A Brief Survey on Taxonomy Construction Techniques

Luyu Gao, Xinwei He, Yunyi Zhang
University of Illinois at Urbana Champaign
April 26th, 2018
Introduction

- What is taxonomy?
  - A directed acyclic structure that organizes words in hypernym-hyponym relation (Is-A relation)

- Automatic taxonomy construction from text corpus
  - Is-A relation extraction:
    - Pattern-based methods
    - Distributional Methods
  - Taxonomy Induction:
    - Use extracted relations to construct taxonomy structure

- Recent Trend
  - Poincare Embedding
Pattern-based Methods

- Extract candidate hypernym pairs if two terms in the same sentence match with some pre-defined pattern, called lexico syntactic patterns (Hearst, 1992):
  - Pairs can be extracted with little pre-encoded knowledge.
  - Example: “The bow lute, such as the Bambara ndang, is...”
- However, using Hearst pattern can be risky:
  - **Wrong candidate pairs** can be extracted due to polysemy, idiomatic expression, parsing error, etc.
    - One simple example: “Domestic animals other than dogs such as cats...”
    - (“cat”, “dog”) will be extracted by Hearst Patterns.
  - Pattern-based methods always have to **sacrifice recall for precision** due to the ambiguity in language and the low quality of syntactic patterns
Pattern-based Methods: Techniques

- Hearst Patterns extension:
  - A list of 59 patterns and their precisions are collected from past literatures (Seitner et al., 2016)

- Automatically pattern generalization:
  - Parse the sentences that contain candidate hypernym pairs and automatically learn new patterns according to the dependency path (Snow et al., 2004)
  - “Star Patterns” (Navigli and Velardi, 2010):
    - Replacing non-frequent words with wildcards in each sentence – star patterns
    - Cluster sentence based on their star patterns
Pattern-based Methods: Techniques

- Iterative learning algorithm:
  - “doubly-anchored pattern” (Kozareva et al., 2008)
  - “cars such as Ford and *” and “Presidents such as Ford and *”
  - Hyponym candidates extracted by patterns, and new candidate can be used to generate more patterns
  - Increase the extraction precision and eliminate terms ambiguity
- Semantic iteration employed by Probase (Wu et al., 2012)
  - Use probabilistic model to capture semantic knowledge
  - The frequency of (“cat”, “animal”) is much higher than (“cat”, “dog”)
  - Bootstrap to acquire more knowledge to increase precision
Pattern-based Methods: Techniques

- **Hypernym Inference:**
  - Use linguistic head: (“white shark”, “shark”)
  - Word similarity (Ritter et al., 2009)
    - Use SVM to filter hypernym pairs extracted by Hearst Patterns
  - Similar terms tend to share hypernym
  - Similarity is modeled by Hidden Markov Model
- **Syntactic Contextual Subsumption** (Luu et al., 2014)
  - Capture other contextual information, e.g. “things tigers can eat”
  - Eat (“tiger”) ⊆ Eat (“animal”) => (“tiger”, “animal”)
- **Others:** capitalization, length difference, pattern order, sibling, ...
Distribution Method: Unsupervised

- Earlier unsupervised distributional method trend: Measure Construction
  - **Symmetric** measure construction
  - **Asymmetric** measure construction

- Symmetric measures
  - (e.g. Cosine, Jaccard and Jensen-Shannon divergence)
  - Most widely used measure: LIN (LIN, 1998)
  - Here \((x, y)\) is a hyponym hypernym pair, \(F_x\) and \(F_y\) are features for \(x, y\) and \(w_x(f)\) represents weight for feature \(f\) of \(x\)

\[
LIN(x, y) = \frac{\sum_{f \in F_x \cap F_y} w_x(f)}{\sum_{f \in F_x} w_x(f) + \sum_{f \in F_y} w_y(f)}.
\]
Distribution Method: Unsupervised

- Asymmetric measures
  - is-a relationship is non-symmetric

- Distributional Inclusion Hypothesis (DIH) (Geffet and Dagan, 2005; Zhitomirsky-Geffet and Dagan, 2009): a hyponym only appears in some of its hypernyms contexts, but a hypernym appears in all contexts of its hyponyms.

  - e.g. the concept fruit has a broader spectrum of contexts than its hyponyms, such as apple and banana.

- Among them SLQS by Santus et al. (2014) and Selective Distributional Inclusion Hypothesis by Roller et al. (2014) are the most influential ones.
Distribution Method: Supervised

- Supervised distribution-based method
  - Classification & Ranking
  - using part of valid hypernym-hyponym pairs to train a classifier or a ranking model
- Classification task
  - Word representation
    - Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and ivLBL (Mnih and Kavukcuoglu, 2013). SensEmbed (Iacobacci et al., 2015)
  - Methodology: Feature engineering & Classifier Training
    - vector difference (Rimell, 2014; Weeds et al., 2014; Fu et al., 2014)
    - Use both vector difference and squared vector difference as features (Roller et al., 2014)
Distribution Method: Supervised

- Ranking task
  - ranking is rarely used compared to classification due to its low recall
  - Chinese language specific taxonomy construct tasks
  - Fundamentally challenging due to expressive language expression
  - A more recent work (non-Chinese related) trying to construct taxonomy based on ranking measure has gained much attention. *(ExTaSem! extending, taxonomizing and semantifying domain terminologies Luis et al., 2017.)*
Sometimes, beyond simple Is-As, people want to see the full taxonomy structure.

Problem: building taxonomy structure from Is-A relations, place the words on the correct level and position.

Types of problems:
- Extracted Is-A relations
- Existing Knowledge
Taxonomy Induction: Graph Approach

- Defining the positioning of words between root and leaf.
- Solution: Graph approach
  - Naturally, all the Is-A relation extracted can be represented as a directed graph with edges representing Is-A relation.
  - Define the best path from each leaf node to root
    - Longest path (Kozareva and Hovy 2010), for each non-root node, the longest path to root will be adopted as the real path.
    - Relevant Path (Anke et al. (2016)), leverage SENSEMBED to assign vector representations to node that are relevant to current domain, use semantically close pairs as candidates to build paths.
Taxonomy Induction: Probabilistic Approach

- Probabilistic approaches to model taxonomy based on current knowledge.
- Noisy Is-A relations
  - WebIsA dataset, 400 million noisy Is-A relations
  - Use probabilistic framework to extract longer Is-A subsequences from noisy Is-A relations by modeling their probability. Build taxonomy based on high probability sequences. (Gupta et al. 2017)
- Incremental learning
  - Extend existing taxonomy structure such as WordNet
  - Model the probability of an extended taxonomy based on the evidence of is-a and cousin relations harvested from texts. (Snow et al. 2006)
  - You can also use SetExpan to enrich siblings.
Poincare Embedding

- Taxonomy Embedding
- Fundamental limitation of traditional embedding models
  - The number of dimensions grows exponentially as the height of the tree grows.
  - Linear embedding of graphs may require a prohibitively large dimensionality to model certain types of relations
- Solution: Euclidean space -> Hyperbolic space \((\text{Nickel and Kiela, NIPS 2017})\)
  - Constant negative curvature
  - Hyperbolic space can be thought of as a continuous version of trees and as such, it is naturally equipped to model hierarchical structures.
  - Number of dimensions needed drops significantly
Poincare Embedding: Poincare ball model

- Poincare ball model:
  - Explicitly capturing the hierarchy in the embedding space
  - Better suit for gradient-based optimization than other hyperbolic space model
  - Differentiable distance function:

\[ d(u, v) = \text{arcosh} \left( 1 + 2 \frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)} \right) \]

- Root node placed at the origin:
  - Zero Euclidean norm => small distance to all other nodes
- Leaf nodes placed close to the boundary:
  - Distance grows fast with a Euclidean norm close to one
Poincare embedding: Optimization

- Loss function:

\[
\mathcal{L}(\theta) = \sum_{(u, v) \in D} \log \frac{e^{-d(u,v)}}{\sum_{v' \in N(u)} e^{-d(u,v')}}
\]

- \( D = \{(u, v)\} \) is the set of observed hypernym relations between noun pairs
- \( N(u) = \{v \mid (u, v) \not\in D\} \cup \{u\} \) is the set of negative examples sampled from training data

- Training Details:
  - Randomly Initialize all embeddings close to the origin
  - Burn-in: good initial angular layout can be helpful to find good embeddings
Poincare Embedding: Evaluation

- Data: WORDNET noun hierarchy
- Taxonomy reconstruction and link prediction:

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Euclidean</strong></td>
<td>Rank</td>
<td>3542.3</td>
<td>2286.9</td>
<td>1685.9</td>
<td>1281.7</td>
<td>1187.3</td>
</tr>
<tr>
<td></td>
<td>MAP</td>
<td>0.024</td>
<td>0.059</td>
<td>0.087</td>
<td>0.140</td>
<td>0.162</td>
</tr>
<tr>
<td><strong>Translational</strong></td>
<td>Rank</td>
<td>205.9</td>
<td>179.4</td>
<td>95.3</td>
<td>92.8</td>
<td>92.7</td>
</tr>
<tr>
<td></td>
<td>MAP</td>
<td>0.517</td>
<td>0.503</td>
<td>0.563</td>
<td>0.566</td>
<td>0.562</td>
</tr>
<tr>
<td><strong>Poincaré</strong></td>
<td>Rank</td>
<td>4.9</td>
<td>4.02</td>
<td>3.84</td>
<td>3.98</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>MAP</td>
<td>0.823</td>
<td>0.851</td>
<td>0.855</td>
<td>0.86</td>
<td>0.857</td>
</tr>
</tbody>
</table>

| **Euclidean**  | Rank | 3311.1 | 2199.5 | 952.3 | 351.4 | 190.7 | 81.5 |
|                | MAP  | 0.024 | 0.059 | 0.176 | 0.286 | 0.428 | 0.490 |
| **Translational** | Rank | 65.7 | 56.6 | 52.1 | 47.2 | 43.2 | 40.4 |
|                | MAP  | 0.545 | 0.554 | 0.554 | 0.56 | 0.562 | 0.559 |
| **Poincaré**   | Rank | 5.7 | 4.3 | 4.9 | 4.6 | 4.6 | 4.6 |
|                | MAP  | 0.825 | 0.852 | 0.861 | 0.863 | 0.856 | 0.855 |

- Successful in the embedding of large taxonomies – both with regard to representation capacity and generalization performance
Thank you!

Full survey report and reference: