Exploiting Parallel Treebanks in Phrase-Based Statistical Machine Translation

John Tinsley

National Centre for Language Technology
Dublin City University
Ireland

Collaborators: Mary Hearne and Andy Way

CICLing 2009
05/03/2009

Overview

Experimental Setup

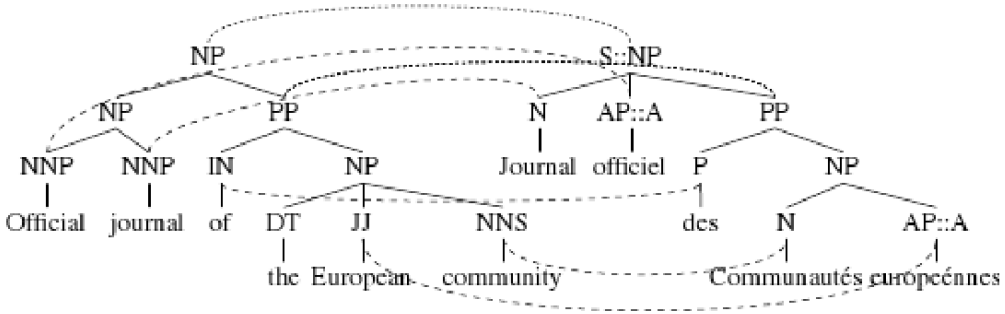
Experiments

Conclusions and Future Work

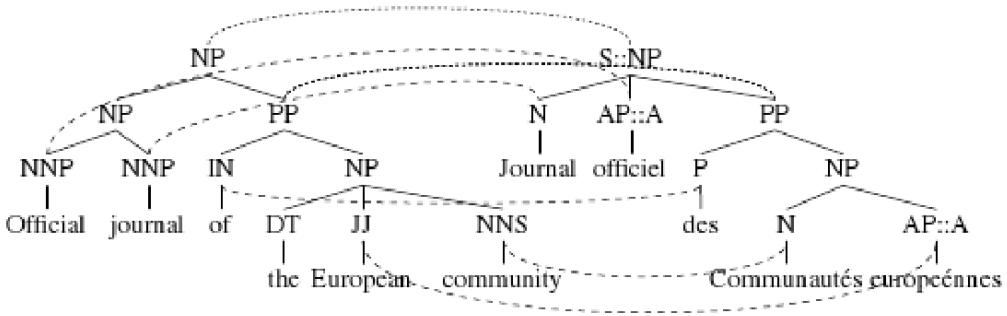
Overview

- Phrase-based SMT systems contain purely statistically induced translation models
- We have demonstrated on small scale that translation accuracy can be improved by supplementing these models with linguistically motivated phrase pairs extracted from parallel treebanks
- Here we test this hypothesis on a large-scale MT task
- We investigate further ways to exploit parallel treebanks in this MT framework

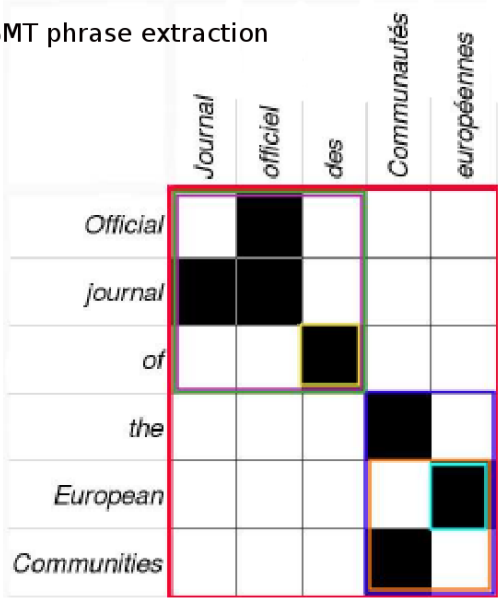
Aligned Parallel Treebank



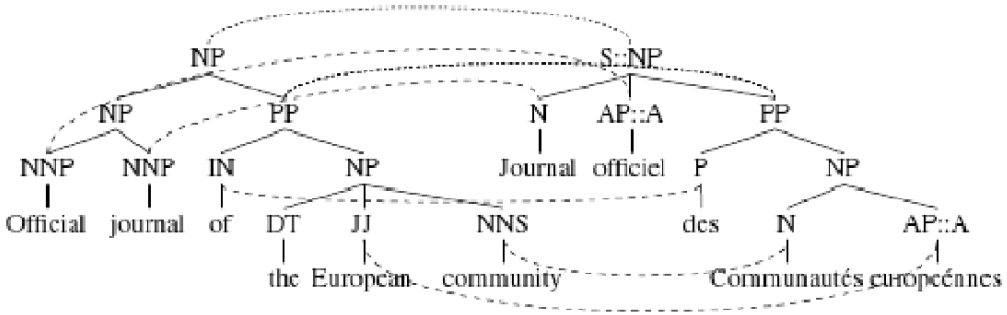
Aligned Parallel Treebank



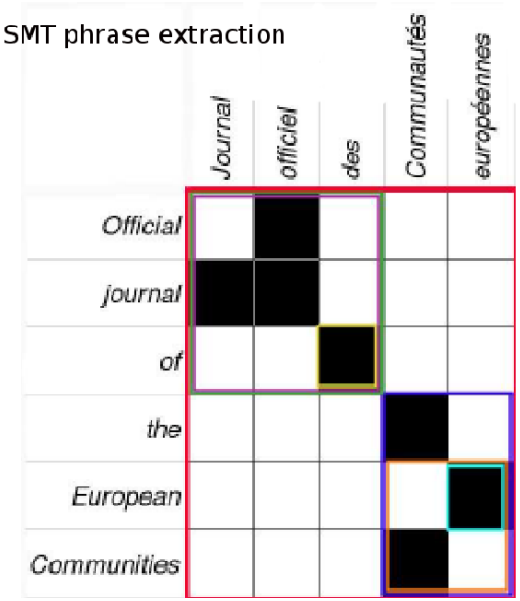
SMT phrase extraction



Aligned Parallel Treebank



SMT phrase extraction



† 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

Data

- 729,891 sentence pairs from English–Spanish Europarl (v2)
- 1,000 sentence devset and 2,000 sentence testset

Data

- 729,891 sentence pairs from English–Spanish Europarl (v2)
- 1,000 sentence devset and 2,000 sentence testset

Parallel Treebank

- Parse both sides monolingually: Berkeley for En; Bikel for Es
- Align using DCU subtree alignment tool

Data

- 729,891 sentence pairs from English–Spanish Europarl (v2)
- 1,000 sentence devset and 2,000 sentence testset

Parallel Treebank

- Parse both sides monolingually: Berkeley for En; Bikel for Es
- Align using DCU subtree alignment tool

MT System

- Baseline PB-SMT system built with Moses
- 5-gram language model (SRILM)
- Minimum error-rate training on devset
- Automatic evaluation using Bleu, Nist and Meteor

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**	**Treebank**
Unique Types	23,261,022	4,985,266
Overlap	1,447,505	
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 - Treebank-Driven Phrase Extraction

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

Experiment II - 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 II - 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	5639
Union_extr	0.3277	6.9587	56.79
+Tree	0.3384	7.0508	**57.88**

Experiment II - Treebank-Driven Phrase Extraction

An interesting observation

- Model Union_extr+Tree gives comparable translation performance to the highest scoring system
- Its phrase table is 56% smaller

Experiment II - Treebank-Driven Phrase Extraction

An interesting observation

- Model Union_extr+Tree gives comparable translation performance to the highest scoring system
- Its phrase table is 56% smaller

Word Alignment	**#Phrases**	**#Phrases+Tree**
Baseline	24.7M	29.7M
Treebank	88.5M	92.89M
Union	7.5M	13.1M

Further Experiments

1. Giving additional weight to treebank phrase pairs in the model
2. Filtering longer phrase pairs from the model
3. Using treebank word alignments to calculate lexical weighting feature in translation model

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 lexical alignments to extract smaller translation models with competative translation quality

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 lexical alignments to extract smaller translation models with competative translation quality

Future Work

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

Thank you

http://computing.dcu.ie/∼jtinsley

http://nclt.dcu.ie/mt

References

- Tinsley, J., V. Zhechev, M. Hearne and A. Way. 2007a.
  Robust Language Pair-Independent Sub-Tree Alignment.
  In Machine Translation Summit XI. Copenhagen, Denmark. p.467–474
- Hearne, M., J. Tinsley, V. Zhechev, and A. Way. 2007.
  Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner.
  In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation. Skvde, Sweden. p.83–94
- Tinsley, J., M. Hearne and A. Way. 2007b.
  Exploiting Parallel Treebanks for use in Statistical Machine Translation.
  In Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT '07). Bergen, Norway. p.175–187
- Hearne, M., S. Ozdowska, J. Tinsley, 2008.
  Comparing Constituency and Dependency Representations for SMT Phrase-Extraction.
  In Actes de la 15éme Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN '08), Avignon, France.

Experiment III - 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 III - 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 III - 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 III - 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 IV - 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 IV - 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 IV - 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