

Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation

John Tinsley

National Centre for Language Technology
Dublin City University
Ireland

Collaborators: Mary Hearne and Andy Way

NCLT Seminar Series
23/02/2009

Introduction

Experimental Setup

Experiments

Conclusions and Future Work

Introduction

Phrase pairs in translation models of PB-SMT system are induced using statistical models and heuristics. There is no linguistic motivation.

- ▶ shift in the field towards more syntactically aware models
- ▶ parallel treebanks are a linguistically rich resource
- ▶ phrase pairs extracted from parallel treebanks can improve translation

Introduction

Phrase pairs in translation models of PB-SMT system are induced using statistical models and heuristics. There is no linguistic motivation.

- ▶ shift in the field towards more syntactically aware models
- ▶ parallel treebanks are a linguistically rich resource
- ▶ phrase pairs extracted from parallel treebanks can improve translation

Can parallel treebank phrase pairs help translation in large-scale tasks?

How else can we use the information encoded in parallel treebanks within in the PB-SMT framework?

Parallel Treebanks

What is a parallel treebank?

- ▶ Linguistically annotated sententially aligned parallel data
- ▶ Alignments also at sub-sentential level
- ▶ Alignments hold implication of translational equivalence between linked constituents

Parallel Treebanks

What is a parallel treebank?

- ▶ Linguistically annotated sententially aligned parallel data
- ▶ Alignments also at sub-sentential level
- ▶ Alignments hold implication of translational equivalence between linked constituents

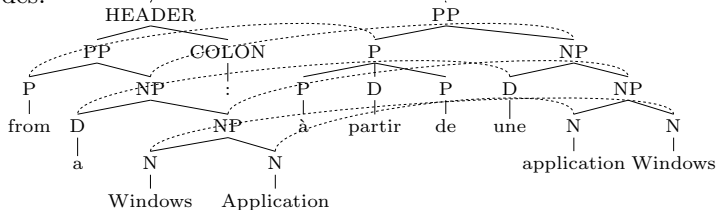
In our case we are dealing with context-free phrase structure parses. Sub-sentential alignments exist across both non-terminal and terminal nodes.

Parallel Treebanks

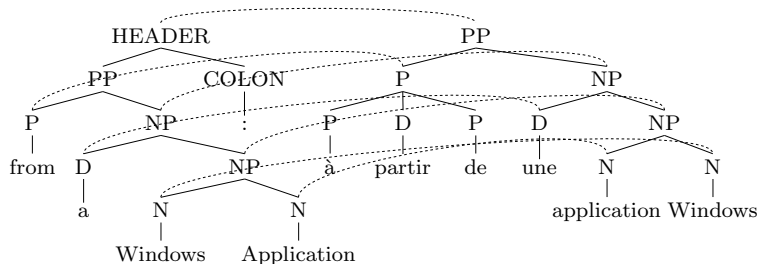
What is a parallel treebank?

- ▶ Linguistically annotated sentimentally aligned parallel data
- ▶ Alignments also at sub-sentential level
- ▶ Alignments hold implication of translational equivalence between linked constituents

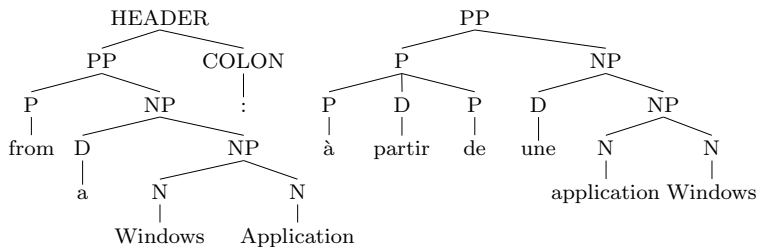
In our case we are dealing with context-free phrase structure parses. Sub-sentential alignments exist across both non-terminal and terminal nodes.



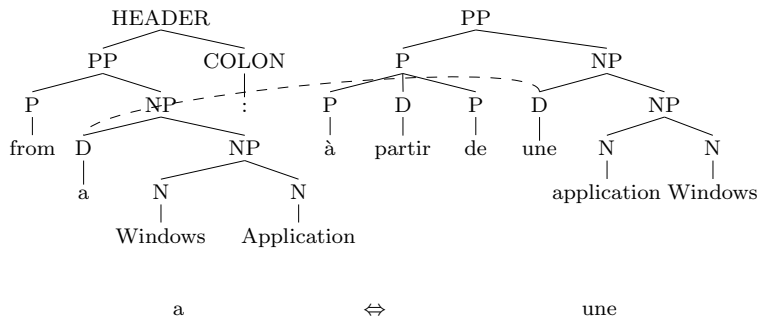
Trebank Phrase Extraction



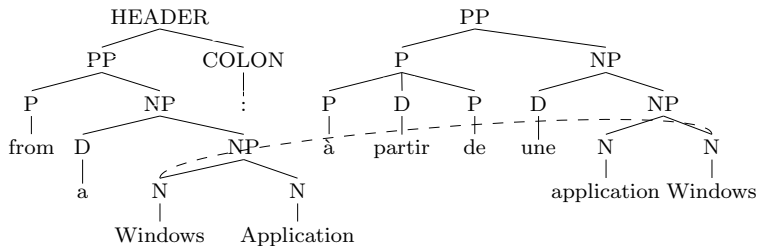
Trebank Phrase Extraction



Trebank Phrase Extraction

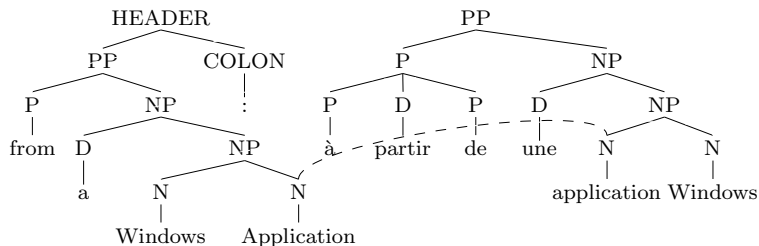


Treebank Phrase Extraction



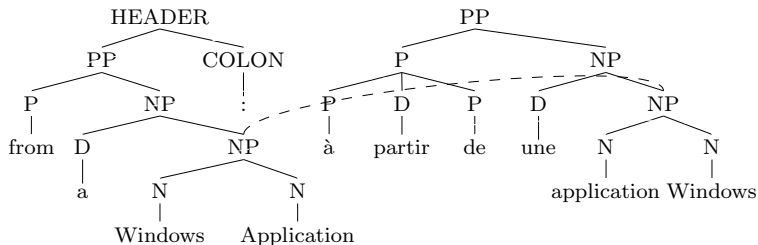
| | | |
|---------|---|-------------|
| a | ⇔ | une |
| from | ⇔ | à partir de |
| Windows | ⇔ | Windows |

Treebank Phrase Extraction



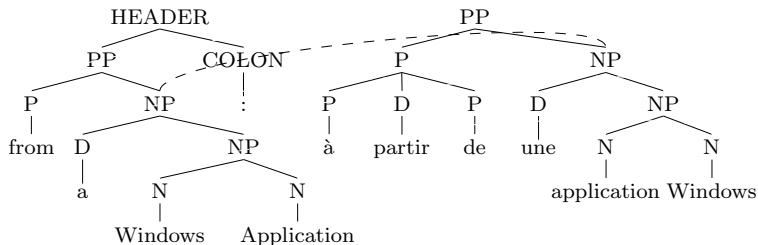
| | | |
|-------------|---|-------------|
| a | ⇔ | une |
| from | ⇔ | à partir de |
| Windows | ⇔ | Windows |
| Application | ⇔ | application |

Trebank Phrase Extraction



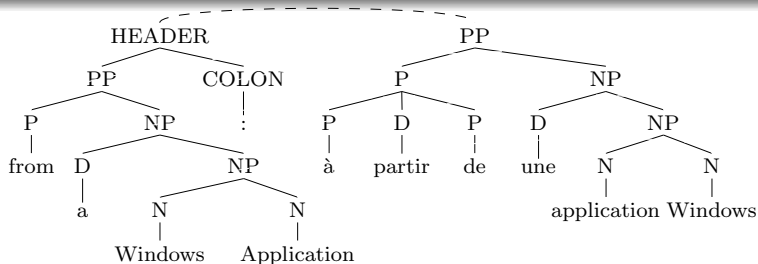
| | | |
|---------------------|---|---------------------|
| a | ⇔ | une |
| from | ⇔ | à partir de |
| Windows | ⇔ | Windows |
| Application | ⇔ | application |
| Windows Application | ⇔ | application Windows |

Trebank Phrase Extraction



| | | |
|-----------------------|---|-------------------------|
| a | ⇔ | une |
| from | ⇔ | à partir de |
| Windows | ⇔ | Windows |
| Application | ⇔ | application |
| Windows Application | ⇔ | application Windows |
| a Windows Application | ⇔ | une application Windows |

Trebank Phrase Extraction



| | | |
|------------------------------|---|-------------------------------------|
| a | ⇔ | une |
| from | ⇔ | à partir de |
| Windows | ⇔ | Windows |
| Application | ⇔ | application |
| Windows Application | ⇔ | application Windows |
| a Windows Application | ⇔ | une application Windows |
| from a Windows Application : | ⇔ | à partir de une application Windows |

Introduction

Experimental Setup

Experiments

Conclusions and Future Work

Data

- ▶ 729,891 sentence pairs from English–Spanish Europarl (v2)
- ▶ extract 1,000 sentence devset and 2,000 sentence testset
- ▶ parse both sides monolingually
- ▶ align using in-house subtree alignment tool

MT System

- ▶ Baseline PB-SMT system built with Moses
- ▶ 5-gram language model
- ▶ Minimum error-rate training on devset
- ▶ Automatic evaluation using BLEU, NIST AND METEOR

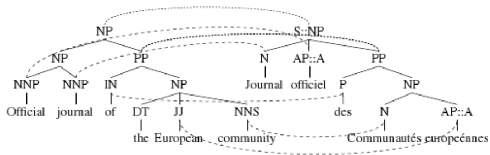
MT System

- ▶ Baseline PB-SMT system built with Moses
- ▶ 5-gram language model
- ▶ Minimum error-rate training on devset
- ▶ Automatic evaluation using BLEU, NIST AND METEOR

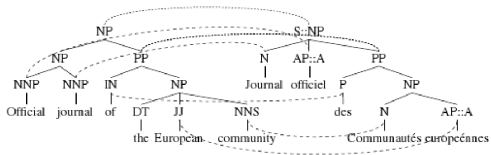
Phrase tables

- ▶ Baseline phrase pairs extracted from word alignments using Moses
- ▶ Phrase pairs extracted from parallel treebank based on node alignments
- ▶ Various combinations are used to build different translation models

Aligned Parallel Treebank



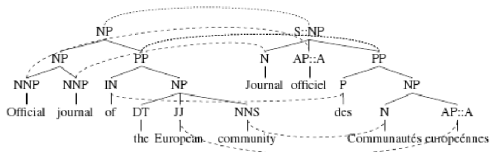
Aligned Parallel Treebank



SMT phrase extraction

| | Journal | officiel | des | Communautés | européennes |
|-------------|---------|----------|-----|-------------|-------------|
| Official | | ■ | | | |
| journal | ■ | | | | |
| of | | | ■ | | |
| the | | | | ■ | |
| European | | | | | ■ |
| Communities | | | | ■ | |

Aligned Parallel Treebank



SMT phrase extraction

| | Journal | officiel | des | Communautés | européennes |
|-------------|---------|----------|-----|-------------|-------------|
| Official | | ■ | | | |
| journal | ■ | | | | |
| of | | | ■ | | |
| the | | | | ■ | |
| European | | | | | ■ |
| Communities | | | | ■ | |

- † Official journal ↔ Journal officiel
- † Official journal of ↔ Journal officiel des
- * Official journal of the European Communities ↔ Journal officiel des Communautés européennes
- * of ↔ des
- * of the European Communities ↔ des Communautés européennes
- * the European Communities ↔ Communautés européennes
- * European ↔ européennes
- ◇ Communities ↔ Communautés
- ◇ Official ↔ officiel
- ◇ journal ↔ Journal

- Phrases extracted from the SMT system only
- ◇ Phrases extracted from the parallel treebank only
- * Phrases extracted from both the SMT system and the parallel treebank

Introduction

Experimental Setup

Experiments

Conclusions and Future Work

Experiment I - Direct Combination

We build three translation models

- ▶ SMT phrase pairs only (Baseline)
- ▶ Parallel treebank phrase pairs only (Tree only)
- ▶ Union of the above two models (Baseline+Tree)

Experiment I - Direct Combination

We build three translation models

- ▶ SMT phrase pairs only (Baseline)
- ▶ Parallel treebank phrase pairs only (Tree only)
- ▶ Union of the above two models (Baseline+Tree)

| Config. | Bleu | NIST | %METEOR |
|-----------|---------------|---------------|--------------|
| Baseline | 0.3341 | 7.0765 | 57.39 |
| +Tree | 0.3397 | 7.0891 | 57.82 |
| Tree only | 0.3153 | 6.8187 | 55.98 |

Experiment I - Direct Combination

| Resource | Baseline | Trebank |
|-----------------|-----------------|----------------|
| Unique Types | 23,261,022 | 4,985,266 |
| Overlap | 1,447,505 | |
| Ave Src Length | 4.28 | 8.56 |
| Ave Tgt Length | 4.39 | 9.02 |
| 1-to-1 | 1.54% | 15.91% |
| 1-to-n | 3.51% | 4.43% |

Experiment I - Direct Combination

We noticed issues with some treebank word alignments

- ▶ Constitute 20.3% of total extracted pairs
- ▶ 7.35% were high-frequency alignments between function words and punctuation
- ▶ Filtered these from model and rerun translation with this model (Strict phrases)

Experiment I - Direct Combination

We noticed issues with some treebank word alignments

- ▶ Constitute 20.3% of total extracted pairs
- ▶ 7.35% were high-frequency alignments between function words and punctuation
- ▶ Filtered these from model and rerun translation with this model (Strict phrases)

| Config. | BLEU | NIST | %METEOR |
|----------------|---------------|---------------|--------------|
| Baseline | 0.3341 | 7.0765 | 57.39 |
| +Tree | 0.3397 | 7.0891 | 57.82 |
| Strict phrases | 0.3414 | 7.1283 | 57.98 |

Experiment II - Weighting Treebank Data

We build three new translation models in which we directly combine the two sets of phrases but we count the treebank phrase pairs 2, 3 and 5 times respectively

Experiment II - Weighting Treebank Data

We build three new translation models in which we directly combine the two sets of phrases but we count the treebank phrase pairs 2, 3 and 5 times respectively

| Config. | BLEU | NIST | %METEOR |
|---------------|---------------|---------------|--------------|
| Baseline+Tree | 0.3397 | 7.0891 | 57.82 |
| +Tree x2 | 0.3386 | 7.0813 | 57.76 |
| +Tree x3 | 0.3361 | 7.0584 | 57.56 |
| +Tree x5 | 0.3377 | 7.0829 | 57.71 |

Experiment II - Weighting Treebank Data

We use a feature of the MT system which allows us to supply the two phrase tables separately. In this case the decoder will select phrases from either table for translation as is deemed appropriate by the model.

Experiment II - Weighting Treebank Data

We use a feature of the MT system which allows us to supply the two phrase tables separately. In this case the decoder will select phrases from either table for translation as is deemed appropriate by the model.

| Config. | BLEU | NIST | %METEOR |
|---------------|---------------|---------------|--------------|
| Baseline+Tree | 0.3397 | 7.0891 | 57.82 |
| Two Tables | 0.3365 | 7.0812 | 57.50 |

Exploiting Word Alignments

Given a parallel treebank, we also have a set of word alignments between the sentence pairs i.e. alignments between pre-terminal nodes. Word alignments are vital to core tasks in SMT.

Exploiting Word Alignments

Given a parallel treebank, we also have a set of word alignments between the sentence pairs i.e. alignments between pre-terminal nodes. Word alignments are vital to core tasks in SMT.

We use treebank based word alignments in place of statistical word alignments in MT for

- ▶ phrase translation model extraction
- ▶ lexical weight scoring

Experiment III - Treebank-Based Lexical Weights

- ▶ Lexical weights are calculated bidirectionally for each phrase pair based on the word alignment between the source and target phrases.
- ▶ Done using the lexical translation probability distribution produced by Giza++

Experiment III - Treebank-Based Lexical Weights

- ▶ Lexical weights are calculated bidirectionally for each phrase pair based on the word alignment between the source and target phrases.
- ▶ Done using the lexical translation probability distribution produced by Giza++
- ▶ We substitute this with a distribution calculated over the word alignments in the parallel treebank
 - ▶ treebank word alignment only (Treebank_weights)
 - ▶ union of SMT and treebank word alignments (Union_weights)

Experiment III - Treebank-Based Lexical Weights

- ▶ Lexical weights are calculated bidirectionally for each phrase pair based on the word alignment between the source and target phrases.
- ▶ Done using the lexical translation probability distribution produced by Giza++
- ▶ We substitute this with a distribution calculated over the word alignments in the parallel treebank
 - ▶ treebank word alignment only (Treebank_weights)
 - ▶ union of SMT and treebank word alignments (Union_weights)

| Config. | BLEU | NIST | %METEOR |
|------------------|---------------|---------------|--------------|
| Baseline+Tree | 0.3397 | 7.0891 | 57.82 |
| Treebank_weights | 0.3356 | 7.0355 | 57.32 |
| Union_weights | 0.3355 | 7.0272 | 57.41 |

Experiment IV - Treebank-Driven Phrase Extraction

- ▶ Phrase pairs are extracted using heuristics over the statistical word alignment

Experiment IV - Treebank-Driven Phrase Extraction

- ▶ Phrase pairs are extracted using heuristics over the statistical word alignment
- ▶ We create new models by running the heuristics over two different word alignments:
 - ▶ treebank word alignment only (Treebank_extr)
 - ▶ union of SMT and treebank word alignments (Union_extr)

Experiment IV - Treebank-Driven Phrase Extraction

- ▶ Phrase pairs are extracted using heuristics over the statistical word alignment
- ▶ We create new models by running the heuristics over two different word alignments:
 - ▶ treebank word alignment only (Treebank_extr)
 - ▶ union of SMT and treebank word alignments (Union_extr)

| Config. | BLEU | NIST | %METEOR |
|---------------|---------------|---------------|--------------|
| Baseline | 0.3341 | 7.0765 | 57.39 |
| +Tree | 0.3397 | 7.0891 | 57.82 |
| Treebank_extr | 0.3102 | 6.6990 | 55.64 |
| +Tree | 0.3199 | 6.8517 | 56.39 |
| Union_extr | 0.3277 | 6.9587 | 56.79 |
| +Tree | 0.3384 | 7.0508 | 57.88 |

Experiment IV - Treebank-Driven Phrase Extraction

An interesting observation

- ▶ Model Union_{extr}+Tree sees an insignificant drop in translation scores against the highest scoring system
- ▶ Its phrase table is 56% smaller

Experiment IV - Treebank-Driven Phrase Extraction

An interesting observation

- ▶ Model Union_{extr}+Tree sees an insignificant drop in translation scores against the highest scoring system
- ▶ Its phrase table is 56% smaller

| Word Alignment | #Phrases | #Phrases+Tree |
|-----------------------|-----------------|----------------------|
| Moses | 24.7M | 29.7M |
| Treebank | 88.5M | 92.89M |
| Union | 7.5M | 13.1M |

Introduction

Experimental Setup

Experiments

Conclusions and Future Work

Conclusions

- ▶ improving SMT by supplementing models with treebank phrase pairs scales
- ▶ treebank word alignments lack sufficient recall to have a positive impact within the SMT framework
- ▶ we can use treebanks to help extract smaller translation models with minimal loss of translation accuracy

Conclusions

- ▶ improving SMT by supplementing models with treebank phrase pairs scales
- ▶ treebank word alignments lack sufficient recall to have a positive impact within the SMT framework
- ▶ we can use treebanks to help extract smaller translation models with minimal loss of translation accuracy

Future Work

- ▶ play with different ways to combine the two phrase resources
- ▶ investigate filtering further
- ▶ apply treebanks to more syntactically-aware MT paradigms e.g. Stat-XFER