The POSTECH's statistical machine translation system

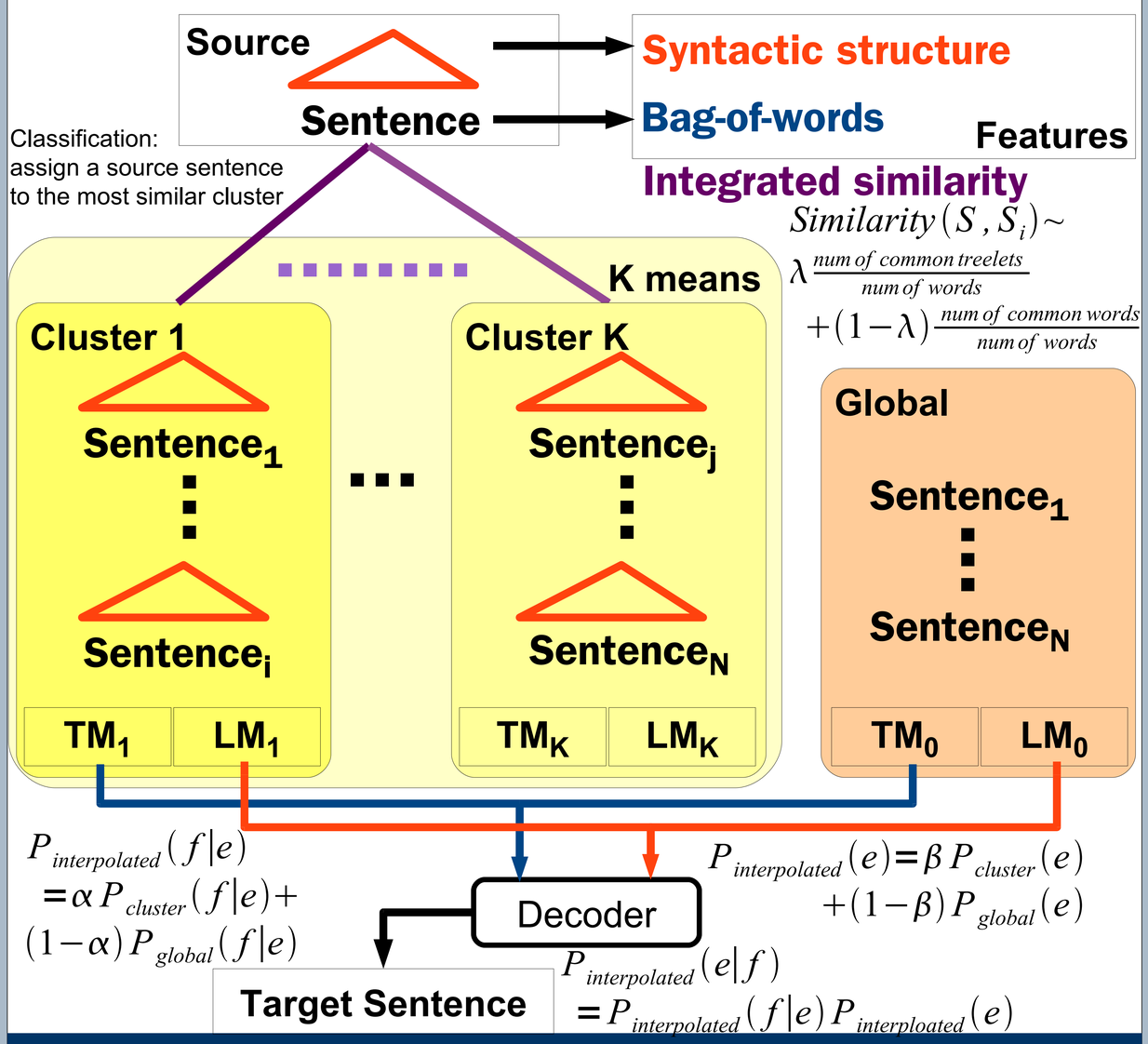
Hwidong Na, Hankyoung Kim, and Jong-Hyeok Lee
Pohang university of science and technology (POSTECH), Pohang, Korea

Introduction

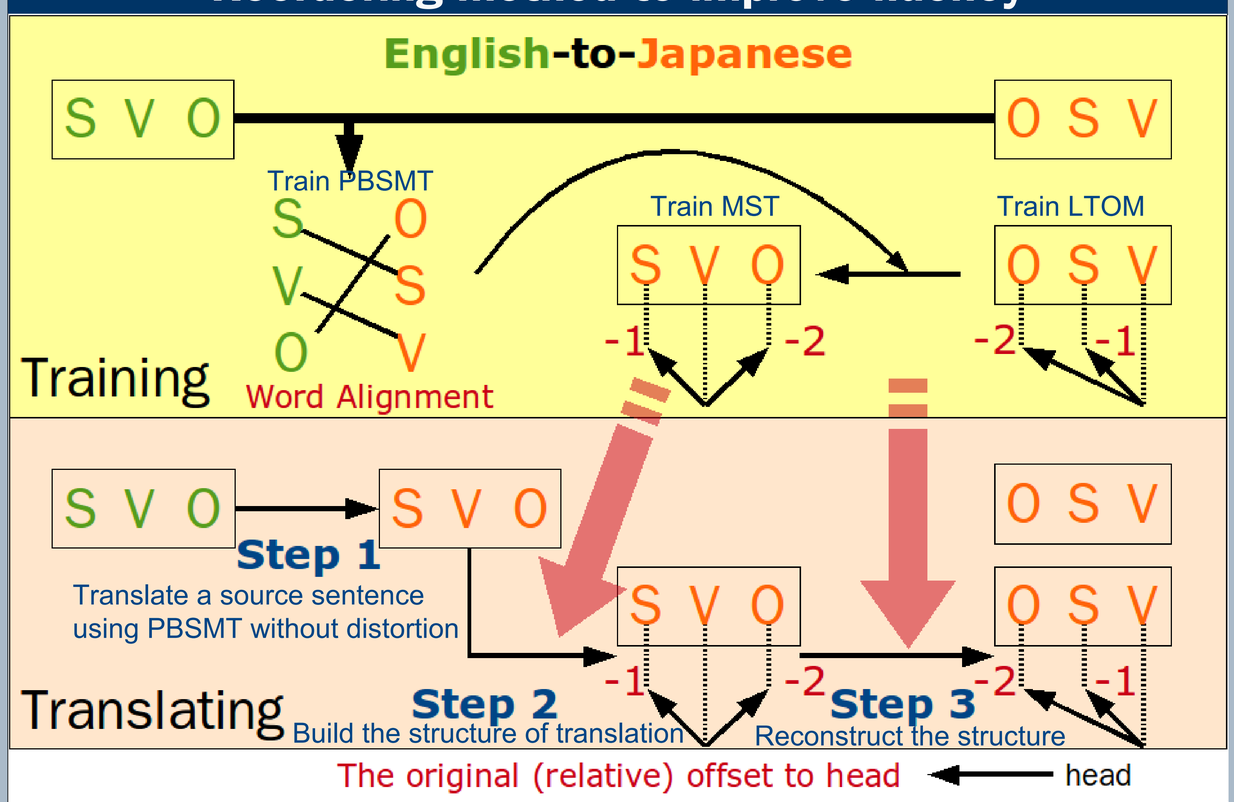
Domain specific model: SMT consists of **translation model**s and **language model**s. Translation models determine **correct** lexcial choice. Language models generate **natural** target sentences. **Specific model**s describe its domain better than the general model.

Improving fluency: Good translation requires **adquacy** and **fluency**. PBSMT tends to give correct lexical choice, but fail to order them naturally. A global reordering method that improves the fluency eventually enhance the overall quality of the translation.

Cluster-based SMT to model specific domain

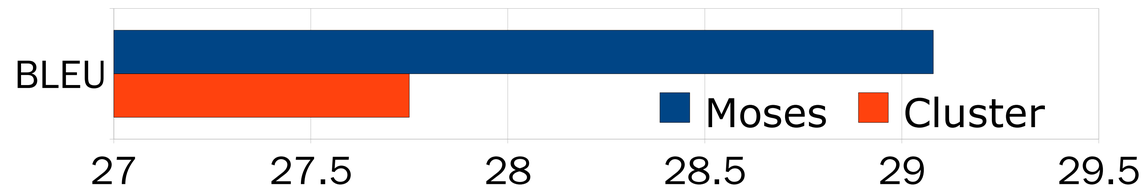


Reordering method to improve fluency

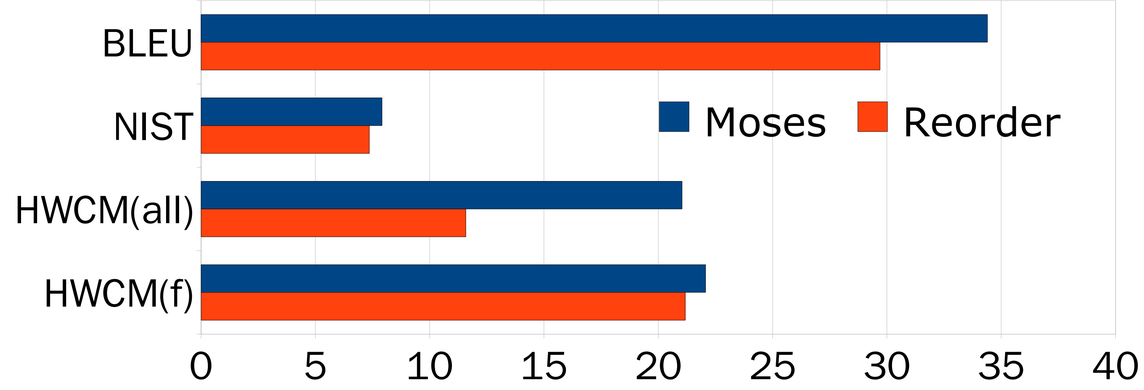


This figure shows the **general architecture** of reodering method for improving fluency. For training, the reordering method utilizes **word alignment**s. For translating, the reordering method utilizes **structure of the reordered target sentence**.

Results



The official results in **Japanese-to-English** translation show that cluster-based SMT achieve a lower score than PBSMT. The patent corpus has **less variation** in sentence type than the other domain such as conversation or news content.
Independent similarity measures for each translation model and language model are more appropriate than the integrated one. Using **mixed crietria** could make the **classification harder**.
Sentence-level features for the word similarity are **too sparse** to compute the cosine similarity. Smoothing techniques seem to be required to resolve the sparsity.
Target language models **based on source sentence** classification assume that the target structures also construct clusters. **N-gram based language model**s would not reveal the syntactic structure.



Various automatic evaluation metrics show that the reordering method achieve a lower score than PBSMT in **English-to-Japanese** translation.

Source	an operation display section 42 is provided on the upper surface of the main body of the copying machine .
Moses	操作 表示 部 42 が 設け られ て いる の 上面 に は 、 複 機 本 で ある 。
Reorder	複 機 本 の 上面 に は 操作 表示 部 42 が 設け られ て いる 。
Target	また 、 複 機 本 の 上面 に は 、 操作 表示 部 42 が 設け られ て いる 。

This table shows the translation results from both Moses and our reordering method. The source sentence has SVO structure. On the other hand, the referenced structure has OSV structure.

References

Cluster-based SMT to model specific domain

1.A.Finch and E.Sumita. Dynamic model interpolation for statistical machine translation, In Proc. of the Third Workshop on Statistical Machine Translation, pages 208-215, 2008.

2.G. Foster and R. Kuhn. Mixture-model adaptation for SMT, In Proc. of the Second Workshop on Statistical Machine Translaion, pages 128-135, 2007.

3.H. Yamamoto and E. Sumita. Bilingual cluster based models for statistical machine translation. IEICE transaction on information and systems E series D, 91(3):588, 2008.

Reordering method to improve fluency

1.H. Na, J.-J. Li, J. Kim, and J.-H. Lee. Improving fluency by reordering target constituents using mst parser in English-to-Japanese phrase-based smt. In Proc. of the twelfth Machine Translation Summit, pages 276–283, August 2009.