Edinburgh Statistical Machine Translation Group
Edinburger statistische Übersetzungsgruppe

Philipp Koehn (pkoehn@inf.ed.ac.uk) *Miles Osborne (miles@inf.ed.ac.uk)*
http://www.statmt.org

Moses bringing the Tablets to the People

Introduction

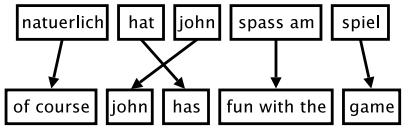
- The SMT Group consists of a dozen people (Faculty, Post-Docs, Graduate Students and Visitors)
- We work on frameworks for automatically translating any language to any other language.
- Current language pairs:
  - Arabic-English, Chinese-English.
  - All European languages to each other.

Sample Arabic-English Translations

The Palestinian Legislative Council ratified at the end of its outgoing one of the items to amend the Constitutional Court. Under the amendment would be appointed President of the Constitutional Court judges by the head of the Palestinian Authority without reference to the Legislative Council.Hamas has described this as a white paper on the Legislative Council. The meeting held by the Legislative Council came out of the celebration.

Statistical Machine Translation

- We use Phrase-based Statistical Machine Translation



An aligned sentence pair

- Expressed as Log-Linear Models, with parameters Λ set using Minimum Error Rate Training:

$$E^* = \text{argmax}_E \sum_i f_i(E,F)\lambda_i$$

- We decode using *Moses* (which is an open source replacement for the other decoder which we also support, Pharaoh).
- Our SMT systems are trained on parallel corpora:

Arabic-English	151 million words, 6.3 million sentence pairs
Chinese-English	224 million words, 10.0 million sentence pairs
Europarl	30 million words, 1 million sentence pairs

- For language modelling (LM), we currently use the SRI LM Toolkit.
- Our LMs are based on a billion word corpus (but we are scaling to a trillion word corpus).

Achievements

- Moses is widely used internationally by the research community.
- At the annual NIST international machine translation competitions, we have outperformed many other universities (eg Berkeley, CMU) and other companies (such as Microsoft and Systran).
- We achieved the best English-Spanish results at the TC-STAR Evaluation Campaign.
- We have a spin-off company (Linear-B: http://www.linearb.co.uk/).

Current Research Themes

- Factored Translation Models (FTMs):
  - FTMs grew out of a recent Johns Hopkins Workshop (2006).
  - Idea: allow multiple sources of information to be integrated into Phrase-based SMT.
  - An Open Source Decoder supporting FTMs is available
    http://www.statmt.org/moses/
  - Funded by DARPA (GALE) and the EU (EuroMatrix).
- Large-scale discriminative modelling:
  - SMT systems are largely generative models, with discriminatively trained scaling factors.
  - The next generation of translation systems will use millions of features and will be fully discriminative.
  - Research funded by EPSRC.
- Trillion-word Language Models (LMs):
  - Using larger LMs improves fluency.
  - We are working on deploying 5-gram (and upwards) LMs.
  - Novel randomised LMs (representing a fraction of the English portion of the Web).
  - LMs based on peer-to-peer techniques running over a cluster of machines.
- Better support for low-density languages:
  - SMT systems perform poorly with modest volumes of parallel data.
  - We have been investigating ways to automatically eliminate aspects of translation which cannot be well modelled using a given volume of data.
  - Adding syntax should improve subjective evaluation.
  - We have been extending Factored Translation Models with light-weight syntax.

Collaboration

- Universities: MIT, USC/ISI, Cambridge, LIMSI, Saarbruecken, Charles University Prague.
- Companies: Systran, Google, CELCT, Morphologic, Globalware.