Corpus-based Learning of Lexical Resources for German Named Entity Recognition

Marc Rössler

Computational Linguistics
University Duisburg-Essen
Duisburg – Germany
marc.roessler@uni-duisburg.de

Abstract

This paper explores the use of unlabeled data in a knowledge-poor approach to German NER. German is especially interesting for NER since not only names but all nouns are capitalized. Therefore, large and reliable lexical resources are necessary to develop and adapt systems for NER. Motivated by a model of word form observance, distinguishing three levels of different granularity, a method for the automatic creation of domain-sensitive lexical resources for NER is proposed. The approach uses linear SVMs and is based solely on an annotated corpus of reasonable size and a large amount of unlabeled data.

1 Introduction

In Named Entity Recognition (NER) proper names are detected and classified into predefined categories. Frequently used categories are PERSON (Anna, Condoleezza Rice), ORGANIZATION (United Nations, IBM), and LOCATION (Mississippi, Lisbon).

NER is a key part of information extraction but high-performance systems also facilitate the annotation of corpora. Systems for NER can be built based on handcrafted rules or on machine learning algorithms. Both utilize the so-called *internal* evidence, taken from within the NE, and the *external* evidence provided by the context in which a name appears. Given a set of labeled examples, external evidence, i.e. contexts and trigger words and internal evidence in the form of morphological or surface features such as capitalization can be learnt. Yet, it is not possible to learn sufficiently large lists of NEs due to the costs of manual labeling.

MUC (MUC-6, MUC-7) evaluations show that systems are able to score precision and recall values higher than 90% for English within a restricted domain. The remaining issues of NER are: i) techniques for a cheap adaptation to new domains and new categories and ii) the development of effective systems for other languages, especially for languages where the characteristics of NER strongly differ from NER for English.

These issues are addressed by working out a knowledge-poor approach for German. In Section 2 we specify the characteristics of NER for German and the usage of lists involved in it. Section 3 proposes a three-level model of word form observance useful for NER as will be shown in Section 4. Section 5 introduces important characteristics of the embedding NER-system and Section 6 demonstrates the automatic creation of domain-sensitive lexical resources necessary for the approach. Experiments on the category PERSON are described in Section 7 and discussed in Section 8.

2 NER for German and the usage of Lists

NER for German texts strongly differs from the same task in English. In German all nouns and not only names are capitalized. Therefore, the number of word forms that must be considered as potential NEs is much larger. Additionally, German is a language with partially free word order. This has an effect on the reliability of the external evidence. For instance, due to the strict subject-verb-object structure of English, a capitalized entry in front of a verb of communication is usually an NE belonging to the category PERSON or ORGANIZATION. For German this clue is much weaker since the finite verb occurs at three different positions within the clause and the subject has only a tendency to occur at the first position.

Mikheev et al. (1999) investigated the role of lists of NEs and showed that reasonable results are possible with small or even no lists. We believe that such results are only possible for languages where capitalization is sufficient to detect NEs and where lists are only needed to support the categorization of the names.

For languages with other characteristics, such as German, NER is heavily dependent on substantial and reliable lists. A lack of coverage lowers recall while unreliable entries, especially frequent ones, dramatically degrade precision.

The reliability of a list is verified by evaluating the fact of being member of a list. But consider the German first name "Mark", which is very likely to appear in any list of German person names but is also part of the currency "Deutsche Mark". It will downgrade the reliability of all the other, possibly highly reliable entries of the list. This is of course not intended but inevitable with simple lists since grouping items to lists is the only possibility to deal with words never seen in the annotated corpus.

To overcome this issue we propose a suitable and intuitive model to observe word forms, distinguishing three levels of granularity. The model helps to clarify the demands on lexical resources and allows the automatic extraction of substantial and reliable lists.

3 Three levels to observe word forms

NER is a form of semantic tagging assigning labels for the predefined NE categories and a label for “not belonging to any predefined category”. For German, every capitalized word form has to be classified whether it is used literally or as a name belonging to one of the predefined NE classes.

The model we propose distinguishes three levels of granularity to observe word forms and the semantic labels assigned to them: (1) A local level, i.e. a single occurrence of a word form in context, (2) a discourse level, i.e. all occurrences of a word form within a text unit and (3) a corpus level, i.e. all occurrences of a word form within all texts available for the application.

(1) On the local level we observe a single occurrence of a word form in context and the semantic label assigned to it. The deliberate meaning of a word form, i.e. the semantic label is unambiguous, apart from intended ambiguity aiming at comic or poetic effects. Some of the word forms occur in a predictive context and can be tagged with NE labels with high reliability.
(2) On the discourse level we observe all occurrences of a word form within a text unit and the semantic labels assigned to them. Addressing word-sense disambiguation, Gale et al. (1992) introduced the idea of a word sense located on the discourse-level and observed a strong one-sense-per-discourse tendency, i.e. several occurrences of a polysemous word form have a tendency to belong to the same semantic class within one discourse. We tested the one-sense-per-discourse tendency for our task of assigning NE-labels to word forms and measured a tendency of 93.5% by using the complete CoNLL-03 Corpus (2003). The word forms tagged with different labels within one discourse unit can be explained with organization names consisting partially of locations ("Deutsche Bank"), persons ("Philip Morris") or regular nouns ("Sport Factory").
(3) On the corpus level we observe all occurrences of a word form within all texts available for the application. The larger the corpus the more likely a particular word form was seen as member of two or more semantic classes. Within the CoNLL-03 Corpus, we measured an increase from 13% on a 50.000 word corpus to 24% ambiguous word forms on a 200.000 word corpus. I.e. for a real-world application dealing with millions of words, almost every noun, at least theoretically, is ambiguous on this level. Of course, this is only true for languages without valuable syntactic capitalization of names, but also for NEs not flagged with capitalization, as in the biomedical domain.

4 The three-level model for NER

The proposed three-level model is directly related to the task of NER. The actual NE-tagging is located on the local level while the discourse and the corpus level are used to support the tagging on the local level.
It is common practice in NER to utilize the discourse level to disambiguate items in non-predictive contexts (see e.g. Mikheev et al., 1999; Neumann & Piskorksi, 2002; Volk & Clematide, 2001). Within one discourse unit, all the NEs classified in predictive contexts are stored and used to disambiguate the NEs in non-predictive contexts.
Lists of NEs are located on the corpus level. Every list, whether manually or automatically compiled, claims that all the word forms contained are likely to appear as NEs within the corpus. How problematic this claim is, was illustrated with the German word "Mark" in Section 2, but is even more evident when recalling the ambiguity going along with observations on the corpus level.
Within ML-approaches lists are integrated in the form of a feature describing that a word form is on a specific list. The reliability of this feature is evaluated on an annotated corpus. Therefore, all entries of the list, - ambiguous and unambiguous word forms - are rated with the same reliability. To overcome this, an individual treatment of every particular word form is necessary, representing the probability of a particular word form to occur with a particular label. Providing information for all predefined NE categories, this results in a list of all word forms seen within all texts available for the application containing the probabilities for all categories.
Unfortunately, such a list and the probabilities can only be calculated based on a very large, fully annotated corpus. A corpus of the size of common training data is evidently too small. To overcome this, we propose a form of lexical bootstrapping. We assume that the probabilities calculated on the basis of a weak classifier applied to a large unlabeled corpus are sufficient for our task.
After introducing the characteristics of the embedding NER-system, the creation of such resources is described in Section 6 and evaluated in Section 7.

5 A knowledge-poor approach to NER

The optimal practice in NER yields efficient and highly reliable results based only on cheaply available resources like an annotated corpus of reasonable size and non-annotated data. Approaches rich of handcrafted knowledge or dependent on other language technology tools suffer from several limitations: They are laborious when adapted to new domains, especially w.r.t the creation and evaluation of the domain-sensitive named-entity lists. Furthermore, the application of additional tools like part-of-speech tagger, syntactic chunker etc. increases processing time.
In order to build an efficient and easy to adapt system we developed a knowledge-poor approach. We refrain from

- any additional linguistic tools, like morphological analyser, part of speech tagger or syntactic chunker
- any handcrafted linguistic resources, like dictionaries
- any handcrafted knowledge providing lists, like gazetteers, lists of NEs or lists of trigger words.

From a linguistic point of view, NEs are phenomena located on the phrase-level. Nevertheless, for the sake of straightforwardness, we restrict our model to single words.
To overcome the knowledge-sparseness we utilize methods based on the three-level model of word form observance described in Section 4. The model allows the automatic extraction and creation of corpus-level knowledge necessary for the detection of NEs, based only on unlabeled data.
Additionally, the common strategy of utilizing the discourse-level is applied: All items of a discourse unit classified as NEs are stored in a dynamic lexicon. Then, the processed discourse unit is matched against the dynamic lexicon in order to detect NEs in non-predictive contexts. In the same step we also try to handle phrases coordinated with commas, hyphens, etc, e.g. the members of a soccer team. When two or more phrases are classified as belonging to the same NE class, the other coordinated phrases are tagged the same way.
The approach is based on a linear SVM classifier. SVM (Vapnik, 1995) is a powerful machine learning algorithm for binary classification able to handle large numbers of parameters efficiently. It is common within the NLP community to use SVMs with non-linear kernels. Mayfield et al. (2003), Takeuchi & Collier (2002) or Isozaki & Kazawa (2002) used polynomial kernel function for NER. Besides the good classifier capabilities of non-linear kernels they are very expensive in terms of processing time for training and applying. Therefore, we favor linear SVMs[1] not suffering from these limitations. For tasks comparable to NER, only a few approaches employed linear SVMs, e.g. Giménez & Márquez (2003) scored very good results for POS-tagging.

[1] All experiments were conducted with the SVM^{light} software package, freely available at: http://svmlight.joachims.org.

Although German has a rich morphology, we do not abstract word forms to lemmas. Still we consider morphology by representing word forms with their positional character n-grams. See Table 1 for an example of this feature set. The representation is capable to capture simple morphological regularities of NEs and the context words surrounding them.
Additionally, we use word-surface features comparable to the ones used in e.g. Borthwick et al. (1998) indexing for instance whether a word form is capitalized, consists of numbers, contains capitals, etc. We also consider word-length and map it to one dimension. See Table 1 for all the features used within our knowledge-poor approach.
To capture the context of the word to classify, we set a 6-word window, consisting of the three preceding, the current, and the two succeeding words. All the features mentioned in Table 1 are extracted for all words of the defined window.

f1	Word-surface feature like e.g. "4-digit number", "Capitalized", "Uppercase only" etc.
f2	Character-based word length
f3	Sub-word-form representation with positional character n-grams. The word "Hammer" is represented as: "r", "er", "mer" at last position, "ham" at first, "amm" at second position etc.
f4	Corpus-lexicon representing how often and how confidential a word was seen as NE of a particular category.

Table 1: The table shows the feature sets f1-f4 extracted for all words of a 6-word window.

6 Creating lexical resources based on unlabeled data

To overcome the problems with simple enumeration lists, we propose sophisticated lists representing the reliability of every particular item. We assume that the output of a weak classifier is sufficient to approximate the tendency of a word form to occur with a particular label if the weakness of the classifier fulfils certain requirements: The classifier can be weak with relation to recall but not to precision and the classifier's weakness should not be biased towards particular lexical units.
Therefore, we set up a classifier trained only on external evidence, i.e. a classifier with fairly good predictions on some contexts or trigger words but without access to internal evidence provided by the current word.
The assumption that the weakness of such a classifier will not be biased towards particular word forms will be violated to a certain extent: We cannot guarantee that particular contexts only occur with particular NEs and the classification of single words leads to the effect that parts of NEs, consisting of more than one word, are considered as context of the current token.
Nevertheless, we assume that the output of such a classifier applied to a large corpus can be used to estimate the probability of a particular word form to occur with a particular label. The first idea to approximate the probabilities is simply to calculate the relative frequency of every label assigned to a particular word form. Yet, in our experiments better results were scored when integrating a confidential value of the classifier's assignment. Therefore, we used the discretized decision value of the SVM classifier, indexing the distance to the separating hyper-margin. The resulting corpus lexicon contains all word forms that are potential NEs, i.e. mainly all capitalized word forms.

7 Experiments for the category PERSON

In order to evaluate our approach, experiments on the category PERSON were conducted. To create a corpus lexicon, a weak SVM classifier was set up to classify tokens as belonging to the category PERSON or to the category NIL, i.e. not PERSON. It had no access to the word form of the current token (feature-set f3 in Table 1). It was trained on the 200.000 words training data of the CoNLL-03 corpus for German. In terms of precision (0.89) and recall (0.40) on the token level, it fulfilled the requirement specified in Section 6. This weak classifier was applied to a 40-million word corpus (Frankfurter Rundschau Corpus, 1994) and the output was used to compile a corpus-specific lexicon. For the 320.000 word forms considered as potential NEs we extracted the total frequency of being tagged as PERSON or as NIL and the relative frequency of being tagged with a particular decision value by the SVM classifier.
The entries of this corpus lexicon (feature set f4 in Table 1) were added to all the words of the 6-word window in combination with the feature sets f1-f3. The effect of this corpus lexicon was compared to a baseline classifier.
Using the resulting classifier to bootstrap the lexical resources scored the best results. Therefore, the classifier based on feature set f1-f4 without any restrictions was applied once again to the 40-million word corpus and the output was used to recompile the corpus lexicon.
The baseline classifier was set up with the feature set f1-f3 (see Table 1) for all words of the 6-word window and trained on the training data of the CoNLL-03 corpus. The evaluation was performed on the 50.000 words test data[2] of the CoNLL-03 corpus using the CoNLL evaluation software.

Classifier	**P**	**R**	**F**
Weak Classifier: feature set f1-f3, but no f3 for the word to classify (per token P: 89, R: 40)	62.3	44.6	52.1
Baseline: Feature set f1-f3	78.6	72.6	75.5
Feature set f1-f4 (includes the corpus lexicon)	89.2	86.2	87.6
Feature set f1-f4 (includes the bootstrapped corpus lexicon)	89.4	88.4	88.9
Volk & Clematide (2001).	92	86	88.8
Neumann & Piskorksi (2002).	95.9	81.3	88.0

Table 2: Results of our experiments and two other approaches for German in terms of Precision, Recall and F-Measure for the category PERSON. See Table 1 for the feature sets f1-f4.

As shown in Table 2 the results scored with our approach are competitive to the state of the art approaches to NER for German texts. On the development data an F-measure even higher than 91 was scored. The other approaches are based on rules and several knowledge sources. It is not clear how far these results are comparable since the other systems were evaluated on different and much smaller test data.

8 Discussion and related work

We have shown a knowledge-poor approach to NER for German texts. Experimental results on the category PERSON show state of the art results. The approach is especially interesting since it addresses the automatic creation of

[2] After correcting few manifest errors in the annotation.

domain-sensitive lists of NEs. These resources are created based solely on resources available at ordinary costs: An annotated corpus of reasonable size and a large amount of unlabeled data. Especially the usage of unlabeled data to create the domain-sensitive lists is an important step towards adaptive systems.

Our work is related to other approaches utilizing unlabeled data. They all have in common to start with a set of seed lists and/or seed rules. Buchholz & Bosch (2000) applied voluminous seed lists on large corpus of Dutch newspaper texts. Seed lists are used to extract external evidence, i.e. contexts and trigger words are learned, while seed rules extract internal evidence, i.e. lists of NEs. Evidence previously learnt can be used to find and verify new evidence in a bootstrapping-cycle. Collins & Singer (1999) start with 7 simple rules to build an NE classifier for English. Riloff & Jones (1999) and Thelen & Riloff (2002) learn syntactic patterns predicting the semantic class. Biemann & Quasthoff (2002) work on German texts and focus on the creation of NE-lists based on seed rules and small seed lists in very large corpora. Lin et al. (2003) present an algorithm that simultaneously learns multiple semantic classes. Yarowsky & Cucerzan (1999, 2002) report a language-independent approach using Expectation Maximation-style bootstrapping to learn internal and external evidence.

Our approach also applies seed rules. However, as in Yarowsky's and Cucerzan's approach (2002), they are learnt from an annotated corpus. It is not clear at the moment whether our intended abandonment of internal evidence is necessary or should be given up. Especially the additional bootstrapping of the lexical resources, integrating internal evidence but still enhancing performance, might contradict our idea. The development of a model for the integrated learning of internal and external evidence is one of the most challenging issues for future research. First experiments with the categories LOCATION and ORGANIZATION also indicate that the approach must be enhanced to score or even outperform state of the art solutions.

References

Biemann, C. & Quasthoff, U. (2002). Named entity learning and verification: EM in large corpora. In *Proceedings of CoNLL-2002, The Sixth Workshop on Computational Language Learning. Taipei, Taiwan.* San Francisco: Morgan Kaufmann. pp. 8-14.

Buchholz, S. & van den Bosch, A. (2000). Integrating seed names and n-grams for a named entity list and classifier. In *Proceedings of LREC-2000.* Athens, Greece. pp. 1215-1221.

Borthwick, A., Sterling, J., Agichtein, E., Grishman, R. (1998). NYU: Description of the MENE Named Entity System as Used in MUC-7. In *MUC*-7.

Collins, M. & Singer, Y. (1999). Unsupervised models for named entity classification. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. New Brunswick, NJ: Association for Computational Linguistics.

CoNLL-03 Corpus (2003). Data provided for the German part of the shared task Language-Independent NER. For a description see Tjong Kim Sang & De Meulder (2003). The corpus is online available: http://cnts.uia.ac.be/conll2003/ner/

Cucerzan, S., & Yarowsky, D. (2002). Language independent NER using a unified model of internal and contextual evidence. In *Proceedings of CoNLL-2002, The Sixth Workshop on Computational Language Learning*. Taipei, Taiwan. San Francisco: Morgan Kaufmann.

Cucerzan, S. & Yarowsky, D. (1999). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In: *Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC.* College Park. pp. 132-138.

Frankfurter Rundschau Corpus (1994). Published on the ECI Multilingual Text CD, distributed by the Linguistic Data Consortium. LDC Catalog Number LDC94T5.

Gale, W. A., Church, K. W., Yarowsky, D. (1992). One sense per discourse. In *Proceedings of DARPA speech and Natural Language Workshop.* Harriman, NY.

Giménez, J. & Márquez L. (2003). Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. *Proceedings of RANLP – 2003.* Borovets, Bulgary.

Isozaki, H, & Kazawa, H. (2002). Efficient Support Vector Classifiers for Named Entity Recognition. In: *Proceedings of COLING-2002.* pp. 390-396, 2002.

Mayfield, J, McNamee, P., Piatko, C. (2003). Named Entity Recognition using Hundreds of Thousands of Features. In Proceedings of CoNLL-2003, *Edmonton, Canada.* pp. 184-187.

Andrei Mikheev, A., Moens, M., Grover, C. (1999)*: Named Entity recognition without gazetteers.* In *EACL'99, Bergen, Norway*. pp. 1-8.

MUC-6. Proceedings of the Seventh Message Understanding Conference. Morgan Kaufmann. Columbia, Maryland. 1995.

MUC-7. Proceedings of the Seventh Message Understanding Conference. Morgan Kaufmann. Fairfax, Virginia. 1998.

Neumann, G. & Piskorksi, J. (2002). A Shallow Text Processing Core Engine. In *Journal of Computational Intelligence, Volume 18, Number 3*. pp. 451-476.

Riloff, E. & Jones, R. (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In: *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99),* pp. 474-479.

Takeuchi, K. & Collier, N. (2002). Use of support vector machines in extended named entity recognition. In Proceedings of CoNLL-2002, The Sixth Workshop on Computational Language Learning. Taipei, Taiwan. San Francisco: Morgan Kaufmann. pp. 119-125.

Thelen, M. & Riloff, E. (1999). A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing.*

Tjong Kim Sang, E. & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 142-147.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer Verlag.

Volk, M. & Clematide, S. (2001). Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition. In: *Proceedings of 6th International Workshop on Applications of Natural Language for Information Systems.* Madrid.

Lin, W., Yangarber, R., Grishman, R. (2003). Bootstrapped Learning of Semantic Classes from Positive and Negative Examples. In: *Proceedings of the ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining.* Washington, D.C.