Recognition of chemical entities in patents using LeadMine

Daniel M. Lowe*[1], Roger A. Sayle[2]

NextMove Software Ltd, Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge, United Kingdom

*[1]daniel@nextmovesoftware.com;
[2]roger@nextmovesoftware.com

Abstract. LeadMine is a dictionary/grammar based approach to entity recognition. For chemical entities, hand-written grammars are used to recognize systematic chemical names and formulae. Trivial names are found using dictionaries, some derived from public sources and some hand curated. A rule-based method is used to detect abbreviations of identified entities. To improve the system's performance on patents, improvements were made to the grammars for families of chemical compounds, and for chemical formulae, especially those containing R groups. Additionally a step was added where short ambiguous formulae, frequently used in Markush descriptions, are recognized e.g. C, N, O. Post-recognition certain terms are trimmed from entities for better agreement with the annotation guidelines e.g. "heterocyclic" instead of "heterocyclic ring". For genes/proteins, a dictionary-based recognizer was developed, using terms from Uniprot, EntrezGene and HGNC. Our system achieved F_1-scores of 85.2% for chemicals and 75.2% for genes/proteins on the test set.

Keywords. LeadMine; CHEMDNER-Patents; grammars

1 Introduction

Chemical patents contain a wealth of chemical and biochemical knowledge. CHEMDNER-Patents is a community challenge to evaluate and encourage the development of tools to extract information from patents, with a specific focus on the extraction of chemicals, genes and proteins. The corpus is composed of 21,000 manually annotated patent abstracts, of which 7,000 form the test corpus. Further information about the challenge is available in the challenge task paper[1].

2 Discussion

Our submission for CHEMDNER-Patents builds on our submission[2] to BioCreative IV's CHEMDNER task. As compared to PubMed abstracts, patents far more frequently mention novel compounds, especially families of novel compounds. Within patent abstracts these families of compounds may be eluded to by systematic names in which some parts are replaced by generic groups e.g. alkyl, heteroaryl etc. Alternatively the abstract may include a Markush structure to fully define the scope of the family. In this case the groups/atoms allowed in embodiments of the invention are frequently described using formulae and/or systematic substituents. These in turn can also be chemical classes e.g. NHR_1, C_1-C_4 alkoxy.

2.1 Formula recognition

We have developed a grammar capable of recognition and parsing systematic chemical line formulae e.g. CH_3CH_2OH. The grammar knows the expected valency of atoms, hence allowing determination of whether a group is a complete molecule (e.g. CH_4), substituent (e.g. CH_3) or a linker (e.g. CH_2). Recognized formulae can be converted to a parse tree in which the morphemes of the formula are related to SMILES[3] and nodes in the tree indicate the operation required to interpret the tree. This parse tree is then used to construct the structure described by the formula. Some of the supported features are demonstrated in Figure 1.

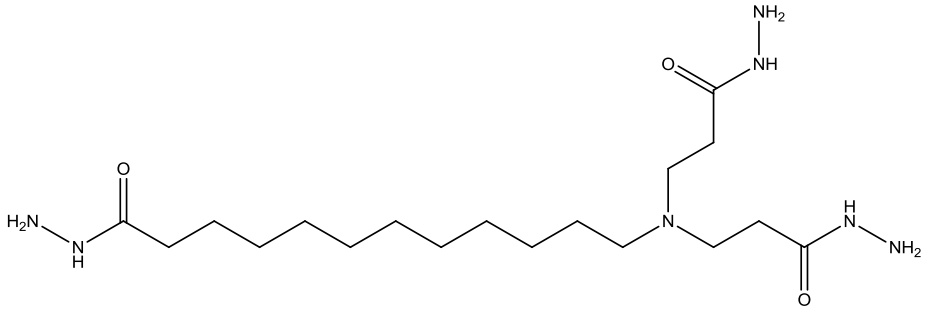

Figure 1: A more complicated formula demonstrating a repeated substituent, repeated linker and inferred double bonds: $(NH_2NHCOCH_2CH_2)_2N(CH_2)_{11}CONHNH_2$

In formulae encountered in patent abstracts, it is not uncommon for there to be references to another definition or a class of group e.g. R1, Ar, Het, X. These are supported by the grammar but interpreted as a

placeholder atom when the formula is converted to SMILES[3]. Similarly variable length alkyl groups are supported, with the same limitation.

2.2 Ambiguous formula recognition

Markush descriptions frequently contain lists of substituents, linkers or heteroatoms. These terms are often short formulae e.g. F, H, O, N and S. As these terms are ambiguous out of context, we use a grammar that finds cases where an ambiguous formula is preceded or followed by a certain term. If the formula is preceded by a key phrase (e.g. "chosen from", "selected from", "preferably"), then the formula is annotated. If the formula is preceded by a term indicating that it appears to be in a list (e.g. ",", "or"), then the formula is noted as a potential formula. Potential formulae are annotated if they are within 4 characters of a recognized chemical entity. As typically the first item in a list will be preceded by a key phrase this confidence will be passed onto each adjacent term in the list allowing recognition of entities in text like Figure 2.

R1 is selected from the group **consisting of** H, F, Cl, Br, I and NO2

Figure 2: Key phrase (bold), ambiguous formulae (red), unambiguous formula (green)

2.3 Final Chemical Entity Mentions (CEM) system

Two include and two stop lists were trained from the union of the training and development sets. The methodology used was to iterate through each false positive and false negative in turn and measure the effect on F_1-score. If a term improved F_1-score it was added into the case insensitive include/stop list, otherwise this was repeated with the term matched case sensitively. If the term then improved F_1-score it was added to the case sensitive include/stop list.

After recognition entities were trimmed for better agreement with the annotation guidelines. Sometimes the result is semantically equivalent e.g. "heterocyclic" implies "heterocyclic ring". However in others (e.g. "lower alkyl", "branched chain alkyl", "substituted alkyl"; trimmed to "alkyl") the resultant entity contains less structural information. We

suggest these cases should be reconsidered in future versions of the annotation guidelines.

The five runs submitted were:
1. Default system
2. Include list trained from training/development sets
3. Stop list trained from training/development sets
4. Include and stop lists
5. Include list with manual corrections

As in the CHEMDNER task our 2nd run gave the best result, the 5th run is a variant of the 2nd in which the include list was manually inspected for terms that should not have been annotated in the corpus e.g. Diuretic (pharmacological effect), Bevacizumab (antibody), isopropyl. (punctuation after entity), podophyllum (plant). The lattermost of these errors stems from a poor translation in the abstracts derived from Chinese patents. Current machine translation (Google Translate) correctly determines that a compound derived from the plant, rather than the plant itself, is being described. Other issues of this type are relatively frequent in text from Chinese patents highlighting that results may be quite dependent on the software used to produce the translation.

2.4 Chemical passage detection (CPD)

Chemical passage detection is a by-product of chemical entity recognition i.e. if text is annotated with at least one chemical it is a chemical passage. Confidence scores were calculated using the formula $\frac{EntityCount \times 25}{Characters\ of\ text}$ for chemical passages and $\frac{Characters\ of\ text}{2000}$ for non-chemical passages, in both cases capped at 1. A typical chemical entity is 25 characters and a long abstract is 2000 characters i.e. a document consisting entirely of chemical entities, or a long abstract with no chemicals are considered to be the examples of perfect confidence, for chemical and non-chemical containing, respectively. Runs were the same as for CEM.

2.5 Gene/Protein recognition (GPRO)

To allow comparison to the current state of the art in gene/protein recognition, we prepared a system using dictionary-based recognition.

The dictionary was derived from terms in Uniprot[4], EntrezGene[5] and HGNC[6]. A mixture of stop words and regexes were used to exclude incorrect terms e.g. genetic diseases, common English words. In cases where a term clashed with an English word, but differed in case, it was added to a case sensitive dictionary. In cases that were still expected to be ambiguous, the term with the word "protein" or "gene" appended was used. Simple variants of terms were generated e.g. hyphenation, Greek characters.

The case insensitive and case sensitive dictionaries contained 14,615,035 and 49,643 distinct terms, respectively. As many terms are very similar to each other the representation used for matching is significantly smaller than the original terms. While not required for this task, use of this system for normalization is straightforward as each term is directly related to a database identifier.

A dictionary of 1354 protein class names was collected in an ad hoc manner from Wikipedia and the training/development sets. This dictionary was used as a stop list. Include and stop lists were trained from the training/developments sets in the same manner as was done for the CEM task. The same abbreviation detection algorithm as was used for CEM was used to find abbreviations entities.

The five runs submitted were:
1. Default system
2. Include list trained from training/development sets
3. Stop list trained from training/development sets
4. Include and stop lists
5. Include and stop lists with more aggressive spelling correction

6. Evaluation

Chemical Runs	Precision	Recall	F_1-score	CPD AUC_PR
Run 1	83.15%	85.39%	84.25%	94.91%
Run 2	82.90%	87.68%	85.22%	94.64%
Run 3	85.01%	83.59%	84.29%	95.52%
Run 4	84.42%	85.87%	85.14%	95.27%
Run 5	82.93%	87.66%	85.23%	94.66%

As anticipated the best performing runs used include lists trained from the training/development corpora (Runs 2 and 5) with manual cleanup of these lists having a negligible positive effect (Run 5). Chemical passage detection favored the improved specificity offered by the stop lists (Run 3).

Gene/Protein Runs	Precision	Recall	F_1-score
Run 1	71.99%	68.19%	70.04%
Run 2	72.91%	76.42%	74.63%
Run 3	78.80%	63.57%	70.37%
Run 4	78.68%	72.03%	75.20%
Run 5	78.53%	72.20%	75.23%

For genes/proteins the best performing runs used both include and stop lists trained from the training/development corpora (Runs 4 and 5), with more aggressive spelling correction having a negligible positive effect (Run 5). While most improvement came from the include lists, the stop lists did improve performance implying that a significant number of names that are either ambiguous, incorrect, or refer to classes of proteins, remain in the dictionary. As all terms are intended to refer to specific genes/proteins, investigation of these cases will be useful to avoid normalizing a class of proteins to a specific protein's identifier.

REFERENCES

1. Krallinger M, et al. (2015) Overview of the CHEMDNER patents task. Proceedings of the fifth BioCreative challenge evaluation workshop

2. Lowe DM, Sayle RA (2015) LeadMine: A grammar and dictionary driven approach to entity recognition. Journal of Cheminformatics 7:S5.

3. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36.

4. UniProt. http://www.uniprot.org/. Accessed 3 Sep 2015

5. NCBI EntrezGene. http://www.ncbi.nlm.nih.gov/gene/. Accessed 3 Sep 2015

6. HUGO Gene Nomenclature Committee HGNC database of human gene names. http://www.genenames.org/. Accessed 3 Sep 2015