ChemSpot: a hybrid system for chemical named entity recognition

Tim Rocktäschel, Michael Weidlich and Ulf Leser

Department of Computer Science, Humboldt-Universität zu Berlin, Rudower Chaussee 25, 12489 Berlin, Germany

ABSTRACT

Motivation: The accurate identification of chemicals in text is important for many applications, including computer-assisted reconstruction of metabolic networks or retrieval of information about substances in drug development. But due to the diversity of naming conventions and traditions for such molecules, this task is highly complex and should be supported by computational tools.

Results: We present ChemSpot, a named entity recognition (NER) tool for identifying mentions of chemicals in natural language texts, including trivial names, drugs, abbreviations, molecular formulas and International Union of Pure and Applied Chemistry (IUPAC) names. Since the different classes of relevant entities have rather different naming characteristics, ChemSpot uses a hybrid approach combining a Conditional Random Field with a dictionary. It achieves an F1 measure of 68.1% on the AGS corpus, outperforming the only other freely available chemical NER tool, OSCAR4A, by 10.8 percentage points.

Availability: ChemSpot is freely available at:
http://www.informatik.hu-berlin.de/wbi/resources

Received on January 23, 2012; revised on March 20, 2012; accepted on April 4, 2012

1 INTRODUCTION

Metabolic and signaling networks representing complex physiological processes play an essential role in systems biology and drug research (Bolouri and Palsson, 2011). For instance, simulation results derived from recently published human metabolic networks provided substantial insight into functional biochemical relationships at the systems level (Bailly et al., 2009). Such networks are typically built by a group of biological experts that systematically scan relevant publications and extract the important information, which is a particularly tedious and time consuming task requiring considerable expertise (Alex et al., 2006). Natural language processing (NLP) can accelerate this process, especially by automatically pre-annotating network components (e.g. chemicals and proteins) and their interactions (Ananado et al., 2006). Such annotations, if of high quality, can considerably help to speed-up literature curation (Alex et al., 2009).

Accordingly, most work has been invested in the development of tools for named entity recognition (NER) of biomedical entities (Krallinger et al., 2008). While these tools mainly focus on identifying genes and protein names, in this work we address chemical names, a task which has not received much attention yet.

To whom correspondence should be addressed.

Cohen and Hersch 2006; Birnberg et al., 2006: Finding mentions of chemicals in text is hindered by the fact that there exist various and highly heterogeneous ways of naming them. This includes trivial names (e.g. water), brand names (e.g. Volaren®), systematic International Union of Pure and Applied Chemistry (IUPAC) names (e.g. adenosine 3′,5′-(hydrogen phosphate)), generic or family names (e.g. alcohols), company codes (e.g. ICI204636), molecular formulas (e.g. COOH) and identifiers of various databases. On top, many of these names are used in abbreviated form (e.g. DMS for dimethyl sulfate).

A number of nomenclature organizations exist and strive for systematic naming in the biochemical field, such as the IUPAC and the International Union of Biochemistry and Molecular Biology (IUBMB). However, most of their rules are only recommendations, leaving ample room for variation in their application (Bolouri, 2004). For instance, separating digits in systematic chemical names using dashes or commas is equally valid. In contrast, both, the existence as well as the non-existence of brackets and whitespaces can be crucial for the correct identification of chemicals. For instance, the placement of spaces between methyl, ethyl and malonate, results in four different chemical structures (Bolouri, 2006). This situation accounts for a high amount of possible synonyms for one chemical entity. Sometimes these synonyms do not even share a single pair of adjacent letters, e.g. in the case of phthalonitrile and o-dicyanobenzene (Bolouri, 2006). Chemical NER also tends to be sensible to spelling errors, which is especially crucial in long formulas, and errors during document transformations, for instance through inappropriate tokenization or sentence splitting (Bolomey et al., 2010). Even small errors may change the meaning of a chemical name completely; for instance, Brecher (1999) points out that several pairs of different structures differ only by one single character (e.g. methyamine and menthylamine). On top of these problems, also homonyms are common-place, especially when it comes to abbreviations.

Despite this heterogeneity, names for chemical structures in text can roughly be divided into two classes: a rather closed (finite) class for brand and trivial names, and an open (infinite) class for names following rule-based conventions (e.g. IUPAC names). In this article, we show that using a proper method for recognizing entities in each of these two classes enables the construction of a high-quality chemical NER system. We built ChemSpot, a tool which combines into a single system the two most prominent methods in NER: machine learning and dictionary matching. ChemSpot uses a conditional random field (CRF) to achieve high quality in recognizing IUPAC names. As dictionary, ChemSpot uses ChemIDPlus (http://www.nlm.nih.gov/pubs/factsheets/chemidplusfs.html), which allows, when applied with a proper matching algorithm, tokenization method, post-processing rules, high-quality annotation of trivial and brand boundaries.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
names, molecular formulas and abbreviations. By bundling both components into a single system, ChemSpot clearly outperforms OSCAR3/4 (Corbett and Murray-Rust 2008; Jessop et al. 2011), the only other freely available NER system addressing all classes of chemicals, and MetaMap (Assmann et al. 2008) on both the comprehensive SCAI corpus (Kolshik et al. 2009) and the automatically annotated DDI corpus (Segura-Bedmar et al. 2011). For instance, ChemSpot outperforms OSCAR4 by >10 percentage points F1 measure on the SCAI corpus.

2 RELATED WORK

NER in the biomedical domain has mainly focused on protein or gene names, where a wealth of systems have been developed [e.g. BANNER (Beanan and Gonzalez 2008), ProMiner (Fluck et al. 2007) or GNAT (Hakenberg et al. 2011)]. In contrast, recognition of chemicals has received much less attention. The Open-Source Chemistry Analysis Routines (OSCAR) software (Corbett and Murray-Rust 2008) is a system for the recognition of chemical entities based on Maximum Entropy Markov Models (MEMMs) McCauley and Freitag 2006; Corbett and Copostakia 2008; evaluated OSCAR3 on a corpus consisting of 42 chemistry publications (Sciborg corpus) and a corpus consisting of 500 PubMed abstracts (PubMed corpus). They reported an F1 measure of 80.7% for the former and 83.2% for the latter corpus. Unfortunately, both corpora are, to this day, not publicly available (PM Murray-Rust, personal communication). Jessop et al. 2011] recently refactored OSCAR3, providing a new version, OSCAR4, which in our evaluation (see Section 3.3) yielded a minor increase in performance compared with OSCAR3.

Klinger et al. 2008 used a CRF for extracting IUPAC and IUPAC-like chemical entities. They reported an F1 measure of 85.6% on their IUPAC test corpus (see Section 3.3 for an overview of the feature set used in their work). This tool is not freely available and does not cover drugs and trivial names. Note that CRFs are widely used for NER in various domains. For instance, ABINER (Settle 2009) and BANNER (Beanan and Gonzalez 2008) are both CRF-based NER tools for extracting protein mentions. BANNER is based on the CRF library Mallet McCauley 2008, and achieves competitive results on the BioCreative III corpus (http://banner.sourceforge.net/ last accessed 2012-01-20). In Section 3.3 we shall describe how we use BANNER’s API to employ a CRF for the recognition of IUPAC entities.

Another common approach for NER is using a dictionary of the terms of interest. Bletzin et al. 2005 built a combined dictionary for names of small molecules, drugs and abbreviations using name lists from the Unified Medical Language System (UMLS), MeSH, ChEBI, DrugBank, KEGG, HMDB and ChemDplus. They increased the quality of the dictionary by applying rule-based term filtering and manually reviewing frequent terms. The performance of all dictionaries as well as the combined dictionary was evaluated on the SCAI corpus (see Section 3.3) using the Peregrine dictionary-matching software Schummer et al. 2008. The combined dictionary achieved an F1 measure of 50%, but ChemDplus alone already achieved 49%. Bletzin et al. used Peregrine with a configuration that performs case-insensitive matching and favors the longest match. They adjusted Peregrine’s tokenizer to perform coarse tokenization, i.e. they did not use periods, commas, plus signs, hyphens, single quotation marks and parentheses as word delimiters. Additionally,

| Table 1. Annotated text corpora for training and assessment of chemical NER tools |
|---------------------------------------------|---------------------------------------------|
| Corpus                                      | Focus                                      | Available |
| PubMed corpus                               | General chemicals                          | No         |
| Sci-borg corpus                             | General chemicals                          | No         |
| IUPAC training corpus                       | IUPAC entities                             | Yes        |
| IUPAC test corpus                           | IUPAC entities                             | Yes        |
| SCAI corpus                                 | General chemicals                          | Yes        |
| DDI corpus                                  | Drugs                                     | Yes        |

*Corpus was annotated using MetaMap, thus, it is not a real gold-standard.

post-filters were applied to remove characters and common suffixes that are not part of chemical entities.

Segura-Bedmar et al. 2008 introduced DrugNER, a system for drug name recognition. This system combines the UMLS MetaMap Transfer (M1T) program and nomenclature rules by the World Health Organization International Nonproprietary Names (INNs) Program. They reported a precision of 99.1% and a recall of 99.8% on their DrugNER corpus. However, drugs in this corpus were automatically annotated using the same tools and, thus, cannot be considered as gold-standard entities (I.Segura-Bedmar, personal communication).

A common problem in chemical NER is the sparsity of annotated corpora for training and evaluation. Many of the corpora mentioned in this section actually are not available publicly, focus only on a restricted class of chemicals, or cannot be considered as gold-standards (Table 1). In this work, we use all available corpora we are aware of and provide evaluation results for ChemSpot and other tools on the IUPAC test corpus, the SCAI corpus and the DDI corpus.

3 METHODS

3.1 ChemSpot

ChemSpot’s main innovation is the combination of a CRF and a dictionary to explicitly cover the different naming conventions for entities commonly subsumed under the term ‘chemical’. IUPAC entities are morphological more complex than other chemical entities, calling for a classification-based tool, whereas brand names, drugs and small molecules follow hardy any rule and are best captured by an exhaustive dictionary (Kolshik et al. 2009). In contrast to previous approaches, which tried to cover both of these name classes with a single approach, ChemSpot uses a specific technique for each class.

Figure 4 illustrates the architecture of ChemSpot’s annotation and post-processing components. First, a CRF (left branch in the figure) and a dictionary (right branch) are independently used to annotate the input text. Dictionary matches are post-processed by expanding partial matches, correcting the boundaries of these matches and truncating common suffixes. Entities extracted by the dictionary may overlap, but they will cover the same span of text after match expansion. Hence, only one entity is kept. Finally, ChemSpot keeps the union of all entities extracted by the dictionary or the CRF. However, both approaches may extract the same entity or substrings of the same entity. In such cases, ChemSpot resolves these overlaps by favoring a match from the CRF over one from the dictionary.

We decided to use this rule, because we observed that in most cases of an overlap the dictionary match is a substring of the CRF match or the CRF’s
3.2 Corpora

Table 1 lists all corpora with annotations for chemicals we are aware of. Only few sentences contain a chemical entity. A NER tool often will be applied to arbitrary biomedical publications where in the training corpus. We consider this a realistic scenario since a chemical containing a chemical entity is much lower in the IUPAC test corpus than of tokens that are part of an entity. Note that the proportion of sentences

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of sentences</th>
<th>Number of tokens</th>
<th>Number of entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>IUPAC</td>
<td>3712</td>
<td>151</td>
<td>391</td>
</tr>
<tr>
<td>PARTIUPAC</td>
<td>322</td>
<td>0</td>
<td>92</td>
</tr>
<tr>
<td>MODIFIER</td>
<td>1040</td>
<td>14</td>
<td>104</td>
</tr>
<tr>
<td>FAMILY</td>
<td>0</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>SUM</td>
<td>0</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>TRIVIAL</td>
<td>0</td>
<td>0</td>
<td>414</td>
</tr>
<tr>
<td>ABBREVIATION</td>
<td>0</td>
<td>0</td>
<td>161</td>
</tr>
<tr>
<td>SCAI</td>
<td>161561</td>
<td>12424</td>
<td>15734</td>
</tr>
</tbody>
</table>

Entity density (%) 25.6 0.7 17.6

Note that there is no overlap between entities.

3.3 CRF

A CRF is a probabilistic undirected graphical model. In contrast to generative models such as hidden Markov models (HMMs), CRFs do not need to make assumptions about the underlying observation distribution. McCallum and Freitag (2000). Furthermore, a huge number of arbitrary and non-independent features can be used to describe the input data. McCallum et al. (2006). In contrast to CRFs, MEMMs suffer from the label bias problem. Ratnani et al. (2006), i.e. more probability mass is assigned to states with fewer outgoing transitions. By overcoming this disadvantage, CRFs are well suited for sequence labeling tasks such as NER. For further information on arbitrary and linear-chain CRFs, we refer to Klinger and Tomanek (2007).

ChemSpot uses Mallet (McCallum et al. 2008) as an underlying CRF implementation through the convenient API provided with BANNER (Klinger and Tomanek, 2007). We use BANNER’s data structures, methods for training and inference, as well as its configuration for Mallet.

To adapt to chemical NER, we turned off BANNER’s tokenizer, POS-tagger, lemmatizer and post-processing components and replaced its feature set with a subset of the one published by Klinger et al. (2006). For the recognition of IUPAC and IUPAC-like chemical names. This set includes the following features classes:

- morphological features (regular expressions)
  - all of the token’s characters are capitalized
  - token represents a real number
  - token is a dash, quote or slash
  - bag-of-words
  - token prefix of length two
  - token suffix of length two
  - token is preceded or succeeded by a whitespace.

ChemSpot configures BANNER to employ a second-order CRF and an offset conjunction of two. Offset conjunction of k preceding and succeeding tokens to the token’s features, thus, providing the CRF with more contextual information.

Tagging with a CRF is performed using the Viterbi algorithm, which is linear in the number of tokens and quadratic in the number of labels (Klinger and Tomanek, 2007). For instance, tagging the SCAI corpus with our CRF takes on average 2 ms per sentence.
When faced with such a huge number of terms, it is crucial to convert the list provided by Hettne to a whole entity. Finally, we remove certain suffixes at the end of entities using Hettne et al. [2009], which converts the dictionary into deterministic finite-state automata resulting in linear time complexity. In contrast to the dictionary-matcher Peregrine (used by Hettne et al.), LINNAEUS has no need for a tokenizer, i.e., mentions are directly extracted from text using the finite-state automaton rather than a dictionary look-up for sequences of tokens. Tagging the SCAI corpus with this dictionary as automaton takes 18 ms per sentence on average.

After the matching phase, we apply a number of post-processing rules. First, we keep only terms with a character length > 2, since one-letter and two-letter words are highly ambiguous. Second, all terms matching a regular expression for real numbers are removed. Third, since the dictionary may find partial matches, every extracted entity is expanded until its boundaries lie next to a whitespace, tab or line-break character (Fig. 2). As proposed by Hettne et al. [2009], we check whether the entity’s boundaries are correct (no full stop or wrongly placed bracket) and remove brackets if they surround the whole entity. Finally, we remove certain suffixes at the end of entities using the list provided by Hettne et al. This list consists of common suffixes that certainly are not part of chemical entities (e.g., ‘-induced’, ‘-inhibitor’ and ‘-related’).

4 RESULTS

First, we compare ChemSpot with OSCAR3 (current Alpha 5 release) and OSCAR4 on the SCAI corpus using standard configurations for OSCAR3 and OSCAR4: The MEMM is used with a confidence threshold of 0.2. Only chemical annotations are considered; the reaction, adjective, enzyme and protein annotations of OSCAR3 and OSCAR4 are out of scope, as well as the MODIFIER annotations provided by the SCAI corpus. We then separately evaluate the performance of ChemSpot’s dictionary component. Furthermore, we provide a detailed comparison of ChemSpot’s CRF component with that of Peregrine et al. [2012] on the IUPAC test corpus. Results are shown in Figure 3 and Table 3. Finally, we compare ChemSpot to MetaMap, as other tools, such as DrugNER, are largely based on MetaMap. Further analysis will be provided in Section 5.

4.1 Evaluation on the SCAI corpus

ChemSpot achieves a precision of 67.3%, a recall of 68.9% and an F1 measure of 68.1% on the SCAI corpus. This is an increase of 10.8 percentage points in F1 measure compared with OSCAR4 (Fig. 1). The CRF component of ChemSpot alone yields a rather low recall of 28.1% on chemical entities of the SCAI corpus. This is not surprising since the CRF is focused solely on IUPAC entities. As the CRF is able to extract IUPAC entities with a very high precision of 88.3%, it is worthwhile to use it in addition to the dictionary. We shall analyse the benefit from using this hybrid approach in more detail in Section 4.2.

One reason for the comparably weak performance of OSCAR4 may be that OSCAR4 uses its own tokenizer. Kolluru et al. [2011] investigated the impact of different tokenizers on the performance of OSCAR3 and observed an increase up to 2.09 percentage points in F1 measure on the Sciborg corpus. As the performance difference to ChemSpot is 10.8 percentage points in F1 measure for OSCAR4 and 13.2 for OSCAR3 on the SCAI corpus, we expect that neither of the two would outperform ChemSpot with a different tokenization on this corpus.

4.2 Dictionary alone

Compared with Hettne et al. [2009], our dictionary-matching component achieves an increase of 9.3 percentage points in F1 measure when using the ChemIDPlus dictionary for extracting chemical entities on the SCAI corpus (Table 3). We attribute this increase in performance to our post-processing using the match expansion explained in Section 3.4. In contrast to Peregrine, ChemSpot does not perform word sense disambiguation, which leads to an increase in recall but also lowers precision. Furthermore, we suspect that the different matching mechanism of LINNAEUS combined with our match expansion is a reason for this increase in performance. Since LINNAEUS directly extracts mentions in text rather than looking up sequences of tokens in the dictionary, often partial matches from longer, unknown entities are extracted. These partial matches would probably be missed by a dictionary look up relying on coarse tokenization. In the subsequent expansion and boundary-correction step, these partial matches often lead to the extraction of the correct entire chemical entities. This is emphasized by the fact that we observed a decrease for the dictionary component of 14.9 percentage points to an F1 measure of 43.4% (precision: 42.7%, recall: 44.0%) on the SCAI corpus when turning match expansion off.

Our dictionary matcher alone already performs 1 percentage points in F1 measure better than OSCAR4 on the SCAI corpus (Fig. 1).
Precision (P), recall (R) and \( F_1 \) measure (F) for the various chemical NER approaches evaluated on the SCAI corpus. The value below the score denotes the difference to the performance of OSCAR4.

Table 3. Precision (P), recall (R) and \( F_1 \) measure (F) for the various chemical NER approaches evaluated on the IUPAC test corpus and the SCAI corpus.

<table>
<thead>
<tr>
<th></th>
<th>IUPAC training corpus</th>
<th>IUPAC test corpus</th>
<th>SCAI corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P R F</td>
<td>P R F</td>
<td>P R F</td>
</tr>
<tr>
<td>OSCAR3 (Kolářík et al.)</td>
<td>52 72 60</td>
<td>45 82 58</td>
<td></td>
</tr>
<tr>
<td>OSCAR3 (Hettne et al.)</td>
<td>41.4 81.6 54.9</td>
<td>45.7 76.5 57.3</td>
<td></td>
</tr>
<tr>
<td>OSCAR4</td>
<td>2.3 81.5 4.4</td>
<td>45.7 76.5 57.3</td>
<td></td>
</tr>
<tr>
<td>CRF (Klinger et al.)</td>
<td>X 86.5 84.8 85.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF (our impl.)</td>
<td>X 61.7 80.1 69.7</td>
<td>88.3 28.1 42.6</td>
<td></td>
</tr>
<tr>
<td>Dictionary (Klinger et al.)</td>
<td>71 37 49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dictionary (our impl.)</td>
<td>X 60.8 56 58.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChemSpot</td>
<td>X 67.3 68.9 68.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Methods in bold were evaluated by us. ‘X’ denotes a training on the corpus.

This is surprising as a dictionary matcher has the disadvantage that its performance relies mostly on up-to-date dictionary entries, an issue especially important for chemical entities where new names appear with high frequency. The high performance of the dictionary suggests that the entities in the dictionary are up-to-date or that the corpus on which OSCAR4’s MEMM was trained is outdated. As the dictionary was composed in 2009 and the SCAI corpus was published in 2008, it is more likely that the SCAI corpus is too outdated to highlight the advantage of the MEMM over the dictionary. This emphasizes again that the lack of sufficient training and evaluation corpora in the chemical domain is a severe problem for evaluating NER tools.

4.3 Evaluation of the CRF on the IUPAC test corpus

As described in Section 3, the CRF component of ChemSpot is trained on the same corpus (IUPAC training corpus) with nearly the same feature set as the system by Klinger et al. (2008). Surprisingly, our performance is much lower than that reported in Klinger et al. (2008). Since this system is not freely available, we can only speculate about the reasons. We obtain a precision of 61.7%, a recall of 80.1% and a \( F_1 \) measure of 69.7%. This is a difference of 15.9 percentage points \( F_1 \) measure compared with Klinger et al. In contrast to their third-order CRF, we use a second-order CRF and a different training method (label likelihood instead of stochastic gradient). Moreover, we do not apply bootstrapping for meta-parameter optimization (R. Klinger, personal communication). However, meta-parameter optimization increases the fit of the model to the corpus, often leading to worse performance on unseen instances (Tikk et al., 2010). Since the test corpus contains only few entities, these differences can lead to large deviations in \( F_1 \) measure.

4.4 Comparison with MetaMap

To estimate the performance difference of ChemSpot to MetaMap-based tools (e.g. DrugNER), we run MetaMap in standard configuration on the SCAI corpus using the UMLS chemical branch. MetaMap achieves a comparably low \( F_1 \) measure of 30% (precision: 23.6%, recall: 41.4%). A large number of errors, when using MetaMap out-of-the-box, are due to the fact that only partial matches of entities get extracted. This is the same problem we encountered when using LINNAEUS and that we were able to tackle by applying match expansion as explained in Section 3.4. Although MetaMap is highly configurable (Aronson, 2001), we do not believe that other configurations could bridge this performance gap. Moreover, finding good configurations would require optimization to the corpus thus increasing the danger of overfitting. Both OSCAR4 and ChemSpot clearly outperform MetaMap in their standard configuration.

5 DISCUSSION

5.1 Combining the strengths of a dictionary and a CRF

We investigated in which way ChemSpot combines the strengths of the CRF trained for the extraction of IUPAC entities and the dictionary. To analyze the coverage of both approaches, we counted...
the number of true positives and false negatives for different classes of chemical entities on the SCAI corpus and calculated the recall that was caused by the dictionary and by the CRF, respectively. Furthermore, we calculated in how many true positives both approaches agreed.

Table 4 shows that the majority of IUPAC entities can be found using the CRF. However, more than half of the IUPAC entities found by the CRF were also found using the dictionary. This is due to the fact that the dictionary also contains few IUPAC entities, in particular those that appear frequently in texts. Furthermore, our matching strategy for dictionary entries is useful to expand partial matches to IUPAC entities. Surprisingly, the dictionary extracted 60 IUPAC entities that were missed by the CRF.

Clearly, the CRF alone should solely be used to extract IUPAC entities. Matches for other classes of chemical entities are sparse (9 matches for FAMILY and 11 for TRIVIAL) and most of them are also found by the dictionary. The dictionary covers a wide range of trivial names. However, its recall for sum formulas, abbreviations and family names is considerably lower than that for trivial names.

5.2 Results for OSCAR

Comparing the results of our evaluation of OSCAR3 with results previously published by others on the same corpus shows considerable differences (Table 3). The $F_1$ measures of OSCAR3 evaluated on the SCAI corpus range from 60% (Koldaïk et al. 2009) to 58% (Heinem et al. 2009), whereas we obtained an $F_1$ measure of only 54.9%. All previous evaluations used—as we do—the standard configuration with the same scope on chemical entities, i.e. disregarding other annotations. Therefore, we believe that the difference most likely is caused by using different versions of OSCAR3.

In contrast to ChemSpot, which uses a dictionary and a CRF, OSCAR relies solely on a MEMM model. Hence, ChemSpot only requires a IUPAC-annotated corpus for training, whereas OSCAR needs one annotated corpus covering all classes of chemical entities. Training both systems on the same corpus would be an interesting experiment, but, apart from the SCAI corpus, no publicly available corpus containing IUPAC and other chemical names exists. The SCAI corpus is thus held back for evaluation purposes and is, in our opinion, too small to be divided into separate training and test sets.

Table 4. TP, FN and R of the CRF and the dictionary component for different classes of chemical entities

<table>
<thead>
<tr>
<th>Class</th>
<th>CRF TP</th>
<th>CRF FN</th>
<th>CRF R (%)</th>
<th>Dictionary TP</th>
<th>Dictionary FN</th>
<th>Dictionary R (%)</th>
<th>Shared TP</th>
<th>Shared FN</th>
<th>Shared R (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IUPAC</td>
<td>263</td>
<td>128</td>
<td>67.3</td>
<td>203</td>
<td>188</td>
<td>51.9</td>
<td>143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARTIUPAC</td>
<td>63</td>
<td>29</td>
<td>68.5</td>
<td>27</td>
<td>65</td>
<td>29.3</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAMILY</td>
<td>9</td>
<td>90</td>
<td>9.1</td>
<td>25</td>
<td>74</td>
<td>25.3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUM</td>
<td>0</td>
<td>49</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>46.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRIVIAL</td>
<td>11</td>
<td>403</td>
<td>2.7</td>
<td>317</td>
<td>97</td>
<td>76.6</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABBREVIATION</td>
<td>0</td>
<td>161</td>
<td>0</td>
<td>80</td>
<td>81</td>
<td>49.7</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>346</td>
<td>860</td>
<td>28.7</td>
<td>675</td>
<td>531</td>
<td>56</td>
<td>180</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The last column denotes the number of true positives shared by both approaches.

The results previously published by others on the SCAI corpus show considerable differences (Table 3). The $F_1$ measures of OSCAR3 evaluated on the SCAI corpus range from 60% (Koldaïk et al. 2009) to 58% (Heinem et al. 2009), whereas we obtained an $F_1$ measure of only 54.9%. All previous evaluations used—as we do—the standard configuration with the same scope on chemical entities, i.e. disregarding other annotations. Therefore, we believe that the difference most likely is caused by using different versions of OSCAR3.

In contrast to ChemSpot, which uses a dictionary and a CRF, OSCAR relies solely on a MEMM model. Hence, ChemSpot only requires a IUPAC-annotated corpus for training, whereas OSCAR needs one annotated corpus covering all classes of chemical entities. Training both systems on the same corpus would be an interesting experiment, but, apart from the SCAI corpus, no publicly available corpus containing IUPAC and other chemical names exists. The SCAI corpus is thus held back for evaluation purposes and is, in our opinion, too small to be divided into separate training and test sets.

Table 5. Error analysis of 50 randomly sampled false negatives missed by ChemSpot on the SCAI corpus

<table>
<thead>
<tr>
<th>Error type</th>
<th>False negatives, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial match</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Annotation error</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Not in dictionary/recognized</td>
<td>36 (72)</td>
</tr>
<tr>
<td>Tokenization error</td>
<td>4 (8)</td>
</tr>
</tbody>
</table>

Table 6. Error analysis of 50 randomly sampled false positives extracted by ChemSpot on the SCAI corpus

<table>
<thead>
<tr>
<th>Error type</th>
<th>False positives, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial match</td>
<td>15 (30)</td>
</tr>
<tr>
<td>Annotation error</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Out of corpus scope</td>
<td>23 (46)</td>
</tr>
<tr>
<td>Not a chemical</td>
<td>8 (16)</td>
</tr>
</tbody>
</table>

5.3 Evaluation on the DDI corpus

We also compared ChemSpot and OSCAR4 on the entities of the DDI corpus (Segura-Bedmar et al. 2010). We found that ChemSpot (precision 80.1%, recall 55.7% and F1 measure 65.7%) outperforms OSCAR4 (precision 70.7%, recall 50.0% and F1 measure 58.6%) by 7.1 percentage points F1 measure even when the latter’s confidence threshold is optimized for the SCAI corpus.

5.4 Error analysis

To assess the frequency and types of errors for ChemSpot, we randomly sampled 50 false negatives (Table 5) and 50 false positives (Table 6) on the SCAI corpus and performed a manual error classification using the scheme from Heinem et al. (2009).

The main reasons why ChemSpot missed entity mentions were either their absence in the dictionary or the fact that they were not recognized by the CRF (72%). Those false negatives mainly fall into

Table 4. TP, FN and R of the CRF and the dictionary component for different classes of chemical entities
We applied ChemSpot to all abstracts from PubMed (as of November chemidplus/chemidlite.jsp last accessed 2012-01-20). ID are linked to ChemIDPlus Light (http://chem.sis.nlm.nih.gov/hu-berlin.de/ last accessed 2012-01-20). Furthermore, GeneView SNPs, histone modifications, species, etc (http://bc3.informatik).PubMed with automatically derived annotations of genes/proteins, entities, whereas a dictionary is useful for extracting the often short treating different classes of chemical entities is highly beneficial. ChemSpot’s boundary correction from removing the surrounding brackets.

We found 46% of the false positives to be larger molecules that commonly are not considered as chemicals, especially protein drugs (Table 8). These were not annotated in the corpus and are therefore out of corpus scope. However, we believe that entities in this error class might be of interest for some chemical NER applications. A partial match could be achieved for 30% of the false positives (e.g. “tetratoze” instead of “tetratoze acid”), which shows that post-processing deserves further attention in future versions of ChemSpot. 8% are annotation errors, i.e. chemicals that in our opinion were omitted or marked incorrectly in the corpus. We found 16% of the false positives to be not a chemical, of which most were caused by homonymous abbreviations. For instance, “CPT” was extracted as abbreviation for ‘camptothecin’, whereas the abstract was in fact about ‘…cumulative prospect theory’. Clearly, such errors call for including appropriate methods for word sense disambiguation.

5.5 GeneView: application of ChemSpot on PubMed

We applied ChemSpot to all abstracts from PubMed (as of November 2011), yielding 73 883 960 entities in 9 861 936 publications. Of these entities, 61 316 472 (83%) could be mapped to 59 255 distinct CAS Registry IDs. All annotations are publicly available through GeneView (Thomas et al., 2010), a tool for searching PubMed with automatically derived annotations of genes/proteins, SNPs, histone modifications, species, etc (http://bc3.informatik.hu-berlin.de/ last accessed 2012-01-20). Furthermore, GeneView can be used to search MEDLINE publications by CAS Registry IDs. Additionally, chemical annotations with a CAS Registry ID are linked to ChemIDPlus Light (http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp last accessed 2012-01-20).

6 CONCLUSIONS AND FUTURE WORK

We introduced ChemSpot, a hybrid system for extracting chemical entities from natural language texts. ChemSpot is based on a CRF trained for identifying IUPAC entities and a dictionary built from ChemIDPlus for extracting drugs, abbreviations, molecular formulas and trivial names. Evaluations showed a major performance advantage compared with the only other freely available NER tool for chemical entities, OSCAR4. Thus, we believe that ChemSpot sets a new state-of-the-art in the recognition of chemical entities.

We conclude that using a hybrid NER approach for adequately treating different classes of chemical entities is highly beneficial. CRF's are suitable for extracting morphologically rich IUPAC entities, whereas a dictionary is useful for extracting the often short and erratically structured names of drugs, trivial names, etc. By combining the strengths of both approaches, ChemSpot achieves a major increase in performance and a broad coverage of chemical entities.

Future work will focus on the normalization of chemical entities to known identifiers. So far, ChemSpot assigns a CAS Registry ID only to entities that were extracted by the dictionary component and not changed during match expansion. We plan to address this drawback in future releases of ChemSpot. Furthermore, we aim at improved performance for abbreviations and sum formulas by applying rule-based NER methods.

ACKNOWLEDGEMENTS

We would like to thank Philippe Thomas for fruitful discussions and for importing ChemSpot’s annotations into GeneView. We thank him and Samira Jaeger for their helpful comments on the manuscript. Furthermore, we thank Christoph Jacob for providing us with MetaMap annotations for the SCAI corpus.

Funding: This work was undertaken as part of the Virtual Liver Network, funded by the German Ministry for Education and Research (BMBF) [0315746].

Conflict of Interest: none declared.

REFERENCES


