Extended Feature Set for Chemical Named Entity Recognition and Indexing

Torsten Huber[1]*, Tim Rocktäschel[2]*, Michael Weidlich[1], Philippe Thomas[1], Ulf Leser[1]
* *These two authors contributed equally to this work*

[1]Humboldt-Universität zu Berlin,Knowledge Management in Bioinformatics, Unter den Linden 6, Berlin, 10099, Germany
[2]Department of Computer Science, University College London, United Kingdom

{thuber,weidlich,thomas,leser}@informatik.hu-berlin.de
ucabroc@ucl.ac.uk

Abstract. The BioCreative IV CHEMDNER Task provides participants with the opportunity to compare their methods for chemical named entity recognition (NER) and indexing in a controlled environment. We contributed to this task with our previous conditional random field based system [1] extended by a number of novel general and domain-specific features. For the latter, we used features derived from two existing chemical NER systems, ChemSpot [2] and OSCAR [3], as well as various external resources. In this paper, we describe our approach and present a detailed ablation study that underlines the positive effect of domain-specific features for chemical NER.

1 Introduction

Patents, scientific articles and medical reports are the main source of information for chemical entities and drugs [4]. The accurate recognition of these entities in such texts is a crucial prerequisite for a variety of applications, including the reconstruction of metabolic pathways, the manual curation of chemical databases (*e.g.* PubChem, ChEBI) and the retrieval of information about substances in drug development and their interactions [5]. A multitude of naming conventions (such as trivial names, systematic names or brand names) and their incoherent usage even in scientific articles makes chemical NER a fairly complex task, interesting for both academia from a research perspective and industry from an application perspective.

The BioCreative IV CHEMDNER task aims to promote research in this area by asking participants to develop a system that accurately identifies and indexes chemicals in scientific texts. The task organizers provided a training corpus and a development corpus, each consisting of 3500 documents and ∼29000 annotations, which are divided into classes `SYSTEMATIC`, `IDENTIFIER` (*i.e.* IUPAC-like), `FORMULA`, `TRIVIAL`, `ABBREVIATION`, `FAMILY` or `MULTIPLE` (multiple mentions of chemicals, *e.g. 'bis- and tris-xylosides'*). As these classes are diverse, identifying and indexing chemical entities requires a versatile approach capable of dealing with the heterogeneous conventions and varying levels of abstraction. In this paper we describe our approach based on a linear chain conditional random field (CRF) [6] and an extended feature set of domain-independent and domain-specific features.

2 Methods

To cope with the diversity of chemical entities, our system ChemSpot [2] uses a range of different methods, including a CRF trained for identifying IUPAC entities, dictionaries for trivial names and regular expressions for sum formulae. However, the basic approach used in ChemSpot is not suitable for recognizing heterogeneous classes of entities as they occur in this task, since its static methods do not make use of pertinent information of surrounding tokens. In [1] we therefore trained a CRF using features derived from the output of the individual components used in ChemSpot as well as other chemical resources. In this work, we additionally use features derived from The Open-Source Chemistry Analysis Routines (OSCAR) [3]. OSCAR is based on a maximum Markov entropy field and tends to have a higher coverage of chemical entities while being less precise than ChemSpot. A summary of the feature classes we added to our previous system is shown in Table 1. With our current approach we essentially aim at training a meta-model that incorporates existing NER tools as well as domain-specific resources and adapts to their individual confidences and biases.

2.1 General Features

The domain-independent features are based on a number of different feature sets covering several morphological, syntactic and orthographic properties such as token surface forms, n-grams, prefixes and suffixes, part-of-speech tags, syntactic patterns and token windows (see [1] for details). In addition to this general feature set, we added the following domain-independent features. Skip-gram features inspired by [7] are comprised of consecutive strings of the surrounding tokens (*e.g. "by injection of"*). Subsequently, we also mask intermediate tokens in this string with the intent of generalizing towards common phrases near chemical entity mentions (*e.g. "by* `#WORD#` *of"*). Furthermore, we added stemming features derived from token surface forms.

2.2 Domain-specific Features

As additional domain-specific resources, we included the Drugs@FDA list and ATC nomenclature list released by the World Health Organization (WHO). If an occurrence of a term from either of these resources is found, tokens in a sequence of length six or less receive a feature with the name of the resource that the sequence was found in. Furthermore, we generate features from the output of OSCAR. Specifically, we

Table 1: Overview of new features classes.

Feature Class	**Description**	**Feature Class**	**Description**
ONTOLOGY	is part of term found in PHARE, Drugs@FDA or WHO ATC list	OSCAR	is part of an entity found by OSCAR
SKIP-GRAM	consecutive string of surrounding tokens with inner token blinded	OSCAR-TYPE	type of chemical returned by OSCAR
STEMMING	stemmed token	OSCAR-CONFIDENCE	binned confidence returned by OSCAR

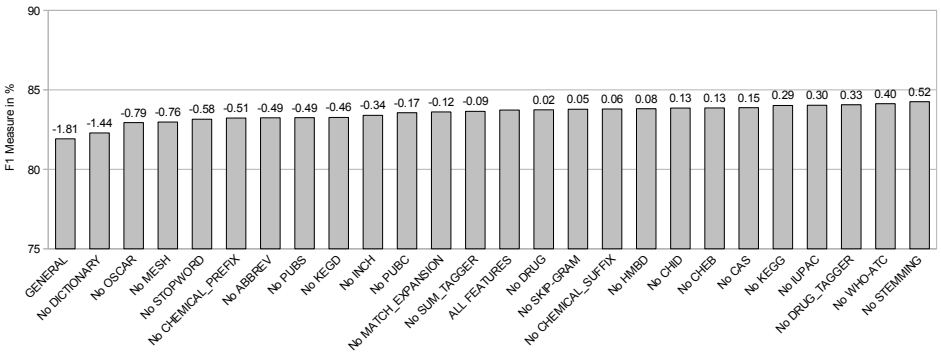

Fig. 1: Results of the ablation study (*i.e.* removing one feature class) trained on the training corpus and tested on the development corpus. The numbers atop the bars show the F_1 percentage point difference to the baseline (All Features). The General feature set contains only domain-independent features.

determine whether OSCAR recognizes a token as part of a chemical entity, the entity type assigned to it as well as the confidence score, binned using 25 percentage point steps. Note that contrary to ChemSpot, OSCAR does recognize mentions of chemical families, providing a higher coverage for entities of this class.

2.3 Document Indexing

Since a CRF in its standard form does not provide confidence estimates for entity mentions, we derive a ranked list of chemicals based on the TF-IDF score of entities extracted by the CRF for the BioCreative IV CHEMDNER chemical document indexing sub-task. First, we calculate the number of occurrences of each chemical in the document. Subsequently, we divide this value by the total number of occurrences of the entity in the entire corpus, yielding a score that is used for compiling a ranked list of entities for each document.

3 Results and Discussion

To measure the contribution of individual feature classes, we performed an ablation study on the development corpus (see Fig. 1). We used the entire feature set as a baseline and removed a single feature class at a time to evaluate its impact on the overall

Table 2: Results for three runs submitted to the BioCreative IV CHEMDNER Task trained on the training corpus and tested on the development corpus.

Features	**Precision**	**Recall**	$\boldsymbol{F_1}$
General features	86.84	77.54	81.92
General and domain-specific features (all features)	89.29	79.11	83.89
Best (all features except stemming)	88.89	80.08	84.25

NER performance. We found that among the domain-specific features, ChemSpot's dictionary tagger and OSCAR contribute most to the overall performance. The least useful feature classes are those based on the WHO ATC list and the stemming component. We furthermore found that removing two or more feature classes from the set consistently reduced the overall performance, regardless of their impact in the initial evaluation. We attribute the sharp performance loss without the OSCAR system to the fact that it is the only component that annotates chemical families.

Our results on the development corpus are shown in Table 2. We tested different feature combinations, such as the general, domain-independent feature set, which already achieves an F_1 measure of 81.92 %. The use of all features further increases this score by 1.97 percentage points, emphasizing the relevance of domain-specific features for chemical NER applications. The best F_1 was achieved by using all features except those generated from the stemming component.

We believe that the lower recall (compared to the precision) in all three runs is due to the fact that the development corpus contains chemical entities that are not part of the training corpus and hence can only be identified with the aid of the semantic context of the surrounding tokens. Chemicals belonging to classes like `ABBREVIATION` only provide few of such clues and are thus harder to identify, leading to a decrease in recall for them.

Acknowledgements

Funding: This work was supported by the Federal Ministry of Economics and Technology (BMWi) [KF2205209MS2 to T.R.] and the German Ministry for Education and Research (BMBF) [0315746 to T.H. and M.W.].

References

1. T. Rocktäschel, T. Huber, M. Weidlich, and U. Leser, "WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs.," in *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval 2013). 2013.*, 2013.
2. T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: A Hybrid System for Chemical Named Entity Recognition," *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, 2012.
3. D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust, "OSCAR4: a flexible architecture for chemical text-mining," *Journal of cheminformatics*, vol. 3, no. 1, p. 41, 2011.
4. M. Vazquez, M. Krallinger, F. Leitner, and A. Valencia, "Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications," *Mol. Inf.*, vol. 30, no. 6-7, pp. 506–519, 2011.
5. P. Thomas, M. L. Neves, I. Solt, D. Tikk, and U. Leser, "Relation extraction for drug-drug interactions using ensemble learning," 2011.
6. J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proc. of ICML-2001*, pp. 282–289, 2001.
7. D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling," in *Proc. of the Fifth international Conference on Language Resources and Evaluation (LREC-2006)*, (Genoa, Italy), 2006.