Active learning for spatial information extraction

Supervision:
Information: Parisa Kordjamshdi
Promoter: Marie-Francine Moens, Luc De Raedt
Supervision: Parisa Kordjamshdi

Research Group:
Declarative Languages and Artificial Intelligence (DTAI) group and Language Intelligence and Information Retrieval (LIIR) team

One of the essential functions of natural language is to talk about spatial relationships between objects. Extraction of spatial information, which we call spatial role labeling, is a key task for applications that are required to answer questions about, or have to reason about, spatial relationships. Examples include systems that perform text-to-scene conversion or inversely generate of textual descriptions from visual data, robots that understand directional or navigational instructions, geographical information systems (GIS) and many others.

Machine learning methods were only recently applied for the recognition of spatial information in language. Hence there are no publicly available corpora that are appropriate for training the machine learning models. For the English language the available labeled corpora are sparse, domain dependent and of limited size.

As a matter of fact labeling data to be used for machine learning is a very expensive task, so that choosing the most informative examples for training is a challenging problem in many applications. Intelligent selection of the examples by the learner is called active learning. This proposal concerns choosing the best examples in the context of spatial information extraction from natural language and efficiently building a corpus to perform this task. Different criteria will be considered including diversity and redundancy of the selected examples, and uncertainty of a current classifier trained on the labeled data. Linguistic knowledge can give additional guidance in the selection process.

Goal:
The purpose of this thesis is
1) To collect sparse small existing corpora which are fully or partially labeled with spatial concepts.
2) To unify and extend the collected corpora via active learning models.
3) To investigate the learning progress by increasing the training data.

Work Outline:
1. Perform a literature study on active learning and spatial information extraction.
2. Implement an active learning setting to choose most informative examples based on the existing corpora.
3. Selective sampling and labeling more data using the final implemented active learning model.

Relevant Literature:
2. Active Learning of Event Detection Patterns, Randolf Altmeyer and Ralph Grishman
Department of Computer Science, New York University, 2009.
3. Spatial Role Labeling: Task Definition and Annotation Scheme. In Proceedings LREC 2010,
KORDJAMSHIDI, Parisa, VAN OTTERLO, Martijn, MOENS, Marie-Francine.

Profile:
Knowledge or interest in machine learning and natural language processing. Programming skills.
Number of students: one student.