Prototype of Semantic Model of Polish for Automatic Speech Recognition

Abstract

Bartosz Ziółko
Department of Electronics
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059, Kraków, Poland
bziolko@agh.edu.pl

Introduction

A model of semantics of natural language can greatly improve efficiency of large vocabulary automatic speech recognition system. However, the task of preparing such model faces problems of data sparsity. The quality of language models depends strongly on the amount of text available during the training. This is the cause for a trade-off of quality and time spent on calculations. The high performance computers serve by obtaining the linguistic rules from the huge amount of texts written in Polish.

Computers used in the project

The prototype of the model works in Matlab, what unfortunately increase time necessary for all calculations because data cannot be processed in cache memory.

Three computers were used in this task. Two of them are Cyfronet machines, Mars and Zeus. The third one is wavelet2, a small server for calculations belonging to Signal Processing Group at AGH. Preparing the model is a process which takes months. Mars and Zeus were updated and switched off during this time. The usage of Matlab on this machines changed which caused it became more complicated to conduct our calculations there. This is why most of work was moved to wavelet2.

The idea of semantic model

Word statistics can be applied in order to provide probabilities of appearance of particular words in Polish. To achieve it, millions of sentences were analysed to search for possible contexts of words from our testing corpus, namely 639 words. In Polish language, which is non-positional and highly inflected, the information is carried by the words of the whole sentences, which can be reordered in many ways, with no crucial change in meaning. The assumption of bags-of-words model is that semantics of a sentence can be described as a sum of semantics of particular words from this sentence. For example, a meaning of expression *brown cow* can be extracted from separate, independent meanings of *brown* and *cow*. However, there are exceptions. A good one was given centuries ago by Voltaire: "*This body, which called itself and still calls itself the Holy Roman Empire, was neither holy, nor Roman, nor an Empire*".

Following this idea, we treated each sentence, where any of our test words appeared, as a pattern. These patterns were grouped using on-line k-means clustering. The patterns kept information of how many sentences were used to create this pattern, which gave weights for each pattern. Each time 1000 already existing patterns and new 1000 sentences from the corpora were extracted . These were combined into 1000 patterns updating weights and treating both sets in the same way. The task is to create a matrix of size 639 * 1000 describing how the tested words can be used. Such matrix is then further processed to use it more efficiently in automatic speech recognition. This processing is computationally quite fast, so it is out of the scope of the conference.

Calculations and data

Instead of multithreading, big corpora were split to smaller parts and processed separately, at the same time. Then results were joined together using k-means clustering. Joining is also a time consuming task. In total, calculations took several months, partly because of losing processes during breakdowns, partly because of a bottleneck, which was access time to large text files, and partly because of using Matlab. Several corpora were analysed including Rzeczpospolita, Wikipedia, literature books, Parliament and Solidarność meetings transcriptions.

This work was supported by MNISW grant number OR00001905.