Enhancing Statistical Machine Translation For Low-Resource Languages Using Semantic Similarity

Highlights

1. **Semantic similarity models**
   (a) Cosine similarity
   (b) Longest common subsequence
   (c) WordNet
   (d) Word2Vec
2. **Improving sentence alignment by extending word alignment based on word similarity**
   - Word similarity is learnt from monolingual data using Word2Vec
3. **Improving conventional pivot methods by similarity between pivot phrases**
   - Similarity between pivot phrases is extracted using the four semantic similarity methods: cosine similarity, longest common subsequence, WordNet, Word2Vec
4. **A new framework was introduced to enhance SMT for low-resource languages**
   - Combining the two proposed methods: pivot translation and sentence alignment with a baseline model trained on an existed small bilingual corpus
5. **Using the proposed sentence alignment algorithm to build bilingual corpora from Wikipedia**
   - Achieving bilingual corpora for low-resource Southeast Asian language pairs: Indonesian-Vietnamese (78K parallel sentences), Malay-Vietnamese (58K parallel sentences), Filipino-Vietnamese (11K parallel sentences)