Conservatism and Over-conservatism in Grammatical Error Correction

Leshem Choshen & Omri Abend

Hebrew University Jerusalem Israel

July 17 2017

Overview

Plan

the task

- Input: a text which is perhaps ungramatical
  - Focus learner language (LL)
- Output: a grammatical text saying the same meaning/content.

Example: However , there are both sides of stories

The task

- Input: a text which is perhaps ~~ungramatical~~ ungrammatical
  - Focus learner language (LL)
- Output: a grammatical text ~~saying~~ conveying the same meaning/content.

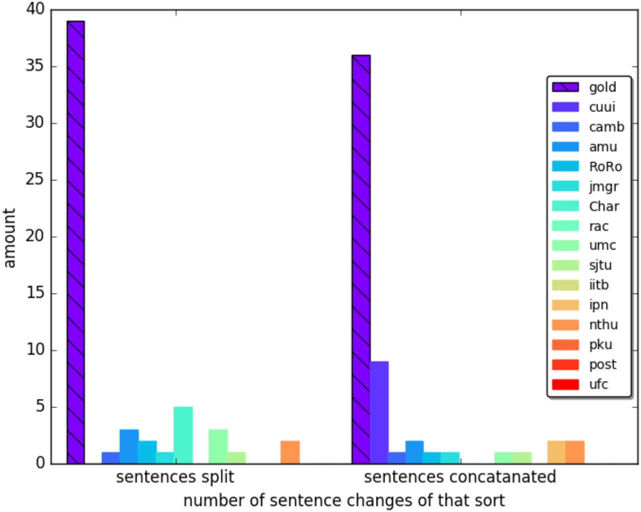
Example: However , there are ~~both sides of stories~~ $\rightarrow$
However , there are two sides to every story.

Plan

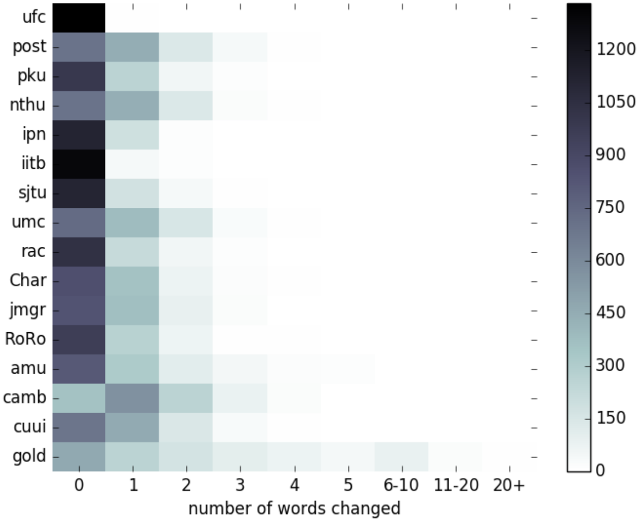
Conservatism? Over-conservatism?

It is a virtue to avoid bad corrections,
but the goal is still to correct...

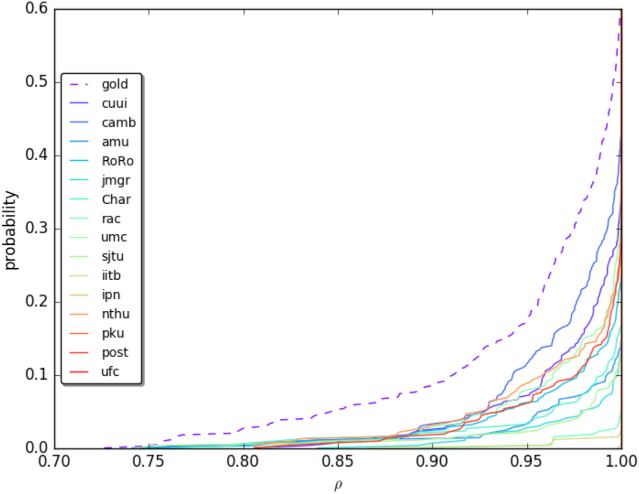
Current systems hardly change *sentence boundaries*



Current systems hardly change *words*



Current systems hardly change *word order*



Plan

The task

Over conservatism

Reference based measures - (RBM)s

- Background and motivation
- Corrections as distribution
- RBMs under estimation as a function of M

Reference-less semantic measure

What exists

Several Evaluation measures were suggested based on a source and a set of references.

F-score M^2 GLUE I-measure

To Train and validate 1 reference per source sentence.

Corrections as distribution

- Each sentence x has a set of valid corrections $correct_x$
- $\mathcal{D}_x$ a distribution of human corrections
- For testing - $Y \sim \mathcal{D}_x^M$ a sample of M references
- $P_{coverage}$ - $P_{y \sim \mathcal{D}_x}(y \in Y)$

Analytical worries

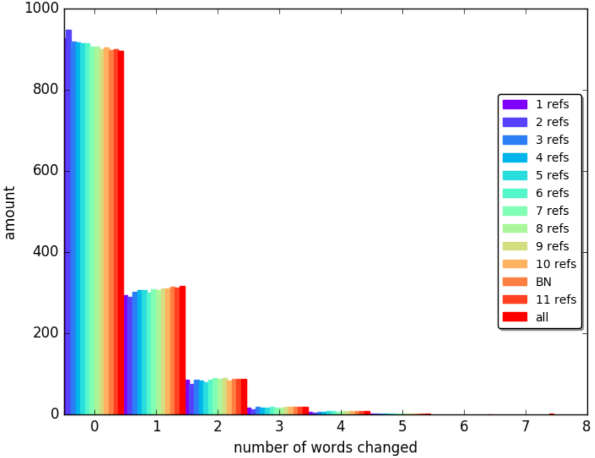
If a system detected a mistake it is incentivized to correct if

$$p_{correct} \cdot p_{coverage} > 1 - p_{detect}$$

If there is α punishment for wrong corrections

$$p_{correct} \cdot p_{coverage} - \left(1 - p_{correct} \cdot p_{coverage}\right)\alpha > 1 - p_{detect}$$

Empirical confirmation



Estimating $\mathcal{D}$

To estimate $\mathcal{D}$ we use UnseenEst. It estimates the histogram minimizing earthmover distance.

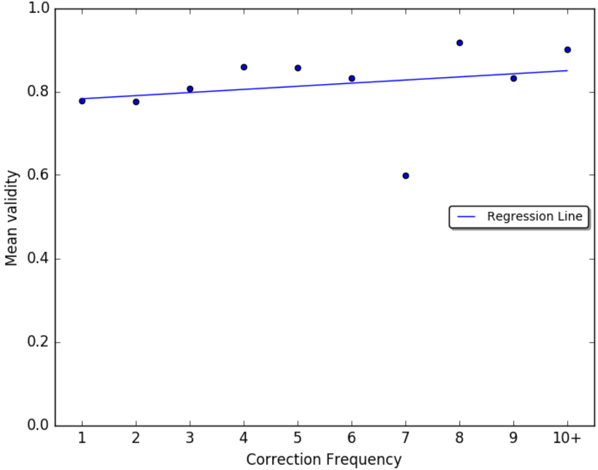
Findings

	Frequency Threshold (γ)			
	0	0.001	0.01	0.1
Variants	1351.24	74.34	8.72	1.35
Mass	1	0.75	0.58	0.37

dists

Yet more findings

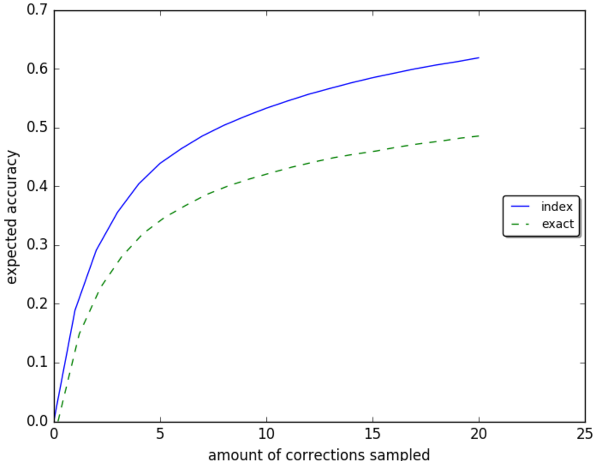
Rare corrections still count



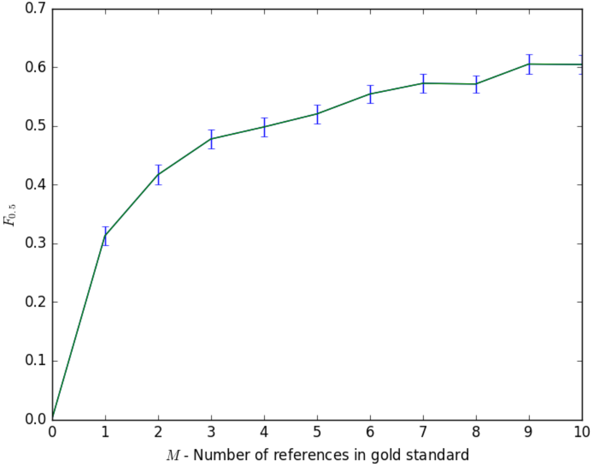
Accuracy - analysis

Given a perfect corrector, how well will it do?

$\frac{1}{N}\sum_{i=1}^{N} P_{Y\sim\mathcal{D}_i^M, y\sim\mathcal{D}_i}\left(y \in Y\right)$



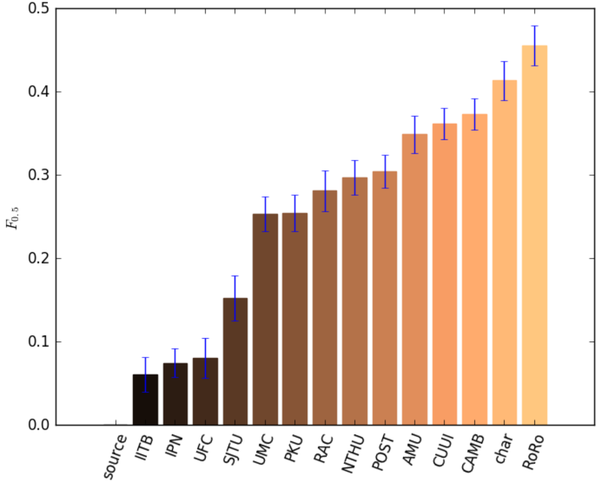
F-score - empirical



Note: with repetition it is more or less the same

Significance

human vs. machine



Plan

Reference-less evaluation

Input: Corrected sentences and Source sentences ~~and references in the form of sentences~~.
Output: A score, but which?!

Reference-less evaluation

Compare the source and the reference

- Suggestion: compare grammar annotations
- Grammar is ill defined with ungrammatical text
- Some define grammar on ungrammatical text as reference some as source

Reference-less evaluation

Combine two measures (worked for MT)

1. faithfulness – semantic similarity of the correction and the source. [1]
2. grammaticality – error detection over the source [2]

[1] Leshem Choshen and Omri Abend. "Conservatism and Over-conservatism in Grammatical Error Correction" - this work

[2] Napoles Courtney, Keisuke Sakaguchi, and Joel Tetreault. "There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction." arXiv preprint arXiv:1610.02124 (2016).

UCCA

- Semantic annotation scheme that builds on typological and cognitive linguistic theories
- Provides a coarse-grained, cross-linguistically applicable representation
- Structures are DAGS, words are leaves
- Text is a collection of *Scenes* and relations between them

Measures

- IAA - percentage of Nodes with same label and leaves
- UCCASim - percentage of Nodes with same label and most matched leaves
- Top down - size of the biggest cut
- Token - Consider only main entities
- (Labeled) Tree edit - tree distance when ordered by tokens alignment

distances table

LL hypotheses

- LL can be annotated using UCCA
- Corrections change grammar, not semantics

all annotation

Corrections preserve meaning

	UCCASIM			DISTSIM	
	s→r	r→s	Avg	A+D	Scene
Different	0.85	0.83	0.84	0.96	0.93
Same	0.92	0.91	0.92	0.97	0.96
IAA	0.85	0.81	0.83	-	-
SAR15	-	-	-	0.95	0.96

Works also automatically (TUPA parser)

	UCCASim		
	s→r	r→s	Avg
TUPA	0.7	0.7	0.7
Different	0.85	0.83	0.84

Any more questions?

Plan

Motivations

Data

Grammatical error correction approaches

- Classifier
- Machine Translation

Evaluation

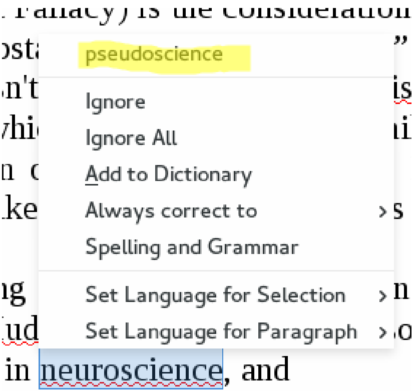
- Naive
- M^2-scorer
- I-measure
- GLEU
- reference-less

Motivation - Natural Language Processing view

Spelling correction is a solved problem, this is the next step.

Motivation - Natural Language Processing view

Spelling correction is a solved problem, this is the next step.



Motivation - Practice

English is a second language for the majority of English speakers. It can be used to enhance learning, as a tool (e.g. the green line in Word) etc.

Motivation - Computational Linguistics

Understanding grammatical errors and the way they can be corrected may lead to better understanding of innate processes and language behaviour.

- what errors do people do? why?
- what do we need to know in order to correct a language?
- what mistakes people will never do?
- Does learners' languages differ from native languages?

Plan

What data is there?

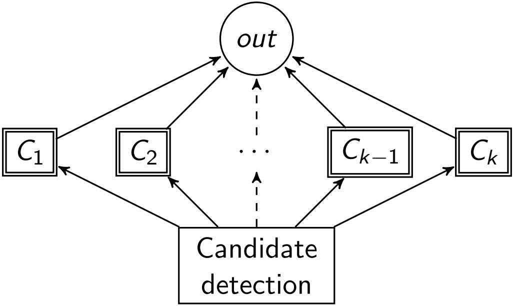
Field research and linguistic evidence
Two types of corpora:

- native or learner language corpora
  - large
  - cheap
- parallel corpora
  - both learner language and their corrections
  - corrections tend to be in the form of edits

Plan

Classifier based

- Different types of errors are chosen (e.g. Noun number errors)
- For each a set of possible corrections are chosen (e.g. $\{s, \emptyset\}$)
- A Classifier is built for each error type
- These are combined with rule based components (e.g. add e before the s when...)



Classifier based - pro con

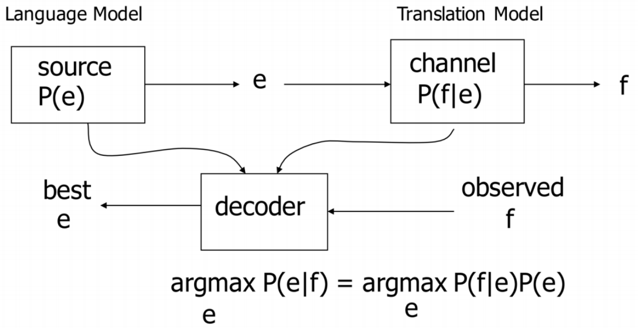
- Can only correct the chosen errors
- Complex mistakes and interleaving mistakes can not be handled properly
- Can generalize to similar problems with unseen words
- Useful in an unsupervised scenario

Machine Translation - motivation

A learner language is a consistent language, we can learn how to translate from it to the proper language.

Machine Translation - main idea

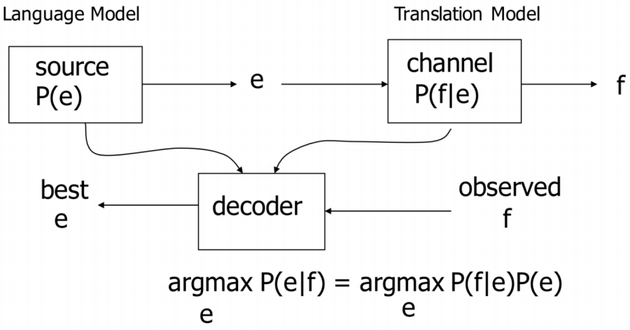
The main idea behind MT is the noisy channel.



Machine Translation - components

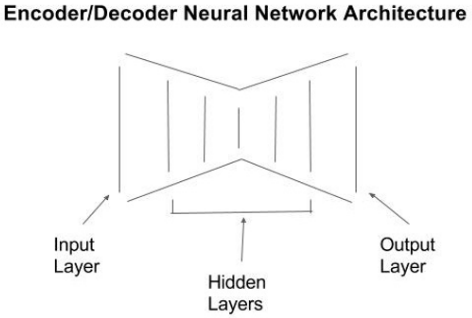
Language model – a model assigning a probability for a sentence to appear in the language $p(e)$

Translation model – uses a parallel corpus to assign probabilities to $p(f|e)$



And (of course) Neural Networks

Standard neural machine translation methods has started to immigrate to grammatical error correction too.[3][4]



[3] Chollampatt, Shamil, Kaveh Taghipour, and Hwee Tou Ng. "Neural network translation models for grammatical error correction." arXiv preprint arXiv:1606.00189 (2016).

[4] Yuan, Zheng, and Ted Briscoe. "Grammatical error correction using neural machine translation." Proceedings of NAACL-HLT. 2016.

MT - pro con

- Can correct the various errors
- Complex mistakes and interleaving mistakes can be handled properly
- Have problems generalizing to similar problems with unseen words (or phrases)
- Less useful in an unsupervised scenario

hybrid

Overall MT is good for many errors but the classifiers are better on the specific classes of errors chosen.
A pipeline starting with classifiers and applying MT over the results. This approach was shown to get the benefits of both models.[5]

[5] Alla Rozovskaya, and Dan Roth. "Grammatical error correction: Machine translation and classifiers." Urbana 51 (2016): 61820.

Plan

Motivations

Data

Grammatical error correction approaches
- Classifier
- Machine Translation

Evaluation
- Naive
- M^2-scorer
- I-measure
- GLEU
- reference-less

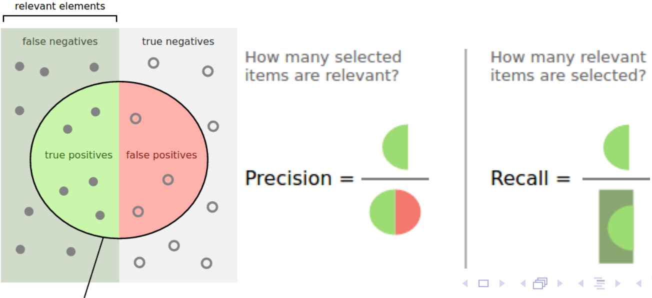
Edit F-score

Input: Corrected phrase-edits, references in the form of gold phrase edits.[6]
Output: phrase edit F-score Back to RBMs.

[6] Dale, Robert, and Adam Kilgarriff. "Helping our own: The HOO 2011 pilot shared task." Proceedings of the 13th European Workshop on Natural Language Generation. Association for Computational Linguistics, 2011.

Phrase edits F-score

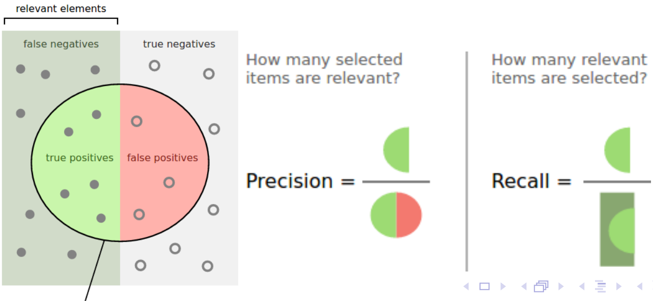
A correction is True iff the same *edit* is found in a reference.
For each sentence the best matching reference is used.



Phrase edits F-score

No-correction should be preferred over wrong correction, thus $F_{0.5}$, emphasizing precision, is used

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$



Edit F-score

But we do not expect correctors to actually mark what was the edit. Especially not in the same way humans "do".
$\{$a $\rightarrow \emptyset$ $\}$ or $\{$a words $\rightarrow$ *words*$\}$ Back to RBMs.

M^2-scorer

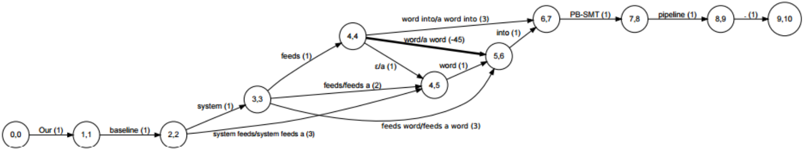
Input: Corrected sentences, Source sentences and references in the form of gold phrase edits.[7]
Output: phrase edit F-score Back to RBMs.

[7] Dahlmeier, Daniel, and Hwee Tou Ng. "Better evaluation for grammatical error correction." Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012.

M^2-scorer – corrections are sentences, not edits

An up to n words edit distance is computed dynamically, and a lattice is made. A negative value is assigned to every edit that is shown in the reference.

		Our	baseline	system	feeds	a	word	into	PB-SMT	pipeline	.
	0	1	2	3	4	5	6	7	8	9	10
Our	1	0	1	2	3	4	5	6	7	8	9
baseline	2	1	0	1	2	3	4	5	6	7	8
system	3	2	1	0	1	2	3	4	5	6	7
feeds	4	3	2	1	0	1	2	3	4	5	6
word	5	4	3	2	1	1	1	2	3	4	5
into	6	5	4	3	2	2	2	1	2	3	4
PB-SMT	7	6	5	4	3	3	3	2	1	2	3
pipeline	8	7	6	5	4	4	4	3	2	1	2
.	9	8	7	6	5	5	5	4	3	2	1



Back to RBMs.

I-measure

Input: Corrected sentences, Source sentences and references in the form of gold word edits.[8]
Output: Token-level weighted accuracy Back to RBMs.

[8] Felice, Mariano, and Ted Briscoe. "Towards a standard evaluation method for grammatical error detection and correction." HLT-NAACL. 2015.

I-measure

- Phrase edit choices are prone to errors (partial match, lack of TN count)

$$W_{acc} = \frac{w \cdot TP + TN}{w \cdot (TP + TN)\ TN + FN - (w+1) \cdot \frac{FPN^{9}}{2}}$$

~~Back to RBMs .~~

[9] places where correction, source and reference all differ

I-measure

- Phrase edit choices are prone to errors (partial match, lack of TN count)
- Use correction-source-reference word alignment maximizing Sum of Pairs

$$W_{acc} = \frac{w \cdot TP + TN}{w \cdot (TP + TN)\, TN + FN - (w+1) \cdot \frac{FPN[9]}{2}}$$

Back to RBMs .

[9] places where correction, source and reference all differ

I-measure

- Phrase edit choices are prone to errors (partial match, lack of TN count)
- Use correction-source-reference word alignment maximizing Sum of Pairs
- *F*-score ignores TN (choices not to correct)

$$W_{acc} = \frac{w \cdot TP + TN}{w \cdot (TP + TN)\ TN + FN - (w+1) \cdot \frac{FPN^{9}}{2}}$$

~~Back to RBMs .~~

[9] places where correction, source and reference all differ

I-measure

- Phrase edit choices are prone to errors (partial match, lack of TN count)
- Use correction-source-reference word alignment maximizing Sum of Pairs
- *F*-score ignores TN (choices not to correct)
- Use accuracy

$$W_{acc} = \frac{w \cdot TP + TN}{w \cdot (TP + TN)\, TN + FN - (w + 1) \cdot \frac{FPN}{2}}$$ [9]

Back to RBMs .

[9] places where correction, source and reference all differ

I-measure

- Phrase edit choices are prone to errors (partial match, lack of TN count)
- Use correction-source-reference word alignment maximizing Sum of Pairs
- *F*-score ignores TN (choices not to correct)
- Use accuracy
- Wrong corrections and no correction is the same for accuracy.

$$W_{acc} = \frac{w \cdot TP + TN}{w \cdot (TP + TN)\ TN + FN - (w+1) \cdot \frac{FPN^{9}}{2}}$$

Back to RBMs .

[9] places where correction, source and reference all differ

I-measure

- Phrase edit choices are prone to errors (partial match, lack of TN count)
- Use correction-source-reference word alignment maximizing Sum of Pairs
- F-score ignores TN (choices not to correct)
- Use accuracy
- Wrong corrections and no correction is the same for accuracy.
- Weighting is introduced

$$W_{acc} = \frac{w \cdot TP + TN}{w \cdot (TP + TN)\, TN + FN - (w+1) \cdot \frac{FPN^{9}}{2}}$$

~~Back to RBMs .~~

[9] places where correction, source and reference all differ

I-measure

- Phrase edit choices are prone to errors (partial match, lack of TN count)
- Use correction-source-reference word alignment maximizing Sum of Pairs
- F-score ignores TN (choices not to correct)
- Use accuracy
- Wrong corrections and no correction is the same for accuracy.
- Weighting is introduced
- Comparison with the source is made possible.

$$W_{acc} = \frac{w \cdot TP + TN}{w \cdot (TP + TN)\, TN + FN - (w+1) \cdot \frac{FPN^9}{2}}$$

Back to RBMs.

[9] places where correction, source and reference all differ

GLEU

Input: Corrected sentences, Source sentences and references in the form of sentences. [10]
Output: Weighted precision of n-gram (BLEU-like) Back to RBMs.

[10] Napoles, Courtney, et al. "Ground truth for grammatical error correction metrics." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Vol. 2. 2015.

GLEU - details

- Without weights the source has the second best score
- Extra weight to valid corrections (overlap with R not S)
- Penalty for no correction (overlap with S not R)

Back to RBMs.

GLEU - Human Rankings

Shown to achieve higher correlation with humans.

Metric	r	ρ
$\mathbf{GLEU_0}$	**0.542**	**0.555**
M^2	0.358	0.429
$GLEU_{0.1}$	0.200	0.412
I-measure	-0.051	-0.005
BLEU	-0.125	-0.225

Back to RBMs.

Findings

Only a handful of references are used, While even a short sentence tends to have hundreds of different valid corrections. It leads to under estimations.

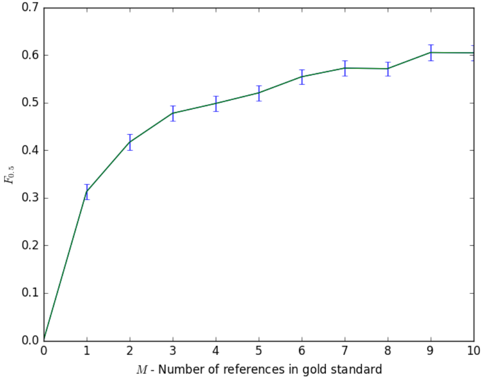


Figure: M^2 scores of a perfect corrector by the number of references

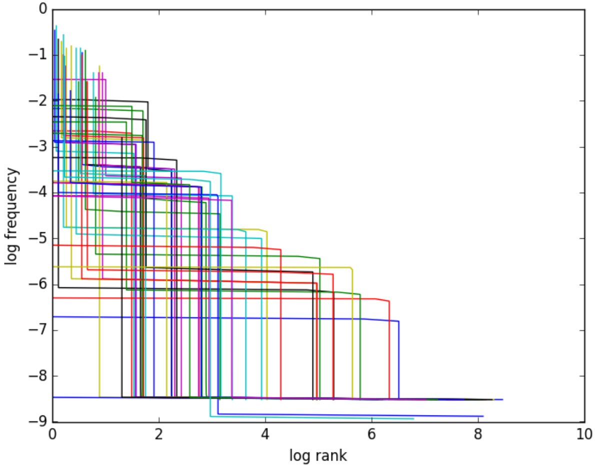
				Token		
	Tree	UCCASim	Top down	Bottom up	Top down	UCCASim
Different	324.57	0.83	0.75	0.74	0.72	0.83
Same	211.50	0.88	0.85	0.83	0.79	0.88
IAA	285.69	0.88	0.78	0.80	0.75	0.87

Back to details.

Annotator-id	NUCLE-id	type
1	2	corrected
2	2	corrected
1	2	learner
2	2	learner
1	3	corrected
2	3	corrected
1	3	learner
2	3	learner
1	5	corrected
2	5	corrected
1	5	learner
2	5	learner
1	6	learner
2	6	learner
2	7	corrected
2	7	learner
1	8	corrected
1	8	learner
1	10	corrected
1	10	learner

Back to UCCA.

Distributions



Back to Dists

Distributions

Back to Dists