Statistical Analysis of Alignment Characteristics for Phrase-based Machine Translation

Patrik Lambert[1], *Simon Petitrenaud*[1], *Yanjun Ma*[2] *and Andy Way*[2]

1. LIUM (Computing Laboratory)
University of Le Mans
France
–
2. CNGL and School of computing
Dublin City University
Ireland

EAMT 2010

Questions

Alignment quality metrics (AER or F-score) do not correlate well with MT metrics (such as BLEU score)

1. Are there other alignment characteristics that may help to improve MT metrics ?
2. How such characteristics depend on parameters such as:
   - The language pair
   - The corpus size or type
   - The type of MT system

Questions

Alignment quality metrics (AER or F-score) do not correlate well with MT metrics (such as BLEU score)

1. Are there other alignment characteristics that may help to improve MT metrics ?
2. How such characteristics depend on parameters such as:
   - The language pair
   - The corpus size or type
   - The type of MT system

⇒ we performed a statistical analysis of alignment characteristics for:

- 2 language pairs: Spanish–English (EsEn), Chinese–English (ZhEn)
- 2 distinct tasks (Europarl, BTEC)
- 3 different corpus sizes for EsEn
- only one MT system (phrase-based SMT)

Main Alignment Characteristics Investigated

- Recall (R), Precision (P), F-Score (F)
- Distortion (dist) (average diff. between source and target positions of a link)
- Percentage of crossing links (crosspl), crossing link distortion (clen)
- Number of links (links), Number of unlinked words (unlnk)
- Distribution of words involved in 1-to-1, 1-to-many (1-to-n) or any-to-many (1n-to-m: 1-to-n+n-to-m)
- Some other variables which had to be discarded in the statistical analysis stage

Alignment systems used

- Discriminative alignment system implementing a log-linear combination of features functions such as:
  - word association features based on IBM1 probabilities
  - unlinked word penalty feature
  - distortion features
  - link bonus feature
- Features weights tuned alternatively according to: alignment F-score, BLEU of phrase-based SMT (Moses), BLEU of n-gram-based SMT (MARIE)
- 2 optimisations for each tuning criterion: 2x3 different alignments+initial set of weights=7
- Combination of IBM model 4 source-target and target-source alignments (Intersection, Union, grow-diag-final heuristic)

$\Rightarrow$ 10 systems in total

Alignment Tuning Procedure

- With a given set of weights: perform alignment and
  - compare alignments to manually aligned reference to calculate F-score or
  - build an SMT system from alignments and compare output to a reference to calculate BLEU score
- Use an optimisation algorithm (here Simultaneous Perturbation Stochastic Approximation) to find the (locally) optimal set of weights
- Objective function to be maximised:
  - $F(\lambda_1, \ldots, \lambda_N)$ or
  - $BLEU(\lambda_1, \ldots, \lambda_N)$ (for the phrase-based and n-gram-based system)

Data

Spanish–English Europarl task

- TC-STAR OpenLab European Proceedings parallel corpus: 1200k, 100k and 20k sentence pairs ($<$30 words/sentence)
- 14k English word alignment reference divided in dev and test sets
- for MT evaluation:
  - 28k word dev corpus (2 references) for internal SMT MERT
  - 25k word dev corpus (2 refs) to calculate BLEU during alignment tuning
  - 23k word test corpus (2 refs)

Data

Chinese–English BTEC task

- Basic Travel Expression Corpus (41.5k sentence pairs, 9.5 words/sentence for English)
- 2k English word alignment reference divided in dev and test sets
- for MT evaluation:
  - 6.1k word dev corpus (7 references) for internal SMT MERT
  - 5.7k word dev corpus (7 refs) to calculate BLEU during alignment tuning
  - IWSLT 2007 test set (3.2k words, 6 refs)
- Simulated easier task by removing sentences with OOV words from dev and test set

Translation Results

- Alignments tuned according to F-score (F), n-gram-based SMT system BLEU (NB), phrase-based system BLEU (PB).
- From these alignments, build a *phrase-based* SMT system

	Discriminative Aligner		Giza++	
	worst	best	worst	best
EsEn full	55.8 (F, NB)	56.3 (PB)	55.6 (I)	56.7 (U)
EsEn 100k	50.9 (NB)	51.4 (PB)	50.7 (I)	51.2 (GDF)
EsEn 20k	45.9 (NB)	46.5 (PB)	46.0 (I)	46.2 (U,GDF)
ZhEn Easy	37.1 (NB)	38.2 (NB)	35.2 (U)	36.1 (I)
ZhEn	34.7 (NB)	35.6 (PB)	33.1 (U)	34.0 (GDF)

Translation Results

- Alignments tuned according to F-score (F), n-gram-based SMT system BLEU (NB), phrase-based system BLEU (PB).
- From these alignments, build a *phrase-based* SMT system

	Discriminative Aligner		Giza++	
	worst	best	worst	best
EsEn full	55.8 (F, NB)	56.3 (PB)	55.6 (I)	56.7 (U)
EsEn 100k	50.9 (NB)	51.4 (PB)	50.7 (I)	51.2 (GDF)
EsEn 20k	45.9 (NB)	46.5 (PB)	46.0 (I)	46.2 (U,GDF)
ZhEn Easy	37.1 (NB)	38.2 (NB)	35.2 (U)	36.1 (I)
ZhEn	34.7 (NB)	35.6 (PB)	33.1 (U)	34.0 (GDF)

- Discriminative aligner:
  - best systems tuned according to PB BLEU (except ZhEn Easy)
  - better than Giza++ when the corpus used for alignment tuning was not a subset of the SMT system training corpus (all except "EsEn full")
- moderate impact of word alignment in terms of BLEU score

Statistical Analysis Methodology

- a number of alignment variables, 10 individuals (alignment systems)
- Principal Component Analysis ; correlation between BLEU score or number of untranslated words and other variables ; linear regression
- Correlation tests which consist in choosing between the null hypothesis (H_0) and the alternative hypothesis (H_1)
- Correlation coefficient:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_X s_Y}, \qquad (1)$$

- if $\alpha \in]0,1[$ is the risk of rejecting hypothesis H_0 by mistake,
- and S is a threshold such that: if $|r_{XY}| < S$ we accept H_0, otherwise, we reject H_0
- for 10 systems and a risk of 0.05, the threshold is about 0.63.

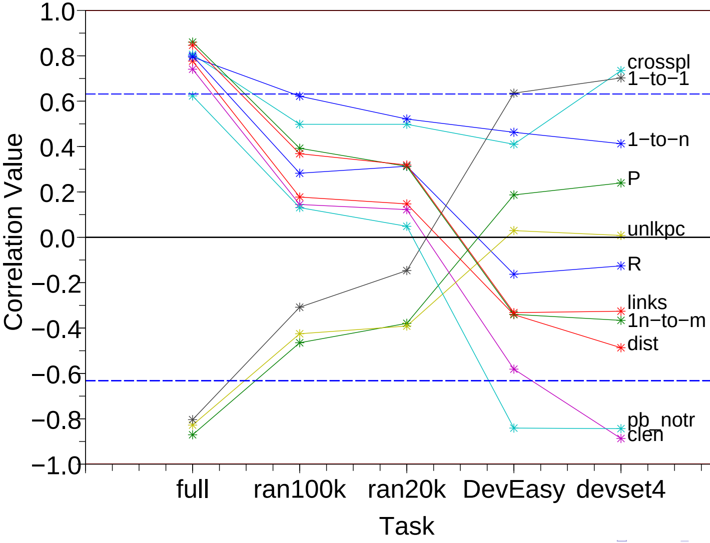
Statistical Analysis

- The hypothesis testing for correlation between two random variables X and Y requires the assumption that both variables are distributed normally.
- Kolmogorov-Smirnov test.
- Example: EsEn full corpus, % of many-to-many alignments or number of gaps in alignment did not pass the test

Statistical Analysis

- The hypothesis testing for correlation between two random variables X and Y requires the assumption that both variables are distributed normally.
- Kolmogorov-Smirnov test.
- Example: EsEn full corpus, % of many-to-many alignments or number of gaps in alignment did not pass the test
- In order to limit the error introduced by MERT, we ran 4 MERT instances
- we computed an interval of possible correlations in a Monte-Carlo way

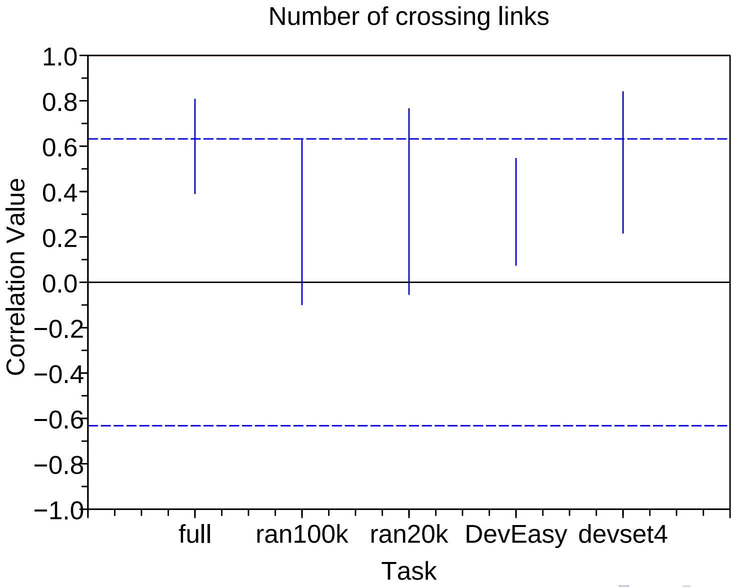
Characteristics helping to help BLEU Score



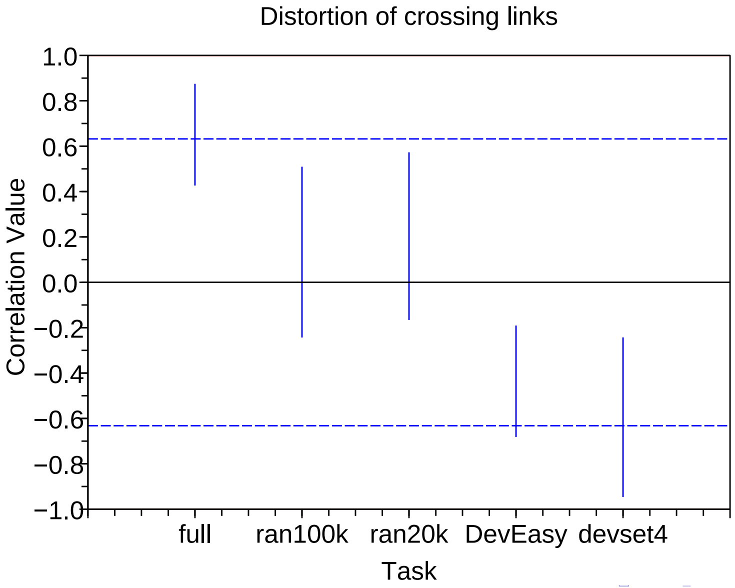
Characteristics helping to help BLEU Score

- the correlation value for most variables is significant (for a risk of 0.05) only in the ‘full’ task.
- trend of the correlation value seems interesting. A number of variables range from negatively correlated to positively correlated with BLEU score depending on the task. ⇒ the impact on BLEU score of these variables greatly depends on the size of the corpus
- note that if the correlation value is decreased below the threshold, it means that the error risk is increased.
- no variable is significantly correlated (positively or negatively) with BLEU score for all corpora (variable most correlated: 1-to-n)
- variables positively correlated with BLEU score in the full task take higher values in dense alignments (and conversely) ⇒ with larger corpora, dense alignments are better for phrase-based SMT, while with smaller corpora, more precise, sparser alignments are required

Characteristics helping to help BLEU Score

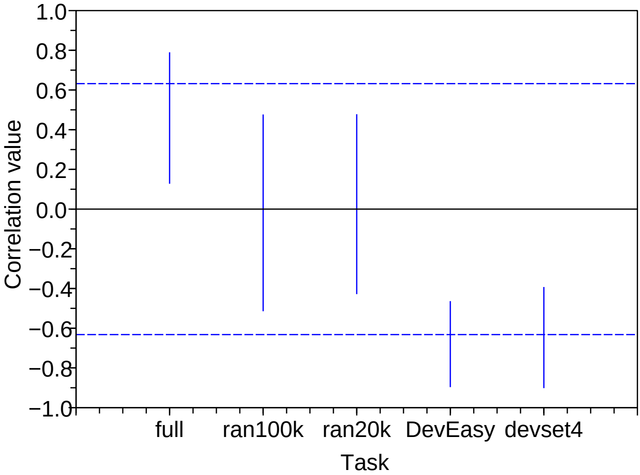


Characteristics helping to help BLEU Score



Number of Untranslated Words

Correlation between the BLEU score and the number of untranslated words.



Number of Untranslated Words

Correlation between the number of untranslated words and a number of alignment variables.

	EsEn full	EsEn 100k	EsEn 20k	ZhEn Easy	ZhEn
1n-to-m	0.585	0.929	0.906	0.724	0.721
links	0.541	0.932	0.897	0.711	0.704
R	0.493	0.944	0.866	0.579	0.551
dist	0.479	0.952	0.965	0.381	0.827
...	...	...	...	...	...
unlnk	-0.461	-0.884	-0.812	-0.452	-0.436
P	-0.564	-0.885	-0.857	-0.582	-0.613
1-to-1	**-0.744**	**-0.957**	**-0.969**	**-0.919**	**-0.918**

- only variable above the significance threshold in all tasks: % of words involved in 1-to-1 alignments

Conclusions

- We performed a statistical analysis of alignment characteristics for 10 alignment systems
- We studied sample correlation coefficients between characteristics and the number of untranslated words as well as the BLEU score

Conclusions

- We performed a statistical analysis of alignment characteristics for 10 alignment systems
- We studied sample correlation coefficients between characteristics and the number of untranslated words as well as the BLEU score
- Limiting the number of untranslated words may improve BLEU score for small tasks like ZhEn BTEC
- This number can be reduced via a higher percentage of one-to-one alignments

Conclusions

- For most tasks no variable is highly correlated with BLEU score

Conclusions

- For most tasks no variable is highly correlated with BLEU score
  However:
- With larger corpora, dense alignments are required while with smaller corpora, more precise, sparser alignments are better for phrase-based SMT
- Crossing links themselves do not seem to be problematic, but avoiding some long-distance crossing links may improve BLEU score when using small corpora

Conclusions

- For most tasks no variable is highly correlated with BLEU score
  However:
- With larger corpora, dense alignments are required while with smaller corpora, more precise, sparser alignments are better for phrase-based SMT
- Crossing links themselves do not seem to be problematic, but avoiding some long-distance crossing links may improve BLEU score when using small corpora
- Main conclusion: the alignment characteristics which help in translation greatly depend on the corpus size

Thank you for your attention !