Development of the Multilingual Semantic Annotation System

Scott Piao[1], *Francesca Bianchi*[2], *Carmen Dayrell*[1], *Angela D'Egidio*[2] and *Paul Rayson*[1]
[1]*Lancaster University, UK;* [2]*University of the Salento, Italy*

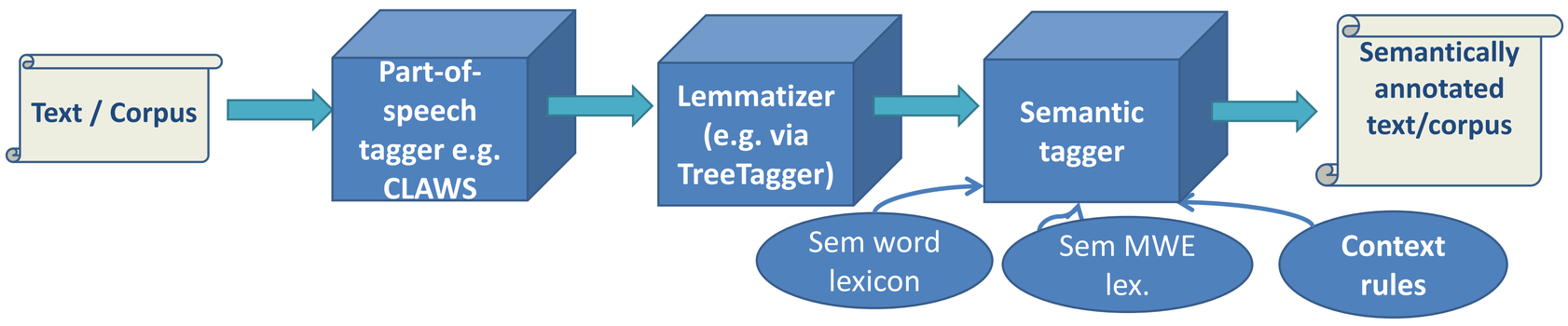
USAS

The UCREL Semantic Analysis System (USAS) is a software framework for automatic semantic analysis of natural language data. Its semantic classification scheme is based on Tom McArthur's Longman Lexicon (1981), consisting of 21 major discourse fields and 232 sub-categories (see http://ucrel.lancs.ac.uk/usas/).

Going Multilingual

Our aim is to make USAS more multilingual and it currently covers a number of languages, including Finnish, Russian, Italian, Portuguese, Chinese, Dutch etc., and work is underway to cover more languages. It provides a multilingual semantic analysis system under a unified semantic classification scheme.

Architecture of USAS



Multilingual Lexical Resource Construction

- In this work, English semantic lexicons are ported for Italian, Portuguese and Chinese.
- Bilingual lexical resources are used for transferring semantic tags of English words/MWEs to that of the above three languages.
- Bilingual lexical resources include: Chinese/English and Portuguese/English Dictionaries, LDC and FreeLang bilingual wordlist.
- Semantic tags of English words are automatically transferred to translation equivalents in the three languages using the bilingual lexicons.
- As a result, generated semantic lexicons including
  - Ita - 33,100 single words & 5,622 MWEs.
  - Chi - 64,413 single words & 19,039 MWEs.
  - Por - 13,942 single words & 1,799 MWEs.
- They contain errors, need further cleaning.

Multilingual USAS Evaluation Statistics

Language	Number of words	Tagged words	Lexicon coverage (%)
Italian	1,479,394	1,265,399	85.53
Chinese	975,482	786,663	80.64
Portuguese	1,705,184	1,251,579	73.40
Average			79.86

Table 1: Lexical Coverage

Language	Sample text size	Tagged words	Correct	Partially correct
Italian	4,510	3,266	1,826 (55.91%)	672 (20.58%)
Chinese	1,053	813	616 (75.76%)	97 (11.93%)
Portuguese	1,231	953	787 (82.58%)	68 (7.14%)
Average			71.42%	13.22%

Table 2: Precision

Experiment and Evaluation

- Automatically generated semantic lexicons are incorporated into USAS software framework and evaluated.
- Evaluated for lexical coverage and precision.
- Tables 1 and 2 above show the results.

Potential Applications

Multilingual USAS can potentially have various applications to practical tasks, such as multilingual information/knowledge extraction and retrieval, multilingual social computing, multilingual data science, multilingual corpus linguistics etc.

The UCREL Semantic Annotation System (USAS) website: *http://ucrel.lancs.ac.uk/usas/*

Contact: *{p.rayson,s.piao}@lancaster.ac.uk*