Approximating a Deep-Syntactic Metric for MT Evaluation and Tuning

Matouš Macháček, Ondřej Bojar; {machacek, bojar}@ufal.mff.cuni.cz; Charles University in Prague, MFF, ÚFAL

Overview

SemPOS metric

- Introduced by Kos and Bojar (2009), inspired by Giménez and Márquez (2007).
- Counts overlapping of deep-syntactic lemmas (t-lemmas) of content words.
- Lemmas are matched only if semantic parts-of-speech (Sgall et al. 1986) agree.
- Does not consider word order and auxiliary words.

Issues

- SemPOS requires full parsing up to the deep syntactic layer.
  => SemPOS is computational costly.
- There are tools assigning t-lemmas and semposes only for Czech and English.
  => SemPOS is difficult to adapt for other languages.

Proposed Solution:

- Approximate t-lemmas and semposes using only tagger output.
  => Faster and more adaptable for other languages.
  => More suitable for MERT tuning.

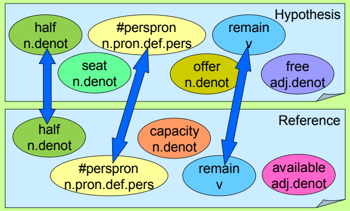
Overlapping

Src: Polovina míst v naší nabídce zůstává volná.
Ref: Half of our capacity remains available.
Hyp: Half of the seats in our offer remains free.

boost-micro (Giménez, Márquez, 2007)

$$O = \frac{\sum_{t\in T}\sum_{w\in r_i}\operatorname{cnt}(w,t,c_i)}{\sum_{t\in T}\sum_{w\in r_i\cup c_i}\max(\operatorname{cnt}(w,t,r_i),\operatorname{cnt}(w,t,c_i))}$$

$$O = \frac{\boxed{3}}{\boxed{1}+\boxed{1}+\boxed{1}+\boxed{1}+\boxed{1}+\boxed{1}+\boxed{1}+\boxed{1}} = \frac{3}{8}$$



Hypothesis
half n.denot
#perspron n.pron.def.pers
remain v
seat n.denot
offer n.denot
free adj.denot

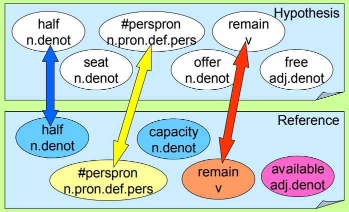
Reference
half n.denot
capacity n.denot
#perspron n.pron.def.pers
remain v
available adj.denot

cap-macro (Bojar, Kos, 2007)

$$O(t) = \frac{\sum_{w\in r_i}\min(\operatorname{cnt}(w,t,r_i),\operatorname{cnt}(w,t,c_i))}{\sum_{w\in r_i}\operatorname{cnt}(w,t,r_i)}$$

$$O = \frac{\sum_{t\in T} O(t)}{|T|}$$

$$O = \frac{\frac{\boxed{1}}{\boxed{2}}+\frac{\boxed{1}}{\boxed{1}}+\frac{\boxed{1}}{\boxed{1}}+\frac{0}{\boxed{1}}}{4} = \frac{5}{8}$$



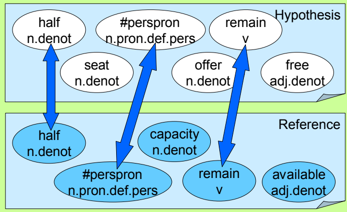
Hypothesis
half n.denot
#perspron n.pron.def.pers
remain v
seat n.denot
offer n.denot
free adj.denot

Reference
half n.denot
capacity n.denot
#perspron n.pron.def.pers
remain v
available adj.denot

cap-micro (our)

$$O = \frac{\sum_{t\in T}\sum_{w\in r_i}\min(\operatorname{cnt}(w,t,r_i),\operatorname{cnt}(w,t,c_i))}{\sum_{t\in T}\sum_{w\in r_i}\operatorname{cnt}(w,t,r_i)}$$

$$O = \frac{\boxed{3}}{\boxed{5}}$$



Hypothesis
half n.denot
#perspron n.pron.def.pers
remain v
seat n.denot
offer n.denot
free adj.denot

Reference
half n.denot
capacity n.denot
#perspron n.pron.def.pers
remain v
available adj.denot

Approximations

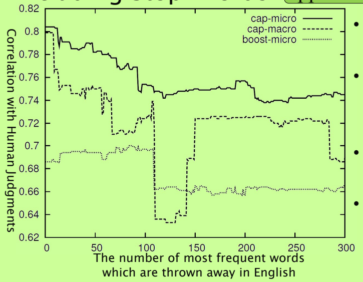
Sempos from Tag approx

- Morphological tag determines sempos.
- CzEng corpus (Bojar and Žabokrtský, 2009) used to create dictinoray which maps morphological tag to most frequent sempos.
- Surface lemmas are used instead of t-lemmas.
- Accuracy on CzEng e-test:
  - 93.6 % for English
  - 88.4 % for Czech

Morph. Tag	Sempos	Rel. Freq.
NN	n.denot	0.989
VBZ	v	0.766
VBN	v	0.953
JJ	adj.denot	0.975
NNP	n.denot	0.999
PRP	n.pron.def.pers	0.999

Example of used dictionary in English

Excluding Stop-Words approx-stopwords



- Deep syntactic layer does not contain auxiliary words.
- Assumed that auxiliary words are the most frequent words, we exclude a certain number of most frequent words.
- Stopwords lists were obtained from CzEng corpus.
- Exact cut-offs:
  - 100 words for English
  - 220 words for Czech

Restricting the Set of Semposes approx-restr

- Contribution of each sempos type to the overall performance can differ a lot.
- We assume that some sempos types raise the correlation and some lower it.
- We restrict the set of considered semposes to the better ones.

Semposes with the highest correlation in English.
This is also the restricted set of semposes used in English.

Tag	Min	Max	Avg
v	0.403	1.000	0.735
n.denot	0.189	1.000	0.728
adj.denot	0.264	0.964	0.720
n.pron.indef	0.224	1.000	0.639

Custom Tagger tagger

- We use sequence labeling algorithm to choose the t-lemma and sempos tag.
- The CzEng corpus served to train two taggers (for English and Czech).
- At each token, the tagger use word form, surface lemma and morphological tag of the current and previous two tokens.
- Tagger chooses sempos from all sempos tags which were seen in corpus with the given morphological tags.
- The t-lemma is often the same as the surface lemma, but it could also be surface lemma with an auxiliary word (*kick off*, *smát se*). The tagger can also choose such t-lemma if the auxiliary word is present in the sentence.
- The overall accuracy on CzEng e-test:
  - 97.9 % for English
  - 94.9 % for Czech

Tunable Metric Task

- We optimized towards linear combination (equal weights) of BLEU and Approx + Cap-micro.
- BLEU chooses sentences with correct morfology and word order, while SemPOS prefers sentences with correctly translated content words.

Our final result heavily depends on the interpretation of human rankings. Out of 8, we are:

≥ others	> others
the fifth	the first

Results

Tested on newstest2008, test2008, newstest2009, newsyscombtest2010.

English as a target language

Approximation	Overlapping	Min	Max	Avg
approx	cap-micro	0.409	1.000	0.804
orig	cap-macro	0.536	1.000	0.801
approx	cap-macro	0.420	1.000	0.799
tagger	cap-micro	0.409	1.000	0.790
orig	cap-micro	0.391	1.000	0.784
approx+cap-micro and BLEU		0.374	1.000	0.754
tagger	cap-macro	0.118	1.000	0.669
BLEU		-0.143	1.000	0.628

Czech as a target language

Approximation	Overlapping	Min	Max	Avg
approx-restr	cap-macro	0.400	0.800	0.608
tagger	cap-macro	0.143	0.800	0.428
orig	cap-macro	0.143	0.800	0.423
approx-restr	cap-micro	0.086	0.769	0.413
tagger	cap-micro	0.086	0.769	0.413
orig	cap-micro	0.086	0.741	0.406
approx	cap-micro	0.086	0.734	0.354
approx+cap-micro and BLEU		0.086	0.676	0.340
approx	cap-macro	0.086	0.469	0.338
BLEU		0.029	0.490	0.279

Overlapping performance

	Average rank in our experiments	
Overlapping	in English	in Czech
boost-micro	12	13
cap-macro	6.6	5
cap-micro	5.4	6

Boost-micro is not suitable for sempos-based metrics.

Full text, acknowledgement and the list of references in the proceedings of EMNLP 2011 Workshop on Statistical Machine Translation (WMT11).