7th International Conference on Spoken Language Processing
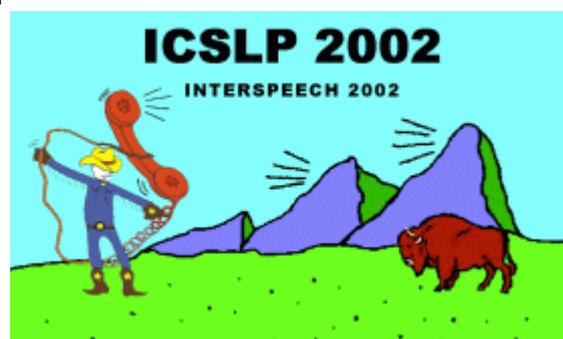
September 16-20, 2002 Denver, Colorado, USA

Bilingual Corpus Cleaning Focusing on Translation Literality

Kenji Imamura, Eiichiro Sumita

ATR Spoken Language Translation Research Laboratories, Japan

When we automatically acquire translation knowledge from a bilingual corpus, redundant rules are generated due to translation variety. To overcome this problem, we propose bilingual corpus cleaning based on translation literality. Word-level correspondence and phrase-level correspondence are applied as the criteria of literality. Using these criteria, a bilingual corpus was cleaned, and translation knowledge for a pattern-based MT system was acquired from the cleaned corpus. As a result, the translation quality of the MT was improved despite reductions in the the corpus size to about 81% and 87% by using word-level and phrase-level literality scores, respectively.

[Full Paper](#)

Bibliographic reference. Imamura, Kenji / Sumita, Eiichiro (2002): "Bilingual corpus cleaning focusing on translation literality", In *ICSLP-2002*, 1713-1716.