Tackling Sparse Data Issue in MT Evaluation

Ondřej Bojar, Kamil Kos, David Mareček
{bojar,marecek}@ufal.mff.cuni.cz, kamilkos@email.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

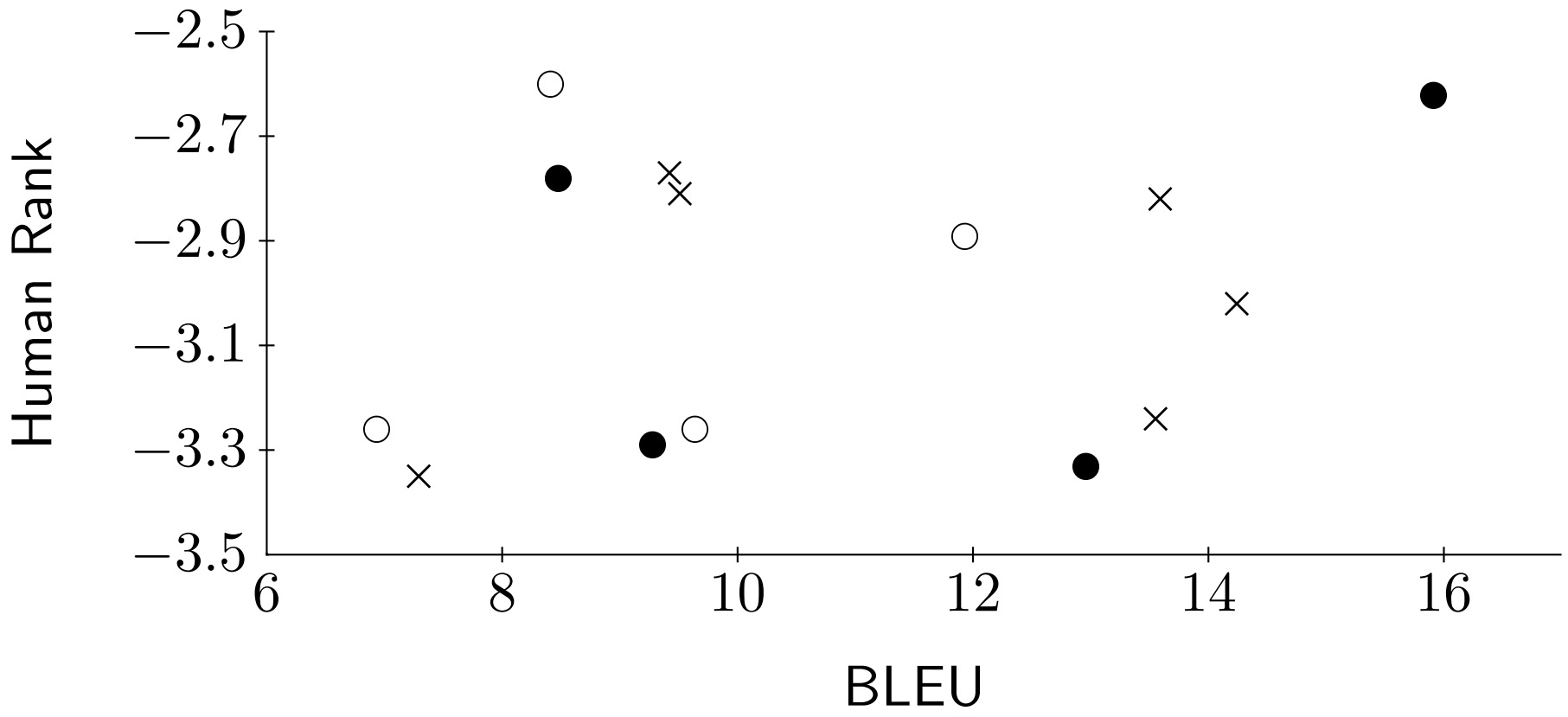
Outline

- Issues of BLEU score.
  - Especially for morphologically rich languages.
- SemPOS
  - Coarser representation of words.
  - Performs better for both Czech and English.

Main Message

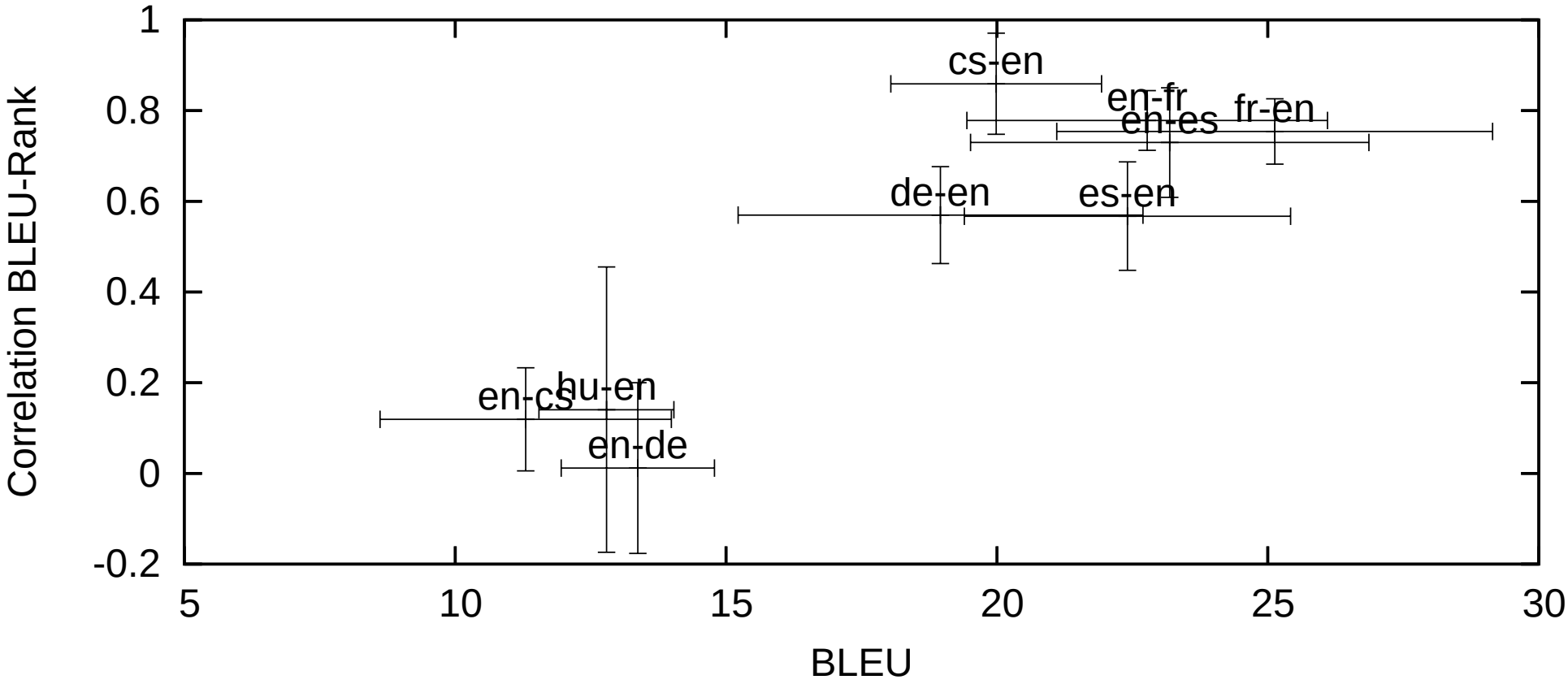
BLEU vs. Human Rank

- English → Czech Translation
  - Evaluation campaigns WMT08 and WMT09.



. . . across Languages

- Various Language Pairs
  - Evaluation campaigns WMT08, WMT09, MetricsMATR.



Reason 1: Focus on Forms

SRC	Prague Stock Market falls to minus by the end of the trading day
REF	pražská burza se ke konci obchodování propadla do minusu
cu-bojar	praha stock market klesne k minus na konci obchodního dne
pctrans	praha trh cenných papírů padá minus do konce obchodního dne

- Only a single unigram in each hyp. confirmed by the reference.
- Large chunks of hypotheses are not compared at all.

Confirmed by Reference	Yes	Yes	No	No
Contains Errors	Yes	No	Yes	No
Running words	6.34%	36.93%	22.33%	**34.40%**

Reason 2: Sequences $\rightsquigarrow$ SemPOS

BLEU overly sensitive to sequences:

- Gives credit for 1, 3, 5 and 8 four-, three-, bi- and unigrams,
- Two of three serious errors not noticed,

  $\Rightarrow$ Quality of cu-bojar overestimated.

SRC	Congress yields: US government can pump 700 billion dollars into banks
REF	kongres ustoupil : vláda usa může do bank napumpovat 700 miliard dolarů
cu-bojar	kongres výnosy : vláda usa může čerpadlo 700 miliard dolarů v bankách
pctrans	kongres vynáší : us vláda může čerpat 700 miliardu dolarů do bank

SemPOS (Kos and Bojar, 2009) gives credit for 8 lemmas:

REF	kongres/n ustoupit/v :/n vláda/n usa/n banka/n napumpovat/v 700/n miliarda/n dolar/n
cu-bojar	kongres/n výnos/n :/n vláda/n usa/n moci/v čerpadlo/n 700/n miliarda/n dolar/n banka/n
pctrans	kongres/n vynášet/v :/n us/n vláda/n čerpat/v 700/n miliarda/n dolar/n banka/n

SemPOS Performs Well

- We evaluated:
  - several variants of SemPOS (e.g. Void, include synt. structure),
  - adaptation of SemPOS for English,
  - linear combination of n-grams and SemPOS.

Correlation with human judgments (selected results):

To English (10 test sets)

Metric	Avg Correlation
Void_{par}	0.75
GTM	0.71
$4{\cdot}\text{SemPOS}{+}1{\cdot}\text{BLEU}_2$	0.70
SemPOS	0.69
BLEU	0.66
TER	0.63

To Czech (3 test sets)

Metric	Avg Correlation
$3{\cdot}\text{SemPOS}{+}1{\cdot}\text{BLEU}_4$	0.55
SemPOS	0.53
GTM	0.35
Void	0.33
BLEU	0.33
TER	0.07

Summary

- BLEU does not correlate with human ranks if below ~20.
  - Too much focus on exact forms and sequences.
    $\Rightarrow > \frac{1}{3}$ of correct tokens not confirmed by reference.

- Suggested SemPOS:
  - Evaluates bags of lemmas
    instead of sequences of words.
- Evaluated several variants of SemPOS.
  - Improvement over several MT evaluation metrics.

MERT optimization towards SemPOS fails, see our poster at WMT10.

References

Kamil Kos and Ondřej Bojar. 2009. Evaluation of Machine Translation Metrics for Czech as the Target Language. Prague Bulletin of Mathematical Linguistics, 92.