Paraphrasing and Translation

Chris Callison-Burch

16 March 2006

School of informatics

Talk Overview

- Paraphrases
  - What they're useful for
  - How other people generate them
  - How we do it
- Applying Paraphrases to Translation
  - Problem of unseen words in SMT
  - Using paraphrases to alleviate this
  - Evaluation

Usefulness of paraphrases

- Paraphrases are alternative ways of conveying the same information
- Useful in NLP application such as:
  - *Generation* - producing paraphrases allows for the creation of more varied and fluent text
  - *Multidocument summarization* - identifying paraphrases allows information repeated across documents to be condensed
  - *Question answering* - paraphrasing is important when going beyond simple keyword matching to find answers
  - *Machine translation* - as we will see later

Paraphrasing with monolingual parallel data

- Previous work by Regina Barzilay and others has focused on *monolingual* parallel corpora
- Monolingual parallel data comes from multiple translations of the same thing:
  - Multiple translations of classic French novels into English
  - Evaluation data for Bleu method of scoring MT systems
- People have also used comparable corpora (encyclopedia articles on the same topic)

Paraphrasing with monolingual parallel data

- Methodology:
  - Align sentences across translations
  - Identify similar contexts in aligned sentences
  - Phrases that appear in similar contexts may be paraphrases
- Example:

Emma burst into tears **and he tried to** comfort **her,** saying things to make her smile.
Emma cried, **and he tried to** console **her,** adorning his words with puns.

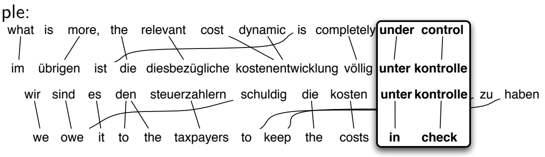
- Extract *burst into tears* = *cried* and *comfort* = *console*

Potential problems with this method

- Parallel monolingual texts are relatively uncommon
- Limits what paraphrases we can generated
  - Limited number of paraphrases
  - Constrained to a few genres

Paraphrasing with bilingual parallel corpora

- Our Methodology:
  - Use statistical MT techniques to align a bilingual parallel corpus
  - Get foreign phrases aligned to the English phrase we want to paraphrase
  - Find other English phrases that foreign phrases align with
  - Treat those English phrases as potential paraphrases, and rank them
- Example:



what is more, the relevant cost dynamic is completely **under control**

im übrigen ist die diesbezügliche kostenentwicklung völlig **unter kontrolle**

wir sind es den steuerzahlern schuldig die kosten **unter kontrolle** zu haben

we owe it to the taxpayers to keep the costs **in check**

More examples

- **military force** → armed forces, defence, force, forces, peace-keeping personnel, military forces
- **sooner or later** → at some point, eventually
- **great care** → a careful approach, greater emphasis, particular attention, specific attention, special attention, very careful
- **at work** → at the workplace, employment, held, holding, in the work sphere, organised, operate, taken place, took place, working

Paraphrase Probability

- Since we have multiple paraphrases, we rank them with a paraphrase probability

$$\begin{aligned}\hat{e}_2 &= \arg\max_{e_2 \neq e_1} p(e_2|e_1) && (1)\\ &= \arg\max_{e_2 \neq e_1} \sum_f p(f|e_1)p(e_2|f) && (2)\\ &= \arg\max_{e_2 \neq e_1} \sum_f \frac{count(f,e_1)}{\sum_f count(f,e_1)} \frac{count(e_2,f)}{\sum_{e_2} count(e_2,f)} && (3)\end{aligned}$$

- Can also rank paraphrases **in context** by weighting paraphrase probability by language model score

Judging paraphrase quality

- Substituted each paraphrase into **2** - **10** sentences which contained original phrase

Under control
What is more, the relevant cost dynamic is completely **in check**.
What is more, the relevant cost dynamic is completely **checked**.
What is more, the relevant cost dynamic is completely **slow down**.
What is more, the relevant cost dynamic is completely **curb**.
What is more, the relevant cost dynamic is completely **curbed**.
What is more, the relevant cost dynamic is completely **limit**.

- Judged whether new sentences preserved meaning and grammaticality

Results

	Meaning and	
Condition	Grammaticality	Meaning
automatic alignments	49%	55%
+ language model	55%	65%
+ multiple corpora	57%	65%
+ word sense disambiguation	62%	70%
manual alignments	75%	85%

Using paraphrases to improve SMT

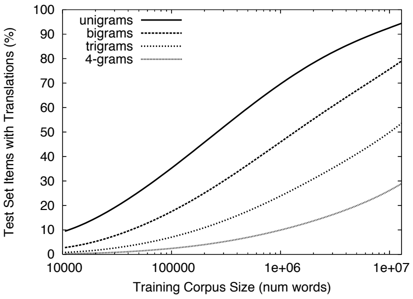
- Statistical machine translation learns the translations of words and phrases from examples

- Currently if a word is unseen then SMT will be unable to translate it

- If a phrase is unseen, but its individual words are, then SMT won't be as likely to produce a correct translation for it

We will try to use paraphrases to alleviate this problem

The extent of the problem



Behavior on unseen words

- A system trained on 10,000 sentences (≈200,000 words) may translate

  Es positivo llegar a un acuerdo sobre los procedimientos, pero debemos encargarnos de que este sistema no sea susceptible de ser usado como arma política.

  as

  It is good reach an agreement on procedures, but we must *encargarnos* that this system is not susceptible to be *usado* as political weapon.

- Since the translations of *encargarnos* and *usado* were not learned, they are either reproduced in the translation, or omitted entirely.

Substituting paraphrases then translating

encargarnos	
garantizar velar procurar asegurarnos	
usado	
utilizado empleado uso utiliza	

It is good reach an agreement on procedures, but we must *encargarnos* that this system is not susceptible to be *usado* as political weapon.

Substituting paraphrases then translating

encargarnos	?
garantizar velar procurar asegurarnos	guarantee, ensure, guaranteed, assure, provided ensure, ensuring, safeguard, making sure ensure that, try to, ensure, endeavour to ensure, secure, make certain
usado	?
utilizado empleado uso utiliza	used, use, spent, utilized used, spent, employee use, used, usage used, uses, used, being used

It is good reach an agreement on procedures, but we must **guarantee** that this system is not susceptible to be **used** as political weapon.

Improvements in coverage

Coverage of	Before Parapreasing	After Paraphrasing
Unique 1-grams	48%	92%
Unique 2-grams	25%	73%
Unique 3-grams	10%	41%
Unique 4-grams	3%	20%

For a Spanish-English SMT system trained in 10,000 sentence pairs (approx. 210,000 words in each language), with paraphrases generated from parallel corpora between Spanish and Danish, Dutch, Italian, French, Finnish, German, Greek, Portuguese, and Swedish,

Average quality of translated paraphrase

Corpus size (sentences)	Single word Paraphrases	Multi-word Paraphrases
10,000	47%	48%
20,000	61%	52%
40,000	58%	55%

Prior to paraphrasing *none* of the unseen words were translating correctly.

Final thoughts

- The data for statistical MT can be used for other tasks, such as paraphrasing
- Paraphrases can be applied to many natural language processing tasks
- Paraphrases can help to overcome the lack of generalization in SMT