A Probabilistic Model for Measuring Grammaticality and Similarity of Automatically Generated Paraphrases of Predicate Phrases

Atsushi FUJITA and Satoshi SATO
Nagoya Univ., Japan
Overview

Abstract pattern

Paraphrase Generation (Instantiation)

Employment shows a sharp decrease

Employment decreases sharply

Paraphrase candidate

Quality Measurement

- Grammaticality
- Similarity

Score (How likely to be paraphrase)
Automatic Paraphrasing

- Fundamental in NLP
  - Recognition: IR, IE, QA, Summarization
  - Generation: MT, TTS, Authoring/Reading aids

- Paraphrase knowledge
  - Handcraft
    - Thesauri (of words) [Many work]
    - Transformation rules [Mel’cuk+, 87] [Dras, 99] [Jacquemin, 99]
  - Automatic acquisition
    - Anchor-based [Lin+, 01] [Szpektor+, 04]
    - Aligning comparable/bilingual corpora [Many work]
Representation of Paraphrase Knowledge

- Fully-abstracted
  - $X \text{ V } Y$ → $X$'s $V$-ing of $Y$
  - $X \text{ V } Y$ → $Y$ be $V$-PP by $X$
  - $X$ show a $A$ $Y$ → $X$ $v(Y)$ adv($A$)
  - $X$ wrote $Y$ → $X$ is the author of $Y$
  - $X$ solves $Y$ → $X$ deals with $Y$
- Fully-lexicalized
  - burst into tears → cried
  - comfort → console

Nominalization
Passivization
Removing light-verb

[Harris, 1957]
[Barzilay+, 2001]
[Lin+, 2001]
Instantiating Phrasal Paraphrases

- Over-generation leads to spurious instances
  - cf. filling arguments [Pantel+, 07]
  - cf. applying to contexts [Szpektor+, 08]

\[
X \text{ show a } A \ Y \quad \rightarrow \quad X \ v(Y) \ \text{adv}(A)
\]

- Employment shows a sharp decrease  \rightarrow \text{ Employment decreases sharply} \quad \text{OK}
- Statistics show a gradual decline  \rightarrow \text{ Statistics decline gradually} \quad \text{Not equivalent}
- The data show a specific distribution  \rightarrow \text{ The data distribute specifically} \quad \text{Not grammatical}
Task Description

- Measuring the quality of paraphrase candidate

  **Input:** Automatically generated phrasal paraphrases

  $s$: Employment shows a sharp decrease

  $t$: Employment decreases sharply

  **Output:** Quality score $[0,1]$
Quality as Paraphrases

Three conditions to be satisfied

1. Semantically equivalent
2. Substitutable in some context
3. Grammatical

Approaches

- Acquisition of instances
  - 1 and 2 are measured, assuming 3
- Instantiation of abstract pattern (our focus)
  - 1 and 2 are weakly ensured
  - 3 is measured, and 1 and 2 are reexamined
Outline

1. Task Description
2. Proposed Model
3. Experiments
4. Conclusion
Proposed Model

Assumptions
- $s$ is given and grammatical
- $s$ and $t$ do not co-occur

Formulation with a conditional probability

$$P(t|s) = \sum_{f \in F} P(t|f)P(f|s)$$

$$= \sum_{f \in F} \frac{P(f|t)P(t)}{P(f)} P(f|s)$$

$$= P(t) \sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}$$

Grammaticality  Similarity
Grammaticality Factor

- Statistical Language Model
  - Structured $N$-gram LM
  - Normalized with length

\[
P(t) = \left[ \prod_{i=1 \ldots |T(t)|} P_d(c_i \mid d_{i,1}^1, d_{i,2}^2, \ldots, d_{i,N-1}^{N-1}) \right]^{1/|T(t)|}
\]
Grammaticality Factor: Definition of Nodes

For Japanese

- What present dependency parsers determine
  - **Bunsetsu**: \{Content word\} + \{Function word\} *

- **Bunsetsu** dependencies

- **Bunsetsu** can be quite long (so not appropriate)

(He will surely not come to today’s meeting.)
Grammaticality Factor: MDS

Morpheme-based Dependency Structure [KURA, 01]

- Node: Morpheme
- Edge:
  - Rightmost node → Head-word of its mother *bunsetsu*
  - Other nodes → Succeeding node

(He will surely not come to today’s meeting.)
Grammaticality Factor: CFDS

Content-Function-based Dependency Structure

- **Node:** Sequence of content words or of function words
- **Edge:**
  - Rightmost node → Head-word of its mother *bunsetsu*
  - Other nodes → Succeeding node

(He will surely not come to today’s meeting.)
Grammaticality Factor: Parameter Estimation

- MLE for 1, 2, and 3-gram models
  - Mainichi (1.5GB)

<table>
<thead>
<tr>
<th>Node Type</th>
<th># of alphabets</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDS</td>
<td>320,394</td>
</tr>
<tr>
<td>CFDS</td>
<td>14,625,384</td>
</tr>
<tr>
<td><em>Bunsetsu</em></td>
<td>19,507,402</td>
</tr>
</tbody>
</table>

- Linear interpolation of 3 models
  - Mixture weights were determined via an EM
  - Yomiuri (350MB) + Asahi (180MB)
Similarity Factor

- A kind of distributional similarity measure

\[ \sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)} \]

- Contextual feature set \((F)\)

  - **BOW**: Words surrounding \(s\) and \(t\) have similar distribution
    \[ \Rightarrow s\ and\ t\ are\ semantically\ similar \]
  
  - **MOD**: \(s\) and \(t\) share a number of modifiers and modifiees
    \[ \Rightarrow s\ and\ t\ are\ substitutable \]
Similarity Factor: Parameter Estimation

- Employ Web snippets as an example collection
  - To obtain sufficient amount of feature info.
  - Yahoo! JAPAN Web-search API
    - “Phrase search”
    - 1,000 snippets (as much as possible)
MLE

- \( P(f|p) \)
  - Based on snippets

- \( P(f) \)
  - Based on static corpus

Similarity Factor: Parameter Estimation (cont’d)

WebCP (42.7GB)  
Mainichi (1.5GB)  

[Kawahara+, 06]
Summary

What is taken into account
- Grammaticality of $t$
- Similarity between $s$ and $t$

You do not need to enumerate all the phrases
- cf. $P(ph \mid f)$, $pmi(ph, f)$

Options

Grammaticality

$$P(t \mid s) = P(t) \sum_{f \in F} \frac{P(f \mid t)P(f \mid s)}{P(f)}$$

Similarity

max # of snippets (1,000 / 500)

MDS / CFDS

Mainichi / WebCP

BOW / MOD
Outline

1. Task Description
2. Proposed Model
3. Experiments
4. Conclusion
Overview

Paraphrase candidate

Quality Measurement

• Grammaticality
• Similarity

Score (How likely to be paraphrase)

Paraphrase Generation (Instantiation)

Employment shows a sharp decrease

Employment decreases sharply

Abstract pattern

X show a A Y

X v(Y) adv(A)
Test Data

- Extract input phrases
  - 1,000+ phrases × 6 basic phrase types
  - Mainichi (1.5GB)
  - Referring to structure

- Paraphrase generation [Fujita+, 07]
  - 176,541 candidates for 4,002 phrases

- Sampling
  - Candidates for 200 phrases
  - Diverse cases (see column Y)

<table>
<thead>
<tr>
<th>Phrase type</th>
<th>All</th>
<th>Sampled</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>s</td>
<td>〈s, t〉</td>
<td>Y</td>
</tr>
<tr>
<td>N:C:V</td>
<td>489</td>
<td>18</td>
<td>57</td>
</tr>
<tr>
<td>N_1:N_2:C:V</td>
<td>966</td>
<td>57</td>
<td>4,596</td>
</tr>
<tr>
<td>N:C:V_1:V_2</td>
<td>982</td>
<td>54</td>
<td>4,767</td>
</tr>
<tr>
<td>N:C:Adv:V</td>
<td>523</td>
<td>16</td>
<td>51</td>
</tr>
<tr>
<td>Adj:N:C:V</td>
<td>50</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>N:C:Adj</td>
<td>992</td>
<td>53</td>
<td>173</td>
</tr>
<tr>
<td>Total</td>
<td>4,002</td>
<td>200</td>
<td>9,652</td>
</tr>
</tbody>
</table>
Overview

X show a A Y

X v(Y) adv(A)

Employment decreases sharply

Paraphrase Generation (Instantiation)

Employment shows a sharp decrease

Employment decreases sharply

Paraphrase candidate

Quality Measurement

- Grammaticality
- Similarity

Score (How likely to be paraphrase)
**Viewpoint**

- How well a system can rank a correct candidate first?

**Models evaluated**

- **Proposed model**
  - All combination of options
  - $P(t) \times P(f) \times \text{Feature set} \times \max \# \text{of snippet}$
    
    $\begin{array}{cccc}
    2 & 2 & 2+1 & 2 \\
    \end{array}$

- **Baselines**
  - Lin’s measure [Lin+, 01]
  - $\alpha$-skew divergence [Lee, 99]
  - HITS

HAR: harmonic mean of BOW and MOD scores

Similarity only

Grammaticality only
Results (max 1,000 snippets)

- # of cases that gained positive judgments
  - Models except CFDS+Mainichi << the best models

<table>
<thead>
<tr>
<th>Model \ Feature</th>
<th>Strict</th>
<th>Lenient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BOW</td>
<td>MOD</td>
</tr>
<tr>
<td>CFDS+Mainichi</td>
<td>79</td>
<td>82</td>
</tr>
<tr>
<td>Lin</td>
<td>79</td>
<td>88</td>
</tr>
<tr>
<td>α-skew</td>
<td>84</td>
<td>89</td>
</tr>
<tr>
<td>HITS</td>
<td></td>
<td>84</td>
</tr>
</tbody>
</table>

XXX: best

XXX: significantly worse than the best (McNemar’s test, p<0.05)
Results (max 1,000 snippets, HAR)

- Lenient precision and score
  - Best candidate ∧ Relatively high score ⇒ High precision
Considerations

- Harnessing the Web led to accurate baselines
  1. Looking up the Web … Feature retrieval + Grammaticality check
  2. Comparing feature distributions … Similarity check

- Two distinct viewpoints of similarity are combined

  **Constituent similarity:**
  - Syntactic transformation + Lexical derivation [Fujita+, 07]

  **Contextual similarity:**
  - Bag of words / Bag of modifiers
Diagnosis shows the room of improvement

Grammaticality

\[ P(t|s) = P(t) \sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)} \]

Similarity

max # of snippets

(1,000 / 500 / 200 / 100)

MDS < CFDS

Mainichi > WebCP

BOW < MOD \equiv \text{HAR}

A2: MDS cannot capture collocation of content words

A4: Linguistic tools are trained on newspaper articles

A3: Combining with P(t) dismisses the advantage

A5: No significant difference (Even Web is not sufficient?)
Conclusion & Future work

Measuring the quality of paraphrase candidates

**Input:** Automatically generated phrasal paraphrases

**Output:** Quality score [0,1]

- Semantically equivalent
- Substitutable in some context
- Grammatical

\[ \text{Similarity} \quad \text{Grammaticality} \]

- Overall: 54-62% (cf. Lin/skew: 58-65%, HITS: 60%)
- Top 50: 80-92% (cf. Lin/skew: 90-98%, HITS: 70%)

Future work

- Feature engineering (including parameter tuning)
- Application to non-productive paraphrases