Distributional Measures of Concept-Distance

A Task-oriented Evaluation

Saif Mohammad and Graeme Hirst
Department of Computer Science
University of Toronto

EMNLP, Sydney, Australia (22–23 July 2006)

Copyright ©2006, Saif Mohammad and Graeme Hirst
Concept-Distance

SALSA  DANCE

CLOWN  BRIDGE

Uses: machine translation, information retrieval, word sense disambiguation, correcting real-word spelling errors, . . .
Word-Distance

Words that co-occur strongly with salsa and dance
Words that co-occur strongly with *salsa* and *dance*
Semantic Measures of Concept-Distance

- Structure of a network or resource
  - The nodes represent senses or concepts

- Examples
  - MeSH: Rada et al. (1989)
Distributional Measures of Word-Distance

- Rely only on raw text
- Consider words with similar contexts close
  - Create distributional profiles (DPs)
    - Strength of association with co-occurring words
      - *salsa*: dance (.28), fun (.2), spicy (.18), shine (.1), chips (.07), ...
  - Measure distance between DPs
## Example Measures

### Strength of association
- conditional probability (cp)
- pointwise mutual information (pmi)

### DP distance
- $\alpha$-skew divergence (ASD)
- cosine (cos)
- Jensen–Shannon divergence (JSD)
- Lin (Lin)

### Typical combinations:
- **ASD** and *cp*
- cos and *cp*
- **JSD** and *cp*
- Lin and *pmi*
The Distributional Hypothesis

Words in similar contexts tend to be semantically related.

- Distributional measure as proxy for a semantic measure
The Distributional Hypothesis

Words in similar contexts tend to be semantically similar.

- Distributional measure as proxy for a semantic measure
- Word sense ambiguity reduces accuracy
Focus: DP of Concepts

- Different senses of a word
  - Different “company” or distributional profiles (DPs)
    - **SALSA (the dance)**: dance (.34), fun (.27), grace (.18), partner (.11), ...
    - **SALSA (the dip)**: chips (.38), tortilla (.31), tomato (.23), hot (.17), ...

- Use of distributional profile of concepts (DPCs)
  - Intuitive and useful
Capturing DPCs

- Method
  - Direct: sense-annotated data
  - Alternative: Mohammad and Hirst (EACL-2006)
    - Combining raw text and a knowledge source

- Sense inventory
  - Published thesaurus
Published Thesauri

- E.g., *Roget’s* (English), *Macquarie* (English), *Cilin* (Chinese), *Bunrui Goi Hyou* (Japanese)

- Vocabulary divided into about 1000 categories
  - Words in a category are closely related.
  - A category can be thought of as a very coarse-grained concept (Yarowsky, 1992).
    - Represents senses of the words in it

- One word, more than one category
  - *bark* in **ANIMAL NOISES** and **MEMBRANE**.
Precomputing Distances

Distributional word–word distance matrix
≈ $100,000 \times 100,000$

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>...</th>
<th>$w_j$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td>$m_{11}$</td>
<td>...</td>
<td>$m_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$w_i$</td>
<td>$m_{i1}$</td>
<td>...</td>
<td>$m_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
| $w_k$ | $m_{kj}$ | ... | ... | ...

WordNet-based concept-concept distance matrix
≈ $75,000 \times 75,000$

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>...</th>
<th>$c_j$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>$m_{11}$</td>
<td>...</td>
<td>$m_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$c_i$</td>
<td>$m_{i1}$</td>
<td>...</td>
<td>$m_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$c_k$</td>
<td>$m_{kj}$</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Distributional Measures of Concept-Distance. Saif Mohammad and Graeme Hirst. 13
Why a Thesaurus?

- Computational ease: concept–concept distance matrix is much smaller (roughly 1000 x 1000 i.e., 0.01%).

- Coarse senses: WordNet is much too fine grained.

- Availability: Thesauri are available in many languages.

- Words for a sense: Each sense can be represented unambiguously with a set of (possibly ambiguous) words.
**Method**

**Step 1. Creating DPCs**

Word–Category Co-occurrence Matrix (WCCM)

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>...</th>
<th>$c_j$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td>$m_{11}$</td>
<td>$m_{12}$</td>
<td>...</td>
<td>$m_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>$w_2$</td>
<td>$m_{21}$</td>
<td>$m_{22}$</td>
<td>...</td>
<td>$m_{2j}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$w_i$</td>
<td>$m_{i1}$</td>
<td>$m_{i2}$</td>
<td>...</td>
<td>$m_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **WCCM**: categories (thesaurus) vs. words (vocabulary)
- **Cell $m_{ij}$**: number of times word $w_i$ co-occurs with a word listed in category $c_j$
First Pass

- Cell (space, CELESTIAL BODY) incremented by 1
- Cell (space, CELEBRITY) incremented by 1
First Pass (continued)

$X$: star, nova, constellation, sun
Word–Category Matrix

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>...</th>
<th>BODY</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td>$m_{11}$</td>
<td>$m_{12}$</td>
<td>...</td>
<td>$m_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>$w_2$</td>
<td>$m_{21}$</td>
<td>$m_{22}$</td>
<td>...</td>
<td>$m_{2j}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>space</td>
<td>$m_{i1}$</td>
<td>$m_{i2}$</td>
<td>...</td>
<td>$m_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Distributional Measures of Concept-Distance.** Saif Mohammad and Graeme Hirst. 18
Contingency Table for $w$ and $c$

<table>
<thead>
<tr>
<th></th>
<th>$c$</th>
<th>$\neg c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w$</td>
<td>$n_{wc}$</td>
<td>$n_{w\neg}$</td>
</tr>
<tr>
<td>$\neg w$</td>
<td>$n_{\neg c}$</td>
<td>$n_{\neg \neg}$</td>
</tr>
</tbody>
</table>

Applying a statistic gives the strength of association

- Conditional probability
- Pointwise mutual information
Evidence for the Senses

SoA → CELESTIAL BODY

• space • • star • • •

SoA → CELEBRITY
Base WCCM

- Matrix created after the first pass of unannotated text
  - Noisy
  - Captures strong associations
- Words that occur close to a target word
  - Good indicators of intended sense
Second Pass

- Cell (space, CELESTIAL BODY) incremented by 1
- New, more accurate, bootstrapped WCCM
  - Word sense dominance (Mohammad and Hirst, EACL-2006)
Method

Step 2. Calculate Concept-Distance

- Two concepts are close if their DPs are close.
  - Strength of association between a concept and co-occurring words: bootstrapped WCCM

- Any distributional measure can now be used to measure concept-distance.
Example: cosine

Before: word-distance

\[ \text{Cos}_{cp}(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} P(w|w_1) \times P(w|w_2)}{\sqrt{\sum_{w \in C(w_1)} P(w|w_1)^2 \times \sqrt{\sum_{w \in C(w_2)} P(w|w_2)^2}}} \]

\( C(x) \): set of words that co-occur with word \( x \)

Now: concept-distance

\[ \text{Cos}_{cp}(c_1, c_2) = \frac{\sum_{w \in C(c_1) \cup C(c_2)} P(w|c_1) \times P(w|c_2)}{\sqrt{\sum_{w \in C(c_1)} P(w|c_1)^2 \times \sqrt{\sum_{w \in C(c_2)} P(w|c_2)^2}}} \]

\( C(x) \): set of words that co-occur with concept \( x \)
Evaluation

1. Rank Closeness of Word Pairs

- Automatic measures rank word pairs
  - From near-synonyms to unrelated

- Correlation with human ranking
  - Rubenstein and Goodenough (1965)
Concept-Distance Approach

\[
distance(star, film) = \min (distance(CELEBRITY, MOTION PICTURE), \\
\hspace{1cm} distance(CELEBRITY, THIN MEMBRANE), \\
\hspace{1cm} distance(CELESTIAL BODY, MOTION PICTURE), \\
\hspace{1cm} distance(CELESTIAL BODY, THIN MEMBRANE))
\]
Results

<table>
<thead>
<tr>
<th>Measure</th>
<th>Word Distance</th>
<th>Concept Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASD and cp</td>
<td>.45</td>
<td>.60</td>
</tr>
<tr>
<td>Cos and cp</td>
<td>.54</td>
<td>.69</td>
</tr>
<tr>
<td>JSD and cp</td>
<td>.48</td>
<td>.61</td>
</tr>
<tr>
<td>Lin and pmi</td>
<td>.52</td>
<td>.71</td>
</tr>
</tbody>
</table>

- WordNet-based measures: .78 to .84 (Hirst and Budanitsky, 2005)
- WordNet-based concept-distance > Distributional concept-distance > Distributional word-distance
2. Correct Real-Word Spelling Errors

Method (Hirst and Budanitsky, 2005):

- No semantically close neighbors: suspect
  ...
  \textit{interest} \textit{money} \textit{band} \textit{loan} \textit{deposit} ...

- Suspect has spelling variant semantically close to a word in context: alarm
  ...
  \textit{interest} \textit{money} \textit{bank} \textit{loan} \textit{deposit} ...

- Two words are semantically close: distance measure
Evaluation (continued)

- Data: 500 articles from the *Wall Street Journal*
  - Every 200th word is replaced by a spelling variant.

- Evaluation metric (Hirst and St-Onge, 1998):
  
  $\text{correction ratio} = \frac{\text{probability of an error being corrected}}{\text{probability of a correct word raising the alarm}}$
## Results

<table>
<thead>
<tr>
<th>Correction ratio (distributional measures)</th>
<th></th>
<th>Correction ratio (WordNet measures)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>measure</td>
<td>word distance</td>
<td>concept distance</td>
<td>measure</td>
</tr>
<tr>
<td><em>ASD</em> and <em>cp</em></td>
<td>5.03</td>
<td>9.49</td>
<td><em>Hirst–St-Onge</em></td>
</tr>
<tr>
<td><em>Cos</em> and <em>cp</em></td>
<td>4.06</td>
<td>9.05</td>
<td><em>Jiang–Conrath</em></td>
</tr>
<tr>
<td><em>JSD</em> and <em>cp</em></td>
<td>4.88</td>
<td>7.87</td>
<td><em>Leacock–Chodrow</em></td>
</tr>
<tr>
<td><em>Lin</em> and <em>pmi</em></td>
<td>6.52</td>
<td>6.87</td>
<td><em>Lin</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Resnik</em></td>
</tr>
</tbody>
</table>

- Distributional concept-distance measures are markedly better than word-distance measures.
- Only Jiang–Conrath of the WordNet measures outperforms the best distributional concept-distance measure.
Discussion

Distributional Measures

• Distributional concept-distance measures superior
  ■ Sense-ambiguity problem of word-distance measures
  ■ Best measures
    • Ranking word pairs: Lin, Cos
    • Correcting spelling errors: ASD, Cos

• Both concept- and word-distance measures
  ■ Adapt to changes in language
  ■ Geared towards specific domains
Discussion
WordNet-based Measures

- Do better in the word-pair-ranking task
  - Small data-set

- Only Jiang-Conrath outperforms the best distributional concept-distance measures in correcting spelling errors

- Rely on the extensive noun hyponymy hierarchy
  - Both evaluation tasks are on noun–noun pairs
  - Performance on other pairs expected to be poor
Discussion

Distributional Concept-Distance Measures

- Combine knowledge source and text corpora

- Rely on the flat structure of a thesaurus
  - The use of hierarchy and links between categories is still to be explored.

- Very coarse sense inventory (about a 1000 concepts)
  - Pre-computing the complete distance matrix is much easier.
Summary

Provided a framework that allows distributional measures to estimate concept-distance

- Used raw text and a published thesaurus
- Created and used distributional profiles of concepts
- Evaluated in comparison with word-distance measures and WordNet-based measures on two tasks
Current and Future Work

- Use sense inventory of intermediate coarseness
  - Paragraphs of the thesaurus

- Create more accurate WCCMs
  - Weight membership of words in categories

- Explore more applications
  - Compositionality of multi-word expressions

- Extend the ideas to determine senses from text
  - Eliminate reliance on a published thesaurus