# Open Source Proposal : Continuous Space Language Models

*Holger Schwenk, University of Le Mans*

Language models play an important role in machine translation. There are an essential part of statistical machine translation systems and also play an important role in syntax systems. Recently there has been a trend to improve language models by using ever larger amounts of training data, for instance several billions of words of English newspaper texts. However, such enormous amounts of data are not necessarily available for many European languages, and we need to deploy methods that take the best advantage of a limited quantity of training data.

The continuous space language model (CSLM) is a possible solution to this problem. The basic idea of this approach is to project the word indices onto a continuous space and to use a probability estimator operating on this space. Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown n-grams can be expected. A neural network can be used to simultaneously learn the projection of the words onto the continuous space and to estimate the n-gram probabilities. This is still a n-gram approach, but the LM probabilities are ”interpolated” for any possible context instead of backing-off to shorter contexts.

This approach has consistently achieved significant improvements when used is speech recognition or statistical machine translation system. The following table gives an overview of some recent results (BLEU score on test data):

| | NIST08 Arabic/English | WMT08 French/English | WMT08 English/French |
|---|---|---|---|
| baseline | 41.90 | 32.75 | 31.59 |
| with CSLM | 42.98 | 33.13 | 32.34 |

Discussions on conferences and meetings seem to indicate that there is quite some interest to use this approach, but to the best of my knowledge it was not yet “reprogrammed”.

I have developed a set of tools to train and use CSLM in a very efficient way. Unfortunately, this code currently uses some proprietary libraries preventing it from being released as open source. It is even not possible to distribute it in binary form. I propose to completely rewrite this software in C++ and to release it as open source to the research community. I will develop a new modular design (the current version is in a bad shape since it contains many incremental additions over several years of development). Hopefully, this will allow other people to extend the tools, for instance with a factored representation of the words. I will also provide interfaces to integrate the CSLM into existing tools, for instance the SRILM toolkit. It will be also possible to call it directly from a decoder like Moses, although this might be computationally too expensive (I’m currently using n-best list rescoring).

The CSLM toolkit will be available in about six months. I propose to release it before the next MT Marathon and to give a tutorial on its usage. Finally, this new tool will be the starting point of the development of continuous space translation models.