1. Introduction

Term recognition, especially domain term recognition, is an important part of NLP tasks in practical usage scenarios. The use of mainstream features such as boundary words, contextual conjunctions is not sufficient to solve the current problem of term recognition in the chemical domain. Diverse word formation and large word lists have been identified in this filed. However, by observing a large number of scientific and technical terms, especially the names of chemical substances, we found that their naming has a stable composition pattern as well as the limited size chemical words which showed a one-to-one correspondence with morphemes. From a linguistic point of view, morpheme is the smallest unit of sound and meaning, and the main way to construct a word is to combine them. Therefore, this paper introduces the concept of morpheme into chemical term recognition, and applies the rules of semantics to the construction of terms to terminology form chemical morphemes. The identification of chemical terms. Specifically, we summarize the constituent elements of chemical terms and their morphological classes through the statistical analysis of compound naming conventions and domain corpus, and then learn the morphological patterns of the elements and their morphological classes from the domain lexicon to obtain the HMM model reflecting the morphological patterns. Thus we transform the problem of identifying chemical terms into a process of calculating the effective morphological probability of chemical elements, and solve the remaining problems of heavy labeling task, low-frequency terms and over-long terms in the existing work.

1. Literature Review

At present, the mainstream method of named entity and technical terms recognition is combining rules and statistics. Liang Liang (2002) Identified drug names etc. in commodity texts based on constructive features and word frequency distribution; Song Dan et al. (2009) classified chemical-specific words into three categories and propose a rule-based approach for identifying chemical substance names ; Nan Li et al. (2010) used rule-based methods and CRF models to translate term recognition into a process of valid word sequence recognition; Jianhong Ma (2018) identified initially using statistical methods based on the linguistic patterns and features of the chemical text and then corrected using lexicographic and rule-based methods.

Solving the problem of term recognition from the perspective of statistical learning requires specific annotation resources and corresponding algorithms, which makes information integration difficult. Or we can use the context information, but it still cannot solve the long-tail problem of many low-frequency terms in chemistry. From the perspective of knowledge rules, there are various ways of morphological rule in the field of chemistry, and the word list is huge. The rules of word formation are not strict and can not be exhausted. Long low-frequency terms and generic lexicons also bring interference, so it is difficult to build a complete database of proper names.

In addition, analyzing the features of constructs from a linguistic point of view can bring some help in term recognition. According to Yu et al. (2014), complex term constructs in specific disciplines can be described as disciplinary primitives, intrinsic morpheme groups, complex Terminology. Nan Li (2003) et al. categorized morphemes in chemical substance naming; Qianqian Wang et al. (2015) defined and classified chemical morphemes, and established the knowledge calculation model of chemical terms by rule-based method.

The existing recognition of chemical terms from the perspective of morphemes are mostly based on rules, with a certain accuracy, but it can not solve the problems of flexible combination and long words. In this paper, we take morphemes as basic units to classify chemical elements, and use model learning and prediction of relationships between different chemical elements to transform the problem of chemical term recognition into the problem of calculating the probability of chemical element construction, which solves the remaining problems in the existing work to some extent.

1. Chemical Terms and Morphemes
   1. Chemical Terms

The broad chemical terms include basic concept terms, chemical elements, chemical formulas, chemical instrument names, chemical reaction principle names, and physical chemical terms and biochemical terms formed by intersection with other disciplines, etc. And in this paper, chemical terms refer to chemical substance terms represented by chemical formulas, which are often large in length and number, highly expansive, and cannot be completely included in the chemical lexicon, and are the difficulties in automatic identification of chemical terms.

At the same time, these terms also have some regular characteristics. By observing a large number of them, it is found that there is a strong regularity in their word-building. Here are some rules for naming organic compounds [6]:

(1) Tiangan + substituent group/ substance generic words; For example, "methyl"(Jiaji), "butane"(Yiwan), "decane"(Guiwan), etc;

(2) Chemical element + substance generic words; For example, "boric acid"(Pengsuan) and "bromic acid"(Xiusuan);

(3) Chinese numbers + tiangan + substituent + tiangan + substance generic words, such as: “trimethylbutane”(Sanjiajidingwan);

(4) Chinese numbers + chemical element + “hua” + Chinese numbers + chemical element name; for example: "carbon dioxide"(Eryanghuatan), "vanadium trioxide"(Sanyanghuaerfan);

(5) Structure words + Chinese numbers / tiangan + chemical element name For example, "polytetrafluoroethylene"(Jusifuyisi), "p-toluene"(Duijiaben), "m-bromophenylacetone"(Jianxiubenbingtong), etc;

(6) Arabic numeral + '-' + Chinese numbers + tiangan + chemical elements / substituent + tiangan + chemical elements; for example: "2,7,8-trimethyldecane"(2,7,8-Sanjiajiguiwan)。

In addition, from the above rules, we can find that the words used in chemical terms are limited and exhaustible, including chemical elements, generic words of chemical substances ("suan", "zhi", "chun"), Arabic numerals, Chinese numerals, tiangan, structure words(dui, ju, pian, ya) and connective symbols; Moreover, the smallest unit of chemical terms and morphemes show a one-to-one correspondence. Therefore, this paper will training language model to learn the combination relationship between different chemical morphemes, and then judge the probability of candidate string sequence becoming chemical terms.

* 1. Chemical morpheme and classifications
     1. **chemical morpheme**

In modern Chinese grammar, the term "语素" is the Chinese translation of the English word "morpheme". This concept was proposed by the American structural linguist Bloomfield. In Zhang Bin's (2017) *New Modern Chinese*, the morpheme is defined as, "Morpheme is the smallest combination of sound and meaning in language and the smallest linguistic unit capable of distinguishing meaning. The role and function of the morpheme is primarily to construct words [19]." In terms of the relation of its expression, content and meaning, it it can no longer be divided, otherwise it will completely change or destroy the meaning of its original grammar and vocabulary. In Chinese, a character is usually a morpheme. Some morphemes can be made into words on their own, and morphemes can also be made into words by combining them with each other. In fact, it is true that the combination of monophonic morphemes does produce a lot of new words, and even becomes one of the most important ways of generating new words.

The basic semantic functions of morphemes in chemical terms are consistent with those of general morphemes, both refer to the basic unit of meaning, and are the smallest word formation unit. Most morphemes in the field of chemistry are monosyllabic morphemes with clear and professional meanings, such as "alkane", "hydrocarbon", "acid" indicate the species of chemical substances. In addition, In addition, there are many transliterated words in chemical terms, and when these words are subdivided into single words, they no longer have practical significance. We treat these transliterated words as a morpheme, such as "pyridine" and "furan", etc. Therefore, the chemical morphemes discussed in this paper refers to the smallest word formation unit with the characteristics of chemical terms or auxiliary naming of chemical substances. It can be a number, a letter, a Chinese character or a word.

* + 1. **chemical morpheme classifications**

According to the position and function of chemical morphemes when they constitute chemical terms, chemical morphemes are divided into four categories: core morpheme, definite morpheme, auxiliary morpheme and general morpheme.

The core morpheme is a kind of morpheme which can clearly indicate the nature or category of substances in chemical terms, so as to distinguish them from other types of substances, and play the most important role of identification and core. These morphemes only appear in the texts related to chemistry. The core morphemes consist of two subcategories, chemical elements and chemical proper noun, which are denoted by letters A and B, respectively. Among them, chemical elements are a fixed set, and there are 118 chemical elements published at present. Chemical proper nouns refer to morphemes that indicate the category of chemical substances and some polysyllabic morphemes, such as "acid", "fat", "salt" and "pyridine" and so on.

The definite morpheme refers to the morpheme that restricts the nature, type or structure of the substance based on the core morpheme.It can further limit the type or scope of the chemical substance, which is conducive to clarifying the concept of the substance. It is represented by the letter C. It mainly includes the restrictive morphemes that often appear at the beginning of words in chemical terms, such as "正(zheng), 原(yuan), 偏(pian), 交(jiao), 重(zhong), 亚(ya), 次(ci), 高(gao), 联(lian)" and so on, and the morphemes that indicate the method of chemical substance formation, such as "化(hua), 合(he), 聚(ju), 缩(suo), 代(dai)" and so on.

The auxiliary morpheme refers to the capital numbers, Roman numerals, Arabic numerals, Heavenly stems, English letters, Latin letters and some chemical connection symbols in chemical terms. It is used together with core morphemes and definite morphemes to improve the specificity of chemical terms. It is represented by the letter D.

The general morpheme refers to a morpheme that can appear in terms of chemistry and other fields. It has a single meaning, has no relevance to the field, and has no effect on the specificity of chemical terms. It mainly includes the morphemes that constitute the basic terms of chemistry, such as "电(dian), 解(jie), 溶(rong), 液(ye)", etc., which are represented by the letter E; and morphemes in the general domain, such as "水(shui), 子(zi), 白(bai)",which are represented by the letter F.

Table 1 Classification of chemical morphemes

|  |  |  |
| --- | --- | --- |
| categories | | chemical morpheme |
| core morpheme | Chemical elements (A) | hydrogen, helium, lithium, beryllium, boron, nickel…… |
| Chemical proper noun (B) | acid, salt, alkali, fat, alcohol, ether…… |
| pyridine, furan, pyrimidine, indole…… |
| definite morpheme | Chemical preposition (C) | 化(hua), 合(he), 聚(ju), 缩(suo), 代(dai)…… |
| Specific prefix (C) | 正(zheng), 偏(pian), 交(jiao), 亚(ya), 次(ci)…… |
| auxiliary morpheme (D) | | Capital numbers, Roman numerals, Arabic numerals, Heavenly stems, English letters, Latin letters and chemical connection symbol |
| general morpheme | Basic morpheme of chemistry (E) | 电(dian), 解(jie), 溶(rong), 液(ye)…… |
| Universal basic morpheme (F) | 水(shui), 子(zi), 白(bai)…… |

* + 1. **Construction of chemical morpheme list**

In order to train the chemical term recognition model, a chemical term vocabulary was constructed in this paper. The vocabulary mainly includes Sogou’s chemical vocabulary and Wikipedia’s organic and inorganic chemistry terms, with a total of more than 22,000 chemical term words. After that, the chemical morpheme list is constructed by combining statistics with artificial method. First, 2045 monosyllabic morphemes were obtained by statistics of chemical term list. Then, 105 polysyllabic morphemes were collected by manual annotation. Because the number of terms collected in the chemical vocabulary is limited, it cannot cover all chemical morphemes, and the the set of core and auxiliary morphemes is relatively closed, so they are artificially supplemented, and the monosyllabic morphemes are selected according to the morpheme frequency. Finally, a total of 718 morphemes were obtained. According to statistics, the number of chemical morphemes is much smaller than the number of chemical terms. It can be seen that the combination of chemical morphemes is flexible and the word formation ability is strong. It is simple and feasible to recognize chemical terms by using the combination rules of chemical morphemes.

Table 2 The 20 most frequent morphemes in chemistry

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| morpheme | frequency | morpheme | frequency | morpheme | frequency | morpheme | frequency |
| 酸(suan, acid) | 6707 | 基(ji, base) | 3421 | 二(er, two) | 2721 | 化(hua, transform) | 2510 |
| 甲(jia, A) | 2368 | 苯(ben, benzene) | 2176 | 乙(yi, B) | 1805 | 氯(lv, chlorine) | 1649 |
| 氧(yang, oxygen) | 1312 | 胺(an, amines) | 1246 | 醇(chun, alcohol) | 1181 | 氨(an, [ammonia](http://youdao.com/w/ammonia/" \l "keyfrom=E2Ctranslation) | 1120 |
| 硫(liu, sulfur) | 1077 | 酰(xian, [acyl](http://youdao.com/w/%EF%BC%88%E9%85%B0%E5%9F%BA%EF%BC%89%20acyl/" \l "keyfrom=E2Ctranslation)) | 1046 | 酯(zhi, ester) | 1019 | 丙(bing, C) | 921 |
| 三(san, three) | 913 | 盐(yan,salt) | 871 | 酮(tong, ketone) | 866 | 烷(wan, alkane) | 841 |

Table 2 shows the 20 most frequent morphemes in chemistry, of which 16 are core morphemes, and the morpheme "acid" appears most frequently. with a total of 6707 occurrences. Figure 1 shows the frequency distribution of all the morphemes in the chemical term vocabulary, which also shows the same trend. That is, the top of the distribution is basically the core morphemes, followed by the definite morphemes.

Fig. 1 Frequency distribution of chemical morphemes (partial)

After obtaining the chemical morpheme table, the chemical morpheme was manually classified according to the above-mentioned chemical morpheme classification instructions, and the number of chemical terms formed by different types of chemical morpheme was counted. According to Table 3, the number of chemical proper noun (B) and definite morphemes (C) are the largest, accounting for 22.98% and 20.75% of all morphemes, respectively, and the number of general morphemes (F) is the least, accounting for 10.58% of all morphemes, and the numbers of other morphemes are close. However, categories A, B, C and D are relatively closed sets with limited number of morphemes. Therefore, the use of morphemes for chemical term recognition can achieve a multiplier effect, while ensuring the recognition effect, reducing manpower investment.

Table 3 Statistics on the classification of chemical morphemes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Chemical elements  (A) | Chemical proper noun  (B) | Definite morpheme  (C) | Auxiliary morpheme  (D) | Basic morpheme of chemistry  (E) | Universal basic morpheme  (F) |
| number | 109 | 165 | 149 | 101 | 118 | 76 |
| percent | 15.18% | 22.98% | 20.75% | 14.08% | 16.43% | 10.58% |

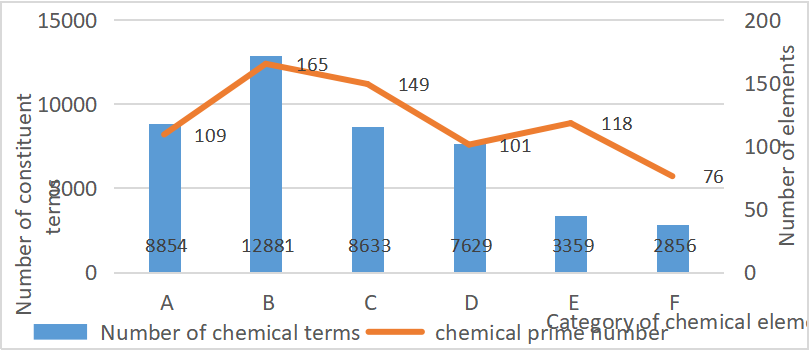


Figure 2: number of chemical morphemes vs number of chemical terms composed by the morpheme

Figure 2 shows the number of chemical morphemes in each category versus the number of chemical terms formed by the chemical morphemes in that category. It has been observed that the number of chemical morphemes in categories C, E, and F is relatively large, but the number of chemical terms it constitutes is relatively small. By analyzing the chemical term vocabulary, we know that category C is a chemical definite morpheme, and there may be many different definite morphemes in a chemical term, such as "亚(ya)/C 铁(tie)/A 氰(jing)/B 化(hua)/C 钙(gai)/A (Calcium ferrocyanide)"; category E and F are relatively common morphemes with various types. They only play an auxiliary role in chemical terms, and its occurrence is irregular. Therefore, the number of morphemes is large, and the number of chemical terms formed by these morphemes is small.

1. Approaches

We take the features of fixed set of chemical morphemes, less semantic ambiguity, regular searchable combinations of morphemes, and high degree of morphological adhesion as the starting point for term recognition, and use chemical morphemes as the basis to model patterns of morphemes in the chemical lexicon.

Specifically, HMM is used to fit **the morphological patterns**, and then **the improved forward algorithm** calculates the probability of morphemes forming the term, and then the probability is used to determine whether to merge the morphemes or not, and finally the merging result is the recognition result. The probability of several morphemes constituting a term uses threshold to binary, which is used to classify whether several morphemes can constitute a term; different probability thresholds for different numbers of morphemes in a term.

* 1. Modeling the morphological patterns

We use Hidden Markov Model(HMM) to model the morphological patterns. HMM is a formal foundation for making probabilistic models of linear sequence 'labeling' problems, we apply HMM to the recognition of chemical terms in Chinese. Formally, given the morphological pattern model λ and λ is specified by the following components:

我是公式而且是熟悉的中文哇哦

Here, morpheme classification is the hidden state, 这里还是公式, including six tags; morpheme is the observation state, 这里还是公式, including 718 morphemes; an initial probability distribution π over morphemes; a transition probability matrix A, each a-i-j representing the probability of moving from morpheme tag i to tag j; a sequence of observation likelihoods B, each expressing the probability of a morpheme being generated from a morpheme tag as following:

我是公式而且是熟悉的中文哇哦

We simply use add-one smoothing for ZERO item in B.

* 1. The improved forward algorithm

Given the morphological pattern model λ, forward algorithm can be used to compute the probability of the morphological patterns. But in process of iterative computation, the forward probability tends to zero as the length of sequence increases.

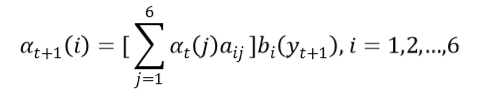
To ensure the final probability does not arithmetically underflow, we introduce the improved forward algorithm, which is a treatment of observation sequence length and can be used to compute the probability safety. We can give a formal definition of the algorithm as follows.

Given model λ = (A,B,pi), the observation sequence 问我是谁我是公式 is computed as:

1. Initialization:

C:\Users\xia18\AppData\Local\Temp\1594047608(1).png

1. Recursion:



1. Termination:



As we compute each probability to a new state for the next observation, the final probability P(O|λ) lead us to a chemical term.

1. Experiments
   1. DataSets

Datasets including a chemical morphology list, a chemical lexicon, and a recognition test set. The chemical morphology list containing 718 morphemes and the chemical lexicon containing The average word length of the lexicon is 4.3 elements out of 22,356 words, see Section 2.2 for details.

Test set is randomly selected from Chinese patent documents under C01 (inorganic chemistry) and C07 (organic chemistry). Patent terminological identification is one of the most important scenarios for NLP. And the International Patent Classification (IPC classification) is the basis for patent classification. The test set contained 531 sentences with an average term length of 6.998, maximum term length of 48, average sentence length of 24.537, maximum sentence length 172.

Data and Code is available through Github.

* 1. Implementation and Metric

In our experiments, there is two parts of implementation: preparation and recognition. Calculating HMM model parameters and determine probability threshold in preparation. And then merging morphemes according to the probability to complete recognition.

//Definition of return values and temporary arrays of morphemes

//Fitting the parameters of the training model and determine probability thresholds

//to be a term

//or not to be

The threshold for term length 2 is 15, for length 3 is 15.8, for term length 4 is 18.7. Reported metrics are precision, recall, F1-score, which evaluates the performance of recognition.

* 1. Results and Analysis

As shown in Graph 3 and Graph 4, our approaches achieve impressive performance of long and complex chemical term recognition with modeling the morphological patterns.

图3

图4

For example, in Graph 3 “将丁烷四羧酸、磷酸二氢铵依次溶于去离子水中”(Butane tetracarboxylic acid and ammonium dihydrogen phosphate were dissolved in deionized water.), linked morphemes such as “丁烷四”“烷四羧”“四羧酸” can merge into a term“丁烷四羧酸”(Butane tetracarboxylic acid), “磷酸二氢铵”( ammonium dihydrogen phosphate ) filtered by probability thresholds. Furthermore, in Graph 4 we get the result correct including complex term with hyphen mark“丙烯酸-2-乙基己酯”(2-ethyl hexyl acrylate).

表4

As the experimental results summarized in Table 4, we see that

1. In practice, term length are often higher than 5. And term length increases, term complexity increases even rapidly.
2. Term morphemes keep stable while various chemical term length.
3. Terms have high degree of adhesion and low morpheme ambiguity.

In conclusion, we can clearly see the benefit of modeling the morphological patterns, which make chemical term recognition more reliability, efficiency and interpretable.

(这里好像有语法错误)

In terms of running efficiency, the test set takes 3500ms for a single thread on experimental machine, with an average of 7ms per sentence. Roughly estimated, it can support at least 30+ connection with response time less than 300ms.

However, after observing results, it is found that poor recognition effect mainly have the following types:

从化学语素类型上看，主要是含有多个“-"、数字的术语以及含有括号的化学术语，如“2-氯-6-三氯甲基吡啶”“四溴邻苯二甲酸双（2-乙基己基）酯”，主要是因为数字和括号具有通用性，领域性较弱，训练数据中该类型的术语数量少，模型学习不充分，且在进行计算时，没有考虑语素的位置信息；从化学术语长度上看，本文的模型在一定程度上解决了识别过程中术语过长的问题，对10字以内的化学术语识别效果较好，大于12字的术语效果也达到了57.69%，相对短术语识别效果较差，可增加长术语的训练数据规模，针对长术语进一步优化模型。

However,

Furthermore, by comparing the performanceson embedding metrics, we find that all models ob-tain decent scores, but none of the models outper-form the others significantly.

We evaluate three main architectures under two scenarios

achieve high accuracy than fixed vocabulary approaches.DST-SC achieves state-of-the-art performance on

We conducted further related-slot tests to verify the effectiveness of DST-SC in solving the related-slot problem.

For further comparisons, we also use the crowd-sourcing labeling resources of our organization tomanually evaluate the relevance and the personaof generated responses.