Improving Statistical Machine Translation
With Paraphrases and Generalization

Chris Callison-Burch

December 5, 2006
The Game Plan

- The problem of coverage in SMT
- Paraphrases as a potential solution, and how we generate them
- Experimental design and evaluation methodologies
- Experimental results
- Generalization beyond paraphrases
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

![Graph showing the extent of the problem](chart.png)
Behavior on unseen words

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

   \textit{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 \textit{encargarnos} that this system is not susceptible to be \textit{usado} as political weapon.

• Since the translations of \textit{encargarnos} and \textit{usado} were not learned, they are either reproduced in the translation, or omitted entirely.
### Substituting paraphrases then translating

<table>
<thead>
<tr>
<th>encargarnos</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>usado</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 | ? |
| garantisar  |   |
| vela         |   |
| 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

<table>
<thead>
<tr>
<th>Word</th>
<th>Translation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>encargarnos</td>
<td>?</td>
</tr>
<tr>
<td>garantizar</td>
<td>guarantee, ensure, guaranteed, assure, provided</td>
</tr>
<tr>
<td>velar</td>
<td>ensure, ensuring, safeguard, making sure</td>
</tr>
<tr>
<td>procurar</td>
<td>ensure that, try to, ensure, endeavour to</td>
</tr>
<tr>
<td>asegurarnos</td>
<td>ensure, secure, make certain</td>
</tr>
<tr>
<td>usado</td>
<td>?</td>
</tr>
<tr>
<td>utilizado</td>
<td>used, use, spent, utilized</td>
</tr>
<tr>
<td>empleado</td>
<td>used, spent, employee</td>
</tr>
<tr>
<td>uso</td>
<td>use, used, usage</td>
</tr>
<tr>
<td>utiliza</td>
<td>used, uses, used, being used</td>
</tr>
</tbody>
</table>

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

• Much recent work has focused on extracting paraphrases from monolingual parallel corpora, or multiple translations

• Example:

<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>

• Extract burst into tears = cried and comfort = console
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
what is more, the relevant cost dynamic is completely under control.

im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle.

we owe it to the taxpayers to keep the costs in check.
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 tables entries were augmented with **additional feature function** that incorporated the paraphrase probability
Paraphrase probability

- Defined in terms of translation model probability
- Naturally falls out of parallel corpus

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

\[
= \arg \max_{e_2 \neq e_1} \sum_f p(f | e_1) p(e_2 | f)
\]

\[
\approx \arg \max_{e_2 \neq e_1} \sum_f \frac{\text{count}(f, e_1)}{\text{total}(e_1)} \times \frac{\text{count}(e_2, f)}{\text{total}(f)}
\]
Additional feature function

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

\[
h(e, f_1) = \begin{cases} 
p(f_2 | f_1) & \text{if phrase table entry } (e, f_1) \\
1 & \text{is generated from } (e, f_2) \\
1 & \text{Otherwise}
\end{cases}
\]
## Expanded phrase table

<table>
<thead>
<tr>
<th>paraphrases</th>
<th>existing translations</th>
<th>new translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>encargarnos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>garantizar</td>
<td>0.07</td>
<td>encargarnos</td>
</tr>
<tr>
<td>velar</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>procurar</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>velar</td>
<td>0.21  1.0</td>
<td>ensure</td>
</tr>
<tr>
<td>make sure</td>
<td>0.11  1.0</td>
<td>make sure</td>
</tr>
<tr>
<td>to ensure</td>
<td>0.09  1.0</td>
<td>to ensure</td>
</tr>
<tr>
<td>protect</td>
<td>0.03  1.0</td>
<td>protect</td>
</tr>
<tr>
<td>ensuring</td>
<td>0.03  1.0</td>
<td>ensuring</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental 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 by paraphrasing *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
Results: Bleu Scores
Results: Bleu Scores
Results: Bleu Scores

![Graph showing Bleu Scores vs. Training Corpus Size (num sentences)](image)

- **Baseline**
- **Single word paraphrases**
- **Multi-word paraphrases**

<table>
<thead>
<tr>
<th>Bleu Score</th>
<th>Training Corpus Size (num sentences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>10000</td>
</tr>
<tr>
<td>23</td>
<td>20000</td>
</tr>
<tr>
<td>24</td>
<td>30000</td>
</tr>
<tr>
<td>25</td>
<td>40000</td>
</tr>
<tr>
<td>26</td>
<td>50000</td>
</tr>
<tr>
<td>27</td>
<td>60000</td>
</tr>
<tr>
<td>28</td>
<td>70000</td>
</tr>
<tr>
<td>29</td>
<td>80000</td>
</tr>
</tbody>
</table>
Results: Improvements in coverage

![Bar chart showing improvements in coverage for unigrams, bigrams, trigrams, and 4-grams before and after paraphrasing.]

- Before paraphrasing:
  - Unigrams: 48
  - Bigrams: 25
  - Trigrams: 10
  - 4-grams: 3

- After paraphrasing:

---

Chris Callison-Burch

Improving SMT With Paraphrases and Generalization

December 5, 2006
Results: Improvements in coverage

- **unigrams**: 48% before paraphrasing, 90% after paraphrasing
- **bigrams**: 25% before paraphrasing, 67% after paraphrasing
- **trigrams**: 10% before paraphrasing, 37% after paraphrasing
- **4-grams**: 3% before paraphrasing, 16% after paraphrasing
Focused manual evaluation

REFERENCE TRANSLATION:
The article combats discrimination of citizens for the reasons listed therein.

SOURCE SENTENCE:
El artículo combate la discriminación de los ciudadanos por las causas enumeradas en el mismo.
Focused manual evaluation

REFERENCE TRANSLATION:
The article combats discrimination of citizens for the reasons listed therein.

SOURCE SENTENCE:
El artículo combate la discriminación de los ciudadanos por las causas enumeradas en el mismo.

HYPOTHESIS TRANSLATION 1:
The article fights against of citizens for the reasons mentioned in the same.
Focused manual evaluation

REFERENCE TRANSLATION:
The article combats discrimination of citizens for the reasons listed therein.

SOURCE SENTENCE:
El artículo combate la discriminación de los ciudadanos por las causas enumeradas en el mismo.

HYPOTHESIS TRANSLATION 1:
The article fights against of citizens for the reasons mentioned in the same.
Focused manual evaluation

The reasons listed therein.

The same reasons mentioned in the same.
Results: Focused Manual Evaluation

<table>
<thead>
<tr>
<th>Training Corpus Sizes (num sentences)</th>
<th>Accuracy of Newly Translated Phrases (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10k</td>
<td>48</td>
</tr>
<tr>
<td>20k</td>
<td>53</td>
</tr>
<tr>
<td>40k</td>
<td>57</td>
</tr>
<tr>
<td>80k</td>
<td>71</td>
</tr>
</tbody>
</table>

- **Single word paraphrases**
- **Multi-word paraphrases**
Results: Focused Manual Evaluation

Bleu was insensitive to improvements more than half of the time.

Accuracy of Newly Translated Phrases (%) for different training corpus sizes:

- Single word paraphrases
- Multi-word paraphrases

Accuracy values:
- 10k: 48%
- 20k: 53%
- 40k: 57%
- 80k: 67%

The graph shows the accuracy of newly translated phrases for different sizes of the training corpus, with single word and multi-word paraphrases. Bleu was found to be insensitive to improvements more than half of the time.
Summary of Results

- 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; it is precisely the situation that arises when adding new official languages to the EU.
Reflection → New Direction

- Paraphrasing improves translation because it introduces generalization.
- However, uses external data to do so, and generalization is limited.
- Currently fail to take advantage of the wealth of information in our data.
- We need models capable of generalization.
- Augment models and training data with linguistic information.
Reflection → New Direction

- Paraphrasing improves translation because it introduces generalization
- However, uses external data to do so, and generalization is limited
- Currently fail to take advantage of the wealth of information in our data
- We need models capable of generalization
- Augment models and training data with linguistic information

So let’s do that...
<table>
<thead>
<tr>
<th>Current Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain declined to confirm that Spain declined to aid Morocco.</td>
</tr>
</tbody>
</table>

| L' Espagne a refusé de confirmer que l' Espagne avait refusé d' aider le Maroc. |

| We see that the French government has sent a mediator. |

| Nous voyons que le gouvernement français a envoyé un médiateur. |
Phrase Extraction
Phrase Extraction

Spain declined to confirm that Spain declined to aid Morocco.

L'Espagne a refusé de confirmer que l'Espagne avait refusé d'aider le Maroc.

Spain declined L' Espagne a refusé de

Spain declined l' Espagne avait refusé d'

Chris Callison-Burch
Improving SMT With Paraphrases and Generalization
December 5, 2006
### Phrase Extraction

<table>
<thead>
<tr>
<th></th>
<th>Spain</th>
<th>L' Espagne</th>
</tr>
</thead>
<tbody>
<tr>
<td>declined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>confirm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>declined</td>
<td>L' Espagne a refusé de</td>
</tr>
<tr>
<td>declined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to confirm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>declined to confirm</td>
<td>L' Espagne a refusé de confirmer</td>
</tr>
<tr>
<td>declined</td>
<td></td>
<td>a refusé de</td>
</tr>
<tr>
<td>to confirm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>declined to confirm that</td>
<td>a refusé de confirmer que</td>
</tr>
<tr>
<td>declined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to confirm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>declined to confirm</td>
<td>a refusé de confirmer que</td>
</tr>
<tr>
<td>declined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to confirm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>declined</td>
<td>L' Espagne</td>
</tr>
<tr>
<td>declined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>to aid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morocco</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Spain declined to confirm that Spain declined that Spain declined to aid Morocco.
Phrase Translation Probabilities

- Translation probability can be estimated as

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

\[ p(\text{"l' Espagne a refusé de} | \text{Spain declined}) = 0.5 \]
\[ p(\text{"l' Espagne avait refusé d'} | \text{Spain declined}) = 0.5 \]

\[ \text{count}(\text{"l' Espagne a refusé de, Spain declined}) = 1 \]
\[ \text{count}(\text{"l' Espagne avait refusé d', Spain declined}) = 1 \]
\[ \text{count}(\text{Spain declined}) = 2 \]
Problems with Current Models

• Words and phrases that are unseen, or that have sparse counts
  – *Cannot* learn translations of *unseen words*, even if *other forms* occur
  – *Problematic* for *morphologically rich* languages

• Lack of linguistic descriptions
  – *Cannot* learn that a language is *subject-object-verb*
  – *Cannot* penalize hypothesis translations that lack a *verb*
  – *Cannot* specify simple things like *adjective-noun reordering*
  – *Unable* to do *inflectional* or *derivational morphology*

• And others...
Factored Translation Models

- Represent phrases with multiple levels of information

- These **linguistic factors** can include
  - fully-inflected words
  - lemmas
  - part of speech tags
  - morphological info (case, number, gender, tense)
  - semantic classes
  - **and more ambitious things!**

- Estimate probabilities for translation between factors other than words
Factorized Parallel Corpus

words: Spain declined to confirm that Spain declined to aid Morocco .

POS: NNP VBD TO VB IN NNP VBN TO VB NNP .

words: L' Espagne a refusé de confirmer que l' Espagne avait refusé d' aider le Maroc .

POS: DET NN AUX VPP PREP VINF PREP DET NN AUX VPP PREP VINF DET NN .

lemmas: la Espagne avoir refuser de confirmer que la Espagne avoir refuser de aider le Maroc .

words: We see that the French government has sent a mediator .

POS: PRP VBP IN DT JJ NN VBP VBN DT NN .

words: Nous voyons que le gouvernement français a envoyé un médiateur .

POS: PRON VBP PREP DET NN ADJ AUX VBG DET NN .

lemmas: Nous voyons que le gouvernement français a envoyé un médiateur .

Chris Callison-Burch  Improving SMT With Paraphrases and Generalization  December 5, 2006
Alignment Between Factors
Alignment Between Factors
Alignment Between Factors
Factorized Phrase Extraction
### Factorized Phrase Extraction

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Part of Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>L' Espagne</td>
<td>Spain, NNP</td>
</tr>
<tr>
<td>L' Espagne a refusé de</td>
<td>Spain declined, NNP VBD</td>
</tr>
<tr>
<td>a refusé de</td>
<td>declined, VBD</td>
</tr>
<tr>
<td>a refusé de confirmer</td>
<td>declined to confirm, VBD TO VB</td>
</tr>
<tr>
<td>confirmer</td>
<td>to confirm, TO VB</td>
</tr>
<tr>
<td>confirmer que</td>
<td>to confirm that, TO VB IN</td>
</tr>
<tr>
<td>que</td>
<td>that, IN</td>
</tr>
<tr>
<td>que l' Espagne</td>
<td>that Spain, IN NNP</td>
</tr>
<tr>
<td>que l' Espagne avait refusé d'</td>
<td>that Spain declined, IN NNP VBN</td>
</tr>
<tr>
<td>l' Espagne</td>
<td>Spain, NNP</td>
</tr>
<tr>
<td>l' Espagne avait refusé d'</td>
<td>Spain declined, NNP VBN</td>
</tr>
<tr>
<td>aider</td>
<td>declined to aid, NNP VBN TO VB</td>
</tr>
<tr>
<td>aider d'</td>
<td>declined to, VBN TO</td>
</tr>
<tr>
<td>aider d' aider</td>
<td>declined to aid, VBN TO VB</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

---

Chris Callison-Burch  
Improving SMT With Paraphrases and Generalization  
December 5, 2006
Translation Between Factors

- We can now calculate translation probability between sequences of factors
- For instance \( p(\bar{f}_{\text{pos}} | \bar{e}_{\text{pos}}) \) or \( p(\bar{f}_{\text{words}} | \bar{e}_{\text{words}}, \bar{e}_{\text{pos}}) \)
- Can calculate these using MLE again:

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

\[
p(\bar{f}_{\text{words}} | \bar{e}_{\text{words}}, \bar{e}_{\text{pos}}) = \frac{\text{count}(\bar{f}_{\text{words}}, \bar{e}_{\text{words}}, \bar{e}_{\text{pos}})}{\text{count}(\bar{e}_{\text{words}}, \bar{e}_{\text{pos}})}
\]
Linguistic Context

- Translation probabilities between factors let us distinguish between things we were previously unable to:

\[ p(l' \text{ Espagne a refusé de} | \text{Spain declined}) = 0.5 \]
\[ p(l' \text{ Espagne avait refusé d'} | \text{Spain declined}) = 0.5 \]

\[ p(l' \text{ Espagne a refusé de} | \text{Spain declined, NNP VBN}) = 0 \]
\[ p(l' \text{ Espagne } \textbf{avait} \text{ refusé d'} | \text{Spain declined, NNP VBN}) = 1 \]
\[ p(l' \text{ Espagne a refusé de} | \text{Spain declined, NNP VBD}) = 1 \]
\[ p(l' \text{ Espagne } \textbf{avait} \text{ refusé d'} | \text{Spain declined, NNP VBD}) = 0 \]
Better Modeling

- They also let us model phenomena we were previously unable to:

\[ p(\text{NN ADJ|JJ NN}) > p(\text{ADJ NN|JJ NN}) \]

- Can also calculate “language model” probabilities over longer sequences of part of speech tags:

\[ p(\text{DET NN AUX VPP PREP VINF}) \]

- Factors with small tag sets have richer counts
Richer Counts

- With morphologically rich languages, $p(\bar{f}_{words} | \bar{e}_{words})$ has sparse counts.
- Break the translation down into steps:
  - Translate from English lemmas into foreign lemmas
  - Translate English morph into foreign morph
  - Use foreign morph markers + lemmas to **generate** foreign words
- Better counts in parallel corpus for $p(\bar{f}_{lemmas} | \bar{e}_{lemmas})$ and $p(\bar{f}_{morph} | \bar{e}_{morph})$
- Use monolingual corpus to estimate $p(\bar{f}_{words} | \bar{f}_{lemmas}, \bar{f}_{morph})$
The Way Forward

• Build **multi-level models** that make **better use of data**
• Incorporate **more sophisticated linguistic factors**
• Many challenges
  – Which linguistic factors are **relevant** for translation
  – Do **proper estimation** of many parallel, overlapping sequences
  – Design **efficient data structures** and **search** strategies
• **Multi-level models** are **not limited** to statistical machine translation
  – **Natural language generation** including paraphrasing
  – **Evaluation** of MT and summarization
  – And others!
Selected Publications


Thank you!
Example English 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
Structural paraphrases

In addition to lexical paraphrases we can learn structural paraphrases:

- **JJ PRP MD VB TO VB** → **MD PRP VB IN VBG** (.09)
- **JJ PRP MD VB TO VB** → **PRP VBP TO VB IN VBG** (.07)
- **JJ PRP MD VB TO VB** → **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.
Usefulness of FF

<table>
<thead>
<tr>
<th>Corpus size</th>
<th>10k</th>
<th>20k</th>
<th>40k</th>
<th>80k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>22.6</td>
<td>25.0</td>
<td>26.5</td>
<td>26.5</td>
</tr>
<tr>
<td>Single word with FF</td>
<td>23.1</td>
<td>25.2</td>
<td>26.6</td>
<td>28.0</td>
</tr>
<tr>
<td>Multi-word with FF</td>
<td>23.3</td>
<td>26.0</td>
<td>27.2</td>
<td>28.0</td>
</tr>
<tr>
<td>Single word without FF</td>
<td>23.0</td>
<td>25.1</td>
<td>26.7</td>
<td>28.0</td>
</tr>
<tr>
<td>Multi-word without FF</td>
<td>20.6</td>
<td>22.6</td>
<td>21.9</td>
<td>24.0</td>
</tr>
</tbody>
</table>