Mixing Statistical and Symbolic Approaches for Chemical Names Recognition

Florian Boudin[♮], Juan Manuel Torres-Moreno[♮,♭] and Marc El-Bèze[♮]

[♮]Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228
84911 Avignon Cedex 9, France
[♭] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada.
{florian.boudin,juan-manuel.torres,marc.elbeze}@univ-avignon.fr
http://www.lia.univ-avignon.fr

Abstract. This paper investigates the problem of automatic chemical Term Recognition (TR) and proposes to tackle the problem by fusing Symbolic and statistical techniques. Unlike other solutions described in the literature, which only use complex and costly human made ruled-based matching algorithms, we show that the combination of a seven rules matching algorithm and a naïve Bayes classifier achieves high performances. Through experiments performed on different kind of available Organic Chemistry texts, we show that our hybrid approach is also consistent across different data sets.

Key words: Term Recognition, Text Mining, Chemical Informatics.

1 Introduction

Over one million new chemical compounds are discovered and published annually. As in many scientific domains, the Organic Chemistry (OC) data are not published coherently but scattered through thousands of different journal articles. Identifiying and extracting chemical compounds is a critical task for chemical information retrieval. Information extraction technology arose in response to the need for efficient processing of documents in specialized domains. Classical Natural Language Processing (NLP) tools such as parsers, taggers or chunkers achieve very poor on OC documents. This is due to the specificity of the domain, a very wide vocabulary, long sentences containing a high quantity of "hapax legomen"[1]. Scientists, especially chemists, want to be able to search for articles related to particular chemical compounds. Nowadays, search engines mainly depend on the "classical" title, author(s) and keywords scheme searching. Extracting chemicals from texts and using them to classify, organize and accelerate the information access fit to a wide range of possible applications.

[1] Terms which only appears once in a text.

Chemical compounds are, in articles, identified by verbal depictions (i.e. name, identifiers, formulae) but also pictorial depictions (chemical structure representations). From the analysis of several articles we have found that most chemical compounds can be automatically extracted by examining chemical texts and verifying the presence of specific patterns. In this work, we propose an hybrid approach combining pattern matching and probabilistic classification. This paper is organized as follows. Section 2 overviews the related work, section 3 defines what we consider as a chemical compound. The two approaches and their combination are described in section 4. Experimental settings are presented in section 5 followed by the results while the section 7 concludes this paper.

2 Related Work

Nowadays, the majority of information extraction approaches in the life sciences have focused on molecular biology and genomics information so far [2]. Only a very limited number of named entity recognition approaches are described in the literature for the recognition of chemical compounds. A rule-based method was introduced by [4]. This approach was tested only on a very small benchmark set (158 chemical terms to be identified, f-measure between 0.7619 and 0.8169, see section 5.3 for details on performance measures). Other systems used simple dictionary matching without any evaluation of the performance [9]. Chemical Formulae extraction using Support Vector Machines (SVM) classification [10] and reconstruction of molecular structure by analyzing chemical terminology [6] have also been tried. These approaches tackle the issue of a different problem. As far as we know, there are no current published works on the adaptation of such statistical text mining techniques to process organic chemical papers.

3 What is a Chemical Compound?

One of the most difficult part is to define what is a chemical compound and what it is not. We have to cope with a large variety of syntactical and semantically different compound description. The International Union of Pure and Applied Chemistry (IUPAC)[2] is mostly well-known as the recognized authority in developing standards for the naming of the chemical elements and their compounds, through its Interdivisional Committee on Terminology, Nomenclature and Symbols (ICTNS). The IUPAC nomenclature is a useful resource for naming chemical compounds and for describing the science of chemistry in general. Chemicals can be described in literature by trivial names (e.g. brand or trade names), by registry numbers (e.g. database identifiers), by systematic naming schemes (e.g. nomenclature such as IUPAC [5] or formal descriptions like SMILES [11]) and by chemical structure depictions. Rules for naming organic compounds are contained in one publication, known as the Blue Book [7]. Compounds are named by

[2] http://www.iupac.org

using a number of prefixes, suffixes and infixes that support very precise information about them (i.e. type and position of functional groups, priority, etc...). For example, the compound `2-methylpropane` is composed by the root names `prop-` and `meth-` corresponding to the number of carbons in the main chain and the attached chain respectively. The main chain is a propane chain and a methyl group is bonded (attached) to the middle (2) carbon, these specifications give the systematic name: `2-methylpropane`. In articles, `2-methylpropane` is commonly called as `isobutane` but can also be `(CH3)2CHCH3`. To illustrate the large variety of synonyms, the chemical `2-methylpropane` has officially 12 synonyms (in which `Trimethylmethane`; `1,1-Dimethylethane`; `iso-C4H10`; `i-Butane`; `Isobutane mixtures`; `tert-Butane`; `Methylpropane`; `2-methyl-isobutane` `Propane`) but the number of variants can be as high as several hundred. All these variants correspond to the same compound and have to be identified. This example gives a flavour of the tremendous difficulty of the task.

4 An Hybrid Approach

In this section, we describe two different approaches we used for chemical names identification and we explain why we choose to combine them.

4.1 Pattern Matching

The first approach consists in manually writing a small pool of patterns based on the Blue Book nomenclature. The system skims through the document verbatim and tries to capture the chemical compounds. The presence of specific prefixes, suffixes, infixes, numbers and special characters (such as brackets or Greek letters) in a term allows our system to identify facile terms (e.g. high probability to be a chemical name). We consider a term T as a token separated by two spaces. The score S_{pm} of a term T to be a chemical compound is calculated as:

$$S_{pm}(T) = \sum_{j=0}^{N} Match_j(T)$$

$$Match_j(T) = \begin{cases} \omega_j \text{ if the pattern } j \text{ match the term } T \\ 0 \text{ else} \end{cases}$$

N is the total number of rules/patterns, $\sum_j \omega_j = 1$ and $\omega_j \in [0, 1]$. Assuming a uniform weights distribution (i.e. weights ω_j are equally spread according to the number of rules), a term is considered to be a chemical compound if at least one rule is matching (i.e. if $S_{pm}(T) \geq 0$). The higher is $S_{pm}(T)$, the higher is the number of patterns matching with the term T and as a result the higher is the likelihood to be a chemical compound. The seven rules given below compose the pool of patterns implemented in our system.

1. Presence of a morpheme indicating the number of carbon atoms (40 patterns): (`*meth*`, `*eth*`, `*propa*`, `*buta*` ...)

2. Presence of a specific suffix (58 patterns):
   (`*ane`, `*yne`, `*thiol`, `*oate`, `*amine` ...)
3. Presence of a numbering prefix/infix (locant):
   (`1,3-*`, `2,3,5-*`, `*-2-*`, `[4,5-b]*` ...)
4. Presence of a multiplying prefix (10 patterns):
   (`tri*`, `tetra*`, `penta*` ...)
5. Presence of a ambiguity prefix (3 patterns):
   (`iso*`, `sec*`, `tert*` ...)
6. Presence of a specific infix (46 patterns):
   (`*chlor*`, `*phosphor*`, `*amin*` ...)
7. Presence of specific Caps and Numbers patterns:
   (`AcOH`, `NH4OAc`, `DMFDMA` ...)

4.2 The Bayes Classifier

The second approach uses a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions [8]. The instances to be classified are described by attribute vectors $\overrightarrow{a} = (a_1, a_2..., a_n)$. The overlaping n-grams of letters ($n = 3$) are used to train the classifier. For example, the term `2-methylpentane` will be splitted in thirteen 3-grams (e.g. `2-m`, `-me`, `met`, `eth`, `thy`, `hyl`, `ylp`, `lpe`, `pen`, `ent`, `nta`, `tan` and `ane`). The use of 3-grams representing the first/last two characters of a term (respectively `**2`, `**2-`, `ne*` and `e**` for the example above) have been experimented but finally not retained[3]. The Bayes classifier assigns to an instance the most probable –or maximum *a posteriori*– classification from a finite set C of classes:

$$C_{map} \equiv \underset{c \in C}{argmax}\ P(c|\overrightarrow{a}) \tag{1}$$

Which after applying Bayes' theorem can be written

$$C_{map} = \underset{c \in C}{argmax}\ P(c)P(\overrightarrow{a}|c) \tag{2}$$

We choose to define each attribute a_i as one of the 3-grams that compose the term T. The finite set C is composed by two classes: c and $\neg$c (e.g. chemical and not chemical). We need to estimate the probability of a certain 3-gram $a_i = (w_{i-2}, w_{i-1}, w_i)$ occurring in a class c.

$$P(w_i|w_{i-2}w_{i-1}c) \tag{3}$$

The posterior probabilities could be estimated directly from the training data using Laplace smoothing to avoid zero probabilities. With this assumption, Equation (2) becomes the Bayes classifer.

$$C_{map} = \underset{c \in C}{argmax}\ P(c) \prod_i P(w_i|w_{i-2}w_{i-1}c) \tag{4}$$

[3] These 3-grams being not discriminant introduce misclassifications

4.3 Combination of the approaches

Although successful, the first approach (c.f section 4.1) is limited by the tremendous variety of chemical names in literature. As a consequence, the overall performance is below the Bayes classifier. The classification approach is more accurate and achieves good results (see section 6). Since our main goal is to produce a system with a very high precision, the choice was hence made to try a combination of the two approaches. The basic idea implemented by the hybrid method is that of "voting" or "recommendation". When one term is classified as chemical compound and at the same time is matched by at least one rule then the term is validated as chemical compound. The combination is hoping to increase the precision to a very high score by removing misclassification errors. The price to paid for an increase of precision will be a fall of recall, only the intersection of the two term classes is considered.

5 Experimental Settings

The method described in the previous section has been implemented and evaluated on a testing corpus. In the following subsections, details of the experimental settings are described.

5.1 Classifier Training

Training the parameters requires (i) creating two vocabulary sets, e.g., a chemical coumpound name set V_c and a non-chemical set $V_{\neg c}$, (ii) estimating the n-gram probabilities by calculating the n-gram occurrences. The chemical compound name vocabulary V_c was created from a CAS[4] database of about 10K compounds. For each chemical compound a query has been sent to the online database: http://webbook.nist.gov and by parsing web pages all differents names (synonyms) have been obtained. The resulting V_c is composed by nearly 65K compound names. The non-chemical vocabulary $V_{\neg c}$ was created using the SCOWL (Spell Checker Oriented Word Lists) corpus [5]. The reasons of using the SCOWL corpus are (1) to avoid the non-chemical vocabulary to contain any chemical compound names or errors, and (2) to easily gather a large quantity of n-grams. The n-gram probabilities were estimated from the occurrence frequencies inside the vocabulary sets. Two training data sets for chemical names (called `Small Voc` for 10K and `Large Voc` for 65K) and ten of increasing size for non-chemical words have been experimented.

5.2 Test Data

In order to evaluate our approach across real-life data sets, we have constructed a test data set composed by abstracts and plain articles. The test corpus is composed by

[4] Chemical Abstracts Service (CAS), a division of the American Chemical Society, assigns these identifiers to every chemical that has been described in the literature. CAS registry numbers are unique numerical identifiers for chemical compounds, polymers, biological sequences, mixtures and alloys.

[5] http://wordlist.sourceforge.net/

12 annotated abstracts extracted from the Beilstein Journal of Organic Chemistry[6] RSS feed and 8 plain articles coming from different journals (Organic Letters and Accounts of Chemical Research[7]) of different years (respectively 2000-2002 and 2005-2007), different authors and topics. The corpus has been annotated by two different annotators and validated by a domain specialist. Corpus size is approximately 20,000 terms in which 850 chemical compounds were manually identified. For the abstracts, there are 2,700 words in which 170 chemical compounds and for the plain articles there are 17,300 words in which 680 chemical compounds.

5.3 Performance Measures

The following performance measures are considered relevant.

Precision. It is the proportion of retrieved and relevant chemical compounds to all the compounds retrieved.

Recall. It is the proportion of retrieved and relevant chemical compounds, out of all relevant compounds.

f-measure. It is the weighted harmonic mean of precision and recall. The traditional f-measure or balanced f-score is:

$$f\text{-measure} = \frac{2 \cdot (Precision \cdot Recall)}{(Precision + Recall)}$$

6 Experimental Results

Figure 1 shows the results of Precision, Recall and f-measure for the expression based pattern matching (c.f section 4.1) according to the rule used and their incremental combination (i.e. the combination in rule 3 means using rules 1,2 and 3). The observed results confirm the limitations of the approach. Indeed, the huge variety of different writing schemes used for chemical compounds makes impossible to obtain a full recall. We observe that each rule allow an increase of the f-measure, this means that all rules are "useful" (allow to increase the classification performance). Rules 6, 2 and 1 are the best-score rules. This is interesting because it is not the logical order according to the number of patterns contained in each rule (58 for rule 2, 48 for rule 6 and 40 for rule 1). It indicates that the presence of a specific infix (rule 6) is more discriminant than the presence of a morpheme indicating the number of carbon (rule 1) or the presence of specific suffixes (rule 2).

One might expect that the performance of the classifier would improve as the size of the training corpus increases, because a larger training corpus usually leads to a better estimation of the n-gram probabilities. In fact, Figure 2 shows that once the corpus size reaches 40% (5850 different 3-grams), f-measures of both Small and Large chemical training sets (respectively `Small Voc` and `Large Voc`) remain obviously at

[6] The Beilstein Journal of Organic Chemistry is an Open Access, peer-reviewed online journal that will encompass all aspects of organic chemistry. The journal covers organic chemistry in its broadest sense, including: organic synthesis, organic reactions, natural products chemistry, supramolecular chemistry and chemical biology. http://bjoc.beilstein-journals.org/home/

[7] http://pubs.acs.org

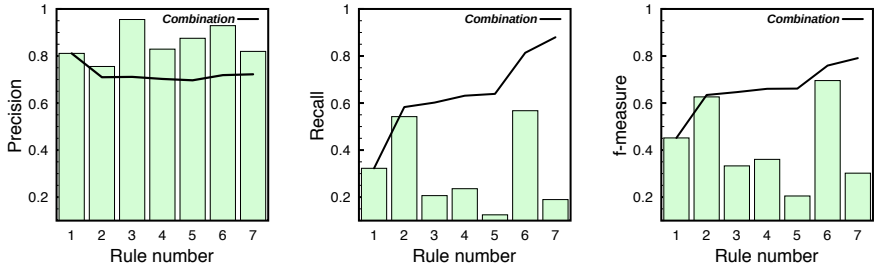

Fig. 1. Performance of the pattern matching approach in relation to the rule used. The performance of the incremental combination is also shown (black line).

the same values. This is due to the fact that the terms containing in `Large Voc` have been obtained automatically (c.f see section 5.1) and so non-chemical terms have been introduced in the chemical training set.

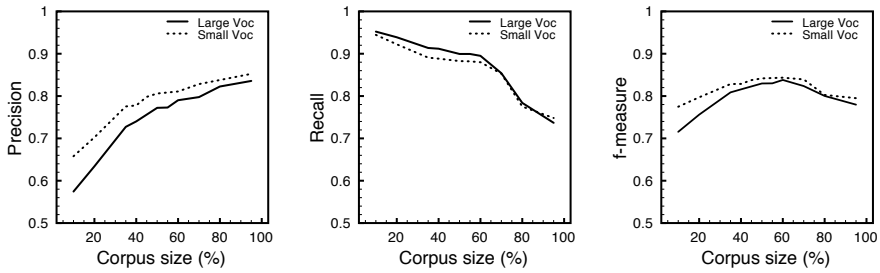

Fig. 2. Performance of the classifier vs. the size of the training corpora.

Figure 3 shows the results of Precision, Recall and f-measure of the rule-based and classifier approaches compared to their combination. We can observe that the combination significantly increase the precision (0.92839 against 0.72224 for the rule-based and 0.79015 for the classifier) and what ensued logically outperforms the best approach alone in f-measure (0.87701 against 0.79099 for the rule-based and 0.83801 for the classifier). This is a very interesting result because we can infer that approaches are complementary and can be combined without any consequent decrease of recall. We can extrapolate and suppose that the Entity Recognition in chemical texts may be broken up into sub-tasks solvable by slightly different but complementary approaches.

We have performed experiments on the two different kind of available corpus, i.e. abstracts and articles. Table 1 compares the performance of the hybrid method on the abstracts and on the articles. Our hybrid approach is consistent across the different data sets, the precision being in both case very high. The lack of recall in articles can be explained by the high proportion of trivial names (e.g. brand or trade names) that are not well recognized by our rule-based approach and also by the difference of

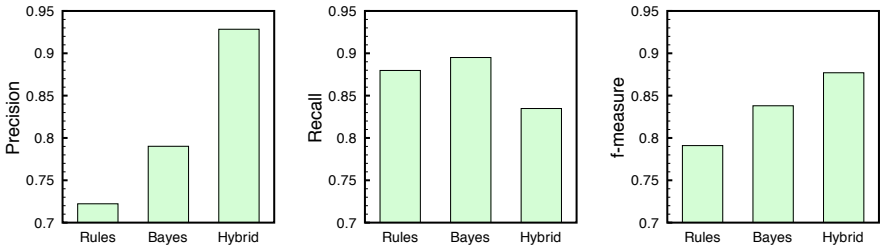

Fig. 3. Performance of the rule-based and classifier approaches compared to their combination.

chemicals proportion in data sets (6.29% for abstracts and 3.93% for articles). An *a priori* adaptation of the Bayes classifier's training corpora by tuning the ratio between the probabilities of the two vocabulary sets ($P(c)$ and $P(\neg c)$) has shown to increase the scores, this technique will be developed in further works.

	Precision	Recall	f-measure
Abstracts	**0.88333**	**0.93529**	**0.90857**
Articles	*0.93402*	*0.82221*	*0.87306*

Table 1. Performance score of the hybrid method on the two kinds of corpus, i.e abstracts and articles.

We have made an *a posteriori* error analysis and have observed that the terms not detected by our systems are essentially historical/common/brand names such as alumina, salt or pipecolate. These names are very difficult to be recognized because of their belonging in the two classes (c and $\neg c$) and because of their structures (not containing discriminant patterns/structures).

7 Conclusion and Future Work

We have described an hybrid method for chemical entity recognition that combines a simple rule-based pattern matching (seven rules) and a naïve Bayes classifier. Through experiments performed on different kind of available Organic Chemistry texts, we have showed that our hybrid approach is also consistent across different data sets. These results are promising, and represent a good starting point for future research but do show a critical point: the unstoppable growth of the number of different chemical compounds in the literature. As a consequence, Information Extraction (IE) approaches are more than ever required by life scientists to ensure an optimal sharing of the information. Among the others, there are several points that would be worthy of further investigation:

– Improve the estimation of probabilities by using smoothing techniques for unseen n-grams [3].
– Run experiments on different kinds of non-chemical corpora or different n-grams sizes and measure their impacts.
– Explore the usage of alternative combinations: combining the approaches in another way.
– Fuse chemical entity recognition with a domain-specialized automatic summarization system [1] as a domain-specialized weighted metric (i.e. the number of chemical compounds within a sentence is used as a parameter by the sentence scoring algorithm).

Acknowledgment

We are grateful to Patricia Velázquez-Morales for making available the test collection and for her help with this data set and to Pr. Alain Krief and Julie Henry for our useful talks. This work was partially supported by the *Laboratoire de chimie organique de synthèse*, FUNDP (*Facultés Universitaires Notre-Dame de la Paix*), Namur, Belgium.

References

1. F. Boudin and J.M. Torres-Moreno. NEO-CORTEX: A Performant User-Oriented Multi-Document Summarization System. *Computational Linguistics and Intelligent Text Processing*, pages 551–562, 2007.
2. J. Fluck, M. Zimmermann, G. Kurapkat, and M. Hofmann. Information Extraction Technologies for the Life Science Industry. *Drug Discovery Today–Technologies*, 2(3):217–224, 2005.
3. C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.
4. M. Narayanaswamy, KE Ravikumar, and K. Vijay-Shanker. A biological named entity recognizer. *Pac Symp Biocomput*, 427:38, 2003.
5. R. Panico, WH Powell, and J.C. Richer. *A guide to IUPAC nomenclature of organic compounds(recommendations 1993)*. Blackwell Science, 1993.
6. U. Reyle. Understanding chemical terminology. *Terminology(Amsterdam)*, 12(1):111–136, 2006.
7. J. Rigaudy and SP Klesney. *Nomenclature of organic chemistry(sections A, B, C, D, E, F, and H)*. Pergamon Press, 1979.
8. I. Rish. An empirical study of the naive Bayes classifier. *Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence*, 335, 2001.
9. S.B. Singh, R.D. Hull, and E.M. Fluder. Text Influenced Molecular Indexing (TIMI): A Literature Database Mining Approach that Handles Text and Chemistry. *J. Chem. Inf. Comput. Sci*, 43(3):743–752, 2003.
10. Bingjun Sun, Qingzhao Tan, Prasenjit Mitra, and C. Lee Giles. Extraction and search of chemical formulae in text documents on the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 251–260, New York, NY, USA, 2007. ACM Press.
11. D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.