Conceptualization for Short Text Understanding

Zhongyuan Wang (王仲远)

*Joint work with Haixun Wang, Jun Yan, Yanghua Xiao, Ji-Rong Wen, and many interns
Short Text

- Search
- Ad keywords
- Anchor text
- Document Title
- Caption
- Question

Short text is *sparse, noisy, and ambiguous*
The big question

• How does the mind get so much out of so little?

• Our minds build rich models of the world and make strong generalizations from input data that is *sparse, noisy, and ambiguous* — in many ways far too limited to support the inferences we make.

• How do we do it?
How to Grow a Mind: Statistics, Structure, and Abstraction

Joshua B. Tenenbaum,1,2 Charles Kemp,2 Thomas L. Griffiths,3 Noah D. Goodman4

MIT, CMU, Berkeley, Stanford
If the mind goes beyond the data given, *another source of information* must make up the difference.
Knowledge Base Efforts

http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.png
1. “Python Tutorial”
2. “Who was the U.S. President when the Angels won the World Series?”

Semantic Network: knowledge → question (internal representation) → understanding → answer

knowledgebase
## Semantic Network vs. Knowledgebase

<table>
<thead>
<tr>
<th>Semantic Network</th>
<th>Knowledgebase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common/linguistic knowledge</td>
<td>Entities</td>
</tr>
<tr>
<td>isA</td>
<td>Facts</td>
</tr>
<tr>
<td>isPropertyOf</td>
<td></td>
</tr>
<tr>
<td>co-occurrence</td>
<td>DayOfBirth</td>
</tr>
<tr>
<td></td>
<td>LocatedIn</td>
</tr>
<tr>
<td></td>
<td>SpouseOf</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Typicality, basic level of</td>
<td>Black or White</td>
</tr>
<tr>
<td>categorization</td>
<td>Precision</td>
</tr>
<tr>
<td>知能之All, Probase</td>
<td>Freebase, Yago</td>
</tr>
</tbody>
</table>
Probase: A Semantic Network

http://research.microsoft.com/probase/

**Nodes:**

- Concepts: 
  - (“Spanish Artists”)
- Entities: 
  - (“Pablo Picasso”)
- Attributes: 
  - (“Birthday”)
- Verbs/Adjectives: 
  - (“Eat”, “Sweet”)

**Edges:**

- isA: 
  - (concept, entities)
- isPropertyOf: 
  - (attributes)
- Co-occurrence: 
  - (isCEOof, LocatedIn, etc)
Probase Concepts (2.7 million+)

- Basic watercolor techniques
- Celebrity wedding dress designers

Probase isA error rate: <1% and <10% for random pair
Research Roadmap

Conceptualization layer

- Term Similarity
- Short Text Similarity
- Head-Modifier Detection for Short Texts

Focus on sense annotation and disambiguation

Application layer

- NER
- Text Annotation
- Ads semantic match
- Query Recommendation
- Text/table Understanding

Conceptualization layer

Semantic network layer

The Semantic Network: Probase

- VLDB’11
- SIGMOD’12, ICDE’13
- EDBT’14, TACL’14, TKDE’15

SIGKDD’12
WSDM’13
CIKM’14

IJCAI’11, 13, 15
CIKM’13, 15
ICDE’14, 15
(Best paper)
What is short text understanding?
Add Common Sense to Computing

Pablo Picasso | 25 Oct 1881 | Spanish
China     Brazil     India

emerging market
The engineer is eating an apple.
Conceptualization On Knowledge Engineering (COKE)

- **Conceptualization**: An *explicit* representation for the *short text*

  \[ P(\text{concept} \mid \text{short text}) \]

  - a domain millions of concepts used in day to day communication
  - search query, anchor text, twitter, ads keywords, ...

- **Short text** is *sparse, noisy, and ambiguous*

- **Explicit** means
  - Conceptualization results can be *easily understood* by human beings
  - Conceptualization model can be *easily customized* for different scenarios
Conceptualization On Knowledge Engineering (COKE)

- Conceptualization: An explicit representation for the short text

```
ShortText: pear apple
Show Parameters
Elapsed Time = 00:00:00.140014

25/fruit          25/fruit
|    pear    |    apple    |
fruit   | 0.572469    | 0.571808    |
fresh fruit | 0.0562256   | 0.0396546   |
true fruit  | 0.03260949  | 0.01607355  |
dried fruit | 0.01165293  | 0.01593271  |
seasonal fruit | 0.01160144  | 0.01593194  |
juice    | 0.01052348  | 0.01546959  |
hard fruit | 0.01011644  | 0.01309386  |
climacteric fruit | 0.009254614 | 0.01297349  |
fruit juice | 0.000948666 | 0.10036025  |
sweet fruit | 0.006294312 | 0.01033749  |
9405/food | 0.1058999   | 0.1241537   |
food     | 0.03120783  | 0.06844533  |
high fiber food | 0.008663075 | 0.01628461  |
ingredient | 0.007349517 | 0.00757149  |
high fiber food | 0.004695997 | 0.004435756 |
fresh food | 0.004608723 | 0.004405118 |
hard food  | 0.003553713 | 0.003733333 |
```

```
ShortText: ipad apple
Show Parameters
Elapsed Time = 00:00:00.6670667

15/mobile device/device          1/technology company/company
|                  |                  |
ipad               | apple            |
mobile device      | 0.8072805       | technology company | 0.9623328 |
apple device       | 0.01989746      | computer manufacturer | 0.005182603 |
tablet device      | 0.01723156      | tech company        | 0.00650604 |
tablet device      | 0.01718074      | innovative company  | 0.004826203 |
ios device         | 0.01706666      | computer company    | 0.00475833 |
gadget             | 0.01549841      | tech gig            | 0.00445252 |
handheld device    | 0.0105837      | technology gig       | 0.00445252 |
digital device     | 0.01037961      | successful company  | 0.00621911 |
wireless device    | 0.00999995      | tech stock           | 0.00441483 |
software company   | 0.00118703      |                      | 0.00313031 |
```

“pear apple”

“ipad apple”
Recap: Conceptualization

Conceptualization 1.0
[IJCAI'11, CIKM'13]:
*mapping terms to concept space based on Bayesian Inference*

Conceptualization 2.0
[ICDE’14, CIKM’14, ICDE’15(Best Paper)]:
*incorporating co-occurrence network*

Conceptualization 2.5
[IJCAI’15]: *leveraging verbs, adjective, attribute, etc.*

Conceptualization 3.0
[CIKM’15]: *learning-based conceptualization/leverage embedding*

Production Impacts (Shippings):

- Ads relevance (2012)
- MSN Query Recommendation (2012)
- Bing Image Search (2013)
- Table understanding in Power Query (2013)
- Definition Answer in EQnA (2014, 2015)
What we resolved?

• Short Text Understanding

Diagram:

- DNN Tool
- Python Tutorial
  - Head
  - Modifier
  - Single Instance
  - adjective, verb, ...
  - Context

- language
- programming language
- snake
- animal
Short Text Understanding

- If the short text is a **single instance**...
  - *SIGMOD 2012, CIKM 2015*

- If the short text has **context** for the instance...
  - *IJCAI 2011/2013, ICDE 2015*

- If the short text contains **verb, adjective**...
  - *IJCAI 2015*

- If the short text contains **multiple instance**...
  - *ICDE 2014*

- **Applications**
  - *WSDM 2013, CIKM 2013/2014*
If the short text is a single instance...

“Python”

• Zhongyuan Wang, Haixun Wang, Ji-Rong Wen, and Yanghua Xiao, *An Inference Approach to Basic Level of Categorization*, in *ACM International Conference on Information and Knowledge Management (CIKM)*, October 2015.
Statistics of Search Queries

(a) By traffic

- 1 Entity: 44%
- 2 Entities: 29%
- 3 Entities: 17%
- 4 Entities: 7%
- 5 Entities: 2%
- More than 5 Entities: 1%

(b) By # of distinct queries

- 1 Entity: 26%
- 2 Entities: 19%
- 3 Entities: 10%
- 4 Entities: 7%
- 5 Entities: 4%
- More than 5 Entities: 1%

Entity 1

Entity 2
A Concept View of “Microsoft”
Basic-level Conceptualization (BLC)

<table>
<thead>
<tr>
<th>Category Level</th>
<th>Informative?</th>
<th>Distinctive?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superordinate</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Basic-level</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Subordinate</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

- **Basic-level conceptualization**

software company

company ... ... largest desktop OS vendor

- KFC
- BMW
- Microsoft
Using $Rep(e, c)$ for BLC

- **Our measure** $Rep(e, c) = P(c|e) \times P(e|c)$ means:
  
  Given $e$, the $c$ should be its typical concept *(shortest distance)*
  
  Given $c$, the $e$ should be its typical entity *(shortest distance)*

  A process of finding *concept nodes* having shortest expected distance with $e$

- **(With PMI)** If we take the logarithm of our scoring function, we get:

  $$\log Rep(e, c) = \log P(c|e) \times P(e|c) = \log \frac{P(e,c)}{P(e)} \times \frac{P(e,c)}{P(c)} = \log \frac{P(e,c)^2}{P(e)P(c)} = PMI(e, c) + \log P(e,c) = PMI^2$$

- **(With Commute Time)** The commute time between an instance $e$ and a concept $c$ is:

  $$Time(e, c) = \sum_{k=1}^{\infty} (2k) * P_k(e, c) = \sum_{k=1}^{T} (2k) * P_k(e, c) + \sum_{k=T+1}^{\infty} (2k) * P_k(e, c)$$

  $\geq \sum_{k=1}^{T} (2k) * P_k(e, c) + 2(T + 1) * (1 - \sum_{k=1}^{T} P_k(e, c)) = 4 - 2 * Rep(e, c)$
Precision@K & NDCG@K

• **Metrics**
  
  • $\text{Precision@K} = \frac{\sum_{i=1}^{K} rel_i}{K}$ (for correctness of concepts)
  
  • $\text{nDCG}_K = \frac{\text{rel}_1 + \sum_{i=2}^{K} \frac{\text{rel}_i}{\log i}}{\text{ideal}_\text{rel}_1 + \sum_{i=2}^{K} \frac{\text{ideal}_\text{rel}_i}{\log i}}$ (for ranking of concepts)

• **Results**

<table>
<thead>
<tr>
<th>Metric</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI(e)</td>
<td>0.769</td>
<td>0.692</td>
<td>0.705</td>
<td>0.685</td>
<td>0.719</td>
<td>0.705</td>
<td>0.690</td>
</tr>
<tr>
<td>PMI_p(e)</td>
<td>0.885</td>
<td>0.769</td>
<td>0.756</td>
<td>0.800</td>
<td>0.754</td>
<td><strong>0.733</strong></td>
<td><strong>0.721</strong></td>
</tr>
<tr>
<td>NPMI(e)</td>
<td>0.692</td>
<td>0.692</td>
<td>0.667</td>
<td>0.638</td>
<td>0.627</td>
<td>0.610</td>
<td>0.610</td>
</tr>
<tr>
<td>Typicality P(c</td>
<td>e)</td>
<td>0.462</td>
<td>0.577</td>
<td>0.603</td>
<td>0.577</td>
<td>0.569</td>
<td>0.564</td>
</tr>
<tr>
<td>Typicality P(e</td>
<td>c)</td>
<td>0.500</td>
<td>0.462</td>
<td>0.526</td>
<td>0.523</td>
<td>0.523</td>
<td>0.510</td>
</tr>
<tr>
<td>Rep(e)</td>
<td>0.846</td>
<td><strong>0.865</strong></td>
<td><strong>0.872</strong></td>
<td><strong>0.862</strong></td>
<td><strong>0.758</strong></td>
<td>0.731</td>
<td>0.719</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI(e)</td>
<td>0.516</td>
<td>0.531</td>
<td>0.519</td>
<td>0.531</td>
<td>0.562</td>
<td>0.574</td>
<td>0.594</td>
</tr>
<tr>
<td>PMI_p(e)</td>
<td>0.725</td>
<td>0.664</td>
<td>0.652</td>
<td>0.660</td>
<td>0.628</td>
<td>0.631</td>
<td>0.646</td>
</tr>
<tr>
<td>NPMI(e)</td>
<td>0.599</td>
<td>0.597</td>
<td>0.579</td>
<td>0.554</td>
<td>0.540</td>
<td>0.539</td>
<td>0.549</td>
</tr>
<tr>
<td>Typicality P(c</td>
<td>e)</td>
<td>0.297</td>
<td>0.380</td>
<td>0.409</td>
<td>0.422</td>
<td>0.438</td>
<td>0.446</td>
</tr>
<tr>
<td>Typicality P(e</td>
<td>c)</td>
<td>0.401</td>
<td>0.386</td>
<td>0.396</td>
<td>0.398</td>
<td>0.401</td>
<td>0.410</td>
</tr>
<tr>
<td>Rep(e)</td>
<td><strong>0.758</strong></td>
<td><strong>0.771</strong></td>
<td><strong>0.745</strong></td>
<td><strong>0.723</strong></td>
<td><strong>0.656</strong></td>
<td><strong>0.647</strong></td>
<td><strong>0.661</strong></td>
</tr>
</tbody>
</table>

• Overall, our measure $\text{Rep}$ performs well in both Precision and NDCG.
• Most important, it’s well interpreted in theory.
If the short text has context for the instance...

“Python Tutorial”

- Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou, Short Text Understanding Through Lexical-Semantic Analysis, in International Conference on Data Engineering (ICDE), April 2015. (Best Paper Award)
- Dongwoo Kim, Haixun Wang, and Alice Oh, Context-Dependent Conceptualization, in IJCAI, 2013.
- Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen, Short Text Conceptualization using a Probabilistic Knowledgebase, in IJCAI, 2011.
Conceptualization Framework (ICDE 2015 Best Paper)

Co-occurrence network

Parsing
Term clustering by isA
Concept filtering by co-occurrence
Head/modifier analysis
Concept orthogonalization

Is-A network

短文本

Conceptualization

Concept Vector

Is-A network

“ipad apple”

ipad

apple

device

product

co-occur

filtering

company

brand

food

fruit

product

...
If the short text contains verb, adjective...

“Dangerous Python”

- Zhongyuan Wang, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen, Query Understanding through Knowledge-Based Conceptualization, in *IJCAI*, 2015.
• Watch *Harry Potter*
• Read *Harry Potter*
Mining Lexical Relationships

• Lexical knowledge represented by the probabilities

\[ p(\text{instance}|\text{watch}) \]

\[ p(\text{verb}|\text{watch}) \]

\[ p(\text{book}|\text{harry potter}) \]

\[ p(\text{movie}|\text{harry potter}) \]

\[ p(\text{watch}|\text{harry potter}) \]

\[ p(\text{movie}|\text{watch, verb}) \]

\[ p(\text{book}|\text{harry potter}) \]

\[ p(\text{movie}|\text{harry potter}) \]

\[ p(c|e) = p(c|t, z = \text{instance}) \]

\[ p(c|t, z) \]

\[ e: \text{instance} \]

\[ t: \text{term} \]

\[ c: \text{concept} \]

\[ z: \text{role} \]
Deriving Probabilities

• Deriving $p(z|t)$: $p(z|t) = \frac{n(t,z)}{n(t)}$

• Deriving $P(c|t, z)$
  
  • Case 1: $z=$instance
    $P(c|t, z = \text{instance}) = p(c|e)$
  
  • Case 2: $z=$attribute
    $P(c|t, z = \text{attribute}) = p(c|a)$
  
  • Case 3: $z=$verb
    
    $P(c|t, z = \text{verb}) = \sum_{e \in c} p(e, c|t, z = \text{verb}) = \sum_{e \in c} p(c|e) \times p(e|t, z = \text{verb})$
  
  • Case 4: $z=$adjective
    
    $P(c|t, z = \text{adjective}) = \sum_{e \in c} p(c|e) \times p(e|t, z = \text{adjective})$
Constructing an offline semantic network

\[ P(c|t) = \sum_z p(c|t, z) \times p(z|t) \]

\[ P(c_2|c_1) = \frac{\sum_{e_i \in c_1, e_j \in c_2} n(e_i, e_j)}{\sum_c \sum_{e_i \in c_1, e_j \in c} n(e_i, e_j)} \]
Understanding Queries

• **Goal**: to rank the concepts and find:

\[ \arg \max_c p(c|t, q) \]

The offline semantic network

Random walk with restart [Sun et al., 2005] on the online subgraph
If the short text contains multiple instance...

“DNN Tool Python”

Mining Concept Patterns

Building concept pattern dictionary:

Get entity pairs from query log

Conceptualization

Concept Patterns for each prepositions

Query Logs

- cover for iphone 5s
- battery for sony a7r
- wicked on broadway

Extract Patterns

- A for B, A of B,
- A with B, A in B,
- A on B, A at B ...

entity1/head

entity2/constraint

concept1,1

concept1,2

concept1,3

concept1,4

concept2,1

concept2,2

concept2,3

(entity1, concept1,2), (concept1,1, concept2,2)

(entity1, concept1,1, concept2,3)

Concept Pattern Dictionary
Why Concepts Can’t Be Too General

- It may cause too many concept pattern conflicts: can’t distinguish head and modifier for general concept pairs

<table>
<thead>
<tr>
<th>Derived Concept Pattern</th>
<th>Head</th>
<th>Modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>device</td>
<td>company</td>
</tr>
<tr>
<td>Supporting Entity Pairs</td>
<td>iphone 4</td>
<td>verizon</td>
</tr>
<tr>
<td></td>
<td>modem</td>
<td>comcast</td>
</tr>
<tr>
<td></td>
<td>wireless router</td>
<td>comcast</td>
</tr>
<tr>
<td></td>
<td>iphone 4</td>
<td>tmobile</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Derived Concept Pattern</th>
<th>Head</th>
<th>Modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>company</td>
<td>device</td>
</tr>
<tr>
<td>Supporting Entity Pairs</td>
<td>amazon books</td>
<td>kindle</td>
</tr>
<tr>
<td></td>
<td>netflix</td>
<td>touchpad</td>
</tr>
<tr>
<td></td>
<td>skype</td>
<td>windows phone</td>
</tr>
<tr>
<td></td>
<td>netflix</td>
<td>ps3</td>
</tr>
</tbody>
</table>
Why Concepts Can’t Be Too Specific

• It may generate concepts with less representation

<table>
<thead>
<tr>
<th>...</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>device</td>
<td>largest desktop OS vendor</td>
</tr>
<tr>
<td>device</td>
<td>largest software development company</td>
</tr>
<tr>
<td>device</td>
<td>largest global corporation</td>
</tr>
<tr>
<td>device</td>
<td>latest windows and office provider</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• Concept level may regress to entity level
  • Large storage space: up to (million * million) patterns

We should use Basic-level Conceptualization (BLC)
<table>
<thead>
<tr>
<th>Cluster size</th>
<th>Sum of Cluster Score</th>
<th>head;modifier;score</th>
</tr>
</thead>
<tbody>
<tr>
<td>615</td>
<td>21146.91</td>
<td>breed;state;3572.98460224501</td>
</tr>
<tr>
<td>296</td>
<td>7752.357</td>
<td>game;platform;627.40347671856</td>
</tr>
<tr>
<td>153</td>
<td>3466.804</td>
<td>accessory;vehicle;533.937050949809</td>
</tr>
<tr>
<td>70</td>
<td>1182.59</td>
<td>browser;platform;132.612807637391</td>
</tr>
<tr>
<td>22</td>
<td>1010.993</td>
<td>requirement;school;271.407526294823</td>
</tr>
<tr>
<td>34</td>
<td>948.9159</td>
<td>drug;disease;154.602405333541</td>
</tr>
<tr>
<td>42</td>
<td>899.2995</td>
<td>cosmetic;skin condition;81.4659415003929</td>
</tr>
<tr>
<td>16</td>
<td>742.1599</td>
<td>job;city;279.0372555528</td>
</tr>
<tr>
<td>32</td>
<td>710.403</td>
<td>accessory;phone;246.513830851194</td>
</tr>
<tr>
<td>18</td>
<td>669.2376</td>
<td>software;platform;210.126322725878</td>
</tr>
<tr>
<td>20</td>
<td>644.4603</td>
<td>test;disease;239.774028397537</td>
</tr>
<tr>
<td>27</td>
<td>599.4205</td>
<td>clothes;breed;98.73396282851</td>
</tr>
<tr>
<td>19</td>
<td>591.3545</td>
<td>penalty;crime;200.544192793488</td>
</tr>
<tr>
<td>25</td>
<td>584.8804</td>
<td>tax;state;240.0818612579</td>
</tr>
<tr>
<td>16</td>
<td>546.5424</td>
<td>sauce;meat;183.95286321553</td>
</tr>
<tr>
<td>18</td>
<td>480.9389</td>
<td>credit card;country;142.91908792152</td>
</tr>
<tr>
<td>14</td>
<td>473.0792</td>
<td>food;holiday;145.54140330924</td>
</tr>
<tr>
<td>11</td>
<td>453.6199</td>
<td>mod;game;257.163856882439</td>
</tr>
<tr>
<td>29</td>
<td>435.0954</td>
<td>garment;sport;47.1533326845442</td>
</tr>
<tr>
<td>23</td>
<td>399.4886</td>
<td>career information;professional;73.2726483731257</td>
</tr>
<tr>
<td>15</td>
<td>386.065</td>
<td>song;instrument;128.189481818135</td>
</tr>
<tr>
<td>18</td>
<td>378.213</td>
<td>bait;fish;78.0426541113169</td>
</tr>
<tr>
<td>22</td>
<td>372.2948</td>
<td>study guide;book;50.8339765053921</td>
</tr>
<tr>
<td>19</td>
<td>340.8953</td>
<td>plugins;browser;55.032607627126</td>
</tr>
<tr>
<td>14</td>
<td>330.5753</td>
<td>recipe;meat;88.2779863422951</td>
</tr>
<tr>
<td>18</td>
<td>321.4226</td>
<td>currency;country;110.825444188352</td>
</tr>
<tr>
<td>13</td>
<td>318.0272</td>
<td>lens;camera;186.081673263957</td>
</tr>
<tr>
<td>9</td>
<td>316.973</td>
<td>decoration;holiday;130.055844126533</td>
</tr>
<tr>
<td>16</td>
<td>314.875</td>
<td>food;animal;73.3854366514</td>
</tr>
</tbody>
</table>

**Detection**

<table>
<thead>
<tr>
<th>Game (Head)</th>
<th>Platform (Modifier)</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry birds</td>
<td>android</td>
</tr>
<tr>
<td>angry birds</td>
<td>ios</td>
</tr>
<tr>
<td>angry birds</td>
<td>windows 10</td>
</tr>
</tbody>
</table>
Combine Explicit and Implicit Semantics (First Attempt):
Learning-based Conceptualization

- Contextual Text Understanding in Distributional Semantic Space (CIKM2015)
Previous Conceptualization Framework (ICDE 2015 Best Paper)

Model Size: \( \sim 100 \text{GB} \)

Concept Vector

\[
\begin{pmatrix}
    c_1, p_1 \\
    c_2, p_2 \\
    c_3, p_3 \\
    \vdots
\end{pmatrix}
\]

Short Text

Embedding?

Conceptualization

Is-A network

Parsing

Term clustering by isA

Concept filtering by co-occurrence

Head/modifier analysis

Concept orthogonalization

“ipad apple”

“ipad”

“apple”

Is-A

product

company

brand

device

food

fruit

co-occur

filtering

Is-A

Embedding?
Basic Idea of Learning-based Conceptualization

• **Target**: best representation for the short text:

```
Conceptualization + Embedding = ?
```

• **Learning-based Conceptualization**: given a term $t$, with its context $q$, we want to find the probability of concept $c$

$$f_{conceptualization}(c, t, q) = p(c|t) \cdot p(c|q) = p(c|t) \cdot \frac{\cos(q, c)}{\sum_{c_i \in c(t)} \cos(q, c_i)}$$

**Traditional conceptualization** based on **Is-A network**

**Contextual knowledge** based on **Embedding**
Learning-based Conceptualization Framework

Word-Concept Joint Embedding

Model Size: From 100GB to ~1GB

- Parsing
- Term clustering by isA
- Concept filtering by co-occurrence
- Head/modifier analysis
- Concept orthogonalization

Short Text

Conceptualization

- Is-A network
- Concepts with scores in context (Explicit Representation)
- Word vectors in context (Implicit Representation)
- Text vectors in context (Implicit Representation)

Key Problems:
1. Joint word-concept embedding: project **words** and **concepts** to the same semantic space
2. Refined **concept inference process**
Joint Word-Concept Embedding

- **Word embedding – Skip gram model**
  \[ \sum_{i=1}^{N} \sum_{o=i-c}^{i+c} \log P(w_o|w_i) \]

- **How to incorporate concept**
  - Concept as input
    \[ \sum_{i=1}^{N} \sum_{o=i-c}^{i+c} \log P(w_o|c_i) \]
  - Concept as output
    \[ \sum_{i=1}^{N} \sum_{o=i-c}^{i+c} \log P(c_o|w_i) \]
  - Concept as both input and output
    \[ \sum_{i=1}^{N} \sum_{o=i-c}^{i+c} \log P(c_o|c_i) \]
Approach 1: Parallel Joint-Embedding Models

- Assume conditionally independent between the word and concept

\[
\sum_{i=1}^{N} \sum_{o=i-c}^{i+c} \log P(w_o|w_i)P(c_o|w_i)
\]

PM-1

\[
\sum_{i=1}^{N} \sum_{o=i-c}^{i+c} \log P(w_o|w_i)P(w_o|c_i)
\]

PM-2

\[
\sum_{i=1}^{N} \sum_{o=i-c}^{i+c} \log P(w_o|w_i)P(c_o|w_i)P(c_o|c_i)
\]

PM-3
Approach 2: Generative Joint-Embedding Models

- Assume conditionally dependent between output word and output concept: A word is selected by firstly select the class it belongs to.

\[ \sum_{i=1}^{N} \sum_{t=i-c}^{i+c} \log P(w_o|w_i) = \sum_{i=1}^{N} \sum_{t=i-c}^{i+c} \log P(c_o|w_i) P(w_o|c_o) \]  

\[ \sum_{i=1}^{N} \sum_{t=i-c}^{i+c} \log P(w_o|w_i, c_i) = \sum_{i=1}^{N} \sum_{t=i-c}^{i+c} \log P(c_o|w_i) P(c_o|c_i) P(w_o|c_o) \]
Evaluation: Word Similarity in Context

• Evaluation Setting
  • **Public dataset**: Eric Huang et al. (2012) Improving Word Representations via Global Context and Multiple Word Prototypes.
  • **Task**: Given *word1*, *word2* and their contexts, compute \( \text{similarity}(\text{word1}, \text{word2}) \)

• Metric: the **Spearman’s correlation** between human judgement and embedding similarity score.

\[
\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}
\]
Word-Concept Embedding Evaluation Results

**Spearman’s correlation**

- **PM**: Parallel Joint-Embedding Model
- **GM**: Generative Joint-Embedding Model

**Non-contextual representations**

<table>
<thead>
<tr>
<th></th>
<th>PM-1</th>
<th>PM-2</th>
<th>PM-3</th>
<th>GM-1</th>
<th>GM-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.5</td>
<td>60</td>
<td>60.5</td>
<td>61</td>
<td>61.5</td>
<td>62</td>
</tr>
<tr>
<td>60</td>
<td>60.5</td>
<td>61</td>
<td>61.5</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>60.5</td>
<td>61</td>
<td>61.5</td>
<td>62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>61.5</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61.5</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conceptualization**

- PM-1
- PM-2
- PM-3
- GM-1
- GM-2
Conceptualization with Word-Concept Embedding

• **Concept Inference**: given a term \( t \), with its context \( q \), we want to find the probability of concept \( c \)

\[
f_{\text{conceptualization}}(c, t, q) = p(c|t) \cdot p(c|q) = p(c|t) \cdot \frac{\cos(q, \hat{c})}{\sum_{c_i \in c(t)} \cos(q, \hat{c}_i)}
\]

• **Conceptualization**: \( p(c|t, q) = \frac{f_{\text{conceptualization}}(c, t, q)}{\sum_{c_i \in c(t)} f_{\text{conceptualization}}(c_i, t, q)} \)

• **Example**: “apple, microsoft and google are world’s most valuable brands.”

\[
p(\text{fruit}|\text{apple}) \quad p(\text{company}|\text{apple})
\]

Concept “fruit”  Concept “company”

context

\[
p(\text{company}|\text{context}) \quad p(\text{fruit}|\text{context})
\]
Learning-based Conceptualization Evaluation Results

SCWS contextual word similarity dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Spearman’s correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>54</td>
</tr>
<tr>
<td>EH</td>
<td>64</td>
</tr>
<tr>
<td>PM-1</td>
<td>64</td>
</tr>
<tr>
<td>PM-2</td>
<td>64</td>
</tr>
<tr>
<td>PM-3</td>
<td>64</td>
</tr>
<tr>
<td>GM-1</td>
<td>66</td>
</tr>
<tr>
<td>GM-2</td>
<td>66</td>
</tr>
</tbody>
</table>

$W_{o1}$, $W_{o2}$, $c_{o1}$, $c_{o2}$, $w_i$

$f_{conceptualization}(c, t, q) = p(c|t) \times p(c|q)$

- **SG**: Skip-Gram (Word2Vec)
- **EH**: Eric Huang’s Sense Embedding
- **PM**: Parallel Joint-Embedding Model
- **GM**: Generative Joint-Embedding Model
Applications

• Short text understanding
• Short text similarity
• Ads/search semantic match
• Q/A system
• Query recommendation based on channels and articles
• Web table understanding
• ...
~Thank You~

http://research.microsoft.com/probase/

Contact: Zhongyuan Wang
(email: zhy.wang # microsoft.com )