Taxonomy Construction Using Syntactic Contextual Evidence

Luu Anh Tuan¹, Jung-jae Kim¹, Ng See Kiong²

¹School of Computer Engineering, Nanyang Technological University, Singapore
²Institute for Infocomm Research, A*STAR, Singapore
Outline

• Introduction
• Related work
• Methodology
• Experiments
• Conclusion and future work
Taxonomy

• Useful for many areas:
  • question answering
  • document clustering

• Some available hand-crafted taxonomies: WordNet, OpenCyc, Freebase
  • time-consuming
  • more general, less specific

→ demand for constructing taxonomies for new domains
Outline

• Introduction
• Related work
• Methodology
• Experiments
• Conclusion and future work
Taxonomic relation identification

- **Statistical approach:**
  - Co-occurrence analysis (Budanitsky, 1999), term subsumption (Fotzo, 2004), clustering (Wong, 2007).
  - Less accurate, heavily depend on feature types and dataset

- **Linguistic approach:**
  - Hand-written patterns: (Kozareva, 2010), (Wentao, 2012)
  - Automatic bootstrapping: (Girju, 2003), (Velardi, 2012)
  - Lack of contextual analysis across sentences → low coverage
Our contribution

• Propose syntactic contextual subsumption method:
  • Utilize contextual information of terms in syntactic structures by evidence from the Web
  • Infer taxonomic relations between terms in different sentences

• Introduce graph-based algorithm for taxonomy induction:
  • Utilize the evidence scores of edges
  • Base on graph’s topological properties
Outline

- Introduction
- Related work
- Methodology
- Experiments
- Conclusion and future work
Workflow

1. Term extraction and filtering
2. Taxonomic relation identification
3. Taxonomy induction
Term extraction and filtering

• Term extraction:
  • Apply Stanford parser \(\rightarrow\) extract all noun phrases
  • Remove determiners, do lemmatization

• Term filtering:
  • TF-IDF
  • Domain relevance, domain consensus (Navigli and Velardi, 2004)

\[
TS(t, D) = \alpha \times \text{TFIDF}(t, D) + \beta \times \text{DR}(t, D) + \gamma \times \text{DC}(t, D)
\]
Taxonomic relation identification

• Combine three methods:
  • Syntactic contextual subsumption
  • String inclusion with WordNet
  • Lexical-syntactic pattern matching
Syntactic contextual subsumption (SCS)

• Find relations across different sentences
• Utilize syntactic structure (Subject, Verb, Object)

• Observation 1: (terrorist, attack, people),
  (terrorist, attack, American)
  \[ \Rightarrow \text{people} \gg \text{American} \]
• But from (animal, eat, meat) and (animal, eat, grass)?
Observation 2:

\[ s_1 \gg s_2 \]

- $S(\text{animal, eat}) = \{\text{meat, wild boar, deer, buffalo, grass, potato, insects}\}$
- $S(\text{tiger, eat}) = \{\text{meat, wild boar, deer, buffalo}\}$

$\rightarrow \text{animal} \gg \text{tiger}$
Syntactic contextual subsumption (SCS)

For terms $s_1$, $s_2$:
- Find most common relation $v$ between $s_1$ and $s_2$. Suppose $s_1$ and $s_2$ are both subjects
- Submit query “$s_1 v$” to search engine, collect first 1000 results, find $S(s_1, v) = \{o | \exists (s_1, v, o)\}$
- Similar for $S(s_2, v)$
- Calculate:

$$Score_{SCS}(s_1, s_2) = \left[ \frac{|S(s_1, v) \cap S(s_2, v)|}{|S(s_2, v)|} + \left( 1 - \frac{|S(s_1, v) \cap S(s_2, v)|}{|S(s_1, v)|} \right) \right]$$

$$\times \log(|S(s_1, v)| + |S(s_2 v)|)$$
String inclusion with WordNet (SIWN)

- SIWN method:

  \[ t_1 = w_{11} w_{12} w_{13} \]
  \[ t_2 = w_{21} w_{22} w_{23} w_{24} w_{25} \]

  \( t_1 \gg t_2 \) or \( t_1 \approx t_2 \)

  \( \gg: \) is hypernym of

  “suicide attack” \( \gg \) “self-destruction bombing”

- attack \( \gg \) bombing

- suicide \( \approx \) self-destruction

\[
Score_{SIWN}(t_1, t_2) = \begin{cases} 
1 & \text{if } t_1 \gg t_2 \text{ via SIWN} \\
0 & \text{otherwise}
\end{cases}
\]
Lexical-syntactic pattern (LSP)

- Use following patterns to query on Google:
  
  
  “t₁ such as t₂”
  “t₁, including t₂”
  “t₂ is [a|an] t₁”
  “t₂ is a [kind|type] of t₁”
  “t₂, [and|or] other t₁”

\[
\text{Score}_{LSP}(t₁, t₂) = \frac{\log(WH(t₁, t₂))}{1 + \log(WH(t₂, t₁))}
\]
Combined method

\[
Score(t_1, t_2) = \alpha \times Score_{SIWN}(t_1, t_2) \\
+ \beta \times Score_{LSP}(t_1, t_2) \\
+ \gamma \times Score_{SCS}(t_1, t_2)
\]
Taxonomy induction

- Step 1: Initial hypernym graph with a ROOT node
- Step 2:
  \[ w(e(t_1, t_2)) = \begin{cases} 
  1 & \text{if } t_1 = \text{ROOT} \\
  \text{Score}(t_1, t_2) & \text{otherwise}
\end{cases} \]
- Step 3: apply Edmonds’ algorithm to find maximum optimum branching of weighted directed graph
Outline

• Introduction
• Related work
• Methodology
• Experiments
• Conclusion and future work
Constructing new taxonomies

- **Terrorism domain:**
  - 104 reports of the US state department “Patterns of Global Terrorism (1991-2002)”
  - Each report ~1,500 words

- **Artificial Intelligence (AI) domain:**
  - 4,119 papers extracted
    - the IJCAI proceedings from 1969 to 2011
    - the ACL archives from 1979 to 2010
Taxonomy construction

- Compare constructed AI taxonomy with that of (Velardi et al., 2012)

|                        | Our system | Velardi’s system |
|------------------------|------------|------------------|
| #vertex                | 1839       | 1675             |
| #edge                  | 1838       | 1674             |
| Average depth          | 6.2        | 6                |
| Max depth              | 10         | 10               |
| Term coverage          | 83%        | 76%              |
## Taxonomy construction

- Number of taxonomic relations extracted by different methods

| Method          | Terrorism domain | AI domain  |
|-----------------|------------------|------------|
| SCS             | 484              | 1308       |
| SIWN            | 301              | 984        |
| LSP             | 527              | 1537       |
| SIWN + LSP      | 711              | 2203       |
| SCS + SIWN + LSP| 976              | 3122       |
Taxonomy construction

• Estimated precision of taxonomic relation identification methods in 100 random extracted relations

|                  | Percentage of correct relations |
|------------------|-------------------------------|
|                  | Terrorism domain | AI domain |
| SCS              | 91%               | 88%       |
| SIWN             | 96%               | 91%       |
| LSP              | 93%               | 93%       |
| SCS + SIWN + LSP | 92%               | 90%       |
Evaluate against WordNet

- Three domains: Animals, Plants and Vehicles:
  - Use the bootstrapping algorithm described in (Kozareva, 2008)
  - Compare the results with (Kozareva, 2010) and (Navigli, 2011)

|                 | Animals domain | Plants domain | Vehicles domain |
|-----------------|----------------|---------------|-----------------|
|                 | Our | Kozareva | Navigli | Our | Kozareva | Navigli | Our | Kozareva | Navigli |
| Term coverage   | 96% |   N.A.  | 94%     | 98% |   N.A.  | 97%     | 97% |   N.A.  | 96%     |
| Precision       | 95% | 98%    | 97%     | 95% | 97%    | 97%     | 93% | 99%    | 91%     |
| Recall          | 56% | 38%    | 44%     | 53% | 39%    | 38%     | 69% | 60%    | 49%     |
| F-measure       | 71% | 55%    | 61%     | 68% | 56%    | 55%     | 79% | 75%    | 64%     |
## Syntactic structures

- Comparison of three syntactic structures: \textit{S-V-O} (Subject-Verb-Object), \textit{N-P-N} (Noun- Preposition-Noun) and \textit{N-A-N} (Noun-Adjective- Noun)

|                      | \(S-V-O\) | \(N-P-N\) | \(N-A-N\) |
|----------------------|-----------|-----------|-----------|
| **Animals domain**   |           |           |           |
| Precision            | 95\%      | 68\%      | 72\%      |
| Recall               | 56\%      | 52\%      | 47\%      |
| F-measure            | 71\%      | 59\%      | 57\%      |
| **Plants domain**    |           |           |           |
| Precision            | 95\%      | 63\%      | 66\%      |
| Recall               | 53\%      | 41\%      | 43\%      |
| F-measure            | 68\%      | 50\%      | 52\%      |
| **Vehicles domain**  |           |           |           |
| Precision            | 93\%      | 59\%      | 60\%      |
| Recall               | 69\%      | 45\%      | 48\%      |
| F-measure            | 79\%      | 51\%      | 53\%      |
Dataset link

- All dataset and experiment results are available at
  [http://nlp.sce.ntu.edu.sg/wiki/projects/taxogen](http://nlp.sce.ntu.edu.sg/wiki/projects/taxogen)
Outline

• Introduction
• Related work
• Architecture
• Experiments
• Conclusion and future work
Conclusion

• Proposed a novel method of identifying taxonomic relations using contextual evidence from syntactic structure and Web data
• Presented a graph-based algorithm to induce an optimal taxonomy from a given taxonomic relation set
• Generally achieve better performance than the state-of-the-art methods
Future work

• Build the probabilistic model for taxonomy
• Consider the time stamp of information
• Apply to other domains and integrate into other frameworks such as ontology learning or topic identification
THANK YOU

Q & A
References

1. W. Wentao, L. Hongsong, W. Haixun, and Q. Zhu. 2012. *Probase: A probabilistic taxonomy for text understanding*. In proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 481-492.

2. Z. Kozareva, E. Riloff, and E. H. Hovy. 2008. *Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs*. In proceedings of the 46th Annual Meeting of the ACL, pp. 1048-1056.

3. R. Navigli, P. Velardi and S. Faralli. 2011. *A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch*. In proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1872-1877.

4. P. Velardi, S. Faralli and R. Navigli. 2012. *Ontolearn Reloaded: A Graph-based Algorithm for Taxonomy Induction*. Computational Linguistics, 39(3), pp.665-707.

5. J. Edmonds. 1967. *Optimum branchings*. Journal of Research of the National Bureau of Standards, 71, pp. 233-240.

6. M. A. Hearst. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. In proceedings of the 14th Conference on Computational Linguistics, pp. 539-545.
7. Z. Kozareva, E. Riloff, and E. H. Hovy. 2008. *Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs*. In proceedings of the 46th Annual Meeting of the ACL, pp. 1048-1056.

8. W. Wong, W. Liu and M. Bennamoun. 2007. *Tree-traversing ant algorithm for term clustering based on featureless similarities*. Data Mining and Knowledge Discovery, 15(3), pp. 349-381.

9. A. Budanitsky. 1999. *Lexical semantic relatedness and its application in natural language processing*. Technical Report CSRG-390, Computer Systems Research Group, University of Toronto.

10. H. N. Fotzo and P. Gallinari. 2004. *Learning “Generalization/Specialization” Relations between Concepts-Application for Automatically Building Thematic Document Hierarchies*. In proceedings of the 7th International Conference on Computer-Assisted Information Retrieval.

11. D. Widdows and B. Dorow. 2002. *A Graph Model for Unsupervised Lexical Acquisition*. In proceedings of the 19th International Conference on Computational Linguistics, pp. 1-7.

12. R. Girju, A. Badulescu, and D. Moldovan. 2003. *Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations*. In proceedings of the NAACL, pp. 1-8.