Dictionary-Based Part-of-Speech Tagging of Polish

Stanisław Galus
Wyższa Szkoła Finansów i Administracji, ul. Władysława IV 22, Sopot, Poland

Abstract. A set of 70 lexical symbols is defined which covers written Polish. Under the assumption of the second order Markov process, a dictionary-based method of tagging parts of speech is presented. Instead of having to be trained on earlier tagged text, parameters of the method are estimated on frequencies of unambiguously classifiable unigrams, bigrams and trigrams of lexical symbols found in untagged literary text. The method is being proved to tag correctly 88% of all words in text.

1 Introduction

Part-of-speech tagging is an act of assigning part of speech to each word in a text. It is one of the tasks of natural language processing.

There are two major applications of automatic tagging. It can be used in an early phase of parsing, that is stating the grammatical functions of all the words in a sentence. It can also be used as a research tool for searching texts for sequences of words of given parts of speech. Automatic tagging has numerous practical applications as well.

Models being used to formulate algorithms of tagging may be classified as probabilistic and non-probabilistic. The most extensively used probabilistic models, namely Markov models and hidden Markov models, are based on Markov processes. The most common non-probabilistic approach is transformation-based tagging.

To estimate parameters of a model, supervised and unsupervised estimation can be used. Supervised estimation requires the correct tagging of a large sample of text, while unsupervised learning does not. A kind of unsupervised estimation is based on a dictionary providing information on parts of speech admissible for each word.

Accuracy of tagging is easy to evaluate. A common measure of accuracy is the fraction of correctly tagged words. The percentage of correctly tagged ambiguous words is also used.

The most simple Markov models are \( n \)-gram models which assume that the \( n \)-th word depends only on the previous \( n - 1 \) words. We shall be proving that with probability 0.99, the accuracy of dictionary-based trigram part-of-speech tagging of Polish exceeds 88%.

\(^1\) See [4, chapter 10].
As far as we know, there has already been an attempt to build an automatic part-of-speech tagger of Polish. Namely, a trigram morphosyntactic tagger presented in [2], disambiguates a set of tags including inflexional categories. Trained on an annotated corpus, it achieves accuracy of 90.6%. Several attempts have also been made to build a lexical analyser, for example LEXAN, described in [10, p. 50-56].

2 Methods

Let $\Sigma$ be a finite alphabet. A language $L$ is a certain set of finite strings over the alphabet $\Sigma$. A text over the language $L$ is a finite sequence of strings from $L$. The set of all texts over the language $L$ will be denoted by $L^*$.

Let us assume that the language $L$ is a sum of $m > 1$ classes of lexical symbols $L_1, L_2, \ldots, L_m$. Given a text

$$a = (a_i)_{i=1}^n \in L^*, \ n \geq 1,$$  

we define its tagging as any sequence $t = (t_i)_{i=1}^n$ such that $t_i \in \{1, 2, \ldots, m\}$ and $a_i \in L_{t_i}$ for $i = 1, 2, \ldots, n$. We assume that among all taggings of the text $a$, there exists exactly one tagging $t$ which may be called a true tagging. In context of tagging, the numbers $1, 2, \ldots, m$ are called tags.

2.1 The alphabet and lexical symbols

Written Polish language uses letters, punctuation marks, a hyphen, digits and graphical signs. The Polish alphabet consists of twenty four letters of the Latin alphabet, two non-Latin letters and nine letters of the Latin alphabet with diacritical marks: A,Ą,B, C, Ć, D, E,Ę, F, G, H, I, J, K, L, Ł, M, N,Ń, O,Ó, P, Q, R, S,Ś, T, U, V, W, X, Y, Z, Ź, Ż, a,ą,b,ć,d,ę,e,ę,f,g,h,i,j,k,l,m,n,ń,o,ó,p,q,r,s,ś,t,u,v,w,x,y,z,ź,ż. Since numerous foreign proper names and borrowings are written without a transcription, the set of letters is often extended. There are ten punctuation marks: ., ;, ,, :, −, ..., ?, !, brackets and quotation marks. [5, p. 1656] names five kinds of brackets: ( ), [ ], / /, { }, < > and three kinds of quotation mark: ” ”, « », ’ ’. Typewriter quotation mark: "" may also sometimes occur [7, p. CXXII].

The hyphen: - is used for spelling compounds and dividing words at ends of lines. Ten digits are: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Graphic signs include basic mathematical symbols, section sign and computer symbols: =, +, %, §, #, $, &, *, @, \, ^, \_, |, −.

We distinguish six classes of lexical symbols: words, punctuation marks, hyphen, numbers, graphical signs and unknown characters. These classes may be further subdivided. The most detailed division used in this paper, possessing $m = 70$ classes, is presented in table 1.

\[^2\] For a more comprehensive description of morphological analysers for Polish see [3].