Recognizing Chemical Compounds and Drugs: a Rule-Based Approach Using Semantic Information.

Sara Lana-Serrano\textsuperscript{1,4}, Daniel Sanchez-Cisneros\textsuperscript{2}, Leonardo Campillos\textsuperscript{3}, and Isabel Segura-Bedmar\textsuperscript{2}

\textsuperscript{1} Universidad Politécnica de Madrid, slana@diatel.upm.es,\textsuperscript{2} Universidad Carlos III de Madrid, \{dscisner, isegura\}@inf.uc3m.es,\textsuperscript{3} Universidad Autónoma de Madrid, leonardo.campillos@uam.es \textsuperscript{4} DAEDALUS, S.A

Abstract. This paper presents a system for recognizing chemical compounds and drug names. It is a rule-based system that utilizes semantic information from the ChEBI ontology and the MeSH Metathesaurus. It also integrates the MetaMap tool, the ANNIE PoS tagger, and pharmacological databases such as DrugBank. We used this system for the CHEMDNER task 2013, and an outcome of this work is the development of non-existing resources for recognizing chemical entities (e.g., gazetteers and a list of biochemical affixes), which are available for the research community.

Key words: Drug Named Entity Recognition, Information Extraction

1 Introduction

Dictionary-based methods and Supervised Machine Learning (SML) methods are the main approaches that have been used for Drug Named Entity Recognition (DNER). Dictionary-based methods focus on matching terms (which are compiled from dictionaries) to text. Their main advantage is that they are relatively easy to implement. Two major drawbacks of dictionary-based approaches are their domain dependency and their inability to recognize terms that are not included in the system dictionaries. Supervised Machine Learning methods use manually tagged corpora to automatically build rule-based systems or sequence labeling algorithms. These techniques have produced state-of-the-art results [1]. However, the most important drawback of supervised approaches is that the manual annotation of corpora is a painstaking, time-consuming process, and therefore, annotated resources are scarce. In this paper we present a hybrid approach that combines the ChemSpot tool [2] and a set of hand-written rules based on lexical and semantic information for constraints.

We used this system for the CHEMDNER task 2013. The task consisted of two subtasks: a chemical document indexing subtask (CDI) and a chemical entity mention recognition subtask (CEM). Further information can be found in [5].
2  System description and methods

The architecture of the system consists of a pipeline with eight stages (Fig. 1). The main engine is made up of a rule-based system developed with GATE tool \(^5\). The rules were composed on the basis of the CHEMDNER annotation guidelines for the task\(^5\), and the error analysis performed in the training phase.

![System architecture diagram](Image)

*Fig. 1. System architecture.*

2.1  Entity recognition with the ChemSpot tool

In the first phase, the corpus is processed by ChemSpot. This tool identifies mentions of chemicals in texts, including trivial names, drugs, abbreviations, and molecular formulas. Afterwards, each entity found is expanded with semantic information (e.g. UMLS semantic types\(^6\) or ChEBI [3] ancestors, among others).

2.2  Expanding semantic information from ChEBI

In this step, the information is expanded using semantic knowledge extracted from the ChEBI ontology. The system finds ancestor paths of each identified entity to extract semantic knowledge. We used the following relations to find out the ancestors: *is-a*, *has-role*, *is-conjugate-acid-of*, and *is-conjugate-base-of*.

\(^5\) http://gate.ac.uk/

\(^6\) http://www.nlm.nih.gov/mesh/meshhome.html
Then, the annotation of each entity is expanded to include the labels of its ancestors (e.g., molecule, mineral, metal, macromolecule, lipid etc). The labels provide useful evidence to evaluate the annotation rules defined in the CHEMDNER annotation guidelines. These labels are also used later in the rule-based processing stages.

2.3 Expanding semantic information from MeSH

In this stage, the annotation is expanded with knowledge from the MeSH Metathesaurus, by adding the following features:

- **MeSH_SemanticType**: classification according to the UMLS semantic types.
- **MeSH_Type**: descriptor record or supplementary chemical record.
- **MeSH_TreeNumbers**: the MeSH tree-structure indexes of the entity. This information is only associated to the descriptor record.

The output of this phase is a XML document containing ChemSpot annotations that are enriched with semantic information from the ChEBI and the MeSH resources. This XML document is the input file for the following stages of the pipeline, which is processed through the GATE tool.

Fig. 2 shows an example of semantic expansion. The first chemical, 3-carboxylic acid, was recognized by ChemSpot, but does not contain any annotation from MeSH or CheBi, since these resources do not include 3-carboxylic acid. The second one, ciprofloxacin, contains semantic information that was expanded using MeSH, but was not found in ChEBI. Finally, the third entity, boron, contains semantic information that was expanded from the MeSH and the ChEBI resources.

![Fig. 2. Example of semantic information expansion.](image)

2.4 Analysis with the MetaMap tool

The corpus is processed with the MetaMap tool [4], which finds concepts from the UMLS Metathesaurus by a shallow syntactic analysis. The output is a set
of noun phrases that are then used to generate candidate variants of the UMLS concepts. MetaMap greatly enriches the annotation. Nevertheless, we only select those candidates with score lower than -700 and that are tagged with at least one of the following semantic types: aapp, antb, bodm, carb, crbs, chvs, clus, drdd, eico, elii, enzy, grup, hops, horm, inch, irda, lbtr, lipd, mosq, nsba, nnon, opco, orch, phsu, strd, vita, imft. These semantic types were proposed on the basis of the observation of the training dataset. We analyzed this dataset with the MetaMap tool and selected the most frequent semantic types which were assigned to the chemical compounds and drugs entities in the training dataset.

2.5 Gazetteers

This phase aims to identify false positive and false negative instances (Fps and Fns, respectively). We apply a gazetteer tagger that is based on the rules put forward in the CHEMDNER guidelines. This phase collects 27 gazetteers that contain more than 340,000 entries. They have been compiled from lists of chemical and gene entities included in the following resources: MeSH, DrugBank, ChEBI, Wikipedia, ChemSpot. The texts were processed by these gazetteers in order to rule out FPs and to annotate FNs that were not recognized by the previous modules.

2.6 PoS tagging

The corpus is processed with the ANNIe PoS tagger. PoS tags may be useful to rule out FPs (e.g. verbs in past participle and present forms should not be annotated as entities). Thus, PoS tags were used in the definition of advanced lexical rules (see Section 2.7).

2.7 Lexical-semantic rules

This phase applies a set of positive and negative rules that were developed considering the CHEMDNER guidelines. There are two kinds of rules: semantic rules, and lexical-semantic rules.

The semantic rules classify the FP and FN instances according to the labels that contain the semantic information (expanded from ChEBI and MeSH resources). An example of a semantic rule is that all instances labeled with the semantic type 'plant' (e.g. cactus) should be ruled out.

The lexical-semantic rules identify CEM entities by applying the following processes:

- Discard CEM entities that contain FN evidence. FN evidences are recognized through gazetteers (see Section 2.5), regular expressions, and rules established in the CHEMDNER guidelines.

7 http://www.drugbank.ca/
8 http://wikipedia.org
9 http://gate.ac.uk/sale/tao/splitch6.html#chap:annie
– Discard CEM entities that are included in other CEM entity.
– Identify the longest nested CEM entities.

Some examples of lexical-semantic rules are shown below.

\[
\begin{align*}
\text{IF (Entity.ChEBI\_Class CONTAINS "monosaccharide") THEN add(Entity)} \\
\text{IF (Entity.ChEBI\_Class CONTAINS "peptidomimetic") THEN add(Entity)} \\
\text{IF (Entity.MeSH\_SemanticTypes NOT CONTAINS "aapp") THEN add(Entity)} \\
\text{IF (Entity.MeSH\_SemanticTypes NOT CONTAINS "eico") THEN add(Entity)} \\
\text{IF (Entity MATCH [a-z][RDXGTHLP]NA[a-z]*) THEN delete(Entity)} \\
\text{IF (Entity1 CONTAINS (Entity2)) THEN delete(Entity2)}
\end{align*}
\]

2.8 Advanced lexical rules

Linguistic methods (e.g. lexicons, ontologies, and roots and affixes) have already been applied for processing information in the pharmacological domain [6] [7]. Our system also applies a collection of advanced lexical rules that classify the entities according to PoS analysis, affix processing, and multi-word processing.

In this stage, the advanced lexical rules filter FP instances that were identified in the error analysis phase. These FP instances may be present participle and past participle verbs. For example: "[...] heart tissue of the rats poisoned by aluminum phosphide.", "[...] then complexed with anti-Ask1 shRNA". The advanced lexical rules also use a list of 663 affixes (considering spelling variations: e.g. sulf- and sulfa-). This list includes the following data:

– Roots and affixes that refer to chemical (e.g. propyl-, -phosphate) and biochemical entities (e.g. -sterol). We did not include affixes with less than 2 characters (e.g. -yl) to prevent possible FP.
– Stems for the recognition of pharmacological substances (e.g. -cavir).

These affixes were curated from lists approved by the American Medical Association (AMA) for the nomenclature of clinical compounds [10], [11] and stems proposed by the World Health Organization (WHO) [8], [9], and other listings of biochemical affixes [12]. However, we ruled out certain affixes to avoid FPs. For example, those that match substances excluded from the CHEMDNER task: e.g. -ase (for enzymes such as kinase or ATPase) and -teplase (which would match large-size proteins such as alteplase). Other affixes that were rejected for that reason were -globin, insulin-, prote-, renin-, thrombin-, tryps-, or -uplase.

Very general affixes (e.g. trans-) were also discarded. For example, trans- would successfully recognize trans-stilbene, but this prefix would also match other words such as transformation or translate. Other general affixes that were not included were hemi-, iso-, proto- or mono-.

In fact, FPs are the main disadvantage of using a list of affixes. For example, the prefix but- may match butyric, but also the conjunction but. Another example

\[10\] http://www.ama-assn.org/resources/doc/usan/stem-list-cumulative.pdf
\[12\] These lists were gathered by Michael Quinion (2008): www.axes.org
is the suffix acetyl-/acetyl, which may identify acetylsalicylic acid or chloroacetil chloride, but also the enzyme acetyltransferase, which is a type of biological entity excluded from the CHEMDNER task. Other affixes from our list that may select FPs are glyco-/glyco (that match glycerol, but glycoprotein), -oidal (steroidal, but colloidal), or cyclo- (cyclosporin, but cyclohexrin).

Finally, the advanced lexical rules process multi-word entities (especially, formulas) that are formed by a combinations of other entity names concatenated with whitespaces, numbers, and symbols (+, -). For example: tertiary 2-(3-hydroxyphenyl)-2-phenethylamine

The following are some examples of rules:

\[
\text{IF ( Entity.ChEBI\_Role CONTAINS "CHEBI:27311" AND Entity.ChEBI\_Class CONTAINS "acid") THEN add(Entity, "ChEBI")}
\]

\[
\text{IF ( Token.kind == "word" AND NEXT IS SpaceToken AND NEXT IS (Token.string CONTAINS "glycoside") THEN add(Token, "Merge")}
\]

The system assigns a score to each entity depending on the rules satisfied. For example, if ChemSpot recognizes an entity whose labels satisfy the identification criteria with semantic expansion, that entity obtains a score of 1. In contrast, if the expanded labels gather only evidence from affixes, the classification is rather uncertain. In this case, the entity may obtain a score of 0.4.

Finally, the output of the rule-based module is processed to create the CDI and CEM format file.

3 Discussion

3.1 Description of the runs

The following are the runs that we carried out, and the combination of modules for each run.

- Run 1: ChemSpot + ChEBI + MeSH + lexical-semantic rules
- Run 2: ChemSpot + ChEBI + MeSH + gazetteers + lexical-semantic rules
- Run 3: ChemSpot + ChEBI + MeSH + MetaMap + gazetteers + lexical-semantic rules
- Run 4: ChemSpot + ChEBI + MeSH + gazetteers + lexical-semantic rules + advanced lexical rules
- Run 5: ChemSpot + ChEBI + MeSH + MetaMap + gazetteers + lexical-semantic rules + advanced lexical rules

3.2 Results

The task involved two subtasks: a Chemical document indexing subtask (CDI evaluation) and a Chemical entity mention recognition subtask (CEM evaluation).
Tables 1 and 2 show the results for the CDI and CEM evaluation, respectively. Our system achieved similar scores in both evaluations. In the CDI evaluation, our results (Micro F-scr) ranged between 0.452 and 0.562 in the development dataset. In the CME evaluation, the results (Micro F-scr) ranged between 0.469 and 0.573 on the development dataset.

<table>
<thead>
<tr>
<th>Run</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8475</td>
<td>5583</td>
<td>7620</td>
<td>0.60286</td>
<td>0.52656</td>
<td>0.56213</td>
</tr>
<tr>
<td>2</td>
<td>8907</td>
<td>6772</td>
<td>7188</td>
<td>0.56808</td>
<td>0.55340</td>
<td>0.56065</td>
</tr>
<tr>
<td>3</td>
<td>9999</td>
<td>18056</td>
<td>6096</td>
<td>0.35641</td>
<td>0.62125</td>
<td>0.45296</td>
</tr>
<tr>
<td>4</td>
<td>8773</td>
<td>7631</td>
<td>7322</td>
<td>0.53481</td>
<td>0.54508</td>
<td>0.53989</td>
</tr>
<tr>
<td>5</td>
<td>9262</td>
<td>15559</td>
<td>6833</td>
<td>0.37315</td>
<td>0.57546</td>
<td>0.45273</td>
</tr>
</tbody>
</table>

Table 1. Results from the system on the CHEMDNER 2013 corpus over the CDI evaluation.

<table>
<thead>
<tr>
<th>Run</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15099</td>
<td>8019</td>
<td>14427</td>
<td>0.65313</td>
<td>0.51138</td>
<td>0.57363</td>
</tr>
<tr>
<td>2</td>
<td>15890</td>
<td>9679</td>
<td>13636</td>
<td>0.62146</td>
<td>0.53817</td>
<td>0.57682</td>
</tr>
<tr>
<td>3</td>
<td>17553</td>
<td>24775</td>
<td>11973</td>
<td>0.41469</td>
<td>0.59449</td>
<td>0.48857</td>
</tr>
<tr>
<td>4</td>
<td>15500</td>
<td>10559</td>
<td>14026</td>
<td>0.59480</td>
<td>0.52496</td>
<td>0.55770</td>
</tr>
<tr>
<td>5</td>
<td>15493</td>
<td>20990</td>
<td>14033</td>
<td>0.42466</td>
<td>0.52472</td>
<td>0.46942</td>
</tr>
</tbody>
</table>

Table 2. Results from the system on the CHEMDNER 2013 corpus over the CEM evaluation.

In both evaluations, the best performance (Run 1) was obtained using only the ChemSpot tool, the semantic expansions from ChEBI and MeSH, and the lexical-semantic rules. Using additional information from the other stages (analysis with MetaMap, gazetteers, and PoS tagging) decreased the performance.

On the other hand, it seems that the analysis performed by the MetaMap tool (run 3 and run 5) allowed increasing the recall of the system, but with a large decrease in the precision. This may be due to the wide range of semantic types proposed to recognize chemical compounds and drugs.

We performed an error analysis on the development test in order to obtain information about the accuracy of each module. For example, the semantic expansion with ChEBI achieved to recognize 44.1% of the entities, while MeSH provided a better coverage with more than 50%. Both semantic expansions produced a low number of FPs (60 FPs were proposed by ChEBI and 74 by MeSH). In contrast, ChemSpot generated 405 FPs.

3.3 Conclusions

In this paper we have described a rule-based system to detect chemical compounds and drug names. The system was evaluated at the CHEMDNER task 2013. A demo of the system is available via web.

Our main contribution has been the development of an important amount of resources (27 gazetteers with more than 340,000 entries and a list of 663 affixes, http://multimedica.uc3m.es:8080/biocreative2013demo/
among others). All these resources are available on the MultiMedica project web site.\footnote{http://labda.inf.uc3m.es/multimedica/CHEMDNER2013team225resources.html}

We have explored the use of semantic information through domain-specific resources such as ChEBI and MeSH. Nevertheless, we realized that expanding semantic knowledge without using a domain-specific filter caused a considerable decrease in precision in our results.

\textbf{Acknowledgments.} This work was funded by the MA2VICMR (S2009/TIC-1542) and the MultiMedica projects\footnote{http://labda.inf.uc3m.es/multimedica/} (TIN 2010-20644-C03-01).

\textbf{References}