random Fields and Linguistic Resources. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany*.

Daniel Weber and Josef Pötzl. 2014. NERU: Named Entity Recognition for German. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany*.