AIM-WEST Project - San Carlos Meeting

LIG (Grenoble), Lattice (Paris), LIF (Marseille)
UFRGS (Porto Alegre), UFSCar (São Carlos)

Funded by CNRS, FAPERGS and FAPESP

October 29-31, 2014

San Carlos Meeting : Agenda

1. Wednesday 29th
   - Project overview
   - Team presentations
   - Previous and ongoing research presentations
2. Thursday 30th
   - AIM-WEST Working session – towards a shared task on MWE translation
   - Task, corpora, evaluation protocols, MT systems
   - + Visioconf with ALine (UFRGS) - 2PM-4PM
3. Friday 31th
   - Workplan for 2015
   - Meeting with local international relations

AIM-WEST Keywords

- Natural Language Processing
- MultiWord Expressions
- Machine Translation
- Automatic Speech Recognition
- Language Technology
- Non Compositionality
- Cognitive Lexical Models

Goal

This project aims to investigate techniques, resources and protocols for ***evaluating*** and ***integrating*** models of ***MWE processing*** into ***MT and ASR technology***.

Partners

- LIG: Speech translation (machine translation and speech recognition), multilingual systems
- LIF: Multiword expressions, speech recognition, natural language processing
- LATTICE: Multiword expressions, cognitive models of the lexicon
- UFRGS: Multiword expressions, cognitive models of the lexicon
- UFSCar: Machine translation, automatic learning of computational-linguistic resources

Previous interactions

- CAMELEON project `http://cameleon.imag.fr`
- LATTICE - UFRGS (book, workshops...)
- UFRGS - UFSCar (students, papers...)
- LIG - LIF

AIM-WEST Funding

- From jan 2014 - dec 2016
- Budget managed by LIG (France), UFRGS and UFSCar (Brazil)
- 10k€ / year on French side - cannot be transfered to year N+1
- Missions in both directions + joint workshop
- Funded by CNRS, FAPERGS and FAPESP
- Website, mailing list
- Paper acknowledgements

Tasks (to be refined during **this meeting**)

1. Corpora construction
2. Protocols for translation evaluation
3. Pre-processing of corpora
4. Lexical and ontological resource construction for French, Portuguese and English
5. Cognitive and data intensive models of MWE learning
6. Lexical models of MWEs
7. Syntactic models of MWEs
8. Semantic models of MWEs
9. Automatic speech recognition systems
10. Machine translation
11. Reports, papers and articles

Tasks 1 and 2

- Tomorrow's discussion...
- Corpora construction
- Protocols for translation evaluation

Proposal

Shared task on MWE translation
Requires test sets and evaluation measures
Internal, then possibly propose as SEMEval or WMT task

Current Internship Subjects (co-supervised Fr. and Br.)

1. Better handling Multi-Word-Expressions (MWEs) in Statistical Machine Translation
2. Extending the MWEtoolkit with token identification
3. Extending the MWE annotation of parallel corpora to noun compounds/idioms
4. Modelling the variability of support verb constructions
5. Creating a dataset for phrasal verb compositionality based on wordnet synonyms
6. Evaluation protocol for assessing MWEs in Automatic Speech Recognition

Quick Presentation of AIM-WEST Web Site

- http://aim-west.imag.fr