TC-Star OpenLab on Speech Translation
Trento, 30/3 - 1/4/2006

Improving Word Alignment and Translation Quality using Morpho-syntactic Information

Maja Popović

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University
D-52056 Aachen

Overview

- **Introduction**
- **Improving Alignment Quality**
- **Improving Translation Quality**
- **Conclusion**

Introduction

Use morphological and syntactic knowledge to improve

- **word alignment quality**
- **translation quality**

Sparseness of training data

- **the quality of a translation system usually increases with an increasing size of the bilingual training corpus**
- **small corpora have certain advantages**

⇒ **use morpho-syntactic information in order to achieve acceptable translation quality with a minimal amount of bilingual training data**

Corpus Statistics
- EPPS Corpus -

Training:		Spanish	English
full	Sentences	1281427	
	Running Words+Punctuation	36578514	34918192
	Vocabulary	153124	106496
	Singletons [%]	35.2	36.2
reduced	Sentences	13360	
	Running Words+Punctuation	385198	366055
	Vocabulary	22425	16326
	Singletons [%]	47.6	43.7
Develop:	Sentences	1008	
	Running Words+Punctuation	25778	26070
	Distinct Words	3895	3173
	OOVs (full) [%]	0.15	0.09
	OOVs (reduced) [%]	2.7	1.7
Test:	Sentences	840	1094
	Running Words	22774	26917
	Distinct Words	4081	3958
	OOVs (full) [%]	0.14	0.25
	OOVs (running words) [%]	2.8	2.6

Improving Alignment Quality

Investigating different types of morpho-syntactic transformations:

- **replacing Spanish adjectives with base forms**
- **replacing Spanish verbs with base forms and reduced POS tags**
- **reordering of Spanish nouns and adjectives**

Evaluation: compare produced alignment with a reference alignment:

- **few sure (S) connections for unambiguous alignments**
- **a lot of possible (P) connections for ambiguous alignments ($S \subseteq P$)**

Error types:

- **recall error: sure alignment is not found**
- **precision error: a found alignment is not possible**
- **alignment error rate (AER): derived from F-measure**

Alignment Error Rates (AER)

full corpus	e→s	s→e	union
baseline	18.99	18.62	17.68
adjective base	18.77	18.39	17.50
reduced verbs	18.77	18.70	17.80
reorder adjective	19.25	18.73	17.97

small corpus	e→s	s→e	union
baseline	30.43	28.78	29.12
adjective base	29.33	28.31	28.25
reduced verbs	29.13	28.44	28.51
reorder adjective	30.08	28.04	28.61

Improving Translation Quality

- **Differences in word order between two languages cause translation errors**

  ⇒ use morpho-syntactic information to “harmonise” word order

- **Spanish↔English**
  - **local reordering of adjectives and nouns**

Spanish

original: la Unión Europea y la Europa Central y Oriental

reordered: la Europea Unión y la Central y Oriental Europa

English

original: the European Union and the Central and Eastern Europe

reordered: the Union European and the Europe Central and Eastern

Popović & Ney 06:

POS-based Word Reorderings for Statistical Machine Translation

To appear in Proceedings of LREC 2006

Translation Results
Spanish→English

full corpus	dev			test		
	WER	PER	BLEU	WER	PER	BLEU
baseline	33.0	24.2	57.5	34.5	25.5	54.7
reorder adjective	32.4	23.9	58.3	33.5	25.2	56.4

small corpus	dev			test		
	WER	PER	BLEU	WER	PER	BLEU
baseline	39.2	28.4	48.7	41.8	30.7	43.2
reorder adjective	37.9	28.3	50.7	38.9	29.5	48.5
+adjective base	37.4	28.0	51.2	41.0	29.6	45.6

Translation Results
English→Spanish

full corpus	dev			test		
	WER	PER	BLEU	WER	PER	BLEU
baseline	39.8	30.2	50.5	39.7	30.6	47.8
reorder adjective	39.7	30.2	50.9	39.6	30.5	48.3

small corpus	dev			test		
	WER	PER	BLEU	WER	PER	BLEU
baseline	48.3	35.8	40.6	49.6	37.4	36.2
reorder adjective	47.4	35.6	41.7	48.1	36.5	37.7
+align adjective base	47.2	35.5	41.8	47.9	36.4	38.0

Word Reordering - Detailed Translation Results
- Spanish → English -

Spanish→English			dev			test		
			WER	PER	BLEU	WER	PER	BLEU
full corpus	reordered	baseline	34.3	24.4	57.1	35.0	25.1	54.9
		reorder adjective	33.4	24.1	58.2	33.8	24.8	57.0
	rest	baseline	30.1	24.0	58.4	33.1	26.6	54.2
		reorder adjective	30.0	23.6	58.4	32.9	26.5	54.4
small corpus	reordered	baseline	40.8	28.6	47.7	42.6	30.6	42.8
		reorder adjective	38.5	28.2	50.8	39.0	29.1	49.2
	rest	baseline	35.5	28.1	51.2	39.2	31.2	44.2
		reorder adjective	35.4	28.2	51.2	38.6	30.7	46.3

Word Reordering - Detailed Translation Results - English → Spanish -

English→Spanish			dev			test		
			WER	PER	BLEU	WER	PER	BLEU
full corpus	reordered	baseline	41.1	30.6	49.8	40.7	30.9	47.0
		reorder adjective	40.7	30.4	50.5	40.6	30.8	47.6
	rest	baseline	36.2	29.1	52.5	36.8	29.9	50.2
		reorder adjective	36.4	29.3	52.3	36.8	29.8	50.3
small corpus	reordered	baseline	49.4	36.0	40.0	50.5	37.3	35.4
		reorder adjective	48.2	35.7	41.5	48.7	36.4	37.3
	rest	baseline	44.6	35.4	42.4	46.4	37.3	38.9
		reorder adjective	44.5	35.4	42.4	45.8	36.8	39.2

Translation Examples

original Spanish sentence: reordered Spanish sentence:	...la situación claramente insatisfactoria de los derechos humanos en dicho país... ...la claramente insatisfactoria situación de los humanos derechos en dicho país...
Generated English sentence without reordering: with reordering:	...the situation clearly unsatisfactory of human rights in the country... ...the clearly unsatisfactory situation of human rights in the country...
reference English sentence:	...the clearly unsatisfactory situation of human rights in that country...

Conclusion

- Improvements of alignment quality by going beyond the full forms of the words
- Improvements of translation quality by:
  - harmonising the sentence structure between the two languages
  - enabling acceptable translation quality for sparse training data

References

- Collins & Koehn^{+} 05
  Clause Restructuring for Statistical Machine Translation.
  43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL)
  pages 531–540, Ann Arbor, Michigan, June.
- Niessen & Ney 00
  Improving SMT quality with morpho-syntactic analysis.
  18th International Conference on Computational Linguistics (CoLing)
  pages 1081–1085, Saarbrücken, Germany, July.
- Niessen & Ney 01a
  Morpho-syntactic analysis for Reordering in SMT.
  Proc. MT Summit VIII,
  pages 247–252, Santiago de Compostela, Galicia, Spain, September.
- Niessen & Ney 04
  Statistical machine translation with scarce resources using morpho-syntactic information.
  Computational Linguistics, 30(2):181–204

References

- **Popović & Ney 04b**
  Improving Word Alignment Quality using Morpho-Syntactic Information.
  20th International Conference on Computational Linguistics (CoLing) **,**
  pages 310–314, Geneva, Switzerland, August.

- **Popović & Vilar$^+$ 05**
  Augmenting a Small Parallel Text with Morpho-syntactic
  Language Resources for Serbian-English.
  ACL Workshop on Building and Using Parallel Texts:
  Data-Driven Machine Translation and Beyond**,**
  pages 41–48, Ann Arbor, Michigan, June.

- **Popović & Ney 06**
  POS-based Word Reorderings for Statistical Machine Translation
  To appear: ***5th Int. Conf. on Language Resources and Evaluation (LREC)*****,**
  Genoa, Italy, May.