CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Sensitivity of automated models for MT evaluation: proximity-based vs. performance-based methods

Bogdan Babych, Anthony Hartley
{b.babych,a.hartley}@leeds.ac.uk
Centre for Translation Studies
University of Leeds, UK

UNIVERSITY OF LEEDS

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Overview

- Classification of automated MT evaluation models
  - Proximity-based vs. Performance-based vs. Hybrid
- Some limitations of MT evaluation methods
- Sensitivity of automated evaluation metrics
  - Declining sensitivity as a limit
- Experiment: measuring sensitivity in different areas of the adequacy scale
  - BLEU vs. NE-recognition with GATE
- Discussion: can we explain/predict the limits?

Classification of MT evaluation models

- Reference proximity methods (BLEU, Edit Distance)
  - Measuring distance between MT and a "gold standard" translation
    - *"...the closer the machine translation is to a professional human translation, the better it is"* (Papineni et al., 2002)
- Performance-based methods (X-score, IE from MT...)
  - Measuring performance of some system which uses degraded MT output: *no need for reference*
    - *"...can someone using the translation carry out the instructions as well as someone using the original?"* (Hutchins & Somers, 1992: 163)

Performance-based evaluation

- Metrics rely on the assumptions:
  - MT errors more frequently **destroy** contextual conditions which trigger rules
  - rarely **create** spurious contextual conditions
    - Language redundancy: it is easier to destroy than to create
    - E.g., (Rajman and Hartley 2001)

    X-score = (#RELSUBJ + #RELSUBJPASS – #PADJ – #ADVADJ)
      - sentential level (+) vs. local (–) dependencies
  - contextual difficulties for automatic tools are ~ proportional to relative "quality"
    - (the amount of MT "degradation")

CENTRE FOR TRANSLATION STUDIES at the University of Leeds

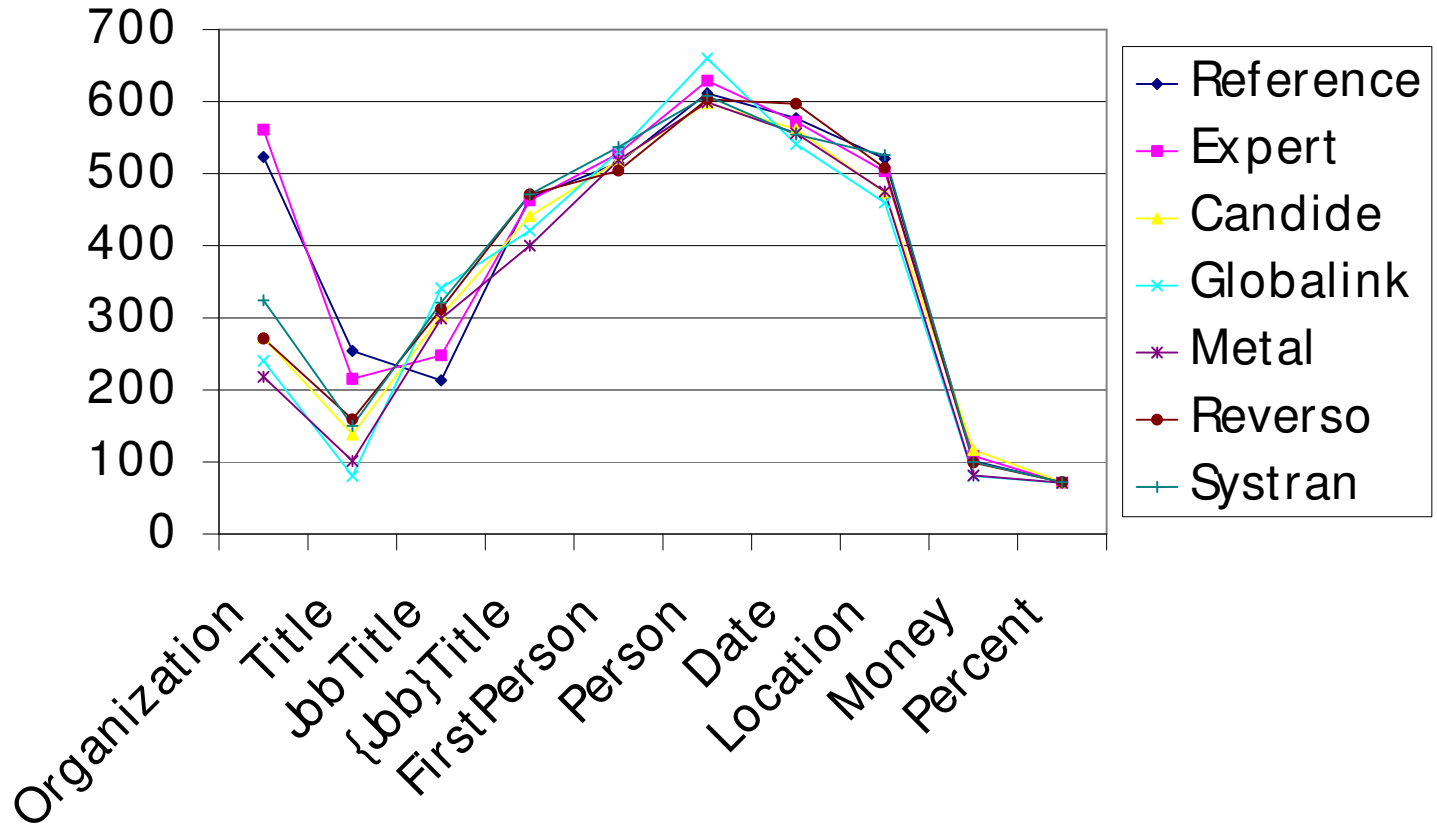
Performance-based evaluation with NE recognition

- NER system (ANNIE) www.gate.ac.uk:
  - the number of extracted Organisation Names gives an indication of Adequacy
  - ORI: … *le chef de la diplomatie égyptienne*
  - HT: *the* <Title>Chief</Title> *of the* <Organization>*Egyptian Diplomatic Corps*</Organization>
  - MT-Systran: *the* <JobTitle> *chief* </JobTitle> *of the Egyptian diplomacy*

UNIVERSITY OF LEEDS

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Performance-based evaluation: number of NEs extracted from MT



CENTRE FOR TRANSLATION STUDIES at the University of Leeds

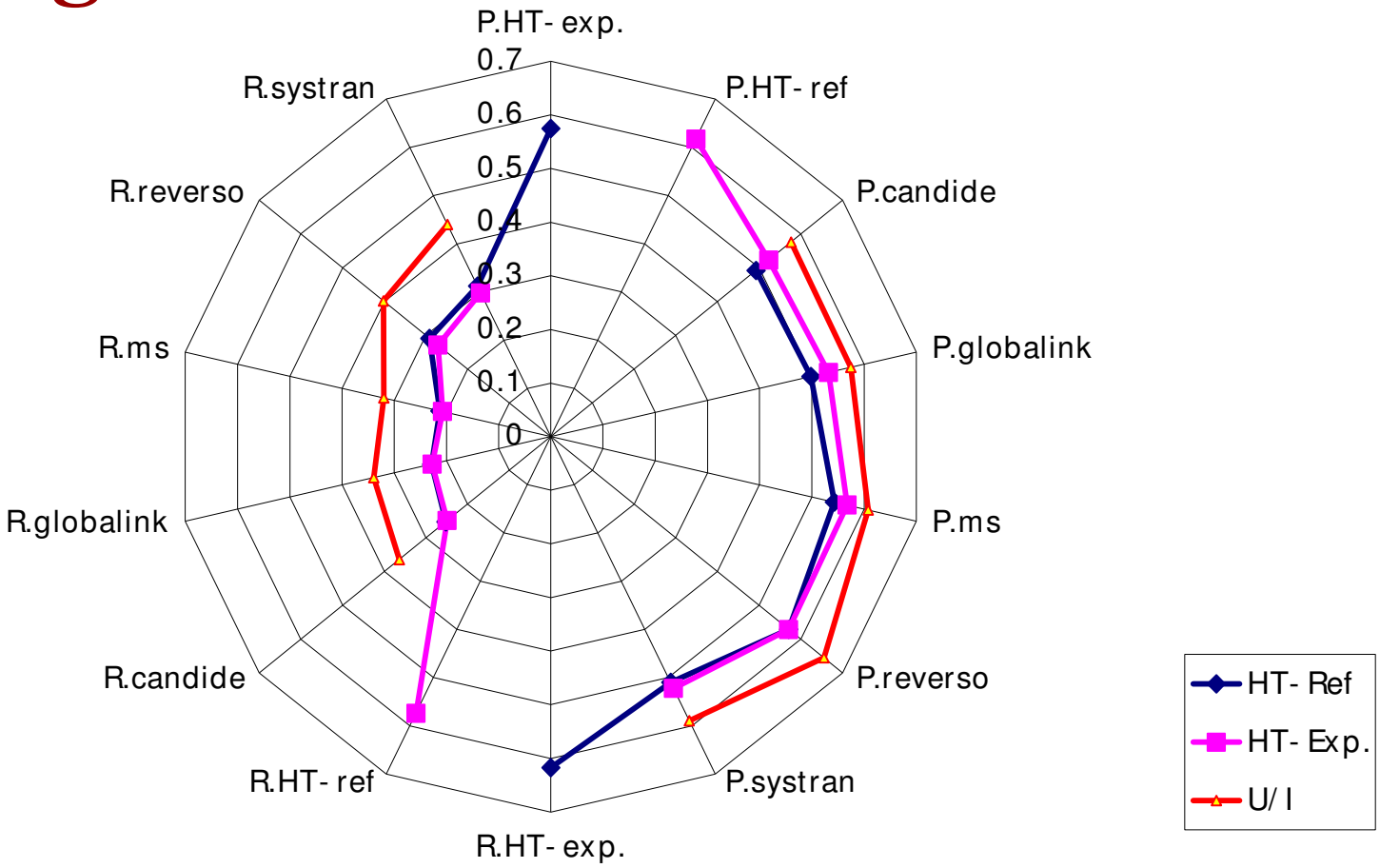
Classification of MT evaluation models: hybrid methods

- Performance+proximity-based
  - Comparing **performance** of an MT system measured by a **proximity** metric (e.g., BLEU) on texts with varying difficulty
    - *need ref.system & correlated automated scores*
    - e.g., Difficulty Slope: shows how systems cope with increasing difficulty of segments / texts (Babych, Hartley, Sharoff, 2007)
      - presentation tomorrow
  - Computing **proximity** to **performance** figures on gold standard translation
    - e.g., Recall of NE extracted from HT vs. MT

UNIVERSITY OF LEEDS

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Hybrid evaluation: Recall of Organisation names in MT vs. HT



Some limits of automated MT evaluation metrics

- Automated metrics useful if applied properly
  - E.g., BLEU: Works for monitoring system's progress, but not for comparison of different systems
    - doesn't reliably compare systems built with different architectures (SMT, RBMT…)

      (Callison-Burch, Osborne and Koehn, 2006)
  - Low correlation with human scores on text/sent. level
    - min corpus ~7,000 words for acceptable correlation
    - not very useful for error analysis

CENTRE FOR TRANSLATION STUDIES at the University of Leeds

… limits of evaluation metrics – beyond correlation

- High correlation with human judgements not enough
  - End users often need ***to predict human scores*** having computed automated scores *(MT acceptable?)*
  - Need regression parameters: Slope & Intercept of the fitted line
- Regression parameters for BLEU (and its weighted extension WNM)
  - are different for each Target Language & Domain / Text Type / Genre
  - BLEU needs re-calibration for each TL/Domain combination

(Babych, Hartley and Elliott, 2005)

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Sensitivity of automated evaluation metrics

- 2 dimensions not distinguished by the scores
  - A. there are stronger & weaker systems
  - B. there are easier & more difficult texts / sentences
- A desired feature of automated metrics *(in dimension B):*
  - To distinguish correctly the quality of different sections translated by the same MT system
- Sensitivity is the ability of a metric to predict human scores for different sections of evaluation corpus
  - easier sections receive higher human scores
  - can the metric also consistently rate them higher?

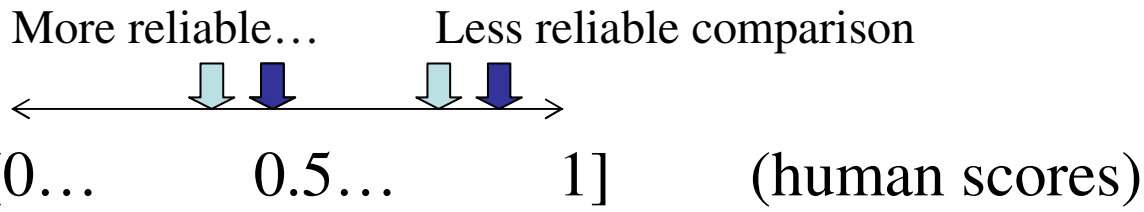
CENTRE FOR TRANSLATION STUDIES at the University of Leeds

Sensitivity of automated metrics – research problems

- Are the dimensions A and B independent?
- Or does the sensitivity *(dimension B)* depend on the overall quality of an MT system *(dimension A)* ?
  - (does sensitivity change in different areas of the quality scale)
- Ideally automated metrics should have homogeneous sensitivity across the entire human quality scale
  - for any automatic metric we would like to minimise such dependence

Varying sensitivity as a possible limit of automated metrics

- If sensitivity declines at a certain area on the scale, automated scores become less meaningful / reliable there
  - For comparing easy / difficult segments generated by the same MT system
  - But possibly also– for distinguishing between systems at that area:



CENTRE FOR TRANSLATION STUDIES at the University of Leeds

Experiment set-up: dependency between Sensitivity & Quality

- Stage 1: Computing approximated sensitivity for each system
  - BLEU scores for each text correlated with human scores for the same text
- Stage 2: Observing the dependency between the sensitivity and systems' quality
  - sensitivity scores for each system (from Stage 1) correlated with ave. human scores for the system
- Repeating the experiment for 2 types of automated metrics
  - Reference proximity-based (BLEU)
  - Performance based (GATE NE recognition)

CENTRE FOR TRANSLATION STUDIES at the University of Leeds

Stage 1: Measuring sensitivity of automated metrics

- Task: to cover different areas on adequacy scale
  - We use a range of systems with different human scores for Adequacy
  - DARPA-94 corpus: 4 systems (1 SMT, 3 RBMT) + 1 human translation, 100 texts with human scores
- For each system the sensitivity is approximated as:
  - r-correlation between BLEU / GATE and human scores for 100 texts

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Stage 2: capturing dependencies: system's quality and sensitivity

- The sensitivity may depend on the overall quality of the system
  - is there such tendency?
- *System-level* correlation between
  - sensitivity (*text-level* correlation figures for each system)
  - and its average human scores
- Strong correlation not desirable here:
  - E.g., strong negative correlation: automated metric looses sensitivity for better systems
  - Weak correlation: metric's sensitivity doesn't depend on systems' quality

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Compact description of experiment set-up

$$rCorrel(System)\left[\frac{humanScore(ade)}{rCorrel(Text)\left[\frac{humanScore(ade)}{bleuScore(n4r1) \vee neGate(organisation)}\right]}\right]$$

- Formula describes the order of experimental stages
- Computation or data + arguments in brackets (in enumerator & denominator)
- Capital letters = independent variables
- Lower-case letters = fixed parameters

CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Results

	BLEU/ade	BLEU/flu	GATE/ade	GATE/flu
system-correl	0.9535	0.995	0.8682	0.9806
sensitivity-correl	**-0.7614**	-0.1265	**-0.2188**	-0.2384

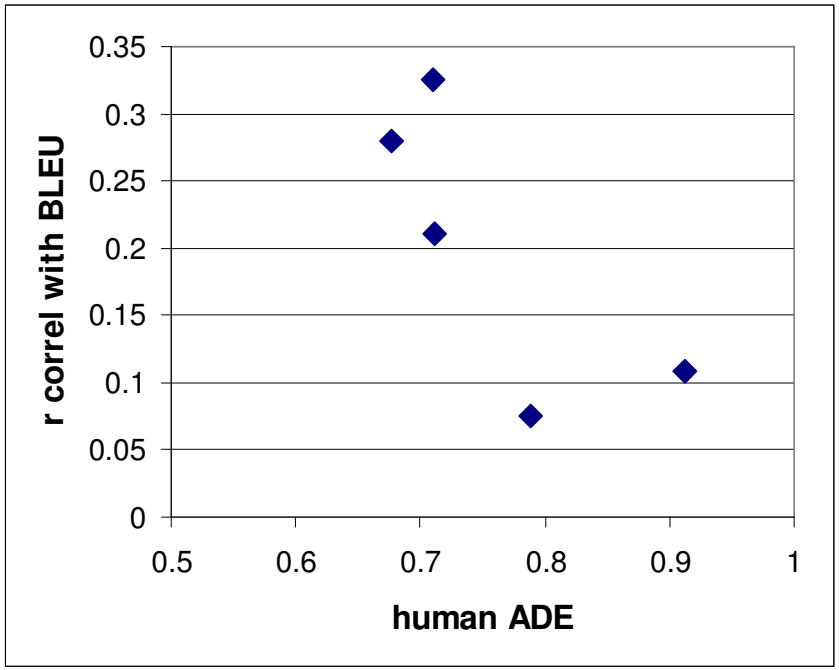
- R-correlation on the system level lower for NE-Gate
- BLEU outperforms GATE
  - But correlation is not the only characteristic feature of a metric …

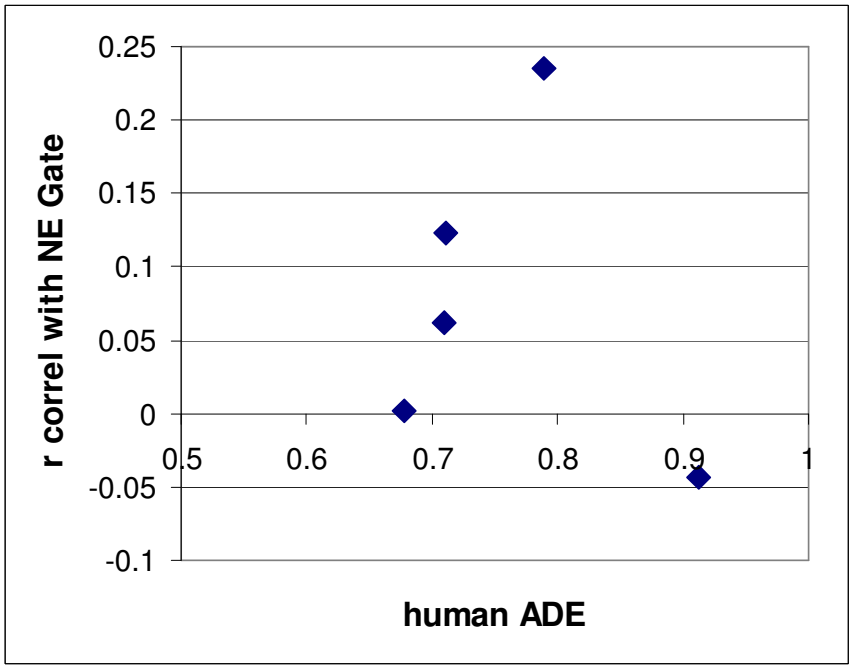
Results

	BLEU/ade	BLEU/flu	GATE/ade	GATE/flu
system-correl	0.9535	0.995	0.8682	0.9806
sensitivity-correl	**-0.7614**	-0.1265	**-0.2188**	-0.2384

- Sensitivity of BLEU is much more dependent on MT quality
  - BLEU is less sensitive for higher quality systems

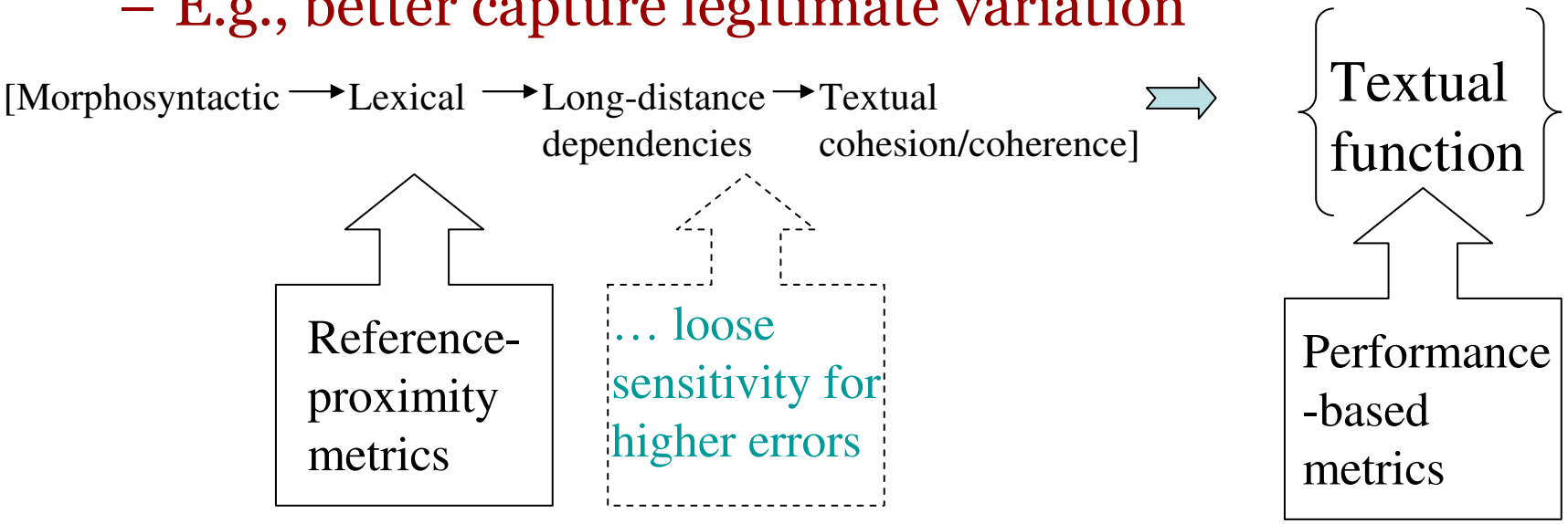
Results (contd.)





Discussion

- Reference proximity metrics use *structural* models
  - Non-sensitive to errors on higher level (better MT)
  - Optimal correlation for certain error types
- Performance-based metrics use *functional* models
  - Potentially can capture degradation at any level
  - E.g., better capture legitimate variation



Conclusions and future work

- Sensitivity can be one of limitation of automated MT evaluation metrics:
  - Influences reliability of predictions at certain quality level
- Functional models which work on textual level
  - can reduce the dependence of metrics' sensitivity on systems' quality
- Way forward: developing performance-based metrics using more adequate functional models
  - E.g., non-local information (models for textual coherence and cohesion…)