УДК 811.111' 367.7 : 659.3

О.В. Гирин
Житомирський державний університет
імені Івана Франка
канд. філол. наук, доцент

CHALLENGES IN AUTOMATIC SYNTACTIC ANALYSIS OF AN ENGLISH SENTENCE

Abstract. *The article deals with the automatic syntactic analysis (parsing) of an English sentence. The article describes 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. The article provides tips of how automatic syntactic analysis can improve the solution of a particular task within the analyzed application spheres. The analysis identifies a number of linguistic issues that will contribute to the development of an improved model of automatic syntactic analysis: lexical and grammatical synonymy and homonymy, hypo- and hyperonymy, lexical and semantic fields, anaphora resolution, ellipsis, inversion etc.*

Key words: parsing; natural language processing; stylometry; formal logics; *automatic syntactic analysis model.*

Introduction. While the use of digital technologies is ever becoming a more integral part of our lives, 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.

Daily we use applications, which presuppose NLP, perhaps without noticing it: we write messages using a T9 dictionary, we hear synthesized speech when

using public transport (train stations, metro, etc), use a search engine for instant access to information, check spelling in a word processor, use automatic translation services for foreign languages. The performance of all these programs and applications is based on a certain level of understanding of how the language works.

On the other hand, we would like to use programs for which the developers still lack knowledge: we cannot, for example, communicate with the ticket selling machine/program about the shortest route to the place, the address of which we do not remember, just recollecting the shops and institutions close to it; we are unable to derive an automatic logical conclusion, based on a processed amount of text; automatic translation still contains serious errors in spelling, style, vocabulary, phraseology, and especially grammar.

Theoretical Background. NLP has been studied in numerous works in foreign linguistics since 1967. However, with the development of the technological possibilities, the approaches to NLP are constantly changing and improving. The issues, related to automatic speech analysis has been reflected in the works of the following scholars: Fleiss J. L. [7], Hollingsworth Ch. [9], Kovar V. [11] etc. In Ukrainian linguistics, however, the Ukrainian language is the main survey object (Darchuk N. P. [2], Cheilytko N.H. [6]). Yet, the experience and theoretical results in the field of English grammar, in particular from the generative perspective (Buniyatova I. G. [1], Polkhovska M. V. [3; 4], Snisarenko I. E. [5]), can frame a basis to the applied use thereof.

The **aim** of this article is to present the status analysis results in the field of NLP as well as to define specific problems in it and to suggest possible improvement directions.

Methods. This analysis 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 single out the

main application areas of automatic syntactic analysis today and to define problematic issues, which have not been sufficiently solved yet.

Syntactic analysis of a natural languages (parsing) is often called the cornerstone of NLP, a necessary basis for in-depth analysis and understanding of any natural language. Syntactic analysis of sentences is aimed at identifying the sentence structure, meaningful language units, and the relationship between them. However, parsing is often replaced with statistical [7] or even stochastic methods in the modern “intelligent” applications. There is even an opinion that parsing is not necessary for applications.

Results and Discussion. We have reviewed challenges in the field of NLP by the spheres of its use, and defined the tasks that have not been sufficiently solved (although there are partial solutions in most cases), as well as defined the ways for automatic syntactic analysis to improve the solution of a particular task within the specified application areas.

Information search involves the search of documents/sites, web pages related to the submitted query. In particular, Google (and other search engines) solve this task by a combination of indicators to assess the relevance of a document to a particular query based on the query key words (their derivatives and synonyms) within, in particular, the PageRank algorithm [12] and the like.

There are still situations, however, where a more complex syntactic query processing would help obtain more accurate results, for example, in the case of queries like "*science fiction novel about Asimov*" preposition *about* is ignored, and consequently the search results will contain links to the

Another improvement direction for the purposes of successful information search is question processing, e.g. "*Who revealed Snowden*?". By typing in such a query we would like to find the information where Snowden is semantic object of exposing. But the search results will provide a number of links to the documents, where Snowden is the semantic subject of exposing. Of course, the search engines have an advanced search option, in particular the use of quotes in the query, but in this case, the search results will not provide a link to the information that we need.

Question answering aims at receiving an answer on the basis of a knowledge base (including the Internet). In some cases, this direction is viewed at like information search, with the only difference that the query is a question in a natural language, rather than the query containing the keywords. Thus, most of the current question answering systems function on this principle – they "pull out" the key words from the input questions, and then use a search engine and just provide the user with the information found by the search engine.

Question answering in our interpretation is a more complex process. The answer must contain exactly the information requested in the question. The answer should be as concise as possible, without any loss of semantics and without redundant information. For example, a certain knowledge base contains a sequence of the following sentences:

(1) *Mary is a student. She is 20. She majors in English.*

The correct answer to the question "*What does Mary study?*" would be

(1') *Mary studies English.*

The sentence sequence in example (1) contains redundant information, which is not the matter of the request (*She is 20*). The sentence, containing the name *Mary*, does not contain the answer to the question; and the sentence with *English* contains neither the name nor the verb *study*. Receiving answer (1') needs the replacement of anaphora *She* with its antecedent *Mary*, as well as substitution of *majors* with the synonym from the question — *does study > studies*. Search engines in similar cases would offer either the whole sequence of sentences (which would contain excessive information) or the found documents/pages would contain keywords from the query, however, this does not mean that they would contain the answer to the question, as shown in the example of Snowden or examples of similar queries in the form of questions such as "*Who supports Bill Gates*?", where the search engine would provide links to articles, documents, etc. describing who B. Gates supports, but not supported by.

The development of a parsing algorithm aimed at receiving accurate, laconic answers to the questions-is a promising direction of our activities in the field of

NLP. Obviously, certain knowledge base pre-processing is necessary: determining semantic elements of sentences and marking them accordingly (e.g. grammatical and semantic subject and object, modifier of time, place, etc.). Complete solution to the range of problems encountered while developing the question answering algorithm, of course, involves consideration of a large number of additional linguistic phenomena, including l*exical and grammatical synonymy and homonymy, hypo- and hyperonymy, lexical and semantic fields, anaphora resolution, ellipsis, inversion etc*.

Logical judgement in the context of computer science and logic usually involves the creation of new formulas in particular those of formal logic, based on certain assumptions. This task means generation of a number of new sentences in natural language according to other sentences in natural language, called knowledge base. An example can be represented as follows:

(2) knowledge base: *Joseph Conrad (1857 – 1924) is a British writer born near Berdychiv.*

possible generated sentences:

(2') *Joseph Conrad is no longer alive;*

(2") *Joseph Conrad wasn't born in England;*

(2"') *Joseph Conrad is famous in literature,* etc.

In addition, such use of automatic syntactic (and lexical) analysis enables validity of a specific formula, that is, whether this formula is in the set of valid formulas created on the basis of the input information. For example:

(3) *civil war > unrest;*

(3') *unrest in Ukraine ≻ civil war in Ukraine*

Such formula analysis can be used to refute so-called "fake" news and messages.

There are currently several experimental approaches to the creation of a system of logical judgment, for example: automatic translation of natural language sentences into the language of predicate logic is succeeded by the generation of inferences on the basis of the predicate logic and consequent translation of the

results back to the natural language [8, 117]; use manual or stochastic transformation rules straight within the natural language without the intermediate use of formal logic [13, 126].

All these approaches require flexible automatic parsing. However, this task still remains unresolved, and there exist many promising directions for research in this respect.

Authorship verification is aimed at reliable automatic authorship detection for an anonymous piece of text (for example, in forensic linguistics). Namely, automatic verification of both vocabulary and especially syntax can conclude with high probability whether an anonymous piece of text and a known text have been written by the same author. Whereas a similar task, *text authenticity verification enables* defining whether a piece of text was written by a specific author.

Obviously, these two tasks simultaneously display similar and opposite features. They are similar, because they serve one purpose — authorship attribution. They are opposite, because authorship verification presupposes a certain number of texts (minimum 2 texts, the maximum is determined by the scope of available author's works), and high matching percentage is indicative of the authorship. T*ext authenticity verification, on the contrary*, usually involves one text, which is mapped to an unlimited number of texts available in the databases (including the Internet), and high matching percentage suggests authorship violation.

Lexical analysis, which is currently used for t*ext authenticity verification* has low reliability results, because it does not rely on lexical and especially grammatical synonymy, such as: active / passive voice Moreover, mere lexical analysis does not take into account the possibility of using translated material while writing the text.

Besides, a so-called stylometric text analysis has branched out from these tasks. In this regard, instead of defining a particular author of the text, its certain features are analyzed, focusing on different characteristics, such as age, education level or gender of the author. In most cases, the same methods can be used for all

tasks of this type. A number of studies [10] have been performed in this area to single out stylomes in a text – a combination of features, characterizing an author (or the selected features category). Such features are generally linguistically motivated, e.g.: frequency of short words, the use of specific words or parts of speech, grammatical structures, and the like.

Grammar check is one of the most important tasks of the NLP. Check of spelling and of simple grammatical phenomena have become habitual in word processors. More complex linguistic phenomena, however, still represent a problem for the automatic analysis. Grammar check, which is carried out through the usage of software packages such as Microsoft Office, can check a limited number of grammatical errors, for example: coordination and governing. But the packages are far from being able to find all the errors.

In natural languages mistakes in subject-predicate agreement obviously belong to the most serious ones. Automatic grammar check fails to spot mistakes in the sentences like:

(4) **The list of items are on the desk.*

In addition, each language has its "own" most common mistakes. Automatic grammar check cannot currently resolve a significant number of such errors. The complexity is intensified by the fact that a large number of grammatical rules include not only morpho-syntactical but also semantic and pragmatic aspects, which render their formalization for automatic check problematic. Thus an automatic grammar check will not find errors in a well-known example (sentence (5)):

(5) *One morning I shot an elephant in my pajamas.*

Since the amount of grammatical, lexical, semantic, pragmatic aspects to be taken into account are numerous, automatic grammar check does not currently involve full parsing; instead it uses the approach that involves consideration of the most common mistakes or a lightweight modification of full syntactic formalisms.

Natural language synthesis. Though synthesized speech is already in use in the modern world, its scope of application is limited by its reproductivity. The

speech generator at railway stations, subways, in various software reproduces the language units that have been previously entered into the program. In this case we cannot be talking about speech generation proper, but only about a playback. A promising direction in NLP is the development an algorithm for relatively independent speech generation. Such generation will anyway remain relatively independent, because it will presuppose input at some stage, though the input text may be further processed and transformed. This task includes some other ones, mentioned above, in particular: information search, question answering, logical judgment.

Associated tasks. In addition to the mentioned parsing application purposesf, there are various tasks associated with language processing, which are normally hidden from users, but they are necessary for NLP and can be used for solving tasks in several application directions.

One of them is defining a morphological class of a word – selecting the correct morphological marker from the set of all possible tokens for individual words (for example, to decide whether "*point*" is a verb or a noun). Many currently existing programs, depend on this task. Other major tasks which enable successful syntactic analysis of a natural language are as follows: anaphor resolution (finding antecedents), restoring elliptical constructions. Another important functional task, which can be treated as a separate task, is the recognition of ungrammatical speech.

Conclusions. Thus, the scope of NLP is wide, however, it reveals obvious directions for the improvement of parsing models. The improvement will consequently expand the scope and improve the results in areas that already employ automatic parsing. Indispensable achievements in vocabulary and morphology processing shall not be neglected while improving automatic syntactic analysis mechanisms for natural languages.

Perspectives. The the following set of linguistic questions will help to develop more sophisticated parsing models: lexical and grammatical synonymy and homonymy, hypo- and hyperonymy, lexical and semantic fields, anaphora

resolution, ellipsis, inversion etc. The following list is not exhaustive and may be supplemented.

REFERENCES

1. Буніятова І. Р. Еволюція гіпотаксису в германських мовах (IV – XIII ст.) : монографія / І.Р. Буніятова. – К. : Вид. центр КНЛУ, 2003. – 327 с.
2. Дарчук Н. П. Автоматичний синтаксичний аналіз текстів корпусу української мови / Н. П. Дарчук //Українське мовознавство. – КНУ ім. Т. Шевченка, 2013. – № 43. – С. 11-19.
3. Полховська М. В. Аналіз англійських медіальних конструкцій з позиції генеративної граматики / М. В. Полховська // Studia philologica. – 2013. – Вип. 2. – С. 32-36.
4. Полховська М. В. Критерії розрізнення медіальних та ергативних конструкцій в англійській мові / М. В. Полховська // Наукові записки [Національного університету "Острозька академія"]. Сер. : Філологічна. – 2012. – Вип. 26.– С. 277-280.
5. Снісаренко І.Є. Позиційні характеристики інфінітивного ад'юнкта мети (на матеріали пам'яток середньоанглійського періоду) / І.Є. Снісаренко // Вісник Харківського національного університету ім. В. Н. Каразіна. – Харків : Видавництво ХНУ ім. В.Н. Каразіна. – N636. – 2004. – С.185-187.
6. Чейлитко Н. Г. Корпусне дослідження зон зв'язків словоформ в українському реченні / Н. Г. Чейлитко // Лінгвістичні студії : [зб. наук. праць]. – Донецьк, Вид-во ДонНУ, 2009. – Вип. 18. – С. 268–275.
7. Fleiss J. L. Statistical methods for rates and proportions / J. L. Fleiss, B. Levin, Ch. P. Myunghee. – John Wiley & Sons, 2013. – 800 p.
8. Hobbs J. R. Discourse and inference: Magnum opus in progress / J. R. Hobbs. – Marina del Rey, 2014 – 168p.
9. Hollingsworth Ch. Using dependency-based annotations for authorship identification / Ch. Hollingsworth // Proceedings of Text, Speech and Dialogue, 15th International Conference. – Berlin. – v. 7499 . – 2012. – P 314–319.
10. Koppel M. Computational methods in authorship attribution / M. Koppel, J. Schler, Sh. Argamon // Journal of the American Society for Information Science and Technology. – No 60(1). – 2009. – P. 9–26.
11. Kovar V. Information extraction for Czech based on syntactic analysis / V. Baisa, V. Kovar // Proceedings of the 5th Language & Technology Conference. - Poznan : Funcacja Universytetu im. A. Mickiewicza, 2011. – P. 466–470.
12. Page L. The PageRank citation ranking: Bringing order to the web. Technical Report // L. Page, S. Brin, R. Motwani, T. Winograd 1999-66, Stanford : InfoLab, 1999. – 128 p.
13. Stern A. The BIUTEE research platform for transformation-based textual entailment recognition / A. Stern, I. Dagan // Linguistic Issues in Language Technology. – No 9. – 2013. – P. 120–146.