Learning Semantic Relations from Text

Preslav Nakov
Qatar Computing Research Institute, HBKU
Qatar Foundation
Tornado Tower, floor 10, PO box 5825
Doha, Qatar
pnakov@qf.org.qa

Diarmuid Ó Séaghdha
Computer Laboratory
University of Cambridge
Cambridge, UK
do242@cam.ac.uk

Vivi Nastase
Human Language Technologies
Research Unit
Fondazione Bruno Kessler
Trento, Italy
nastase@fbk.eu

Stan Szpakowicz
School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, Canada
szpak@eecs.uottawa.ca

1 Summary

Every non-trivial text describes interactions and relations between people, institutions, activities, events and so on. What we know about the world consists in large part of such relations, and that knowledge contributes to the understanding of what texts refer to. Newly found relations can in turn become part of this knowledge that is stored for future use.

To grasp a text's semantic content, an automatic system must be able to recognize relations in texts and reason about them. This may be done by applying and updating previously acquired knowledge. We focus here in particular on semantic relations which describe the interactions among nouns and compact noun phrases, and we present such relations from both a theoretical and a practical perspective. The theoretical exploration sketches the historical path which has brought us to the contemporary view and interpretation of semantic relations. We discuss a wide range of relation inventories proposed by linguists and by natural language processing people. Such inventories vary by domain, granularity and suitability for downstream applications.

On the practical side, we investigate the recognition and acquisition of relations from texts. In a look at supervised learning methods, we present available datasets, the variety of features which can describe relation instances, and learning algorithms found appropriate for the task.

Next, we present weakly supervised and unsupervised learning methods of acquiring relations from large corpora with little or no previously annotated data.

We show how enduring the bootstrapping algorithm based on seed examples or patterns has proved to be, and how it has been adapted to tackle Web-scale text collections. We also show a few machine learning techniques which can perform fast and reliable relation extraction by taking advantage of data redundancy and variability.

2 Outline

The tutorial is based on (Nastase et al., 2013).

1. Introduction
   (a) motivating example: the Mars rovers
   (b) what the tutorial is about
   (c) historical overview: semantic relations in linguistics, logic and artificial intelligence
   (d) examples, applications
2. Semantic Relations
   (a) relations between concepts vs. relations between nominals (OR paradigmatic vs. syntagmatic relations)
   (b) examples of fixed semantic inventories, with focus on relations in noun-noun compounds
   (c) the opposite view: that it is not possible to design a small set of semantic relations, but one can use, e.g., paraphrases
   (d) semantic relations in ontologies
   (e) properties of semantic relations: ontological vs. idiosyncratic, binary vs. n-ary, targeted vs. emergent, first-order vs. higher-order, general vs. domain-specific

(f) properties of semantic relation schemata: coarse-grained vs. fine-grained, flat vs. hierarchical, closed vs. open
(g) general clues for extracting semantic relations

3. Features
   (a) methods for learning semantic relations: supervised vs. unsupervised
   (b) entity vs. relational features
   (c) entity features: basic, syntactic, semantic clusters, co-occurrences, distributional representations, features from semantic networks
   (d) relational features: basic (bag of words, sequence of words, dependency paths, dependency graphs, constituency trees), paraphrases

4. Supervised Methods
   (a) task formulation
   (b) datasets: MUC, ACE, SemEval
   (c) algorithms for relation learning: kernels, sequential labeling models
   (d) non-binary relations
   (e) practical considerations

5. Unsupervised Methods
   (a) mining very large corpora
   (b) historical overview
   (c) mining dictionaries
   (d) Hearst's lexico-syntactic patterns
   (e) bootstrapping: basic algorithm, tackling semantic drift
   (f) distant supervision
   (g) unsupervised relation extraction: is-a relations, emergent relations
   (h) self-supervised relation extraction
   (i) Web-scale relation extraction: machine reading, never ending learning

6. Wrap-up
   (a) lessons learned
   (b) the bigger picture
   (c) hot research topics and future directions

3 Instructors

Preslav Nakov, a Senior Scientist at the Qatar Computing Research Institute, part of Qatar Foundation, holds a Ph.D. from the University of California at Berkeley. His research interests include computational linguistics and NLP, machine translation, lexical semantics, Web as a corpus and biomedical text processing.
⟨http://people.ischool.berkeley.edu/∼nakov/⟩

Vivi Nastase is a researcher at the Fondazione Bruno Kessler in Trento, working mainly on lexical semantics, semantic relations, knowledge acquisition and language evolution. She holds a Ph.D. from the University of Ottawa, Canada.
⟨http://hlt.fbk.eu/people/profile/nastase⟩

Diarmuid Ó Séaghdha, Senior NLP Researcher at VocalIQ and Visiting Industrial Fellow at the University of Cambridge, holds a Ph.D. from the University of Cambridge. His research interests include discourse and dialog, lexical and relational semantics, machine learning for NLP, scientific text mining and social media analysis.
⟨http://www.cl.cam.ac.uk/∼do242/⟩

Stan Szpakowicz, an emeritus professor of Computer Science at the University of Ottawa, holds a Ph.D. from the University of Warsaw and a D.Sc. from the Polish Academy of Sciences. He has been active in NLP since 1969. His recent interests include lexical resources, semantic relations and emotion analysis.
⟨http://www.eecs.uottawa.ca/∼szpak/⟩

References

[Nastase et al.2013] Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz. 2013. *Semantic Relations Between Nominals*, volume 19 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.