PLLM - Language modeling using SOM network

Leszek Gajecki[1], Ryszard Tadeusiewicz[2]

[1]University of Information Technology and Management, ul. Sucharskiego 2, Rzeszów, Poland
[2]AGH University of Science and Technology, ul. Mickiewicza 30, Kraków, Poland
[1]lgajecki@wsiz.rzeszow.pl; [2]rtad@agh.edu.pl

Abstract

Our tool –PLLM (Language Model for Polish Language) consist two parts: – *learn* , which is application for learning of neural network, and *latitice-tool_GLM* application that perform recognition task –apply the language model into word lattice.

learn – is java application, command line tool for language model learning. Learning examples are taken from IPI PAN corpus of Polish language. Program performs SOM network (Self Organised Map) learning. Neurons are clusters that are modelling relation between POS (Part of Speech) of two words. Parameter max_wordspan (>=2)– defines maximal distance of two words position (Fig 1.)., these words are taken to learning relation (value 2 means two neighbor words). We can choose the number of first and last file of corpus – to establish training set (since corpus is divided into many files), number of iteration, and *max_wordspan*. This tool computes also the coverage of neural network clusters (how often each neuron is used to modeling relation). Coverage is needed to establish, which cluster will be positive rule (coverage higher than MinCov) , any other defines negative rule (Fig 2.).

latitice-tool_GLM is modification of SRILM *latitice-tool*. This command line tool perform searching task on word lattice (vitterbi algorithm), with application of our language model. The parameters are language model scale, acoustic model scale, *logzero* (score in case there is no known word) or *MinCov* –minimal coverage parameter. To work with several lattice files efficiently (and WER computation) bash scripts are provided : *recnn2.sh* which use our model and *rec2.sh* for Knesser – Ney models application.

Since SRILM is C++ application while learn and associate language model classes are written on Java, we used Java Native Inerface and Java Invocation API for class loading and further using them.

In case of both tools *learn* and *latitice-tool_GLM* option *–help* documents syntax.

Details about language modeling we provide in our paper "Language modeling using SOM network".

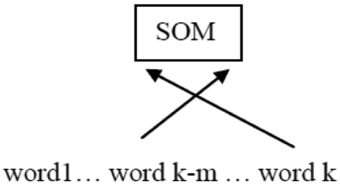

Fig 1.The idea of binary relations (max_worspan=m)

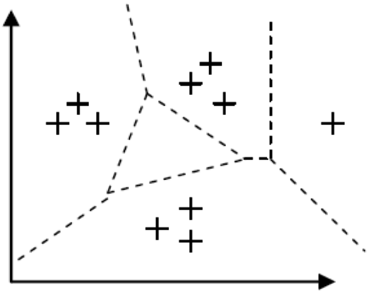

Fig 2. Clusters of positive rules (language rules) and negative clusters (no language rules)