

From: "Qin Gao" <qing@cs.cmu.edu>  
Subject: [11-734] Topic selection - Large scale architecture  
Date: Wed, January 23, 2008 12:52 pm  
To: "Alon Lavie" <alavie+@cs.cmu.edu>,"Stephan Vogel" <stephan.vogel@cs.cmu.edu>

---

#### Large Scale Architecture for MT Systems

As a statistical approach, Statistical Machine Translation always ask for more data, therefore more space and computational resources. In order to make the MT system practical, a large scale system must be considered.

SMT system often consists with two important parts, namely training and decoding. Both stages require fast processing, and large amount of memory. The major problem can be listed as follows:

#### Training stage:

\*

As the amount of training corpora increase, the training (aligning) time also grows.

\*

The vocabulary size is also increasing, therefore more memory is required.

\*

In order to manipulate large language model, both the speed and storage need to be optimized

#### Decoding stage:

\*

The search space is spawned by several dimensions, for example, language model history, word position and possible translations of each word. Therefore the large the model is, the slower the decoding algorithm becomes.

\*

The storage of models is also a big issue, and if the model is far too large to fit into memory, it will greatly impact the speed of processing.

Therefore, there are several important difficulties to overcome:

1.

Parallel processing, both training and decoding, which may accelerate the processing speed.

2.

Distributed model storage and look-up. It will be a balance

between speed and size, when the model (language model or translation model) is too large to fit into a single machine's memory, it must be distributed amount multiple machines, how to have a fast access to the model, is a remaining problem.

There are some papers dealing with these issues, some are directly related to SMT, some are not related but is possible to be used.

#### 1. Google language model

Large Language Models in Machine Translation

<http://www.google.com/search?lr=&ie=UTF-8&oe=UTF-8&q=Large+Language+Models+in+Machi>

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, Jeffrey Dean  
<<http://research.google.com/people/jeff/>>, /Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)/, pp. 858-867.

2. Distributed Hash Tables ( There are many papers and mature libraries, and could be used as a possible way of storing models)

A. Rowstron and P. Druschel (Nov 2001). "Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems  
<<http://research.microsoft.com/%7Eantr/PAST/pastry.pdf>>". /IFIP/ACM International Conference on Distributed Systems Platforms (Middleware), Heidelberg, Germany/: 329-350.

Other related work may be general issues on parallelism and code optimization.

-----  
Preferred Date:

Mar 5,  
Feb 27.  
Apr 2

Thank you!

Qin

---