Bilingually Motivated Domain-Adapted Word Segmentation for Statistical Machine Translation

Yanjun Ma Andy Way

National Centre for Language Technology
School of Computing
Dublin City University

EACL 2009, Athens

April 1st, 2009

Word Segmentation for SMT

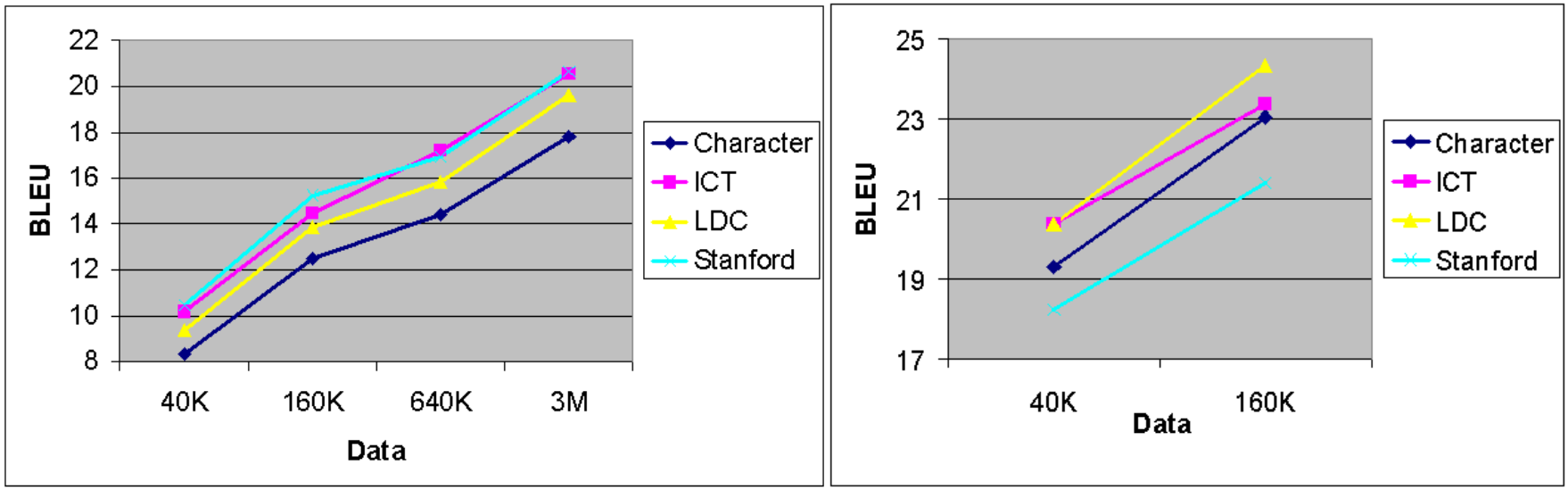

Figure: Word Segmentation for SMT: NIST data vs. IWSLT data

Related Work

Training [Xu et al., 2004, Ma et al., 2007, Xu et al., 2008]

- Use word alignment models to propose the word segmentation for a sentence in a bid to obtain better word alignment and better translation model
- $\Rightarrow$ Multiple segmentations for the same source sentence depending on the target sentence
- Use a dictionary-based approach to segment the test set and use the single best segmentation

Decoding [Xu et al., 2005, Dyer et al., 2008]

- Use n segmenters to segment the training data, train n MT systems and combine the translation models
- $\Rightarrow$ Still rely on monolingual segmenters
- Word lattice decoding

Our approach

Coherent training and decoding without using monolingual segmenters

Word units proposed by alignment

may	可 能	favorite	最 喜 欢
may	可 以	interesting	有 意 思
food	食 物	miami	迈 阿 密
food	食 品	last	最 后 一
july	七 月	block	个 街 区

Figure: Example of 1-to-n word alignments between English words and Chinese characters

Word lattice decoding

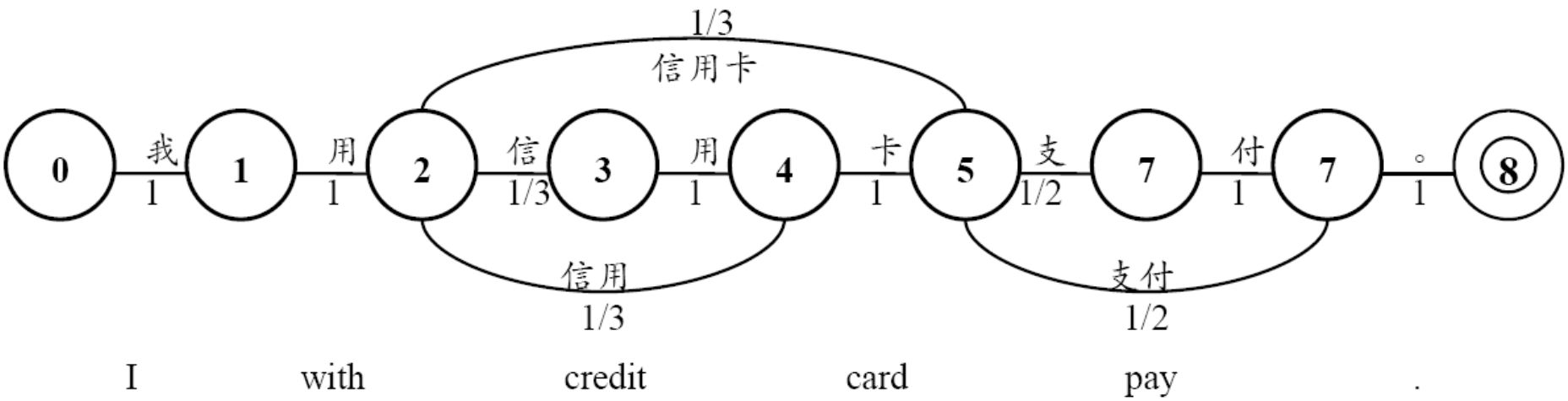

Figure: Example of a word lattice

Experimental Results

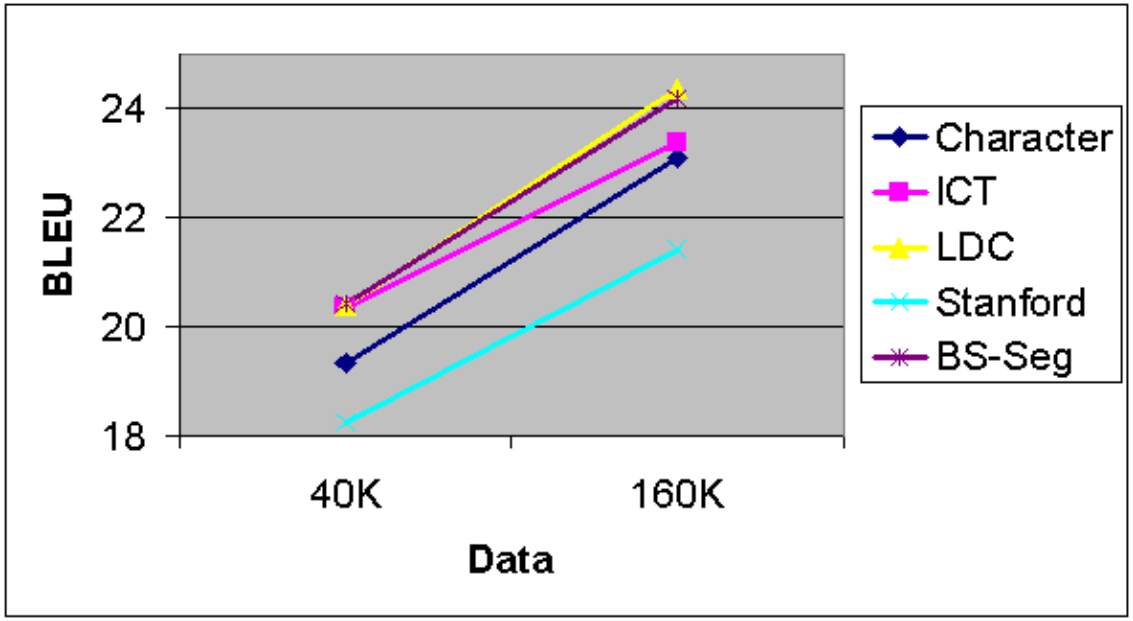

Figure: Bilingually Motivated Word Segmentation for SMT (IWSLT)

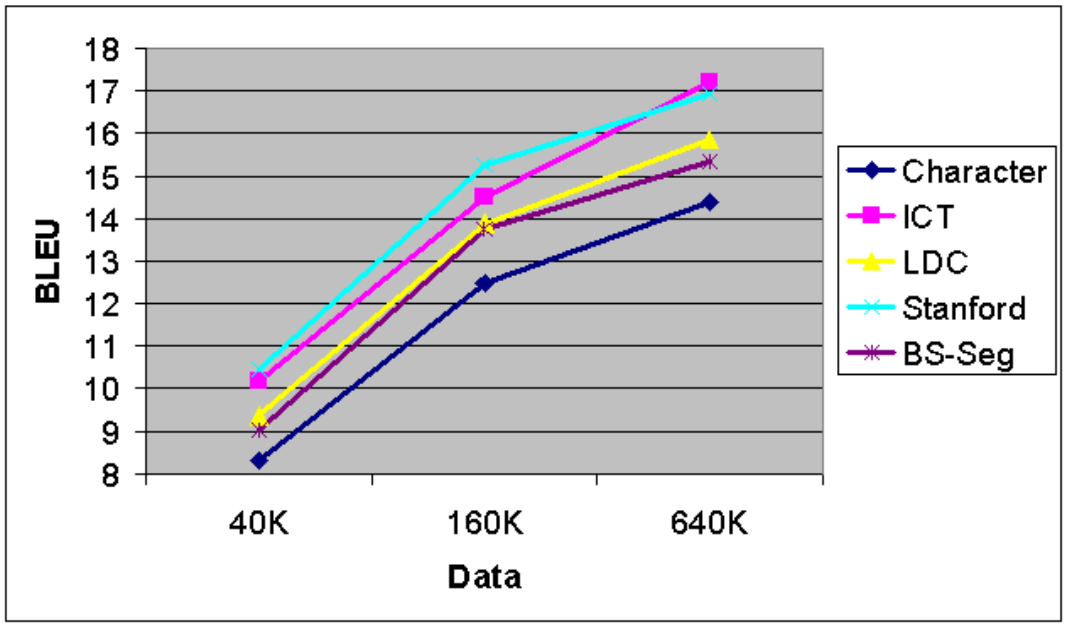

Figure: Bilingually Motivated Word Segmentation for SMT (NIST)

Summary

- Unsupervised word segmentation approach
- Bilingually motivated, i.e. different segmentation for different language pairs
- Domain-adapted, i.e. different segmentation for different data domains
- Competitive and consistent performance when used for Phrase-Based SMT

More information in the poster session!

References

Dyer, C., Muresan, S., and Resnik, P. (2008).
Generalizing word lattice translation.
In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1012–1020, Columbus, OH.

Ma, Y., Stroppa, N., and Way, A. (2007).
Bootstrapping word alignment via word packing.
In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 304–311, Prague, Czech Republic.

Xu, J., Gao, J., Toutanova, K., and Ney, H. (2008).
Bayesian semi-supervised chinese word segmentation for statistical machine translation.
In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1017–1024, Manchester, UK.

Xu, J., Matusov, E., Zens, R., and Ney, H. (2005).
Integrated Chinese word segmentation in statistical machine translation.
In *Proceedings of the International Workshop on Spoken Language Translation*, pages 141–147, Pittsburgh, PA.