Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity

Charley Chan, Chris Callison-Burch, Benjamin Van Durme
Center for Language and Speech Processing, HLTCOE
Johns Hopkins University

GEMS 2011
Paraphrasing

• Goal

To use huge amount of text to identify paraphrases that are semantically similar

<table>
<thead>
<tr>
<th>huge amount of</th>
<th>paraphrase candidate</th>
<th>paraphrase score</th>
</tr>
</thead>
<tbody>
<tr>
<td>huge amount of</td>
<td>huge amount of</td>
<td>1.0</td>
</tr>
<tr>
<td>huge amount of</td>
<td>large quantity of</td>
<td>0.98</td>
</tr>
<tr>
<td>huge amount of</td>
<td>large number of</td>
<td>0.98</td>
</tr>
<tr>
<td>huge amount of</td>
<td>great number of</td>
<td>0.97</td>
</tr>
<tr>
<td>huge amount of</td>
<td>vast number of</td>
<td>0.94</td>
</tr>
<tr>
<td>huge amount of</td>
<td>in large numbers</td>
<td>0.10</td>
</tr>
<tr>
<td>huge amount of</td>
<td>large numbers of</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Our Approach

Bilingual Paraphrase Extraction

(Bilingual pivoting)

+ Monolingual Reranking

(distributional context vector)

Relevant Work

- Clusters phrases through statistical characteristics
- Dependency path similarities (e.g. Lin and Pantel, ‘01)
- Distributional co-occurrence information
- e.g. Ravichandran et al., ‘05, Paşca and Dienes, ’05; Van Durme and Lall, ‘10

- Extract potential paraphrases by grouping English phrases that share the same foreign translations
- e.g. Bannard and Callison-Burch, ‘05

five farmers were thrown into jail in Ireland because they resisted...
**Pros and Cons**

**Bilingual Paraphrase Extraction** (Bilingual pivoting)
- Incorrect word alignment
- High paraphrase probability ≠ good paraphrases
- Produces manageable paraphrase candidate set
  → Scalability
- Less prone to antonym issues
- Multiple languages as pivots

**Monolingual Reranking** (distributional context vector)
- Antonyms, cousins (e.g. *rise*, *fall*; *boy*, *girl*)
- Hard to generate candidate set from corpus
- Prefers grammatically similar phrases
- Training data availability

---

Chan, Callison-Burch, Van Durme

GEMS 2011
Scoring Methods

- **MonoDS** – *Monolingual Distributional Similarity*
  - Distributional context vector built from monolingual n-gram corpus
  - Implemented via locality sensitive hashing (LSH)

- **BiP** – *Bilingual Pivoting*
  - Phrase table and translation model from French-English parallel corpus

- **SyntBiP** – *Syntactically-constrained Bilingual Pivoting*
  - Identical to BiP except paraphrase candidates share the same syntactic type
MonoDS

• Based on distributional similarity
• Baseline approach: Distributional context vector
  – Neighboring unigram frequency
  – For each phrase $x$
    e.g. $x = \text{“dog”}$
    • Every n-gram feature of the form:
      (“the dog”, count = 1707225) or (“dog barks”, count = 8654)
    • Gives rise to a (feature, value) pair:
      $(w_{i-1} = \text{“the”}, 1707225)$ or $(w_{i+1} = \text{”barks”}, 8654)$
    • Vector size: $2|\mathcal{V}|$, $|\mathcal{V}|=$ size of vocabulary
  – Similarity: standard cosine distance
    $\rightarrow$ MonoDS score
    \[
    \text{sim}(x, y) = \frac{|v_x \cdot v_y|}{|v_x||v_y|}
    \]
MonoDS

• Limitations
  – Computational burden in candidate extraction
    → need to compute context vectors for all phrases in n-gram corpus
  – Susceptibility to antonyms
    → Example:
      - Source phrase = reluctant
      - Candidates generated by
        - bilingual paraphrase method
        - hand-selected phrases

\[ \text{sim}(\text{reluctant, willing}) > \text{sim}(\text{reluctant, disinclined}) \]
BiP - Bilingual Paraphrase Extraction

• Idea
  – Identify paraphrases by pivoting through foreign phrases in bilingual parallel corpora
  – Objective:

\[
\hat{e}_2 = \arg \max_{e_2 \neq e_1} p(e_2 | e_1)
\]

\[
p(e_2 | e_1) = \sum_f p(e_2, f | e_1)
\]

\[
= \sum_f p(e_2 | f, e_1) p(f | e_1)
\]

\[
\approx \sum_f p(e_2 | f) p(f | e_1)
\]
SyntBiP – Syntactic Bilingual Paraphrasing

- In addition,
  - Identify paraphrases of the same syntactic type by pivoting through foreign phrases with the specific syntactic type in bitext
  - Objective:

\[
\hat{e}_2 = \arg \max_{e_2 \neq e_1} p(e_2 | e_1, s)
\]

\[
p(e_2 | e_1, s) = \sum_f p(e_2, f | e_1, s) = \sum_f p(e_2 | f, e_1, s) p(f | e_1, s)
\]

In last week five farmers were thrown into jail in Ireland because they were imprisoned, tortured and killed last week five farmers were festgenommen in Ireland because they were festgenommen, weil sie verhindern wo verschwunden oder wurden disappeared or have been, gefoltert und getötet.

Syntactic type

Paraphrase Probability

Translation Probability

Chan, Callison-Burch, Van Durme

GEMS 2011
Bilingual Paraphrase Extraction

• Less susceptible to antonyms
  – Should alleviates antonym problem
    → Antonyms don’t usually share foreign translations
  – Previous example:

<table>
<thead>
<tr>
<th>reluctant</th>
<th>MonoDS\textsubscript{hand-selected}</th>
<th>BiP</th>
</tr>
</thead>
<tbody>
<tr>
<td>*willing, .99</td>
<td>not, .56</td>
<td></td>
</tr>
<tr>
<td>loath, .98</td>
<td>unwilling, .04</td>
<td></td>
</tr>
<tr>
<td>*eager, .98</td>
<td>reluctance, .03</td>
<td></td>
</tr>
<tr>
<td>somewhat reluctant, .98</td>
<td>reticent, .03</td>
<td></td>
</tr>
<tr>
<td>unable, .98</td>
<td>hesitant, .02</td>
<td></td>
</tr>
<tr>
<td>denied access, .98</td>
<td>reticent about, .01</td>
<td></td>
</tr>
<tr>
<td>disinclined, .98</td>
<td>reservations, .01</td>
<td></td>
</tr>
<tr>
<td>very unwilling, .97</td>
<td>reticence, .01</td>
<td></td>
</tr>
<tr>
<td>conducive, .97</td>
<td>hesitate, .01</td>
<td></td>
</tr>
<tr>
<td>linked, .97</td>
<td>are reluctant, .01</td>
<td></td>
</tr>
</tbody>
</table>
Paraphrasing Pipeline

Bilingual paraphrase acquisition

Bilingual parallel corpus → Grammar Extraction → Phrase table → Bilingual Pivoting → Paraphrase table

BiP scores (probabilities) → Phrases & paraphrases

Paraphrase model
- Europarl French-English corpus (Koehn, 2005)
  - 1.3M sentences, 34M words on English side
- Word alignment: Berkeley aligner
- Syntactic information: Stanford parser
- Phrase table and probabilities: Thrax grammar extractor
Context vectors
- Latest version of Google n-gram (Lin et al., 2010)
- max 5-gram in corpus $\rightarrow$ source phrases up to 4-gram
- 3.8 Billion n-grams: Scalability problem
$\rightarrow$ Locality Sensitive Hashing
  (Indyk & Motwani, ’98; Charikar, ’02)
Paraphrasing Pipeline

Bilingual paraphrase acquisition

- Bilingual parallel corpus
  - Grammar Extraction
  - Phrase table
    - Bilingual Pivoting
    - Paraphrase table
      - BiP scores (probabilities)
      - Phrases & paraphrases

Monolingual features construction

- Monolingual N-gram corpus
  - Context Vector Construction
  - LSH Signatures
    - Approximate Cosine Similarity
    - MonoDS scores

Chan, Callison-Burch, Van Durme

GEMS 2011
### Example

<table>
<thead>
<tr>
<th>huge amount of</th>
<th>BiP</th>
<th>SyntBiP</th>
<th>BiP + MonoDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>large number of, 0.33</td>
<td><em>large number of, 0.38</em></td>
<td>huge amount of, 1.0</td>
<td></td>
</tr>
<tr>
<td><em>in large numbers, 0.11</em></td>
<td>great number of, 0.09</td>
<td>large quantity of, 0.98</td>
<td></td>
</tr>
<tr>
<td>great number of, 0.08</td>
<td><em>huge amount of, 0.06</em></td>
<td><em>in large numbers, 0.10</em></td>
<td></td>
</tr>
<tr>
<td>large numbers of, 0.06</td>
<td>vast number of, 0.06</td>
<td>vast number of, 0.94</td>
<td></td>
</tr>
<tr>
<td>vast number of, 0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>huge amount of, 0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>large quantity of, 0.03</td>
<td></td>
<td>large numbers of, 0.08</td>
<td></td>
</tr>
</tbody>
</table>

- **BiP-MonoDS**
  - Less weight on grammaticality error

- **Coverage of SyntBiP**
  - Limited coverage due to syntactic type constraint
Human Evaluation

• Substitution test
  • Amazon Mechanical Turk
  • Source phrase substituted by each of the paraphrase candidates
  • Judge separately the preservation of meaning and grammaticality on a 5-point scale
  • 3 repetition per human intelligence task (HIT), results averaged

• 5-point Scale

<table>
<thead>
<tr>
<th>Score</th>
<th>Meaning</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>perfect</td>
<td>perfect</td>
</tr>
<tr>
<td>4</td>
<td>minor differences</td>
<td>okay but awkward</td>
</tr>
<tr>
<td>3</td>
<td>moderate differences</td>
<td>one error</td>
</tr>
<tr>
<td>2</td>
<td>substantially different</td>
<td>many errors</td>
</tr>
<tr>
<td>1</td>
<td>completely different</td>
<td>ungrammatical</td>
</tr>
</tbody>
</table>
Human Evaluation

- Substitution test example

<table>
<thead>
<tr>
<th>Original</th>
<th>the sea level will rise by about 14 inches instead of 39 .</th>
<th>Paraphrase</th>
<th>Meaning</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>that will rise by about 14 inches instead of 39 .</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the rise will rise by about 14 inches instead of 39 .</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>sea level will rise by about 14 inches instead of 39 .</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>the will rise by about 14 inches instead of 39 .</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sea levels will rise by about 14 inches instead of 39 .</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>perfect</td>
</tr>
<tr>
<td>4</td>
<td>minor differences</td>
</tr>
<tr>
<td>3</td>
<td>moderate differences</td>
</tr>
<tr>
<td>2</td>
<td>substantially different</td>
</tr>
<tr>
<td>1</td>
<td>completely different</td>
</tr>
</tbody>
</table>

Chan, Callison-Burch, Van Durme
Human Evaluation

• Dataset
  – 100 source phrases
    • 25 phrases per n-gram, n = 1,2,3,4
    • Randomly sampled from paraphrase tables generated by BiP and SyntBiP
    • 5 sentences per phrase, 500 total

• Each scoring method (BiP, SyntBiP, MonoDS) evaluated with:
  – Averages of meaning and grammaticality scores
  – Rank correlation with human judgment (Kendall Tau-b)
    \[ \tau_B = 1: \text{perfectly correlated} \]
    \[ \tau_B > 0: \text{positively correlated} \]
    \[ \tau_B < 0: \text{negatively correlated} \]
Correlation with Human Judgments

Kendall Tau-b

Meaning
Grammar

BiP
BiP-MonoDS
SyntBiP
SyntBiP-MonoDS

* : p ≥ 0.01
Thresholding by Score

BiP

Grammar: Okay but awkward

Meaning: substantially different

MonoDS

Grammar: Okay but awkward

Meaning: moderate differences

Bilingual ↔ Meaning

MonoDS ↔ Grammar

Chan, Callison-Burch, Van Durme

GEMS 2011
Combining MonoDS and BiP Thresholds

• Joint Thresholding
  – Tighter threshold $\rightarrow$ higher paraphrase quality
  – Additional gain in score than each alone
  – High precision paraphrase retrieved at the expense of coverage

Chan, Callison-Burch, Van Durme

GEMS 2011
Effects of Bilingual Training Corpus

- **British vs American English**
  - `{legalise, legalize}`
  - BiP:
    → trained on corpus with British English

- **Incorrect word alignment**
  - `{considerable changes, caused quite}`
  - Incorrect translation from misalignment propagate through bilingual paraphrase model

  ```
  considerable changes \rightarrow \text{modifie considérablement}  
  modifie considérablement \rightarrow \text{caused quite (misaligned)}
  ```

→ MonoDS alleviates problems caused by bilingual model training
Influence

- \{hauled, delivered\}

*Countries which do not comply with community legislation should be hauled before the court of justice...*

\[\text{MonoDS takes advantage of context information}\]
Influence

- \{fiscal burden, taxes\}

1. ... the member states can reduce the fiscal burden consisting of taxes and social contributions.

2. ... and, in addition, the fiscal burden in Europe has reached an all-time high of 46%.

→ MonoDS is more useful in context that defines the context vector.
MonoDS Limitations

• Feature sparsity
  – \{to deal properly with, address\}
    • Non-zero features for to deal properly with:
      to deal properly with the 709
    • Non-zero features for address
      43120 n-grams with unigram neighbor
      E.g. address lawsuits 36 operators address 247
today address 685 address judicial 341
  – \{you have just stated, you have just suggested\}
    • MonoDS score = 0.05
    • MonoDS score of \{stated, suggested\} = 0.91

→ MonoDS suffers from feature sparsity for higher n-grams

 Chan, Callison-Burch, Van Durme

GEMS 2011
Summary

• Presented a novel paraphrase acquisition method
  – Generate paraphrase candidate using bilingual pivoting
  – Rerank with monolingual distributional similarity and paraphrase probabilities

• Findings through analysis on human judgment tasks:
  ❖ Combining monolingual and bilingual information
    → substantially improves paraphrase quality
  ❖ Monolingual distributional similarity confines structural match
    → MonoDS score correlates strongly with grammaticality
  ❖ Joint threshold on MonoDS and BiP scores
    → high precision paraphrases at expense of coverage

Chan, Callison-Burch, Van Durme

GEMS 2011
Summary

• Future directions
  – Explore better models for MonoDS to enhance meaning preservation
    • Experiment with richer features
      – Increased number of neighboring words
      – Syntactic dependency parsing (Lin, 1997)
      – Mutual information between phrasal co-occurences (Church and Hanks, 1991)
  – Beyond joint thresholding
    • Systematically combine different scores and additional features
      – SVM training
  – Large-scale paraphrase dataset... stay tuned
  – Investigate domain-specific paraphrasing
Thank you

Questions?