Technology Structure Mining from Scientific Literature

Behrang Qasemizadeh, Paul Buitelaar

Unit for Natural Language Processing, DERI

National University of Ireland, Galway

Behrang.Qasemizadeh@deri.org

Introduction

We introduce "Technology Structure Mining" as the task of extracting a labelled digraph, we name it "*Technology Structure Graph"*, from scientific publications in a domain of expertise in a way that:

- Each node of the graph refers to a technological concept
- Each edge of the graph refers to a relational concept that describes interdependencies between the technological concepts.

The proposed research involves several established research challenges in Information Extraction and Natural Language Processing:

- Generalized names learning and technical term extraction
- Relation Extraction
- Semantic Role Identification

and in a broader sense, Natural Language Understanding and Semantic Computing with two emerging research application areas:

- Open (Domain) Information Extraction (OIE)
- Ontology Learning (OL)

Example

Assuming the following sentence as an input:

"*There have been a few attempts to integrate a speech recognition device with a natural language understanding system."*

Then we expect a "Technology Structure Graph" as output with the following elements:

- Technical Term Vocabulary (TTV): {natural language understanding, speech recognition}
- Technical Concepts (TC): {<NLU,TECHNOLOGY>,<SR;TECHNOLOGY>}
- Relational Concept (RC): {MERGE}
- Relation Vocabulary (RV): {integrate with}
- Mapping function between TTV and TC: natural language understanding → <NLU,TECHNOLOGY>
  speech recognition → <SR,TECHNOLOGY>
- Mapping function between RV and RC: integrate with → MERGE
- Partial function that maps TC x TC → RC x M: <<SR,TECHNOLOGY>, <NLU,TECHNOLOGY>> → <MERGE, ◊>

* with M defined as *possible* and *certain* modalities, i.e., {◊, Δ}

Methodology

The proposed research is built upon two major technologies:

- Generic human language technology
- Machine learning techniques

In our proposed methodology we focus on techniques for:

- Automatic compilations of training data sets to develop machine learning models
- Introducing novel linguistic features
- Make use of Knowledge Based approaches for supporting Training -Set Collection, Term Representation, and Identification of common relations
- Reducing the task of technology and relation identifications to classification problems

Experiments and evaluation

We perform our experiments over two different corpora:

- ACL Anthology Reference Corpus (ACL ARC)
- Semantic Web Dog Food Corpus

Evaluation, however, *remains* one of the most complex aspects of our research. We have tried to develop datasets out of the above corpora for evaluation purposes as the research goes on. So far, our efforts have involved:

- Manual annotation of 486 sentences from Section C of ACL ARC
- Manual annotation of 3000 Technical Terms from 3 different sub-corpora

Ongoing and Future Work

- Introducing number of novel linguistic features specially “Factivity features”
- Distilling a training data-set for Relation Extraction by Traversing Wikipedia articles using Yago’s WordNet links Mapping between natural language text and structured information in DBPedia

References

[1] Behrang QasemiZadeh. Towards Technology Structure Mining from Text by Linguistics Analysis, DERI Technical Report, 2010.
[2] Behrang QasemiZadeh and Paul Buitelaar. Developing a Dataset for Technology Structure Mining, IEEE ICSC, 2010.