



PANACEA WP6: Lexical Acquisition

10-11 October 2011
Munich

UCAM

Technical developments

- **Verb classification**

- Goal
 - the development of a new method to acquire hierarchical classifications (instead of flat ones)
 - Integrating different levels of granularity, such classifications can be more useful for practical applications
- Lin Sun and Anna Korhonen. Hierarchical Verb Clustering Using Graph Factorization. 2011. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK.

Technical developments

- **Word sense induction**
- Goal: obtain lexical (e.g. SCF) information specific to word sense
- Method:
 - a factorization model where words, their context words and dependency relations are linked to latent dimensions
 - allows one to determine the dimensions that are important in a particular context, and adapt the dependency-based feature vector accordingly
- Tim Van de Cruys, Thierry Poibeau and Anna Korhonen. 2011. Latent Vector Weighting for Word Meaning in Context. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK.

Technical developments

- **Subcategorization frame acquisition**
 - Goal: compare SCFs in different domains
 - Method: a pilot study applied to the biomedical data
 - Conclusion:
 - significant variation found across sub-domains of biomedicine
 - resources created for a sub-domain may not generalize to the entire domain
 - minimally-supervised SCF acquisition method seems the most sensible way forward
 - Tom Lippincott, Laura Rimell, Karin Verspoor, Anna Korhonen. (to be submitted to JBMI). An investigation of challenges in the automatic acquisition of verb subcategorization information in the biomedical literature.

Other SCF experiments

- Improvements to current Cambridge SCF system (different parsers, combining parsers, work on rules)
- Experiments on using SCF information to improve parsing (i.e. task-based evaluation)
- A novel unsupervised approach to SCF acquisition and verb classification (suitable for domains) using Tensor Factorisation
 - Using method of Van de Cruys (2009)
 - Joint clustering of verbs by SCF and selectional preferences (idea: one form of lexical acquisition supports another)
 - Requires work on evaluation as unsupervised methods can discover information beyond the gold standards



SCF Gold Standard

- ENV/LAB Eng annotation complete (single annotator)
- Subset with 2nd annotator for I-AA due 21st Oct
 - (2 domains, 5 verbs also in general language gold standard)

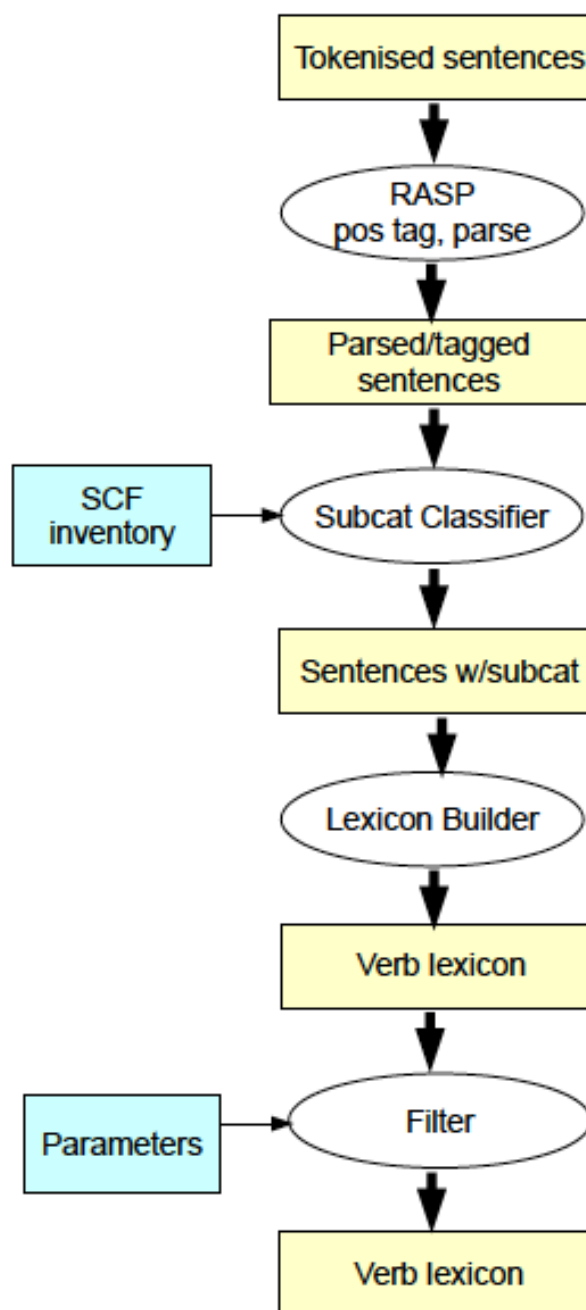


Internal milestone T22

Cf. Flash Meeting: July 15th 2011

UCAM:

- Complete domain annotation.
- Further develop SCF systems using the approach of Van de Cruys (2009).
- Initial results for domain SCF system.





Web services

UCAM:

- RASP will be available as a Web service through the PANACEA platform
- Subcat classifier, lexicon builder and filter

Real world evaluation

- Evaluation of lexical acquisition (for example) subcategorization frames in a MT environment?