Paraphrasing

May 1, 2012
Goals of today’s lecture

• Understand what **paraphrases** are
• Discuss how we can re-use **MT machinery** of for other **text-to-text (T2T) generation tasks**
• Review various data-driven methods for learning paraphrases
• Focus on a method that uses **bilingual pivoting**
• Define a set of modifications that we need to make to the MT pipeline to customize it to new tasks
What are Paraphrases?

<table>
<thead>
<tr>
<th>cup</th>
<th>mug</th>
</tr>
</thead>
<tbody>
<tr>
<td>the king’s speech</td>
<td>His Majesty’s address</td>
</tr>
<tr>
<td>X₁ talks to X₂</td>
<td>X₁ converses with X₂</td>
</tr>
<tr>
<td>NN devoured NP</td>
<td>NP was eaten by NN</td>
</tr>
</tbody>
</table>

Differing textual expressions of the **same meaning**:

Many Republicans' hearts were broken by Chris Christie reiterating his refusal to run for the presidency.

The Garden State governor stated once again that he will not seek the presidential nomination, disappointing Republicans.
What are they good for?

Anything that deals with **text and meaning**, i.e. automatic...

...summarization, translation, MT evaluation, question answering, information retrieval, natural language generation, essay grading, sentiment analysis, linguistic stenography, entailment recognition, etc.

Real question is **where do we get them?**
Many NLP tasks can be viewed as “MT”

• If you have a “source” and a “target” that are aligned on the sentence-level, then you can re-use much of the MT machinery to “translate” between them

• Input this parallel corpus and then re-use
  — Word alignment algorithms
  — Phrase table extraction
  — Decoder + LM

• Example task: Sentence simplification
<table>
<thead>
<tr>
<th>English-Simple</th>
<th>Parallel Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>a synonym for &quot;lolcat&quot; is cat macro, since the images are a type of image macro.</td>
<td>a different word for lolcat is cat macro because it is a kind of image macro.</td>
</tr>
<tr>
<td>genetic engineering has expanded the genes available to breeders to utilize in creating desired germlines for new crops.</td>
<td>new plants were created with genetic engineering.</td>
</tr>
<tr>
<td>the dominant classical dance amongst tamils is bharatanatyam.</td>
<td>bharatanatyam is the main dance of the tamil people.</td>
</tr>
<tr>
<td>a naval mine is a self-contained explosive device placed in water to destroy ships or submarines.</td>
<td>a naval mine is a bomb placed in water to destroy ships or submarines.</td>
</tr>
</tbody>
</table>
a **synonym** for "lolcat" is cat macro, **since** the images **are** a type of image macro.

a **different word** for lolcat is cat macro **because** it is a **kind** of image macro.
<table>
<thead>
<tr>
<th>synonym</th>
<th>different word</th>
</tr>
</thead>
<tbody>
<tr>
<td>, since</td>
<td>because</td>
</tr>
<tr>
<td>are</td>
<td>is</td>
</tr>
<tr>
<td>type</td>
<td>kind</td>
</tr>
</tbody>
</table>

a synonym for "X" is Y | a different word for X is Y
since then they have changed their name to palladium and played alongside amy winehouse.

since then, they have changed their name to palladium and played with amy winehouse.
Done! Right??

• Just need to calculate a BLEU score and then write a paper

• What is wrong with this?
• Where does it get things right and where does it get things wrong?
• (Discuss with your neighbor)
Paraphrasing with parallel monolingual data

• Some work has used parallel monolingual data

• Comparable corpora
  – Encyclopedia articles on the same topic
  – Different newspapers’ accounts of one event

• Multiple translations of the same foreign text
  – Evaluation data for Bleu metric
  – Different translations of classic French novels into English
What a scene! Seized by the tentacle and **glued to** its suckers, the unfortunate man was **swinging in the air** at the **mercy** of this enormous appendage. He gasped, he choked, he yelled: "Help! Help!" I'll hear his **harrowing plea** the rest of my life! The **poor fellow** was **done for**.

What a scene! The unhappy man, seized by the tentacle and **fixed to** its suckers, was **balanced in the air** at the **caprice** of this enormous trunk. He rattled in his throat, he was stifled, he cried, "Help! help!" That **heart-rending cry**! I shall hear it all my life. The **unfortunate man** was **lost**.
Paraphrasing with parallel monolingual data

• Barzilay and McKeown (2001) used identical contexts in aligned sentences:

<table>
<thead>
<tr>
<th>Emma burst into tears and he tried to comfort her, saying things to make her smile.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emma cried and he tried to console her, adorning his words with puns.</td>
</tr>
</tbody>
</table>

• **burst into tears** = **cried** and **comfort** = **console**
Potential problems with these methods

• Multiple translations are relatively uncommon
• This Limits what paraphrases we can generate
  – Limited number of paraphrases
  – Constrained to a few genres
If we consider oculist and eye-doctor we find that, as our corpus of utterances grows, these two occur in almost the same environments. In contrast, there are many sentence environments in which oculist occurs but lawyer does not...

It is a question of the relative frequency of such environments, and of what we will obtain if we ask an informant to substitute any word he wishes for oculist (not asking what words have the same meaning).

These and similar tests all measure the probability of particular environments occurring with particular elements... If A and B have almost identical environments we say that they are synonyms.

–Zellig Harris (1954)
Duty and Responsibility

• To operationalize the Distributional Hypothesis we must define similar environments
• Lin and Panel (2001) used dependency relationships
• Duty and responsibility share a similar set of dependency contexts in large volumes of text:

<table>
<thead>
<tr>
<th>modified by adjectives</th>
<th>objects of verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>additional, administrative, assigned, assumed, collective, congressional, constitutional ...</td>
<td>assert, assign, assume, attend to, avoid, become, breach ...</td>
</tr>
</tbody>
</table>
Problem with distributional similarity

• Distributional methods group related words that are not synonymous:
  – cats and dogs, girls and boys
Paraphrasing with Bilingual parallel corpora

• Bilingual parallel corpora are much more common than monolingual parallel corpora
• However, no longer contain identical contexts
• Use aligned foreign language phrase as pivot
• Less prone to retrieve non-synonymous related words
... 5 farmers were thrown into jail in Ireland ...

... fünf Landwirte festgenommen, weil ...

... oder wurden festgenommen, gefoltert ...

... or have been imprisoned, tortured ...
The establishment of the military force is in their view a tool to realise these aims. It will be a force comprised of various national units.

Die Truppe ist ihrer Auffassung nach ein Mittel zur Durchsetzung dieser Ziele.

It will be involved in peacemaking. The 1000 strong Friedenstruppe soll zur Friedensschaffung herangezogen werden.

1000 Mann starke Friedenstruppe wird zur Friedensschaffung eingesetzt.

The 1000 strong Friedenstruppe will be involved in peacemaking. Die 1000 Mann starke Friedenstruppe soll zur Friedensschaffung herangezogen werden.

The EU may carry out tasks which do not use military force. Die EU sollte Aufgaben durchführen bei denen keine militärische Nutzung erlaubt.

Die EU kann nicht auf militärischen Einsatz verzichten. Die Europäische Union kann nicht auf militärischen Einsatz verzichten.

The EU may carry out tasks which do not use military force. Die EU sollte Aufgaben durchführen bei denen keine militärische Nutzung erlaubt.

The EU may carry out tasks which do not use military force. Die EU sollte Aufgaben durchführen bei denen keine militärische Nutzung erlaubt.

Angola, for example, has powerful armed forces. Angola beispielsweise besitzt starke Streitkräfte.

Angola besitzt starke Streitkräfte. Angola besitzt starke Streitkräfte.

Angola, for example, has powerful armed forces. Angola beispielsweise besitzt starke Streitkräfte.

Angola, for example, has powerful armed forces. Angola beispielsweise besitzt starke Streitkräfte.

Due to reduced defence spending the national forces can currently only supply approximately 20,000 men. Aufgrund eines gekürzten Verteidigungshaushaltes können die Streitkräfte gegenwärtig jedoch nur etwa 20,000 Mann aufbringen.

Due to reduced defence spending the national forces can currently only supply approximately 20,000 men. Aufgrund eines gekürzten Verteidigungshaushaltes können die Streitkräfte gegenwärtig jedoch nur etwa 20,000 Mann aufbringen.

Due to reduced defence spending the national forces can currently only supply approximately 20,000 men. Aufgrund eines gekürzten Verteidigungshaushaltes können die Streitkräfte gegenwärtig jedoch nur etwa 20,000 Mann aufbringen.

Due to reduced defence spending the national forces can currently only supply approximately 20,000 men. Aufgrund eines gekürzten Verteidigungshaushaltes können die Streitkräfte gegenwärtig jedoch nur etwa 20,000 Mann aufbringen.
Many, many alternatives

Paraphrase candidates for “thrown into jail”

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad / Ugly</th>
</tr>
</thead>
<tbody>
<tr>
<td>jailed</td>
<td>being arrested</td>
</tr>
<tr>
<td>arrested</td>
<td>in jail</td>
</tr>
<tr>
<td>imprisoned</td>
<td>put in prison for</td>
</tr>
<tr>
<td>incarcerated</td>
<td>maltreated</td>
</tr>
<tr>
<td>locked up</td>
<td>thrown</td>
</tr>
<tr>
<td>taken into custody</td>
<td>cases</td>
</tr>
<tr>
<td>thrown into prison</td>
<td>custody</td>
</tr>
</tbody>
</table>
Good examples

• **dead bodies** → corpses, carcasses, bodies, skeletons, people

• **military force** → force, forces, peace-keeping personnel, armed forces, military forces, defense

• **sooner or later** → eventually, at some point

• **wish to clarify** → want to make perfectly clear, would like to ask, would like to comment on, would like to mention, would like to deal with, would comment on

• **every other** → any other, all, other, every, all other, everyone else, others, all the others
Bad examples

• **are perfectly entitled** → perfectly entitled, have every right, have the right

• **for small-scale projects** → small-scale projects, small, of, only trifling amounts are at stake, for smaller projects, for smaller-scale projects, to, for smaller projects

• **groundwork for** → for, groundwork, to, basis for, the, basis, preparation, foundations for, that

• **create equal** → equal, to create a, create, to create equality, same, created, conditions
Separating the Good from the Bad

• How could we differentiate good paraphrases from bad ones?

• (Discuss with your neighbor)
• Paraphrase probability

\[
p(e_2|e_1) = \sum_{f} p(e_2, f|e_1) = \sum_{f} p(e_2|f, e_1)p(f|e_1) \approx \sum_{f} p(e_2|f)p(f|e_1)
\]

\[
p(f|e) = \frac{\text{count}(e, f)}{\sum_{f} \text{count}(e, f)}
\]

Log-linear model with additional features
Phrase extraction with unaligned words

- For 3.7m paraphrases of 400k phrases
  - 34% were sub- or super-strings
  - 73% of the paraphrases that were ranked highest by the paraphrase probability
Syntactic Constraints

- Require phrases and their paraphrase to be the same *syntactic type*
- Redefine the paraphrase probability to *condition on syntactic labels*
- Change the *phrase extraction algorithm* so that it enumerates phrase pairs and syntactic labels
• Coverage is significantly reduced
• <25% of phrases that were previously paraphrasable are paraphrasable now
Using complex labels

- Coverage improves 3x over simple labels
- Covers 2/3 of phrases that the baseline does

Cómo podemos crear una igualdad de derechos?
Example improvements

• create equal | equal, to create a, create, to create equality, same, created, conditions

• VP/NNS → create equal | creating equal
• VP/NNS PP → create equal | promote equal, establish fair
• VP/NNS PP PP → create equal | creating equal, provide equal, create genuinely fair
Example improvements

- **equal** | same, equality, equals, equally, the, fair, equal rights

- **JJ** → **equal** | same, fair, similar, equivalent

- **ADJP** → **equal** | necessary, similar, identical, the same, equal in law, equivalent
SCFGs for Paraphrasing

• What does this notation remind you of?
  – \( JJ \rightarrow \text{equal} \mid \text{same} \)

• Synchronous context free grammars!

• If you hadn’t guessed already, we can fuse the idea of pivoting with syntactic MT to get SCFGs for paraphrasing
Meaning preserving transformations

- Adapting our syntactic MT models, we learn structural transformations, like the English possessive rule

\[
\begin{align*}
\text{NP} & \rightarrow \text{NP 's NN} \mid \text{le NN de NP} \\
\text{NP} & \rightarrow \text{the NN of NP} \mid \text{le NN de NP} \\
\text{combine to} & \\
\text{NP} & \rightarrow \text{NP 's NN} \mid \text{the NN of NP}
\end{align*}
\]
<table>
<thead>
<tr>
<th>Rule Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Possessive rule</strong></td>
<td>NP → the NN of the NNP</td>
</tr>
<tr>
<td></td>
<td>NP → the NNS(_1) made by the NNS(_2)</td>
</tr>
<tr>
<td><strong>Dative shift</strong></td>
<td>VP → give NN to NP</td>
</tr>
<tr>
<td></td>
<td>VP → provide NP(_1) to NP(_2)</td>
</tr>
<tr>
<td><strong>Adv./adj. phrase move</strong></td>
<td>S/VP → ADVP they VBP</td>
</tr>
<tr>
<td></td>
<td>S → it is ADJP VP</td>
</tr>
<tr>
<td><strong>Verb particle shift</strong></td>
<td>VP → VB NP up</td>
</tr>
<tr>
<td><strong>Reduced relative clause</strong></td>
<td>SBAR/S → although PRP VBP that</td>
</tr>
<tr>
<td></td>
<td>ADJP → very JJ that S</td>
</tr>
<tr>
<td><strong>Partitive constructions</strong></td>
<td>NP → CD of the NN</td>
</tr>
<tr>
<td></td>
<td>NP → all DT\NN</td>
</tr>
<tr>
<td><strong>Topicalization</strong></td>
<td>S → NP, VP.</td>
</tr>
<tr>
<td><strong>Passivization</strong></td>
<td>SBAR → that NP had VBN</td>
</tr>
<tr>
<td><strong>Light verbs</strong></td>
<td>VP → take action ADVP</td>
</tr>
<tr>
<td></td>
<td>VP → to take a decision PP</td>
</tr>
</tbody>
</table>
Sentential Paraphrasing

• These paraphrasing SCFGs can be used for monolingual text-to-text generation tasks
• Non-naive reuse of SMT machinery
• Adapt translation framework with appropriate
  – Development data
  – Objective function
  – Feature sets
  – Grammar augmentations
Example: Sentence Compression

• Problem: given an input sentence, rewrite it into a shorter sentence while preserving the core meaning:

  and he said that the project will cover the needs of the region in the long term.

  he said the project includes all the district's long-term needs.
SMT Machinery

• What we can directly re-use:
  – Grammar extraction & formalism
  – Decoding & n-gram language model integration
  – Log-linear model formulation
  – MERT for parameter tuning
## SMT Machinery

<table>
<thead>
<tr>
<th>Development Data</th>
<th>Multi-reference sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective Function</td>
<td><strong>BLEU</strong></td>
</tr>
<tr>
<td>Features</td>
<td>$P_{\text{phrase}}(e_1</td>
</tr>
</tbody>
</table>
Reusing SMT for Text-to-Text

- Inter-reference BLEU is typically very high (52.7)
- Resulting paraphrases are almost always identity

<table>
<thead>
<tr>
<th>Input</th>
<th>the election campaign, which did not gain the interest of voters, ended friday.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrase</td>
<td>the election campaign, which did not gain the interest of voters, ended friday.</td>
</tr>
</tbody>
</table>
Adapting SMT Machinery

<table>
<thead>
<tr>
<th></th>
<th>SMT</th>
<th>Sentence Compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development Data</td>
<td>Multi-reference sets</td>
<td>&lt;sentence, compression&gt;</td>
</tr>
<tr>
<td>Objective Function</td>
<td>BLEU</td>
<td>CMPBLEU</td>
</tr>
<tr>
<td>Features</td>
<td>$P_{\text{phrase}}(e_1</td>
<td>e_2)$, $P_{\text{lex}}(e_1</td>
</tr>
<tr>
<td>Augmentation</td>
<td>n/a</td>
<td>Deletion rules</td>
</tr>
</tbody>
</table>
Development Data

• Common compression corpora are deletion-based (e.g. Ziff-Davis)
• We create a development and test sets from reference translations for SMT
• Consists of compressive sentential paraphrases (CR 0.8 to 0.5, 0.73 avg.)

and he said that the project will cover the needs of the region in the long term.

he said the project includes all the district's long-term needs.
Objective Function

- Penalize insufficient compressions
- Reward well-formed language
- Penalize overzealous compressions

\[
\text{CMPBLEU}_{\lambda, \theta}(i, o) = \begin{cases} 
  e^{\lambda(\theta - c)} \cdot \text{BLEU}(o) & \text{if } c > \theta \\
  \text{BLEU}(o) & \text{otherwise}
\end{cases}
\]
• Augment rules with length information
  – Number of words on source & target side
  – Difference in number of words
  – Difference in number of characters
• Added deletion rules for hand-chosen POS
  – JJ, JJR, JJS
  – RB, RBR, RBS
  – DT

\[
\text{JJ} \rightarrow \text{superfluous} \mid \varepsilon
\]
Example sentence compression

Lexical paraphrase:
JJ → offensive | insulting

Reduced relative clause:
NP → NP that VP | NP VP

Pred. adjective copula deletion:
VP → are JJ to NP | JJ NP

Partitive construction:
NP → CD of the NNS | CD NNS
Text-to-text generation tasks

• Sentence compression
• Sentence simplification
• English as a Second Language (ESL) error correction
• Poetry generation
• Legalese to plain English translation
Conclusions

- Paraphrases are useful for a wide range of NLP tasks
- Tempting to think of SMT as a tool that can be used to do anything ... just find “parallel corpus”
- Doesn’t work well if done simplemindedly
- Better to extract paraphrases from bilingual parallel corpora
- Then adapt the SMT machinery in non-naive ways