УДК 811.111' 367.7:659.3

PRINCIPAL PROBLEMS OF NATURAL LANGUAGE PROCESSING SYSTEMS

O. Hyryn,
Zhytomyr Ivan Franko State University,
40, Velyka Berdychivska St., Zhytomyr, 10008, Ukraine
oleg_hyryn@ukr.net
ORCID iD 0000-0002-3641-2440

The article deals with natural language processing, namely that of an English sentence. The article describes the problems, which might arise during the process and which are connected with graphic, semantic, and syntactic ambiguity. The article provides the description of how the problems had been solved before the automatic syntactic analysis was applied and the way, such analysis methods could be helpful in developing new analysis algorithms. The analysis focuses on the issues, blocking the basis for the natural language processing — parsing — the process of sentence analysis according to their structure, content and meaning, which aims to analyze the grammatical structure of the sentence, the division of sentences into constituent components and defining links between them.

Key words: *parsing; natural language processing; statistical machine learning; ambiguity.*

Гирин О.В.

Основні проблеми систем обробки природних мов

Стаття присвячена обробці природної мови, а саме обробці англійських речень. У статті описуються проблеми, які можуть виникнути під час цього процесу, пов'язані з графічною, семантичною, синтаксичною неоднозначністю. У статті наведено опис шляхів вирішення цих проблем до застосування автоматичного синтаксичного аналізу, а також яким чином такі методи аналізу можуть бути корисними для розробки нових алгоритмів аналізу. Аналіз зосереджений на питаннях, які унеможливлюють основу обробки природної мови — парсинг — процес аналізу речень за їх структурою, змістом і значенням, метою якого є аналіз граматичної структури речення, розподіл речень на складові компоненти і визначення зв'язків між ними.

Ключові слова: *синтаксичний аналіз, обробка природної мови, статистичне машинне навчання, неоднозначність.*

Гирин О. В.

Основные проблемы систем обработки естественных языков

Статья посвящена обработке естественного языка, а именно обработке английских предложений. В статье описываются проблемы, которые могут возникнуть во время этого процесса, связанные с графической, семантической, синтаксической неоднозначностью. В статье приведено описание путей решения этих проблем до применения автоматического синтаксического анализа, а также каким образом такие методы анализа могут быть полезны при разработке новых алгоритмов анализа. Анализ сосредоточен на вопросах, которые делают невозможным основу обработки естественного языка — парсинг — процесс анализа предложений по их структуре, содержанию и значению, целью которого является анализ грамматической структуры предложения, распределение предложений на составляющие компоненты и определение связей между ними.

Ключевые слова: *синтаксический анализ, обработка естественного языка, статистическое машинное обучение, неоднозначность.*

Introduction

The use of digital technologies has become an integral part of our lives. Therefore, there arises an urgent need to replace the work performed by people with automatic operation. Natural language processing (NLP) is one of the tasks, which can be performed automatically. The goal of NLP is to study natural language mechanisms (both internal and external) and to use this knowledge in applications and programs that will help facilitate everyday communication with the use of machines.

Theoretical Background

Natural language processing has been studied in numerous works in foreign linguistics since 1967. The issues, related to automatic speech analysis have been reflected in the works of the following scholars: Fleiss J. L. [8], Hollingsworth Ch. [10], Kovar V. [11]

etc. Although in Ukraine the study concerning analysis of an English language has so far been of theoretical character, yet the experience and theoretical results in the field of English grammar, in particular from the generative perspective (Buniyatova I. R. [2], Polkhovska M. V. [5; 6]), can frame a basis to the applied use thereof.

Current application as well as perspectives of natural language processing (NLP) was specified in [4]. The study specifies the use of parsing for the purposes of automatic information search, question answering, logical conclusions, authorship verification, text authenticity verification, grammar check, natural language synthesis and other related tasks, such as analysis of ungrammatical sentences, morphological class definition, anaphora resolution etc [4].

The **aim** of this article is to present the solution status for the problems, which inevitably appear during NLP.

Methods

This research suggests some linguistic issues, which should be considered for the development of syntactic analysis models, as well as the usage of the scientific methods of analysis, synthesis, description and comparison as well as linguistic methods of substitution and transformation in order to solve the main problems arising during the application of automatic syntactic analysis, which have not been sufficiently solved yet.

Results and Discussion

NLP can by no means be called a smooth process. Numerous difficulties arise due to a number of objective reasons, such as the existence of hundreds of natural languages, each possessing syntactic rules as well as variations thereof in a language. Within the same language, there are words that may have different meanings depending on the context of use. Even the graphic level suggests some technical difficulties. Thus NLP has to consider the encoding type, used in a particular document. The text can be stored in different encodings: ASCII, UTF-8, UTF-16 or Latin-1 [14, 74]. Special processing types may be required for punctuation and for numbers. Sometimes it is necessary to handle the use of characters that represent emotions (combinations of characters or special characters), hyperlinks, recurring punctuation marks (... or ---), file extensions and user names containing dots.

Splitting the text into fragments or elements usually means presentation of the text in the form of a words sequence. Should it be the case, the words are referred to as the "lexical element", "lexeme", or just "token", and the process of splitting the text is called "tokenization". This process does not cause particular difficulties in languages that use spacing characters to separate words, but in languages similar to Chinese, this is much more difficult to do, since the characters can denote both syllables and entire words. Moreover, English itself can present some difficulty during the tokenization process, since in English there is a large number of alternative ways of formal representation of the self-same word: it can be spelled together, separately or it can be hyphenated.

Words naturally are combined into phrases and sentences. Determining the boundaries of sentences may also be associated with certain difficulties, although the first glance suggests that it might suffice to find full stops indicating the ends of sentences. However, dots can also occur inside sentences, for example, after abbreviated words etc.

However, grammatical analysis suggests more serious problemsconcerning analysis accuracy than those, connected with text formal representation. Firstly, much depends on the quality of the part-of-speech tagging, which should be very high (97–98 %) [3], but in long sentences it is often possible to encounter an incorrectly recognized part of speech, which leads to further analysis errors. Secondly, existing automatic parsing gives accuracy of about 90–93 % [3], which means that in a long sentence there will almost always be parsing errors. For example, with the accuracy of 90 %, the probability of speech-part tagging without any error for a sentence of 10 words long will be 35 % [3].

The current state of research gives hope for an improvement in the quality of parsing, but often the right syntactic analysis also presupposes understanding the semantics of the sentence. However, there seem to be sentences, which at present can be parsed by a "human" analysis only. Therefore, in the sentence *"I hit a man with a camera",* there can be two different variants of parsing, depending on whether we believe that the hit man had a camera or the camera was used as the instrument for hitting. Of course, to get the most accurate syntactic analysis, it should make sense to leave some of the most likely options, and then determine the correct one by a combination of different factors, including semantic ones.

Sometimes, during the NLP it is essential to determine the relationships between words in different syntactic groups. Such co-reference resolution defines the relationships between specific words denoting the same object, that is, they have the same referent in one or several sentences. For example, in the sentences "*The town is small but beautiful. It is located at the foot of the mountain*". The word "*it*" co-refers to, that is, is referentially identical to the word "*city*". Co-reference phenomena derive from fundamental patterns of text organization. Since the text has a linear structure, and the situation it describes is usually non-linear, the text almost inevitably should contain repeated nomination of elements in the situation described. At each new

reference to the same object, a new nomination of this object is based on what has already been said about this object and on that knowledge which is not verbalized in the text. Although the problem of coherence in linguistics has been thoroughly studied, the practical implementation of this theoretical knowledge is quite complicated [1, 41].

Should a word have several semantic interpretations, in order to determine its meaning in this particular case, it may be necessary to utilize word sense disambiguation (WSD) [14, 77]. Sometimes this means solving some difficulties. For example, in the sentence "*Mary returned home.*" The word "*home*" may mean "*housing that someone is living in*" or "*the state or city where someone lives*".

One of the most open problems in NLP is ambiguity of its units, which can occur at all language levels. It comprises the phenomena of polysemy, homonymy and synonymy. Ambiguity can be either lexical (existence of more than one word meaning, for example, "*bank*"); syntactic, or structural (when one sentence has several possible grammatical options and, accordingly, has a different meaning, such as attachment ambiguity, when a PP can follow both a VP and a NP within the same sentence with the corresponding meaning change: *"The police shot the burglars with guns"*); semantic ambiguity (when the same sentence can be understood differently in different contexts, although lexical or structural polysemy is absent: "*All philologists stick to a theory*"); pragmatic ambiguity (when the same sentence can be understood differently in different contexts, where it may exist "*My brother thinks he is a genius*").

Existing systems of lexical ambiguity solutions have accuracy in the range of 60–70 % [13, 1165] and are more likely to be presented as separate methods. Solving the issue of unambiguity will require the integration of several sources of information and methods.

Thus the primary task for a syntactic analysis is determining whether the sentence is grammatically correct in terms of generally accepted rules for constructing phrases in a particular language. However, the task of understanding the text by the machine is recognition of the grammatical structure of a sentence, which allows a formalized presentation of the text meaning. The syntactic structure can act either as an intermediate result, which is an input for further semantic analysis, or as a convenient representation of natural language text for solving applied problems, for example, in information-analytical systems or machine translation systems.

Despite all the difficulties listed, the technology of natural language processing in most cases is able to successfully handle its tasks, thus it can be applicable in many industries.

A natural language, though structured and systemized appears quite problematic for symbolic algorithms aimed at its processing, therefore, the dominant approaches to the modern NLP are approaches based on statistical machine learning [9, 49]. In about half of homonymy cases, the set of morphological features is insufficient to define syntactic classes of units. It is though possible to reduce the ambiguity by using syntactic and semantic analysis via statistical techniques, which allow rejecting extremely unlikely variants. Natural language, although it is symbolic in its nature, to process it with the help of symbolic, based on logic, rules and objective models is a rather complicated process.

In early 90s, machine-learning methods began to evolve, and parallel to it, a number of studies on statistical linguistics were conducted. In machine learning, the classification algorithms for various tasks proved effective, namely for processing texts: spotting spam, sorting documents by subject, highlighting of named entities. The use of statistical methods in computer linguistics made it possible to determine parts of the language with high preciseness. There appeared parsers based on stochastic context-free grammars, projects on statistical machine translation were created. Fundamentals of in-depth learning have also been laid, which due to progress in high-performance systems and the emergence of large volumes of data used for learning, only recently produced first results [3].

In 2010, a model of lexical probabilistic (stochastic) grammar was suggested, which enabled the increase of grammatical parsing accuracy up to 93%, which, of course, is far from ideal. The parsing precision is the percentage of correctly defined grammatical ties, as well as the likelihood (which is usually very low) that the long sentence will be properly analyzed. At the same time, due to new algorithms and approaches, including deep learning, the speed of grammatical parsing has increased. Moreover, all the leading algorithms and models have become available to a wider range of researchers, and perhaps the most famous work in the field of deep learning for NLP has become the algorithm by Thomas Mikolov [12].

After the appearance of new deep learning methods, it became possible to obtain clear semantic descriptions for words, phrases and sentences, even without the present surrounding of the units. Creation of own semantic dictionaries and databases now requires less effort, so it is easier to develop automatic text processing systems. However, NLP is still far from adequate analysis of interrelated events presented in the form of a sequence of sentences or images, as well as dialogues. All known methods currently work successfully either in solving problems of "surface" understanding of language, or with substantial limitation of the subject area [3].

The deep learning methods are more precise than surface methods that do not attempt to "understand"

the text, but in fact, only very limited subject areas possess required databases for their processing, and therefore, at present, surface methods are often used. Such methods take into account the closest words, using analogous information, by studying the valency of words. The rules can be automatically obtained with a computer by using a text-based learning database of words added with their lexical semantics. In theory, this method is not as effective as deep methods, although in practice it provides better results [7].

Conclusions

The process of understanding and generating natural language with the use of computer technology is extremely difficult. Thus currently the most effective methods of working with language data are machine learning algorithm methods with a "teacher"-operator helping the system distinguish language structures and rules from the annotated corpus data. For example, the task of categorizing documents by categories: sports, politics, economics, and entertainment seems quite simple, because the words used in documents of such subject areas serve a hint. Based on their own experience, a human reader can easily refer the text to a certain topic, but it is unlikely that they can name the specific rules, used for that purpose. Creating a rule or a set of rules for automatic text categorization is complex and laborious. Using machine-learning algorithms with the "teacher" inputting this information, can let the machine determine the language structures that will allow categorizing the documents. This approach might prove effective for limited areas, like sport, law or economics. However, for wider areas like history, politics, sociology etc. this method will prove ineffective due to time-consuming character of its nature.

Perspectives

Thus the need for effective syntactic analysis seems obvious. The analysis of classical as well as contemporary syntactic analysis patters is the perspective for further research.

REFERENCES

1. Boiarskii, K. K. (2013). Vvedenie v kompiuternuiu linhvistiku [Introduction to Computational Linguistics]. Uchebnoie posobie, SPb: NIU ITMO, 72 p.
2. Buniiatova, I. R. (2003). Evoliutziia hipotaksysu v hermanskykh movakh (IV–XIII st.) [Evolution of Hypo-taxis in Germanic Languages (4–13c.)]. Monohrafiia, K.: Vyd. tzentr KNLU, 327 p.
3. Velikhov, P. Mashinnoe obuchenie dlia ponimaniia yestestvennogo yazyka [Machine Learning for Understanding Natural Language Processing].
https://www.osp.ru/os/2016/01/13048649/
4. Hyryn, O. V. (2017). Avtomatychnyi syntaksychnyi analiz anhliiskoi movy: zastosuvannia ta perspektyvy [Automatic Syntactic Analysis of an English Sentence: Application and Perspectives]. Zhytomyr: Vyd-vo ZhDU im. I.Franka*, Visnyk Zhytomyrskoho derzhavnoho universytetu imeni Ivana Franka*, Vyp. 1 (85), pp. 26–30.
5. Polkhovska, M. V. (2013). Analiz anhliiskykh medialnykh konstruktsii z pozytsii heneratyvnoi hramatyky [A Generative Perpective to the Analysis of English Medial Constructions]. *Studia Philologica*, Vyp. 2, 32–36.
6. Polkhovska, M. V. (2012). Kryterii rozriznennia medialnykh ta erhatyvnykh konstruktsii v anhliiskii movi [Defining Criteria of Medial and Ergative Constructions in English]. Naukovi zapysky Natsionalnoho universytetu «Ostrozka akademia»]. Ser.: *Filolohichna*, Vyp. 26, pp. 277–280.
7. Chen, P. A (2009). Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 28–36.
8. Fleiss, J. L. (2013). Statistical Methods for Rates and Proportions. John Wiley & Sons, 800 p.
9. Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers. 309 p.
10. Hollingsworth, Ch. (2012). Using Dependency-based Annotations for Authorship Identification. Berlin, *Proceedings of Text, Speech and Dialogue,* 15th International Conference, V. 7499, pp. 314–319.
11. Kovar, V. (2011). Information Extraction for Czech based on Syntactic Analysis. *Proceedings of the 5th Language & Technology Conference*, Poznan: Funcacja Universytetu im. A. Mickiewicza, pp. 466–470.
12. Mikolov, T. (2013). Efficient Estimation of Word Representations in Vector Space. 12 p.
13. Mohd, S. H. (2013). Word Sense Ambiguity: A Survey. *International Journal of Computer and Information Technology,* Vol. 02, Issue 06, pp. 1161–1168.
14. Reese R. M. (2015). Natural Language Processing with Java. Packt Publishing, 262 p.