
Paraphrasing and Translation

Chris Callison-Burch

November 28, 2007



Talk Outline

- Paraphrasing with bilingual parallel corpora
 - Generating paraphrases from data
 - Ranking paraphrases with probabilities
 - Evaluating quality of paraphrases
 - Results
- Improved statistical MT with paraphrases
 - The problem of coverage in SMT
 - Paraphrases as a potential solution
 - Experimental design and evaluation methodologies
 - Results and conclusions

Usefulness of paraphrases

- Paraphrases are **alternative ways** of conveying the **same information**
- Useful in NLP applications such as:
 - Generation
 - Question answering
 - Multidocument summarization
 - Automatic evaluation of MT and summaries
- How to get them?

Paraphrasing with monolingual parallel data

- Previous work used [monolingual parallel corpora](#)
- Multiple translations of the same source
 - Multiple translations of French novels into English
 - Evaluation data for Bleu metric
- Comparable corpora are similar
 - Encyclopedia articles on same topic
 - Different newspapers' accounts of one event

Paraphrasing with monolingual parallel data

- Find identical contexts in aligned sentences:

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**

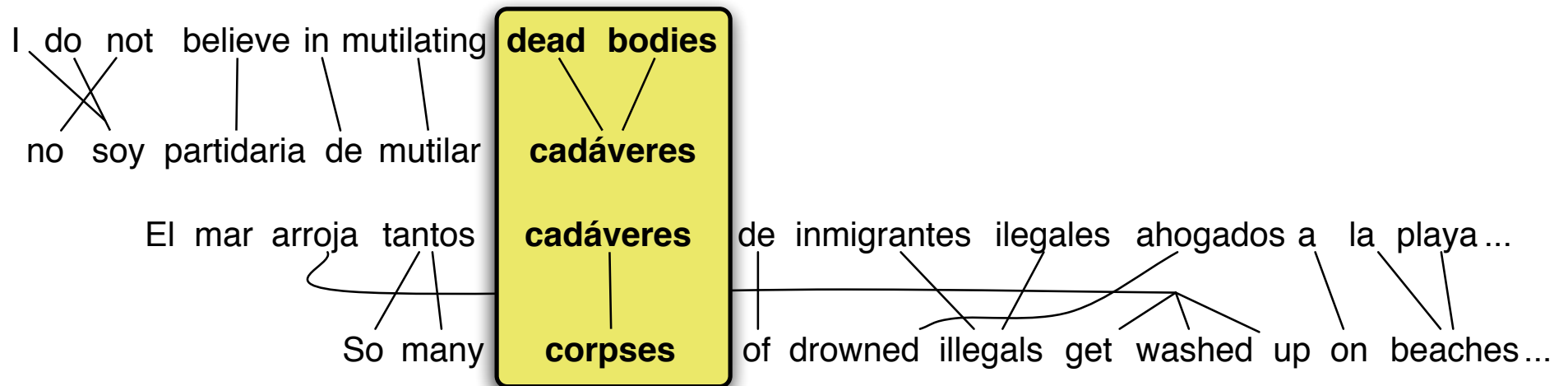
Potential problems with this method

- Monolingual parallel corpora are relatively **uncommon**
- **Limits** what paraphrases we can generated
 - Limited number of paraphrases
 - Constrained to a few genres

Paraphrasing with bilingual parallel corpora

- Bilingual parallel corpora are **much more common**
- However, no longer contain identical contexts
- Adopt techniques from **phrase-based** statistical MT
- Use aligned foreign language phrase as pivot

Paraphrasing with bilingual parallel corpora



Example paraphrases

- **dead bodies** → corpses, carcasses, bodies, skeletons, people
- **military force** → armed forces, defense, 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

Multiple paraphrases

- Generate multiple paraphrases for a given phrase
- Want to be able to **rank** them
- Define a **paraphrase probability**

Paraphrase probability

- Defined in terms of [translation model probability](#)
- [Naturally](#) falls out of parallel corpus

$$\begin{aligned}\hat{e}_2 &= \arg \max_{e_2 \neq e_1} p(e_2|e_1) \\ &= \arg \max_{e_2 \neq e_1} \sum_f p(f|e_1)p(e_2|f) \\ &\approx \arg \max_{e_2 \neq e_1} \sum_f \frac{\text{count}(f, e_1)}{\text{total}(e_1)} * \frac{\text{count}(e_2, f)}{\text{total}(f)}\end{aligned}$$

Context matters

- Further refined method to use [context](#)
- Use surrounding words in context with [language model](#) probability
- Use WSD algorithm to restrict paraphrases to [same word sense](#)

Paraphrase evaluation

- Previous work used **poor** evaluation methodology
- Asked judges whether paraphrases had “approximate conceptual equivalence” or were “roughly interchangeable given the genre”
- Show paraphrases **out of context**, or only substitute paraphrases into original context

Paraphrase evaluation

Is **console** a good paraphrase for **comfort**?

Emma burst into tears and he tried to **comfort** her, saying things to make her smile.

Emma burst into tears and he tried to **console** her, saying things to make her smile.

Paraphrase evaluation

Is **console** a good paraphrase for **comfort**?

Emma burst into tears and he tried to **comfort** her, saying things to make her smile.

Emma burst into tears and he tried to **console** her, saying things to make her smile.

What about now?

George Bush said Democrats provide **comfort** to our enemies.

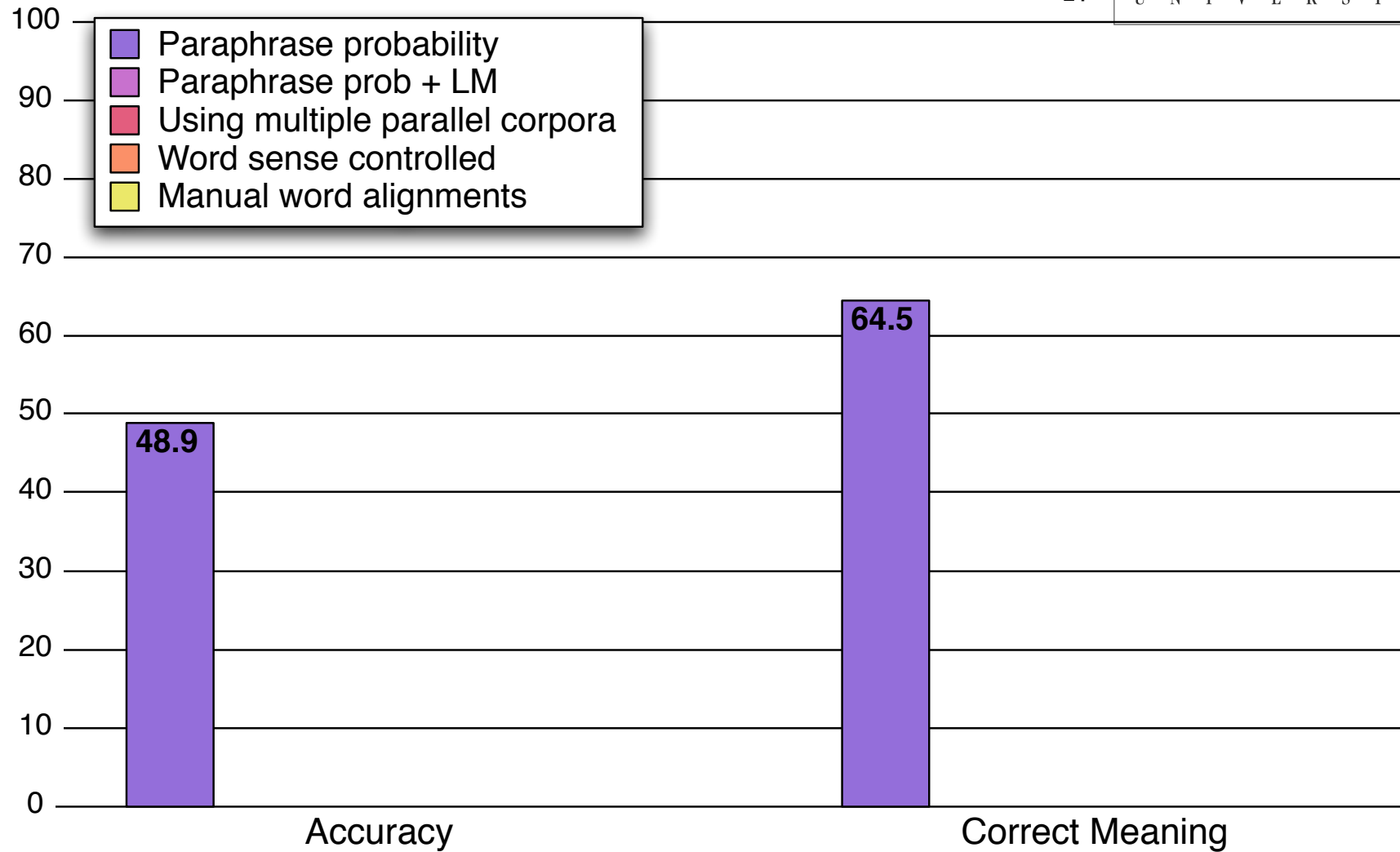
George Bush said Democrats provide **console** to our enemies.

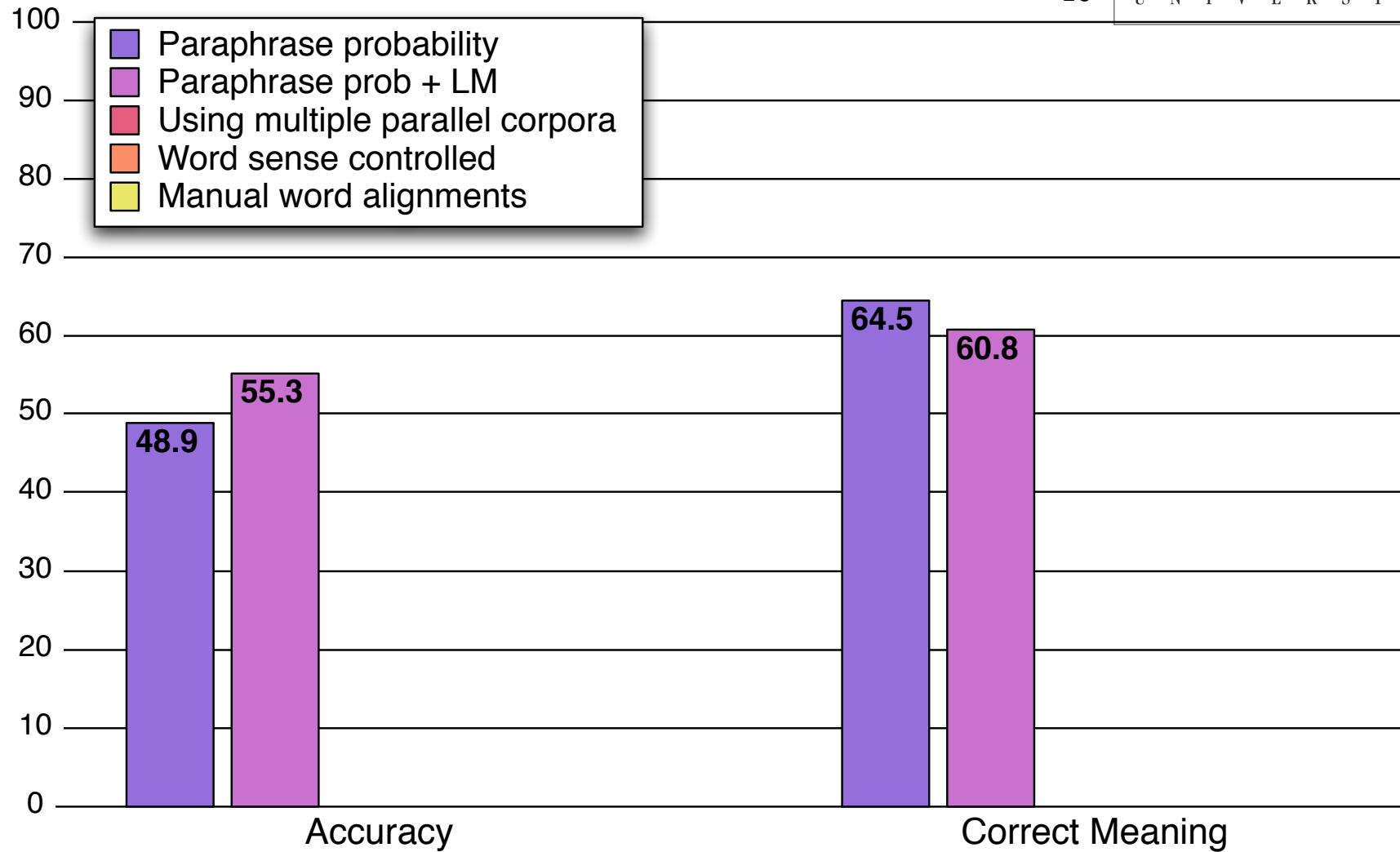
Paraphrase evaluation

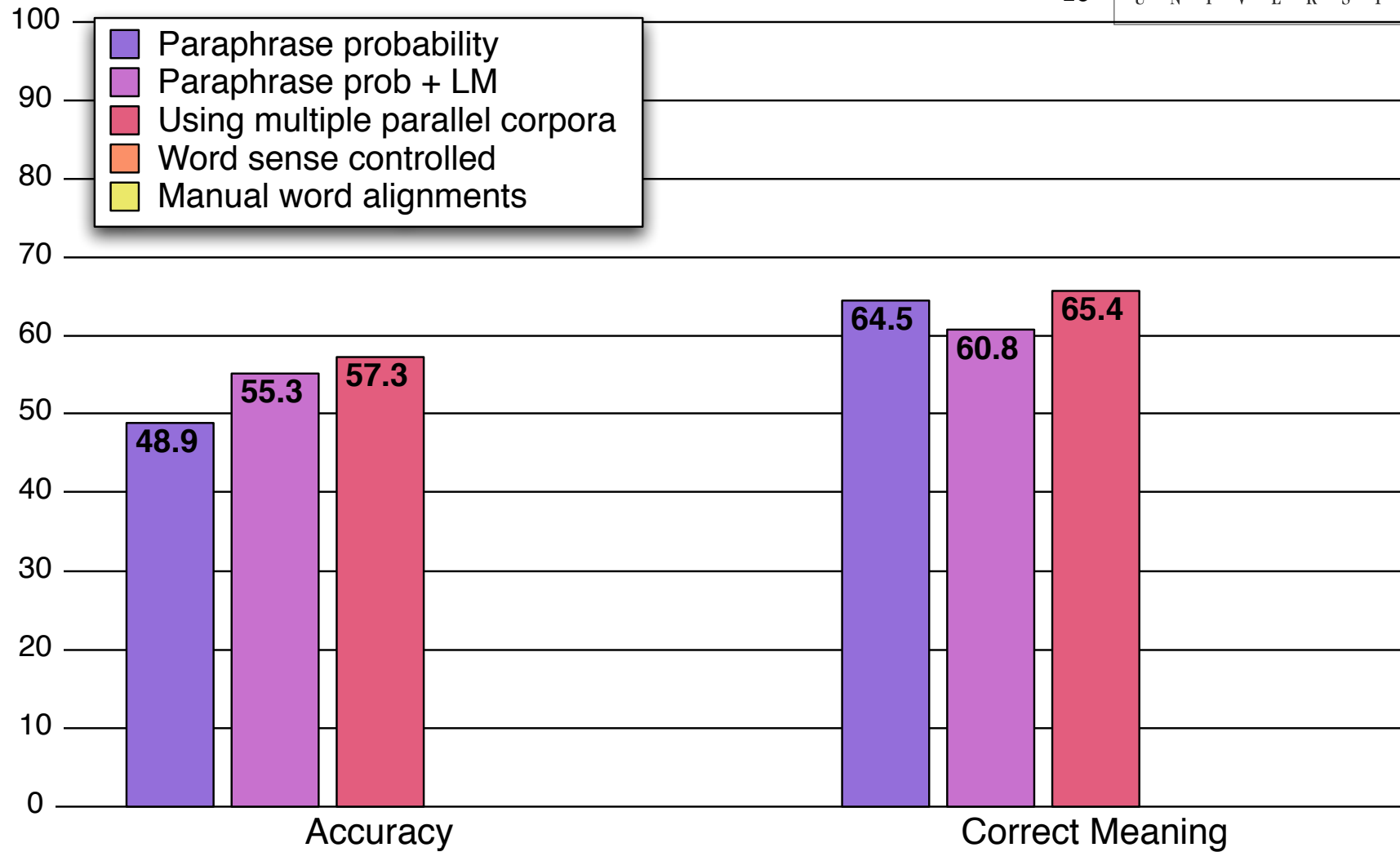
- We substituted paraphrases into a number of **new contexts**
- Asked judges whether paraphrases **preserved meaning** and **remained grammatical**
- Two measures: accuracy and correct meaning

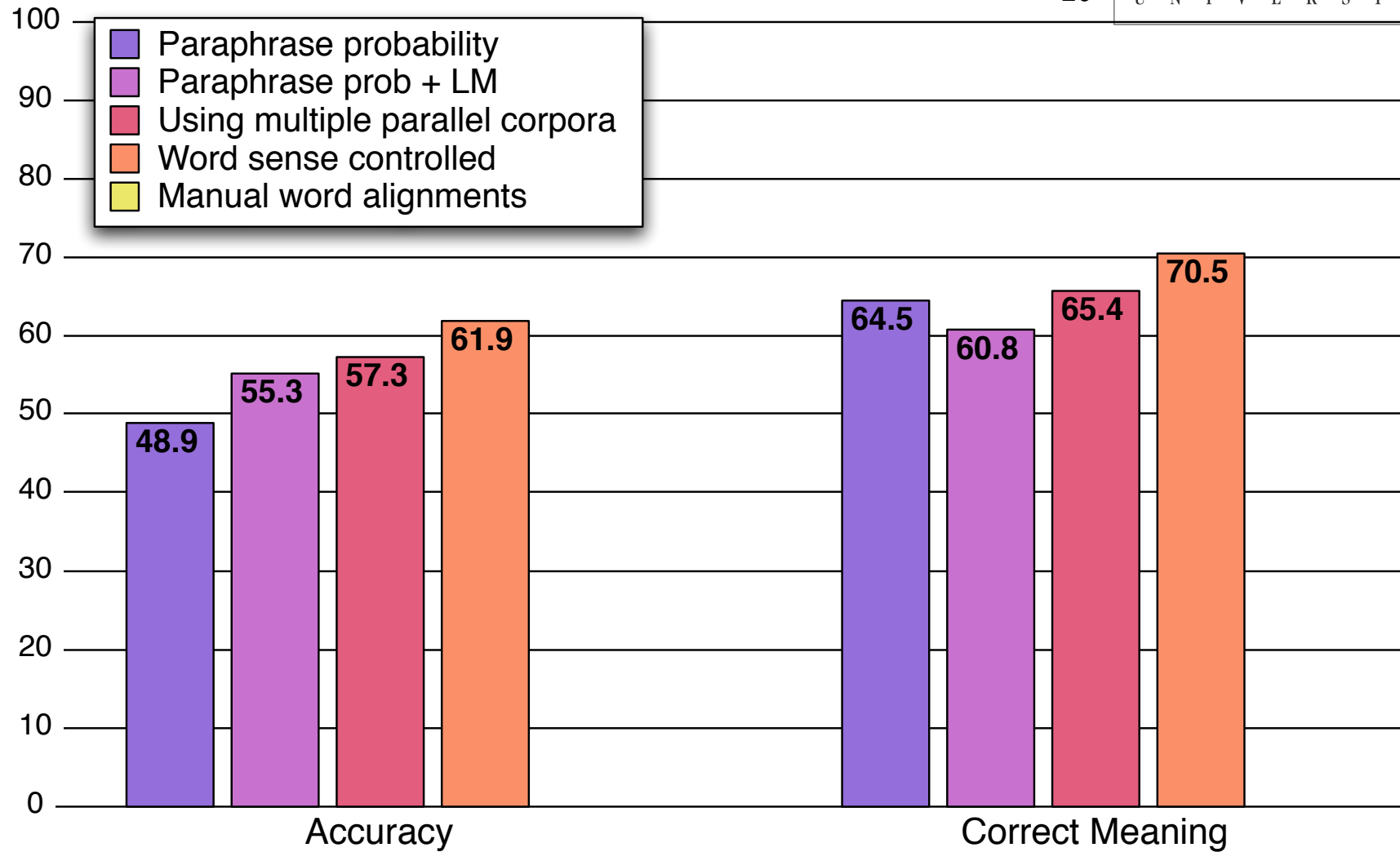
Experimental conditions

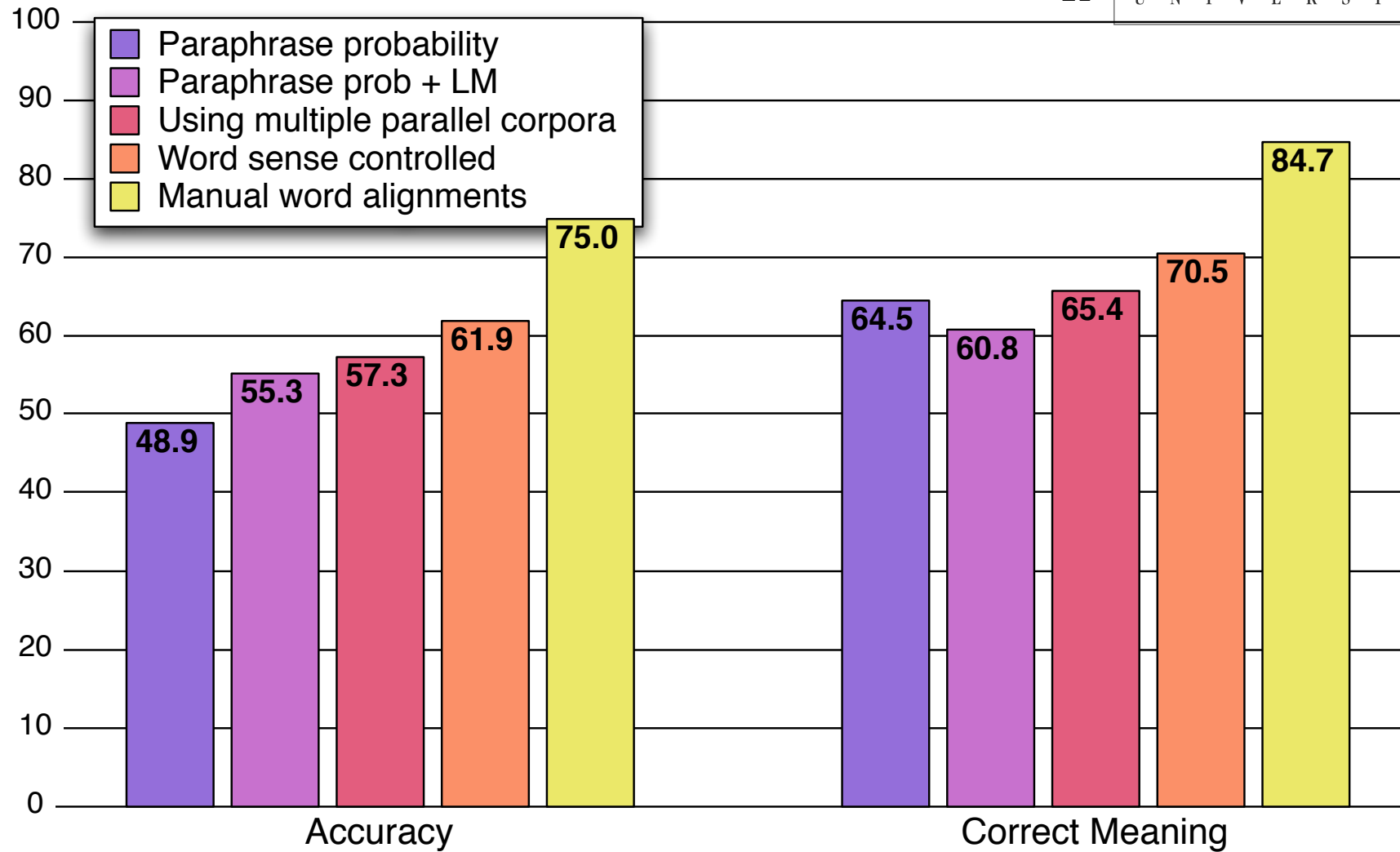
- Extracted paraphrases from parallel corpora
- Evaluated quality under various conditions
 1. Ranking paraphrases with **paraphrase probability**
 2. Ranking paraphrases in context with a **language model**
 3. Using **multiple** parallel corpora
 4. Limiting paraphrases by **word sense**
 5. **Manual** word alignments











Conclusions for Part 1

- Can generate **high quality** paraphrases bilingual parallel corpora
- Quality improvements seen with refined **paraphrase probability**
- Data is **more common** than in previous approaches
 - Can generate more paraphrases for wide variety of genres
 - Can be easily applied to different languages
- How useful are paraphrases when **applied to a task**?

Part 2:

Using Paraphrases to Improve Statistical Machine Translation

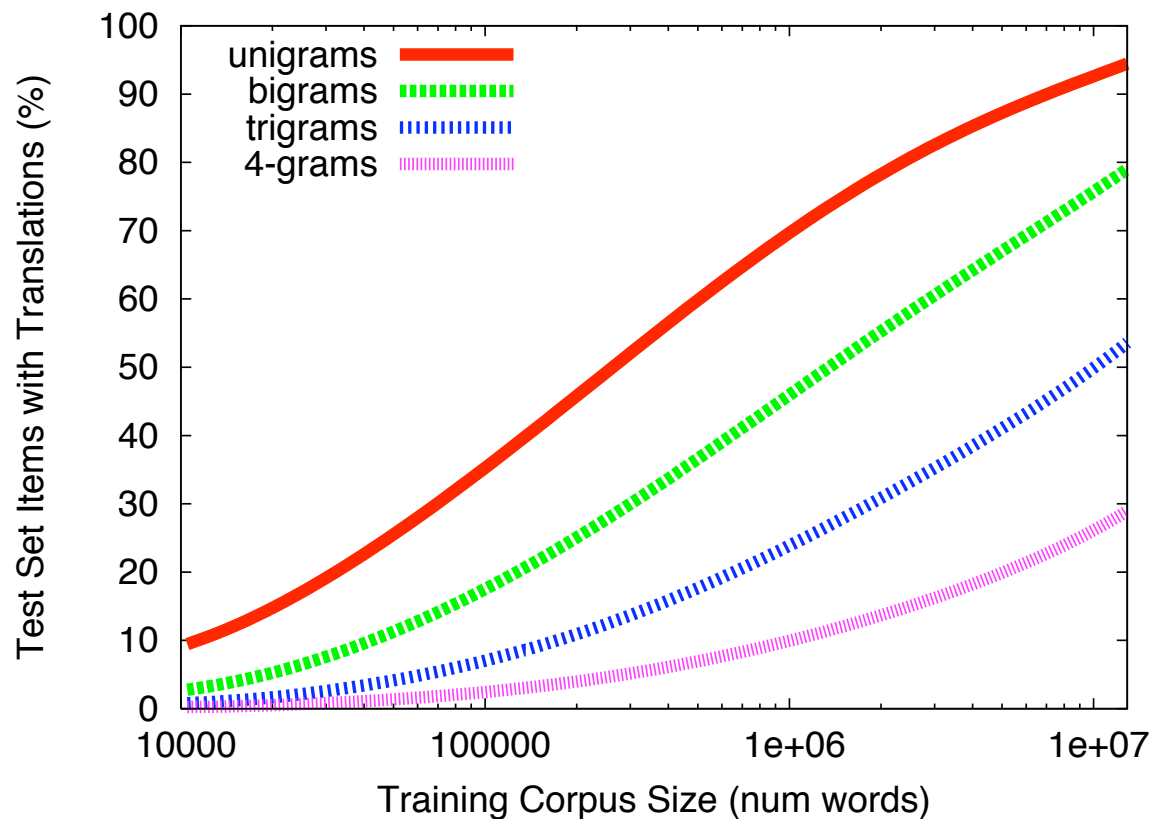
Improving SMT with paraphrases

- Different places paraphrases could be applied
 - Transform source to something **easier to translate**
 - Flexibly match references for **automatic evaluation** [Zhou et al, 2006]
 - Expand references for **parameter tuning** [Madnani et al, 2007]
 - Create equivalence classes to **improve word alignments**
- Here we focus on paraphrasing source to **improve coverage**

Problem of coverage in SMT

- Statistical machine translation learns the translations of words and phrases from **parallel corpora**, which are often **limited in size**
- 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 will be **less likely** to produce a correct translation for it

The extent of the problem



Behavior on unseen words

- A system trained on 10,000 sentences ($\approx 200,000$ 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	?
usado	?

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

Wait a minute... isn't this circular?

- How can we **improve coverage** when paraphrases come from **same data**?
- Paraphrases can come from **other** parallel corpora
- Imagine we're translating from **English into Maltese**, and have a very small English-Maltese parallel corpus
- We can increase coverage by generating **English paraphrases** from
 - **English-Arabic**
 - **English-French**
 - **English-German**
 - etc.

Our Data

- Used various sized portions of the Europarl [Spanish-English parallel corpus](#) to train [translation models](#) / create phrase tables
- Used all [other Spanish parallel corpora](#) to train the [paraphrase model](#)
- Specifically, we used these parallel corpora for [paraphrasing Spanish](#):
 - [Spanish-Danish](#)
 - [Spanish-Dutch](#)
 - [Spanish-Finnish](#)
 - [Spanish-French](#)
 - [Spanish-German](#)
 - [Spanish-Italian](#)
 - [Spanish-Portuguese](#)
 - [Spanish-Swedish](#)

Experimental Setup

- **Generated paraphrases** for every Spanish test phrase f_1 for which we failed to learn any translations
- We **expanded the phrase table** by adding an entry for f_1 with the English translations of each paraphrase f_2 for which we had learned translations
- New phrase table entries were augmented with **additional feature function** that incorporated the paraphrase probability

Additional feature function

- Paraphrase probability can be used as additional feature function in the log linear formulation of SMT

$$h(\mathbf{e}, \mathbf{f}_1) = \begin{cases} p(\mathbf{f}_2|\mathbf{f}_1) & \text{If phrase table entry } (\mathbf{e}, \mathbf{f}_1) \\ & \text{is generated from } (\mathbf{e}, \mathbf{f}_2) \\ 1 & \text{Otherwise} \end{cases}$$

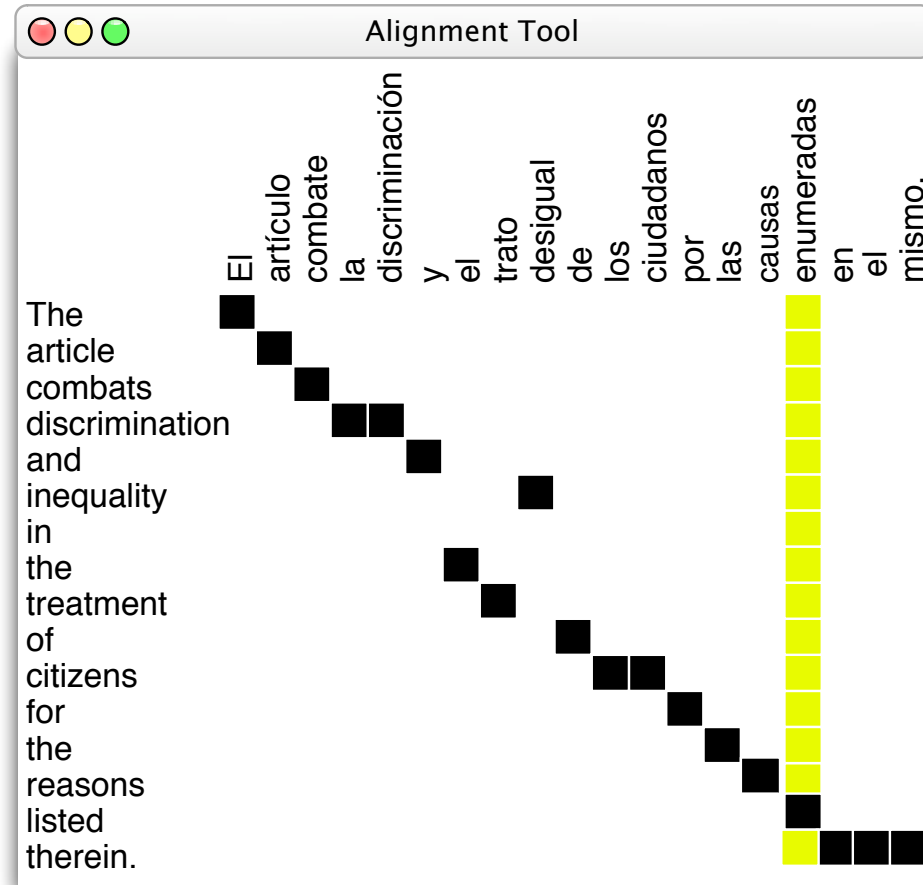
Conditions

- Baseline: Pharaoh decoder, standard feature functions, minimum error rate training, unknown words output as-is
- Single word paraphrases: Pharaoh, standard FFs plus paraphrase prob FF, MERT, phrase table expanded by [paraphrasing unknown words](#) other than names, numbers
- Multi-word paraphrases: As above, but with phrase table expanded with translations of [unknown words and phrases](#)
- Various sized training corpora

Evaluation methodology

- Measured translation quality improvements in terms of
 - Improvements in **Bleu score** over baseline
 - Increase in **coverage** of expanded phrase table
 - **Accuracy** of newly translated phrases
- Judged accuracy of newly translated phrases through **targeted manual evaluation**

Focused manual evaluation



Focused manual evaluation

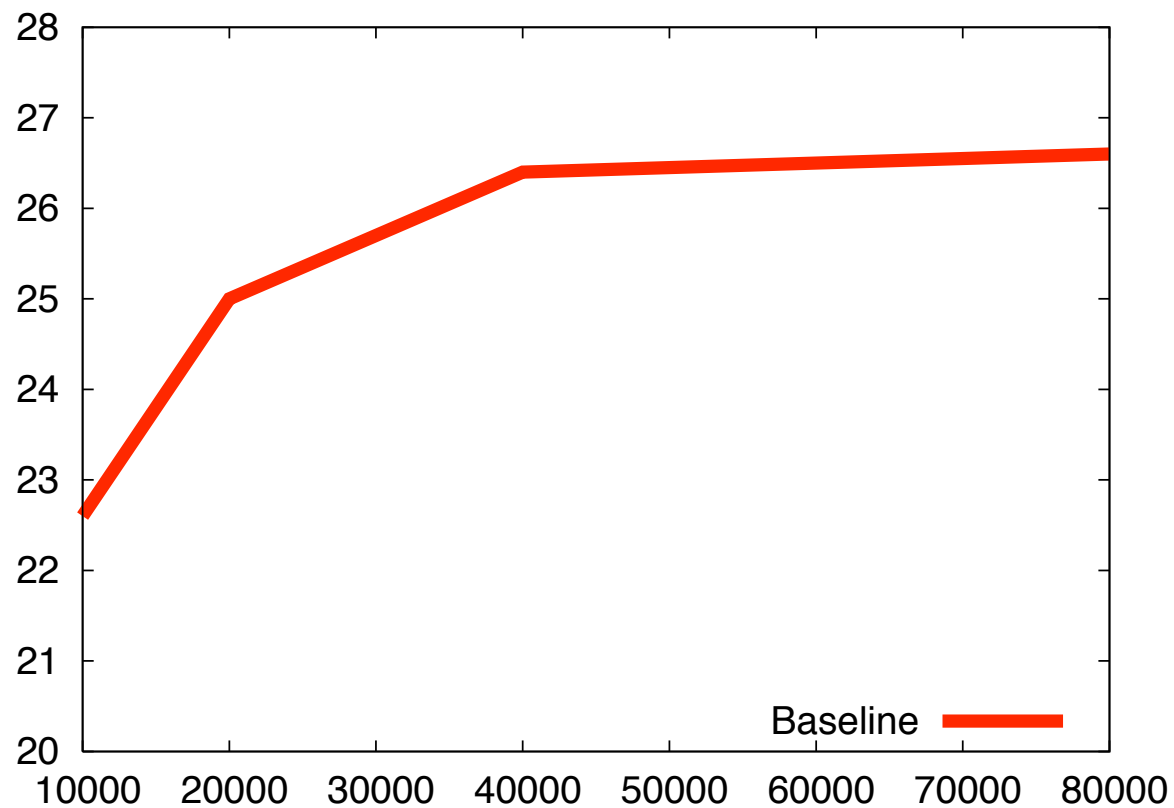
The article combats discrimination and inequality in the treatment of citizens for the reasons **listed** therein.

The article combats discrimination and the different treatment of citizens for the reasons **mentioned** in the same.

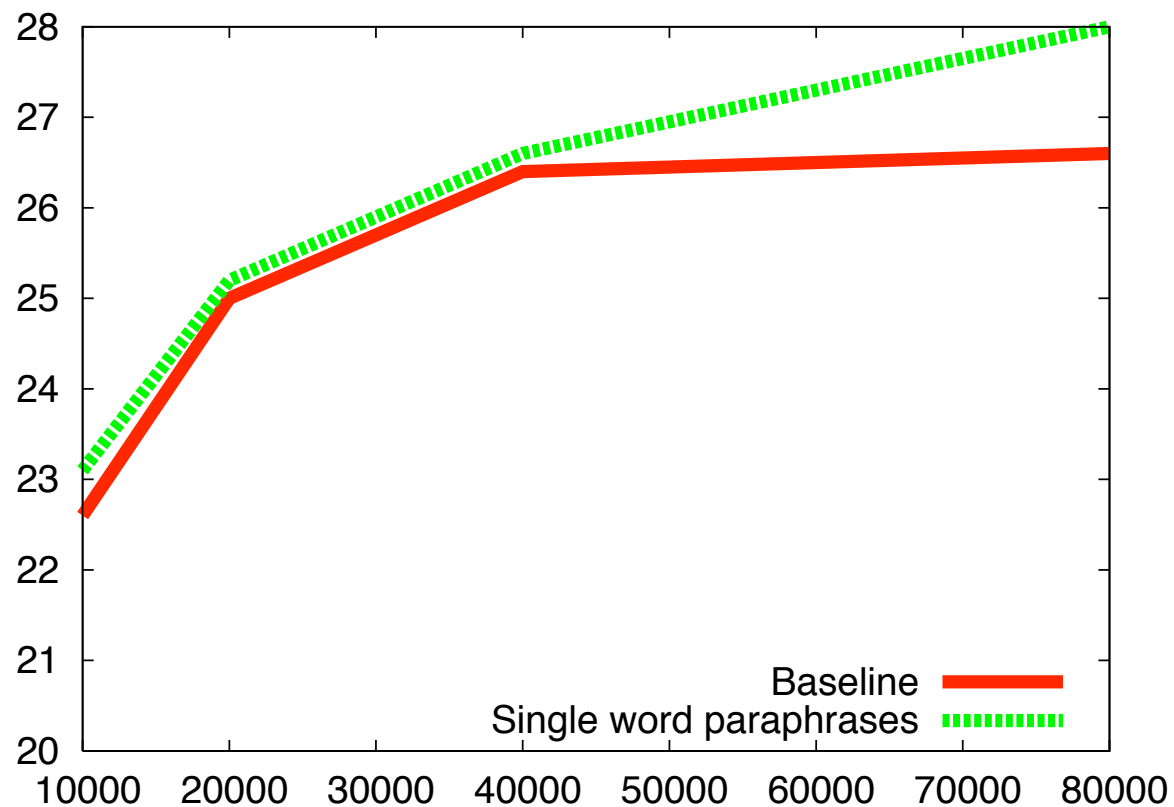
The article fights against uneven and the treatment of citizens for the reasons **set out** in the same.

The article is countering discrimination and the unequal treatment of citizens for the reasons **that** in the same.

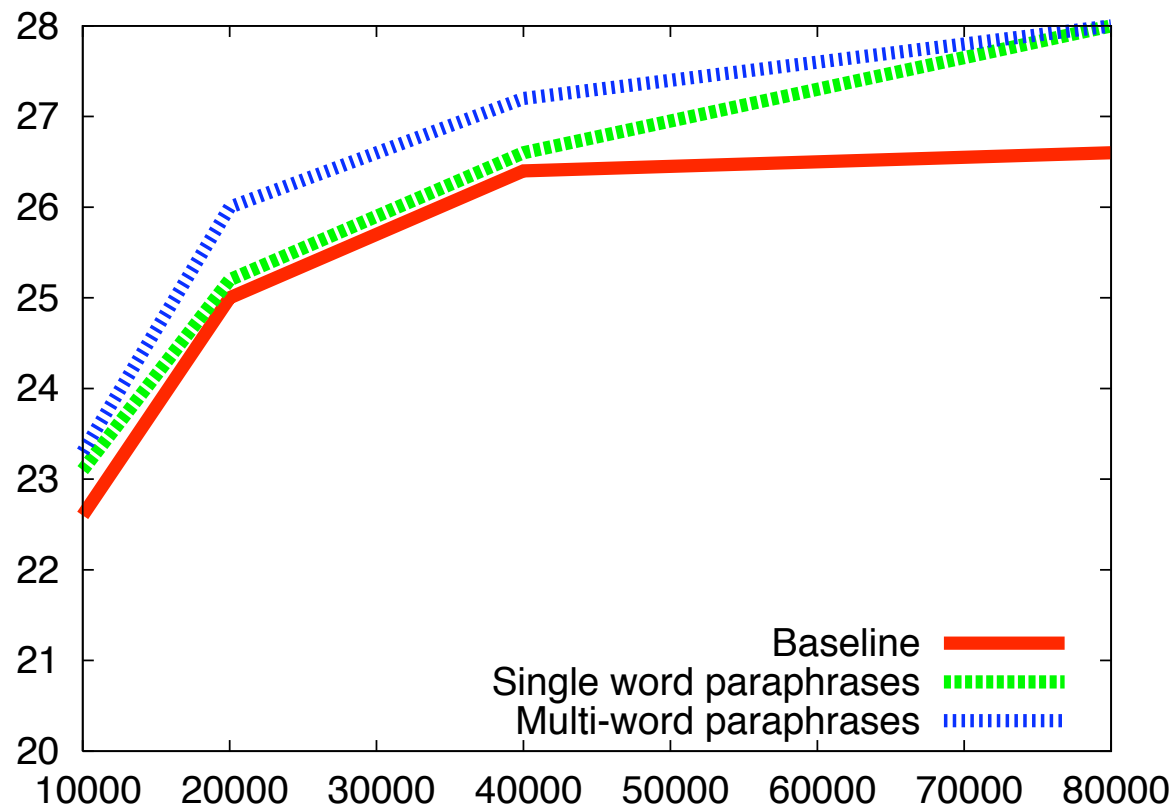
Results: Bleu Scores



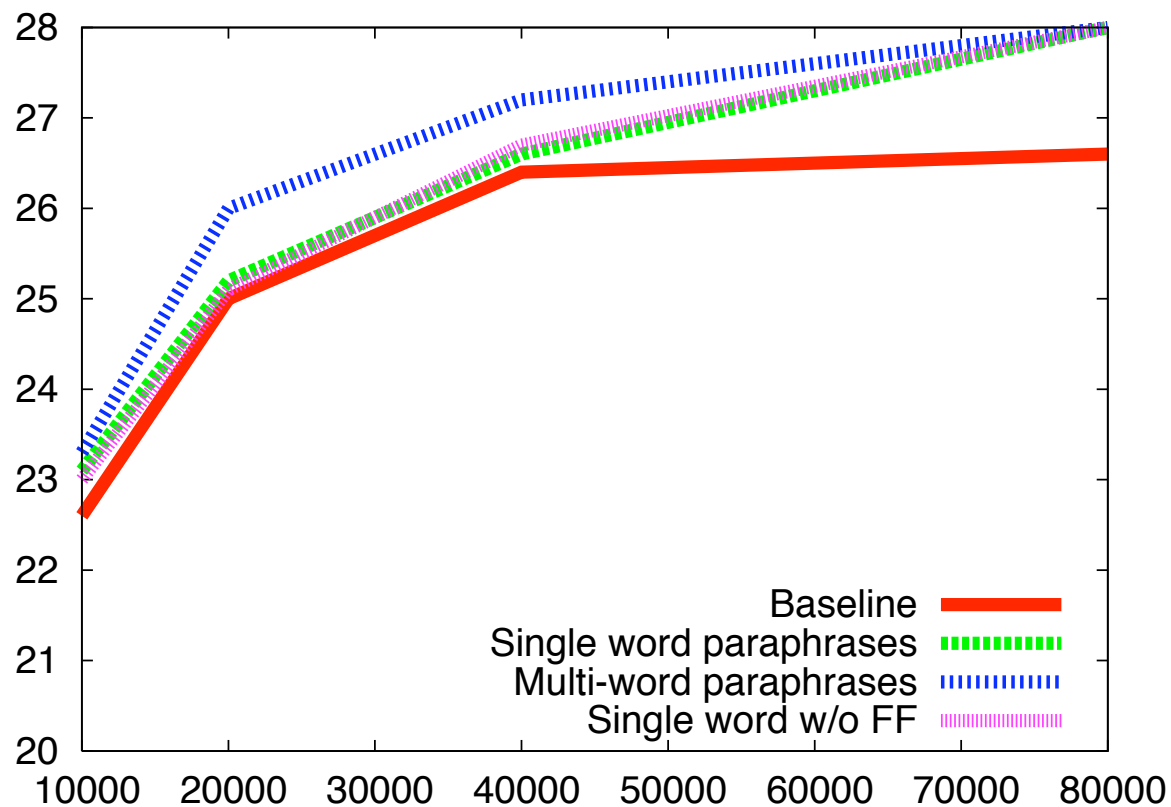
Results: Bleu Scores



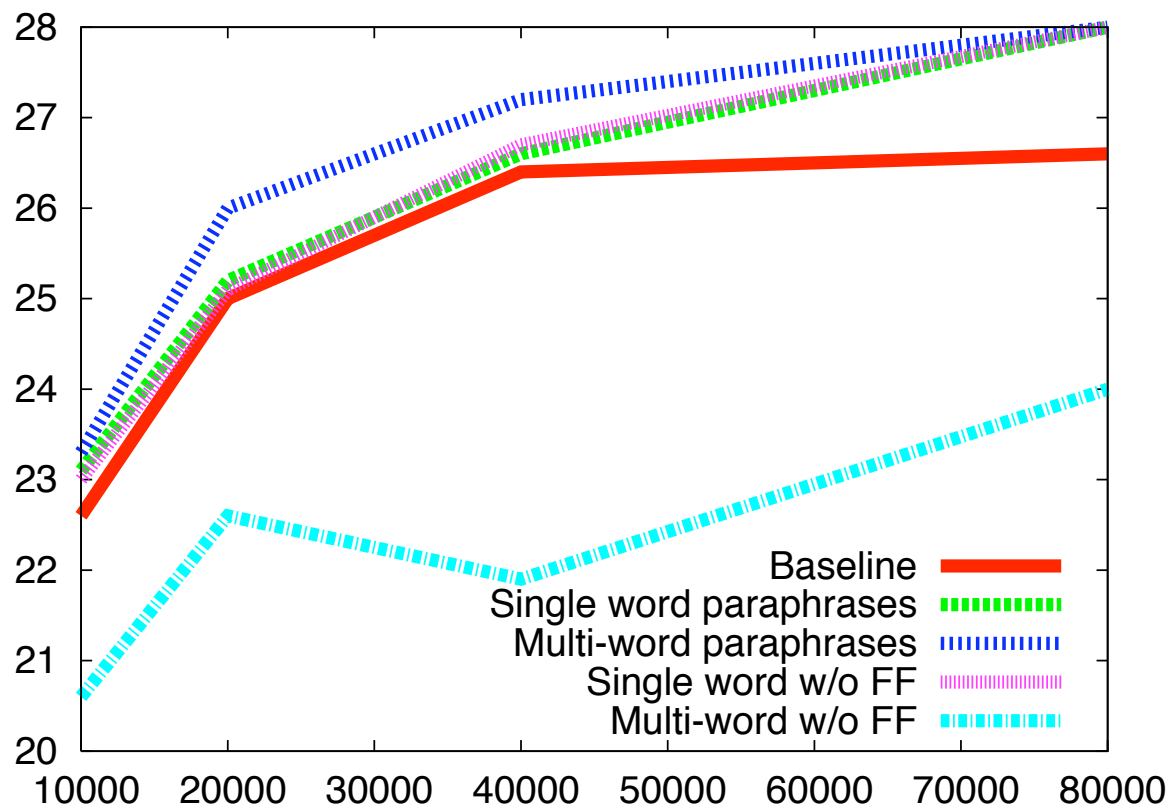
Results: Bleu Scores



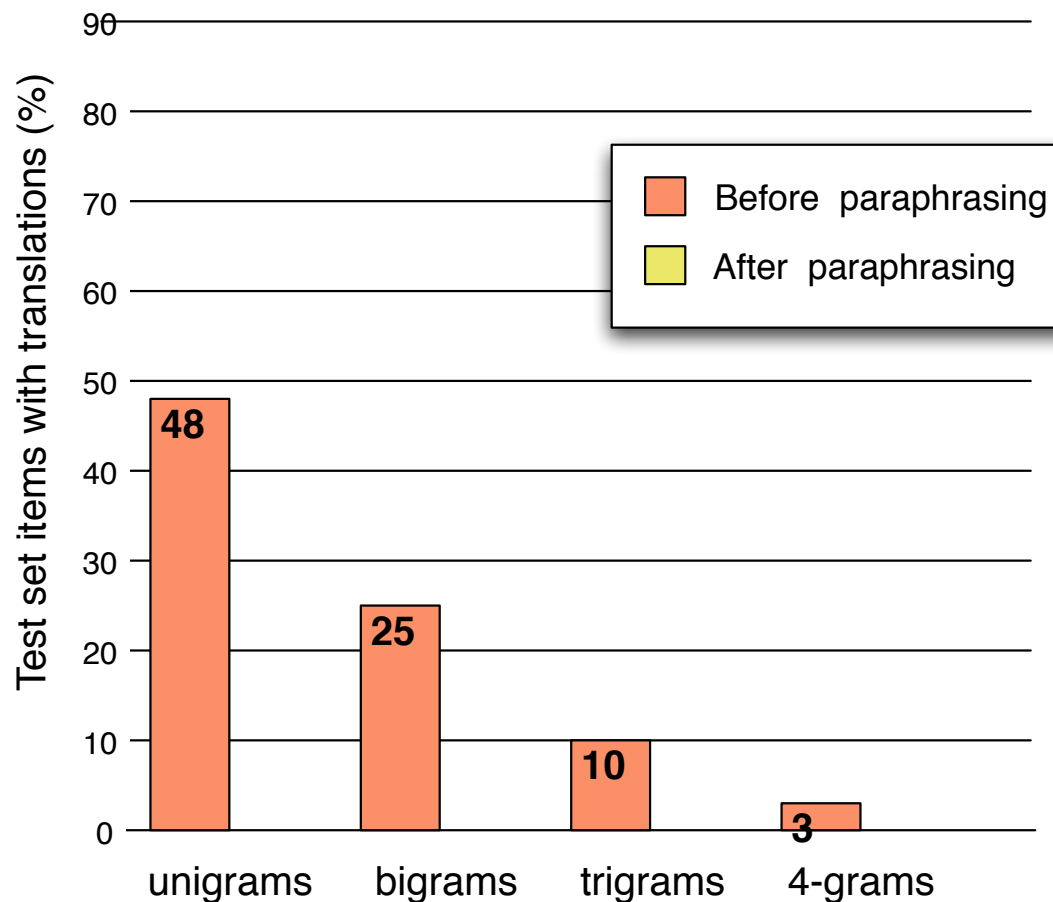
Importance of the paraphrase FF



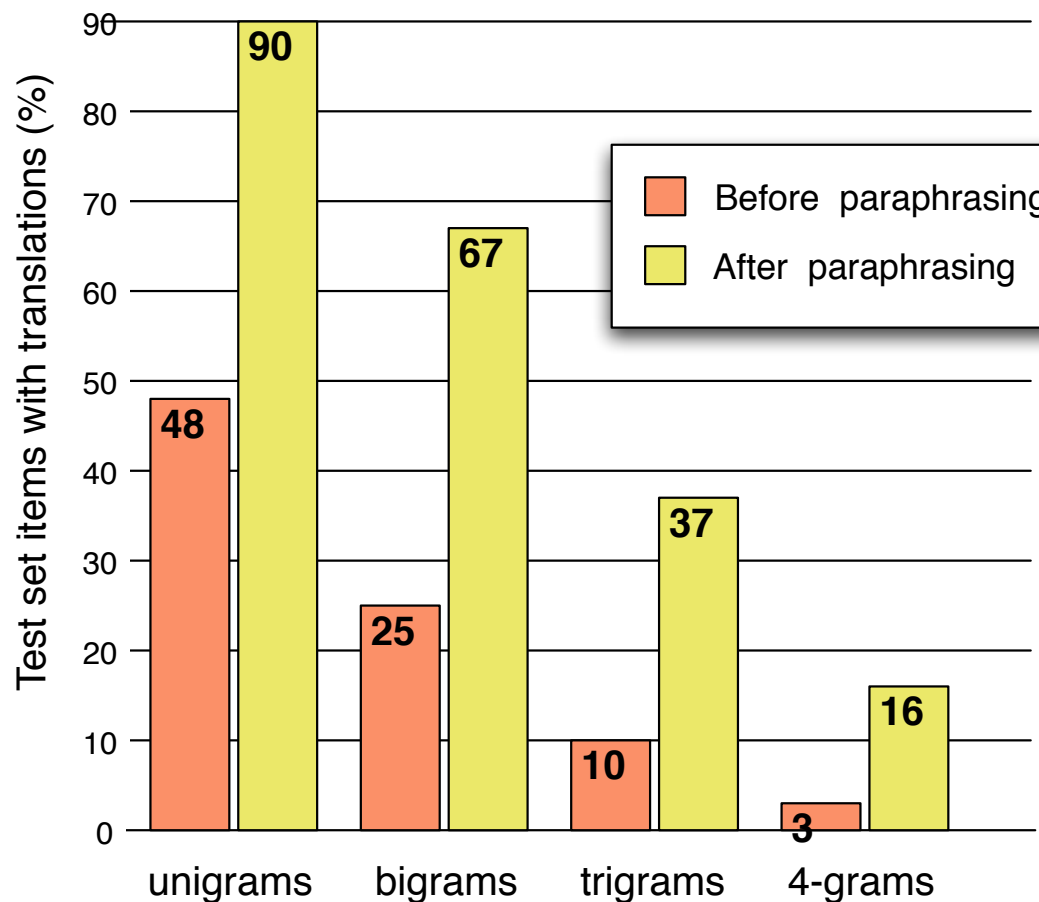
Importance of the paraphrase FF



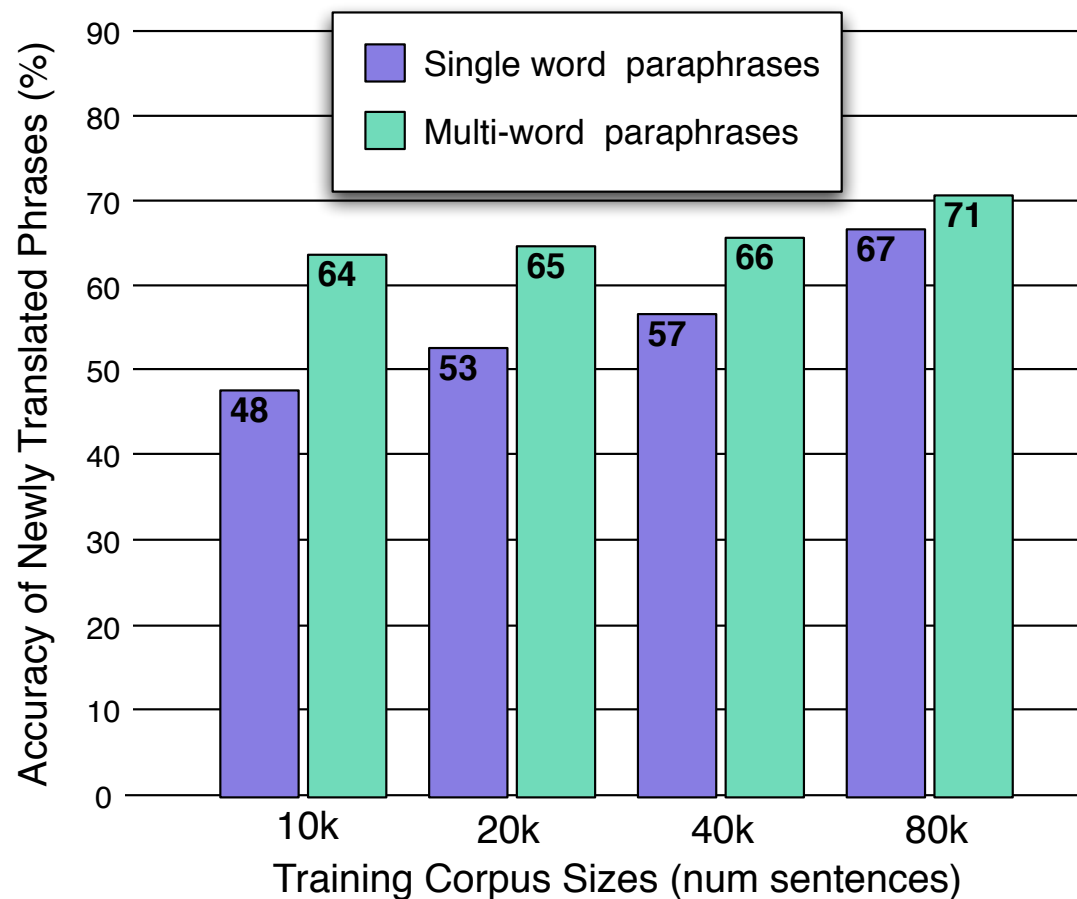
Results: Improvements in coverage



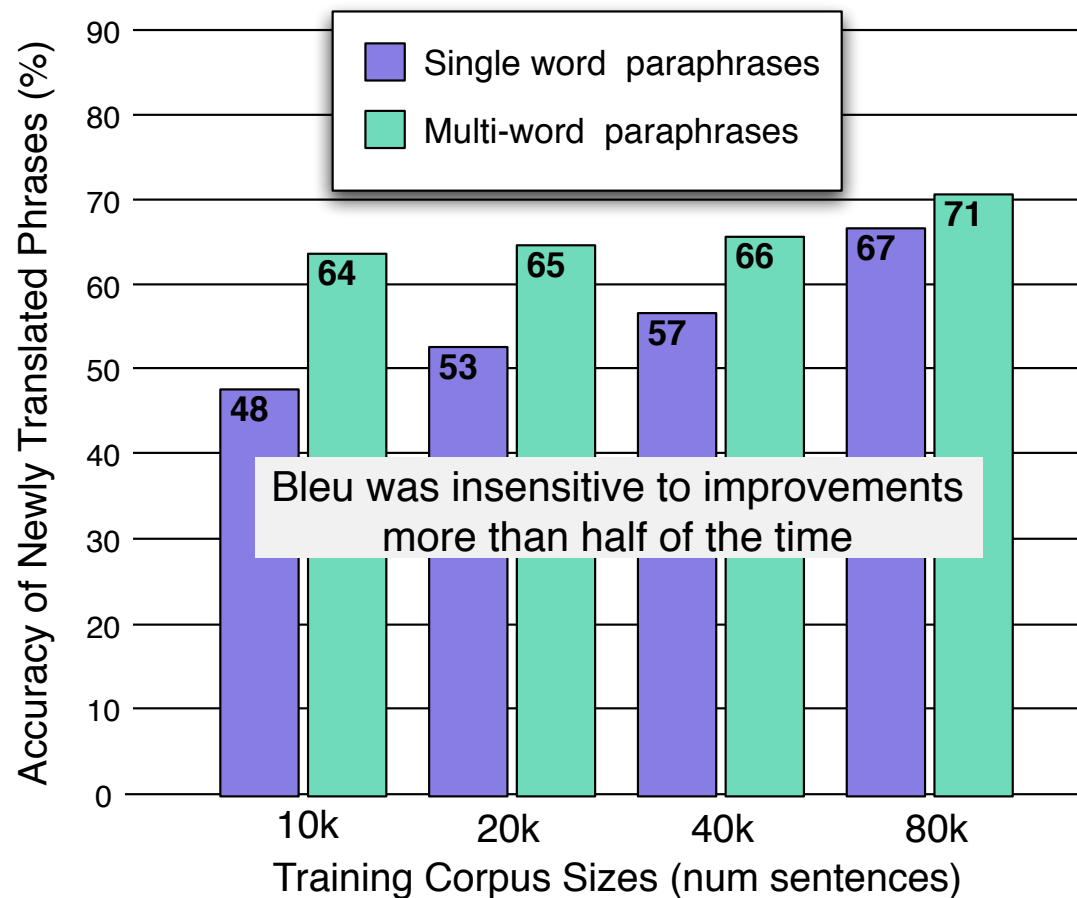
Results: Improvements in coverage



Results: Targeted Manual Evaluation



Results: Targeted Manual Evaluation



Conclusions

- Paraphrasing is an **effective solution** to deal with the **problem of coverage** in statistical machine translation
- Paraphrasing with bilingual corpora **ideally suited** to this task because of its multilinguality, probabilistic formulation, multi-word phrases, and high recall
- For small parallel corpus vocab coverage increases from 48% to 90%, with **more than half** of newly covered items **translating accurately**

Current Work & Broader Vision

- Current Work:
 - Improve statistical models of translation through increased **generalization**
 - Generate **more sophisticated paraphrases** including syntactic transformations
- Broader Vision:
 - Unlike data for other statistical NLP tasks, such as parsing, parallel corpora are created naturally by other human industry
 - Harnessing the data through savvy manipulation will improve translation, and can potentially **improve all other language technologies**

Thank you!

Selected Publications

Callison-Burch et al. “(Meta-) Evaluation of Machine Translation” in WMT-2007.

Callison-Burch, Koehn and Osborne. “Improved Statistical Machine Translation Using Paraphrases” in NAACL-2006.

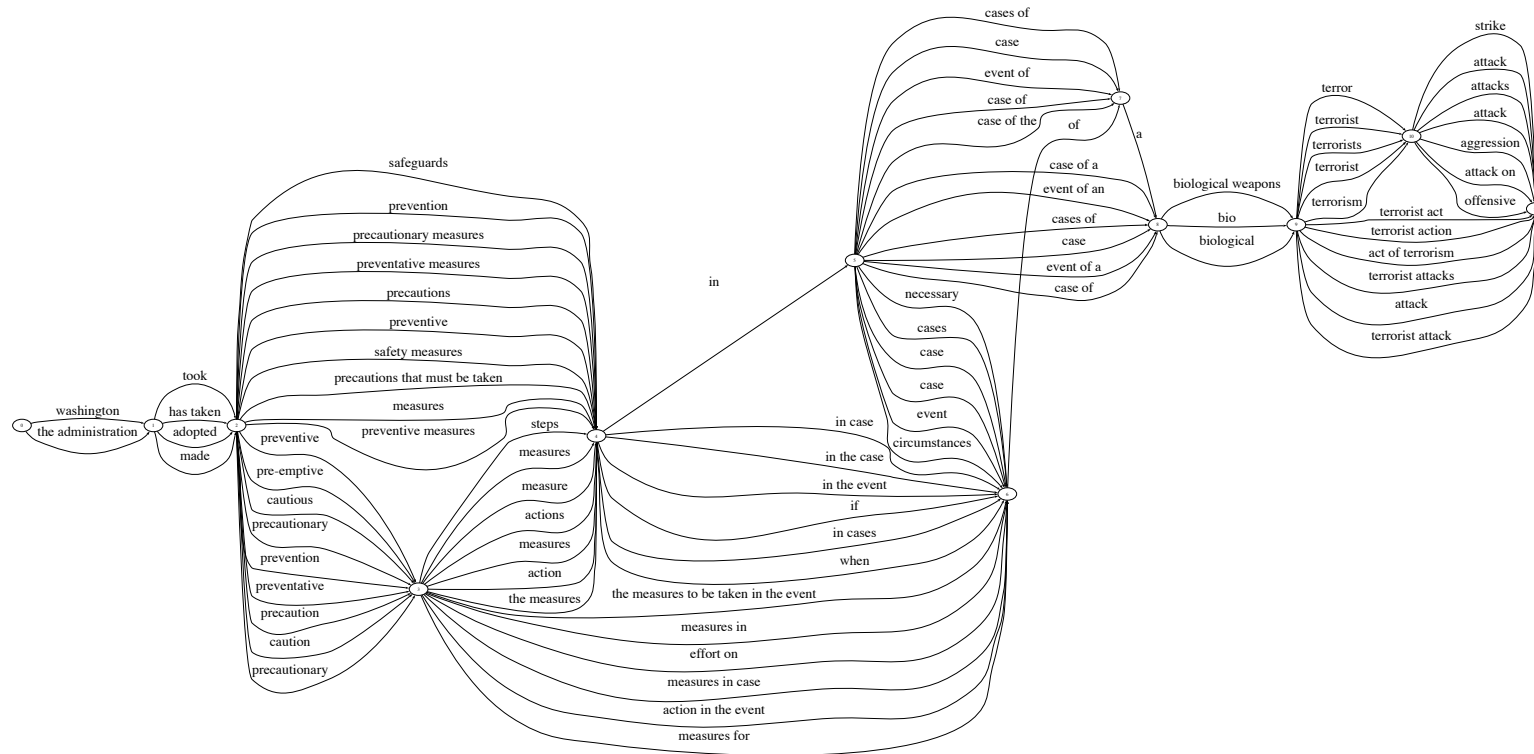
Callison-Burch, Osborne and Koehn. “Re-evaluating the Role of Bleu in Machine Translation Research” in EACL-2006.

Bannard and Callison-Burch. “Paraphrasing with Bilingual Parallel Corpora” in ACL-2005.

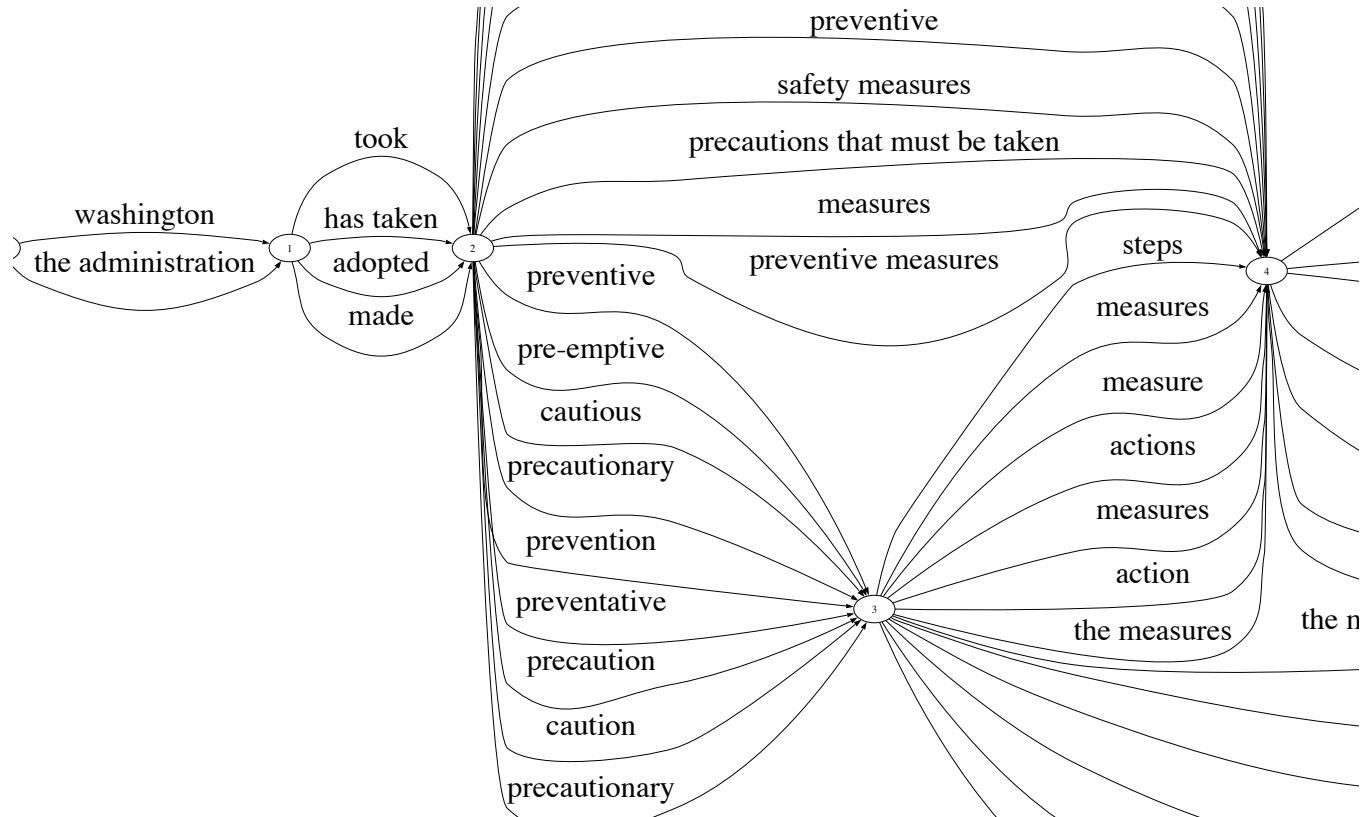
Callison-Burch, Bannard and Schroeder. “Scaling Phrase-Based Statistical Machine Translation to Larger Corpora and Longer Phrases” in ACL-2005.

Callison-Burch, Talbot and Osborne. “Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora” in ACL-2004.

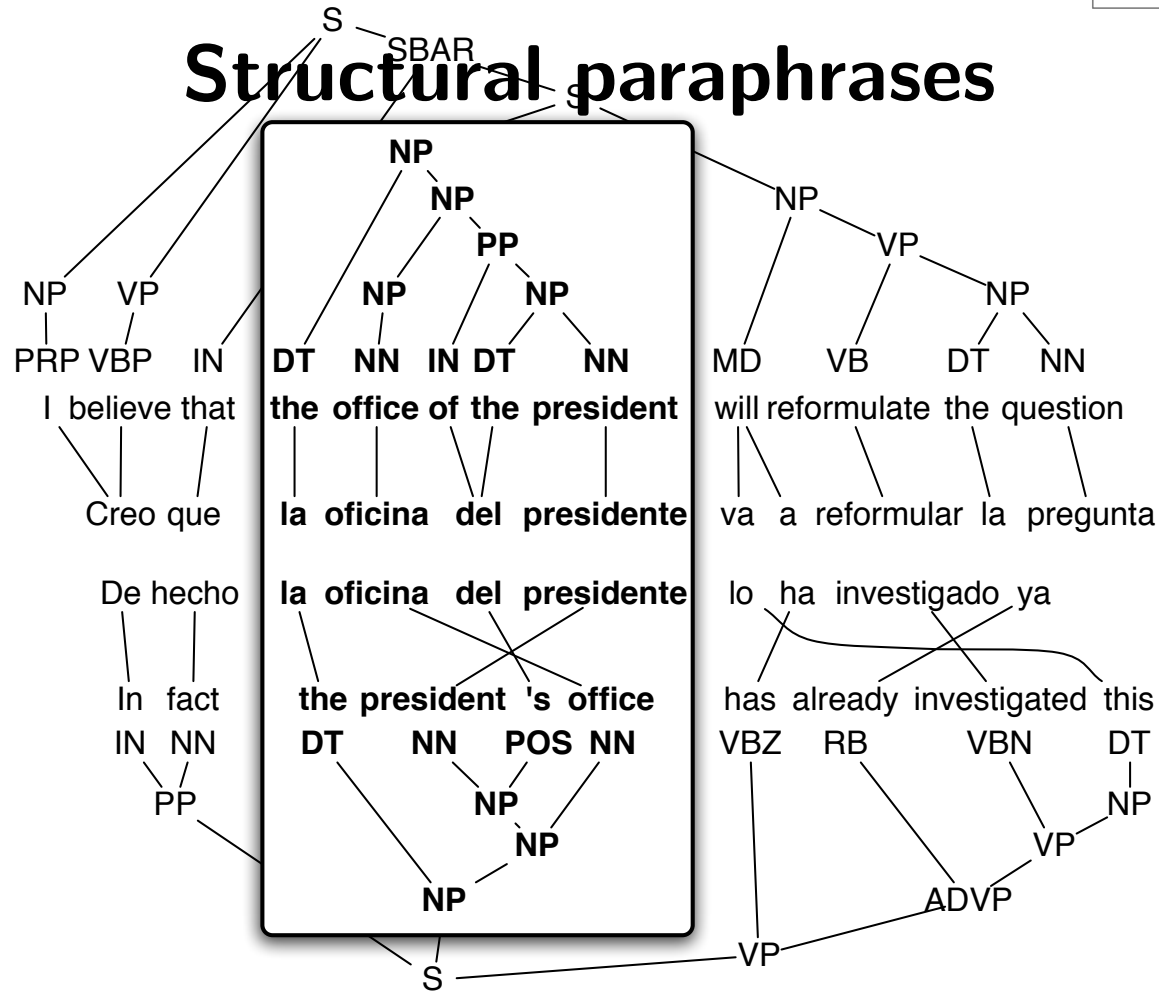
Paraphrase Lattices



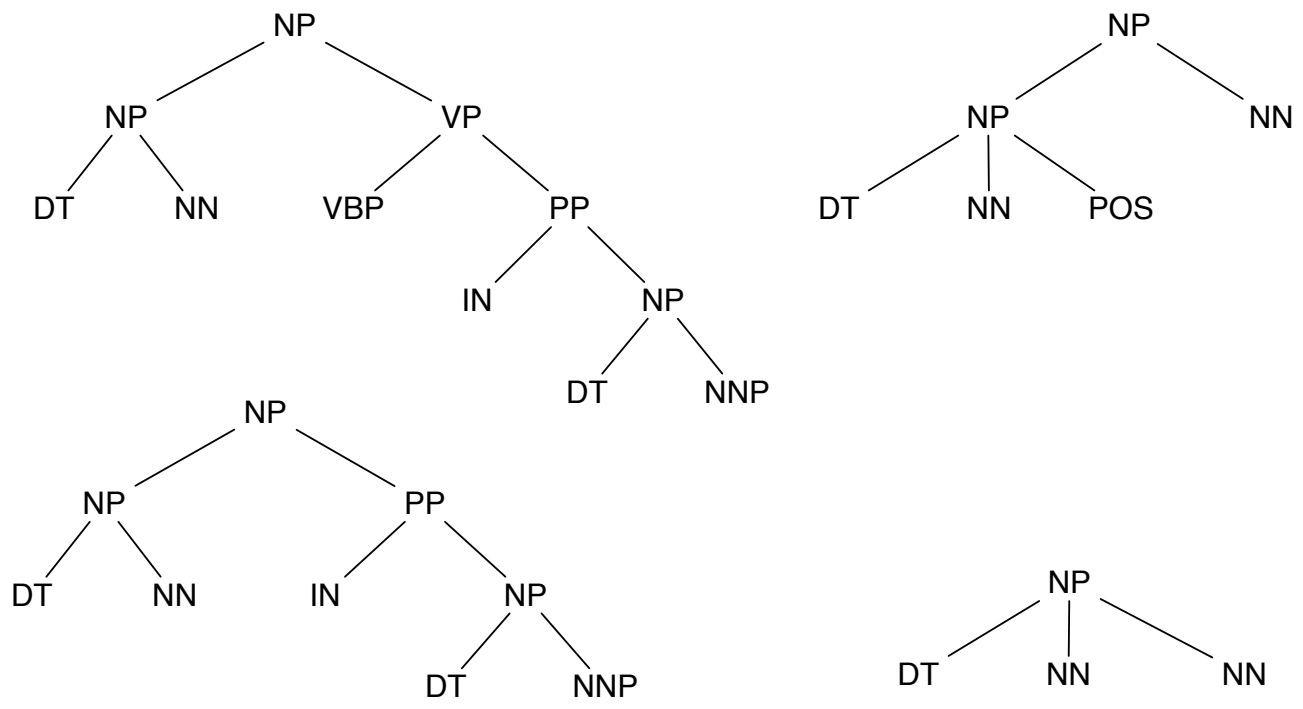
Paraphrase Lattices



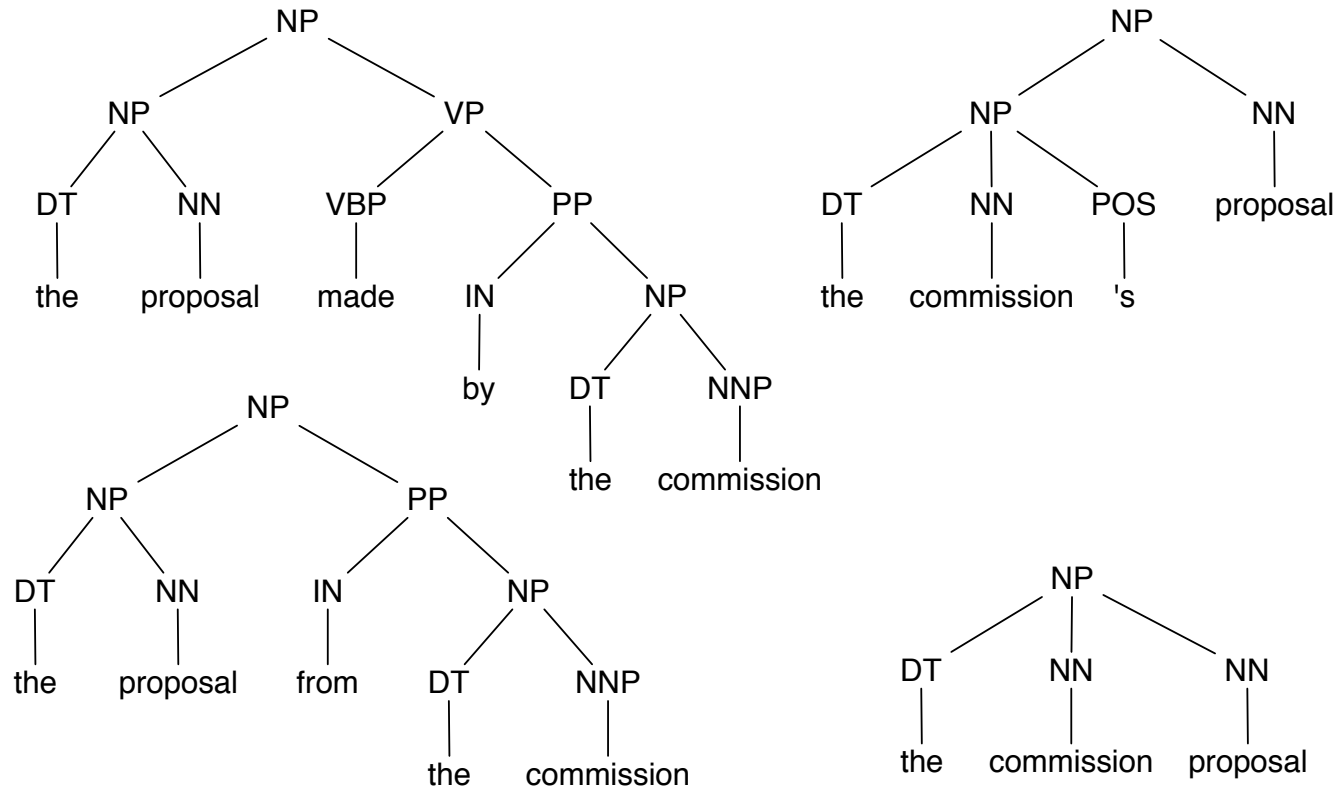
Structural paraphrases



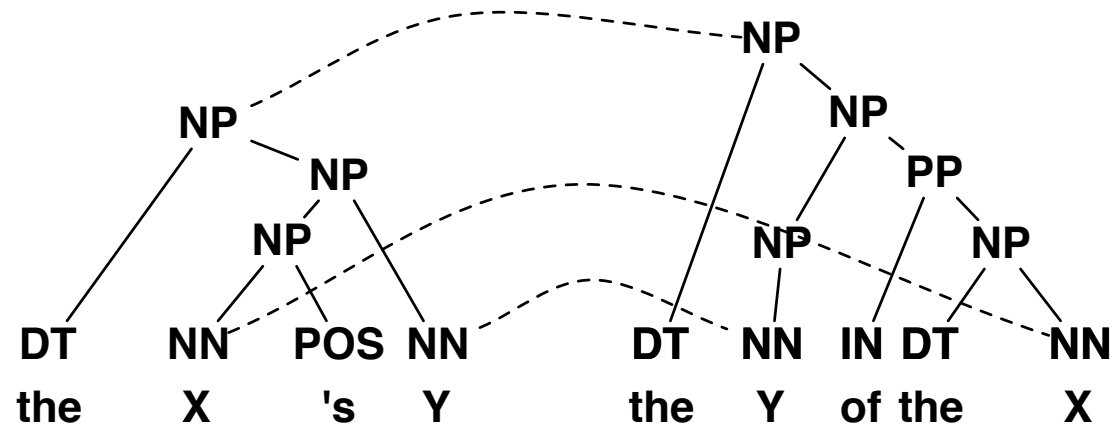
Structural paraphrases



Structural paraphrases



Structural paraphrases



Structural paraphrases

In addition to lexical paraphrases we can learn **structural paraphrases**:

JJ PRP MD VB TO VB → **MD PRP VB IN VBG** (.09)
 → **PRP VBP TO VB IN VBG** (.07)
 → **RB PRP MD VB TO VB** (.06)

JJ - Adjective
PRP - Personal pronoun
MD - Modal
VB - Verb, base form
TO - to
IN - Preposition or subordinating conjunction
VBG - Verb, gerund or present participle
VBP - Verb, non-3rd person singular present
RB - Adverb

Structural paraphrases

- Given **JJ PRP MD VB TO VB** → **MD PRP VB IN VBG** we might be able to transform
- **First I would like to thank** → **May I begin by thanking**
- This is quite a complex transformation:
 - The infinitival verb becomes a gerund
 - The the verb **begin** takes the place of the adverb **first**
 - The selection of which modal verbs is quite complex
- Promising result, which requires further research

Amount of data used to generate paraphrases

- We did use a tremendous amount of data to create our Spanish paraphrases
- While the scenario of using all other Europarl language pairs may seem idealized, it is precisely the situation that arises when adding new official languages to the EU

Future Work & Broader Vision

- Future Work:
 - Improve statistical models of translation through increased **generalization**
 - Generate **more sophisticated paraphrases** including syntactic transformations
- Broader Vision:
 - Unlike data for other statistical NLP tasks, such as parsing, parallel corpora are created naturally by other human industry
 - Harnessing the data through savvy manipulation will improve translation, and can potentially **improve all other language technologies**