Bleu: a Method for Automatic Evaluation of Machine Translation

Johanna Heininger, Eduard Schaf

University of Tübingen

November 13, 2014

Blau? Blue? Blüh?

What does this Bleu mean?!

- BLEU = BiLingual Evaluation Understudy
- automatic evaluation method for machine translation

Automatization

Why do we need an automatic evaluation method?

- human evaluations are extensive, but also expensive and time consuming
- developers need to monitor the effect of daily changes of their translation systems to optimize them
- benefit from a quick, language independent, human like and inexpensive automatic evaluation

Measurement

How to measure the performance:

- closeness of the MT compared to a professional human translation
- numerical value is compared to one or more reference human translations
  $\rightarrow$ this requires
  a) a numerical metric for translation closeness
  b) a corpus of high quality human reference translations
- baseline metric is selected from various weighting schemes (very high correlation with human judgements)

Intro

There are several things to consider when rating a translation:

- a source sentence has many "perfect" translations
- there is variation in word order and word choice
- even then humans can reliably distinguish good translations from bad ones

Example - Candidate 1

Candidate 1:

It is a guide to action which ensures that the military always obeys the commands of the party.

References

- Reference 1:
  It is a guide to action that ensures that the military will forever heed Party commands.
- Reference 2:
  It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference 3:
  It is the practical guide for the army always to heed the directions of the party.

Example - Candidate 2

Candidate 2:

It is to insure the troops forever hearing the activity guidebook that party direct.

References

- Reference 1:
  It is a guide to action that ensures that the military will forever heed Party commands.
- Reference 2:
  It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference 3:
  It is the practical guide for the army always to heed the directions of the party.

BLEU's task

BLEU accomplishes this task:

- Compare candidate n-grams with reference n-grams and count the matches
- Matches are independent from the position
- The more matches, the better the candidate

Precision Example

Consider the following example:

Example 2:

Candidate: the the the the the the the

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Unigram precision: 7/7 (1)

This leads to the conclusion that a reference word should be exhausted after a matching candidate word is found.

Modified n-gram precision

The modified n-gram precision (with any n) conquers this problem:

- collect all candidate n-gram counts and their corresponding maximum reference counts
- clip the candidate counts by their corresponding reference maximum value
- sum the candidate counts
- divide the candidate counts by the total number of n-gram candidates

Modified Precision Example

Lets recall the previous example, clipped words are underlined:

Example 2:

Candidate: the the the the the the the

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram precision: 2/7

Now we see that each match is only counted once. By applying the modified n-gram precision we get a score of 0, because no candidate bigram appears in the references.

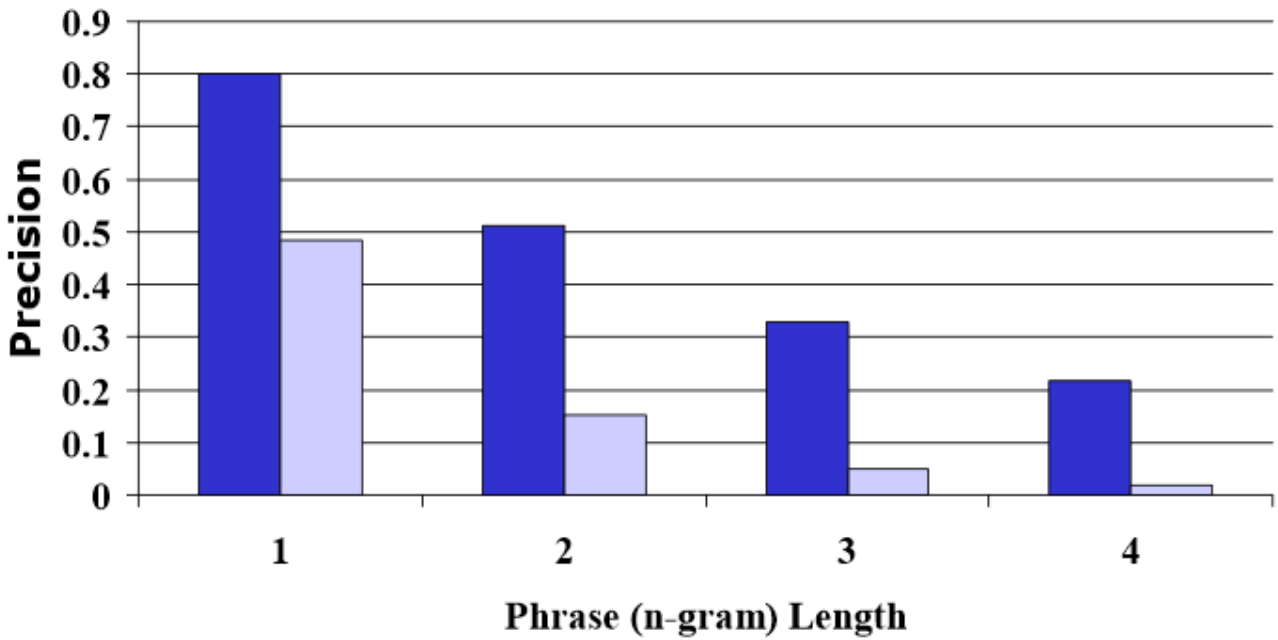
Modified n-gram precision Conclusion

- translations using the same words (unigrams) as in the references tends to satisfy adequacy
- The longer n-gram matches account for fluency

Human - machine distinction results

To check that modified n-gram precision can distinguish a good translation from a bad one a computation was performed:

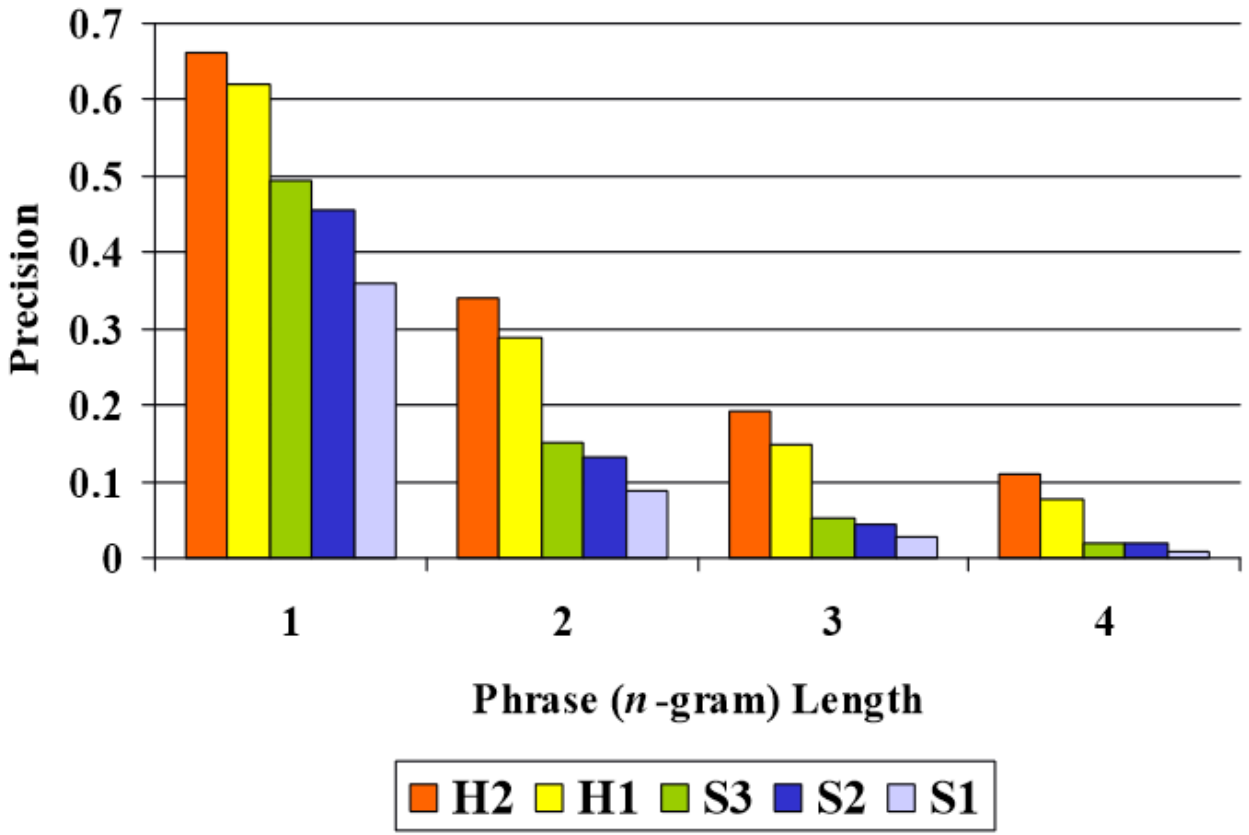
Figure 1: Distinguishing Human from Machine



(4 reference translations for each of 127 source sentences)

Human - machine distinction II results

Figure 2: Machine and Human Translations



Sentence Length

The sentence length needs to be considered as well because a candidate translation should neither be too long nor too short:

- spurious words in the candidate not appearing in the reference are penalized
- rewards using a word as many times as warranted
- penalizes using a word more times than it occurs in the references
- but: fails to enforce the proper translation length

Translation length problem

For the following short candidate we obtain a modified unigram precision of 2/2 and modified bigram precision of 1/1:

Candidate "of the"

- Reference 1:
  It is a guide to action that ensures that the military will forever heed Party commands.
- Reference 2:
  It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference 3:
  It is the practical guide for the army always to heed the directions of the party.

Translation length problem

Traditionally length related problems are handled with recall, but this won't work here:

- BLEU considers multiple reference translations
- each can have a different word choice
- a good candidate translation will use only one choice

Recall problem

Consider the following:

"I always invariably perpetually do." vs "I always do."

- Reference 1: I always do.
- Reference 2: I invariably do.
- Reference 3: I perpetually do.

The problem here is:

- The first candidate recalls more reference words, but the translation is worse
- naive recall over the set of all reference words is not a good measure
- computing recall on concepts rather than words is too complicated

Sentence brevity penalty

However, to overcome the translation length problem, a sentence brevity penalty was introduced:

- penalizes candidates shorter than their references
- the brevity penalty is a multiplicative factor modifying the overall BLEU score
- the penalty is 1 when candidate and any reference length are the same

BLEU formulas

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases}$$

$$BLEU = BP * \exp\left(\sum_{N}^{n=1} w_n \log p_n\right)$$

c = length of the candidate translation
r = effective reference corpus length
p_n = geometric average of the modified n-gram precisions
N = 4
w_n = 1/N

Evaluation

- BLEU metric ranges from 0 to 1 (1 = translation identical to reference)
- comparisons on evaluations with different number of references cannot be made
- test corpus of about 500 sentences (40 general news stories)
- human translator scored
  0.3468 against four references
  0.2571 against two references

Scores

BLEU scores of the 5 systems against 2 references:

Table 1: BLEU on 500 sentences

S1	S2	S3	H1	H2
0.0527	0.0829	0.0930	0.1934	0.2571

The test corpus was divided into 20 blocks of 25 sentences each, BLEU score was computed on each block:

Table 2: Paired t-statistics on 20 blocks

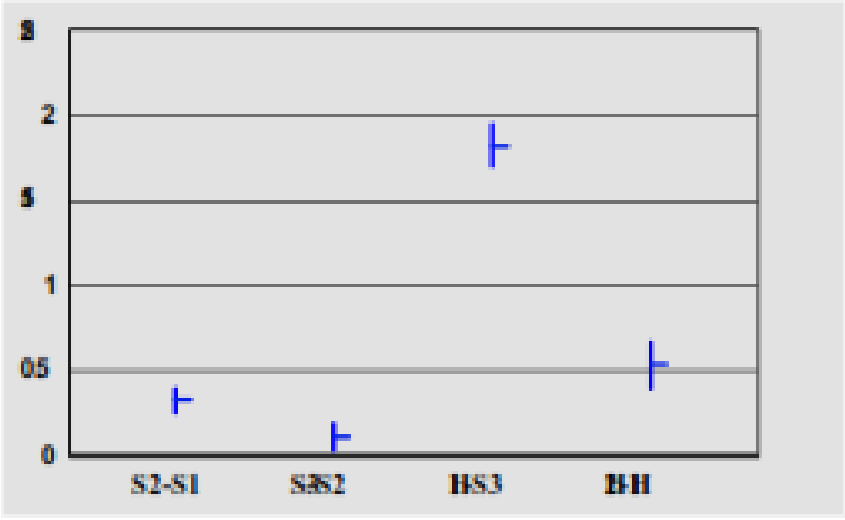
	S1	S2	S3	H1	H2
Mean	0.051	0.081	0.090	0.192	0.256
StdDev	0.017	0.025	0.020	0.030	0.039
t	—	6	3.4	24	11

Evaluation

- 10 native speaker of English, monolingual group
- 10 native speaker of Chinese, who lived in the US for years, bilingual group
- tested on a Chinese sentence subset, total of 250 pairs of Chinese source sentences and English translations
- rated from 1 (very bad) to 5 (very good) based on readability and fluency
- each judge's rating for a sentence was compared across systems

Monolingual

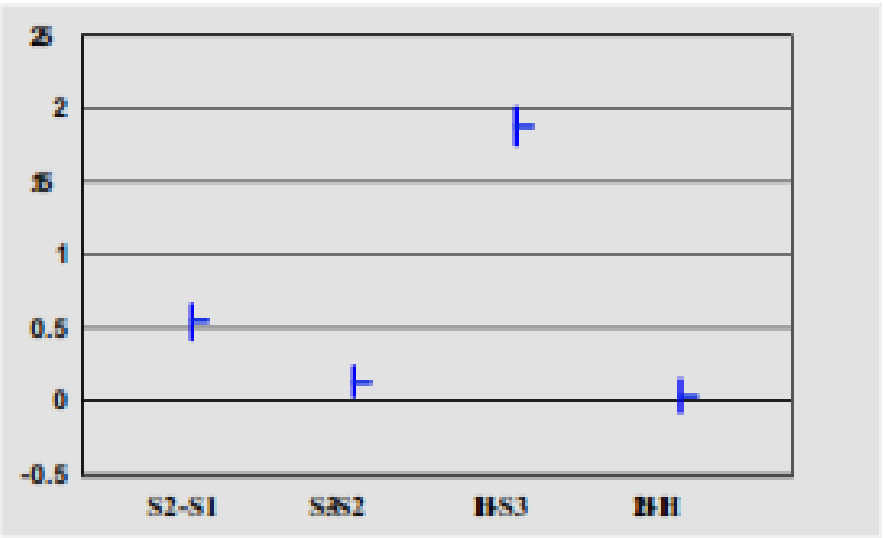
Figure 3: Monolingual Judgments - pairwise differential comparison



+95%	0.400	0.194	1.945	0.670
-95%	0.252	0.034	1.705	0.400
monolingual	0.326	0.114	1.825	0.535

Bilingual

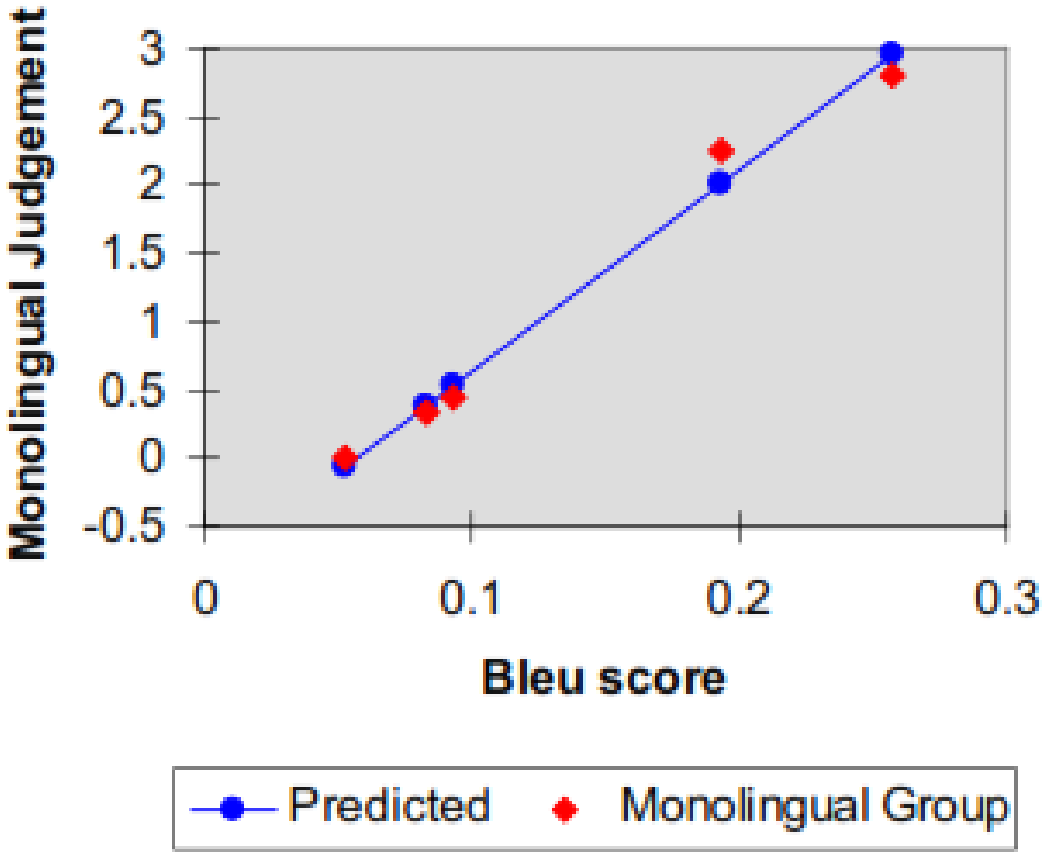
Figure 4: Bilingual Judgments - pairwise differential comparison



+95%	0.667	0.238	2.007	0.145
-95%	0.435	0.042	1.759	-0.069
bilingual	0.551	0.140	1.883	0.038

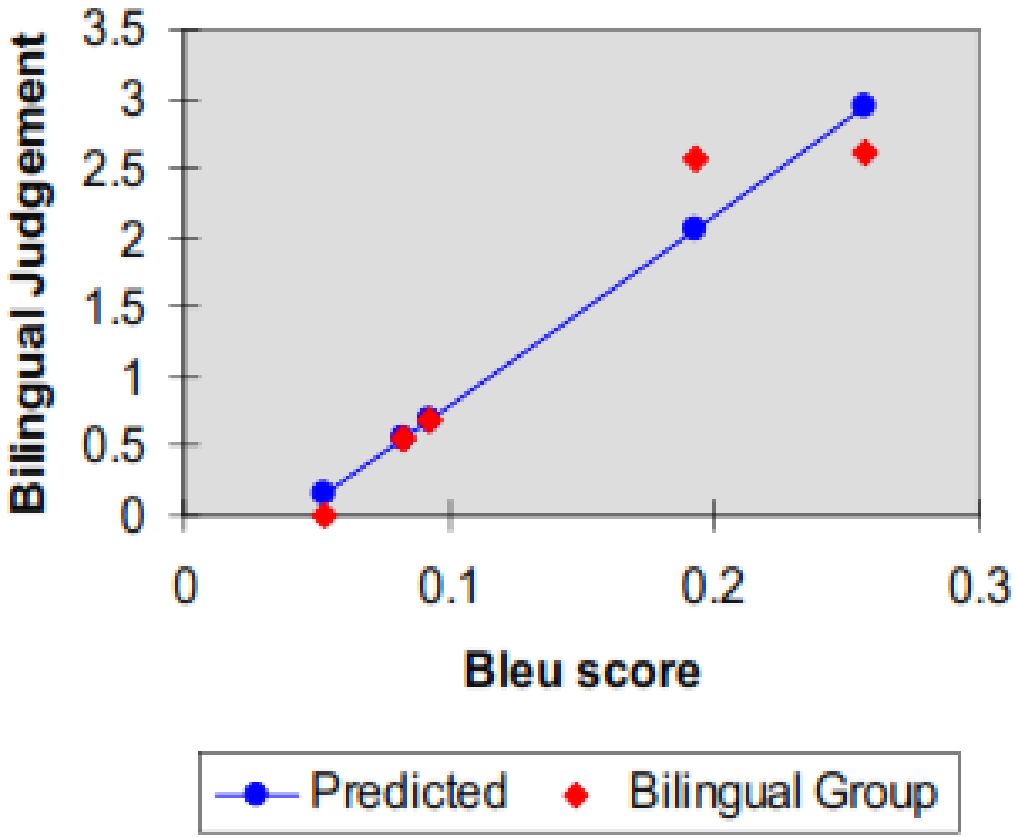
Bleu vs. Human - Monolingual

Figure 5: BLEU predicts Monolingual Judgments



Bleu vs. Human - Bilingual

Figure 6: BLEU predicts Bilingual Judgments

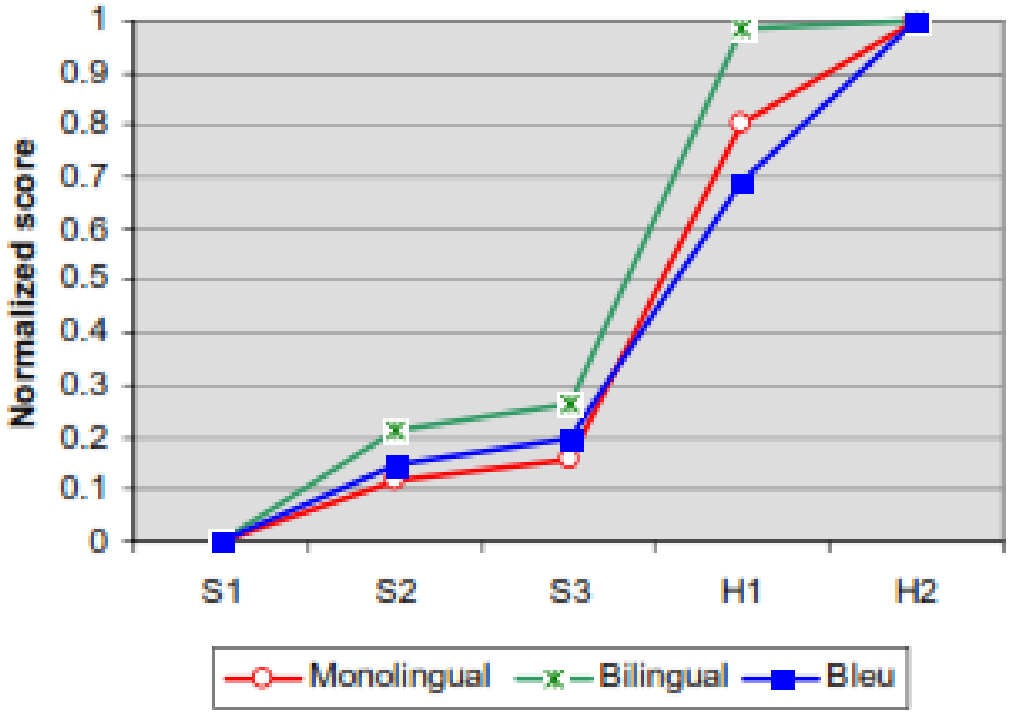


Bleu vs Human

- high correlation (.99 and .96) is an indicator for BLEU tracking human judgments well
- can distinguish between S2 and S3 even though they are quite close to each other

Bleu vs. Human

Figure 7: BLEU vs Bilingual and Monolingual Judgments



Conclusion

- the simplicity of BLEU is a very appealing feature for research and MT systems
- a more highly parameterized form of the brevity penalty, modified n-gram precision or precision averaging expressions could lead to a more accurate estimator of translation quality
- other language pairs could be evaluated
- the n-gram similarity of a candidate to a set of references could be extended to the evaluation of language generation and summarization systems