Multiword Terms and Machine Translation

Serge Potemkin

Lomonosov MSU, Moscow, Russia
prolexprim@gmail.com

Abstract. In this article we discuss, using morphological analysis and foreign equivalent selection, the following issues: the definition of syntax and the logical-semantic structure of terms using morphological analysis and foreign equivalent selection. Texts of articles in the "Voprosy Psichologii" (*Psychological Problems*) journal were used as the source corpus. Statistics of multiword terms in this corpus were calculated as well as the pointwise mutual information (PMI) between terms. We defined the distance between multiword terms in a multidimensional space, and the measure of terms proximity, as the minimal semantic distance between them. Then, we performed multidimensional scaling for visualising the terms space. Machine translation was used as a means to finding equivalents of scientific and technical terms in two arbitrary languages. With the help of forward and reverse translation of terms, using online machine translation tools, the meaningful equivalents were evaluated. We measured the proximity between terms as the Levenshtein distance between the original term and its direct and reverse translation, and tried to minimise this measure.

Keywords: Term Equivalent, Structure of Composite Terms, PMI, Machine Translation, Direct and Reverse Translation.

1 Introduction

The increasing volume of published scientific and technical information is doubling annually or faster, requiring the identification and standardisation of the vocabulary used in science and technology, to allow the readers to correctly understand the essence of the message. This observation fully applies to the terminology used by different authors in the same and related areas of science and technology, as well as in academic and popular literature. Perhaps an even more important issue for development is the need for an exchange of scientific information generated in different countries in different languages. Such an exchange is impossible without the accurate translation of scientific articles which contain terminology.

Terminology problems are also associated with the development of machine translation (MT) systems. Grammar of the scientififc text is simple and its translation depends mainly on the correct translation of nominal constructions, primarily terminological words and phrases. At the moment "it is no longer in doubt that for the correct, scientifically substantiated solution of terminological problems, one should learn the terminology with the recognition of its nature and logical existence. That's why with-

EUROPHRAS 2019, pages 133–139,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_018

in the framework of the above-mentioned, the problems of terminology should be explored by linguists and technologists" [1].

2 The concept of the Scientific Term

The term is a word, phrase, acronym, or other lexical unit, which designates the corresponding extra-linguistic concept in the real world. According to Mikhail Glushko "the term – is a word or phrase to express concepts and a notation of objects having, thanks to its rigorous and precise definitions, clear semantic boundaries within an appropriate classification system" [2]. Generally, the term must be unambiguous, i.e. correspond to only one particular entity. The uniqueness of the term, in contrast to the general-language word, does not depend solely on the surrounding context. Specifically the term can be used in isolation, its belonging to the specific terminology defines the uniqueness of the term within this terminology; in other areas the term may have a completely different denotation. Unfortunately, even in the same subject area, or the same knowledge field, the term may be defined differently by different authors. The object of our study of terminology is the vast area of "Psychology".

For example, consider the definitions given by different authors of the term "*photopsia*":

a) Photopsia, (from the Greek φωτός — light + ὄψις - vision) - subjective light phenomena (feeling), not having the nature of certain figures or objects. Usually these are flashing spots, sparks, and light zigzags etc.; photopsia is caused by the action of the mechanical or toxic stimulation of the visual analyser [3].

b) Photopsia. The emergence of moving shapes, dots, spots, etc., mostly luminous, shining in the field of view. Photopsia is observed in diseases of the retina, and elementary visual hallucinations as a psychopathological phenomenon [4].

The first definition does not contain the notion of a "psychopathic phenomenon", which is important in the field of psychology. Such examples are numerous. Often, even the same author will give several interpretations of the term which reflects the different use of it in the various sub-fields of science. The reader could hardly guess what kind of entity is being described by the term in the context.

Even greater difficulties arise in the process of translation of the term. For example, the term "*слияние*" has at least 2 English equivalents: [4]

СЛИЯНИЕ (symbiotic) (MERGING)

СЛИЯНИЕ (synthesizing) (FUSION)

3 Term definition problems

The use of compound terms - terminological expressions partly cope with such difficulties. In this context, an attempt was made to review the syntax and logical-semantic structure of terms. A very important class of multiword terms (MWT) is one where the meaning/semantics is not obvious from the composition of the meanings of the constituent words. e.g. in *"a double blind expertise"* the individual constituent words have no connotation relation to the actual meaning of the phrase, which is to ask experts to give suggestions about some matter. Contrast this with *blind man*

which has no other interpretation than the literal one obtained from the composition of the constituent words. These kinds of collocations are very common in human language due to the prolific metaphorical and figurative use of language. Handling these kinds of MWT is crucial for robust natural language processing [5].

4 Morphology of Compound Terms

Using the resource glossary [6] containing more than 2,500 entries, we performed an automatic morphological analysis of single-word and compound-word psychological terms. As a result of the morphological analysis each term is attributed to the following type of information:

Table 1. Morphological characteristics of the compound word term

№	word form	lemma	morphology	POS
1	адекватность (adequacy)	адекватность (adequacy)	ж=Ns;Asi;	n
2	ощущения (feelings)	ощущение (feeling)	с=Gs;Np;Api;	n
3	и (and)	и (and)	союз=0;	cnj
4	восприятия (of perception)	восприятие (perception)	с=Gs;Np;Api;	n

where ж - feminine noun; с – neuter noun; N – nominal case; s - singular; A - accusative case; i - the inanimate; n - noun; etc. (Notation corresponds to the electronic version of A.A.Zaliznyak's dictionary [7]).
Counting the morphological characters of all terms in the glossary gave the results shown in the following Table 2.

Table 2. Frequency of syntactic structures of terms depending on the number of constituent words.

POS + case	Term example	# of words	Frequency
nN	абазия (abasia)	1	0.313
aN nN	абсолютный порог (absolute threshold)	2	0.319
nN nG	автоматизация движений (movements automation)	2	0.131
nN nN	ведущая деятельность (leading activity)	2	0.013
nN aN	агнозия зрительная (visual agnosia)	2	0.004
numN nG	двенадцать шагов (twelve steps)	2	0.002
nN aG nG	амнезия раннего детства (early childhood amnesia)	3	0.063
aN nN nG	агрессивное поведение животных (aggressive behavior of animals)	3	0.035
aN aN nN	абсолютная слуховая чувствительность (absolute hearing sensitivity)	3	0.029
nN preL nL	либидо во фрейдизме (libido in Freudianism)	3	0.004
nN preG nG	запечатление у животных (imprinting in ani-	3	0.003

	mals)		
nN nG nG	нарушения восприятия времени (time perception disorders)	3	0.003

The frequency of the other grammatical structures is less than 0.002.

5 Logical and Semantic Structure of Terms

The study of the logical and semantic structure of a term was performed using Pointwise Mutual Information (PMI) metrics.
PMI is defined as:
PMI(wa, wb) = ln(p(wa, wb)/p(wa)p(wb)).
Where wa, wb –terms; p(wa, wb) – the probability of co-occurrence wa and wb; p(wa), p(wb) –probabilities of occurrence wa and wb respectively.
The distributional hypothesis in linguistics is: words that occur in similar contexts tend to have similar meanings [9]. This hypothesis is the justification for applying the PMI to measuring word similarity. A word may be represented by a vector in which the elements are derived from the occurrences of the word in various contexts. Proximate row vectors in the word–context matrix indicate similar word meanings.
We used machine-readable texts from the journal, "Voprosy Psychologii" (*Psychological Problems*) issued in the years between 1980-2010, [8] as the source corpus. It contains more than 13 billion words and 282,000 lemmas. A statistical diachronic study of this corpus was published in 2015 [10]. As its base, the experts in psychology chose 100 specific two-word terms as examples of terminology in this field. We performed the usage count for each of these terms and the PMI method for each pair of them. Some pairs with their counts and PMI are shown in Table 3. The table contains the top part of all pairs ordered according to PMI value. Good collocation pairs have a high PMI because the probability of co-occurrence is only slightly lower than the probability of occurrence of each word. These are:
клиническая психология (clinical psychology) <> консультативная психология (counseling psychology); PMI=3.6016
ценности жизни (values of life) <> экзистенциальный анализ (existential analysis); PMI=4.7447
гендерный анализ (gender analysis) <> женская психология (female psychology); PMI=4.7729
The PMI measure could be used for: ranking of web pages, rare term extraction from NL texts, sentiment classification, etc.
In the next stage we performed the search of proximity measure for two multi-word terms based on MTI values. Each term's *Ta* is a vector in multidimensional space (in our case a 100-dimentional space) with components corresponding to the MTI(Ta, Tbi) values where *Tbi* – the set of all the terms (in our case i={1,2,…,100}).

Table 3. The top part of the list of distances between terms ordered in descending order

A-B distance	Term A	Term B
6.0681	методы исследования (re-	психологические исследования (psy-

	search methods)	chological research)
6.1343	индивидуальные особенности (individual characteristics)	индивидуальные различия (individual differences)

The distance measure could be used for information retrieval, terms clustering, multi-dimensional scaling, etc [11].

6 Terms Translation

Finally, we performed a study of translated terms using the online translator.
The methodology of this research was as follows the analysed term was subjected to machine translation with the help of an online translator[1]. Then, the reverse translation was performed, and the Levenshtein distance[2] (LD) between the original Russian text and the text obtained as a result of direct – reverse translation was calculated[3] [10]. If this distance is equal to zero, the term and its foreign equivalent are accepted. Otherwise we conclude that the machine translation system "does not know" the Russian term, and, accordingly, incorrectly selects its foreign equivalent. This result means that the online translator "knows" the Russian term and gives the adequate translation. Otherwise the Russian term should be changed.
Initial (Russian) = вера в справедливый мир
Russian to English = just-world hypothesis
Back English to Russian = вера в справедливый мир // Levenshtein distance = 0
An example of an incorrect translation of the term:
запечатление у животных и **захватив животных // Levenshtein distance = 2*
While iteratively repeating the procedure of direct and reverse translation the process can sometimes come to zero LD. Then the source and the resulting terms often are synonymous.
An example of an iterative equivalents search:
Initial (Russian) = нарушения восприятия времени
R to E = disorders of perception of time
Back E to R = расстройства восприятия времени // LD = 1
R to E = time perception disorder
Back E to R = Время расстройство восприятия // LD = 3
R to E = Time perception disorder
Back E to R = расстройство восприятия времени // LD = 2
R to E = disorder of time perception
Back E to R = расстройство восприятия времени // LD = 0
One can conclude:
нарушения восприятия времени =sin= *расстройство восприятия времени*

[1] Google translator, available at https://translate.google.ru/, last accessed: 30.02.2019
[2] Levenshtein distance, available at https://en.wikipedia.org/wiki/Levenshtein_distance , last accessed: 30.02.2019
[3] Direct and reverse translation available at http://www.philol.msu.ru/~serge/Translation/form11.php. , last accessed: 30.02.2019

The resulting term, after iterations, may be the converse of the source one: *абсолютная слуховая чувствительность* =conv= *абсолютная чувствительность слуха* (*absolute hearing sensitivity* =conv= *absolute sensitivity of the hearing*); or the original term and the resulting term have a different word order (permutation): *амнезия раннего детства* =perm= *раннего детства амнезия (amnesia of early childhood* = perm = *early childhood amnesia*)
Logical and semantic analysis using machine translation allows the selection of clusters of related terms, a partial order relation according to the "Levenstein distance."

7 Conclusion

We have performed an automatic morphological analysis of terms n the field of psychology on the basis of A.A. Zalizniak's grammar dictionary [7]. The statistics of syntactic structures of composite terms (phrases) is extracted for the later retrieval of such terms in the text. The foreign-language equivalents for the terms are searched by automated means through the procedure of multiple direct and reverse translations. The possibility of clustering a set of terms in a given subject area was mentioned. The results can be used to improve MT systems.

References

1. Gorelikova, S.N.: The nature of the term and some features of term formation in English [Priroda termina i nekototye osobennosti trminoobrasovania v angliiskom iazyke] // Vestnik OGU №6, 129-136 (2002)
2. Glushko, M.M. et al.: Functional style of the popular language and methods of its study [Funkcionalnyi stil obschestvennogo iazyka i metody ego issledovania] Moscow. – pp 1–33 (1974)
3. Mescheriakov, B., Zinchenko, V.: The concise psychological dictionary [Bolshoi psikhologicheskii slovar], ACT Moscow, Prime-Evroznak, -pp.1 – 816 (2009)
4. Bleikher, V, Kruk, I: Explanatory dictionary of the psychological terms [Tolkovyi slovar psikhologicheskih terminov] NPO MODEK, Voronezh, pp. 1 – 640 (1995)
5. Kunchukuttan, A.: Multiword Expression Recognition (2017) . – available at https://pdfs.semanticscholar.org/3e3f/d0173dcb28aa1a11d5342da527a835235ae4.pdf last accessed 15.02.2017
6. Barness, E.M., Bernard, D.F.: Psychoanalitic terms and concepts [Psykhoanaliticheskie terminy I poniatia], Class, M, pp.1 – 304 (2000)
7. Zalizniak, A.A.: Grammatical dictionary of Russian [Grammaticheskii slovar russkogo iazyka] (2017) available at: http://www.speakrus.ru/dict/zdf-win.zip, last accessed: 30.02.2019
8. Problems of psychology [Voprosy psihologii] Scientific journal, Issues 1980-2010 years, Pedagogika, Moscow, (1980) available at http://www.voppsy.ru/ last accessed 16.02.2019
9. Harris, Zellig: Distributional structure. Word 10(23). 146–162 (1954)
10. Potemkin, S.B., Hasin, L.A., Hasina, P.L., Schedrina, E.V.: Analysis of psychology development on the basis of terms frequency dynamics [Analiz tendencii razvitia psikhologii na os-

nove vyiavlenia dinamiki chastoty ispolzovania psikhologicheskikh terminov], Problems of psychology [Voprosy psikhologii], Vol.6, 95-103 (2015)
11. Potemkin, S.B., Kedrova, G.E.: Exploring semantic orientation of adverbs, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011" [Komputernaya Lingvistika i Intellektualnye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2011"], Bekasovo, pp.71–78 (2011)