A Very Very Preliminary Investigation of the Effects of Segmentation on Chinese Phrased-Based Statistical Machine Translation Performance

William Morgan

CS 224N

Problem Statement

- Statistical machine translation is based on words.
  - Phrase-based MT, for example, takes word alignments and builds "phrases" out of them.
- Non-segmented languages such as Chinese must therefore be segmented.
- But there are many ways to segment Chinese.
  - No strong linguistic right or wrong.
- So how should we segment? The *best* way.
  - I.e., the way that maximizes translation quality.

Experimental Setup

- How to build a (fairly good!) MT system in the privacy of your own home:

  GIZA++: word alignment

  SRILM: language model

  Philipp Köhn's "training" package: phrase extraction, model tuning

  Pharaoh: decoding

- Corpora I used:

  LDC2004T08: Hong Kong parallel corpus. 700k sentence pairs of newswire text (plus other stuff I didn't use)

  LDC2006T04: NIST MT 2003 competition eval set. 919 sentences, four reference translations, *2003 state-of-the-art system output*

Experiments

- Four segmentation standards:

  LDC: CRF segmenter trained on LDC's Chinese Treebank.

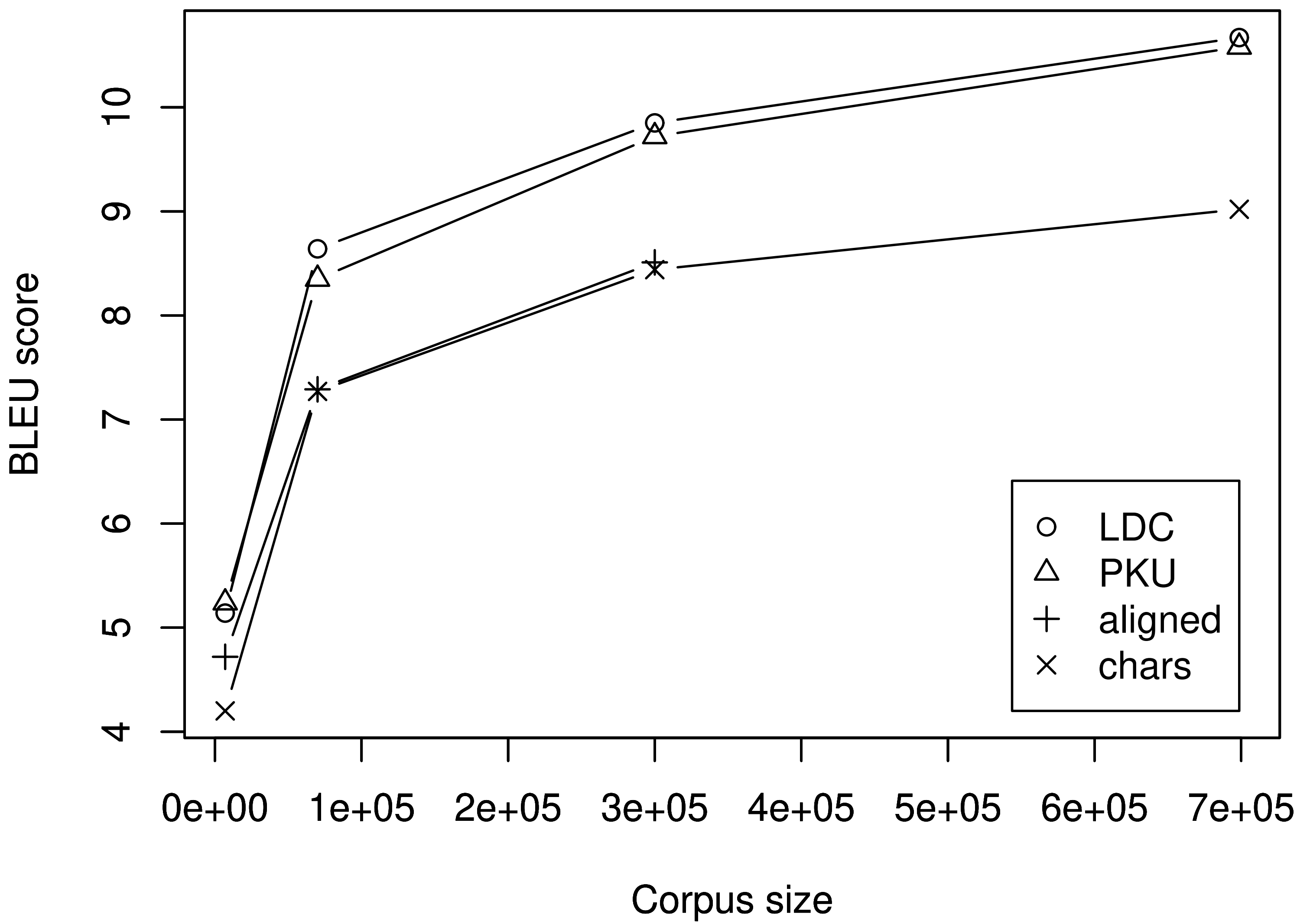
  PKU: CRF segmenter trained on PKU's SIGHAN 2005 data.

  chars: character segmentation (William's favorite).

  aligned: segmentation based on alignment of English words and Chinese characters. Call every sequence of consecutive characters aligned with an English word a segment. Split everything else.

- Process:
  1. Segment training and evaluation data.
  2. Train model. (No tuning!)
  3. Evaluate model with BLEU.

Results



BLEU score
Corpus size
4
5
6
7
8
9
10
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
LDC
PKU
aligned
chars

Conclusion

- Words, sadly, are useful for MT.
- Character results are low, I suspect, because phrase table entries become quite short, and (related) phrase table is smaller: 540k vs 660k entries for LDC segmentation. (Also hella long time to train—44 hours for full corpus!)
- LDC more "verbose" than PKU, in general—might be related to the difference between them?
- Aligned results are low, but I used LDC segmentation for evaluation data, which is definitely bad. Ideally I should train a new segmenter on training data.
- Spent way too much time fixing other people's C++ code.