The Relevance of Using Hidden Markov Chains in Solving the Problem of Removing Homonymy

Alona Zhernovnikova[0000-0001-6887-6234] and Zoia Kochueva[0000-0002-4300-3370]

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

alena.zhernovnikova1998@gmail.com, kochueva@kochuev.com

Abstract. The use of the method of hidden Markov chains for the removal of morphological homonymy is considered. The effectiveness of this method in relation to various languages is analyzed. The current state of work is described and the main results in this direction are presented, conclusions are drawn about the applicability of this resolution method with an assessment of its accuracy.

Keywords: homonymy, hidden Markov chains, removing homonymy, morphological polysemy.

The ambiguity of linguistic forms is one of the natural features of a natural language, contributing to the qualitative development of the vocabulary, thereby "saving" verbal material. The resolution of polysemy (the so-called disambiguation) is one of the most important tasks of the automatic processing of a natural language. Resolution results are used to improve the accuracy of the methods of classification and clustering of texts, improve the quality of machine translation, information retrieval and other applications [1].

There are several types of polysemy of the natural language: morphological, syntactic and lexical-semantic polysemy.

The task of resolving morphological polysemy is to determine for a word the parts of speech and grammatical features that are appropriate for the context. Morphological polysemy is mainly represented by grammatical homonymy, i.e., the coincidence of words in separate grammatical forms.

For a language with poor morphology, the problem of solvability of morphological ambiguity, as a rule, reduces to resolving ambiguities at the level of parts of speech (POS-tagging), which, in turn, significantly simplifies the task. In agglutinative languages, such as Turkish, Hungarian and Tatar, morphemes are added to the number of words, which, in addition to semantics, also define syntactic relations. Morphological polysemy in these languages is manifested in various forms. In some cases, both syntactic and semantic analysis may be required to determine morphological ambiguity.

COLINS'2020, Volume II: Workshop. Lviv, Ukraine, April 23-24, 2020, ISSN 2523-4013
http://colins.in.ua, online

Homonymy problems are currently relevant. In automatic natural language processing, the corpora of morphologically marked texts are often used. The markup is done through pre-engineered morphological analyzers. The main problem of such systems is ambiguity. Because of this, there is a need to choose the only variant of analysis that is right for the given context [2].

There are the following methods for removing morphological homonymy: use of a neural network; rules-based methods; using the hidden Markov model algorithm [3].

To solve the problem of removing morphological homonymy, a well-known probabilistic approach based on the use of Hidden Markov Model (HMM) tagging was chosen.

The algorithm based on the use of the hidden Markov model (HMM) requires preliminary training of the system on the already marked out selection of texts of large volume. Preliminary experimental results showed an accuracy of the algorithm for the Russian language of at least 95% [4].

It is noted that in a comparative analysis of algorithms based on the hidden Markov model and the Markov model of maximum entropy, both algorithms do a good job (accuracy of at least 95%) with the task of frequent disambiguation, but they remove the homonymy with an expanded set of grammatical tags much worse. As a rule, algorithms make mistakes when marking proper names, pronouns, Roman numerals, initials and abbreviations.

To implement the algorithm, it is necessary to perform morphological markup in such a way as to maximize the function:

P (word | tag) * P (tag | previous n tags),

where P (tag | previous n tags) is the conditional probability (calculated by the tagged case) of the occurrence of a given tag tag, provided that the previous n tags are already defined.

P (word | tag) is a conditional probability (also calculated from the corpus) of the word word occurring in the given place, provided that the word has a given grammatical class tag [5].

The Hidden Markov model algorithm has a fairly high accuracy for English, namely 96%. There are difficulties in applying this model to the Ukrainian language, since the large-scale corpora is required, given the richness of Ukrainian word-formation and word-translation in comparison with English.

References:

1. Turdakov, D.Yu.: Metody i programmnyye sredstva razresheniya leksicheskoy mnogoznachnosti terminov na osnove setey dokumentov: avtoref. dis. … kand. tekh. nauk. 05.13.11. Moskva, 2010. 20 p.
2. Jurafsky, Martin: Speech and Language Processing. Upper SaddleRiver, NJ, USA: Prentice-Hall, Inc., 2009.
3. Kobritsov, B.P.: Methods for removing semantic ambiguity. NTI, Ser. 2, Vol. 3, 2004.

4. Lakomkin, E.D., Puzyrevskiy, I.V., Ryzhova, D.A.: Analiz statisticheskikh algorit-mov snyatiya morfologicheskoy omonimii v russkom yazyke. URL: http://aistconf.org/ stuff/aist2013/submissions/aist2013_submission_33.pdf, last accessed 2020/04/09.
5. Jurafsky, Daniel, James, H. Martin: Speech and Language Processing, 2000.