Tolerant BLEU: a Submission to the WMT14 Metrics Task

Jindřich Libovický, Pavel Pecina {libovicky, pecina}@ufal.mff.cuni.cz
Charles University in Prague, Institute of Formal and Applied Linguistics

Source:

I am driving a new red car

Reference:

Jedu novým červeným autem

Translation:

Jedu s novém červeném auto

0 | 1/3 | 1/6 | 2/3

Maximum weighted biparite matching w.r.t. to the affix distance

"Corrected" translation:

(Jedu, 1) (s, 1) (novým, 2/3) (červeným, 5/6) (autem, 1/3)

Unigram precision

Jedu	→ Jedu	1	✓
s	→ s	1	✗
novém	→ novým	2/3	✓
červeném	→ červeným	5/6	✓
auto	→ autem	1/3	✓

Bigram precision

Jedu s	→ Jedu s	avg(1,1) = 1	✗
s novém	→ s novým	avg(1, 2/3) = 5/6	✗
novém červeném	→ novým červeným	avg(2/3, 5/6) = 3/4	✓
červeném auto	→ červeným autem	avg(5/6,1/3) = 7/12	✓

tBLEU unigram precision = $= \frac{11}{6} / 5 \approx 0.367$

BLEU unigram precision = $= 1 / 5 = 0.2$

tBLEU bigram precision = $= \frac{16}{12} / 4 \approx 0.333$

BLEU bigram precision = $= 0 / 4 = 0$

Motivation

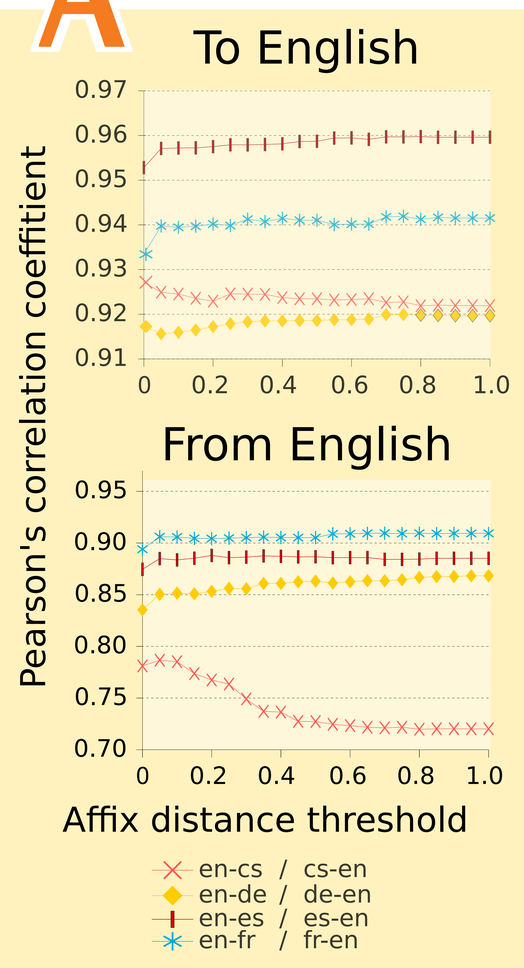
- BLEU treats inflection errors as comletely different words
- wrongly inflected words could be aligned and penalized less than if it were different totally different words

Algorithm

- affix distance – approx. measure of word relatedness
- monolingual alignment of the sentences is found as a maximum weighted bipartite matching w.r.t. the distance
- BLEU algorithm with *n*-gram precision weighted by the alignment

Results

- straightforward, language-independent generalization of the BLEU score
- better correlation with human judgement for translation to morphologically richer languages

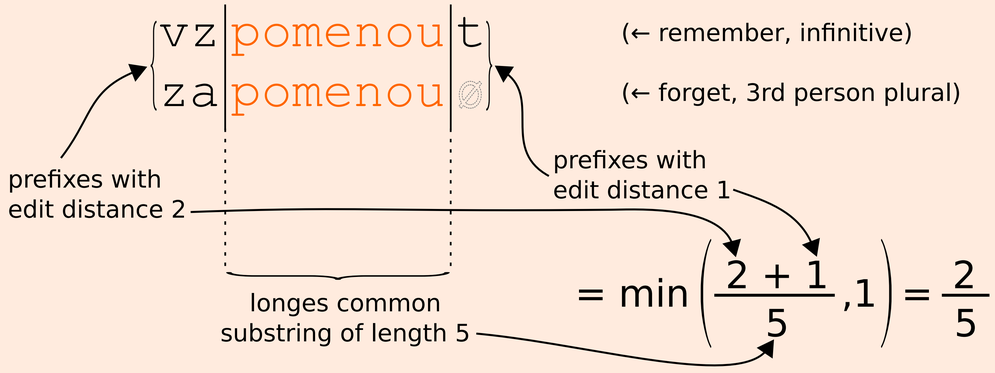


Pearson's correlation of the tBLUE and human judgement on the WMT 13 dataset in comparison with BLEU and METEOR

direction	BLEU	METEOR	tBLEU
en-cs	.781	.860	.787
en-de	.835	.868	.850
en-es	.875	.878	.884
en-fr	.887	.906	.906
from English	.844	.878	.857

direction	BLEU	METEOR	tBLEU
cs-en	.925	.985	.927
de-en	.916	.926	.917
es-en	.975	.968	.953
fr-en	.940	.983	.933
to English	.923	.974	.935

Affix distance computation



Presented at the Ninth Workshop on Statistial Machine Translation
This research has been funded by the Czech Science Foundation (grant n. P103/12/G084) and the EU FP7 project Khresmoi (contract no. 257528).