Learning to Exploit Structured Resources for Lexical Inference
Vered Shwartz, Omer Levy, Ido Dagan and Jacob Goldberger

1. Lexical Inference

Task: Identifying Lexical Inference
- Inferring the meaning of one term from another: cat → animal
- Useful in many NLP applications

Prior Methods:
- Corpus-based: high recall, limited precision
- Resource-based:
  - Usually based on WordNet - manual selection of WordNet relations, high precision
  - Limited recall. Missing:
    - recent terminology (social network)
    - proper-names (Lady Gaga)

Goal: high-precision resource-based method with improved recall
Means: community-built resources:

We look at paths connecting x, y in resource graph

Which edge types are indicative of the target lexical inference relation?
Challenge: manual selection is infeasible — thousands of relations to choose from!
Solution: learn an optimal subset of edge types for a certain target relation

2. Structured Resources

- We use knowledge resources:
  - WordNet
  - Diederik
  - Yago
  - BibNet
  - DBpedia
  - Freebase

- Knowledge resource is a graph:
  - Nodes: terms / concepts
  - Edges: concept-term / semantic edges between concepts

3. Learning

Input: annotated dataset of term-pairs
- Defining a lexical-semantic relation $R$ (e.g. “is a”)

Task: Given a new term-pair $(x, y)$, predict whether $xRy$

Parameter: whitelist of indicative edge types $w = \{\text{instance-of, occupation, subclass-of, ...}\}$

Inference:
- A path is indicative if all its edge types are whitelisted.
- A term-pair is positive if at least one of its paths is indicative.

Training:
- Learn optimal whitelist (optimize $F_{\beta}$ on train set)
- Subset selection problem - Genetic Search

4. Evaluation

1. Can we replicate WordNet-based methods for common nouns?
   - Using WordNet as the only resource
   - Baseline: 4 manual whitelists based on the literature

   - Our method replicates the best whitelist on each dataset

2. Can we extract inferences over proper-names from community-built resources?
   - Improved recall over WordNet on our proper-names dataset
   - Maintains high precision (p=97%, r=29%)

   - Interesting whitelist (44 edge types):

3. How is our method compared to corpus-based methods?
   - Baseline: state-of-the-art distributional methods: word2vec (Mikolov et al., 2013) + concatenation (Baroni et al.,2012)
   - Corpus-based method: higher recall, lower precision

   - Our method complements corpus-based methods

5. Contribution

- Supervised framework for automatically selecting an optimized subset of resource relations for a given target inference task.
- Enables the use of large-scale resources, providing a rich source of high-precision inferences over proper-names.
- Available to download: https://github.com/vered1986/LinKeR