A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances

Chengyu Wang, Xiaofeng He*, Aoying Zhou
School of Computer Science and Software Engineering, East China Normal University

Introduction

- A taxonomy is a semantic hierarchy that organizes concepts by is-a relations, which is extensively used in many NLP and IR tasks.
- To include more domain-specific and long-tailed knowledge, methods have been developed to induce taxonomies from text corpora.
- Our work overviews recent advances on this topic, resources for evaluation and challenges for future research.

Taxonomy Construction Techniques - Is-a Relation Extraction

- **Pattern-based Methods**
  - Pioneer work is *Hearst patterns* (e.g., [C] such as [E]). It is prone to error due to idiomatic expressions, parsing errors, incomplete/uninformative extractions, ambiguous issues, etc.
  - **Methods Improving Recall**
    - **Pattern Generalization**: Learn more abstract patterns from Hearst-style patterns rather than surface matching.
    - **Iterative Extraction**: Iteratively learn patterns and is-a pairs from text corpora automatically. Use more specific patterns to avoid "semantic drift".
    - **Hypernym Inference**: If y is a hypernym of x and another term x' is sufficiently similar to x, there is a high probability that y is a hypernym of x'. Syntactic inference on hyponym modifiers can generate additional is-a relations.
  - **Methods Improving Precision**
    - **Confidence Assessment**: Discard extracted pairs with low confidence scores. Statistical measures include PMI, the ratio of likelihood, the prediction scores of classifiers, etc. Negative evidence and external data sources can be also employed to estimate confidence scores.
    - **Classification-based Validation**: Train a classifier f to predict the correctness of an extracted pair (x, y).
- **Distributional Methods**
  - The first step is to extract targeted term pairs from a domain-specific corpus by *key term extraction* and *domain filtering* to construct the desired taxonomy.
  - **Unsupervised Measures**
    - **Distributional Similarity Measures**: Most measures model the *asymmetric* property of is-a relations, following the *Distributional Inclusion Hypothesis* (DIH). Examples of such measures include WeedsPrec, BalAPInc, ClarkeDE, cosWeeds, invCL and WeightedCosine. A few measures such as SLQS go beyond the DIH.
    - **Features**: Contextual words, the relations between contextual words and the central word, more sophisticated memory frameworks, etc.
  - **Supervised Models**
    - **Classification**: Use word embeddings (e.g., Word2Vec, GloVe, ivLBL and SensEmbed) to represent x and y in a term pair (x, y). The *concat* model uses $\vec{x} \oplus \vec{y}$ as features but suffers from the *lexical memorization* problem. Other models include the *diff* model $\vec{y} - \vec{x}$, *simDiff*, *vector sum* $\vec{x} + \vec{y}$ and *dot product* $\vec{x} \cdot \vec{y}$.
      Is-a relation specific word embeddings can be learned by designing task-specific neural nets (e.g., the distance-margin based neural net, the dynamic weighting neural net).
    - **Hypernym Generation**: It make prediction for a pair (x, y) based on whether the is-a projection model can map $\vec{x}$ to a vector close enough to $\vec{y}$. A pioneer work uses piecewise linear projection model, followed by a number of extensions.
      The negative sampling technique proves effective to enhance projection learning, which considers not-is-a term pairs.
    - **Ranking**: It gives ranks to a collection of candidate hypernyms for an entity. It selects the top-1 term as the most probable hypernym.
- **Comparison** There are some disagreements on which methods are more effective for is-a relation prediction. Pattern-based methods are more precise but have limited recall and are overly language-dependent. Distributional approaches can make predictions based on the entire corpus but are less precise and more related to the training set. Integrating them may improve the performance.

Taxonomy Construction Techniques - Taxonomy Induction

- **Incremental Learning**: These methods construct an entire taxonomy from a "seed" taxonomy. They iteratively extract new entities and their related is-a relations. New entities are continuously attached to the taxonomy. The "seed" taxonomy are acquired by using existing knowledge sources, Hearst pattern matching and heuristic rules.
- **Clustering**: Taxonomy learning can be modeled as a clustering problem where similar terms clustered together may share the same hypernym. Hierarchical clustering is employed to cluster similar terms into a taxonomy.
- **Graph-based Induction**: They model the collection of is-a relations as a graph. The path from the root to each targeted term is constructed. A frequently used algorithm is the optimal branching algorithm. The factor graph model can be also employed to model the taxonomy induction problem as structured learning.
- **Taxonomy Cleansing**: Wrong is-a relations should be removed to improve the quality of the taxonomy.

Resources and Analysis

- **Resources**
  - **High-quality Taxonomies/KBs**: WordNet, YAGO, WiBi, etc.
  - **Test Sets**: (Please find the detailed citations in the paper.)

Contributor/Paper	#Positive	#Negative
Kotlerman et al. (2010)	1,068	2,704
Baroni and Lenci (2011)	1,337	13,210
Baroni et al. (2012)	1,385	1,385
Jurgens et al. (2012)	1,154	1,154
Levy et al. (2014)	945	11,657
Rei and Briscoe (2014)	3,074	-
Weeds et al. (2014)	2,564	3,771
Turney and Mohammad (2015)	920	772
Shwartz et al. (2016) (Lex)	5,659	22,636
Shwartz et al. (2016) (Rnd)	14,135	56,544

  - **Shared Tasks**: TExEval (SemEval-2015 Task 17) and TExEval-2 (SemEval-2016 Task 13).
    Distributional methods are not fully exploited in these tasks.
  - **Domain-specific Corpora**: AI papers, biomedical corpora, Web pages related to animals, plants and vehicles and MH370, terrorism reports, disease reports and emails, etc.
- **Evaluation Metrics**
  - **Is-a Relation Extraction Task**: Precision, Recall and F-Measure.
  - **Taxonomy Construction Task**: Please refer to TExEval track papers.
- **Our Recommendations**
  - **Ensemble Representations and Deep Architectures**: Currently, the power of deep learning is not fully exploited for taxonomy learning. How to take advantage of the deep learning boom for taxonomy induction is worth researching in the future.
  - **Benchmarks and Evaluation**: Benchmarks should contain text corpora, gold standards and evaluation metrics. Some progress has been made in TExEval tracks and other works, but there is more to do for a complete benchmark.
  - **Unambiguous and Canonicalized Terms**: It is desirable to construct taxonomies where each node represents an unambiguous term associated with its possible surface forms and their contexts.
  - **Incorporating Domain Knowledge**: Domain knowledge is essential for term and relation extraction in domain-specific corpora but it is difficult to obtain such knowledge from such limited corpora. It is an important task to construct a taxonomy based on a text corpus and a knowledge base of a specific domain.
  - **Non-English and Under-resourced Languages**: The task has not been extensively studied for under-resourced and non-English languages.

Conclusion

- We overview methods to learn hypernymy from texts, and discuss how to induce taxonomies from is-a relations.
- While there is significant success, this task is still far from being solved.

Contact info: chywang2013@gmail.com School of Computer Science and Software Engineering, East China Normal University