Probase: A Probabilistic Taxonomy for Text Understanding

Wentao Wu ¹, Hongsong Li ², Haixun Wang ², Kenny Q. Zhu ³

¹ University of Wisconsin, Madison, WI, USA
² Microsoft Research Asia, Beijing, China
³ Shanghai Jiao Tong University, Shanghai, China
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion
Text Understanding

- Machines need to understand text to unlock the information confined in Web data.

What’s this?

“Pablo Picasso, 25 Oct 1881, Spain”

“animals other than dogs such as cats”

“cats are animals”? or “cats are dogs”?
Conceptualization

- A little piece of *knowledge* makes the difference.
  - “Pablo Picasso is a person”
  - “cats are animals”

- Can machines know this?
  - They can’t.
  - We need to pass this piece of knowledge to them.
Taxonomies

A *hierarchical* structure showing the *isA* relationships among concepts.

![Hierarchical structure diagram](image)
Limited Size of Concept Space

“How do we compete with the largest companies in US?”

<table>
<thead>
<tr>
<th>Existing Taxonomies</th>
<th>Number of Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probase</td>
<td>2,653,872</td>
</tr>
<tr>
<td>YAGO</td>
<td>352,297</td>
</tr>
<tr>
<td>WordNet</td>
<td>25,229</td>
</tr>
<tr>
<td>Freebase</td>
<td>1,450</td>
</tr>
<tr>
<td>DBPedia</td>
<td>259</td>
</tr>
<tr>
<td>NELL</td>
<td>123</td>
</tr>
</tbody>
</table>
Knowledge is Black and White

“How do we compete with the largest companies in US?”

“Vague” concepts

- “largest companies in US” => Walmart? Microsoft? P&G?
- “beautiful cities” => Seattle? Chicago? Shanghai?

There is inherent uncertainty inside these concepts!
Probase

- Automatically constructed from 1.6 billion web pages (with 92.4% precision).

- The largest concept space so far (2.6 million).

- Use probabilistic approach to model the uncertainty inside the concepts.
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion
Previous Work

- Syntactic Iteration (*KnowItAll, TextRunner, NELL*)

  e.g., Hearst Patterns (as seeds):
  
  $$NP \text{ such as } \{NP,\}^*\{(or|and)\} NP$$
Problems of Syntactic Iteration

- Syntactic patterns have limited extraction power.
  - “… animals other than dogs such as cats …”

- High quality syntactic patterns are rare.
  - **Good** patterns: “x is a country” => x = “China”
  - **Bad** patterns: “war with x” => x = “planet Earth”

- Recall is sacrificed for precision.
  - E.g., some methods only focus on extracting *proper nouns*. 
Our Approach

- Semantic Iteration

Syntactic Iteration

Semantic Iteration
An Example

\( s: \ldots \text{companies other than oil companies such as IBM, Walmart, Proctor and Gamble}, \ldots \)
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion
Goal

- Build a taxonomy graph from the edges ("isA" pairs) from the previous data extraction stage.

(organisms, animals)
(organisms, plants)
(plants, trees)
(plants, grass)
Challenges

- Should we merge the two “apple” here?
  - \( e_1 = (\text{fruit, apple}), e_2 = (\text{companies, apple}) \)

- Should we merge the two “plants” here?
  - \( e_1 = (\text{plants, tree}), e_2 = (\text{plants, steam turbines}) \)

Words such as “apple” and “plants” have multiple meanings (senses).
Properties & Operations(1)

Example:

- ... plants such as trees, grass, and herbs ...
- ... plants such as steam turbines, pumps, and boilers ...

Local Taxonomy Construction
Properties & Operations (2)

Example:

a) ... plants such as trees, grass, and herbs ...
b) ... plants such as trees, grass, and shrubs ...

Horizontal Merge
Properties & Operations (3)

- Example:
  a) ... organisms such as plants, trees, grass and animals ...
  b) ... plants such as trees, grass, and shrubs ...
  c) ... plants such as steam turbines, pumps, and boilers ...

Vertical Merge
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion
Plausibility

How likely is that the claim “y is an x” is true?

\[ P(x, y) = 1 - p(E) = 1 - p(\prod_{i=1}^{n} s_i) = 1 - \prod_{i=1}^{n} (1 - p_i) \]

- \( s_i \): evidence (or sentence) that supports \((x, y)\)
- \( p_i \): the probability that the evidence \( s_i \) is true
Typicality

- Which one is more typical for the concept “bird”? A robin or ostrich?

\[ T(i \mid x) = \frac{n(x, i) \cdot P(x, i)}{\sum_{i' \in I_x} n(x, i') \cdot P(x, i')} \]

An instance of “big company” is also an instance of “company”.

\[ T(i \mid x) = \frac{\sum_{y \in D(x)} \tilde{P}(x, y) \cdot n(y, i) \cdot P(y, i)}{\sum_{i' \in I_x} \sum_{y \in D(x)} \tilde{P}(x, y) \cdot n(y, i') \cdot P(y, i')} \]

\( \tilde{P}(x, y) \) is the plausibility that \( y \) is a descendant concept of \( x \).
Application of Typicality (1)

- Semantic Web Search (ER’12)
Application of Typicality (2)

- Understanding Web Tables (ER’12)

![Image of a table showing American politicians' birthdays]

<table>
<thead>
<tr>
<th>Birth order</th>
<th>Birthplace</th>
<th>Century</th>
<th>Order of office</th>
<th>U.S. Vice President</th>
<th>Birthdate</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>Yorba Linda, California</td>
<td>26th</td>
<td>36</td>
<td>Richard Nixon</td>
<td>January 9, 1913</td>
</tr>
<tr>
<td>28</td>
<td>New York City, New York</td>
<td>19th</td>
<td>25</td>
<td>Theodore Roosevelt</td>
<td>October 27, 1858</td>
</tr>
<tr>
<td>46</td>
<td>Indianapolis, Indiana</td>
<td>26th</td>
<td>44</td>
<td>Dan Quayle</td>
<td>February 4, 1947</td>
</tr>
<tr>
<td>39</td>
<td>Wallace, South Dakota</td>
<td>26th</td>
<td>38</td>
<td>Hubert Humphrey</td>
<td>May 27, 1911</td>
</tr>
<tr>
<td>40</td>
<td>Omaha, Nebraska</td>
<td>26th</td>
<td>40</td>
<td>Gerald Ford</td>
<td>July 14, 1913</td>
</tr>
<tr>
<td>42</td>
<td>Milton, Massachusetts</td>
<td>26th</td>
<td>43</td>
<td>George H. W. Bush</td>
<td>June 12, 1924</td>
</tr>
<tr>
<td>44</td>
<td>Lincoln, Nebraska</td>
<td>26th</td>
<td>46</td>
<td>Dick Cheney</td>
<td>January 30, 1941</td>
</tr>
<tr>
<td>45</td>
<td>Scranton, Pennsylvania</td>
<td>26th</td>
<td>47</td>
<td>Joseph Biden</td>
<td>November 20, 1942</td>
</tr>
<tr>
<td>9</td>
<td>Kinderhook, New York</td>
<td>18th</td>
<td>8</td>
<td>Martin Van Buren</td>
<td>December 5, 1782</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>Senator</th>
<th>Party</th>
<th>Date of birth</th>
<th>Term</th>
<th>Age (Years/Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>Hillary Clinton</td>
<td>Democratic</td>
<td>October 26, 1947</td>
<td>2001 - 2009</td>
<td>2001 19 26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Application of Typicality (3)

- Short Text Understanding (IJCAI'11)
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion
A concept is *relevant* if it appears at least once in the top 50 million popular queries in Bing’s query log.
IsA Relationship Space (1)

- The Concept-Subconcept Relationship Space

<table>
<thead>
<tr>
<th></th>
<th># of isA pairs</th>
<th>Avg # of children</th>
<th>Avg # of parents</th>
<th>Avg level</th>
<th>Max level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probase</td>
<td>4,539,176</td>
<td>7.53</td>
<td>2.33</td>
<td>1.086</td>
<td>7</td>
</tr>
<tr>
<td>WordNet</td>
<td>283,070</td>
<td>11.0</td>
<td>2.4</td>
<td>1.265</td>
<td>14</td>
</tr>
<tr>
<td>WikiTaxonomy</td>
<td>90,739</td>
<td>3.7</td>
<td>1.4</td>
<td>1.483</td>
<td>15</td>
</tr>
<tr>
<td>YAGO</td>
<td>366,450</td>
<td>23.8</td>
<td>1.04</td>
<td>1.063</td>
<td>18</td>
</tr>
<tr>
<td>Freebase</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
IsA Relationship Space (2)

- The Concept-Instance Relationship Space

![Concept Size Distribution in Probase vs. Freebase](chart.png)

- Concept Size Distribution in Probase v.s. Freebase

5/13/2019
Precision of the Extracted Pairs

- 92.4% precision in average over the 40 benchmark concepts.
Outline

- Overview
- Iterative Extraction
- Taxonomy Construction
- Probabilistic Modeling
- Evaluation
- Conclusion
Conclusion

- We present a novel iterative extraction framework to extract the isA relationships from text.

- We present a novel taxonomy construction framework based on merging concepts by their senses.

- We use the above techniques to build Probase, which is currently the largest taxonomy in terms of concepts.

- We present a novel probabilistic approach to model the plausibility and typicality of the facts in Probase, and demonstrate its effectiveness in important text understanding applications.
Thank you 😊

Please visit our website:
http://research.microsoft.com/probase/
for more information about Probase!
**Algorithm Outline (Extraction)**

- **Input**: $S$, the set of sentences matching Hearst Patterns
- **Output**: $\Gamma$, the set of *isA* pairs

Repeat

\[
\text{foreach } s \text{ in } S \text{ do}
\]

\[
X_s, Y_s \leftarrow \text{SyntacticExtraction}(s);
\]

\[
\text{if } |X_s| > 1: X_s \leftarrow \text{SuperConceptDetection}(X_s, Y_s, \Gamma);
\]

\[
\text{if } |X_s| = 1: Y_s \leftarrow \text{SubConceptDetection}(X_s, Y_s, \Gamma);
\]

\[
\text{add valid } \text{isA} \text{ pairs to } \Gamma;
\]

end

Until no new pairs added into $\Gamma$;

Return $\Gamma$;
Syntactic Extraction

- Challenges
  - ... animals other than dogs such as cats ...
  - ... classic movies such as Gone with the Wind ...
  - ... companies such as IBM, Nokia, Proctor and Gamble ...

- Strategy
  - Use “,” as the delimiter to obtain the candidates.
  - For the last element, also use “and” and “or” to break it down.
Super-Concept Detection

• Find the most likely super-concept among the candidates.

\[ r(x_1, x_2) = \frac{p(x_1 | Y_s)}{p(x_2 | Y_s)} = \frac{p(Y_s | x_1)p(x_1)}{p(Y_s | x_2)p(x_2)} \]

Pick \( x_i \) if \( r(x_i, x_2) > \varepsilon \)

Assuming independence of \( y_i \)'s

\[ r(x_1, x_2) = \frac{p(x_1) \prod_{i=1}^{n} p(y_i | x_1)}{p(x_2) \prod_{i=1}^{n} p(y_i | x_2)} \]

1) \( Y_s \) is the set of sub-concepts of the sentence \( s \).
2) \( p (y_i | x_i) = p(x_i, y_i) / p(x_i) = n(x_i, y_i) / n(x_i) \).

We maintain a count \( n(x, y) \) for each \( (x, y) \) in \( \Gamma \).
Super-Concept Detection (Ex)

\[ r(x_1, x_2) = \frac{p(x_1 | Y_s)}{p(x_2 | Y_s)} = \frac{p(Y_s | x_1) p(x_1)}{p(Y_s | x_2) p(x_2)} \]

\[ p(y_i | x_i) = p(x_1, y_i) / p(x_1) = n(x_1, y_i) / n(x_1) \]

\( r \) (companies, oil companies)

Diagram:

- Syntactic Extraction
- Knowledge \( \Gamma \)
- Super-concept Detection
- Sub-concept Detection

Companies: IBM, Walmart, Proctor, Gamble, Proctor and Gamble
Sub-Concept Detection (1)

- Find the valid sub-concepts among the candidates.

Observation 1. The closer a candidate sub-concept is to the pattern keywords, the more likely it is a valid sub-concept.

Observation 2. If we are certain a candidate sub-concept at the $k$-th position from the pattern keywords is valid, then most likely candidate sub-concepts from position 1 to position $k-1$ are also valid.

E.g., ... representatives in North America, Europe, the Middle East, Australia, Mexico, Brazil, Japan, China, and other countries.
Sub-Concept Detection (2)

Strategy

- Find the largest scope wherein sub-concepts are all valid:
  \[ \text{find the maximum } k \text{ s.t. } p(y_k \mid x) > \varepsilon' \]
- Address the ambiguity issues inside the scope \( y_1, \ldots, y_k \):

\[
r(c_1, c_2) = \frac{p(c_1 \mid x, y_1, \Lambda, y_{j-1})}{p(c_2 \mid x, y_1, \Lambda, y_{j-1})}
\]

Suppose that \( y_j \) is ambiguous with two candidates \( c_1 \) and \( c_2 \).

Assuming independence of \( y_i \)'s

\[
r(c_1, c_2) = \frac{p(c_1 \mid x) \prod_{i=1}^{j-1} p(y_i \mid c_1, x)}{p(c_2 \mid x) \prod_{i=1}^{j-1} p(y_i \mid c_2, x)}
\]

Pick \( c_1 \) if \( r(c_1, c_2) > \varepsilon'' \)
Sub-Concept Detection (Ex)

\[ r(c_1, c_2) = \frac{p(c_1 \mid x, y_1, \Lambda, y_{j-1})}{p(c_2 \mid x, y_1, \Lambda, y_{j-1})} \]

\( r \) (Proctor and Gamble, Proctor)
Properties of “Such As” (1)

Property 1. Let $s = \{(x, y_1), \ldots, (x, y_n)\}$ be the isA pairs derived from a sentence. Then, all the $x$’s in $s$ have a unique sense, that is, there exists a unique $i$ such that $(x, y_j) \models (x^i, y_j)$ holds for all $1 \leq j \leq n$.

- Example:
  - ... plants such as trees and grass ...
  - ... plants such as steam turbines, pumps, and boilers ...

But sentences like “... plants such as trees and boilers ...” are extremely rare.
Properties of “Such As” (2)

Property 2. Let \( \{(x^i, y_1), \ldots, (x^i, y_m)\} \) denote pairs from one sentence, and \( \{(x^j, z_1), \ldots, (x^j, z_n)\} \) from another sentence. If \( \{y_1, \ldots, y_m\} \) and \( \{z_1, \ldots, z_n\} \) are similar, then it is highly likely that \( x^i \) and \( x^j \) are equivalent, that is, \( i = j \).

- Example:
  a) ... plants such as trees and grass ...
  b) ... plants such as trees, grass and herbs ...

The “plants” in a) and b) are highly likely to have the same sense.
Properties of “Such As” (3)

Property 3. Let \{(x^i, y), (x^i, u_1), ..., (x^i, u_m)\} denote pairs obtained from one sentence, and \{(y^k, v_1), ..., (y^k, v_n)\} from another sentence. If \{u_1, u_2, ..., u_m\} and \{v_1, v_2, ..., v_n\} are similar, then it is highly likely that \((x^i, y) |= (x^i, y^k)\).

Example:

a) ... organisms such as plants, trees, grass and animals ...
b) ... plants such as trees, grass, and shrubs ...
c) ... plants such as steam turbines, pumps, and boilers ...

The “plants” in a) and b) are highly likely to have the same sense, but not the “plants” in a) and c).
Local Taxonomy

- Based on Property 1
Horizontal Merge

- Based on Property 2
Vertical Merge (1)

- Single Sense Alignment (Based on Property 3)
Vertical Merge (2)

- Multiple Sense Alignment (Based on Property 3)
Similarity Function

- We favor the similarity $f(A, B)$ to be measured by the absolute overlap of the two sets $A$ and $B$.
  - Similarity based on relative overlap such as Jaccard similarity will raise weird results (see the paper for an example).

- More generally, the similarity function is desired to have the following closure property:

  **Property 4.** If $A, A', B,$ and $B'$ are any sets s. t. $A \subseteq A'$ and $B \subseteq B'$, then $\text{Sim}(A, B) \Rightarrow \text{Sim}(A', B')$. 
Algorithm Outline (Construction)

- **Input**: $S$, the set of sentences with extracted *isA* pairs
- **Output**: $T$, the taxonomy graph

**Stage 1**: For each $s$ in $S$, construct a *local taxonomy*.

**Stage 2**: Perform all possible *horizontal* merges.

**Stage 3**: Perform all possible *vertical* merges.

**Return** the graph $T$ after the 3 stages
Theoretical Results

**Theorem 1.** Let $T$ be a set of local taxonomies. Let $O^\alpha$ and $O^\beta$ be any two sequences of horizontal and vertical merge operations on $T$. Assume no further operations can be performed on $T$ after $O^\alpha$ or $O^\beta$. Then, the final graph after performing $O^\alpha$ and the final graph after performing $O^\beta$ are identical.

**Theorem 2.** Let $O$ be the set of all possible sequences of operations, and let $M = \min\{|O| : O \in O\}$. Suppose $O^\sigma$ is the sequence that performs all possible horizontal merges first and all possible vertical merges next, then $|O^\sigma| = M$. 
Applications of Typicality (1)

- Semantic Web Search

ACM fellows working on semantic web

database conferences in asian cities

Are you interested in the text or instances of “ACM fellows”, “database conferences” and “asian cities”? 
Applications of Typicality (2)

- Short Text Understanding (Y. Song et al. *IJCAI’11*)
  - Conceptualize from a set of words by performing Bayesian analysis based on the (inverse) typicality $T(x|i)$.

<table>
<thead>
<tr>
<th>Example:</th>
</tr>
</thead>
<tbody>
<tr>
<td>India =&gt; country / region</td>
</tr>
<tr>
<td>India, China =&gt; Asian country / developing country</td>
</tr>
<tr>
<td>India, China, Brazil =&gt; BRIC / emerging market</td>
</tr>
</tbody>
</table>

- Cluster Twitter messages based on conceptualization signals of words.
Concept Space (1)

- Probase contains more then 2.6 million concepts. Are they useful?

- Evaluate this using the top 50 million popular queries in Bing’s query log from a 2-year period.

- Metrics in the evaluation
  - Relevance
  - Taxonomy Coverage
  - Concept Coverage
Concept Space (2)

- Relevance: A concept is relevant if it appears at least once.
Concept Space (3)

- Taxonomy Coverage: A query is covered if it contains at least one concept or instance in the taxonomy.

![Graph showing the number of queries covered by different taxonomies across different top k queries.](image_url)
Concept Space (4)

- Concept Coverage: A query is covered if it contains at least one concept in the taxonomy.

![Chart showing the number of queries for different taxonomies over a range of top k queries.]