Ann. For. Res. 65(1): 501-507, 2022
DOI: 10.5281/zenodo.7190929

ANNALS OF FOREST RESEARCH
https://www.e-afr.org/

MAIN APPLICATIONS OF COMPUTER LINGUISTICS

Barotova Mubashira Barotovna
Bukhara engineering-technological institute

Abstract: The definition of the term "computational linguistics" is given. The basic concepts of the subject area are considered, the classification of linguistic software is given. Particular attention is paid to the relationship between the terminology of computational linguistics and theoretical linguistics.

Key words: Native language(NL), computer linguistics (CL), ontology, WordNet system, antonyms, hyponymy (genus-species relation), Crosslexic.

Introduction. The development of linguistic processors requires an appropriate presentation of linguistic information about the processed NL. This information is displayed in a variety of computer dictionaries and grammars. Dictionaries are the most traditional form of representing lexical information; they differ in their units (usually words or phrases), structure, coverage of vocabulary (dictionaries of terms of a specific problem area, dictionaries of general vocabulary, etc.).

Materials and experiments. A dictionary unit is called a dictionary entry; it contains information about a lexeme. Lexical homonyms are usually presented in different dictionary entries. The most common in Computer Linguistics are morphological dictionaries used for morphological analysis; their dictionary entry contains morphological information about the corresponding word - part of speech, inflectional class (for inflectional languages), list of word meanings, etc. Depending on the organization of the linguistic processor, grammatical information, for example, word control models, can also be added to the dictionary. There are dictionaries that provide more information about words. For example, the linguistic model "Meaning-Text" essentially relies to an explanatory-combinatorial dictionary, in the dictionary entry of which, in addition to morphological, syntactic and semantic information (syntactic and semantic valences), information about the lexical compatibility of this word is presented. A number of linguistic processors use dictionaries of synonyms. A relatively new type of dictionaries is paronym dictionaries, i.e. outwardly similar words that differ in meaning, for example, alien and alien, editing and reference [1;23]. Another type of lexical resources is phrase bases, in which the most typical phrases of a particular language are selected. Such a base phrases of the Russian language (about a million units) is the core of the Cross Lexica system [1;34]. More complex types of lexical resources are thesauri and ontology. Thesaurus is a semantic dictionary, i.e. dictionary, which semantic connections of words are presented - synonymous, gender-species relations (sometimes called the relation above-below), part-whole, associations. The spread of thesauri is associated with the solution of information retrieval problems [2,4].

Discussion and results. The concept of ontology is closely related to the concept of thesaurus [3,5]. Ontology is a set of concepts, entities of a certain field of knowledge, focused on reuse for various tasks. Ontology can be created on the basis of the vocabulary existing in the language - in

this case they are called linguistic. Such a linguistic ontology is the WordNet system [4,43], a large lexical resource that contains the words of the English language: nouns, adjectives, verbs and adverbs, and their semantic several types of connections. For each of the indicated parts of speech, the words are grouped into groups of synonyms (synsets), between which the relations of antonymy, hyponymy (genus-species relation), meronymy (part-whole relation) are established. The resource contains approximately 25 thousand words, the number of hierarchy levels for the relation species is on average 6-7, sometimes reaching 15. The upper level of the hierarchy forms general ontology - a system of basic concepts about the world. According to the English WordNet scheme, similar lexical resources for other European languages, grouped under a common name EuroWordNet. A completely different kind of linguistic resources is NL grammars, such as which depends on the syntax model used in the processor. In the first approximation, grammar is a set of rules that express the general syntactic properties of words and groups of words. The total number of grammar rules also depends on the syntax model, varying from several tens to several hundreds. In essence, such a problem manifests itself here as the relationship between grammar and vocabulary in the language model: the more information is presented in the dictionary, the shorter the grammar can be and vice versa.

It should be noted that the construction of computer dictionaries, thesauri and grammars is a voluminous and time-consuming work, sometimes even more time-consuming than the development of a linguistic model and the corresponding processor. Therefore, one of subordinate tasks of CL is the automation of the construction of linguistic resources [6, 15]. Computer dictionaries are often formed by converting plain text dictionaries, but often their construction requires a much more complex and painstaking work. This usually happens when building dictionaries and thesauri for rapidly developing scientific fields - molecular biology, informatics and etc. The source material for extracting the necessary linguistic information can be collections and corpora of texts. A corpus of texts is a collection of texts collected according to a certain principle of representativeness (by genre, authorship, etc.), in which all texts are marked up, i.e. are equipped with some linguistic markup (annotations) - morphological, accent, syntactic, etc. [3]. Currently, there are at least a hundred different corpora - for different NL and with different markup, the most famous being the National Corpus of the Russian, English and Uzbek Languages. Labeled corpora are created by linguists and are used both for linguistic research and for tuning (training) models and processors used in CL using well-known mathematical methods of machine learning. So, machine learning is used to tune resolution methods lexical ambiguity, part of speech recognition, resolution of anaphoric references. Since corpora and collections of texts are always limited in terms of the linguistic phenomena represented in them (and corpora, in addition, are created for quite a long time), recently more and more often as a more complete linguistic resource Internet texts are considered [8, 35]. Undoubtedly, the Internet is the most representative source of modern speech samples, but its use as a corpus requires the development of special technologies. The area of application of computational linguistics is constantly expanding, so we characterize the most famous applied tasks solved by its tools. Machine translation [9,6] is the earliest application of CL, with which this field itself arose and developed. The first translation programs

were built over 50 years ago and were based on the simplest word-for-word translation strategy. However, it was quickly realized that machine translation requires a complete linguistic model that takes into account all levels of the language, up to semantics and pragmatics, which has repeatedly hampered the development of this direction. We will note, however, that in the case of a translation into a related language, for example, when translating from Spanish to Portuguese or from Russian to Ukrainian (which have much in common in syntax and morphology), the processor can be implemented based on a simplified model, for example, in based on the same strategy of word-for-word translation.

Currently, there is a whole range of computer translation systems (of varying quality), from large international research projects to commercial automatic translators. Of significant interest are projects of multilingual translation, using an intermediate language, into which encodes the meaning of translated phrases. Another modern direction is statistical translation [5], based on the statistics of the translation of words and phrases (these ideas, for example, are implemented in the search engine translator Google). But despite many decades of development of this whole area, in general The task of machine translation is still very far from a complete solution. Another fairly old application of computational linguistics is information retrieval and related tasks of indexing, summarizing, classifying and categorizing documents [1, 20, 22]. Full-text search of documents in large databases of documents (primarily scientific, technical, business), is usually carried out on the basis of their search images, which are understood as a set of keywords - words that reflect the main theme of the document. At first, only individual words of the SL were considered as keywords, and the search was carried out without taking into account their inflection, which is not critical for weakly inflectional languages such as English. For inflectional languages, for example, for Russian, it was necessary to use a morphological model that takes into account inflection.

The search request was also presented as a set of words, suitable (relevant) documents were determined based on the similarity of the request and the search image of the document. Creating a search image of a document involves indexing its text, i.e. selection of keywords in it [7,8] . Since very often the topic and content of the document are displayed much more accurately than separate words, and phrases, phrases began to be considered as keywords. This significantly complicated the procedure for indexing documents, since it was necessary to use various combinations of statistical and linguistic criteria to select meaningful phrases in the text.

In fact, information retrieval mainly uses a vector text model (sometimes called a bag of words), in which a document is represented by a vector (set) of its keywords. Modern Internet search engines also use this model, indexing texts by the words used in them (at the same time, they use very sophisticated ranking procedures to return relevant documents). The specified text model (with some complications) is also used in the related problems of information retrieval considered below. Abstracting the text - reducing its volume and getting it short summary - abstract (contracted content), which makes it faster to search in collections of documents. A general abstract can also be drawn up for several documents related to the topic. The main method of automatic abstracting is still the selection of the most significant sentences of the abstracted text, for which the keywords of the text are usually calculated first and the coefficient of significance

of the sentences of the text is calculated. The choice of meaningful sentences is complicated by anaphoric links of sentences, the break of which is undesirable - to solve this problem, certain strategies for selecting sentences are being developed. A task close to summarizing is annotating the text of a document, i.e. writing an annotation. In its simplest form, an annotation is a list of the main topics of the text, for the selection of which indexing procedures can be used. When creating large collections of documents, the tasks of classifying and clustering texts are relevant in order to create classes related to the topic. documents [6]. Classification means assigning each document to a certain class with known parameters in advance, and clustering means dividing a set of documents into clusters, i.e. subsets thematically related documents. To solve these problems, machine learning methods are used, and therefore these applied tasks are called Text Mining and belong to the scientific direction known as Data Mining, or intellectual analysis data [6,8]. The task of categorizing a text is very close to classification - its assignment to one of the previously known thematic headings (usually headings form a hierarchical tree of topics). The problem of classification is becoming more widespread, it is solved, for example, when recognizing spam, and a relatively new application is the classification of SMS messages in mobile devices. New and up to date direction of research for the general task of information retrieval - multilingual search in documents. Another relatively new task related to information retrieval is formation of answers to questions (Question Answering) [9]. This task is solved by determining the type of question, searching for texts that potentially contain the answer to this question, and extracting the answer from these texts. A completely different applied direction, which is developing, albeit slowly, but steadily, is the automation of preparation and editing texts on NL. One of the first applications in this direction were programs for automatically detecting word hyphenation and programs for spelling text checks (spellers, or auto-correctors). Despite the seeming simplicity hyphenation problem, its correct solution for many NL (for example, English) requires knowledge of the morphemic structure of the words of the corresponding language, and hence the corresponding dictionary. Spell checking has long been implemented in commercial systems and relies on an appropriate vocabulary and morphology model. An incomplete syntax model is also used, on the basis of which rather frequent all syntactic errors (for example, word agreement errors) are revealed. At the same time in auto-correctors have not yet been implemented to identify more complex errors, for example, misuse of prepositions. Not detected and many lexical errors, in particular errors resulting from typographical errors or incorrect using similar words (for example, weighty instead of weighty). In modern CL research suggests methods for automated detection and corrections of such errors, as well as some other types of stylistic errors [11, 29]. These methods use statistics on the occurrence of words and phrases. An applied task close to supporting the preparation of texts is training natural language, within the framework of this direction, computer systems for teaching languages are often developed - English, Russian, etc. (similar systems can be found on the Internet). Typically, these systems support learning individual aspects of the language (morphology, vocabulary, syntax) and are based on appropriate models, for example, the morphology model. As for the study of vocabulary, electronic analogues of text dictionaries are also used for this (in which, in fact, there are no language models). However,

multifunctional computer dictionaries are also being developed that have no text analogues and are aimed at a wide range of users - for example, Dictionary of Russian phrases Crosslexic [12]. This system covers a wide range of vocabulary - words and their acceptable word combinations, and also provides information on word management models, synonyms, antonyms and other semantic correlates of words, which is clearly useful not only for those who study Russian, but also for native speakers. The next application area worth mentioning is the automatic generation of texts in NL [2]. In principle, this task can be considered a subtask of the machine translation task already considered above, however, within the framework of directions have a number of specific tasks. Such a task is multilingual generation, i.e. automatic construction in several languages of special documents - patent formulas, instructions for the operation of technical products or software systems based on their specification in a formal language. For rather detailed language models are used to solve this problem. An increasingly relevant applied task, often attributed to the direction of Text Mining, is the extraction of information from texts, or Information Extraction [8], which is required when solving problems of economic and industrial analytics. For this, certain objects are selected in the NL test – named entities (names, personalities, geographical names), their relationships and related events with them. As a rule, this is implemented on the basis of partial syntactic text analysis that allows you to process news feeds from news agencies.

Most often, this problem was solved for specialized databases - in this case, the query language is sufficiently restricted (lexically and grammatically) that allows the use of simplified language models. Requests to the base, formulated in NL, are translated into a formal language, after which the search for the necessary information is performed and the corresponding response phrase is built. As the last in our list of CL applications (but not least) Let us indicate the recognition and synthesis of sounding speech. Inevitably arising in these problems, recognition errors are corrected by automatic methods based on dictionaries and linguistic knowledge about morphology. Machine learning will also be applied in this area.

Conclusion. Computer linguistics demonstrates quite tangible results in various applications for automatic processing of texts in NL. Its further development depends both on the emergence of new applications and independent development. Various models of the language, in which many problems have not yet been solved. The most developed are the models of morphological analysis and synthesis. Models syntax has not yet been brought to the level of stable and efficient modules, despite the large number of proposed formalisms and methods. Even less studied and formalized are models of the level of semantics and pragmatics, although automatic discourse processing is already required in a number of applications. Note that the already existing tools of computational linguistics itself, the use of machine learning and text corpora, can significantly advance the solution these problems.

REFERENCES:

1. Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval, Adison Wesley, 1999.

2. Bateman, J., Zock M. Natural Language Generation. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p.304. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.
3. Oakes, M., Paice C. D. Term extraction for automatic abstracting. Recent Advances in Computational Terminology. D. Bourigault, C. Jacquemin and M. L'Homme (Eds), John Benjamins Publishing Company, Amsterdam, 2001, p.353-370.
4. Grishman R. Information extraction. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 545-559.
5. Harabagiu, S., Moldovan D. Question Answering. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 560-582.
6. Hearst, M.A. Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998, p.131-151.
7. Hirst, G. Ontology and the Lexicon. In.: Handbook on Ontologies in Niformation Systems. Berlin, Springer, 2003.
8. Jacquemin C., Bourigault D. Term extraction and automatic indexing // Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. p. 599-615.
9. Mubashira Barotova. (2021). Some Features Of Translating Original Literary Text. JournalNX - A Multidisciplinary Peer Reviewed Journal, *7*(03), 369–373. Retrieved from https://repo.journalnx.com/index.php/nx/article/view/2788
10. Mubashira Barotova. EditorJournals and Conferences. (2021, October 28). Some Specific Features And Difficulties Of The Translation Of The Novel “Jane Air” By Charlotte Brontë. https://doi.org/10.17605/OSF.IO/H4EWG
11. Barotova, M. (2020). The Problems Of Recreating Writers Style In Translation. Theoretical & Applied Science, (2), 189-192.
https://www.elibrary.ru/item.asp?id=42658997
12. Mubashira Barotova. (2019). Successful expressing of phraseological units. Proceedings of The ICECRS, 3. https://doi.org/10.21070/icecrs.v3i0.264
13. Barotovna, B. M. (2018). Visibility as a means of creating communicative motivation on teaching oral communication. Academy, (5 (32)), 74-75.
https://cyberleninka.ru/article/n/visibility-as-a-means-of-creating-communicative-motivation-on-teaching-oral-communication
14. Barotovna, B. M. (2021). Development Of Listening Skills At The Senior Stage Of Teaching English. Web of Scientist: International Scientific Research Journal, 2(10), 119-127.
https://wos.academiascience.org/index.php/wos/article/view/398
15. Barotovna, B. M. (2018). Translating idioms between English and Uzbek from a cultural perspective. Academy, (5 (32)), 77-78.
https://cyberleninka.ru/article/n/translating-idioms-between-english-and-uzbek-from-a-cultural-perspective
16. Khudoyberdievna, S. Z. (2022). Modern Methods of Translating Phraseological Units. Eurasian Research Bulletin, 4, 153-158.

https://geniusjournals.org/index.php/erb/article/view/516.
17. Khudoyberdievna, S. Z. (2017). Didactic games as framework of students in cooperation. Научный журнал, (3 (16)), 48-50. https://cyberleninka.ru/article/n/didactic-games-as-framework-of-students-in-cooperation.
18. Z Saidova., Advantages And Disadvantages Of Modular Object-Oriented Dynamic Learning Environment (Moodle) In The System Of Education. https://journal.buxdu.uz/index.php/journals_buxdu/article/view/43512.
19. Saidova, Z.(2022). Изучение фразеологии и сравнительный анализ фразеологических единиц, выражающих психическое состояние человека. Центр научных публикаций (buxdu. Uz), 8 (8). http://journal. buxdu. uz/index. php/journals_buxdu/article/view/4354.
20. Saidova, Z.(2022). Изучение фразеологии и сравнительный анализ фразеологических единиц, выражающих психическое состояние человека. Центр научных публикаций (buxdu. Uz), 8 (8). извлечено от http://journal. buxdu. uz/index. php/journals_buxdu/article/view/4354.
14.