UDC 81-22
DOI https://doi.org/10.32837/2312-3192-2018-11-53-57

MODIFICATION OF TERM SENSE EMBEDDINGS REGARDING WORD-SENSE DISAMBIGUATION

Rodmonga Potapova[1], Ksenia Oskina[2], Vsevolod Potapov[3]

Abstract

This paper proposes a context-based mechanism which makes it possible to approach the solution of word sense disambiguation with respect to the subject domain of speechology (*spoken language sciences*). Special meanings[4] of terms are decomposed into a multidimensional vector space of context words. Hereafter, on the basis of this expansion the program computes the a posteriori probability that the target term in a particular sentence is used with a special meaning. The proposed mechanism can be integrated into the pre-editing module of the Machine Translation system (MT). This article suggests a mechanism for increasing the significance of context words for more accurate determination of meaning of an ambiguous term. This mechanism consists in modifying the coordinates of vector representation of term meaning which correspond to the most significant context words. Criteria for text analysis of speechology will be (1) preciseness of speechology concepts and their definition, explanation, circumscription, etc.; (2) exactness and consistent use of speechology terminology; (3) indicators of a possible merger between object language and metalanguage in microstructure studies referring to the text under analysis; (4) macrostructure of the given text form. The formula for calculating the probability that a term has a special meaning is derived on the assumption that event s has already occurred. As a perspective, it is necessary to empirically calculate the most optimal value of the "importance weight" as well as the threshold for classification model.

Keywords
Natural language processing (NLP), term sense embedding, automatic pre-editing, domain adaptation, word sense disambiguation (WSD), context analysis.

1. Introduction. Speechology (*spoken language sciences*) is a multidisciplinary science of spoken language which was established as a scientific heterogeneous direction in the second part of the XXth century and which includes a set of convergent sciences which, together with spoken language, anatomy, physiology, psychology, cognition, physics, acoustics, mathematics, sociology, medicine, and speech communication today are closely interconnected[5]. This multidisciplinarity of speechology complicates the lexical and phraseological usage of the terminology in this domain.

Today the scope and target of *terminology* is defined as follows: "Terminology is the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e. lexical items belonging to specialized areas of usage of one or more languages. In its objectives it is akin to lexicography which combines the double aim of generally collecting data about the lexicon of a language with providing information, and sometimes even an advisory, service to language users. The justification of considering it a separate activity from lexicography lies in the different nature of the data traditionally assembled, the different background of the people involved in this work, and to some extent in the different methods used Sager[6]."

"A *term*, by definition, is any conventional symbol representing a defined concept. Term as an entry in a specialized dictionary or glossary is accompanied by a *definition*. The classical pattern of a definition is an equation between the *definiendum* and the *definiens*. The *definiendum* is *the term defined*. The *definiens* is composed of the *genus proximum* (the next higher concept in the notional system's hierarchy) and the *differentia specifica* (distinctive characteristics)"[7].

As a rule, technical vocabulary is associated with *terminology*. But terminology is only one sector of the specialist's wordstock, although it constitutes its core. Another sector is occupied by *nomenclatures*[8], that is designations for physical objects or abstract entities in an ordered and homogeneous system, e.g. the Linnaean nomenclatures of botany and zoology, the medical nomenclatures of anatomy and physiology, the periodical system of chemical elements, etc. Still another sector of the specialist's vocabulary is constituted by *professionalisms* or *jargon words*. These are often colorful everyday expressions which designate tools, materials, vehicles, or particular phases of the working process (e.g. doghouse, a slang expression in geophysics meaning 'the drill-master's shed')[9].

[1] Prof., Sc.D. R.K. Potapova. Institute of Applied and Mathematical Linguistics, Moscow State Linguistic University, Ostozhenka 38, Moscow 119034, Russia, Email: RKPotapova@yandex.ru

[2] Ph.D. K.A. Oskina. Institute of Applied and Mathematical Linguistics, Moscow State Linguistic University, Ostozhenka 38, Moscow 119034, Russia, Email: ksenia.oskina@gmail.com

[3] Sc.D. V.V. Potapov. Faculty of Philology, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia, Email: volikpotapov@gmail.com

[4] Meaning which denotes a term used in a certain subject domain (compared with common words).

[5] Potapova, 2010

[6] Sager, 1990, 2-3

[7] Gläser, 1995, 36-37

[8] Gläser, 1989; Sager, 1990, 90-97

[9] Glaeser, 1995, 36

Terminology is connected with a number of basic targets with terminography. This descriptive and normative branch of the *language for special purposes* (LSP) research focuses attention on harmonizing and codifying terminological systems, and develops guidelines for term formation and for LSP glossaries and dictionaries. It also utilizes the systems of information storage and retrieval in data banks.

The speechology includes in a broad sense both the theory of speech as wave motion, how speech waves are produced and heard, how speech connects with neurophysiology, etc. Classical speechology is first of all articulatory phonetics dealing with an inventory of speech sounds defined with the vocal tract functions[10]. The speechology covers a broad scope of professional fields as follows: modeling of sound structure, vocal tract and some basic vowel features, articulatory correlates of acoustic items, representation of verbal information in memory, rhythm of speech, speech pathology, etc.

The speech communication has wide ranging aspects, from a discussion of how humans produce and perceive speech to details of computer-based speech processing for diverse communication applications[11]. Speech communication as an interdisciplinary subject covers a wide field of problems: speech communication (production, perception, analysis, coding, synthesis, recognition, mathematics for speech processing, signals, filtering, convolution, frequency analysis, etc.; speech production and acoustic parameters; hearing, auditory psychophysics, speech stimuli, perception of distorted speech; coding of speech signals, quantization, etc.; linear predictive coding, speech synthesis, speech recognition, speaker verification and recognition, etc.)[12].

The term terminology denotes:

a) the inventory of technical terms, that is, lexical items which designate a defined concept in a particular subject field, and;

b) the theoretical categories, principles and rules for correlating words and phrases to defined concepts, and the recommendations for the lexical material thought suitable for this naming process.

Criteria for text analysis of speechology will be:

– preciseness of speechology concepts and their definition, explanation, circumscription, etc.;

– exactness and consistent use of speechology terminology;

– indicators of a possible merger between object language and metalanguage in microstructure studies referring to the text under analysis;

– macrostructure of the given text form.

2. Motivation. One of the major problems in Natural Language Processing is Word Sense Disambiguation as well as the problem of out-of-vocabulary words processing. In order to solve these problems word sense embeddings are of current interest. For example, attempts are being made to integrate this approach into neural networks[13]. The details of representation of words in a vector form are considered in works[14].

Irrelevant results on the output of modern MT systems (especially when translating scientific and technical texts) are often associated with the lack of domain adaptation[15]. One way of domain adaptation is to introduce a pre-editing module, as well as consider using a multilingual context-oriented terminology dictionary[16], which being combined will make it possible to achieve correctness when translating scientific texts from source language to the target one.

While using context to determine the meaning of ambiguous words, the main problem of determining the significance of context words is that more significant words are not found often enough (due to their complexity and length). As a result modern metrics, used for evaluating the significance of context words, assign low weight to these context words.

As a solution to this problem, the article proposes to increase the weight (or significance) of less frequent but more significant words, which most likely will increase the accuracy of determining the meaning of an ambiguous term. This mechanism can potentially act as part of a pre-editing module for the MT system. The proposed mechanism will make it possible to achieve more relevant results while using context-based methods for determining the special meaning of an ambiguous term[17] in a sentence.

3. Modification of word sense embeddings. The proposed approach was implemented in several stages. At the first stage, a vector representation of word meanings in the context vector space was formed. For this purpose, a specialized corpus of texts on speechology was compiled with a total volume of 204,000 words. The processing of this corpus, vector space construction, and normalization of the vectors in it were implemented using a script in Perl-language.

The script has excluded the following characters and strings from the corpus: new line characters, Latin words, numbers, and various punctuation marks, since they did not affect the final result. A list of speechology terms was automatically created by finding words in the text missing from the list of lemmas and word-

[10] Fant, 1973; Potapova, 1989; Potapova, 2010

[11] O'Shaughnessy, 1987

[12] O'Shaughnessy, 1987; Potapova, 1989; Potapova, 2010; Potapova, Potapov, 2013; Potapova, Potapov, 2014; Potapova, Potapov, 2015; Potapova, Potapov, 2016a; Potapova, Potapov, 2016b; Potapova, Potapov, 2017a; Potapova, Potapov, 2017b

[13] Rios, 2017

[14] Arefyev, 2015; Pelevina, 2016

[15] Oskina, 2016

[16] Potapova, Oskina, 2015

[17] The article considers terms which have become ambiguous due to being borrowed from subject domains other than speechology or from common vocabulary. As a consequence, one of the meanings of the term is special (or belongs to the subject domain of speechology), any other is not.

forms for Russian. This list was supplemented with speechology terms taken from the corresponding dictionary[18]. Tokenization and normalization were carried out, stop words were removed. After that in the corpus there were found all the sentence, where special terms occurred. All the terms within such sentences were taken for context words. These context words made up the basis of a vector space. For decomposable word the total number of skipping bigrams, made up from a special term and a context word from the space basis, is computed. This number will be considered as a coordinate of the resulting vector with respect to the corresponding basis vector in space.

Further, it is proposed to modify the coordinates of the obtained vectors for special terms. If an ambiguous word is found together with a special term within the same sentence, it is most likely that the meaning of the target term will belong to the subject domain in question. The target vector is proposed to be modified by multiplying the coordinate of the vector (which corresponds to the most significant basis vector, or context word) on a coefficient, thus increasing the angle cosine between the analyzed vector and the vector of the most significant context word. This multiplication enables one to correctly define the position of the investigated vector in the constructed vector space.

After modification the resulting vector is normalized according to formula (1).

$$P_i = \frac{count(w_i)}{\sum_i count(w_i)} \qquad 1)$$

where P_i is a coordinate of normalized vector;
count(w_i) is a coordinate of the entry vector;
i is an ordinal number of the basis vector in space.

The coordinate of the normalized vector can be considered as the probability that context word w_i occurs within the same sentence in context with the target term T_n in case the meaning of the term T_n is special *S*.

The probability of the occurrence of a context word is, in fact, the probability that this word occurs in the sentence in the context of term T_n, if the meaning of T_n is *S*. Thus, using the methods of mathematical statistic, the formula of conditional probability of term occurrence can be derived (2).

$$P_i = P(w_i \mid T_n \in S) \qquad 2)$$

where *S* is a set of special terms,
T_n is an ambiguous term.

Next formula (2) can be modified by using the Bayes theorem (3).

$$P(w_i|T_n \in S) = \frac{P(T_n \in S|w_i) * P(w_i)}{P(S)}. \qquad 3)$$

From formula (3) it is possible to derive a formula for a posteriori probability, i.e. the probability that the meaning of the target term T_n is *S*, if the context word has already occurred (4).

$$P(T_n \in S \mid w_i) = \frac{P(w_i|T_n \in S) * P(S)}{P(w_i)} \qquad 4)$$

where P($T_n \in S$ |w_i) is the a posteriori probability that the meaning of the target term is *S* if the context word has already occurred.

P(S) is the general probability that a class of terms with special meaning will be present in the sample at all, regardless of test sentence. The probability of occurrence of class S is taken to be equal to 1 in the test sample: P(S) = 1.

Then the classification stage follows, where it is necessary to classify the input word as belonging to class *S* or to class $\overline{S}$. Suppose there is an unknown sentence *s*. It is necessary to calculate the a posteriori probability for $T_n \in S$ (i.e. the probability that term T_n in the given sentence will take the *S* meaning if context word w_i is encountered): $P(T_n \in S \mid s)$.

Further, using the composite probability formula for conditional probabilities, it is possible to "extend" the probability $P(T_n \in S \mid s)$ to context words w_i(5).

$$P(T_n \in S|s) = \sum_i P(T_n \in S|s \cap w_i) * P(w_i|s) = \\ = \sum_i P(T_n \in S|s \cap w_i) * P(w_i) \qquad 5)$$

$P(T_n \in S|s \cap w_i)$ can be replaced by $P(T_n \in S|w_i)$, because $w_i \subseteq s$, i.e. if event w_i has already occurred, then *s* has also occurred, because the w_i set is included in *s*, and the probability P(s) = 1.

Hence follows equation (6).

$$P(T_n \in S|s) = \sum_i P(T_n \in S|w_i) * P(w_i|s) \qquad 6)$$

$P(T_n \in S|s)$ is the target probability which is to be calculated, i.e. the probability that the term in the sentence will have the meaning of speechology if event *s* has already occurred.

Then $P(T_n \in S|w_i)$ can be replaced according to formula (3):

$$P(T_n \in S|s) = \sum_i \frac{P(w_i|T_n \in S) * P(S)}{P(w_i)} * P(w_i|s) = \\ = \sum_i P(w_i|T_n \in S) * P(w_i|s) * \frac{P(S)}{P(w_i)} = \\ \sum_i P(w_i|T_n \in S) * P(w_i|s) \qquad 7)$$

where α is the coefficient introduced in this work, which represents the difference between the probability that class S will be present at all and the probability of occurrence of a context word;

$P(w_i|T_n \in S)$ is conditional probabilities of context words;

$P(w_i|s)$ is the probability that context word will occur in sentence s.

The following (8) holds for $P(w_i|s)$.

[18] Potapova, 2010

$$P(w_i|s) = \begin{cases} 1, \text{if context occurred in } s \\ 0, \text{if context word did not occur in } s \end{cases} \quad 8)$$

For α it is necessary to make an approximation that the probability of occurrence of a context word is approximately the same for all w_i, i.e. this probability is a constant, and hence the formula (9) can be derived.

$$P(T_n \in S|s) = \alpha * \sum_i P(w_i|T_n \in S) * P(w_i|s) \quad 9)$$

After that it is necessary to determine the value range of α. Assuming that P(S) ≈ 0,5, and $P(w_i) \approx [0{,}05 \ldots 1]$, α will vary in range $0.5 < \alpha < 10$.

Considering the target probability being equal to the product of α and β and the fact that α is constant, it is possible to estimate the introduced coefficient β according to formula (10).

$$\beta = \sum_i P(w_i|Sp) * P(w_i|S) \quad 10)$$

β can be accurately calculated under the condition of event s. Then, if the value of β exceeds a certain threshold, a decision is made that the probability $P(T_n \in S|s)$ is high enough to classify the meaning of target term T_n as the one belonging to set *S*.

4. Conclusion. This article suggests a mechanism for increasing the significance of context words for more accurate determination of meaning of an ambiguous term. This mechanism consists in modifying the coordinates of vector representation of term meaning which correspond to the most significant context words. The formula for calculating the probability that a term has a special meaning is derived on the assumption that event *s* has already occurred. As a perspective, it is necessary to empirically calculate the most optimal value of the "importance weight" as well as the threshold for classification model.

5. Acknowledgements: This research is supported by the Russian Science Foundation, Project № 18-18-00477.

BIBLIOGRAPHY

Потапова Р.К. Речевое управление роботом. – М.: Радио и связь, 1989. – 248 с.

Потапова Р.К. Речь: коммуникация, информация, кибернетика. – Изд. 4-е. – М.: Либроком, 2010. – 600 с.

Arefyev N., Panchenko A., Lukanin A., Lesota O., Romanov P. Evaluating Three Corpus-based Semantic Similarity Systems for Russian // Dialog, 28. Moscow (2015) (25.05. 2018: http://www.dialog-21.ru/media/1119/arefyevnvetal.pdf)

Arntz R., Picht H. Einführung in die Terminologiearbeit. – 2d ed. – Hildesheim, Zürich, New York: Georg Olms Verlag, 1991. – 331 S.

Fant G. Speech Sounds and Features. – Cambridge, Massachusetts, and London, England: The MIT Press, 1973. – 227 p.

Felber H., Budin G. Terminologie in Theorie und Praxis. – Tübingen: Gunter Narr Verlag, 1989. – 126 S.

Gläser R. Nomenklaturen im Grenzbereich von Onomastik und Fachsprachenforschung // Peterson L., Strandberg S. (eds.) Studia Onomastica. Festskrift till Th. Andersson, 23 februari 1989. Lund, 1989. – P. 105-114.

Gläser R. Linguistic Features and Genre Profiles of Scientific English. – Frankfurt am Main: Peter Lang, 1995. – 250 S.

O'Shaughnessy D. Speech Communication: Human and Machine. – New York: Addison-Wesley Publishing Co., 1987. – 548 p.

Oskina K. Text Classification in the Domain of Applied Linguistics as Part of a Pre-editing Module for Machine Translation Systems // SPECOM 2016. Lecture Notes in Artificial Intelligence. – Vol. 9811. – Heidelberg: Springer, 2016. – P. 691–698.

Pelevina M., Arefyev N., Biemann Ch., Panchenko A. Making Sense of Word Embeddings // Proceedings of the 1st Workshop on Representation Learning for NLP. – Berlin: ACL, 2016. – P. 174–183.

Picht H., Draskau J. Terminology: An Introduction. – Guildford (England): University of Surrey, 1985. – 265 p.

Potapova R., Oskina K. Semantic Multilingual Differences of Terminological Definitions Regarding the Concept "Artificial Intelligence" // SPECOM 2015 . Lecture Notes in Artificial Intelligence. – Vol. 9319. Heidelberg: Springer, 2015. – P. 356-363.

Potapova R.., Potapov V. Auditory and visual recognition of emotional behavior of foreign language subjects (by native and non-native speakers) // Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS. – Vol. 8113. – Heidelberg: Springer, 2013. – P. 62–69.

Potapova R., Potapov V. Associative mechanism of foreign spoken language perception (forensic phonetic aspect) // Ronzhin, A., Potapova, R., Delic, V. (eds.) SPECOM 2014. LNCS. – Vol. 8773. – Heidelberg: Springer, 2014. – P. 113–122.

Potapova R., Potapov V. Cognitive mechanism of semantic content decoding of spoken discourse in noise // Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM'2015. LNCS. – Vol. 9319. –Heidelberg: Springer, 2015. – P. 153–160.

Potapova R., Potapov V. On individual Polyinformativity of speech and voice regarding speaker's auditive attribution (forensic phonetic aspect) // Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS – Vol. 9811. – Heidelberg: Springer, 2016/ – P. 507–514.

Potapova R., Potapov V. Polybasic attribution of social network discourse // Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS. – Vol. 9811. – Heidelberg: Springer, 2016. – P. 539–546.

Potapova R., Potapov V. Cognitive entropy in the perceptual-auditory evaluation of emotional modal states of foreign language communication partner // Karpov, A., Potapova, R., Mporas, I. (eds.). SPECOM 2017. LNAI. – Vol. 10458. – Cham: Springer, 2017 – P. 252–261.

Potapova R., Potapov V. Human as acmeologic entity in social network discourse (multidimensional approach) // Karpov, A., Potapova, R., Mporas, I. (eds.). SPECOM 2017. LNAI. – Vol. 10458. –Cham: Springer, 2017. – P. 407–416.

Rios A., Mascarell L., Sennrich R. Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings // Second Conference on Machine Translation. – Kopenhagen: ACL, 2017. – P. 11–19.

Sager J.C. A Practical Course of Terminology Processing. – Amsterdam, Philadelphia: J. Benjamins Pub. Co., 1990. – 258 p.

REFERENCES

Arefyev N., Panchenko A., Lukanin A., Lesota O., Romanov P. (2015). Evaluating Three Corpus-based Semantic Similarity Systems for Russian. Dialog, 28. Moscow, 2015. (25.05. 2018: http://www.dialog-21.ru/media/1119/arefyevnvetal.pdf)

Arntz R., Picht H. (1991). Einführung in die Terminologiearbeit. 2d ed. Hildesheim, Zürich, New York: Georg Olms Verlag.

Fant G. (1973). Speech Sounds and Features. Cambridge (Massachusetts), L. (England): The MIT Press.

Felber H., Budin G. (1989). Terminologie in Theorie und Praxis. Tübingen: Gunter Narr Verlag.

Gläser R. (1989) Nomenklaturen im Grenzbereich von Onomastik und Fachsprachenforschung. Peterson L., Strandberg S. (eds.) Studia Onomastica. Festskrift till Th. Andersson, 23 februari 1989. Lund, 1989. Pp. 105–114.

Gläser R. (1995). Linguistic Features and Genre Profiles of Scientific English. Frankfurt am Main: Peter Lang.

O'Shaughnessy D. (1987). Speech Communication: Human and Machine. NY: Addison-Wesley Publishing Co.

Oskina K. (2016). Text Classification in the Domain of Applied Linguistics as Part of a Pre-editing Module for Machine Translation Systems. SPECOM 2016. Lecture Notes in Artificial Intelligence. Vol. 9811. Heidelberg: Springer. Pp. 691–698.

Pelevina M., Arefyev N., Biemann Ch., Panchenko A. (2016). Making Sense of Word Embeddings. Proceedings of the 1st Workshop on Representation Learning for NLP. Berlin: ACL. Pp. 174–183.

Picht H., Draskau J. (1985). Terminology: An Introduction. Guildford (England): University of Surrey.

Potapova R.K. (1989). Rechevoe upravlenie robotom [Речевое управление роботом]. Moscow: Radio i svyaz'.

Potapova R.K. (2010). Rech': kommunikatsiya, informatsiya, kibernetika [Речь: коммуникация, информация, кибернетика]. 4th ed. Moscow: Librokom.

Potapova R., Oskina K. (2015). Semantic Multilingual Differences of Terminological Definitions Regarding the Concept "Artificial Intelligence". SPECOM 2015. Lecture Notes in Artificial Intelligence. Vol. 9319. Heidelberg: Springer. Pp. 356–363.

Potapova R.., Potapov V. (2013). Auditory and visual recognition of emotional behavior of foreign language subjects (by native and non-native speakers). Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS. Vol. 8113. Heidelberg: Springer. Pp. 62–69.

Potapova R., Potapov V. (2014). Associative mechanism of foreign spoken language perception (forensic phonetic aspect). Ronzhin, A., Potapova, R., Delic, V. (eds.) SPECOM 2014. LNCS. Vol. 8773. Heidelberg: Springer. Pp. 113–122.

Potapova R., Potapov V. (2015). Cognitive mechanism of semantic content decoding of spoken discourse in noise. Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM'2015. LNCS. Vol. 9319. Heidelberg: Springer. Pp. 153–160.

Potapova R., Potapov V. (2016). On individual Polyinformativity of speech and voice regarding speaker's auditive attribution (forensic phonetic aspect). Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS. Vol. 9811. Heidelberg: Springer. Pp. 507–514.

Potapova R., Potapov V. (2016). Polybasic attribution of social network discourse. Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS. Vol. 9811. Heidelberg: Springer. Pp. 539–546.

Potapova R., Potapov V. (2017). Cognitive entropy in the perceptual-auditory evaluation of emotional modal states of foreign language communication partner. Karpov, A., Potapova, R., Mporas, I. (eds.). SPECOM 2017. LNAI. Vol. 10458. Cham: Springer. Pp. 252–261.

Potapova R., Potapov V. (2017). Human as acmeologic entity in social network discourse (multidimensional approach). Karpov, A., Potapova, R., Mporas, I. (eds.). SPECOM 2017. LNAI. Vol. 10458. Cham: Springer. Pp. 407–416.

Rios A., Mascarell L., Sennrich R. (2017). Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. Second Conference on Machine Translation. Kopenhagen: ACL. Pp. 11–19.

Sager J.C. (1990). A Practical Course of Terminology Processing. Amsterdam, Philadelphia: J. Benjamins Pub. Co.

Анотація

Мовленнєва комунікація як міждисциплінарна галузь знань охоплює широке коло проблем творення мовлення (рос."речепроизводство"), мовленнєвого сприйняття (рос."речевосприятие"), аналізу мови, кодування, синтезу, розпізнавання, математики для обробки мовлення, частотного аналізу та ін. Важливим є вивчення продукування мовлення та акустичних параметрів слуху, слухової психофізики, мовленнєвих стимулів, сприйняття спотвореної мови, кодування мовленнєвих сигналів, квантування, лінійного передбачення кодування, дикторської верифікації та ідентифікації та ін. У статті пропонується контекстно-залежний програмний підхід, який дозволяє підійти до вирішення смислової неоднозначності стосовно предметної галузі процесів мовлення (рос. "речеведение") як міждисциплінарної науки, об'єктом якої є усне мовлення. Спеціальні значення термінів розглядаються як окремі кластери багатовимірного векторного простору, що включає масиви контекстних слів. Надалі, базуючись на даному підході, програма обчислює апостеріорну ймовірність того, що цільовий термін в конкретному реченні використовується зі спеціальним значенням. Запропонований механізм може бути інтегрований в модуль попередньої обробки системи машинного перекладу. В якості вирішення цього завдання в статті пропонується збільшити вагу (або значення) менш частотних, але разом з тим найбільш значущих слів, що, допоможе підвищити точність визначення значення двозначного терміну. Цей механізм може потенційно діяти як частина модуля попереднього редагування для системи машинного перекладу. Запропонований програмний підхід дає змогу досягти більш високих результатів з опорою на використання контекстного підходу при визначенні феномена термінологічної двозначності в уже згадуваному тексті.

Ключові слова

Обробка природної мови, термінологічне смислове вкладення, автоматичне попереднє редагування, адаптація домену, зняття двозначності в значенні слова, контекстуальний аналіз.