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 exzeptí buňka se ke konci obchodního dne klesne k minusu
SYS1 práha teh cerných papírů pada minusu do konce obchodního dne
SYS2 práha stock market klene k minusu 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 34.49 42.91 38.94 33.14

Overly Sensitive to Sequences

SRC Congress yields: US government can pump 700 billion dollars into banks
REF zpráva usnihla vpuknutím 700 milárd dolarů do bank
SYS1 kongres vysohla vláda us může vputovat 700 milárd dolarů do bank
SYS2 kongres vysohla us může vputovat 700 milárd dolarů do bank

• 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; Hajic et al. 2006) and thus limited to Czech.
• The tectogrammatical layer is being adapted for English (Cinková et al., 2004; Hajic et al., 2009), including automatic annotation tools.

SemPOS: Coarser Metric for a Better Performance

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 (nouns, verbs, adjectives).
• Functor uses tectogrammatical (deep syntactic) dependency edge label (ACTor, PATient, DIRfrom, …)
• Void uses a single class only.

Deep Syntactic Relations
• part requires for each token also the lemma of the parent to match
• noun 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 (calculated on n-grams only but including the brevity penalty).

Correlation to Human Judgments; SemPOS and Other Metrics

Metric Avg Best Worst
Voss 0.79 0.90 0.69
Voop 0.75 0.90 0.54
Voul 0.72 0.91 0.39
Functorvouk 0.72 1.00 0.43
GTM 0.71 0.90 0.54
SemPOS + 1 BLEU 0.70 0.93 0.43
SemPOS 0.70 0.93 0.30
SemPOS + 1 BLEU 0.70 0.93 0.26
4 SemPOS + 1 BLEU 0.69 0.93 0.43
NIST 0.69 0.90 0.53
SemPOS 0.69 0.94 0.40
SemPOS + 1 BLEU 0.69 0.95 0.38
2 SemPOS + 1 BLEU 0.68 0.91 0.09
BLEU1 0.68 0.87 0.63
BLEU 0.68 0.90 0.26
BLEU1 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
BLEU1 0.63 0.89 0.31
Functor 0.57 0.83 0.03
Functorvouk 0.55 0.82 0.09

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

Czech-to-English Tested on:
• WMT08 News Articles (en: 267), WMT08 Commentary (en: 243), WMT09 (en: 1243)

English-to-Czech Tested on:
• WMT08 News Articles (cz: 320), WMT08 Commentary (cz: 479), WMT09 (cz: 484, fr: 786, hu: 287)

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