Bllip: An Improved Evaluation Metric for Machine Translation

Michael Pozar **Eugene Charniak**

Brown University Department of Computer Science
Providence, RI, 02912

Abstract

In this paper we present a new automatic scoring method for machine translations. Like the now-traditional BLEU score it maps a proposed translation and a set of reference translations to a real number. This number is intended to reflect the quality of the proposed translation. We present some experiments that indicate that this new metric, the Bllip score (Brown Laboratory for Linguistic Information Processing) correlates better with human judgment than does BLEU.

1 Introduction

Machine Translation systems use automatic evaluation methods to assess translation quality. The problem is to determine a score for a candidate translation, given a set of human-generated reference translations, which indicates how good the candidate translation is. The effectiveness of a metric is how well the metric correlates with human judgment when comparing the candidate to the references. Recently, it has been contended by many that current evaluation metrics are inadequate, and that much improvement on these metrics is possible (Gimenez and Amigo, 2006). Many automatic translators use the BLEU metric, or similar algorithms, which count n-gram matches between the candidate sentence and a set of reference sentences. Such systems, though fast and very simple, do not make use of linguistic information in any way, and thus we believe significant improvements to be possible in this area.

Rather than comparing n-grams, the Bllip metric constructs dependency trees for each candidate, and compares these to those of the references. The intention is to capture more of the structure of the sentence and give a more accurate depiction of the candidate translation quality.

Below, we outline the methods used for the Bllip score algorithm, and then compare results of our metric to our implementation of the BLEU score algorithm as outlined in Papineni et al (2001).

2 The Bllip Metric

There are, of course, many correct translations of a given sentence. There are at least two dimensions along which these translations may differ, two of the most influential being word choice and word order. The Bllip metric attacks the latter.

A problem with comparing n-grams is that sentences in which word ordering has been interchanged, without significantly changing the meaning, will result in fewer matches. Such changes, if they do not change the meaning, usually will not change the phrasal structure of the sentence. Thus, a major advantage of comparing dependency sets is that in these cases, we are able to capture the fact that these sentences do not differ as much in meaning as n-gram matching would indicate. Example 2, below, and the results of our comparison between Bllip and BLEU support this hypothesis.

2.1 Constructing Dependency Sets

Given a sentence, we wish to construct its dependency tree. First, we parse the sentence using the Charniak parser (Charniak, 2000).

The parser can be configured to also report the lexical head of each constituent. We extract the dependency relation from this in the usual fashion. A dependency is an ordered pair of words. Given a parse tree, the dependency set is the set of all lexical head pairs (w1, w2) for which:

- $w1 \neq w2$
- w1 is the lexical head of a constituent which immediately dominates a constituent for which w2 is the lexical head.

An example will help to clarify this:

Example 1:

Sentence: He has a ball

See Figure 1 for the parse tree and the corresponding dependency tree. The subscript of each node in the parse tree is the lexical head of that constituent.

Looking at the parse tree, we can read off all of the dependencies and get the dependency set (actually, it may be called a dependency "bag," since we allow duplicates, but we will take the convention of calling these "sets").

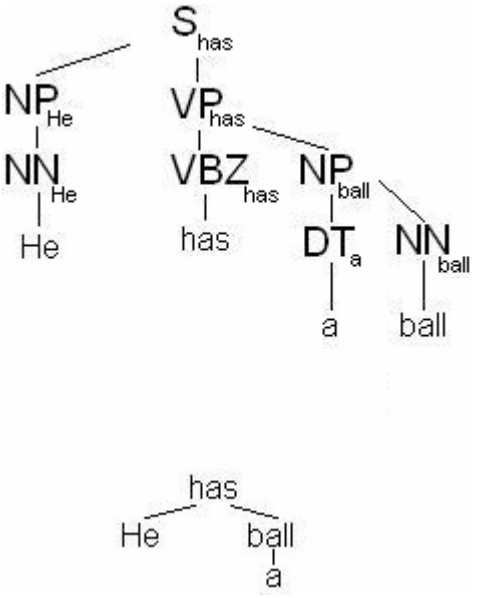


Figure 1: The parse and dependency trees for the sentence "He has a ball."

The dependencies are thus:

(has, He)
(has, ball)
(ball, a)

2.2 Scoring the Sentence

Once the dependency sets have been constructed for the candidate and reference sentences, we wish to somehow compare them. Our scoring system is very straightforward. We only compare the candidate sentence to one reference at a time. We use precision-recall on the dependencies. That is, for each dependency (A, B) in the candidate, we check for its existence in the reference, computing precision. Likewise, for each dependency in the reference, we check for the existence of it in the candidate. The score for our candidate against a reference sentences is then the total number of matches found divided by the number of possible matches (if each dependency in one sentence existed in the other). Then, we compose these scores together to come up with the final BLLIP score. A number of different composition techniques were tried, such as taking the best score, and the average score. The results of these methods are described later on.

Consider the following example, chosen to demonstrate the strong points of Bllip.

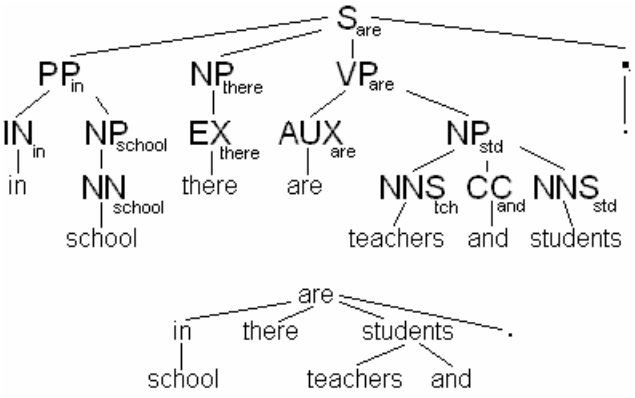
Example 2:

Candidate sentence:
In school there are teachers and students.

Reference sentence:
There are students and teachers in school.

Clearly, these two sentences are identical in their meaning. Any translator outputting one of these when our known, reference translation is the other, is doing his/her/its job.

Below are the parse trees of each sentence, followed by the dependency trees:



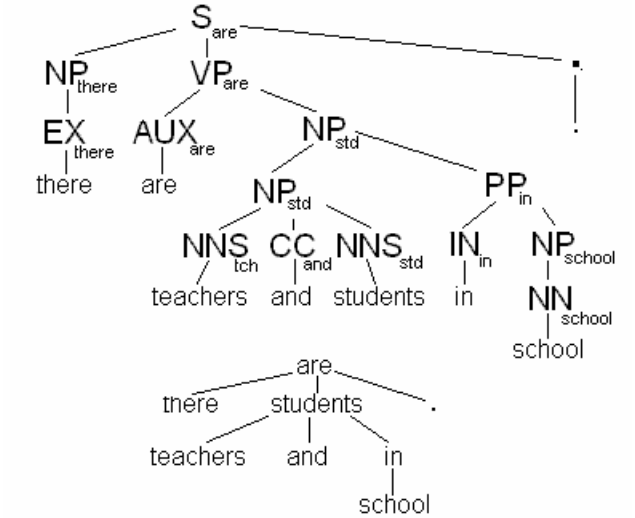


Figure 2: The parse and dependency trees for the sentences in Example 2.

First, examine how Bllip would analyze this example:

BLLIP

Candidate Dependencies:
(are, in)
(are, there)
(are, students)
(are, .)
(in, school)
(students, teachers)
(students, and)

Reference Dependencies:
(are, there)
(are, students)
(are, .)
(students, teachers)
(students, and)
(students, in)
(in, school)

87.5% matches in the dependency sets.

Score: 0.875.

Now, examine how BLEU would analyze this example:

BLEU

BLEU (Papineni et al, 2001) performs modified precision on n-grams of length 1, 2, 3, and 4. It checks to see if an n-gram from the candidate translation exists in **any** of the reference translations. Table 1 shows the number of matches and the number of possibilities for each n-gram length for Example 2.

Table 1:

	Matches / possibilities
Unigrams	7 / 7
Bigrams	2 / 6
Trigrams	0 / 5
4-grams	0 / 4

When measuring n-gram matches for a single sentence rather than across an entire corpus, it is appropriate to take the arithmetic mean of these ratios. If we do this for the above values, we get:

Score: 0.333

In such a short sentence (and with only one reference), it is not unusual to expect a low number of tri- and 4-gram matches. Still, the example emphasizes the way in which the BLLIP metric captures the important phrasal structure of the sentences and recognizes these similarities, while the BLEU score fails to pick up on this at all. See later examples in the results section for similar examples.

3 Results

We ran the Bllip scorer on 1024 and compared the scores to our implementation of BLEU using n=4 for the maximum length for n-grams. Again, to compute the BLEU score for a single candidate sentence, we are using the arithmetic mean of the ratios of n-gram matches to possibilities. We examined the correlation between the two methods. For the most part, candidate sentences which BLEU scored highly were also scored highly by Bllip, and likewise for low-scoring candidates. This is a positive result, as BLEU is already a decent metric, so we would expect many sentences to be scored similarly. For analysis, we examined sentences in which one metric gave a score far (at least half a standard deviation) from the mean score of that metric, and the other metric scored it on the other side of the mean. About 5-10% of sentences fell into this category. Then, we studied the correlation between these scores, and human judgment of those sentences. We saw a much stronger correlation between the Bllip scores and human scores than between BLEU and human.

That is, when Bllip and BLEU disagree, Bllip tends to more accurately depict whether the candidate is a strong translation given the reference sentences. Below is the distribution of BLEU and Bllip scores.

BLEU metric:
Mean: 0.386
Standard Deviation: 0.164

Bllip metric:
Mean: 0.405
Standard Deviation: 0.354.

We use the z scores, which correspond to the number of standard deviations a sentence's score was from the mean, i.e.

$$z = (\text{score} - \text{mean}) / \text{s.d.}$$

In Appendix A, we list the sentences which fell into the above category, where one metric's score for the sentence is at least half a standard deviation from the mean, and the other's score is in the other direction – i.e., either BLEU is scoring high when Bllip is not, or vice versa.

The first 15 sentences are situations where BLEU scores far below the mean and Bllip scores around or above the mean, or where Bllip scores far above the mean and BLEU scores around or below the mean. It is fairly clear to any native English speaker that these sentences should be given high scores. For example:

Candidate: The results were published after the close of the market .
Reference: Results were published after market closure.

The BLEU score is significantly below average, and the Bllip score is above average.

The final 3 sentences of Appendix A are slightly different, and point to a weakness of Bllip. A closer look at Sentences 18-20 show that Bllip is scoring them low because the more important words in the sentence differ.
Sentence 20:
Candidate: The Friday events have been the worst of the worst .
Reference: Friday 's event was the worst of the worst .

Since the subject and verb of the sentences do not match ("have been" versus "was," and "Friday events" versus "Friday's event), the dependency sets end up being very different. BLEU is able to tolerate these mismatches to an extent, and ends up scoring fairly high simply due to the fact that "the worst of the worst" appears in each sentence. However, note that if the Reference were, instead "A B C D the worst of the worst .", this would still result in the same BLEU score. So, the question is one of how tolerable the difference between "Friday's event" and "The Friday events" is – should this sentence be scored highly or not?

Sentences 18-20 illustrate that Bllip does not attack the word choice issue, but only word order. If minor words are replaced by similar words, the Bllip score is hurt slightly, as is the BLEU score. However, when key words of a sentence (e.g., the subject or main verb of a sentence or phrase) are replaced by similar words, Bllip is hurt significantly, due to the fact that these words tend to be a part of many dependencies.

Overall, these results are very positive. In the majority of cases, BLEU and Bllip agree on how a candidate sentence should be scored. In the cases we examined where they disagree, Bllip correlates better with human judgment 17 out of 20 times.

As mentioned previously, we tried a few different methods of scoring the dependency tree comparisons. Averaging the comparisons between the candidate and each reference versus taking the best reference did not yield significantly different results, while taking the worst reference yielded poorer results.

Another possible improvement could be seeing if a candidate dependency relation exists in any of the references' sets (similar to the way BLEU checks for existence of an n-gram across all references), rather than simply scoring the candidate against each reference individually. However, this does not allow for an obvious method of recall. This is the same problem with attempting recall in n-gram matching – if we want all the references' n-grams to match n-grams in the candidate, lengthier, poorer candidates will be scored too highly. To borrow an example from Papineni et al (2001):

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

The first candidate recalls more words from the references, but is obviously a poorer translation than the second candidate.

This could presumably occur just as easily if the same technique were used with dependencies rather than n-grams – the candidates with lots of extra words or phrases get higher recall scores than other, more accurate candidates. Papineni et al (2001) correct this by introducing a Brevity Penalty, a deduction for candidate sentences that are significantly longer than their references. Perhaps a similar penalty could be investigated to make such a scheme work for dependency matching as well.

4 Remarks

There is much further improvement possible in the area of automatic evaluation of machine translation. We have outlined here a metric based on dependency tree comparison which outperforms current metrics. Furthermore, our method is very straightforward. Once parsed, sentences' dependency trees are easy to construct, and the scoring algorithm is very simple and direct. The main bulk of computing time is taken up by the parser. When being used in practice, it would make sense to pre-parse all the reference sentences, and construct all their dependency sets, since these will be used over and over.

More improvement is possible from this base, for example different methods of combining information from the references, or comparing more information within the sets. A major form of improvement would be to address the word choice issue, for example by putting in a tolerance for similar words appearing in the different sentences.

5 References

Chin-Yew Lin and Franz Josef Och, 2004. "Orange: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation." In *Proceedings of COLING*.

Dabbabdie, M., A. Hartley, M. King, K.J. Miller, W. M. El Hadi, A. Popescu-belis, F. Reeder & M. Vanni, 2002. "A Hands-On Study of the Reliability and Coherence of Evaluation Metrics." In M. King (ed.), *Machine Translation Evaluation – Human Evaluators Meet Automated Metrics, Workshop Proceedings, LREC'2002*, pp. 8--16.

Dekang Lin, 1998. "Dependency-based Evaluation of MINIPAR." In *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

Eugene Charniak, 2000. "A Maximum-Entropy-Inspired Parser." In *Proceedings of NAACL'00*, pages 132-139.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev, 2003. "Final Report of Johns Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation." Technical report, Johns Hopkins University.

Harper, Mary, Roark, Brian, Yang Liu, Robin Stewart, Matthew Lease, Matthew Snover, Izhak Shafran, Bonnie J. Dorr, John Hale, Anna Krasnyanskaya, and Lisa Yung, "Sparseval: Evaluation Metrics for Parsing Speech," to appear in Proceedings of the International conference on Language Resources and Evaluation, Genoa, Italy, 2006.

Jesus Gimenez and Enrique Amigo, 2006. "IQMT: A Framework for Automatic Machine Translation Evaluation." TALP Research Center, LSI Department.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2001. "BLEU: a Method for Automatic Evaluation of Machine Translation," IBM Research Report, rc22176. Technical report, IBM T.J. Watson Research Center.

Appendix A. Results

The first sentence is the candidate, and the second the reference. Listed below each is the computed BLEU score (the average of the ratios of matches to possibilities for each length) and the Bllip score, as well as these scores' corresponding z values.

Sentence 1
The profit per share , however , fell only by 2 % from 1.26 to 1.23 dollars per share since the company continued in its plan to buy back shares .
However , profit per share fell by just 2 % , from 1.26 to 1.23 dollars per share , because the company continued repurchasing stock as planned .
BLEU score: 0.356793, z = -0.169495, BLLIP score: 0.61096, z = 0.581849

Sentence 2
It therefore makes absolutely no sense that each market adopts different security precautions .
Therefore , it makes no sense to let each market adopt different safety measures .
BLEU score: 0.199176, z = -1.0705, BLLIP score: 0.483046, z = 0.220845

Sentence 3

The results were published after the close of the market .
Results were published after market closure .
BLEU score: 0.298169, z = -0.504612, BLLIP score: 0.569803, z = 0.465694

Sentence 4
This profit represents a return for Security Pacific on its assets of 0.89 % and a return on the original capital of 18.9 % .
For Security Pacific this profit represents 0.89 % return on assets and 18.9 % return on equity .
BLEU score: 0.264656, z = -0.696187, BLLIP score: 0.565685, z = 0.454074

Sentence 5
LIN closed on the federal market outside the stock exchange at a price of 104.75 dollars , that is , down 2.75 dollars less .
At the federal off-exchange market LIN closed at 104.75 dollars , down 2.75 dollars .
BLEU score: 0.246889, z = -0.797749, BLLIP score: 0.516398, z = 0.314972

Sentence 6
Fidelity Investments placed new ads in the papers yesterday , and created a new advertisement that was released today .
Yesterday Fidelity Investments placed new advertisements in papers and produced another new advertisement which emerged today .
BLEU score: 0.258273, z = -0.732673, BLLIP score: 0.433861, z = 0.0820337

Sentence 7
This time , the companies were prepared .
This time firms were ready .
BLEU score: 0.160714, z = -1.29036, BLLIP score: 0.57735, z = 0.486994

Sentence 8
The PaineWebber company was likewise able to react swiftly thanks to the decline of 1987 .
PaineWeber was also able to react promptly , thanks to the 1987 plunge .
BLEU score: 0.243006, z = -0.819946, BLLIP score: 0.467707, z = 0.177556

Sentence 9
The PaineWebber company even considered a more aggressive sales campaign with the recommendation of certain stocks .
PaineWeber even considered a more aggressive marketing campaign with specific stock recommendations .
BLEU score: 0.281486, z = -0.599977, BLLIP score: 0.470871, z = 0.186485

Sentence 10
Non-credit expenditures grew by only 4 % for the period under consideration .
Non-interest expenses grew by just 4 % in the reported period .
BLEU score: 0.176282, z = -1.20136, BLLIP score: 0.480384, z = 0.213334

Sentence 11
The task of improving the functioning of the market is still not completed , however .
However , the task of improving market operation is not finished .
BLEU score: 0.261195, z = -0.71597, BLLIP score: 0.433013, z = 0.0796398

Sentence 12
The above attitudes represent either a triumph of indifference , or politeness .
The aforementioned views represent a triumph of lethargy or courtesy .
BLEU score: 0.199009, z = -1.07145, BLLIP score: 0.418121, z = 0.037612

Sentence 13
The Canadian government justified these measures on the basis of protection interests .

The Canadian government justified the measure with environmental concerns .
BLEU score: 0.248339, z = -0.789459, BLLIP score: 0.438529, z = 0.0952082

Sentence 14
Mrs. Hills yesterday stated that the arbitration panel had rejected this argument by the Canadian government .
Yesterday Ms. Hills stated the arbitration panel rejected this argument of the Canadian government .
BLEU score: 0.385075, z = -0.00782412, BLLIP score: 0.626224, z = 0.624929

Sentence 15
She stated that the Canadian restriction must be removed before these contracts are concluded .
She stated the Canadian restrictions must be lifted before such contracts are made .
BLEU score: 0.238095, z = -0.848017, BLLIP score: 0.414039, z = 0.0260925

Sentence 16
However , additional trading was stopped completely at 3.45 p.m. , since the futures markets had dropped by another 30 points , which represented the daily limit of price decline .
However , futures trading was completely stopped at 3-45 because futures markets fell by another 30 points which represents a daily limit for price decrease .
BLEU score: 0.229665, z = -0.89621, BLLIP score: 0.457905, z = 0.149893

Sentence 17
In the second game , played on a cold Sunday evening in this land of eternal fall , lots of home runs were hit by Terry Steinbach , the A's catcher .
During a cold Sunday evening in a country of permanent fall when the second game took place a lot of runs were made by A team catcher Terry Steinbach .
BLEU score: 0.255716, z = -0.747291, BLLIP score: 0.419573, z = 0.0417104

Sentence 18
Non-interest income increased by 16 % to 496 million dollars .
Non-interest expenses grew by 16 % , i.e. to 496 million dollars .
BLEU score: 0.528157, z = 0.810082, BLLIP score: 0.0836242, z = -0.906417

Sentence 19
However , the markets can work more or less effectively .
But the markets can function more or less effectively .
BLEU score: 0.505429, z = 0.680165, BLLIP score: 0.381385, z = -0.0660656

Sentence 20
The Friday events have been the worst of the worst .
Friday 's event was the worst of the worst .
BLEU score: 0.511679, z = 0.715892, BLLIP score: 0.381385, z = -0.0660656