Semi-supervised Learning to Machine Translation

Paper by: Nicola Ueffing, Gholamreza Haffari, Anoop Sarkar, 2008

Advanced MT Seminar 2010
Presenter: Vamshi Ambati

Context

- How do we deal with low parallel data scenario-
  - Get more data
    - Pay for more translations
    - Harvest online for parallel data (In domain vs Out-of domain)
    - Obtain Comparable training data
  - Try to do better with what you have
    - Re-define models (factored)
    - Seek annotations to build sharper models (annotate some word-alignments)

Current paper

- Goal:
  - Producing synthesized translations using models built from existing data
  - Self-training applied to MT
  - Focus on domain adaptation

Related Work

- Nicola Bertodi and Marcello Federico: Domain Adaptation for Statistical Machine Translation with Monolingual Resources (WMT 2009)
  - LM and TM adaptation by interpolation (UN corpus to Europarl)
- Holger Schwenk and Jean Senellart: Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training (MT Summit 2009)
  - Large scale adaptation
- Nicola Ueffing, Haffari, Sarkar: Transductive learning for statistical machine translation (2007)

Framework

- Repeat until "stopping criteria"
  - **Estimate** : compute a TM using data in current iteration
  - **Filter**: Sample a set of monolingual sentences that are relevant to the translation task
  - Decode set using MT system trained on to generate Nbest lists
  - **Score:** rate the translations to produce measures of confidence
  - **Select:** Choose a subset of good sentence pairs

Stopping Criteria

- Stopping criteria
  - Fixed set of iterations
  - Score on held out data set
- Effect:
  - Too many iterations introduces noise as can be seen by ‘select’ function later
  - Too few iterations may not obtain required benefit
- Held-out data-set: Does it not make it too specific and closer to Transductive learning?

Filter

- Select from among the monolingual data that is relevant to the development set
  - Assumes DEV and TEST are in-domain
- Average over n-gram coverage (n=1 to 6)

Estimate

- Re-estimation with new data is not done on entire data
- Models trained are combined independently and re-optimized on DEV
- PORTAGE
  - A typical ‘beam-search based’ PBSMT
  - Support for multiple LM
  - Rescoring of N-best lists

Score

- Length-normalized decoder likelihoods
- Confidence Estimation:
  - Word posterior probabilities computed by Levenshtein alignment between hyp and Nbest entries
  - Phrase posteriors (segmentation from SMT system)
  - Sentence posteriors
  - Language model scores
- Log-linear combination of all the above tuned to sentence ‘Classification Error Rate’

Select

- Importance Sampling:
  - Sample with replacement from a distribution of the translations for a sentence (Nbest list)
- Selection using a threshold
- Top K

Experiments

- Fr-En
  - Europarl - 688K (parallel data)
  - Hansards – 1130 K (monolingual data)
- Ch-En
  - NIST 2006 Evaluation corpus: 3.2M +5M (parallel)
  - Subset of Chinese Giga word : 50K (monolingual)

Results
Ch-En

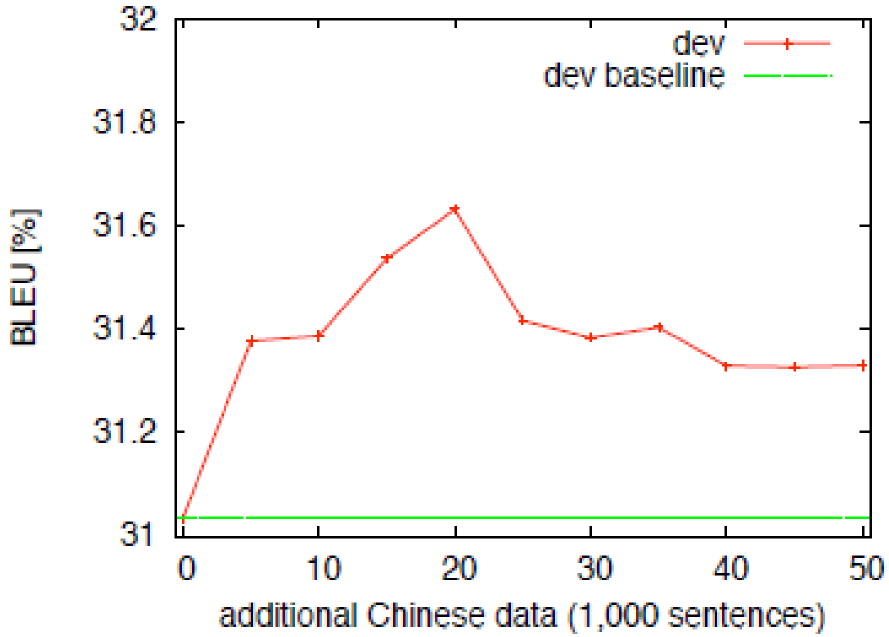
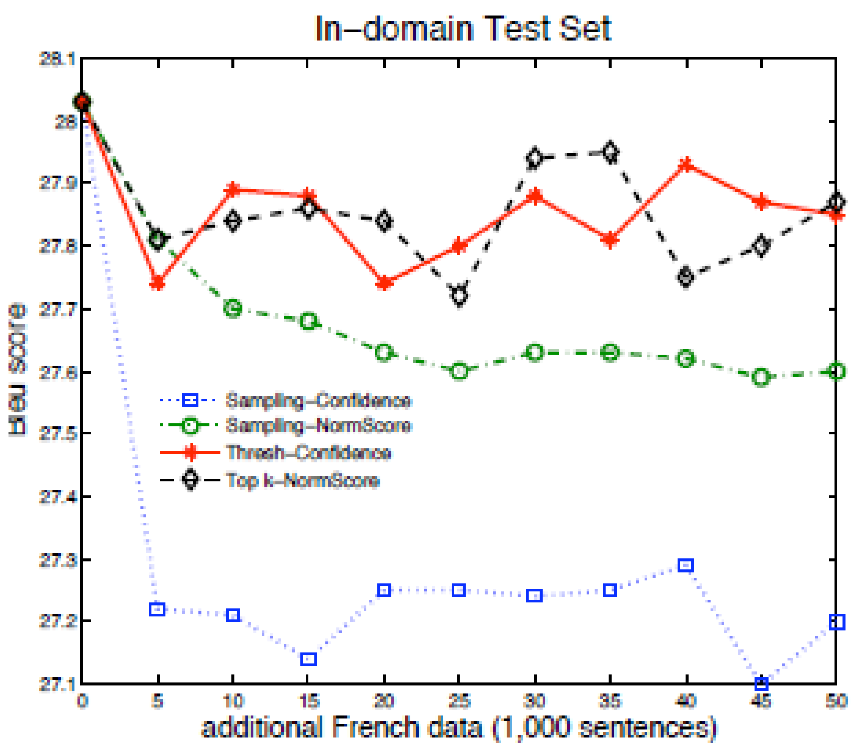
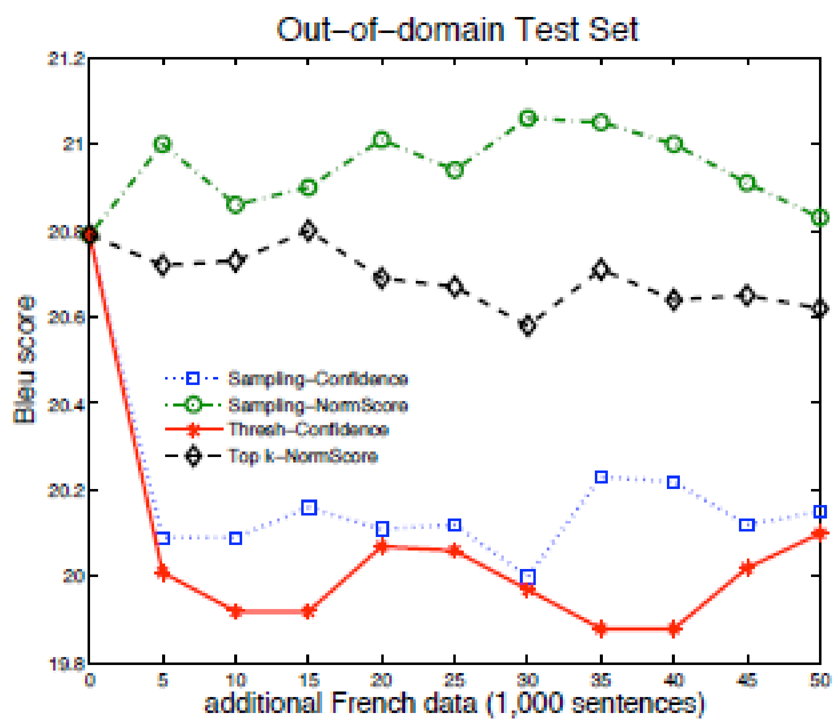


Figure 11.1 Translation quality using an additional phrase table trained on monolingual Chinese news data. Chinese–English development set.

Results
Fr-En





Point for Discussion

- Do Semi-supervised techniques work in NLP?
  - Success stories in MT or other areas of NLP
- Stopping criteria for Semi-supervised training