MTTK: An Alignment Toolkit for Statistical Machine Translation

Yonggang Deng [1] **William Byrne** [1,2]

[1] Center for Language and Speech Processing, Johns Hopkins University, Baltimore MD 21218, US

[2] Machine Intelligence Lab, Cambridge University Engineering Department, Cambridge CB2 1PZ, UK

Introducing MTTK

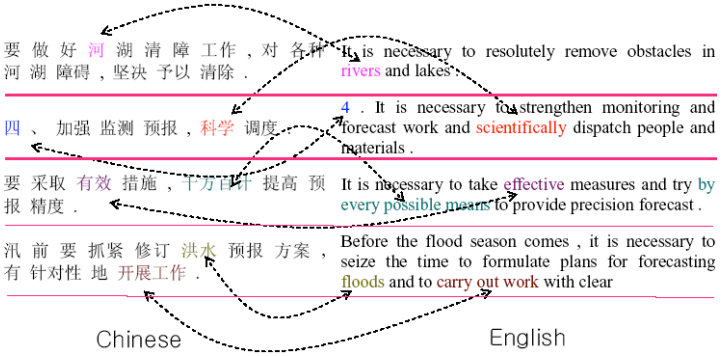
- Build statistical alignment models from parallel text
- Perform sentence/sub-sentence alignment, word and phrase alignment
- Train statistical word alignment models
- Extract word-to-word translation, phrase-to-phrase translation table
- Can be used for Machine Translation, Multi-lingual Text Processing and more ……

Feature Highlights

- Implement IBM Model-1,Model-2, Word-to-word HMM and **Word-to-Phrase HMM**
  - the entire modeling approach is an alternative to the GIZA++ pipeline
  - aligning words to phrases explicitly [Deng & Byrne HLT/EMNLP '05]
  - **good alignment quality** comparable to GIZA++
- Fast training by **parallelism**, memory efficient
- Being able to process **hundred of millions** of words and more
- Language independent, no linguistic knowledge required, models learn from data
- Has been used for
  - 2005 TC-STAR evaluation
  - 2004/2005 NIST MT Arabic-English and Chinese-English evaluation
- ……

Availability

- Free for public use under the Open Source Education Community License
- **Download:** http://mi.eng.cam.ac.uk/~wjb31/distrib/mttkv1/



Typical Usage