GermEval 2014 Named Entity Recognition Shared Task: Companion Paper[*]

Darina Benikova[*] , Chris Biemann[*], Max Kisselew[†], Sebastian Padó[†]
[*] Language Technology, TU Darmstadt, Germany
[†] Institute for Natural Language Processing, Universität Stuttgart, Germany
{darina.benikova@stud,biem@cs}.tu-darmstadt.de
{max.kisselew,pado}@ims.uni-stuttgart.de

Abstract

This paper describes the GermEval 2014 Named Entity Recognition (NER) Shared Task workshop at KONVENS. It provides background information on the motivation of this task, the data-set, the evaluation method, and an overview of the participating systems, followed by a discussion of their results. In contrast to previous NER tasks, the GermEval 2014 edition uses an extended tagset to account for derivatives of names and tokens that contain name parts. Further, nested named entities had to be predicted, i.e. names that contain other names. The eleven participating teams employed a wide range of techniques in their systems. The most successful systems used state-of-the-art machine learning methods, combined with some knowledge-based features in hybrid systems.

1 Introduction

Named Entity Recognition (NER or NERC) is the identification and classification of proper names in running text. NER is used in information extraction, question answering, automatic translation, data mining, speech processing and biomedical science (Jurafsky and Martin, 2000).

The starting point for this shared task is the observation that the level of performance of NER for German is still considerably below the level for English although German is a well-researched language. At least part of the reason is that in English, capitalization is an important feature in detecting Named Entities (NEs). In contrast, German capitalizes not only proper names, but all nouns, which makes the capitalization feature much less informative. At the same time, adjectives derived from NEs, which arguably count as NEs themselves, such as *englisch* ("English"), are not capitalized in German, in line with "normal" adjectives. Finally, a challenge in German is compounding, which allows to concatenate named entities and common nouns into single-token compounds.

This paper reports on a shared task on Named Enitity Recognition (NER) for German held in conjunction with KONVENS 2014. Compared to the only well-known earlier shared task for German NER held more than ten years ago in the context of CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003), our shared task corpus introduces two substantial extensions:

Fine-grained labels indicating NER subtypes. German morphology is comparatively productive (at least when compared to English). There is a considerable amount of word formation through both overt (non-zero) derivation and compounding, in particular for nouns. This gives rise to morphologically complex words that are not identical to, but stand in a direct relation to, Named Entities. The Shared Task corpus treats these as NE instances but marks them as special subtypes by introducing two fine-grained labels: `-deriv` marks derivations from NEs such as the previously mentioned *englisch* ("English"), and `-part` marks compounds including a NE as a subsequence *deutschlandweit* ("Germany-wide").

Embedded markables. Almost all extant corpora with Named Entity annotation assume that NE annotation is "flat", that is, each word in the text can form part of at most one NE chunk. Clearly, this is an oversimplification. Consider the noun phase *Technische Universität Darmstadt* ("Technical University (of) Darmstadt"). It denotes an organization (label `ORG`), but also holds another NE, *Darmstadt*, which is a location (label `LOC`). To account for such cases, the Shared Task corpus is annotated with two levels of Named Entities. It captures at least one level of smaller NEs being embedded in larger NEs.

In summary, we distinguish between 12 classes of NEs: four main classes `PER`son, `LOC`ation, `ORG`anisation, and `OTH`er and their subclasses, annotated at two levels ("inner" and "outer" chunks). The challenge of this setup is that while it technically still allows a simple classification approach it introduces a recursive structure that calls for the application of more general machine learning or other automatically classifying methods that go beyond plain sequence tagging.

2 Dataset

The data used for the GermEval 2014 NER Shared Task builds on the dataset annotated by (Benikova et al., 2014)[1]. In this dataset, sentences taken from German Wikipedia articles and online news were used as a collection of citations, then annotated according to extended NoSta-D guidelines and eventually distributed under the CC-BY license[2].

As already described above, those guidelines use four main categories with sub-structure and nesting. The dataset is distributed contains overall more than 31,000 sentences with over 590,000 tokens. Those were divided in the following way: the training set consists of 24,000 sentences, the development set of 2,200 sentences and the test set of 5,100 sentences. The test set labels were not available to the participants until after the deadline. The distribution of the categories over the whole dataset is shown in Table 1. Care was taken to ensure the even dispersion of the categories in the subsets.

Class	All	Nested[3]
Location	12,204	1,454
Location deriv	4,412	808
Location part	713	39
Person	10,517	488
Person deriv	95	20
Person part	275	29
Organization	7,182	281
Organization deriv	56	4
Organization part	1,077	9
Other	4,047	57
Other deriv	294	3
Other part	252	2
Total	41,124	3,194

Table 1: Distribution of classes in the entire dataset of 31,300 sentences. Counts differ slightly fron what was reported in (Benikova et al., 2014) due to correction of inconsistencies in June 2014.

The entire dataset contains over 41,000 NEs, about 7.8% of them embedded in other NEs (*nested* NEs), about 11.8% are derivations (*deriv*) and about 5.6% are parts of NEs concatenated with other words (*part*).

The tab-separated format used in this dataset is similar to the CoNLL-Format. As illustrated in Table 2, the format used in the dataset additionally contains token numbers per sentence in the first column and a comment line indicating source and data before each sentence. The second column contains the tokens. The third column encodes the outer NE spans, the fourth column the inner ones. The BIO-scheme was used in order to encode the NE spans. In our challenge, further nested columns were not considered.

3 Evaluation method

We defined four metrics for the shared task, but only one was used for the final evaluation ("official metric"). The others were used in order to gain more insight into the distinctions between the

[1] The dataset was updated for this task to fix some inconsistencies.

[2] This license allows to distribute, alter and mix the data in any possible way and to use it for any purpose, including commercial ones (see `https://creativecommons.org/licenses/by/3.0/de/`).

[3] These numbers include all occurrences on the second level, regardless of the class of the first level NE

#	http://de.wikipedia.org/wiki/Manfred_Korfmann		
1	Aufgrund	O	O
2	seiner	O	O
3	Initiative	O	O
4	fand	O	O
5	2001/2002	O	O
6	in	O	O
7	Stuttgart	B-LOC	O
8	,	O	O
9	Braunschweig	B-LOC	O
10	und	O	O
11	Bonn	B-LOC	O
12	eine	O	O
13	große	O	O
14	und	O	O
15	publizistisch	O	O
16	vielbeachtete	O	O
17	Troia-Ausstellung	B-LOCpart	O
18	statt	O	O
19	,	O	O
20	„	O	O
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
26	”	O	O
27	.	O	O

Table 2: Data format illustration. The example sentence contains five named entities: the locations “Stuttgart”, “Braunschweig” and “Bonn”, the noun including a location part “Troia”-Ausstellung, and the title of the event, “Troia - Traum und Wirklichkeit”, which contains the embedded location “Troia”. (Benikova et al., 2014)

different systems.

We follow the pattern of previous evaluation in NER shared tasks using non-recursive data, which used the standard precision, recall and F_1 score metrics, using each individual markable as a datapoint in the P/R calculation. Let P denote the set of NE chunks predicted by a model and G the set of gold standard chunks. Precision, Recall, and F_1 are usually computed on the basis of of true positives and false positives and negatives, defined by set theoretic operations, e.g. $TP = P \cap G$ which in turn build on the definition of matches between predicted chunks and gold standard chunks. Normally, strict match is assumed: $p == g$ iff $label(p) = label(g)$ and $span(p) = span(g)$.

We would like to retain precision and recall as evaluation measures but need to redefine their computation to account for the nested nature of the data. Let P_1 and G_1 denote the set of all “first-level”/“outer” NEs (and P_2 and G_2 denote the set of all “second-level”/“inner” NEs in the predictions and in the gold standard, respectively.

3.1 Metric 1: Strict, Combined Evaluation (Official Metric)

The most straightforward evaluation treats first-level and second-level NEs individually and independently. This can be modeled by combining G and P across levels, but taking the level into account in the match definition:

$$\begin{aligned} P &= P_1 \cup P_2 \\ G &= G_1 \cup G_2 \\ p == g \text{ iff } label(p) &= label(g) \text{ and} \\ span(p) &= span(g) \text{ and} \\ level(p) &= level(g) \end{aligned}$$

Thus, this metric distinguishes all 12 labels (4 NE types, each in base, deriv and part varieties) and treats all markables on a par. It is used to determine the overall ranking of the systems in this challenge.

3.2 Metric 2: Loose, Combined Evaluation

Metric 2 again treats each NE individually but we collapse the label subtypes (base, deriv, part) so that a match on the base NE class is sufficient. For example, PER matches PERderiv:

$$\begin{aligned} P &= P_1 \cup P_2 \\ G &= G_1 \cup G_2 \\ p == g \text{ iff } baseLabel(p) &= baseLabel(g) \text{ and} \\ span(p) &= span(g) \text{ and} \\ level(p) &= level(g) \end{aligned}$$

This metric is useful to quantify the quality of systems at a coarse-grained level. It also makes the scores better comparable to previous NER evaluations, which have mostly used only four labels.

3.3 Metric 3: Strict, Separate Evaluation

Finally, this evaluation computes two sets of P/R/F1 values, one for G_1/P_1 and one for G_2/P_2. This metric considers the first-level and second-level markables separately which allows us to see

System ID	Institution
Nessy	LMU Munich
NERU	LMU Munich
HATNER	LMU Munich
DRIM	LMU Munich
ExB	ExB GmbH
BECREATIVE	LMU Munich
PLsNER	TU Darmstadt
mXS	University of Tours
MoSTNER	Marmara University
Earlytracks	EarlyTracks S.A.
UKP	TU Darmstadt

Table 3: Participants of the GermEval 2014 shared task.

System	HR	GQ	NB	ME	SVM	CRF	NN
NERU	X						
Nessy	X		X				
HATNER	X			X			
DRIM					X		
EarlyTracks	X	X				X	
ExB				X[4]		X	
BECREATIVE		X	X				
PLsNER							X
mXS				X			
MoSTNER						X	
UKP							X

Table 4: Methods used by participating systems
HR = handcrafted rules, GQ = gazetteer queries, NB = Naïve Bayes, ME = Maximum Entropy, SVM = Support Vector Machine, CRF = Conditional Random Field and NN = Neural Networks/Word Embeddings

how well systems do on first-level vs. second-level markables individually. It uses strict matching of labels, and thus uses exactly the traditional match definition (cf. the beginning of Section 3).

4 Participating systems

11 teams listed in Table 3 participated in the GermEval 2014 challenge. In the first subsection their general approaches will be discussed. The second subsection will present the variety of features that was used by the systems. Although many teams experimented with other methods and features, only those used by the respective final system will be mentioned here.

4.1 Methods used by the participants

Table 4 shows the different approaches the teams used for their NER systems. The first two columns describe handcrafted rules or gazetteer queries as an individual processing step, when not used merely as a feature in the overall system.

The NERU (Weber and Pötzl, 2014) system uses handcrafted rules made individually for the classes PERson, LOCation and ORGanization. Hence it is the only participating system not using any machine learning (ML).

The table shows that four systems (Nessy (Hermann et al., 2014), HATNER (Bobkova et al., 2014), EarlyTracks (Watrin et al., 2014), and BECREATIVE (Dreer et al., 2014)) use a hybrid approach, combining a ML method with handcrafted rules or gazetteer queries. All three systems use ML in the first step of their classification and some sort of gazetteer look-up as a post-processing step. Both Nessy and BECREATIVE use NB in the first step of their system, whereas HATNER uses ME. Nessy and HATNER do so only for the part and deriv classification using handcrafted rules.

The goal of the ExB group (Hänig et al., 2014) was to build a system that runs efficiently on mobile devices. They experimented with different ML mechanisms. The result of their experiment was that the system that found more correct NEs made use of CRFs, but recommend to use ME in situations where resources are limited.

All other groups decided for one ML mechanism only. DRIM (Capsamun et al., 2014) uses SVM, ExB Group, and MoSTNER (Schüller, 2014) use CRF, and PLsNER (Nam, 2014) and UKP (Reimers et al., 2014) use NN.

4.2 Features used by the participating systems

Table 5 displays the types of features used by the participating systems. As NERU used gazetteers for its handwritten rules, it made no use of any other features. As shown, all systems except PLsNER made use of gazetteers and POS-tags.

5 Discussion of results

This section provides and discusses the results of the submitted systems.

5.1 Analysis by official metric (M1)

Table 6 shows the results of the systems in terms of M1, the official metric. For the sake of clarity, we

[4] More efficient, but lower prediction quality than CRF

System	G	POS	tok	NE-n	cap	NE	lem	1st	last	tok-n	#span	POS-n	char	WS	KW	SeC	SiC	WE
NERU	X																	
Nessy	X	X	X	X	X	X	X	X	X	X	X							
HATNER	X	X	X				X			X		X						
DRIM	X	X	X				X			X			X	X	X			
EarlyTracks	X	X	X		X					X		X	X	X		X		
ExB Group	X	X	X										X	X			X	
BECREATIVE	X	X	X		X		X			X								
PLsNER			X		X	X												X
mXS	X	X	X				X											
MoSTNER	X	X	X							X		X	X				X	
UKP	X	X			X								X					X

Table 5: Features used by systems. G = gazetteers, POS = part of speech, tok = token, NE-n = NE n-gram, cap = capitalization, lem = lemma, 1st = first word in span, last = last word in span, tok-n = token n-gram, #span = number of tokens in span, POS-n = POS n-gram, char = character-level, including affixes, n-grams, decompounding, WS = word shape, KW=keywords, SeC = semantic class, SiC = similarity clusters, WE = word embeddings

only show the best run submitted for each system, since our analysis has found that the within-system variance across runs is quite small compared to the between-system variance. The table is sorted according to F_1 measure.

It is clearly visible that the systems fall into three tiers: one top tier (ExB, UKP) with F-Scores between 75 and 77; a middle tier (PLsNER, MoSTNER, Earlystracks, DRIM) with F-Scores between 69 and 72; and a third tier with lower F-Scores.

The overall winner is the ExB system. Its victory is mostly due to its excellent recall of almost 4 points higher than that of the next-best system, while its precision is close to, albeit above, the median. Overall, all systems have a considerably higher precision that recall. We interpret this as an indication of the important role of successful *generalization* from the training data to novel, potentially different test data. The systems that were most successful in this generalization were the overall most successful systems in the shared task. Conversely, the system with the highest precision, mXS, does not fare well overall precisely due to its comparatively low recall.

System	Precision	Recall	F_1
ExB	78.07	**74.75**	**76.38**
UKP	79.54	71.10	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 6: Precision, Recall, and F_1 for Metric 1 on the test set (official ranking)

Impact of Methods. Following up on the analysis from Section 4.1, we observe that purely rule-based systems and systems relying heavily on gazetteer queries could not reach competitive performance. In line with general trends in the field, it seems to be beneficial to rather plug in rules, lists and language-specific extractors as features in a machine learning framework than using them verbatim. As for machine learning methods, simple classification approaches that do not exploit information about interdependencies among datapoints are substantially outperformed by CRFs and Neural Networks. See (Hänig et al., 2014) for a direct comparison between ME and CRF using the same features.

Impact of features. Building on the results of Section 4.2, we observe that the three best systems have a comparatively small overlap in features: their intersection contains gazetteer-based, POS-level and character-level features. While gazetteers and parts of speech are used by nearly all the participating systems, the character-level features warrant further exploration. The best system, ExB, used several character query-based features in order to find sequences that are characteristic for NE classes, e.g. *-stadt*, *-hausen* or *-ingen*, which are typical endings for German cities. The

System	Precision	Recall	F_1
ExB	78.85	**75.50**	**77.14**
UKP	80.41	71.88	75.91
PLsNER	78.09	67.31	72.30
MoSTNER	79.94	65.92	72.26
Earlytracks	80.55	65.16	72.04
DRIM	77.53	63.92	70.07
mXS	**81.21**	51.26	62.85
Nessy	64.34	55.31	59.48
NERU	63.61	49.16	55.46
HATNER	66.19	43.58	52.56
BECREATIVE	40.78	35.26	37.82

Table 7: Precision, Recall, and F_1 for Metric 2 (subtypes *base*, *deriv* and *part* collapsed

MoSTNER system used Morphisto (Schmid et al., 2004; Zielinski and Simon, 2008) in order to divide tokens into morphological units at character level, which also may have categorized NE specific affixes. These morphological features can be understood as contributing to the generalization aspect outlined above.

The same is true for the use of *semantic* generalization features, which also can be found in different realizations in each of the three best system. Each used at least one high-level semantic feature, such as *Similarity Clusters* or *Word Embeddings*, that were rarely used by other systems. These features are computed in an unsupervised fashion on large corpora and alleviate sparsity by informing the system about words not found in the training set via their similarity to known words – be it as clusters of the vocabulary (MoSTNER, ExB) or vector representations (UKP, PLsNER). The use of simple semantic generalization to improve recall for NER was demonstrated in previous work (Biemann et al., 2007; Finkel and Manning, 2009; Faruqui and Padó, 2010).

5.2 Analysis by “loose metric” (M2)

Table 7 shows the evaluation results for the Metric 2 which does not distinguish between label subtypes.

Our main observation regarding Metric 2 is that the results are very similar to Metric 1. The three tiers can be identified exactly as for Metric 1, and the ordering in Tiers 1 and 3 is in fact identical. The only reordering takes place in Tier 2, where the differences among systems are so small (<.5% F_1) that this is not surprising. In absolute terms, systems typically do between .5% and 1% F-Score better on M2 than on M1, an improvement equally spread between higher precision and recall scores. Our conclusion is that the subtypes do not constitute a major challenge in the data.

Given that the M2 (four-class) results are most comparable to previous work on four-class NER, it is interesting to note that the best results of this challenge are quite close to the best reported results on the other prominent German dataset, the CoNLL 2003 newswire dataset. It is a question of further work to what extent this is a glass ceiling effect connected to, e.g., annotation reliability.

5.3 Per-Level Analysis (M3)

Finally, Table 8 shows the results according to Metric 3, that is, separately for inner and outer level NEs.

Across all systems, we see a noticeably worse performance on second-level NEs: the best F_1 on first-level NEs is 79, the best one on second-level NEs is 49. The more general observation is that first- and second-level NEs behave substantially differently. On first-level NEs, precision and recall are fairly balanced for most systems, with a somewhat higher precision. This is reflected in the maximum values reached: 82 points precision and 77 points recall, respectively. On second-level NEs, precision tends to be much higher than recall for many systems, often twice as high or even more. The maximum values obtained are 70 points precision and 41 points recall.

Another interesting finding is that the overall best system, ExB, is the best system for first-level NEs by a margin of over 2% F_1 (79% vs. 77%). In contrast, it is merely the median system on second-level NEs (43%) and performs more than five points F_1 below the best system, UKP (49%). Among all systems, UKP performs most consistently across first- and second-level NEs, obtaining second place on both levels. On the second level, is closely pursued by the Earlytracks system which shows a very high precision on second-level NEs (70%) but is hampered by a low recall (37%), resulting on an overall F-Score of 48%.

It is an open question for future analysis to what extent the large differences between first-

	First-level NEs			Second-level NEs		
System	**Precision**	**Recall**	**F$_1$**	**Precision**	**Recall**	**F$_1$**
ExB	80.67	**77.55**	**79.08**	45.20	41.17	43.09
UKP	79.90	74.13	76.91	58.74	**41.75**	**48.81**
MoSTNER	79.71	67.74	73.24	69.14	36.12	47.45
Earlytracks	80.44	66.98	73.10	**70.00**	36.70	48.15
PLsNER	77.93	68.52	72.92	57.86	37.86	45.77
DRIM	77.27	65.93	71.15	64.78	31.07	41.99
mXS	**81.90**	53.63	64.81	51.67	18.06	26.76
Nessy	64.83	56.93	60.62	42.86	27.38	33.41
NERU	63.67	51.33	56.84	33.85	12.62	18.39
HATNER	72.88	44.14	54.98	24.81	32.04	27.97
BECREATIVE	40.14	37.60	38.83	0	0	0

Table 8: Precision, Recall and F_1 for Metric 3, computed separately for first-level NEs and second-level NEs. Systems ranked according to F_1 on first-level NEs.

and second-level NEs reflect actual differences in difficulty (i.e., embedded NEs are more difficult to capture) and to what extent they are simply a result of the substantially smaller number of training examples (compare Table 1).

5.4 Per-NE Type Analysis

Finally, Table 9 shows the F_1 scores of the three best systems on the four NE classes from the data. All systems show the same patterns: best performance on PERson, followed by LOCation, ORGanization and finally on OTHer. The differences between PERson and LOCation are nonexistant to small (2%) while they perform substantially worse on ORG and again substantially worse on OTH. Again, it is interesting to compare the two top systems, ExB and UKP: UKP does slightly better on PER and LOC, the two most frequent classes (cf. Table 1), while ExB excels significantly for the two minority classes ORG and OTH. This complementary behavior indicates that there is a potential for ensemble learning using these systems.

In this comparison of NE types, the same question arises as for the comparison of levels: to what extent are the results a simple function of training set sizes? It is definitely striking that the ranking of the NEs types in terms of performance corresponds exactly to the ranking in terms of training data (cf. Table 1). At the same time, there is also reason to believe that the NE categories ORGanization and, in particular, OTH, are much less internally coherent than PER and LOC and therefore more difficult to model.

	ExB	UKP	MoSTNER
PER	84.05	85.48	82.54
LOC	84.05	84.62	80.47
ORG	76.29	69.60	62.24
OTH	59.46	49.81	48.38

Table 9: Peformance by NE type for top systems (F1 according to M1, outer chunks)

5.5 Comparing systems

An open question at this point is to what extent the submitted systems are complementary: do they make largely identical predictions or not? Given that the methods that the systems use are quite diverse, a large number of identical predictions could indicate problems with the dataset. Conversely, highly complementary output presents an opportunity for ensemble and other system combination methods. Historically, the best CoNLL 2003 system was also an ensemble (Florian et al., 2003).

We first computed the overlap between the predictions of each pair of systems at the word level, i.e., for what portion of words the two systems predicted the same label. We excluded words where both systems predicted O. Only the overall best run of each system was considered. We included the gold standard as a pseudo system (GOLD).

The results are shown in Table 10. The overlap

	UKP	Nessy	BECREATIVE	GOLD	NERU	ExB	DRIM	mXS	MoSTNER	PLsNER	Earlytracks	HATNER
UKP	—	0.447	0.317	0.594	0.406	0.561	0.542	0.448	0.578	0.613	0.568	0.389
Nessy	0.447	—	0.316	0.419	0.406	0.457	0.503	0.441	0.465	0.466	0.487	0.446
BECREATIVE	0.317	0.316	—	0.292	0.286	0.316	0.333	0.312	0.343	0.344	0.343	0.299
GOLD	0.594	0.419	0.292	—	0.392	0.614	0.525	0.418	0.556	0.558	0.553	0.361
NERU	0.406	0.406	0.286	0.392	—	0.431	0.442	0.426	0.432	0.443	0.442	0.448
ExB	0.561	0.457	0.316	0.614	0.431	—	0.550	0.460	0.578	0.572	0.576	0.406
DRIM	0.542	0.503	0.333	0.525	0.442	0.550	—	0.506	0.574	0.572	0.605	0.481
mXS	0.448	0.441	0.312	0.418	0.426	0.460	0.506	—	0.491	0.499	0.503	0.486
MoSTNER	0.578	0.465	0.343	0.556	0.432	0.578	0.574	0.491	—	0.610	0.619	0.437
PLsNER	0.613	0.466	0.344	0.558	0.443	0.572	0.572	0.499	0.610	—	0.595	0.453
Earlytracks	0.568	0.487	0.343	0.553	0.442	0.576	0.605	0.503	0.619	0.595	—	0.447
HATNER	0.389	0.446	0.299	0.361	0.448	0.406	0.481	0.486	0.437	0.453	0.447	—

Table 10: Pairwise word-level overlap of system predictions

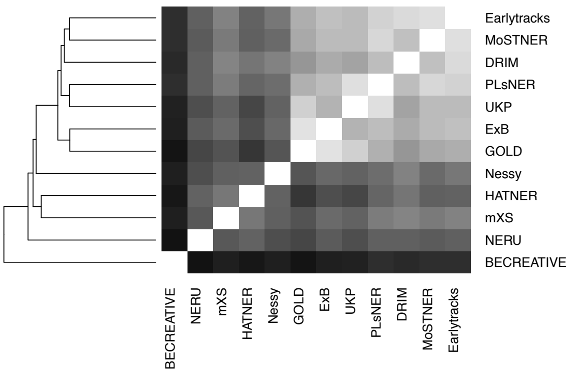

Figure 1: Heat map for pairwise system overlap

is relatively low: only a handful of comparisons yield an overlap of more than 0.5. We visualize the system comparisons as a heatmap in Figure 1. We see that BECREATIVE is very dissimilar to all other systems (it did not make any predictions for second-level NEs), while Earlytracks and MoSTNER have a comparatively high overall similarity to other systems (i.e., they produce a kind of "consensus" annotation). These two systems have also been clustered together, which may be related to the fact that they both use CRFs as their learning framework. Similarly, PLsNER and UKP, which are both based on neural networks, are also grouped together. The overall best system, ExB, has been grouped together with the gold standard.

Overall, these results look promising regarding future work on system combination. Without running a full-fledged analysis, we gauged the concrete potential by performing two simple analyses. The first one follows up on the per-level results from M3 (cf. Table 8), where we found that ExB and UKP show the best results for the first and the second level, respectively. Simply combining the ExB first level with the UKP second level yields a new best system with F_1=77.03 (M1), a further improvement of ΔF=.65 over ExB's previous result (cf. Table 6. The improvement notably is gained in precision (79.40 compared to 78.07) while recall stays about constant (74.79 compared to 74.75).

Finally, we computed an upper bound for the recall of an ensemble of the current systems. We performed this analysis because the fact almost all systems have a lower recall than precision (the best system has a recall of almost 75%, but the median is just at 63%) could be interpreted as an indicator that the corpus annotation is inconsistent or extremely difficult to recover automatically. However, when computing how many NE chunks in the gold standard are found by any of the systems, we determined that an oracle with access to all systems can cover 89.5% of the NE chunks. We take this result as an indication that there are no serious problems with the corpus, and that innovative strategies can hope to substantially improve over the current recall level.

6 Concluding remarks

In this paper, we have described the GermEval 2014 Named Entity Recognition shared task which extends the setup of traditional NER with morphologically motivated subtypes and embedded NEs.

The 11 submissions we received span a wide range of learning frameworks and types of features. The top systems appear to combine expressive machine learning techniques appropriate for the task (sequence classification and neural networks) with features that support intelligent generalization, notably encoding semantic knowledge.

The systems already achieve reasonable predictions on the dataset, in particular for precision-focused scenarios (median precision 76.7%, median recall 63.25%). At the same time, overlap in predictions between systems is surprisingly small, and system or feature combination may be able to further improve on the current results.

References

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

Chris Biemann, Claudio Giuliano, and Alfio Gliozzo. 2007. Unsupervised part of speech tagging supporting supervised methods. In *Proceedings of RANLP-07*, Borovets, Bulgaria.

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

Roman Capsamun, Daria Palchik, Iryna Gontar, Marina Sedinkina, and Desislava Zhekova. 2014. DRIM: Named Entity Recognition for German using Support Vector Machines. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, Hildesheim, Germany.

Fabian Dreer, Eduard Saller, Patrick Elsässer, Ulrike Handelshauser, and Desislava Zhekova. 2014. BE-CREATIVE: Annotation of German Named Entities. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, Hildesheim, Germany.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS*, pages 129–133, Saarbrücken, Germany.

Jenny Rose Finkel and Christopher D Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of HLT-NAACL 2009*, pages 326–334, Boulder, CO, USA.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL*, pages 168–171. Edmonton, Canada.

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

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

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, USA.

Jinseok Nam. 2014. Semi-Supervised Neural Networks for Nested Named Entity Recognition. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, Hildesheim, Germany.

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

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *Proceedings of LREC*, pages 1263–1266, Lisbon, Portugal.

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

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, Sapporo, Japan.

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

Andrea Zielinski and Christian Simon. 2008. Morphisto – An Open Source Morphological Analyzer for German. In *Proceedings of the 7th International Workshop FSMNLP*, pages 224–231.