A New Method for the Study of Correlations between MT Evaluation Metrics

Paula Estrella
Andrei Popescu-Belis
Margaret King

School of Translation and Interpreting
University of Geneva

Introduction

- Correlation with human metrics is a desirable property of automatic metrics
  - Typically adequacy and fluency
- Results are difficult to compare across studies
  - Diversity of results
    - "BLEU correlates 95% with humans" (Papineni et al. 2002) *vs*. "BLEU does not correlate well" (Koehn et al. 2006)
- What factors affect correlation coefficients?
  - Compare two situations: texts from different domains and MT qualities (high vs. low quality)

Plan

- Proposal for computing correlation
- Resources
- General domain
- Specific domain
- High/low translation quality
- Conclusion

Plan

- Proposal for computing correlation
- Resources
- General domain
- Specific domain
- High/low translation quality
- Conclusion

Computing correlation of metrics

- Usually calculated cross-system
  - Final scores of every evaluated system are correlated with fluency or with adequacy scores
    - Small number of sample points
    - Global result for an evaluation
- Our approach: compute a form of correlation for *each system*
  - Use bootstrapping to generate a large number of sample points
    - Artificially generate several samples for each system
  - Hypothesis
    - Correlation should be visible independently of the system, test set, etc
- Why did we choose this approach?
  - Useful if few systems are tested, unlike other forms of correlation
  - Results can be obtained separately for each system

Bootstrapping algorithm

- Statistical method to infer estimators of a variable
  - in MT used for statistical significance tests (Koehn 2004) ; in ASR to estimate c.i. (Bisani & Ney 2004)
- Advantages
  - Applicable to one (or more) system(s)
  - Individual results for each system
- Disadvantage
  - direct comparison with standard correlation not possible

Bootstrapping algorithm (II)

- Given a corpus (set of texts) with *N* segments
  1. Generate a new corpus with *N* segments randomly selected
     - Segments can appear 0 or more times
  2. Apply metrics on the new (= artificial, bootstrapped) corpus
  3. Repeat 1,500 times
  4. Calculate correlation over 1,500 scores

- For consistency of Pearson's *R* coefficients
  - Metrics applied at system level
  - Random numbers fixed for all metrics

- Output: correlation matrixes per system, for any pair of evaluation metrics

Plan

- Proposal for computing correlation
- Resources
- General domain
- Specific domain
- High/low translation quality
- Conclusion

Resources used

- Corpus from the CESTA MTeval campaign
  - 5 systems translating EN → FR
- 1st run: general domain texts from the *Official Journal of the European Communities*
  - 790 segments, ~25 words/segment on average
- 2nd run: systems could adapt to the health domain
  - 288 segments, ~22 words/segment on average

Evaluation metrics

- Human evaluation metrics
  - Fluency and adequacy, average of 2 evaluators
  - 5-point scale, normalized to [0; 1] interval
  - Agreement on 1st run
    - for identical values: fluency 40% | adequacy 37%
    - for 0-1 point difference: fluency 84% | adequacy 78%
  - Agreement on 2nd run
    - for identical values: fluency 41% | adequacy 47%
    - for 0-1 point difference: fluency 84% | adequacy 78%

- Automatic evaluation metrics
  - BLEU, NIST, mWER, mPER, GTM
  - Acceptable cross-system correlations reported by CESTA
    - BLEU/NIST *vs*. adequacy ≈ 0.63
    - BLEU/NIST *vs*. fluency ≈ 0.69

Plan

- Proposal for computing correlation
- Resources
- General domain
- Specific domain
- High/low translation quality
- Conclusion

Texts from general domain

- Correlation calculated on texts from the CESTA “general domain”
- General results
  - Relatively high *R* correlation for metrics of the same family
    - WER *vs*. PER > 0.8, BLEU *vs*. NIST > 0.7, PREC *vs*. REC > 0.76
  - No particular trend between different automatic metrics
    - WER/PER *vs*. BLEU/NIST decrease as system ranking decreases
  - Correlations with human metrics
    - 0.2–0.35 for systems ranked highest or lowest
    - 0.3–0.5 for systems ranked in the middle
    - 0.67–0.71 for adequacy *vs*. fluency
  - NIST has overall lowest correlation scores
- NB: CESTA reports only on adequacy/fluency correlation
  → values are not directly comparable

Plan

- Proposal for computing correlation
- Resources
- General domain
- Specific domain
- High/low translation quality
- Conclusion

Texts from specific domain (health)

- Previously found some low values
  - Specially with human metrics
  - Depends on the system
- Performed experiment on a corpus from a specific domain
  - CESTA corpus for health domain – 288 segments
  - Hypothesis: correlations should improve since systems were specially adapted
- Comparison to previous results
  - NB: slight change in evaluation protocol for humans
  - Majority of systems participating in both campaigns

Results (1/2)

- Values do not change a lot for specific domain
  - Decreased for correlations of adequacy *vs*. fluency
    - E.g. adequacy *vs*. fluency 0.26–0.4 (was 0.6–0.7)
      - Influenced by the change of human evaluation protocol?
- Similar values between automatic metrics
- Special case of system increasing correlations
  - All metrics with adequacy 0.5 – 0.7 but between 0.2 – 0.35 with fluency
  - Only system with better *R* with adequacy than fluency

Results (2/2)

S5

S2 \	WER	BLEU	NIST	ADE	FLU	GTM
WER		-0.82	-0.69	-0.20	-0.28	-0.51
BLEU	-0.87		0.80	0.17	0.21	0.66
NIST	-0.72	0.84		0.21	0.21	0.80
ADE	-0.72	0.68	0.51		0.34	0.16
FLU	-0.27	0.35	0.24	0.27		0.15
GTM	-0.89	0.71	0.62	0.62	0.24	

Plan

- Proposal for computing correlation
- Resources
- General domain
- Specific domain
- High/low translation quality
- Conclusion

High *vs*. low quality translations

- Explore correlation over “good” or “bad” translations
  - Translation quality measured by adequacy/fluency scores
  - Hypothesis: high quality translations should be easier to evaluate → better correlation?
- Empirical threshold for low, respectively high scores
  - Adequacy and fluency > 0.85 and respectively < 0.15
- Analysis performed on output of 2 systems, S2 & S5
  - Extracted 130 low quality segments and 180 high quality segments

Results (1/2)

- S5 outperforms S2 for all metrics on low quality segments
- S2 much better on high quality segments for all metrics applied
- Correlation between adequacy and fluency increases for high quality segments
- Independently of translation quality
  - S2 scores correlate better with fluency
  - S5 with adequacy
  - NIST shows lowest coefficients
  - Correlation still very low despite high inter-judge agreement

Results (2/2)

	High		Low	
	S2	S5	S2	S5
GTM vs. Ade	0.24	0.11	0.24	0.32
GTM vs. Flu	0.41	0.27	0.02	0.13
WER vs. Ade	-0.36	-0.17	-0.1	-0.16
WER vs. Flu	-0.43	-0.25	-0.14	-0.32
BLEU vs. Ade	0.28	0.14	0.18	0.25
BLEU vs. Flu	0.40	0.29	0.06	0.17

- Correlation values for high/low quality segments for S2 and S5

Plan

- Proposal for computing correlation
- Resources
- General domain
- Specific domain
- High/low translation quality
- Conclusion

Conclusions

- Low correlation of human *vs*. automatic metrics
  - Despite high inter-judge agreement
- Stronger correlations remain so regardless of the amount of text used
  - High correlation between automatic metrics of the same family
  - Some acceptable cross-correlations: WER/BLEU, NIST/Prec
- Low quality translations might be more difficult to evaluate
  - They lead to a larger variation of scores
- Coefficients vary depending on system
  - Maybe related to translation algorithms used by systems
  - Could be misleading to present cross-system correlations

Future work

- This work raised even more questions
  - How do we interpret correlations?
  - To what extent should automatic and human metrics correlate?
- We need to further investigate correlation
  - Check our procedure and results
  - Ideally try other setups for human evaluation → costly
- Try metrics that are not n-gram/distance based
  - e.g. METEOR

Thanks for you attention!

Any questions?