Machine Translation of WordNet Glosses

Proposer(s) / Proposatzailea(k): names / izenak

Gorka Labaka & German Rigau

Contact / Kontaktua: german.rigau@ehu.eus

Description / Deskribapena

Several local wordnets are integrated into the MCR: Spanish, Catalan, Basque, Galician. However, most of them lack of a proper definition. Moreover, for the new synsets of the Collaborative ILI we also need to provide a proper English definition of any new local concept.

Goals / Helburuak

Following (Giménez et al. 2005) the main objective of the project is to build a new NMT system for automatically translating English WordNet glosses. We intend to generate a preliminary material which could be utilized to enrich other wordnets lacking of glosses. We also plan to create a NMT system to automatically translate local WordNet glosses to English.

Requirements / Betebeharrak

Basic knowledge of Linux command-line interface:

- execution of programs through the command line
- handling of text files

Framework / Esparrua

NMT systems will be trained using one of the accessible implementations (OpenNMT, Marian, ...). These systems implement more than one encoder-decoder architecture (RNN, Transformer, CNN, ...), and it will be necessary to decide which one to use at the beginning of the project.

Tasks and plan / Atazak eta plana

- Collect publicly accessible general domain corpus
- Training a generic NMT system
- Gather specialized training examples (WordNet glosses, dictionary definitions, ...)
- Train a specialized NMT system with both generic and specialized corpus
- Use additional information about the gloss to be translated (semantic class, Wordnet categories, ...) to generate more accurate translations.

References

Giménez J., Màrquez L. and Rigau G. *Automatic Translation of WordNet Glosses*. Proceedings of the Eurolan workshop "Cross-Language Knowledge Induction", pg. 1-8. 2005.