ИНФОРМАТИКА

COMPUTER SCIENCE

МРНТИ 20.23.19
УДК 004.912

https://doi.org/10.51889/3879.2022.77.24.017

LINGUISTIC ONTOLOGY AS MEANS OF MODELING OF A COHERENT TEXT

Aitim A.K.[1*], *Satybaldiyeva R.Zh.*[2]

[1]*International Information Technology University, Almaty, Kazakhstan*
[2]*Satbayev University, Almaty, Kazakhstan*
[*]*e-mail: a.aitim@iitu.edu.kz*

Abstract

In real time, the size of information resources in natural language is growing rapidly. The processing of these resources urgently requires the presence of linguistic databases and knowledge. Processing of information resources in natural language requires the presence of text corpora and thesauri. To create and process them, markup languages and ontological models of subject areas are required. Insufficient use of linguistic and ontological knowledge used in information retrieval and automatic text processing applications leads to various problems: irrelevant search, poor-quality categorization and referencing of documents. The existing markup languages mainly contain concepts of Romano-Germanic and Slavic language groups. These puzzles are considered burning in the field of computational linguistics. For these purposes, it is proposed to create a metalanguage and an ontological model of the grammar of the Kazakh language.

Keywords: ontological model, Kazakh language, natural language, linguistic, Kazakh grammar, semantic.

Аңдатпа

Ә.Қ. Әйтім[1], *Р.Ж.Сатыбалдиева* [2]
[1]*Халықаралық Ақпараттық Технологиялар Университеті, Алматы қ., Қазақстан*
[2]*Сатпаев Университеті, Алматы қ., Қазақстан*

ЛИНГВИСТИКАЛЫҚ ОНТОЛОГИЯ БАЙЛАНЫСТЫ МӘТІНДІ МОДЕЛЬДЕУ ҚҰРАЛЫ

Қазіргі уақытта табиғи тілдегі ақпараттық ресурстардың көлемі тез өсуде. Бұл ресурстарды өңдеу жедел түрде лингвистикалық мәліметтер базасы мен білімнің болуын талап етеді. Ақпараттық ресурстарды табиғи тілде өңдеу мәтіндік корпус пен тезауристан құралады. Ақпаратты іздеу және мәтінді автоматты өңдеу қолданбаларында қолданылатын лингвистикалық және онтологиялық білімдерді жеткіліксіз пайдалану әртүрлі мәселелерге әкеледі. Қолданыстағы белгілеу тілдерінде негізінен роман-герман және славян тілдері топтарының ұғымдары бар. Оларды жасау үшін белгілеу тілдері және пәндік облыстардың онтологиялық үлгілері қажет болады. Бұл есептеуіш лингвистика саласында кеңінен таралған деп саналады. Сонымен қатар, мәтінді автоматты өңдеудің заманауи әдістеріне тіл мен әлем туралы қосымша білім көлемін енгізу күрделі мәселе болып табылады. Осы мақсатта қазақ тілі грамматикасының метатілі мен онтологиялық моделін жасау ұсынылады.

Түйін сөздер: онтологиялық модель, қазақ тілі, табиғи тіл, лингвистикалық, қазақ грамматикасы, семантикалық.

Аннотация

А.К. Айтим[1], *Р.Ж. Сатыбалдиева*[2]
[1]*Международный Университет Информационных Технологий, г.Алматы, Казахстан*
[2]*Сатпаев Университет, г.Алматы, Казахстан*

ЛИНГВИСТИЧЕСКАЯ ОНТОЛОГИЯ КАК СРЕДСТВО МОДЕЛИРОВАНИЯ СВЯЗНОГО ТЕКСТА

В настоящее время объем информационных ресурсов на естественном языке стремительно растет. Развитие этих ресурсов требует наличия актуальной лингвистической базы данных и знаний. Обработка информационных ресурсов на естественном языке состоит из корпуса текстов и тезауруса. Недостаточное использование лингвистических и онтологических знаний, используемых в приложениях для поиска информации и обработки текстов, приводит к различным проблемам. Существующие языки нотации в основном содержат понятия романо-германской и славянской языковых групп. Для их создания потребуются языки разметки и онтологические модели предметных областей. В то же время в современные методы автоматической обработки текстов трудно внедрить дополнительные знания о языке и мире. Для этого предлагается создать метамодель и онтологическую модель грамматики казахского языка.

Ключевые слова: онтологическая модель, казахский язык, естественный язык, лингвистика, казахская грамматика, семантика.

Introduction

Currently, due to the huge volumes of electronic documents, there is an increasing need for processing unstructured textual information, improving the quality and efficiency of existing text processing methods. The actively developing areas of processing unstructured textual information include tasks such as information retrieval, filtering, categorization, and clustering of documents, searching for answers to questions, automatic annotation of a document and a group of documents, search for similar documents and duplicates, document segmentation and much more.

Modern information retrieval and information analysis systems work with textual information in broad or unlimited subject areas, therefore, a characteristic feature of modern methods of processing textual information has become the minimal use of knowledge about the world and about language, reliance on statistical methods of accounting for the frequency of occurrence of words in a sentence, text, set of documents, the common occurrence of words, etc.

At the same time, the introduction of additional volumes of knowledge about the language and the world into modern methods of automatic text processing is a difficult task. This is since such knowledge must be described in specially created computer resources (thesauruses, ontologies), which must contain descriptions of tens of thousands of words and phrases. When using such resources, it is usually necessary to automatically resolve the ambiguity of words, i.e., choose the correct meaning. In addition, since the management of any resource’s lags the development of the subject area, it is necessary to develop combined methods that consider both knowledge and the best modern statistical methods of text processing.

The metalanguage is intended for marking up texts of natural languages. A metalanguage has the following properties: with the help of its linguistic means, it is possible to express everything that is expressible by means of an object language, and to designate all signs, expressions of an object language for which there are names; in a metalanguage, it is possible to talk about the properties of an expression of an object language and the relations between them; it is possible to formulate definitions, designations, rules of education and transformations for object language expressions. There are well-known marking systems, such as the CLAWS marking system, which is used for marking the British Corps, the Brown Corps marking system, several marking systems are used in the American national Corps, about marking systems. All these systems are mainly used for marking up the English language.

These metalanguages are not adapted to describe the Turkic languages, which have many concepts different from the concepts of the above language groups. Therefore, the creation of a single metalanguage for marking up texts of Turkic languages is an urgent task for processing Turkic languages. Such a language will allow to unify the markup, facilitate their understanding, and use common software, as well as allow for a comparative analysis of linguistic concepts of the Turkic languages.

And also, in addition to the markup language, for computer processing of any natural languages, the formalization of their grammatical (morphological and syntactic) rules, the development of algorithms for the analysis and synthesis of words and sentences according to these rules, the software implementation of all these algorithms, the creation of thesauri by subject areas similar to WordNet, the construction of text corpora (database of marked texts) and other programs for text analysis and processing.

Ontology is used to formalize grammatical rules. Ontology is a conceptual scheme consisting of many concepts and many statements about these concepts, based on which classes, relationships, properties, functions, and individuals can be described. All ontological models are built in the Protégé environment, which makes it possible to simplify the process of creating, downloading, modifying, and transforming the knowledge base, as well as to provide it for general use in the form of joint viewing and editing. Ontology is a knowledge base, because if you add interpretive functions to the structural–semantic model, it will become a knowledge base.

Morphology modeling is related to all applications such as natural language processing and tasks, and includes information intelligence, sentiment testing, spelling correction, generated word detection, part-of-speech tagging, and entity extraction. Morphology is used in linguistics to refer to the study of the structure and formation of texts. Agglutinative languages are languages whose morphological system is characterized by the agglutination of all possible formants. Either prefixes or suffixes act as a formant, and each of them contains its own meaning.

1 Modeling the Structure of a Connected Text

Initially, information-search thesauruses were widely used as resources for information search. But they were created for manual indexing of documents by human indexers, and in recent decades their role has sharply decreased. Then a lot of experiments in the field of automatic text processing and information retrieval were carried out based on the thesaurus WordNet1. However, this thesaurus was created as a test of psycholinguistic theory and does not consider the features of automatic text processing, which is why there are many problems in its use in applied developments. In addition, many researchers have shown insufficient formalization of descriptions in the above types of thesauri, which leads to serious problems with automatic logical inference, necessary in many applications of automatic text processing and information search (search query extension, category output, ambiguity resolution, etc.).

Problems with logical inference are amplified when processing whole texts (as opposed to processing a single sentence), which can contain hundreds or thousands of words and have a complex internal structure.

One of the modern paradigms of computer resources describing knowledge about the world and subject areas is the so-called formal ontologies. At the same time, many researchers in this field see as their goal the development of rather complex formal approaches to the description of practically axiomatized theories. However, automatic processing of unstructured natural language texts with their ambiguity and inaccuracy is difficult to carry out using axiomatized theories. In addition, descriptions within such formalisms do not scale well for knowledge representations in broad unstructured subject areas.

When processing modern news streams, automatic annotation of news clusters, aggregates of messages on the same topic is of great importance. A news cluster is a collection of thematically related documents. Therefore, the thematic structure of a news cluster, as well as a separate element, is revealed by building a thematic representation of this cluster, and this representation is used to control the set proposals to the cluster annotation, namely, to solve such tasks as ensuring completeness, reducing repetitions, as well as ensuring the coherence of the cluster annotation.

To test the proposed model of annotation of the news cluster, the following experiment was conducted. The abstract consisted of a title and four sentences. For news clusters, their thematic representations were obtained.

Further, the manual annotations were marked up for the presence of the main thematic nodes for this cluster and named entities. The result of the analysis was the fact that 83% of the proposals of real manual annotations (out of the total number of proposals) made by linguistic experts satisfy the assumptions made. The peculiarity of the remaining 17% of the sentences is that they were all the last sentences of the manual annotation. The experiment proves that the assumptions made in the method of automatic annotation of news clusters have a high correlation with the structure of human annotations.

A multifactorial model for automatic extraction of terms from texts for automated expansion of the linguistic ontology.

The paper shows that the automatic selection of terms should be based on a set of different features of words and phrases, which should be combined into a multifactorial model. At the same time, such multifactor models should be portable from one subject area to another. In this study, an approach is proposed to identify many signs for automatic extracting terms from texts and combining these features with machine learning methods.

Linguistic ontologies are used in two stages: as one of the sources of signs and to assess the quality of extraction [1].

The proposed model uses three types of features to extract terms:

- features based on the text collection of the subject area.
- signs obtained from the information of the global search engine.
- features obtained based on a given thesaurus of the subject area (Fig.1.).

This is how the situation of the development of the existing thesaurus is modeled, in which the knowledge described in the current version of the thesaurus should be used to automatically extract new terms. The feature sets differ for individual words, two-word phrases, and phrases with many words.

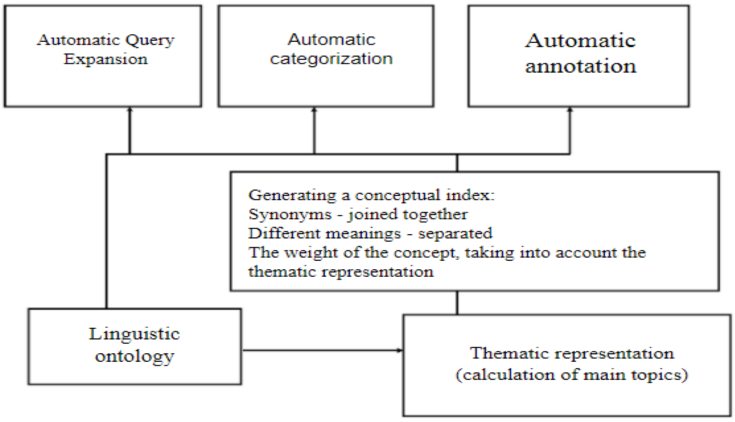
Automatic Query Expansion
Automatic categorization
Automatic annotation
Generating a conceptual index:
Synonyms - joined together
Different meanings - separated
The weight of the concept, taking into account the thematic representation
Linguistic ontology
Thematic representation (calculation of main topics)

Figure 1. Scheme of automatic word processing applications

Machine learning methods are used to combine the selected features for the best extraction of domain terms. The task of applying the methods is to reorder the original list of words (originally ordered as the frequency decreases) so that as many terms as possible get to the top of the list. Thus, the best reordering of the list will reduce the expert's labor costs for entering terms into terminology resources – the expert will view words that are not terms less (Fig.2.).

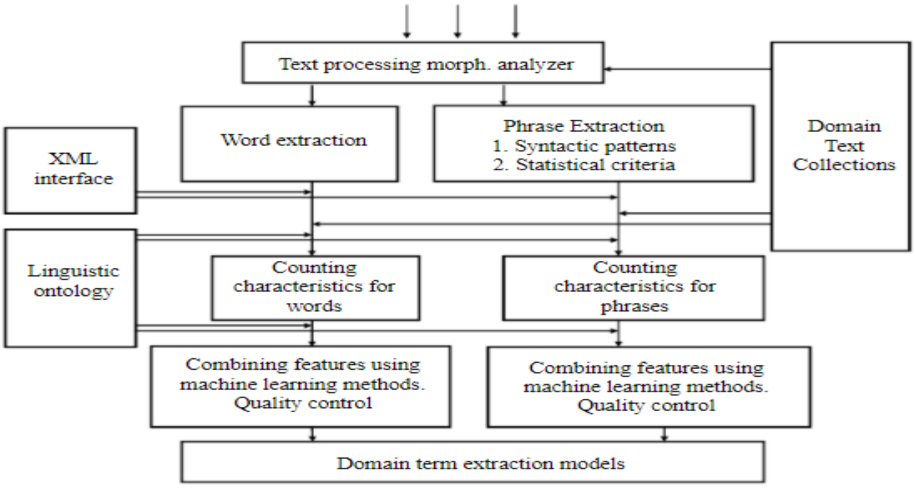
Text processing morph. analyzer
Word extraction
Phrase Extraction
1. Syntactic patterns
2. Statistical criteria
Domain Text Collections
XML interface
Linguistic ontology
Counting characteristics for words
Counting characteristics for phrases
Combining features using machine learning methods. Quality control
Combining features using machine learning methods. Quality control
Domain term extraction models

Figure 2. Scheme of the proposed method for self-acting extraction of definitions

2 The Structure of a Connected Text

Almost all word processing models in the field of information retrieval hope to self-embed texts in a coherent text. Between what is known, in fact, that the word has many texts related in meaning and contains an internal hierarchical structure. There are a fairly large number of possible applications of self-acting word processing that would be able to provide the best results if it were possible to mechanically identify the structure of the content of the connected word. Between them these applications are like self-acting segmentation of words, allowing polysemy, correct intelligence of information, one of the best determinations of the weights of definitions in a document, word categorization, automatic word annotation, etc. The concept of word coherence can be considered in several qualities.

Distinguish between structural connectivity and word connectivity. In practice, we are talking about internal (structural) and external (pragmatic) connectivity. Connectivity is the connection of text components, in which the interpretation of 1 text components is dependent on others. Connectedness is a connectedness brought about by something external to the word, primarily the knowledge of its addressee. Based on this knowledge, the addressee can build a specific list and add links that are obviously not present in the text. From a different point of view, mass and local connectivity of a word is distinguished. The mass coherence of a word is guaranteed by the fact that the word itself contains a single theme. The local connectivity of the discourse takes place in the relationship between adjacent smallest units of the word [2].

Thematic Structure and Thematic Coherence of the Text

Determining the leading topic of a word is considered a necessary step for many information retrieval applications. the concept of the leading (or global) theme of the word is associated with these qualities of the word, as a connection directed to a certain topic and a construction directed to a certain topic. A word can be de jure connected through all sorts of guises of coherence, but if it does not contain a single theme, then it does not have the ability to be considered as a word. The topic of the whole word can be characterized in terms of subtopics, and subtopics in terms of even more local subtopics. Any sentence of the word corresponds to that or another subtopic of the hierarchical structure of the word. The macrostructure of a word determines its mass connectedness. “Without such mass connectivity, it would be impossible to steer local connections. Proposals have every chance of being perfectly connected in terms of local connectivity, but they could be rejected if there were no massive restrictions on their content” [3].

Including the manual processing of the word by specialists, it is not easy to discern the hierarchical structure of the word one by one. Thus, when manually indexing or categorizing documents, different interpretations of side documents by different specialists are considered one of the significant moments of the subjectivity of these processes [4].

In self-acting document processing, the significance of a text or a term for the content of a word, their proximity to the leading topic of the document is evaluated with the support of special weights. It is expected, in fact, that the higher in the hierarchy of the structure directed to a certain subject the text or the term is mentioned, that they are closer to the main thing.

3 Method of Text Polarity Analysis

There are 3 leading ways to determine the polarity of a word:

1. Testing the word by means of vector analysis (often with the introduction of r-gram models), comparing the word with the previously indicated reference body according to the chosen measure of proximity, and classifying (classifying) it as negative or flatter in the end comparison [5].

2. Reconnaissance of psychological vocabulary (lexical polarity) in the text according to polarity dictionaries compiled in advance (lists of templates) with the introduction of linguistic analysis, the word can be assessed on a scale that reflects the number of negative and positive vocabulary. This method can apply as lists of patterns that are inserted into systematic expressions, for example, and criteria for combining a polarity dictionary into a sentence.

3. Mixed method (combination of the first and 2 approaches). In real-time word polarity test produced from several frontiers.

At the first stage, Internet pages are processed according to a special category, which analyzes the pages for the table of contents of the main texts from the database.

At the second stage, the word of selected web pages is processed by morphological analysis to determine the parts of speech and the data of any part of speech.

At the third stage, a lightweight syntactic test works: texts and phrases are combined into chains of polarities; A sentence has a subject, a predicate, and an object.

At the fourth stage, a lightweight self-acting instruction works; it determines the polarity of the word.

Morphological Analysis

Kazakh is an ordinary Turkic language, retaining the bulk of the group's cumulative devil and having some common Kipchak features. structural and typological features of the Kazakh language in the leading language are associated with its adaptation to agglutinative languages [6]. As a rule, the description of an agglutinating similarity is used for a set of symptoms, which includes not only phonetic, but also morphological and syntactic symptoms. The order of completion in the Kazakh language is strictly defined. For example, for nouns at the beginning, completions of a numerous number are added to the base of the text, followed by possessive completions (that is, the object belongs to the person), followed by case completions, and after that, completions of conjugation forms (animate nouns are added). In general, the Kazakh language is a perfectly formalized language [7].

Almost all word processing models in the field of information retrieval hope to self-embed texts in a coherent text. Between what is known, in fact, that the word has many texts related in meaning and contains an internal hierarchical structure. There are a fairly large number of possible applications of self-acting word processing that would be able to provide the best results if it were possible to mechanically identify the structure of the content of the connected word. Between them, these applications, such as self-acting segmentation of words, allowing polysemy, correct intelligence of information, one of the best determinations of the weights of definitions in a document, word categorization, automatic word annotation, etc. The concept of word coherence can be considered in several qualities [8].

Distinguish between structural connectivity and word connectivity. In practice, we are talking about internal (structural) and external (pragmatic) connectivity. Connectivity is the connection of text components, in which the interpretation of 1 text components is dependent on others [9]. Connectedness is a connectedness brought about by something external to the word, primarily the knowledge of its addressee. Based on this knowledge, the addressee can build a specific list and add links that are obviously not present in the text. From a different point of view, mass and local connectivity of a word is distinguished. The mass coherence of a word is guaranteed by the fact that the word itself contains a single theme. The local connectivity of the discourse takes place in the relationship between adjacent smallest units of the word [10].

Determining the leading topic of a word is considered a necessary step for many information retrieval applications. the concept of the leading (or global) theme of the word is associated with these qualities of the word, as a connection directed to a certain topic and a construction directed to a certain topic. A word can be de jure connected through all sorts of guises of coherence, but if it does not contain a single theme, then it does not have the ability to be considered as a word [11]. The topic of the whole word can be characterized in terms of subtopics, and subtopics in terms of even more local subtopics. Any sentence of the word corresponds to that or another subtopic of the hierarchical structure of the word. The macrostructure of a word determines its mass connectedness. “Without such mass connectivity, it would be impossible to control local communications. Proposals have every chance of being perfectly connected in accordance with aspects of local connectivity, but they could be rejected if there were no massive restrictions on their content.

Semantic networks are one of the methods of artificial origin of the mind to represent knowledge in natural language. The semantic network provides an information model of the subject area, which has a picture of a targeted graph, the tops of which correspond to the objects of the subject area, and the edges determine the relationship between them. Objects have every chance of being opinions, actions, qualities, processes. Thus, semantic networks are considered one of the methods for representing knowledge in natural language. In recent years, in the works of many scientists, it is proposed to apply graph dynamic systems as a formal basis for designed mental systems, abstract logical-semantic models of mental systems.

In real time, almost all scientists are working on the creation of systems for morphological and syntactic analysis. The note describes layouts applicable for agglutinative languages, and studies on automating the semantic analysis of words in the Kazakh language are presented in a rather small number.

The task of semantic analysis is carried out in the construction of a semantic graph of a word. In contrast to the morphological and syntactic, at the semantic stage, a formal representation of the meaning of the word is noticed. In the process of semantic analysis, the semantic interpretation of texts and systems is performed, the matters between the larger substances of the word are clarified.

Conclusion

Extracting information from texts is a sufficiently developed area of computational linguistics and automatic text processing, offering a wide range of methods and appropriate tools for building various application systems, as well as demonstrating a sufficiently effective solution to the problems of extracting different types of information.

The relevance of the tasks of the direction remains: the construction of an effective model can significantly facilitate the subsequent processing of the extracted structured data, which is a key moment in the life cycle of accumulation and use of new knowledge for processing the text of the Kazakh language.

In systems of self-acting word processing, 5 leading frontiers of analysis are usually distinguished: graphemic, morphological, fragmentation, syntactic and semantic. It has been noted that, in fact, when processing words in the Kazakh language, it is not easy to clearly divide the boundaries of morphological, syntactic, and semantic analysis, there is their connection. This is due to the peculiarities of word formation in the languages of the given structure. Other than that, it's these 3 steps that cause the biggest problems with automation outside of language dependency. As a result, in the provided note, we restrict ourselves to their consideration.

References:

1 Aitim A., Satybaldiyeva R., Wojcik W., (2020) "The construction of the Kazakh language thesauri in automatic word processing system," Proceedings of the 6th International Conference on Engineering & MIS, pp 1-4, doi:10.1145/3410352.3410789.

2 Satybaldiyeva R., Uskenbayeva R., Moldagulova A., Kalpeyeva Z., and Aitim A., (2019) "Features of Administrative and Management Processes Modeling", World Congress on Global Optimization, pp. 842-849, doi:10.1007/978-3-030-21803-4_84.

3 Satybaldiyeva R.Zh., Aitim A.K., (2020) ”Analysis of methods and models for automatic processing systems of speech synthesis,” International Journal of Information and Communication Technologies 1 (2), pp. 118-123.

4 Aitim A.K., Satybaldiyeva R.Zh., (2020) “ Methods of applying linguistic ontologies in text processing”, Proceedings of the Scientific and Technical Society, Kakhak 1 (72), pp.132-137

5 Barzilay R., Lee L., (2017) “Catching the drift: Probabilistic content models, with applications to generation & summarization,” In 43rd Annual Meeting of ACL, Ann Arbor, MI, pp. 113–120.

6 Bateman J., (2015)“Enabling technology for multilingual nat. lang. generation: the KPML development environment,” Nat. Lang. Engineering, 3 (1), pp. 15–56.

7 Berners-Lee T., Hendler J., Lassila O., (2016)“The Semantic Web,”. Sc. American, pp. 34–43.

8 Bontcheva K., (2015)“Generating tailored textual summaries from ontologies,” In 2nd European Semantic Web Conf., Heraklion, Greece, pp. 56-63.

9 Bontcheva K., Cunningham H., (2016)“The Semantic Web: a new opportunity and challenge for human language technology,” In Workshop on Human Lang. Tech. for the SW and Web Services, 2nd Int. Semantic Web Conf., Sanibel Island, FL, pp. 179-185.

10 Reed S., Lenat D, (2017)“Mapping ontologies into Cyc”, AAAI 2002 Conference Workshop on Ontologies for the Semantic Web, Edmonton, Canada, pp. 206-215.

11 Gomez-Perez A., Manzano-Macho D., (2015)“An overview of method and tools for ontology learning from texts,” Knowledge Eng Rev, pp. 187-212, doi:10.1017/S0269888905000251.

12 Maedche A, Staab S. (2016)“Discovering conceptual relations from text,” In: Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, pp. 321-325.