Unsupervised Bayesian Learning of Natural Language Semantics

Ivan Titov and Alexandre Klementiev, Saarland University

Why semantic representations?

Question answering about knowledge contained in a collection of biomedical publications:

Question: What does cyclosporin A suppress?

Answer: expression of EGR-2

Sentence: As with EGR-3 , expression of EGR-2 was blocked by cyclosporin A .

Question: What inhibits tnf-alpha?

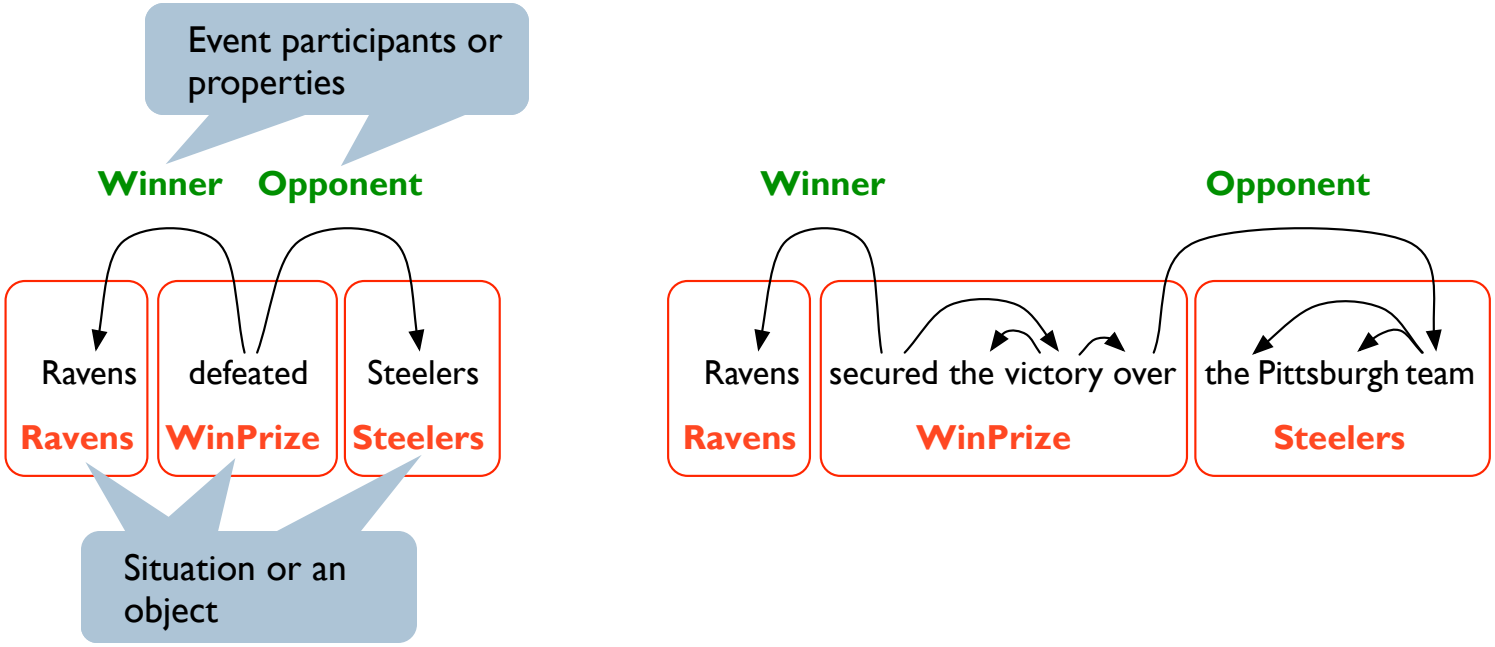
Answer: IL -10

Sentence: Our previous studies in human monocytes have demonstrated that interleukin (IL) -10 inhibits lipopolysaccharide (LPS) -stimulated production of inflammatory cytokines , IL-1 beta , IL-6 , IL-8 , and tumor necrosis factor alpha by blocking gene transcription .

We need to abstract away from specific syntactic and lexical realizations

Semantic Representations

- Given a text automatically annotated with syntactic information induce its semantics representation:



- A decomposition and clustering task on top of a syntactic graph
  - The syntactic graph is shown with black arrows

Our approach

- A joint model of syntax, semantic and lexical information (i.e. words)
- Unsupervised learning
  - Context in which a word (or phrase) appears provides a strong clue about its meaning
  - Possible alternatives (i.e. rule-based and fully supervised systems) do not achieve high coverage of semantic constructions and are domain dependent
- Nonparameteric Bayesian modeling
  - Provides an elegant alternative to model selection heuristics (e.g., deciding on the number of clusters and size of syntactic fragments)
  - More accurately handles uncertainty, as necessarily to account for linguistic ambiguity
  - Provides a natural way to encode prior knowledge as priors on model parameters

Empirical evaluation and extensions

- Application-based evaluation on the question answering task
  - Outperforms alternative techniques which do not tackle the task as induction of semantic representations
- A standard benchmark for shallow semantics: comparison to annotation by linguists
  - The best reported results among unsupervised models
- Induction of crosslingual semantics
  - Inducing semantics for two languages at the same time by exploiting agreement on parallel data (sentences and their translations)
  - The first result to show that crosslingual induction of semantics is beneficial
- A small amount of supervision helps substantially
  - There are some phenomena which are hard to learn in an unsupervised setting

Come to the poster to see details