Sparse Data Issue in MT Evaluation

Ondřej Bojar, Kamil Kos, David Mareček; {bojar,marecek}@ufal.mff.cuni.cz, kamilkos@gmail.com

Charles University in Prague, Institute of Formal and Applied Linguistics (ÚFAL)

Problems of BLEU

Overly Sensitive to Word 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
SYS1 praha trh cenných papírů padá minus **do** konce obchodního dne
SYS2 praha stock market klesne k minus na **konci** obchodního dne

- Only one word of each hypothesis confirmed by the reference.
- Moreover, the match is a false positive in SYS1.

Large Portions Unscored

- About 1/3 of running words not confirmed by the reference despite not containing any errors (based on manual flagging of errors).
- Fortunately only few false positives (6.34% of running words).

Confirmed	Has Errors	1-grams	2-gms	3-gms	4-gms
Yes	Yes	6.34	1.58	0.55	0.29
Yes	No	36.93	13.68	5.87	2.69
No	Yes	22.33	41.83	54.64	63.88
No	**No**	**34.40**	**42.91**	**38.94**	**33.14**

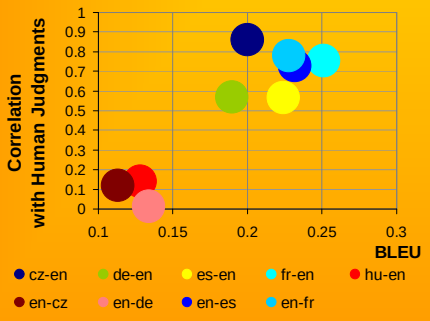
Overly Sensitive to Sequences

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

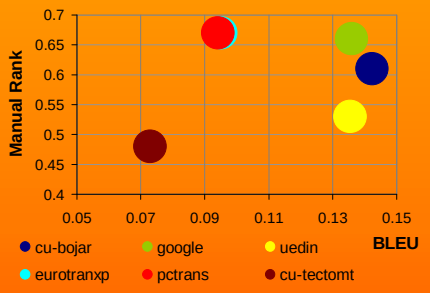
Don't Trust BLEU <20

- cu-bojar scored far better in BLEU thanks to n-gram LM and correct sequences (incl. one 4-gram)
- two important words (mistranslated as nouns instead of verbs by cu-bojar and rendering the sentence incomprehensible) unscored in both cases

…across Languages



for English–Czech



SemPOS: Coarser Metric for a Better Performance

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 (Kos and Bojar, 2009) evaluates the overlapping of lemmas (base forms) of content-bearing words including their semantic part of speech between the reference and the hypothesis.
- In this particular example, pctrans still does not win but at least the quality of cu-bojar is not overestimated.

SemPOS for English

- The original SemPOS (Kos and Bojar, 2009) is based on the tectogrammatical theory as utilized in the Prague Dependency Treebank (Sgall et al., 1986; Hajič et al. 2006) and thus limited to Czech.
- The tectogrammatical layer is being adapted for English (Cinková et al., 2004; Hajič et al., 2009), including automatic annotation tools.

Variants of SemPOS

Various Classifications of Content-Bearing Words
The match in lemma is accepted only within the class of the word:
- SemPOS uses semantic part of speech (noun, verb, …)
- Functor uses tectogrammatical (deep syntactic) dependency edge label (ACTor, PATient, DIRfrom, …)
- Void uses a single class only.

Deep Syntactic Relations
- par requires for each token also the lemma of the parent to match
- sons requires for each token also the set of children's lemmas to match

Combining with BLEU
- SemPOS completely ignores word order.
- We test several linear combinations of SemPOS and $BLEU_x$ (calculated on x-grams only but including the brevity penalty)

Correlation to Human Judgments; SemPOS and Other Metrics

Metric	Avg	Best	Worst
$Void_{par}$	0.75	0.89	0.60
$Void_{sons}$	0.75	0.90	0.54
Void	0.72	0.91	0.59
$Functor_{sons}$	0.72	1.00	0.43
GTM	0.71	0.90	0.54
4 SemPOS + 1 $BLEU_2$	0.70	0.93	0.43
$SemPOS_{par}$	0.70	0.93	0.30
1 SemPOS + 4 $BLEU_3$	0.70	0.91	0.26
4 SemPOS + 1 $BLEU_1$	0.69	0.93	0.43
NIST	0.69	0.90	0.53
$SemPOS_{sons}$	0.69	0.94	0.40
SemPOS	0.69	0.95	0.30
2 SemPOS + 1 $BLEU_4$	0.68	0.91	0.09
$BLEU_1$	0.68	0.87	0.43
$BLEU_2$	0.68	0.90	0.26
$BLEU_3$	0.66	0.90	0.14
BLEU	0.66	0.91	0.20
TER	0.63	0.87	0.29
PER	0.63	0.88	0.32
$BLEU_4$	0.61	0.90	-0.31
$Functor_{par}$	0.57	0.83	-0.03
Functor	0.55	0.82	-0.09

- The classification in English less reliable ⇒ Void performs best.
- Syntactic structure informative for English (par>sons>nothing).
- The combinations of SemPOS and BLEU outperforms both individual metrics and scores best for Czech.
- Automatically assigned functors in Czech deceiving.

Czech-to-English tested on:
(Number of judgments in brackets.)
- MetricsMATR08 (cn+ar: 1652),
- WMT08 News Articles (de: 199, fr: 251),
- WMT08 Europarl (es: 190, fr: 183),
- WMT09 (cz: 320, de: 749, es: 484, fr: 786, hu: 287)

English-to-Czech tested on:
- WMT08 News Articles (en: 267),
- WMT08 Commentary (en: 243),
- WMT09 (en: 1425)

Metric	Avg	Best	Worst
3 SemPOS + 1 $BLEU_4$	0.55	0.83	0.14
2 SemPOS + 1 $BLEU_2$	0.55	0.83	0.14
2 SemPOS + 1 $BLEU_1$	0.53	0.83	0.09
4 SemPOS + 1 $BLEU_3$	0.53	0.83	0.09
SemPOS	0.53	0.83	0.09
$BLEU_2$	0.43	0.83	0.09
$SemPOS_{par}$	0.37	0.53	0.14
$Functor_{sons}$	0.36	0.53	0.14
GTM	0.35	0.53	0.14
$BLEU_4$	0.33	0.53	0.09
Void	0.33	0.53	0.09
NIST	0.33	0.53	0.09
$Void_{sons}$	0.33	0.53	0.09
BLEU	0.33	0.53	0.09
$BLEU_3$	0.33	0.53	0.09
$BLEU_1$	0.29	0.53	-0.03
$SemPOS_{sons}$	0.28	0.42	0.03
$Functor_{par}$	0.23	0.40	0.14
Functor	0.21	0.40	0.09
$Void_{par}$	0.16	0.53	-0.08
PER	0.12	0.53	-0.09
TER	0.07	0.53	-0.23

Full text, acknowledgement and the list of references in the proceedings of ACL 2010.