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

Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output

Maja Popović

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

Overview

- **Introduction**
- **Morpho-syntactic Information and Automatic Evaluation**
- **Error Analysis**
- **Conclusion**

Introduction

- **standard automatic evaluation measures (WER, PER, BLEU, NIST) do not give any details about actual translation errors**
- **subjective evaluation is time-consuming**

$\Rightarrow$ **use morpho-syntactic information in combination with automatic error measures**

Popović & de Gispert$^+$ 06:

Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output

Submitted to HLT/NAACL Workshop on Statistical Machine Translation 2006, FinTal 2006

Error Analysis

Possible problems for translation of the Spanish-English language pair:

- **reordering errors**

$\Rightarrow$ **compare WER and PER**
(large difference $\leftrightarrow$ reordering errors)

- **inflectional errors**

$\Rightarrow$ **compare PER of full forms and PER of base forms**
(large difference $\leftrightarrow$ inflectional errors)

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**

Translation System

- state-of-the-art translation system
- log-linear combination of seven models:
  - phrase-based models (source to target and target to source)
  - single word based models at phrase level (source to target and target to source)
  - language model
  - phrase penalty and word penalty
- results comparable to those obtained in the first TC-Star evaluation campaign

Error Analysis - Evaluation Results
- reordering errors -
relative difference between WER and PER

English output:

$1 - \frac{PER}{WER}$	full corpus		red. corpus	
	dev	test	dev	test
nouns+adjectives	24.7	24.7	27.8	25.7
+reordering	21.6	20.8	21.2	20.1
verbs	4.9	4.1	4.9	4.6
adjectives	8.4	10.2	6.8	8.4
nouns	19.8	20.1	19.8	19.1

Spanish output:

$1 - \frac{PER}{WER}$	full corpus		red. corpus	
	dev	test	dev	test
nouns+adjectives	20.5	21.5	22.9	22.9
+reordering	18.9	20.3	20.6	19.8
verbs	3.2	3.3	3.4	3.9
adjectives	6.0	5.6	6.0	5.4
nouns	18.2	16.9	21.2	19.3

Error Analysis - Evaluation Results
- inflectional errors -
PER for different word classes

English output:

PER	full corpus		red. corpus	
	dev	test	dev	test
verbs	41.0	44.8	51.8	56.1
adjectives	28.2	27.3	38.2	38.1
nouns	22.6	23.0	39.2	31.7

Spanish output:

PER	full corpus		red. corpus	
	dev	test	dev	test
verbs	59.5	61.4	70.4	73.0
adjectives	40.4	41.8	50.0	50.9
nouns	27.8	28.5	35.0	37.0

Error Analysis - Evaluation Results
- inflectional errors -
relative difference between base form PER and full form PER

Spanish output:

$1 - \frac{basePER}{fullPER}$	full corpus		red. corpus	
	dev	test	dev	test
verbs	25.9	26.9	25.9	23.7
adjectives	6.2	9.3	12.8	15.1
nouns	7.5	8.4	7.4	6.5

Conclusion

- framework for automatic analysis of translation errors based on morpho-syntactic information
- results correspond to the results of the manual error analysis reported in [Vilar & Xu$^+$ 06]
- improvements of the baseline system adequately reflected on new measures

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

reduced 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

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

reduced 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

References

- **Banerjee & Lavie 05**
  METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments
  In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation an d/or Summarization*, pages 65–72, 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 & Och$^+$ 00**
  An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research
  ***Proc. 2nd Int. Conf. on Language Resources and Evaluation (LREC)*, pages 39–45, Athens, Greece, May.**
- **Niessen & Ney 04**
  Statistical machine translation with scarce resources using morpho-syntactic information.
  ***Computational Linguistics*, 30(2):181–204**

References

- **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.
- **Popović & de Gispert^{+} 06:**
  Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output
  Submitted to HLT/NAACL Workshop on Statistical Machine Translation 2006, FinTal 2006
- **Vilar & Xu^{+} 06**
  Error Analysis of Statistical Machine Translation Output
  To appear: *5th Int. Conf. on Language Resources and Evaluation (LREC)*, Genoa, Italy, May.