

## Student Research Symposium 2003

---

Alicia Tribble

### Overlapping Phrases in a Statistical Machine Translation System

Competitive research systems for Statistical Machine Translation (SMT) all employ some method of phrase-level translation in addition to or as a replacement for the word-level models originally proposed by (Brown et al., 1993). While extraction methods for these phrases differ among systems (see (Vogel et al., 2003), (Marcu and Wong, 2002), (Zens et al., 2002) for examples), the systems themselves all must combine phrase translation candidates in a useful way during decoding in order to take full advantage of them.

This presentation treats the combination of partial translations in an SMT system currently used at CMU. Specifically, I address the inability of the decoder in this system to combine phrase translations that overlap. As an example, consider the following two translation pairs:

Source side	Target side
a b c	w x y
b c d	x y z

While the phrases a b c and b c d can be translated into w x y and x y z, respectively, the traditional system is unable to use this information to translate a b c d as w x y z.

The planned talk describes a series of experiments on allowing such overlapping phrase translations. I will present translation results on the Arabic-English development set used in the TIDES evaluations this Spring, along with an analysis of the number and quality of overlapping phrases that were generated. Interesting issues raised by this work include the effect of chaining longer and longer phrase rules together, the role of these long phrases in combination with reordering models which may contradict them, and the effect on decoder speed and the number of translation hypotheses generated for a single test sentence.

Overall this presentation represents an effort to help the SMT system move beyond memorization of the training data in its use of phrase-level translations.

## References

- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.
- Daniel Marcu and William Wong. 2002. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of EMNLP-02*, Philadelphia, July.
- Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU Statistical Translation System. To appear in *MT-Summit*, New Orleans, September.
- Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-Based Statistical Machine Translation. In *KI-2002: 25th Annual German Conference on AI*, Springer Verlag, September.