Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations

Lasha Abzianidze\(^1\), Johannes Bjerva\(^2\), Killan Evang\(^1\), Hessel Haagsma\(^1\), Rik van Noord\(^1\), Pierre Ludmann\(^2\), Duc-Duy Nguyen\(^3\) and Johan Bos\(^1\)

\(^1\)CLCG, University of Groningen; \(^2\)École Normale Superieure de Cachan; \(^3\)University of Trento

**Motivation**
- We propose a framework for producing meaning banks for several languages
- Basic assumption: translations have similar meanings
- Method: projection of meaning representations from one language to another
- Driven by a word alignment, syntactic derivations and compositional semantics
- Result: the Parallel Meaning Bank (PMB)
- A multilingual parallel corpus with shared meaning representations
- The PMB provides data for building semantic parsers for languages other than English
- A test bed for cross-lingual compositional semantics

**Languages & Corpora**
- 11.3 M tokens
- 863 K sentences
- 285 K documents

**Universal Semantic Tagging**
- Language neutral (80 tags)
- Generalises over part-of-speech & named entity classes
- PER Name of a person
- NOT Negation
- EXS Untensed simple event
- PST Past tense

**Automatic Annotation Pipeline**
- Human annotation: BoW (Bits of Wisdom)
- Tokenizer
- EasyCCG Supertagger
- Semantic tagger
- Symbolizer

**Cross-Lingual Projection**

**Composed Meaning Representation**

**Conclusion**
- We proposed new language-neutral tagging schemes (e.g., for segmentation/tokenization and universal semantic tagging)
- Promising results in meaning projection
- An online PMB explorer allows collaborative annotation and automatic reprocessing of the documents

**Future Work**
- Enabling systematic studies of the challenges arising from loose translations and different semantic granularities
- Develop methods to automatically detect shifts in meaning that are deliberately made by human translators to obtain better translations

**Supertagger**
- Aligner
- EasyCCG
- BoW (Bits of Wisdom)

**Annotated Documents**
- Manually checked that all tags are correct
- Non-gold with at least one BoW per layer

**Layer**
- Tokens: 7,784, 5,191, 3,182, 1,047
- Tokens: 2,550, 689, 416, 511
- Tokens: 274,953, 198,504, 117,396, 38,878
- Tokens: 241,872

**Symbols**
- Tokens: 2,026, 1,234, 282,109

**Stats as of**
- 27.03.2017

**Ask me for the demo**

**Register now at the PMB site!**