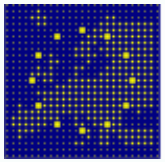
Effiecient SMT by Triangulation

LREC 2008: Marrakech, Morocco

Yu Chen
Andreas Eisele
Martin Kay



EuroMatrix
Statistical and Hybrid Machine Translation
Between All European Languages

Department of Computational Linguistics
Saarland University

May 29, 2008

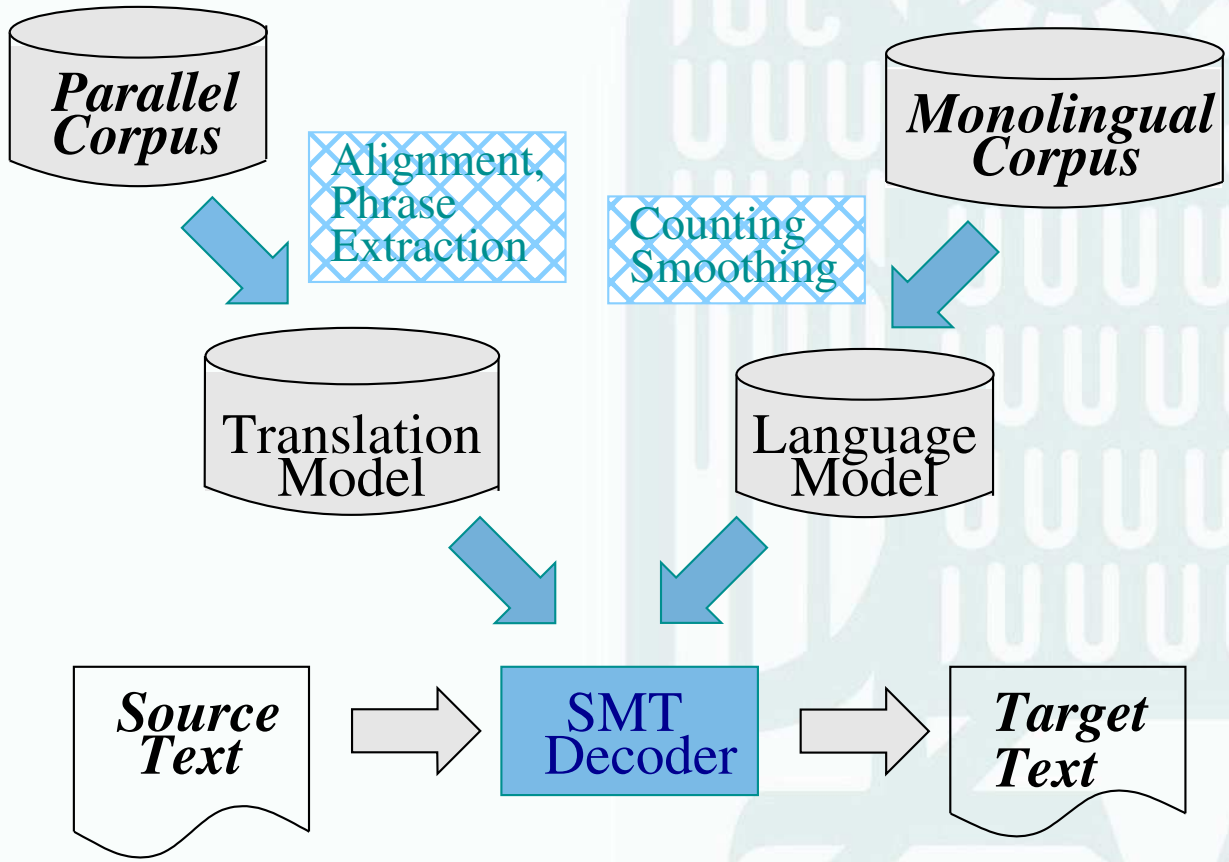
Outline

Outline

SMT architecture

To build a phrase-based SMT system:

- Parallel corpus
- Moses Toolkit, etc.



Problems with Translation Models


```
nicht . ||| ' . ||| 0.00035137 0.00703986 0.000659631 0.0023873 2.718
nicht . ||| 's fault . ||| 0.5 0.0095052 0.000659631 2.87847e-08 2.718
nicht . ||| 't . ||| 0.111111 0.418755 0.000659631 0.000876442 2.718
nicht . ||| 't do ! ||| 1 0.0249022 0.000659631 2.52495e-08 2.718
nicht . ||| 't stick . ||| 1 0.418755 0.000659631 2.87473e-08 2.718
nicht . ||| , as did ||| 0.0102041 6.20073e-05 0.000659631 3.20962e-08 2.718
nicht . ||| , as ||| 3.29272e-05 6.20073e-05 0.000659631 7.5149e-05 2.718
nicht . ||| , no . ||| 0.0714286 0.168673 0.000659631 0.00317554 2.718
nicht . ||| , they do not . ||| 1 0.288859 0.000659631 4.94651e-07 2.718
nicht . ||| , would not . ||| 1 0.70589 0.000659631 0.000160212 2.718
nicht . ||| , ||| 4.89461e-06 6.20073e-05 0.00329815 0.0094167 2.718
```

Problems with Translation Models


```
nicht . ||| ' . ||| 0.00035137 0.00703986 0.000659631 0.0023873 2.718
nicht . ||| 's fault . ||| 0.5 0.0095052 0.000659631 2.87847e-08 2.718
nicht . ||| 't . ||| 0.111111 0.418755 0.000659631 0.000876442 2.718
nicht . ||| 't do ! ||| 1 0.0249022 0.000659631 2.52495e-08 2.718
nicht . ||| 't stick . ||| 1 0.418755 0.000659631 2.87473e-08 2.718
nicht . ||| , as did ||| 0.0102041 6.20073e-05 0.000659631 3.20962e-08 2.718
nicht . ||| , as ||| 3.29272e-05 6.20073e-05 0.000659631 7.5149e-05 2.718
nicht . ||| , no . ||| 0.0714286 0.168673 0.000659631 0.00317554 2.718
nicht . ||| , they do not . ||| 1 0.288859 0.000659631 4.94651e-07 2.718
nicht . ||| , would not . ||| 1 0.70589 0.000659631 0.000160212 2.718
nicht . ||| , ||| 4.89461e-06 6.20073e-05 0.00329815 0.0094167 2.718
```

Mit “The Hunting Party” und “Mörderischer Frieden” beschäftigen sich wieder zwei Filme mit dem Balkankrieg. Doch beide überzeugen nicht.

The Hunting Party and Mörderischer Frieden make two more films dealing with the Balkan War. But neither of them is convincing.

Problems with Translation Models


```
nicht . ||| ' . ||| 0.00035137 0.00703986 0.000659631 0.0023873 2.718
nicht . ||| 's fault . ||| 0.5 0.0095052 0.000659631 2.87847e-08 2.718
nicht . ||| 't . ||| 0.111111 0.418755 0.000659631 0.000876442 2.718
nicht . ||| 't do ! ||| 1 0.0249022 0.000659631 2.52495e-08 2.718
nicht . ||| 't stick . ||| 1 0.418755 0.000659631 2.87473e-08 2.718
nicht . ||| , as did ||| 0.0102041 6.20073e-05 0.000659631 3.20962e-08 2.718
nicht . ||| , as ||| 3.29272e-05 6.20073e-05 0.000659631 7.5149e-05 2.718
nicht . ||| , no . ||| 0.0714286 0.168673 0.000659631 0.00317554 2.718
nicht . ||| , they do not . ||| 1 0.288859 0.000659631 4.94651e-07 2.718
nicht . ||| , would not . ||| 1 0.70589 0.000659631 0.000160212 2.718
nicht . ||| , ||| 4.89461e-06 6.20073e-05 0.00329815 0.0094167 2.718
```

Mit “The Hunting Party” und “Mörderischer Frieden” beschäftigen sich wieder zwei Filme mit dem Balkankrieg. Doch beide überzeugen nicht.

The Hunting Party and Mörderischer Frieden make two more films dealing with the Balkan War. But neither of them is convincing.

“The Hunting Party” and “Mörderischer peace” deal another two films with the Balkans war. But both cases.

Outline

What's **triangulation**?

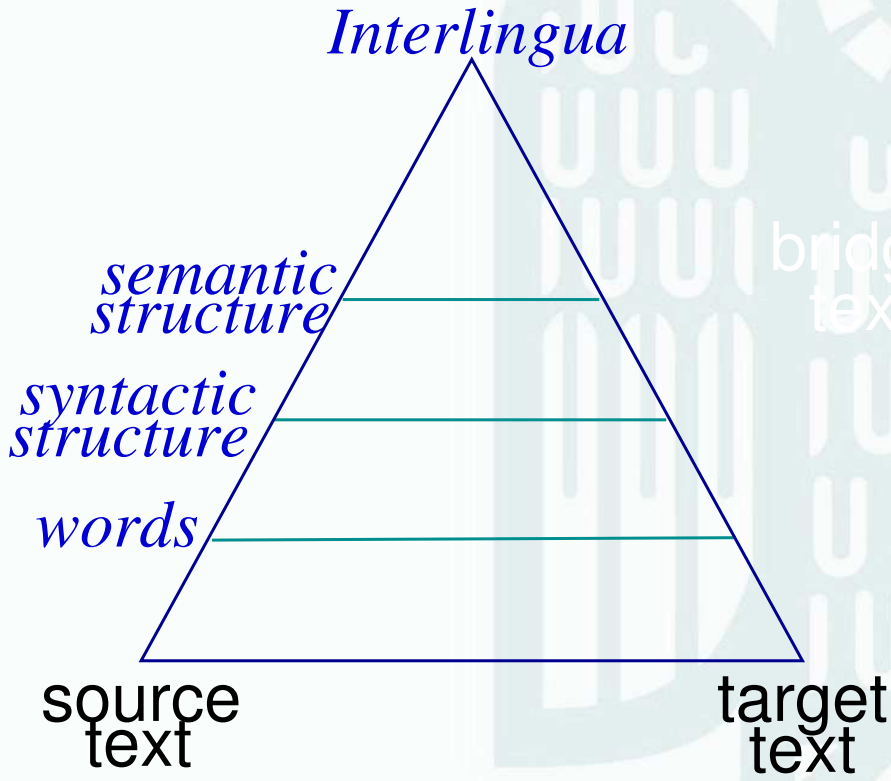
- In the social sciences
  the use of multiple cross-checked sources and methodology
- In qualitative research
  combining methods for more accurate and credible research

What's **triangulation**?

- In the social sciences
  the use of multiple cross-checked sources and methodology
- In qualitative research
  combining methods for more accurate and credible research
- In the context of machine translation
  making use of resources in languages other than the two involving in translation

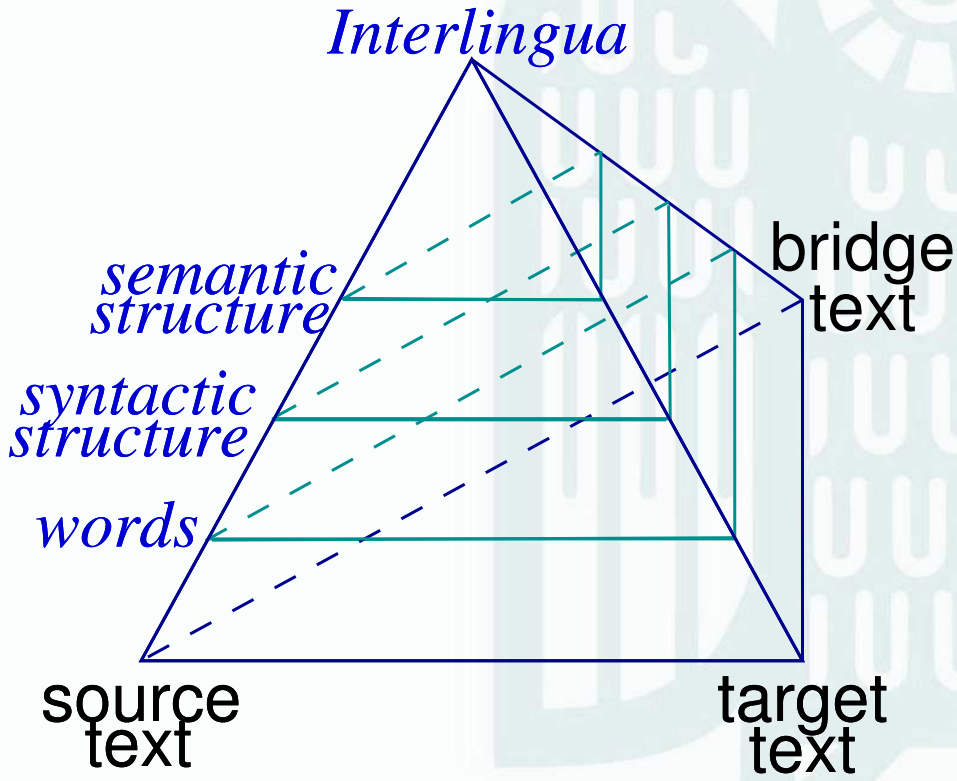
Is *triangulation* possible in MT?

Vauquois...



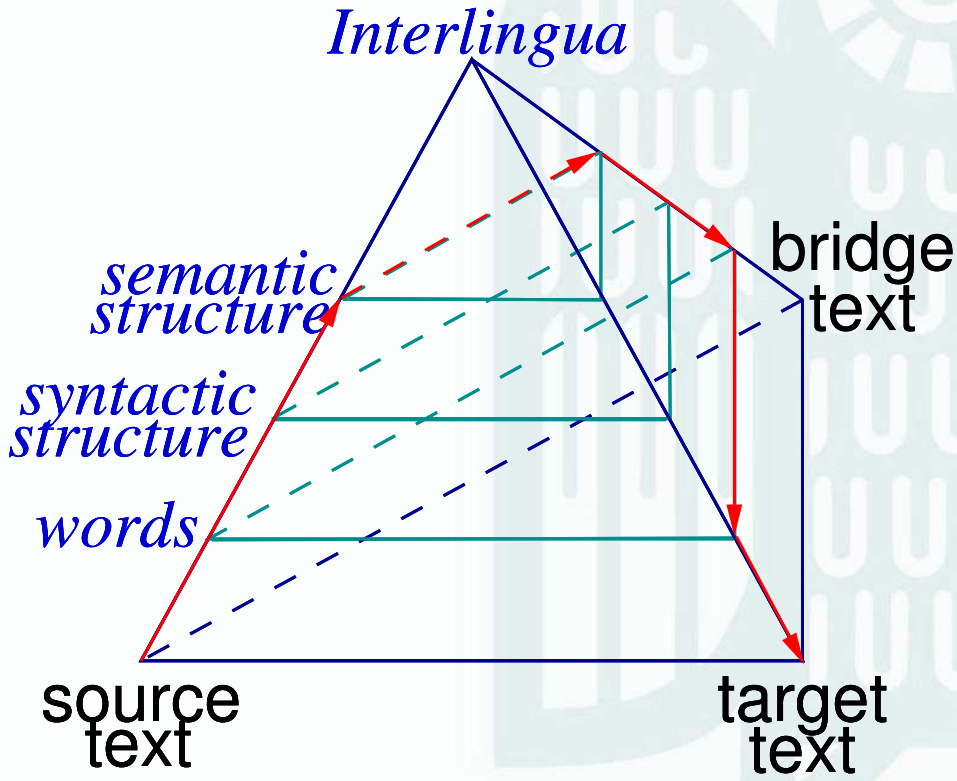
Is *triangulation* possible in MT?

Vauquois...



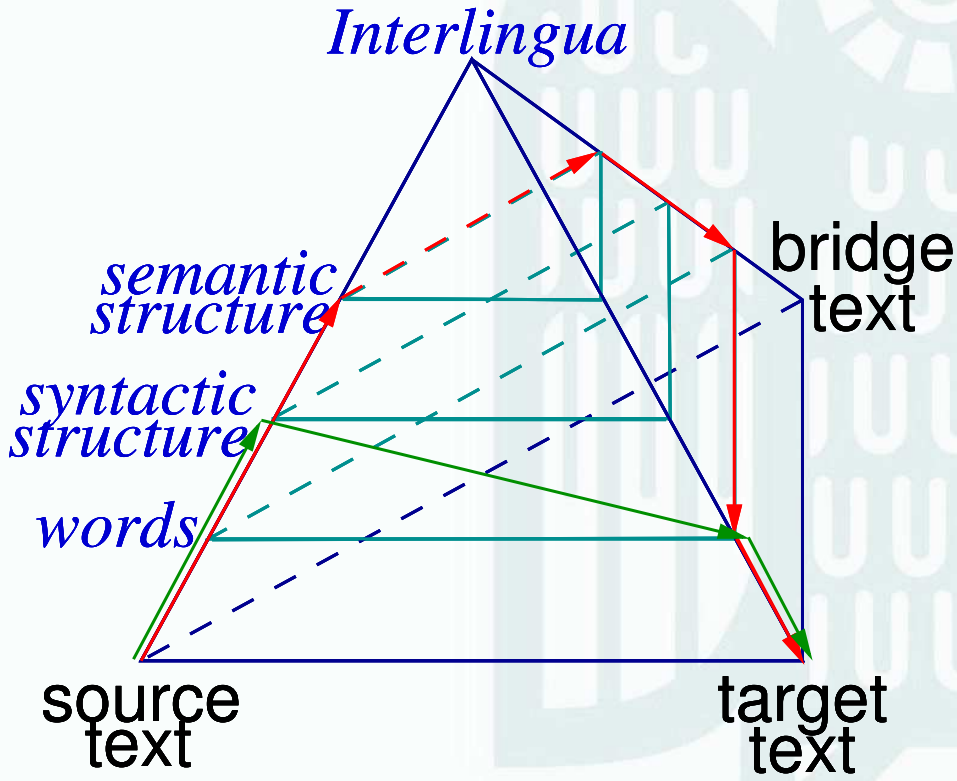
Is *triangulation* possible in MT?

Vauquois...



Is *triangulation* possible in MT?

Vauquois...



Triangulation in Machine Translation

- Need for Triangulation
  - lack of resources for the direct language pairs
  - rich resources for frequent languages
  - difficult language pairs
- Advantages
  - solve ambiguity
  - help with word orders
  - increase lexical coverage

Triangulation in Machine Translation

- Need for Triangulation
  - lack of resources for the direct language pairs
  - rich resources for frequent languages
  - difficult language pairs
- Advantages
  - solve ambiguity
  - help with word orders
  - increase lexical coverage
- Existing approaches
  - Sentence alignment [Simard, 1999]
  - Word alignment [Kumar et al., 2007]
  - Translation model [Cohn and Lapata, 2007]
  - Hypothesis reranking [Och and Ney, 2001]

Outline

General description

- Motivation
  - Phrase tables contain a lot of noise
  - Size of a phrase table is critical for decoder
  - Data in a third language convey extra information
- Use the extra information to filter a phrase table
  - phrases in the 3rd language as linking evidence
  - only keep the most probable phrase pairs

Procedure

- Additional translation models
  1. From source language to bridge language
  2. From target language to bridge language

- Examine the phrase table by entry for each phrase pair:
  - search both additional tables for a common link in the third language corresponding to both source phrase and target phrase in the pair

  Method 1 exact phrase matching
  Method 2 word overlap

  - remove the entry when such a link does not exist
  - keep the probabilities
  - use the reduced table as the original

Method 1: exact match
Example


```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

Method 1: exact match

Example

```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

Method 1: exact match

Example

```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

```
.
.
.
auch automobilhersteller ,
auch automobilhersteller
autohersteller
autoherstellern
automobilhersteller
automobilherstellern
bedeutende hersteller
computerhersteller
damit ihren schnitt
damit ihren
damit
.
.
.
```

Method 1: exact match

Example

```
.
.
.
auch automobilhersteller ,
auch automobilhersteller
autohersteller
autoherstellern
automobilhersteller
automobilherstellern
bedeutende hersteller
computerhersteller
damit ihren schnitt
damit ihren
damit
.
.
.
```

```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

```
die fahrzeughersteller haben
```

Method 1: exact match

Example

```
.
.
.
auch automobilhersteller ,
auch automobilhersteller
autohersteller
autoherstellern
automobilhersteller
automobilherstellern
bedeutende hersteller
computerhersteller
damit ihren schnitt
damit ihren
damit
.
.
.
```

```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

```
die fahrzeughersteller haben
```

Method 1: exact match
Example


```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

Method 2: word overlap
Example


```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

Method 2: word overlap
Example


```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

Method 2: word overlap
Example


```
··· automobilhersteller, hersteller,
herstellern,arzneimittelhersteller,
arzneimittelproduzenten,auch,
automobilhersteller,autohersteller,
autoherstellern, vor,gehen, haben,
ist, verknüpfen, verknüpfen, zu,
bedeutende, den,computerhersteller,
und ···
```

```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers h
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

```
den,hersteller, durch,
ein, einem
```

Method 2: word overlap
Example

··· automobilhersteller, **hersteller**, herstellern, arzneimittelhersteller, arzneimittelproduzenten, auch, automobilhersteller, autohersteller, autoherstellern, vor, gehen, haben, ist, verknüpfen, verknüpfen, zu, bedeutende, **den**, computerhersteller, und ···

fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers h
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers

den, **hersteller**, durch, ein, einem

Method 2: word overlap

Example

$$\frac{|W_{S\to F}(s) \cap W_{F\to E}(e)|}{\min(W_{S\to F}(s), W_{F\to E}(e))}$$

··· automobilhersteller, **hersteller**, herstellern, arzneimittelhersteller, arzneimittelproduzenten, auch, automobilhersteller, autohersteller, autoherstellern, vor, gehen, haben, ist, verknüpfen, verknüpfen, zu, bedeutende, **den**, computerhersteller, und ···

```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers h
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

$\frac{2}{5} = 0.4$

den, **hersteller**, durch, ein, einem

Method 2: word overlap

Example

$$\frac{|W_{S\to F}(s) \cap W_{F\to E}(e)|}{\min(W_{S\to F}(s), W_{F\to E}(e))}$$

```
fabricantes ||| , manufacturers
fabricantes ||| a manufacturer
fabricantes ||| battalions
fabricantes ||| car manufacturers have
fabricantes ||| car manufacturers
fabricantes ||| makers
fabricantes ||| manufacturer
fabricantes ||| manufacturers
fabricantes ||| producers are
fabricantes ||| producers need
fabricantes ||| producers
fabricantes ||| suppliers
```

$\frac{2}{5} = 0.4$

Outline

Experiment Setup

- Language pair: Spanish-English
- Bridge languages: German, French
- Training data: Europarl subsets

Max. Sent. Len.	Sent. Num.
40	950,000
50	1,100,000

- Testset: Europarl testset from WMT 2008
- Baselines: built with `Moses`, MERTed for `BLEU`
- Filtering method: 1 & 2 + baseline weights
- Evaluation:
  - Sizes of the phrase tables
  - Translation quality

Results

Size of filtered phrase-tables

Model	Entries	PT(Byte)	RT(Byte)	Removed
Europarl-40	19M	2.5G	1.9G	
1:French	8M	1.1G	741M	55.21%
2:French	15M	1.9G	1.3G	23.52%
1:German	6M	725M	492M	69.16%
2:German	14M	1.8G	1.2G	29.16%
Europarl-50	54M	7.1G	5.4G	
1:French	24M	3.0G	2.3G	55.77%
2:French	42M	5.5G	4.2G	24.10%
1:German	16M	1.9G	1.5G	70.70%
2:German	38M	5.0G	3.8G	30.42%

Results

Size of filtered phrase-tables

Model	Entries	PT(Byte)	RT(Byte)	Removed
Europarl-40	19M	2.5G	1.9G	
1:French	8M	1.1G	741M	55.21%
2:French	15M	1.9G	1.3G	23.52%
1:German	6M	725M	492M	69.16%
2:German	14M	1.8G	1.2G	29.16%
Europarl-50	54M	7.1G	5.4G	
1:French	24M	3.0G	2.3G	55.77%
2:French	42M	5.5G	4.2G	24.10%
1:German	16M	1.9G	1.5G	70.70%
2:German	38M	5.0G	3.8G	30.42%

Results

Translation Quality (BLEU)

Method 1	None	French	German
Europarl-40	31.43	28.27	31.58
Europarl-50	31.65	31.73	31.92

Method 2	None	French	German
Europarl-40	31.43	28.20	31.38
Europarl-50	31.65	31.69	31.75

Results

Translation Quality (BLEU)

Method 1	None	French	German
Europarl-40	31.43	28.27	31.58
Europarl-50	31.65	31.73	31.92

Method 2	None	French	German
Europarl-40	31.43	28.20	31.38
Europarl-50	31.65	31.69	31.75

Example

src	Como ha señalado el Sr. de Soto, no esperamos que el progreso sea tarea fácil, y el éxito del proceso de las Naciones Unidas no está ni mucho menos garantizado.
ref	As Mr de Soto noted, we do not expect progress to be easy, and the success of the UN process is far from assured.
Baseline	As has been pointed out by Mr de Soto, we hope that progress is not an easy task, and the success of the UN process is far from guaranteed.
1:French	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
2:French	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
1:German	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
2:German	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.

Example

src	Como ha señalado el Sr. de Soto, no esperamos que el progreso sea tarea fácil, y el éxito del proceso de las Naciones Unidas no está ni mucho menos garantizado.
ref	As Mr de Soto noted, we do not expect progress to be easy, and the success of the UN process is far from assured.
Baseline	As has been pointed out by Mr de Soto, we hope that progress is not an easy task, and the success of the UN process is far from guaranteed.
1:French	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
2:French	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
1:German	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.
2:German	As Mr de Soto, we do not expect that progress is easy, and the success of the UN process is far from guaranteed.

Outline

Summary

- Summary
  - More possibility of **triangulation**
  - Filtering reduces the size of phrase tables
  - Filtering preserves the translation quality
  - The approaches work better for larger models
  - Different bridge languages have different effect
- Future Work
  - More thorough experiments
  - Integration with other triangulation approaches

Thank you! Any questions?

References I


Cohn, T. and Lapata, M. (2007).
Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora.
In *the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech.


Kumar, S., Och, F. J., and Macherey, W. (2007).
Improving word alignment with bridge languages.
In *the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech.


Och, F. J. and Ney, H. (2001).
Statistical multi-source translation.
In *MT Summit VIII*, Santiago de Compostela, Spain.


Simard, M. (1999).
Text-translation alignment: Three languages are better than two.
In *Proceedings of EMNLP/VLC-99*, College Park, MD.