AUTOMATIC TERM EXTRACTION FROM COMPARABLE CORPORA

Data Collection & Annotation

1. comparable corpus
2. parallel corpora
3. languages (Dutch, English, French)
4. domains (corruption, dressage, heart failure, wind energy)
5. labels (Specific, Common, and OOD Terms & Named Entities)
6. 534,559 tokens of specialised text annotated
7. 110,444 monolingual annotations (terms & named entities)
8. 11,312 cross-lingual annotations (of equivalents)

Extracting Monolingual Terms

HAMLET
- Hybrid > combination of linguistic and statistical features
- Adaptable > to domains, languages and term types
- Machine Learning approach to forest classifier
- Extract > identify in specialized corpora
- Terminology > specialised, domain-specific linguistic units

TermoStat

Comparison with state-of-the-art on corpus about dressage (nl)

HAMLET: 55.6%
TermoStat: 16.2%
Gold Standard: 299

HAMLET vs TermoStat:
- f1-scores of pilot study:
  - Balanced (50/50)
  - Bit imbalanced (20/80)
  - Very imbalanced (5/95)
- en-fr
- en-nl
- fr-nl
- C: 84% 88% 90% 92% 94% 96%
- B: 86% 88% 90% 92% 94% 96%
- A: 90% 92% 94% 96% 98%

Linking Multilingual Equivalents

Ongoing & Future Work

- annotate multilingual training & test data
- pilot study with binary classifier and existing features
- add distributional & character features
- implement entire pipeline for multilingual term extraction from comparable corpora
- calibrate different components for optimal interaction
- evaluation and validation

Contact
ayla.rigoutsterryn@ugent.be

References