Token-based MWE identification

AIM-WEST meeting Oct 29-31

Previous work - mwetoolkit

- MWE lexicon extraction
- Multi-level patterns
- Different modules
  - Candidate extraction
  - Feature generation
  - Filtering
- Evaluation in lexicography

mwetoolkit

Previous work – SMT and PVs

- Alexander Kobzar – co-directed L. Besacier
- Focus on English phrasal verbs
- Experiments on translation evaluation
  - Build dataset
  - Guidelines for manual evaluation
  - Perform evaluation
  - Compare SMT paradigms

Ongoing work

- Silvio Cordeiro – co-directed A. Villavicencio
- Token-based identification
  - Lexicon projection
  - Mwetoolkit patterns extension
    - Negation
    - Overlapping
    - Disambiguation
  - Measures to evaluate token-based identification
  - Human-readable formats

Future directions 1

- Comparison/integration with SOTA sequence models for token-based evaluation (Constant & Tellier, Schneider et al.)
- Token identificaiton on parallel corpora
  - Monolingual MWEs on both sides
  - Monolingual MWEs and word alignments
  - Word alignments
- Relation with paraphrases (?)

Future directions 2

- Monolingual parsing
  - Discontiguous verbal expressions
    - Variability
    - Semantics
    - Paraphrases
- Ambiguous constructions
  - Complex conjunctions and adverbs – tokenisation
    - “en fait”
  - Idioms – semantics
    - “piece of cake”