Using Active Learning and Semantic Clustering for Noise Reduction in Distant Supervision

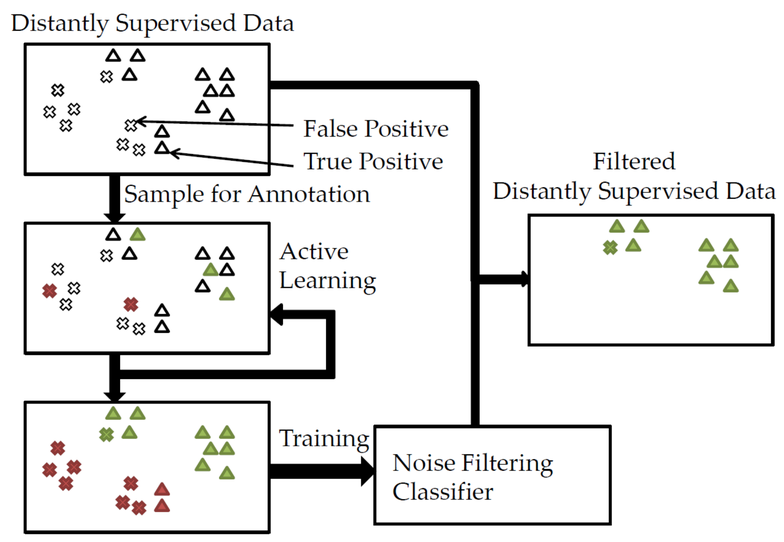
Lucas Sterckx, Thomas Demeester, Johannes Deleu and Chris Develder
Email: lucas.sterckx@intec.ugent.be

Noise Reduction in Distantly Supervised Data

- **Distant Supervision** :
  - **State-of-the-art** for training Knowledge Base Population
    - ➢ Inherently suffers from **noise**
  - We propose **noise reduction** using :
    - **Discriminative classifier** trained on a **small set of labeled examples**
    - **Active learning** strategy and **Semantic Similarity** between the contexts of the training examples

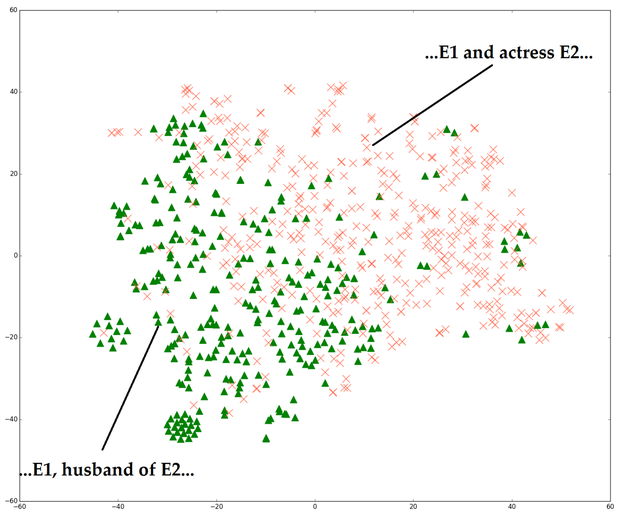
➢ Combination **facilitates the creation of a clean training set** for relation extraction, at a **reduced manual labeling cost**.

Methodology



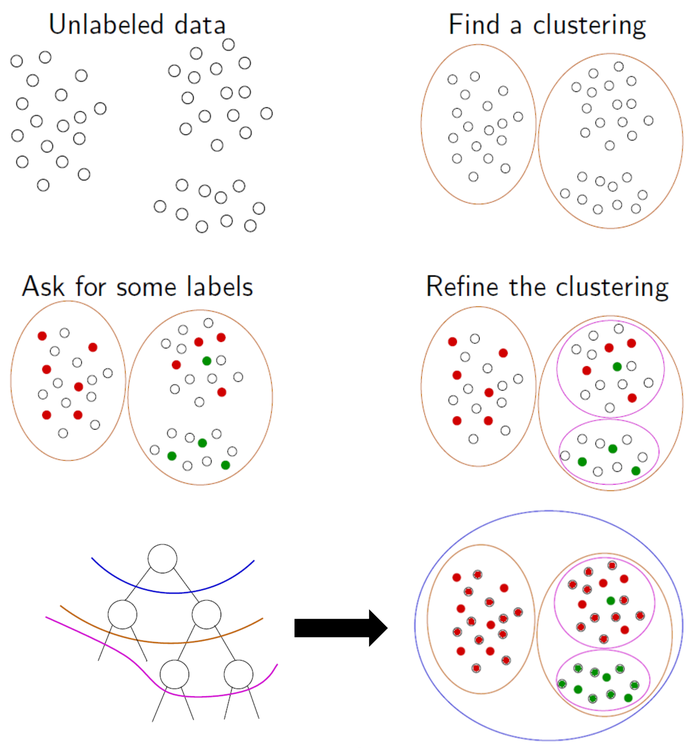
Semantic Clustering

- Distributional Hypothesis
- **Summed Average** of Embeddings of Words in context of relations
- Use **semantic vector representations** in cluster based active learning



Cluster-based Active Learning

- Exploit hierarchical cluster structure in data
- Efficient search through hypothesis space



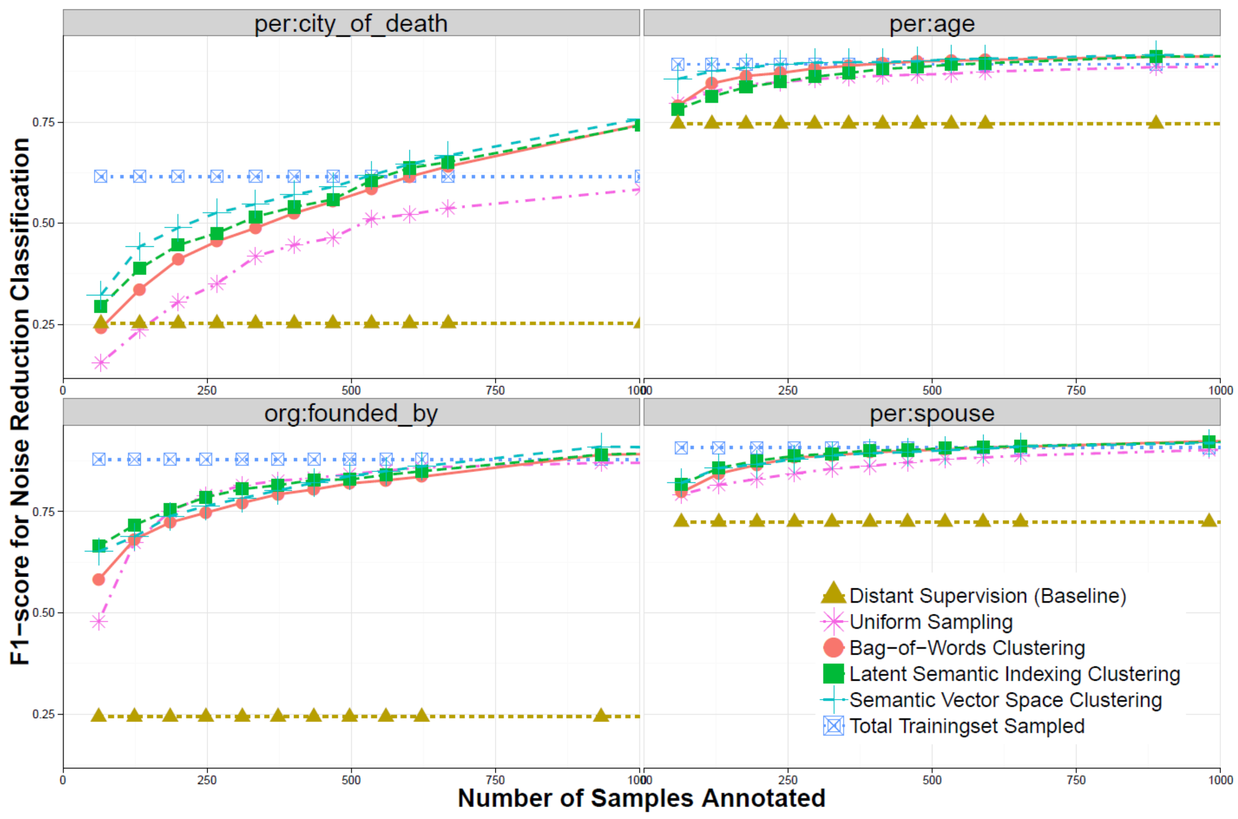
(Dasgupta and Hsu, 2008)

Experimental Setup

- Part of a participation in the TAC-KBP slot filing competition, precision after noise reduction **+ 8%**
- 2,000 training samples assigned with a True or False label with respect to the 2014 TAC-annotation guidelines for a selection of 12 relations with a large quantity of training data

	Precision	Recall	F1
Distant Supervision (Baseline)	51.9	100.0	60.8
Random Sampling	72.0	**72.8**	66.0
Bag-of-Words Clustering	73.4	65.2	66.6
Latent Semantic Indexing Clustering	73.7	68.5	68.3
Semantic Vector Space Clustering	**74.6**	71.4	**71.2**

^ Macro-average filter performance using **70 labeled distantly supervised training examples**



Conclusion

Novel approach for filtering a distantly supervised training set by building a binary classifier to detect true relation mentions. The classifier is trained using a cluster based active learning strategy. Clustering of relation mentions and adding semantic information reduces human effort and makes this a promising approach more feasible to filter a wide variety of relations.

Future work: we suggest the use of more **compositional methods** for transforming to a semantic vector space from the field of paraphrase detection.

iMinds
Internet Based Communication Networks and Services, IBCN • www.ibcn.intec.ugent.be
Department of Information Technology • Gaston Crommenlaan 8 bus 201, 9050 Gent, Belgium
INTEC