Zhytomyr Ivan Franko State University Journal. Philological Sciences. Vol. 1 (99)

Вісник Житомирського державного університету імені Івана Франка. Філологічні науки. Вип. 1 (99)

ISSN (Print): 2663-7642
ISSN (Online): 2707-4463

УДК 811.111' 367.7 : 659.3
DOI 10.35433/philology.1(99).2023.75-82

DETERMINING THE MORPHOLOGICAL CLASS OF A WORD DURING THE AUTOMATIC NATURAL LANGUAGE PROCESSING

O. V. Hyryn*

The article considers a mandatory component of the linguistic provision of any system of automatic natural language processing, i.e. automatic morphological analysis, the tasks of which include: determining for each text unit its place in the morphological system of the corresponding language; identification of word forms of the lexeme. As a result of automatic morphological analysis, each word form of the text is assigned a tag for the part of speech and the meaning of the grammatical categories (gender, number, case, aspect, tense, person, etc.). The nature of this information, its volume, and the methods used to establish morphological information depend on the purpose of the research, within which automatic analysis is carried out with the focus on the nature of the analyzed texts. Morphological analysis is present at all stages of text analysis, because neither morphemic, nor syntactic, nor semantic analysis can be performed without parts-of-speech tagging. With automatic syntactic analysis, only if lexical-grammatical and grammatical information is available for each word form, it is possible to syntactically bind word forms in a sentence. Morphological features of text units further become a tool for researching the relationship between vocabulary and grammar and the use in speech; between paradigmatics (in the aspect of consideration of case forms of declinable words) and syntagmatics (in the aspect of linear relationships of words, text coherence).

The article examines the difficulties that prevent unifying the process of tagging text units, namely lexical-grammatical homonymy, ambiguity of grammatical forms, polysemy. The study considers approaches to resolving morphological ambiguity based on the context analysis of the ambiguous word, which can be divided into statistical and rule-based. Rules can be compiled manually or derived from marked-up corpora. Statistical methods are based on quantitative indicators in large labelled corpora. Morphological ambiguity resolution methods are usually applied after primary tagging, which is usually done using dictionaries.

The article also provides a morphemic analysis algorithm for automatic morphological analysis.

Keywords: *automatic syntactic analysis, morphological analysis, tagging, lemmatization, stemming, parsing.*

* Candidate of Philological Science (PhD), Associate Professor
(Zhytomyr Ivan Franko State University)
oleg_hyryn@ukr.net
ORCID: 0000-0002-3641-2440

ВИЗНАЧЕННЯ МОРФОЛОГІЧНОГО КЛАСУ СЛОВА ПІД ЧАС АВТОМАТИЧНОЇ ОБРОБКИ ПРИРОДНОЇ МОВИ

Гирин О. В.

У статті розглянуто обов'язкову складову лінгвістичного забезпечення будь-якої системи автоматичної обробки природної мови – автоматичний морфологічний аналіз, до завдань якого входять: визначення для кожної одиниці тексту місця вморфологічній системі відповідної мови, ідентифікація словоформ однієї лексеми.

Унаслідок автоматичного морфологічного аналізу кожній словоформі тексту приписується код частини мови та значення граматичних категорій (рід, число, відмінок, вид, час, особа тощо). Характер цієї інформації, обсяг її й методи, за допомогою яких установлюють морфологічну інформацію, залежать від мети дослідження, у межах якого здійснюється автоматичний аналіз, від орієнтації на характер аналізованих текстів. Морфологічний аналіз наявний на всіх етапах аналізу тексту, тому що ані морфемний, ані синтаксичний, ані семантичний аналізи не можуть обійтися без визначення частин мови. У процесі автоматичного синтаксичного аналізу лише за наявності лексико-граматичної та граматичної інформації до кожної словоформи можна синтаксично прив'язати словоформи в реченні.

Морфологічні ознаки одиниць тексту далі стають інструментом дослідження зв'язку між лексикою та граматикою із їх використанням у мовленні, між парадигматикою (в аспекті розгляду відмінкових форм відмінюваних слів) і синтагматикою (в аспекті лінійних зв'язків слів, сполучуваності в тексті). У статті розглянуто труднощі, що заважають уніфікувати процес тегування одиниць тексту, а саме лексико-граматична омонімія, неоднозначність граматичних форм, полісемія.

У статті проаналізовано підходи до вирішення морфологічної неоднозначності, засновані на аналізі контексту неоднозначного слова, які поділяють на статистичні й такі, що визначаються правилами (rulebased). Правила можуть складатися вручну або виводитися з розмічених корпусів. Статистичні методи ґрунтуються на кількісних показниках у великих розмічених корпусах. Методи вирішення морфологічної неоднозначності застосовуються зазвичай після первинного тегування, що виконується, як правило, за допомогою словників.

У статті також наведено алгоритм морфемного аналізу для здійснення автоматичного морфологічного аналізу.

Ключові слова: *автоматичний синтаксичний аналіз, морфологічний аналіз, тегування, лематизація, стемінг, парсинг.*

Defining the problem. Natural language processing (NLP) has become an indispensable part of daily life and currently represents itself as an important tool. It helps people in many scientific and public areas performing a range of tasks, such as information retrieval, machine translation, speech synthesis and recognition, etc.

An important stage and what makes the automatic text processing complex is automatic syntactic analysis, which is impossible without morphological analysis. The latter itself is far from being challenge-free. The fact is that today there is no such language part classification that could satisfy researchers, and the final criteria for the distribution of words by parts of speech have not yet been established, the number of parts of speech is not a constant unit. Therefore the difficulty of classifying words by parts of speech in the process of automatic syntactic analysis determines the relevance of the work.

The object of the research is the process of automatic syntactic analysis, and the study scope is methods determining the allocation of a text unit in a morphological class of a word in the process of automatic syntactic analysis.

The aim of the paper is to study the peculiarities of the part-of-speech tagging in the process of automatic syntactic analysis.

Methods. This research suggests some linguistic issues, which should be considered while tracing morphological ambiguity, as well as the usage of the scientific methods of analysis, synthesis, description, as well as linguistic methods of semantic analysis and substitution in

order to illustrate the challenges natural language processing is currently facing.

Analysis of previous research. Research of N. Chomsky (syntactic structures) [3; 4], J. Backus (syntax of formal languages), J. Nivre (dependency grammar) [10], D. Yurafsky, J. Martin (NLP) [5; 6; 7] has been the basis and a contribution to the development of automatic syntactic analysis of natural language texts.

Currently, a simpler variation of NLP, namely automatic word processing is used to solve a variety of tasks, many of which every person deals with on a daily basis. They include spell checking and autocorrection, spam filtering, and automatic translation of small text fragments. There are also more complex tasks: finding relevant answers to queries, full-fledged machine translation of large texts, anti-plagiarism systems, analysis of the text style, determining the subject of a text, composing an annotation to the document, automatic abstracting and simplification of the text, construction of recommendation systems that would work with large arrays of unstructured data etc.

Various NLP models and effective algorithms for presenting natural language text arrays have been created to automatize the stages of analysis/synthesis of natural language texts. Traditionally, linguistic analysis of arrays of natural language texts is presented as a sequence of processes of morphological, syntactic and semantic analysis/synthesis. Appropriate models, methods and algorithms have been created for each process: focused on specific groups of languages (analysis of lexical morphology), namely system grammars of Holliday, grammars of Noam Chomsky [3; 4], extended networks of transitions (sentence syntax); classical semantic networks and Minsky frame models (text semantics). Significant contributions to the development of automatic syntactic analysis of natural language texts were also made by John Backus (syntax of formal languages), Joachim Nivre (dependency grammar) [10], Daniel Yurafsky and James H. Martin (NLP) [5; 6; 7].

Natural language text arrays of data in oral and written form are the main means of presenting and storing information. Therefore, the effectiveness of using IT depends to a large extent on solving the problem inherent of NLP automatic systems, the ultimate goal of which is to recognize their content. In NLP, two levels are distinguished depending on the depth and complexity of the processing process: formal (transformation of text fragments without referring to the analysis of its semantics) and content (semantic recognition of individual elements and logical-semantic relations between them to present the semantics of the message).

Results and Discussion. The first level (formal processing) is the basis of all existing IT in the operating NLP systems, and the second level is a field for theoretical and experimental research of automatic semantic analysis. A mandatory component of the linguistic support of any NLP system is automatic morphological analysis (AMA), or parsing, the tasks of which include: to determine the place for every information unit of the text in the morphological system of the corresponding language; to identify word forms of a lexeme. Morphological analysis is present at all stages of text analysis, because neither morphemic, nor syntactic, nor semantic analysis can do without the definition of parts of speech. With automatic syntactic analysis, only if lexical-grammatical and grammatical information is available to each word form, it is possible to syntactically bind word forms in a sentence.

As a result of AMA, codes of parts of speech and meanings of grammatical categories (gender, number, case, aspect, tense, person, etc.) are assigned to each word form of the text. The nature of this information, its scope, and the methods used to establish morphological information depend on the purpose of the research within which the AMA is carried out, on the orientation to the nature of the analyzed texts. At the level

of formal text analysis, morphological information provides computer access to content mediated through the correlation of content units with expression units. Morphological features of text units should become a tool for researching the relationship between vocabulary and grammar (i.e. lexical and grammatical semantics), between its use in speech, between paradigmatics (considering grammatical forms of declinable words) and syntagmatics (linear relationships of words, combinability in the text). The function of just such a link between language levels is possible only after allocation of a text information unit into a certain morphological class, or a part of speech.

As a matter of fact, AMA is present in all types of text analysis, since none of them can go without the analysis of word forms, determination of whether a word belongs to a grammatical class. The linguistic explanation for this is the objectively existing close connection between the lexical and grammatical meanings of language units, as well as between the systems of paradigmatic and syntagmatic relations. Morphological information provides the analyzing computer program with access to the content of the text, since until now the only real way of automatic analysis of the lexical semantics remains the indirect way through its correlation with the formal (morphological) units.

Thus it can be stated that the grammatical (morphological) meaning should be understood as the meaning abstracted as a result of the mandatory distinction of at least two identical, constantly repeated signs of a large number of specific words with their inherent lexical meanings.

Thanks to computer analysis, which performs various complex mathematical calculations, there are currently a large number of speech processing systems aimed at at various language levels. The quality of such processes depends on the research models created for this or that language, since the program needs to be trained precisely on formal structured data.

But before the parser can analyze a sentence, it needs to be given information about each word in the sentence. For example, to parse the sentence *The manager is responsible for the project*, the parser has to "know" that *the manager* is a singular noun in the Nominative case, the verb is in the third person singular form, *for the project* is a singular noun in the accusative case (according to the extended case theory), etc. Such information can be obtained from a digital dictionary that just lists all word forms with their part-language affiliation and inflectional categorial information, such as number and tense.

Moreover, words can consist of a root (which carries the main dictionary meaning) and one or more affixes that contain grammatical information. For example: *Cars drive slowly* – *car*+N+PL *drive*+V+PresIndef *slow*+Adj+Adv Affix.

English, like many other languages, has a complex and productive derivational morphology. For example, such derived forms as *accept, receipt, succept, decept*, etc., come from the root *cept*, which does not occur as an unbound morpheme. It is close to impossible to list in the dictionary all derived forms (including new terms or stylistically coloured made-up words or nonce-words) that may occur in a natural text.

Morphological parsing is the problem of extracting the deep lexical form from the surface form (for speech processing it includes the definition of word boundaries.) We must also consider: analytical forms (for example, *will write*, *has gone*), systematic rules (for example exceptions in plural forms, suppletivity etc.). Morphological parsing cannot be looked upon without considering tagging.

Part-of-speech (POS) tagging is the process of assigning a part of speech or other syntactic class marker to each word in the corpus. Since tags can also be often applied to punctuation, part of speech tagging requires punctuation to be separated from words. Therefore, tokenization is performed before or as part of the process of adding tags, separating commas, quotation marks,

etc. from words and removing ambiguity at the end of a sentence. This task is quite complex due ti the factors that we have mentioned. Especially problematic tagging can be due to lexical-grammatic homonyms, which are present in English in abundance. Therefore, part of speech tagging has to take into account the context of the word. For a human mind a context is associated with the meaning both of an isolated sentence or a whole text. There is also "common sense" which hints us in which meaning the word was used. Whereas a computer program can only so far compare the semantic fields, in which a few words to the right and to the left from the can be used. Thus in a well known sentence *The old man the boat,* apart from performing the syntactic analysis, looking into the semantic fields can be of help as it will find matches: *man* – one of possible verbal meanings is to operate a machine or a transport (whereas a noun wouldn't be specifically related to the navigation area); *boat* – a means of transport.

The importance of parts of speech tagging for language processing lies in the large amount of information it provides about the word and its environment. For example, these tag sets distinguish between possessive pronouns (*my, your, his, her, its*) and personal pronouns (*I, you, he, me*). Knowing whether a word is a possessive pronoun or a personal pronoun tells us what words are likely to occur next to it – possessive pronouns are likely to be followed by a noun, personal pronouns are likely to be followed by a verb. This can be useful in a language model for speech recognition.

For direct classification, machine learning methods are used, that is, a special training sample is created, where the input data includes information about the word: whether it is written in capital letters, whether there are hyphens or numbers, whether this word is included in the list of punctuation symbols, etc. The example of the latter can be *slash*. So when one says *students slash scientists,* they can mean the way this is spelled (students-scientists), but not what the students do to the scientists. All these parameters are coded into vector numerical representations, which later go into the machine learning algorithm.

In language processing, each word in a sentence is marked with its own part of speech tag. These tags then become useful for higher-level applications. The process of identifying parts of speech is more complex than just creating a list of words and their parts of speech, because some words can represent several parts of speech at different times, and some parts of speech are complex or unpronounceable. This is not uncommon – in natural language (unlike many artificial ones), a large proportion of word forms are ambiguous.

Although there is general consensus on the basic categories, many borderline cases make it difficult to choose a single "correct" set of tags. For example, it is difficult to say whether *stone* is an adjective or a noun in a well-known *stone wall.*

A second important example is the use of a word as an example, disregarding the part of speech to which it may be attributed, as in the following sentence, where "*five*" can be replaced by a word from any part of speech:

The word "five" has 4 letters.

In English, there are also many cases when parts of speech and "words" do not have an unambiguous correspondence, for example: *as far as, by himself, David's, gonna, don't, vice versa, first-cut, cannot, pre- and post-secondary.*

Many corpora treat *be, have,* and *do* as independent words (as in the Brown Corpus) [2], while some treat them all as usual verbs (e.g. Penn Treebank [11]). Because these particular words have more forms than other English verbs and occur in very different grammatical contexts, treating them just as "verbs" means that the tag will have much less information.

Part-of-speech tags originated from a linguistic approach, but later moved to a statistical approach. Modern models achieve accuracy of more than 97%. Research on part-of-speech markers

conducted with English text corpora has been adapted to many other languages.

A part-of-speech classifier contains a phrase or a sentence and assigns a corresponding tag to each word. In practice, the input text is often preprocessed. One of the common tasks before processing is to tokenize the text so that the classifier can see the sequence of words and punctuation. Other tasks such as removal of irrelevant elements, punctuation removal, and lemmatization can be performed before the tagging.

A set of predefined tags is called a tag set. This is important information about what classes the part-of-speech classifier will classify into. Example tags are NNS for a plural noun, VBD for a past tense verb, or JJ for an adjective. A set of tags can also contain punctuation.

Instead of developing different tag sets for every analyzer, a common practice is to use known tag sets: 87tags from the Brown corpus, 45tags from the Penn Treebank, 61tags from the C5set, or 146tags from the C7set.

If a part-of-speech tagging shows low accuracy, it negatively affects other tasks that follow. To improve accuracy, some researchers have suggested combining POS tagging with other processing. For example, joint POS tagging and dependency resolution is an approach to improve accuracy compared to independent modelling. Sometimes the word itself can give useful clues. For example, *the* is a definiteness marker. The prefix *un* implies an adjective (yes, it can also imply a verb, but in verbal morphology this prefix is non-productive), for example, *unathomable*. The suffix *ly* suggests an adverb, for example, *importantly*. Capitalization can suggest a proper noun. Graphic form is also useful, for example, several hyphens in a word, for example, *35-year-old*, suggest an adjective.

A word can be tagged based on neighbouring words and the possible tags those words may have. Word probabilities also play an important role in choosing the correct disambiguation tag. For example, the above mentioned *man* is rarely used as a verb and is mostly used as a noun.

With the statistical approach, we can count the frequency of word tags in a labelled corpus and then assign the most likely tag. This is called a unigram tag [8]. Bigram tagging is a further approach, where the tag frequency specified by a certain previous tag is taken into account. So the tag depends on the previous tags.

So overall we can distinguish between the following types of algorithms for POS tagging:

◦ lexical method - assigns a POS tag that occurs most often with a word in the training corpus.

◦ rule-based: a dictionary is built with possible tags for each word. The rules regulate the process of tagging and are manual and learned. An example rule might say: "If an ambiguous / unknown word X is preceded by a determiner and followed by a noun, then it is an adjective";

◦ statistical: the text corpus is used to obtain useful probabilities, given a sequence of words, the most probable sequence of tags is selected. They are also called stochastic or probabilistic labels ;

◦ memory-based: a set of cases is stored in memory, each case containing a word, its context, and a corresponding tag. The new sentence is tagged based on the best match to the instances stored in the memory. It is a combination of a rule-based and stochastic methods;

◦ transformation-based: rules are automatically induced from the data. Thus, it is a combination of rule-based and stochastic methods. Tagging is done using broad rules and then refined or transformed using more specific rules;

◦ neural networks: the most recent one type which presupposes deep learning techniques: RNN and bidirectional LSTM are two examples of neural network architecture for POS tagging.

Part-of-speech taggers can be both supervised and unsupervised. Supervised tags rely on the corpus tag to generate a dictionary, rules, or tag sequence probabilities. They work best

when taught and applied in the same genre of text. Unsupervised tagging induce groups of words. This saves the effort of pre-marking the corpus, but clusters of words are often coarse.

A combination of both approaches is also common. For example, rules are automatically induced from an unlabelled corpus. The output from this is corrected by a human and resubmitted to the tagger. The tagger reviews the patch and adjusts the rules. Many iterations of this process may be necessary.

In this article we are making an attempt of offering our own morphological analysis pattern, which is based on the combination of the existing approaches, but with a slight incline onto the morphemic analysis

We propose a parsing pattern which is based on the analysis of the affixes of word forms. It will include an algorithm for a text in which each word usage is assigned a code of a grammatical class (part of speech) and a grammatical subclass (gender, number, case, person, tense).

The algorithm accounts for several partial issues, which are solved in the following sequence:

◦ detection of affix and its separation from the base of the word;

◦ detection of a root morpheme and/or its allomorphs, if any;

◦ combining all forms of the word into one group (paradigm);

◦ selection or reconstruction of the dictionary form of the word.

Since each grammatical class of words that has a word change has its own peculiarities of form formation and combining forms into paradigms, it is advisable to provide the number of blocks in the algorithm that corresponds to the number of grammatical classes characterized by word change. According to the accepted list of grammatical classes of words, the following blocks are distinguished:

◦ compilation of paradigms of nominatives (nouns, gerunds, personal pronouns);

◦ compilation of paradigms of attributive classes (adjectives, participles, ordinal numerals, possessive and demonstrative pronouns);

◦ compilation of verb paradigms;

◦ compilation of adverbial paradigms (abbreviated forms of adjectives and adverbs);

◦ compilation of pronominal paradigms;

◦ compilation of quantitative numeral paradigms;

◦ list of variable forms of prepositions.

◦ list of variable forms of conjunctions.

◦ list of variable forms of exclamations and particles.

So, in order to build a morphological analyzer, we will need the following:

1. dictionary: a list of bases and affixes, together with basic information about them (whether this base is a noun or a verb, etc.);

2. morphotactics: a model of the order of morphemes in a word that explains which classes of morphemes can follow other classes within a word. For example, there is a rule that a plurality morpheme follows a noun rather than precedes it;

3. orthographic rules: this will include spelling rules, that are used to indicate changes that occur in a word, usually when combining two morphemes (for example, *y* > *ie*, the spelling rule which explains why *city* + *-s* > *cities*, not *citys*).

Conclusion. A mandatory component of the linguistic support of any NLP system is automatic morphological analysis, the tasks of which include: determining the place of textual information units in the morphological system of the corresponding language; identification of word forms of one lexeme.

As a result of the work of AMA, codes of parts of speech and meanings of grammatical categories (gender, number, case, aspect, tense, person, etc.) are assigned to each word form of the text. The nature of this information, its scope, and the methods used to establish morphological information depend on the purpose of the research within which the AMA is carried out. It has to take into

consideration the nature of the analyzed texts. Morphological analysis is present at all stages of text analysis, because neither morpheme, nor syntactic, nor semantic analysis can do without the definition of parts of speech. With automatic syntactic analysis, only if lexical-grammatical and grammatical information is available for each word form, it is possible to syntactically bind word forms in a sentence. At the level of formal text analysis, morphological information provides computer access to content, derived from the correlation of content units with formal units. Morphological features of text units should become a tool for researching the relationship between vocabulary and grammar, between its use in speech, between paradigmatics (consideration of case forms of declinable words) and syntagmatics (linear relationships of words, combinability in the text), which altogether constitutes a vast field for **further scientific research**.

REFERENCES

1. British National Corsup URL: https://www.english-corpora.org/ (reference date: 25.01.2023)
2. Brown corpus: Corpus of American English. URL: https://www.sketchengine.eu/brown-corpus/ (reference date: 05.02.2023)
3. Chomsky, N. Formal properties of grammars. Wiley: Handbook of Mathematical Psychology, 1963. 2, Ch. 12. P. 323–418. [in English].
4. Chomsky, N., Miller, G.A. Introduction to the formal analysis of natural languages. Wiley: Handbook of Mathemati-Mathematical Psychology, 1963. 2, Ch. 12. P. 269–322. [in English].
5. Jurafsky, D., Martin, J. H. Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing. Prentice Hall, 2006.
6. Jurafsky, D. From Languages to Information. Stanford, 2020. [in English]. URL: https://web.stanford.edu/class/cs124/lec/Information_Extraction_and_Named_Entity_Recognition.pdf. (reference date: 25.01.2023)
7. Jurafsky, D. Speech and Language Processing. Prentice Hall, 2008. 1044 p. [in English].
8. Kupiec, J. Robust part-of-speech tagging using a hidden markov model. Computer Speech & Language, 1992. Vol. 6, no. 3. P. 225–242. [in English].
9. Natural Language Processing (NLP) IBM Cloud Education, 2020. [in English]. URL: https://www.ibm.com/cloud/learn/natural-language-processing?mhsrc=ibmsearch_a&mhq=nlp (reference date: 25.01.2023)
10. Nivre, J., Hall, J., Nilsson, J. Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, 2007. 13:95. P. 135. [in English].
11. Penn treebank. URL: https://catalog.ldc.upenn.edu/LDC99T42 (reference date: 05.02.2023)

Стаття надійшла до редколегії: 05.02.2023
Схвалено до друку: 28.02.2023