Development of universal detection methods for identifying chronological or pseudo-chronological order of occurrence of terms in a given subject area

Ekaterina Filimonova, Sergey Soloviev, Irina Polyakova
Lomonosov Moscow State University
Moscow, Russia
filkate0903@rambler.ru, glosscom@mail.ru, polyak@cs.msu.su

***Abstract*—This work seeks to develop universal detection methods for identifying chronological or pseudo-chronological order of occurrence of terms in a given subject area. To solve the problem of reconstruction of the chronological order of words and terms, it is proposed to use three methods: the method of word formation, the method of dictionary use, and finally, the method of hyponyms and hyperonyms.**

The method of word formation can be divided into several ways in relation to the problem: the prefixal method, the suffixal method, the prefixal-suffixal method, the non-suffixal method and the merging method. Prefixal method of word formation forms a new word by adding a prefix to the base. The suffixal method of word formation forms new words by adding a suffix to the base. Prefixal-suffixal method is based on the two methods of word formation described above. The non-suffix method forms new words using a zero suffix. The merging method forms new words by adding existing words. The method of using etymological dictionaries makes it possible to identify the exact sequence of the terms according to the available accurate data collected by such people as Max Vasmer.

Each method builds the order of words and terms as they appear and is taken with a certain confidence factor of that order.

I. Introduction

It is known that different words in Russian language can have either direct or indirect connections.

In addition to the usual everyday words in the Russian language is a special category of words, referred to as terms. Terms represent an area of special vocabulary of the language formed as a result of scientific and technological progress.

Terms are created by a person to be able to communicate in various special areas. They should accurately reflect the results of people's experience and practice. Terms should be concise, specific, precise, and unambiguous. Terms can be formed by monosyllabic nouns, complex words, phrases, etc.

For example, monosyllabic nouns could be the word soil ('почва'), politics ('политика'), regions ('регионы'), enterprises ('предприятия'), economy ('экономика').

For example, complex words could be the word agriculture ('земледелие'), engineering ('машиностроение'), biosphere ('биосфера'), pricing ('ценообразование').

For example, phrases could be the word information security pricing ('информационная безопасность'), economic growth pricing ('экономический рост').

Special terms of a particular subject area are usually collectively described in a Glossary, where each term is an object containing both name and definition.

For example, 'деньги - особый товар, выполняющий роль всеобщего эквивалента при обмене товаров, форма стоимости всех других товаров. Деньги выполняют функции: меры стоимости, средства обращения, средства образования сокровищ, средства платежа и мировых денег.'

Glossary is user-friendly, as it contains terms together with their definitions, also including links. Over time, some terms become obsolete and go out of circulation.

For example, 'чеканка - получение рельефных изображений на листовом металле. Чеканка: является одним из древнейших видов художественной обработки металла; выполняется ударами особым молотком по чеканам; ведется по поверхности металлического листа, положенного на эластичную подложку из особой смолы. Различают механизированную и ручную чеканку.'

Obviously, there is a need to rank the terms by the time of appearance relative to each other.

II. Methods

To implement the task of searching for the chronological order of words and terms, we need to use a method that allows us to determine the sequence of their appearance relative to each other by two given terms. We assign to each method a degree of confidence in the correctness of its work. While choosing the final result, we will give preference to the method with the greatest confidence.

The following is the description of three methods for identifying the order of appearance of terms.

The first method is based on comparing morphemic structures of given terms. There is a system that allows dividing the input word by morphemes with high accuracy. It is not limited to the scope of a particular subject area, so the resulting morphemic structure of the term allows application of the word formation rules of the Russian language, in which the greatest interest are: prefixal, suffixal, prefixal-suffixal, non-suffixal. Also, to the above methods, another method of merging is added.

Prefixal method of word formation forms a new word by adding a prefix to the base. In the discussed case, only nouns and adjectives are used as terms, so it is necessary to determine if a certain term originated earlier or later, however it is not difficult. It is worth noting that the prefixal method of word formation does not cause a jump between parts of speech: a noun is obtained from a noun, an adjective is an adjective. For example, the word demobilization ('демобилизация') was formed from the word mobilization ('мобилизация') prefixed way, or the word prediction ('предсказание') is derived from legend ('сказание').

Terms can consist of only two parts of speech-adjective and noun, terms that are verbs, as well as other parts of speech, are not expected to be found, so the scheme of the prefix method in relation to parts of speech is as follows:

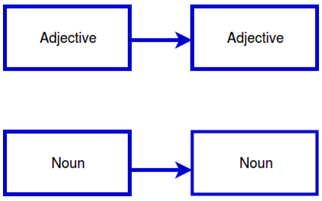

Figure 1. Scheme of the prefixal method in relation to nouns and adjectives.

This scheme does not contain such parts of speech as the verb, adverb, and others, because in this paper the consideration of these cases is not required.

The suffixal method of word formation forms new words by adding a suffix to the base. This method differs from the previous one, because a noun, an adjective and a verb can be formed from a noun, while a verb and an adverb from an adjective. But since we are interested in the formation of nouns and adjectives in this problem, let's consider the cases when either a noun forms a noun or a noun forms an adjective. For example, the word market ('рыночная') formed from the word market ('рынок') suffixal way.

Terms can consist of only two parts of speech-adjective and noun, terms that are verbs and other parts of speech, are not supposed to be found, so the scheme of the suffixal method in relation to parts of speech is as follows:

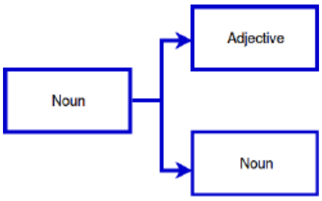

Figure 2. Scheme of the suffixal method in relation to nouns and adjectives.

This scheme does not contain such parts of speech as the verb, adverb, and others, because in this paper the consideration of these cases is not required.

Prefixal-suffixal method is based on the two methods of word formation described above. In the case of using nouns and adjectives with the help of word formation prefixal-suffixal way to get the same parts of speech (adjectives and nouns), just like in the suffixal and prefixal ways. An example of using this method can be seen in a couple of words: armour ('оружие') and disarm ('обезоружить').

The scheme of operation of the prefixal-suffixal method in relation to parts of speech is as follows:

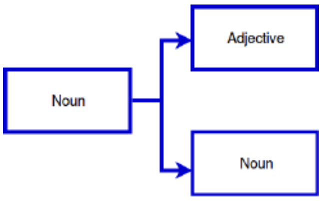

Figure 3. Scheme of the prefixal-suffixal method in relation to nouns and adjectives.

This scheme does not contain such parts of speech as the verb, adverb, and others, because in this paper the consideration of these cases is not required.

The non-suffix method forms new words using a zero suffix. Thus the zero suffix in the letter and in the speech is not expressed in any way. The non-suffix method allows you to change part of speech. Thus, a noun can be formed from a verb, an adjective or a noun, an adjective from a noun, an adjective and a verb. An example of this method of word formation is a pair of words: smooth surface ('гладь') and smooth ('гладкий').

The non-suffix method, like prefixal, prefixal-suffixal and suffixal, very comfortable. It is easy to identify and recognize, but if we know what word from what was formed. But we are faced with the inverse problem-to identify the order of an unknown method, if it can be applied at all. There is a complexity especially in the case of suffixal and non-suffix method. It is not always clear what method is used. It is precisely such cases that give rise to uncertainty. In other words, we either guess with the answer or not, and the probability of hitting about 50 percent. For the study of such cases, this method gives an inaccurate result, which means that the confidence factor will also not reach the highest.

The scheme of the non-suffixal method for parts of speech for our problem is as follows:

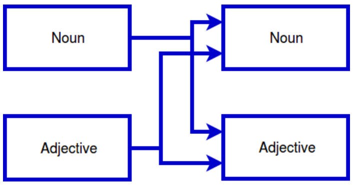

Figure 4. Scheme of the non-suffixal method in relation to nouns and adjectives.

This scheme does not contain such parts of speech as the verb, adverb, and others, because in this paper the consideration of these cases is not required.

The merging method forms new words by adding existing words. In this case, between the words can be put a hyphen. In some cases, between the components of the final word can be put a connecting letter, such as 'o' or 'e'. Often merging words forms a new word by removing the end (sometimes suffix) of the first word and joining the second. Accordingly, part of the speech of the resulting word will be determined by part of the speech of the second word.

As an example, the word 'agriculture' ('земледелие'), which is formed by merging the basics of 'land' ('земл') and 'deeds' ('дел') with the addition of the letter 'e' between them.

The scheme of operation of the merging method in relation to parts of speech for our problem is as follows:

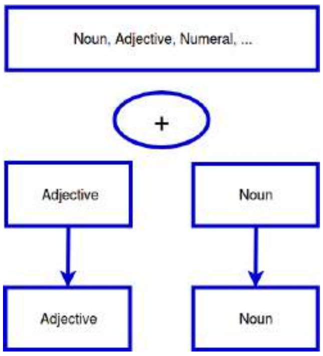

Figure 5. Scheme of the merging method in relation to nouns and adjectives.

This scheme does not contain such parts of speech as the verb, adverb, and others, because in this paper the consideration of these cases is not required.

Certainly when it comes to the methods described above, using the first method can give rise to ambiguity, that is, it will definitely not be clear which term was formed first, the first from the second or vice verse. It is impossible to solve this ambiguity with the help of the first method, leading to potentially inaccurate results, and, therefore, reducing the confidence factor for this particular method. Inaccuracies can be resolved with the following method.

Let us consider the method of identifying the order of terms based on their use in etymological dictionaries. Currently, a large number of dictionaries exists, where for every term there can be found another term, from which the first one was formed. This "other" term is spelled out explicitly, and finding it leads to a correct result. The method does not generate ambiguities, the only problem is that it may not give the desired result when either the term itself or the etymologically original term are not described in the dictionary. That is, the method with the overall final result will be taken into account with a large confidence factor, unlike other methods.

At the moment, created etymological dictionaries, which include a huge amount of words. For each word, you can find information that describes the origin of the word. So, for example, for the term 'policy' ('политик') in etymological dictionaries it is possible to find the word from which it is formed. Let us turn to the etymological dictionary of Max Vasmer translated into Russian. In addition to other information, we can highlight the main related to our task. In relation to our example, we can get information that the term 'policy' ('политик') is formed from the word 'city' ('город'). Or, for example, another example, where the term 'society' ('общество') comes from the word 'general' ('общий'). This information gives a complete and error-free result, because the information in the etymological dictionaries is reliable.

However, in etymological dictionaries is not always found the right word with the necessary information. In this case, you can apply the method of word formation using the above methods, such as prefixal, prefixal-suffixal, suffixal and non-suffixal. A word close to this can also be found in the etymological dictionary.

Finally there is the method of identifying the order of terms based on the allocation of generalization and quotient, better known as the problem of finding hyponyms and hyperonyms.

Hyponym (Greek. $\upsilon\pi o$-under, below + $o\nu o\mu\alpha$ - name) is a concept expressing a particular entity in relation to another, more general concept. Hyperonym (super) - a word with a broader meaning, expressing a general, generic concept, the name of the class (set) of objects (properties, features).

A hyperonym is the result of a logical generalization operation or, in a mathematical sense, a complement to a set.

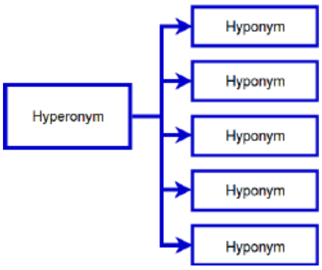

Figure 6. Hyponyms and hyperonym.

If one of the terms is a hyponym, and the other is a hyperonym, and they have a common word, then the hyperonym appeared before the hyponym.

For example, for the pair of terms: art ('искусство') and theatre arts ('театральное искусство'), the term art is the hyperonym and the term theatre arts – hyponym, which suggests that the term theatre arts came later of the term art. The complexity is represented in a situation where hyponym and hyperonym are not syntactically similar. In this case, it is impossible to determine which concept arose earlier without any additional information.

Therefore, this method does not always produce a correct result. Hence, the confidence factor will not be high.

Of course, in the Russian language in addition to the above difficulties, there are other ambiguities and obstacles to solving problems related to computer linguistics in General. Consider one of these problems.

Polysemy, or polysemy of words occurs due to the fact that the language is a system limited in comparison with the infinite variety of reality, so that in the words of academician Vinogradov, " the Language is forced to carry countless meanings in one or another headings of the basic concepts."

This problem could be another difficulty in achieving this goal, but in our task the work is done with a specific Glossary. It is convenient not only because it contains a large number of terms, but also because of the identified links and definitions.

Thus, there is no problem of ambiguity of understanding of terms. Connection with other concepts allows you to analyze the proximity of terms and the order of their appearance.

III. Conclusion

Thus, several methods are proposed to solve the problem of ranking terms by the time of their appearance and to identify the chronological or pseudo-chronological order of occurrence of terms in a given subject area.

References

[1] Ye. Zemskaya, *Sovremennyy russkiy yazyk. Slovoobrazovaniye: ucheb. posobiye*, 3rd-ed. Russia, Moscow: Flinta.

[2] V. Nemchenko, *Sovremennyy russkiy yazyk. Slovoobrazovaniye: Ucheb. posobiye dlya filol*, Russia, Moscow: Vyssh. shk., 1984, p. 255.

[3] Ye. Ilina, YU. Dracheva, *Sovremennyy russkiy yazyk:morfemika i slovoobrazovaniye: uchebno-metodicheskoye posobiye*, Vologda: VoGU, 2015, p. 68.

[4] I. Vasilyeva, D. Fedorov, *Web-tekhnologii: uchebnoye posobiye*, SPb.: SPbGEU, 2014, p. 67.

[5] L. Babenko, *Lexicology of the Russian language. Textbook of the Ural state Unversity named after M. Gorky, faculty of Philology*, Ekaterinburg: Ural state University named after A. M. Gorky, philological faculty, 2008, p. 125.

[6] S. Barkhudarov, *Lexical synonymy*, Moscow: Nauka, 1967, p. 180.

[7] Y. Apresyan, *Lexical semantics. 2nd edition, revised and supplemented*, Moscow: Languages of Russian culture; publishing company "Eastern literature" RAS, 1995, p. 472.

[8] I. Kobozeva, *Linguistic semantics*, M.: editorial URSS, 2000, p. 352.

[9] V. Beloshapkova, *modern Russian language. Second edition, corrected. and dop.*, M.: High school, 1989, p. 800.

[10] Er. Freeman, El. Freeman, *Studying HTML, XHTML and CSS*, Saint-Petersburg, 2012, p. 656.

[11] B. Hogan, *HTML5 and CSS3. Web development according to standards of new generation*, Publishing house "Peter", 2011, p. 272.

Разработка универсальных методов выявления хронологического или псевдохронологического порядка возникновения терминов в заданной предметной области

Филимонова Е. А., Соловьев С. Ю., Полякова И. Н.

В работе ставится задача разработки универсальных методов выявления хронологического или псевдохронологического порядка возникновения терминов в заданной предметной области. Для решения задачи реконструкции хронологического порядка возникновения слов и терминов предлагается использовать три метода: метод словообразования, метод использования словарей, а также метод гипонимов и гиперонимов.

Метод словообразования можно разделить на несколько способов применительно к поставленной задаче: приставочный способ, суффиксальный способ, приставочно-суффиксальный способ, бессуффиксный способ и способ слияния. Приставочный способ словообразования формирует новое слово добавлением приставки к основе. Суффиксальный способ словообразования формирует новые слова добавлением суффикса к основе. Приставочно-суффиксальный способ основан на двух описанных выше способах словообразования. Бессуффиксный способ формирует новые слова при помощи нулевого суффикса. Способ слияния формирует новые слова сложением уже существующих слов. Метод использования этимологических словарей позволяет по имеющимся точным данным, собранным такими людьми, как Макс Фасмер, выявить точную последовательность возникновения терминов.

Каждый из методов строит порядок слов и терминов по мере их появления и берется с определенным коэффициентом уверенности этого порядка.

Received 10.01.19