Adapting an NER-System for German to the Biomedical Domain

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

Abstract

In this paper, we report the adaptation of a named entity recognition (NER) system to the biomedical domain in order to participate in the ”Shared Task Bio-Entity Recognition”. The system is originally developed for German NER that shares characteristics with the biomedical task. To facilitate adaptability, the system is knowledge-poor and utilizes unlabeled data. Investigating the adaptability of the single components and the enhancements necessary, we get insights into the task of bio-entity recognition.

1 Introduction

NER describes the detection and classification of proper names into predefined categories. Beside the distinction between rule-based and automatically trained systems, the approaches can be classified according to the amount of domain- and/or linguistic knowledge they incorporate.

In order to build an efficient and easy to adapt system, we developed a knowledge-poor approach that is successful for German person names (Rössler, 2004). German NER shares some characteristics with bio-entity recognition such as the unreliable capitalization of names, the resulting difficulties of boundary detection and the entailed treatment of homonymic and polysemic items. We believe that the process of adaptation is able to sketch out some interesting aspects of the biomedical domain.

In Section 2 we introduce the design guidelines and the underlying model of our knowledge-poor approach to NER. In Section 3 we describe the adaptation of the system and the modifications and enhancements involved. Section 4 introduces a three-level model to observe word forms that allows further improvements based on discourse units and the utilization of unlabeled data. These techniques were successfully applied to German person names, i.e. they led to more than 10 points increase in f-score, thus exhibiting state of the art performance. However, they completely failed on the bio-entity task. We will discuss what the failure of this technique reveals about the bio-entity task. Section 5 presents and discusses the final evaluation, while Section 6 contains some concluding remarks.

2 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 in handcrafted knowledge or dependent on other language technology tools suffer from several limitations: They are laborious to maintain and to adapt to new domains, especially with respect to the creation and evaluation of the domain-sensitive lists of NEs. Furthermore, the application of additional tools like part-of-speech tagger, syntactic chunker etc. increases processing time, and it is not clear at the moment whether such tools facilitate the task without additional adaptations to the new domain. 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 at the phrase-level. Nevertheless, for the sake of straightforwardness, we restrict our model to single words. To overcome the knowledge sparseness, the so-called three-level model of word form observance was developed and successfully applied to German person names (Rössler, 2004). In Section 4 we discuss our attempts to apply this model to the biomedical domain.

The approach is based on linear SVM classifiers. 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. Takeuchi and Col-

lier (2003) successfully applied a polynomial kernel function for biomedical NER. Beside 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.

Instead of using surface words in combination with morphological analyses and/or handcrafted suffix and prefix lists, we represent words with a set of positional character n-grams. Using the training data, this set is compiled by extracting the last uni- and bigram, three trigrams from the end, and three trigrams from the beginning of every word. All the entries occurring less than four times are removed. Table 1 contains an example of this feature set f3. The representation is capable of capturing simple morphological regularities of NEs and the context words surrounding them. Additionally, we use deterministic word-surface features (feature set f1) commonly used in NER (see Bykel et al., 1997), 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 (feature set f2). To capture the context of the word to classify, we set a six-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”, ”ATCG-sequence”, ”Uppercase only” etc.
f2	Character-based word length
f3	Sub-word form representation with positional character n-grams. ”Hammer” is represented as: ”r”, ”er”, ”mer” at the end, ”ham” at first, ”amm” at second, “mme” at next to last position.
f4	Probabilites of all classes if higher than zero, calculated by the second-order Markov Model.

Table 1: The table shows the feature sets f1-f4 extracted for all words of a 6-word window. Feature set f4 is described in Section 3.

3 Adapting the System

After adding ATCG sequence (see Shen et al. 2003) and GreekLetter (see Collier et al. 2000) as domain-specific deterministic word-surface features, we ran first experiments on the GENIA (2003) corpus. While inspecting the results we noticed that special attention was necessary to address the correct boundary detection of the entities and the transformation of the output of the SVM-classifiers to the IOB-notation.

A first step to improve the boundary detection is based on the output of a second-order Markov Model in order to support the SVMs that are not optimised to tag linear sequences. We trained TnT (Brants, 1998), a Markov Model implemented for POS-tagging on the surface words, and used the probabilities for all classes as features for the SVMs (feature set f4 on Table 1).

The second step was implemented within the post-processing component designed to transform the output of the SVM-classifiers to the IOB-notation. In order to facilitate the multi-class output, we set up a total of seven classifiers: Five of them specific to the five NE-classes and two additional classifiers assigning a general begin-tag and a general outside-tag. Although a dynamic programming approach to resolve the multi-class issue for SVMs is an important desideratum, we implemented a simple heuristic as a first step.

To transform the output of the seven classifiers into the IOB-output, we first applied a simple one-vs-rest method based on the decision values of the SVMs. The general begin tag was used to support the correct detection of the B-tags.

In a second post-processing step, we improved the results based on a definition of the revisability of a label assigned with respect to a competing label. According to this, a label is revisable if the competing label is among the three best labels and has a decision value higher than 0.2, or if the value of the outside-classifier is lower than 0.2, i.e. the label OUTSIDE is not that confident. A label is considered to be competing to the current label if it was assigned to the word before or the word after.

4 Attempting to utilize the three-level model to the biomedical domain

The three-level model described in Rössler (2004) is motivated by the fact that lexical resources in the form of named entity lists deal with surface words, i.e. word forms, thus ignoring the problems of homonymy and polysemy.

To address this issue, we distinguish three different levels to observe word forms and the semantic labels assigned to them and show how they are related to and support the NER:

- The local level describes a single occurrence of a word form. The correct labelling of these occurrences is the actual task of NER.
- The discourse level describes 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 at the discourse-level and observed a

[1] All experiments were conducted with the SVMlight software package, freely available at: http://svmlight.joachims.org.

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. 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).

- The corpus level describes all occurrences of a word form within all texts available for the application. The larger the corpus, the more likely a particular word form is seen as member of two or more semantic classes.

In order to utilize the discourse level, all words tagged as entity within one MEDLINE abstract are stored in a dynamic lexicon. Then, the processed discourse unit is matched against the dynamic lexicon in order to detect entities in non-predictive contexts. To find the correct boundaries the unit is post-processed as described in Section 3.

To reflect the issues concerning polysemy and homonymy of lexical resources, we propose sophisticated word-form based NE lists, representing how likely a particular entry will be tagged with a particular label. These values are specific for a corpus, i.e. they are located at the corpus level.

To create such resources, 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. Therefore, we trained classifiers for all classes and applied them to a 30-million word corpus extracted from MEDLINE (1999), using the search term [”blood cell” or ”transcription factor”]. This automatically annotated corpus was used to create a corpus specific lexicon containing about 95,000 word forms. For all these entries, we extracted the total frequency of *being tagged with a particular label* and the relative frequency of *being tagged with a discretized decision value* by the SVM classifiers, i.e. we set five thresholds and counted how often an item was labelled with a decision value fulfilling a particular threshold.

Both techniques completely failed: Neither the utilization of the discourse-level, nor the lexical bootstrapping had a positive impact when applied to the biomedical domain. This raises the question on the specifics of the biomedical domain.

The utilization of the discourse-level is proved of value in most NE-tasks, thus the failure within the biomedical domain is surprising. The one-sense-per-discourse tendency is obviously weaker in the biomedical domain, since genes and proteins can share the same name and be mentioned in the same abstract. Additionally, the NEs occurring within the GENIA corpus consist in average of more than two words and seem to be diverse in their appearance, even within one document. For almost every word form, even brackets and stopwords can be a part of an NE, it is a great deal of work to develop heuristics improving recall without lowering precision dramatically. Moreover, the method is highly sensitive to precision errors, as it spreads out elements tagged incorrectly. Furthermore, it is questionable if abstracts – due to their enormous density and shortness – are appropriate text units for this method.

The failure of the lexical bootstrapping is more difficult to interpret since this technique is not that well-tested. In our experiments, it was successfully applied to German person names and also had some positive impact on German organization and location names. One source of problems can be seen in the low precision of the classifier used to create the annotated corpus. We assume that a high-precision and low-recall classifier will produce better lexical resources. Another source can be seen in the complexity and the length of biological names. The restriction to single words is probably not appropriate for the bootstrapping process. For future research, we will investigate the bootstrapping of external evidence, i.e. we will not focus on the learning of names, but rather on the units that indicate the beginning or the end of a name-class.

5 Evaluation

All the evaluation was conducted on the corpus made available for the shared task Bio-Entity Recognition. All configurations were trained on the 2000 abstracts provided, i.e. 500,000 words to train and we finally evaluated them on the 100,000 words evaluation data. Table 2 shows the scores for the different classifiers and components in the first rows, and the performance of the best configuration evaluated for each NE-class.

On the basis of the scores in Table 2 it is possible to discuss the impact and values of the different components of the system.

Using the surface words instead of f3, the subword-form representation with positional character n-grams leads to a decrease of more than 2 points in terms of recall and precision.

The f-score of the Markov Model, trained on the word forms, is almost comparable to the basic SVM-configuration f1-f3, but the precision of the SVM is higher.

The post-processing component cannot be applied to the output of the Markov Model, as the definition of the revisability is specifically designed for the output of the seven SVM-classifiers. The post-processing component shows very good results and leads to an increase of 4 points almost

equal for precision and recall, i.e. the component is able to address the boundary detection problem by means of the definition of the revisability of a tag with regard to a competing tag.

	f1	f2	f3	f4	postProc	R	P	F
Overall score	Markov Model only					62.6	54.1	58.0
	x	x	[2]			57.9	54.4	56.1
	x	x	x			61.0	56.2	58.5
	x	x	x		x	65.4	59.9	62.6
	x	x	x	x		66.3	60.1	63.1
	x	x	x	x	x	**67.4**	**60.1**	**64.0**
protein	x	x	x	x	x	72.9	62.0	67.0
cell_line	x	x	x	x	x	55.2	42.9	48.3
DNA	x	x	x	x	x	57.9	52.6	55.1
cell_type	x	x	x	x	x	62.7	70.6	66.5
RNA	x	x	x	x	x	44.1	49.5	46.7

Table 2: Overall scores and scores for each NE class. See Table 1 for the feature sets f1-f4; postProc refers to the second post-processing component described in Section 3.

Combining the basic SVM-configuration f1-f3 with f4, the probabilities calculated by the Markov Model, leads to a slight increase compared to the post-processing component. We are convinced that both the post processing and the Markov Model cover similar phenomena by supporting the SVM to detect the correct boundaries.

The combination of all feature sets f1-f4 with the post-processing leads to a further increase of 1 point, demonstrating the ability of the SVM to optimize its predictions on heterogeneous knowledge sources.

6 Conclusion

We have demonstrated the adaptation of an NE tagger originally developed for German to the biomedical domain. We believe that the process of adaptation is able to sketch out some interesting aspects of the new domain.

The names of the biomedical domain have morphological features that can be covered by the subword-form representation with positional character n-grams.

The failure of the techniques based on the three-level model indicate that the polysemic and homonymic items and the complexity of biological names hamper or even inhibit a further optimization of models based on simple n-grams of words. We believe that the consideration of more complex units and longer distant phenomena will lead to further progress in NE-tagging. For the biomedical domain, the work of Takeuchi and Collier (2003) demonstrates the successful incorporation of shallow parsing.

For future research, we plan to address these issues by focusing on learning external evidence, i.e. triggers and longer-distant phenomena from unlabeled texts.

[2] Instead of the positional character n-grams the system is trained on surface words.

References

D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. 1997. *Nymble: a High-Performance Learning Name-finder*. Proceedings of the Fifth Conference on Applied Natural Language Processing. Washington, DC.

T. Brants. 1998. *TnT - A Statistical Part-of-Speech Tagger*. Saarland University, Computational Linguistics. Saarbruecken. Available at: http://www.coli.uni-sb.de/~thorsten/tnt/.

N. Collier, C. Nobata, and J. Tsujii. 2000. *Extracting the Names of Genes and Gene Products with a Hidden Markov Model*. Proceedings of COLING'2000. Saarbruecken.

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

GENIA Corpus 2003. Available at: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/.

MEDLINE. 1999. The PubMed database is available at: http://www.ncbi.nlm.nih.gov/PubMed/.

A. Mikheev, M. Moens, and C.Grover, C. 1999. *Named Entity recognition without gazetteers.* Proceedings of EACL'99. Bergen.

M. Rössler. 2004. *Corpus-based Learning of Lexical Resources for German Named Entity Recognition*. Proceedings of LREC 2004. Lisboa.

D. Shen, J. Zhang, G. Zhou, J. Su, and C. Tan. 2003. *Effective Adaptation of Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain*. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine. Sapporo.

K. Takeuchi and N. Collier. 2003. *Bio-Medical Entity Extraction using Support Vector Machines*. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine. ACL 2003. Sapporo.

V. Vapnik. 1995. Statistical Learning Theory. Springer. New York.