Named Entity Recognition for German Using Conditional Random Fields and Linguistic Resources

Patrick Watrin
EarlyTracks SA
Louvain-la-Neuve
Belgium

Louis de Viron
EarlyTracks SA
Louvain-la-Neuve
Belgium

Denis Lebailly
EarlyTracks SA
Louvain-la-Neuve
Belgium

Matthieu Constant
Université Paris-Est
Marne-la-Vallée
France

Stéphanie Weiser
EarlyTracks SA
Louvain-la-Neuve
Belgium

Abstract

This paper presents a Named Entity Recognition system for German based on Conditional Random Fields. The model also includes language-independant features and features computed form large coverage lexical resources. Along side the results themselves, we show that by adding linguistic resources to a probabilistic model, the results improve significantly.[1]

1 Introduction

These last few years, models based on Conditional Random Fields (CRF) have shown interesting achievements for Named Entity Recognition (NER) tasks. However, most of the experiences carried out also show a lack of lexical coverage. To counterbalance this lack, two main kinds of strategies have been designed: the use of gazetteers and of clustering techniques. Both lead to a significant improvement of the results. For a review of these techniques, see (Tkachenko and Simanovsky, 2012). In the work presented here, we have opted for a more linguistic approach, close to the gazetteers: we included lexical resources as new features for a model based on CRF and measured their impact. This kind of approach has already been proven successful for a Part-of-Speech tagger by Constant and Sigogne (2011).

This work took place in the framework of the GermEval Named Entity Recognition Shared Task[2] and is therefore applied to German. However, this approach has already been implemented for English, French and Dutch.

The characteristics of the GermEval tagset are presented in section 2. In section 3 is described our system for named entity recognition based on CRF and the adaptations we suggest for this kind of model. Section 4 presents the linguistic resources we added. Finally, our experiments and the results we obtained are presented in section 5.

2 GermEval Characteristics

2.1 Tagset

The tagset defined for the GermEval shared task (Benikova et al., 2014b) consists of four main classes. The class *Person* (1) includes person names but also nicknames and fictional characters names. The class *Organisation* (2) contains all kind of organisations, companies, and also festivals, music bands, etc. The *Location* class (3) is made for all kind of places: cities, countries, planets, churches, etc. The class *Other* (4), is the widest one as it includes a large variety of items: movies and books titles, languages, websites, market indexes etc.

These four main classes have two subclasses each: *deriv* and *part* (LOCderiv, OTHderiv, PERderiv, ORGderiv, LOCpart, OTHpart, PERpart, ORGpart). The *deriv* one is used to tag items that are derived from named entities. Most of the times they are adjectives such as *asiatischen* (*asian*). The *part* one is made for named entities that are included in a larger token, in compound

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

forms. As the German language is agglutinative, this happens quite often, without diacritical marks (*Bundesligaspiele*) .

2.2 Entities Embedding

Another specificity of the GermEval task is that nested entities are allowed. For example, the film title *Shakespeare in Love* must be tagged OTH but it must also contain an inner tag PER for *Shakespeare*. The tagger we developed therefore needed to be adapted to include this particularity.

3 Conditional Random Fields

As presented in (Lafferty, 2001), CRF define a framework for building probabilistic models that are able to split and tag sequences of data. Since they exist, CRF have lead to many works in Natural Language Processing (*e.g.* Constant and Sigogne (2011)) and more specifically in NER (*e.g.* Finkel et al. (2005) and Klein et al. (2003)).

3.1 Standard Approach

In practice, the probability of a sequence of labels depends on a set of features that are representative of the observation sequence (*i.e.* the tokens). Most of these features are language-independent and limited to local observations. CRF systems generally use a set of generic features, that we present in table 1.

These features are language-independent. However, some characteristics of the language can be in conflict with one or more features. For example, the feature that represents the presence or absence of a capital letter is less pertinent for German – where many words begin with a capital letter – than for other languages.

3.2 Hybrid approach

The statistical models are limited to their training corpus and therefore their lexical coverage is often not large enough. Many works have tried to compensate for this weak coverage to help the classification of unseen words. Faruqui and Padó (2010) and Finkel et al. (2005) suggest to add a distributional similarity feature trained on a very large corpus. The hypothesis of a strong correlation between the terms of a same distributional class is the basis of this feature. Faruqui and Padó (2010) show very interesting results for German,

Feature	Explanation
$\dots w_{-1} w_0 w_1 \dots$	tokens
lowercase	token in lowercase
shape	token in a Xx form
isCapitalized	is the token capitalized?
prefix(n)	n first letters of the token (1 to 4)
suffix(n)	n last letters of the token (1 to 4)
hasHyphen	does the token contain hyphens?
hasDigit	does the token contain digits?
allUppercase	is the token uppercase only?

Table 1: Language-independent features

Feature	Explanation
pos	Token PoS-tag
containsFeature(x)	Does the token belong to the semantic class x?
sac	Semantic ambiguity class *i.e.* all possible classes for the token

Table 2: Lexical features

with an increase of 6-7% for precision and 12-13% for recall.

In parallel to this method, other studies suggest the use of external lexical resources (Nadeau and Sekine, 2007; Kazama and Torisawa, 2007; Constant and Sigogne, 2011). Indeed, a simple way to decide if a sequence of tokens corresponds to a named entity is to check in a dictionary. Today, many multilingual encyclopedic resources are available online and facilitate the construction of these dictionaries (DBPedia, Yago, FreeBase...). To integrate the information of these dictionaries in our model, we have defined 3 types of features, that are presented in table 2, where the classes correspond to the different classes of the GermEval tagset. The linguistic resources we used and their impact are presented in section 4 and 5.

4 Adding Linguistic Resources to the Model

The linguistic resources we used are divided into two types: dictionaries (word lists including morphological data) and grammars made of transducers created with the software Unitex[3]. The objective of these resources is to counterbalance the lack of lexical coverage due to the training corpus.

4.1 Dictionaries

We use two kinds of dictionaries. First, we use a general language dictionary of German, that we adapted from the resources created by Daniel Naber[4], using Morphy[5]. It contains lemmas, in-

[3]http://www-igm.univ-mlv.fr/ unitex/

[4]http://danielnaber.de/morphologie/

[5]http://www.wolfganglezius.de/doku.php?id=cl:morphy

Dictionary	Nb. entries
Morphy	749.212
Persons	1.266.390
Places	200.392
Places *deriv*	2.642
Organisations	648.273
Others	2.617.902

Table 3: Number of entries by dictionary

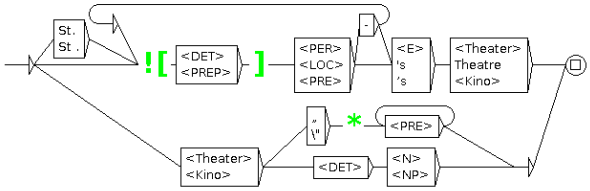

Figure 1: Transducer for matching Theatres such as *Berlin's Theater*

flectional variations and part-of-speech tags. The second type of dictionaries are useful for data that can be fully listed, such as countries for example. We created dictionaries for most of the entities that needed to be extracted using free resources such as Freebase[6]. We also created dictionaries for the *deriv* entities to follow the GermEval guidelines. Table 3 gives the number of entries for each dictionary.

4.2 Local Grammars

Local grammars that we created using Unitex transducers (Paumier, 2003) are efficient for entities that can vary more or are difficult to fully list. For example, a grammar can be defined to describe all kind of universities or theatres names, as it is shown in the figure 1.

These grammars can also handle German specificities such as concatenation of words. Some specific transducers have been made to cover the *part* entities (when an entity is included in a larger token as *Hamiltonoperator* for example). Our grammar library contains 9 main graphs (one for each category, one for each *deriv* category and one for all *part* entities) and around 20 subgraphs.

5 Experiments and Results

In this section, we present our experiments and put our results in balance with those of the other participants to the GermEval task. The table 4 shows the results obtained by all the systems that have participated to the GermEval 2014 Shared Task. We rank number 4, out of 11 models competing. The table 5 presents the results we obtained with two models: the simple CRF model and the model enriched by the lexical resources. The four metrics we use are explained by Benikova et al. (2014a).

[6]http://www.freebase.com/

Model	Precision	Recall	F_1
ExB	78.07	**74.75**	**76.38**
UKP	79.54	71.1	75.09
MoSTNER	79.20	65.31	71.59
EarlyTracks	79.92	64.65	71.48
PLsNER	76.76	66.16	71.06
DRIM	76.71	63.25	69.33
mXS	**80.62**	50.89	62.39
Nessy	63.57	54.65	58.78
NERU	62.57	48.35	54.55
HATNER	65.62	43.21	52.11
BECREATIVE	40.14	34.71	37.23
Median	76.71	63.25	69.33

Table 4: Results obtained by all the participants to the GermEval 2014 NER Shared Task (Strict Metric)

Model	Metric	Precision	Recall	F_1
CRF	M-Strict	77.14	61.56	68.47
	M-Loose	77.89	62.15	69.14
	M-Outer	77.57	63.89	70.07
	M-Inner	68.38	33.59	45.05
CRF+LING	M-Strict	79.92	64.65	71.48
	M-Loose	80.55	65.16	72.04
	M-Outer	80.44	66.98	73.10
	M-Inner	70.00	36.70	48.15

Table 5: Impact of adding linguistic resources to a CRF model

Our results are interesting because they show that by adding lexical resources and grammars as new features to our model, the results are improved by 3.01% for the strict metric, which is significant. This number should keep rising while the resources increase.

Table 6 shows the results obtained for each outer class and each inner class and the improvement made with lexical resources. As the class OTH is very versatile, it obtains less good results than the other classes. Furthermore the entity classes *part* and *deriv*, as well as the inner-classes, are less represented in the training set and therefore also reach less good results. The classes ORG, LOC and PER which can rely on external lexical resources obtain better results.

Entity	M-Outer			M-Inner		
	Occ.	**CRF**	**CRF+**	**Occ.**	**CRF**	**CRF+**
PER	1639	76.63	80.20	82	4.49	10.87
ORG	1150	63.54	66.34	41	8.51	8.89
LOC	1706	75.54	79.36	210	56.09	56.99
OTH	697	50.51	52.46	7	0.00	0.00
PERpart	44	16.00	12.24	4	40.00	40.00
ORGpart	172	56.39	58.61	1	0.00	0.00
LOCpart	109	55.49	54.97	5	0.00	0.00
OTHpart	42	16.33	25.00	1	0	0
PERderiv	11	16.67	0.00	4	0.00	0.00
ORGderiv	8	22.22	22.22	1	0	0
LOCderiv	561	78.31	80.15	159	54.12	59.46
OTHderiv	39	47.46	47.62	0	0	0
Global	**6178**	**70.07**	**73.10**	**515**	**45.05**	**48.15**

Table 6: For each outer and inner entity: number of occurrences in the evaluation corpus and F_1 for the simple CRF and the enriched CRF

6 Conclusion

In this paper, we presented our Named Entity Recognizer for German. We achieve a global F-measure of 71.48% on the GermEval evaluation corpus with the complete tagset. In parallel, we evaluated the impact of using linguistic resources as an input to the statistical model: it improves the results by 3.01% for the strict metric. As a next step, to increase this impact, the dictionaries, that are still in an early stage, should be enhanced: they have been automatically gathered and could use a manual correction to avoid erroneous entries. In addition, we will try to find other precise dictionaries and enlarge the grammars to improve the recall, in particular to cover more completely the *Others* class.

Another possible way of improving our system would be to combine our linguistic approach to a clustering strategy.

References

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.

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).

Matthieu Constant and Anthony Sigogne. 2011. Mwu-aware part-of-speech tagging with a crf model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, pages 49–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Jun'ichi Kazama and Kentaro 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. Citeseer.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 180–183. Edmonton, Canada.

John Lafferty. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289. Morgan Kaufmann.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Sébastien Paumier. 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Université de Marne-la-Vallée, July.

Maksim Tkachenko and Andrey Simanovsky. 2012. Named entity recognition: Exploring features. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 118–127. ÖGAI, September. Main track: oral presentations.