The Limits of N-Gram Translation Evaluation Metrics

Christopher Culy
FXPAL
3400 Hillview Ave, Bldg. 4
Palo Alto, CA 94304
culy@fxpal.com

Susanne Z. Riehemann
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
riehemann@ai.sri.com

Abstract

N-gram measures of translation quality, such as BLEU and the related NIST metric, are becoming increasingly important in machine translation, yet their behaviors are not fully understood. In this paper we examine the performance of these metrics on professional human translations into German of two literary genres, the Bible and *Tom Sawyer*. The most surprising result is that some machine translations outscore some professional human translations. In addition, it can be difficult to distinguish some other human translations from machine translations with only two reference translations; with four reference translations it is much easier. Our results lead us to conclude that much care must be taken in using n-gram measures in formal evaluations of machine translation quality, though they are still valuable as part of the iterative development cycle.

1 Introduction

1.1 Background

Machine translation evaluation is notoriously expensive, requiring multiple human translations and laborious manual judging. Thus, when the BLEU metric (Papineni et al. 2002), an automatically calculated, n-gram based metric, was proposed and shown to correlate well with human judgments of translation quality, it was seen (correctly) as a significant boon to MT evaluation. Human judges are not needed in the BLEU loop.

The authors of BLEU were careful to position it as an *aid* to evaluation. It still remains to be seen exactly how BLEU is best used, but already a variant of BLEU has been adopted by NIST for its MT evaluation effort (NIST Report, 2002). In addition many groups, including those in the DARPA Babylon speech to speech translation program, of which we are a part, are using BLEU or other n-gram based metrics as part of their iterative development process.

While the usefulness of these automatic n-gram based metrics is without doubt, at the same time we need to understand these metrics well if we want to have full confidence in their usefulness (cf. Hovy et al. 2002). The behavior of these n-gram metrics needs to be studied with a full range of translation types, genres, and languages. In this paper, we take a step in that direction, looking at how n-gram metrics score professional translations into German of two literary genres. Though we do not anticipate MT being used for literary texts any time in the near future, the results of this study are nonetheless instructive.

1.2 N-gram metrics

The basic idea of n-gram metrics as measures of translation goodness is that a good translation will have a distribution of n-grams similar to other good translations. More precisely, for each segment of the evaluation text, the algorithm examines the corresponding aligned segments from the reference translations, and compares their n-gram counts. These counts are then tallied over all segments in the translation, and with various weights and factors, combined into a final score. In the case of BLEU, the metric is in the range [0,1].[1] The particular details of BLEU and the NIST n-gram metric are not crucial here, since their behavior on our data is very similar.

Before turning to the actual experiments, it is worth making the observation that the n-gram metrics proposed and used to date are not, strictly speaking, measures of translation goodness. Rather, they are measures of document similarity. Their value as measures of translation goodness comes from the *assumption* that a good translation of a text will be similar to other good translations of the same text. As we will see, that assumption may not always hold, leading to some problems in using these metrics for evaluation.

2 Experiments

2.1 The texts

SRI is developing a bidirectional speech to speech translation system for English and Pashto. As such, we have a strong interest in knowing how well n-gram metrics will perform, not only for English, but for the morphologically richer language Pashto. However, one of the huge challenges in developing an English/Pashto system is the lack of data, and especially the lack of translation data.

German serves as a surrogate for Pashto in this regard. It has approximately the same richness of morphology and syntactic variation as does Pashto. At the same time, there is much more data available in German than in Pashto.

Even so, getting multiple professional translations of a single text is a challenging task, without commissioning them ourselves. Following the lead of other researchers (e.g. Resnik et al. 2000; Yarowsky et al. 2001), we used the Bible as one such text, in particular the first four chapters of Luke (ca. 4500 words). The Bible was the only text for which we found multiple electronic versions (e.g. at http://www.biblegateway.com/). Note that since the n-gram metrics measure document similarity, we did not need to have a definitive source language text for comparison (which in any case would not have been English).

Bible	**Tom Sawyer**
Bengel, 1974	Jacobi, 1930
Elberfelder, 1855	Johannsen, 1900
Hoffnung für Alle, 1999	Krüger, 1985
Luther, 1545	Lorenz, 1994
Schlachter, 1951	Roch, 2001
	Torberg, 1996

Table 1: Translations used

The second text we used was the first two chapters of *Tom Sawyer* (also ca. 4500 words). We were able to find several German translations, though only one electronic one. The others we entered ourselves.

Both the Bible translations and the Tom Sawyer translations were done by professional translators (with the possible exception of the Tom Sawyer translation we found on the web). The list of translations of the two texts that we used is given in Table 1.

We also used two commercially available MT systems, Systran (via http://babelfish.altavista.com/) and L&H Power Translator Pro Version 7.0, to create additional German versions of the two texts. The English source for Luke was the New American Standard version at http://www.biblegateway.com/, while the English source for Tom Sawyer was from http://www.gutenberg.org/.

2.2 The measurements

Our technique was to evaluate each translation against all combinations of human translations. Existing evidence is that the metrics are sensitive to the number of reference translations used. Following the BLEU work, we calculated scores using both 2 and 4 reference translations. Given the 5 Bible translations, there were 10 combinations of 2 reference translations and 5 combinations of 4 reference translations. Similarly for Tom Sawyer, there were 15 combinations of 2 reference translations and 15 combinations of 4 reference translations.

Luke was segmented by verse. Where versification was different across translations, we used the smallest common versification, merging verses where necessary rather than making arbitrary splits. Tom Sawyer was segmented into sentences, merging where necessary. For direct speech we treated each conversational turn, including short introduc-

[1] It is less clear what the range of the NIST metric is. A text compared with itself among the reference translations gets a BLEU score of 1, while the NIST scores for our texts compared to themselves ranged from 12.8744 to 14.5006.

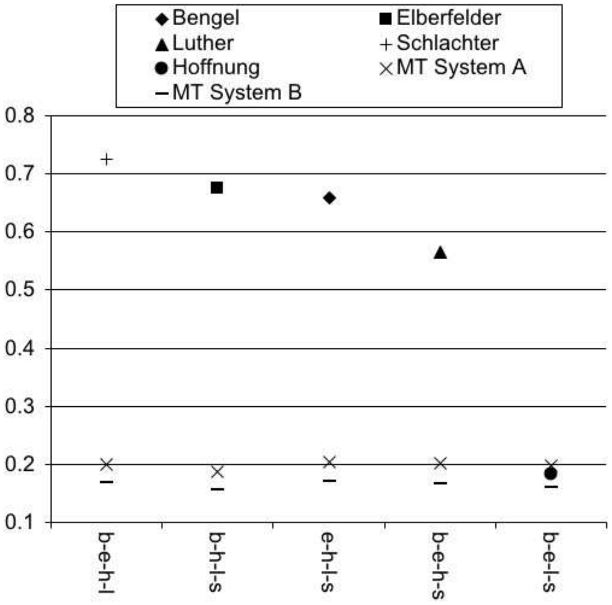

Figure 1: BLEU scores for segmented Luke with 4 reference translations

tions like "So he said:", as a unit even when it contained more than one (usually short) sentence, so that the segment size would be comparable to Luke. In addition, we explored the issue of whether segmentation matters by removing all segmentation from the texts and evaluating again.

3 Results

The BLEU scores for Luke with four reference translations can be seen in Figure 1. The lowest score for a professional human translation (Hoffnung) was 0.1835, which is lower than the score for MT System A (0.1971).[2]

The BLEU and NIST metrics did not differ substantially from each other with respect to the relative ranking of translations in our experiments. Nor was there a substantial difference between considering texts by segments or as undivided units. As an example, we include in Figure 2 the NIST scores for unsegmented Luke with four reference translations. We can note that in this case, the worst human translation (Hoffnung again), was just slightly better than the best machine translation.[3]

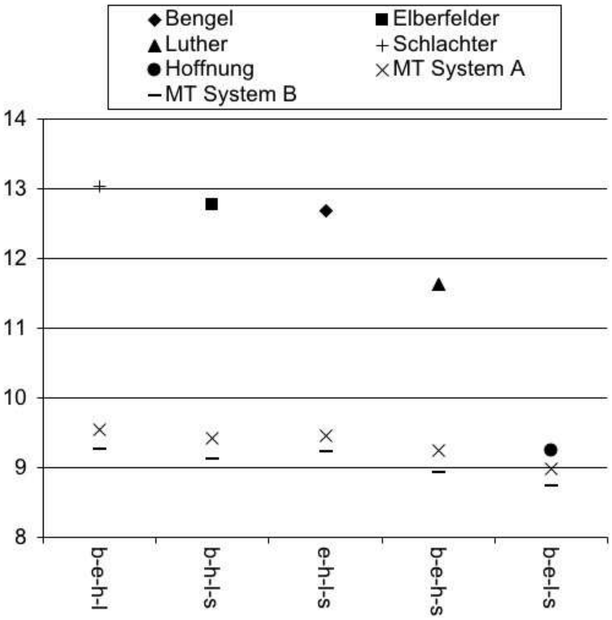

Figure 2: NIST scores for unsegmented Luke with 4 reference translations

The BLEU scores for Luke with two reference translations can be found in Figure 3. Note that the absolute scores a translation received vary quite substantially depending on which two reference translations are used. For example, the Elberfelder translation gets a score of 0.6434 with Bengel and Schlachter as the reference translations, and a score of 0.387 with Hoffnung and Luther as the reference translations. Even the relative ranking of two translations with respect to each other changes. For example, the Schlachter translation receives a much better score than the Bengel translation when Elberfelder and Hoffnung are the reference translations, but a worse score when Hoffnung and Luther are the reference translations. Indeed, Schlachter has the second largest standard deviation of any of the translations (see Table 2).[4]

The BLEU scores for Tom Sawyer with four reference translations are given in Figure 4. Note that the lowest score for a human translation (Lorenz, which incidentally claims to be 'the only authorized translation') was 0.1923, while MT System A scored 0.2216 and MT System B scored 0.1886 with re-

[2] We have anonymized the two MT systems, since their results are similar, and our discussion is not about the quality of any particular system.

[3] Throughout, the NIST scores rated the worst human translations a little lower than BLEU when run on segmented text, but somewhat higher when run on unsegmented text.

[4] Hoffnung and the two MT systems have low standard deviations since they uniformly scored poorly.

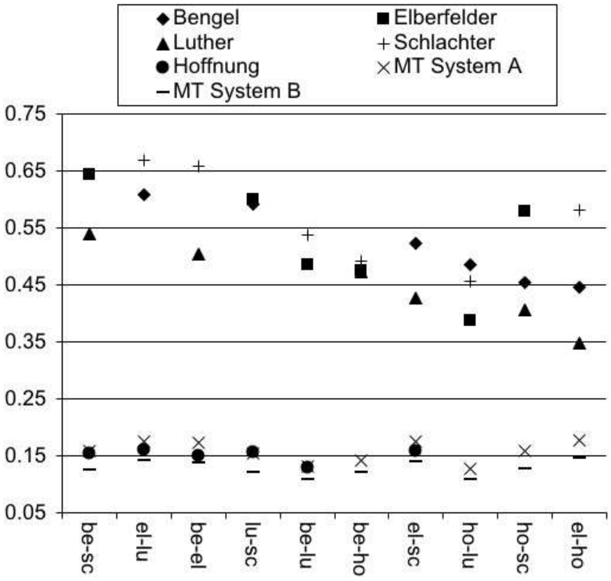

Figure 3: BLEU scores for segmented Luke with 2 reference translations

	Standard Deviation
Bengel	0.070
Elberfelder	0.095
Luther	0.070
Schlachter	0.087
Hoffnung	0.012
MT System A	0.019
MT System B	0.013

Table 2: Standard deviation for the translations in Figure 3 (Luke with 2 reference translations)

spect to the same four reference translations. This is in spite of the fact that both MT systems produce many incomprehensible sentences, struggling with ambiguity, leaving some words in English, and making consistent mistakes like using polite pronouns instead of informal ones. For example, the English sentence *your saying so don't make it so* gets translated by MT System A as (1) and by MT System B as (2):

(1) *ihr Sprichwort macht es deshalb nicht damit*
your proverb makes it therefore not with it
(2) *ihr Saying also bilden es nicht so*
your Saying therefore forms it not so

Figure 5 gives the BLEU scores for Tom Sawyer with two reference translations. Note that it is still possible to see that the human translation by Lorenz gets scores similar to those of the machine translation systems. However, unlike with four reference translations, it is much harder to draw a clear line between the other human translations and the machine translations, at least for some sets of reference translations. The scores for the human translations also vary widely depending on which reference translations are used. For example, with Lorenz and Torberg as the reference translations, Krüger scores 0.3255, substantially better than Jacobi (0.1928). But with Johannsen and Lorenz as the reference translations, Krüger scores only 0.2203, which is worse than Jacobi (0.2694). We can also note that the standard deviations of the translations compared with two reference translations are much larger than the standard deviations of the translations compared with four reference translations (see Table 3).

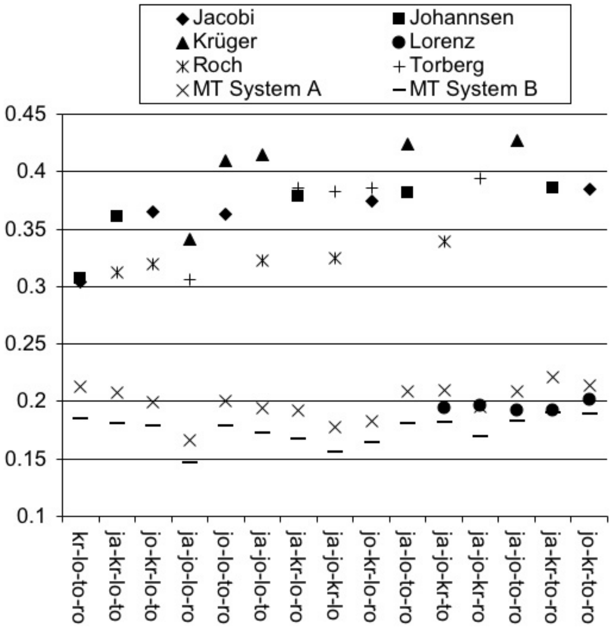

Figure 4: BLEU scores for segmented Tom Sawyer with 4 reference translations

The NIST scores for Tom Sawyer with two reference translations (Figure 6) make it even harder to distinguish some human translations from machine translations. The score for MT System A with Krüger and Torberg as the reference translations is 5.5864, which is not much worse than the score that Johannsen's translation gets for the same ref-

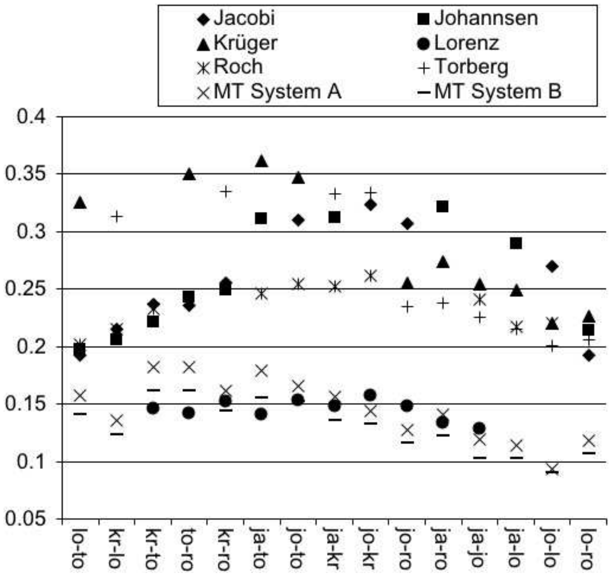

Figure 5: BLEU scores for segmented Tom Sawyer with 2 reference translations

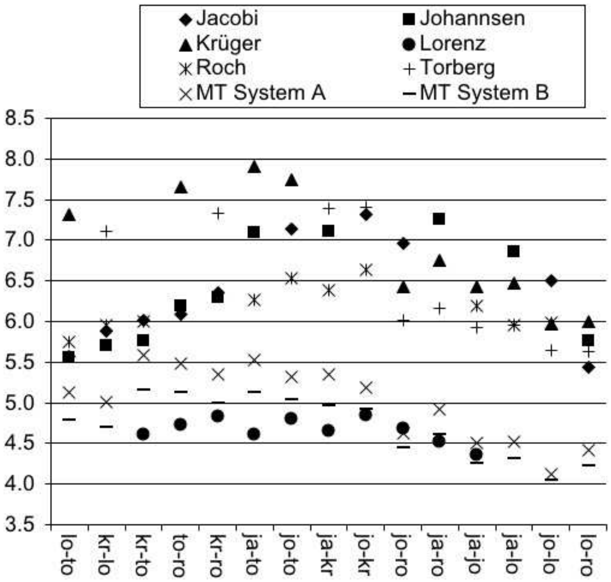

Figure 6: NIST scores for segmented Tom Sawyer with 2 reference translations

erence translations (5.7650). And it is better than the score that Johannsen's (5.5536) and Jacobi's (5.5649) translations get with Lorenz and Torberg as the reference translations.

One reason why some of the human translations get relatively low scores is that the human translators of Tom Sawyer often did not translate particularly faithfully, and made changes even when there was no clear necessity to do so. Consider the famous list of items Tom Sawyer ends up with at the end of Chapter 2:

“He had besides the things before mentioned, twelve marbles, part of a jews-harp, a piece of blue bottle-glass to look through, a spool cannon, a key that wouldn't unlock anything, a fragment of chalk, a glass stopper of a decanter, a tin soldier, a couple of tadpoles, six fire-crackers, a kitten with only one eye, a brass doorknob, a dog-collar – but no dog – the handle of a knife, four pieces of orange-peel, and a dilapidated old window sash.”

	Standard Deviation 4 references	Standard Deviation 2 references
Jacobi	0.031	0.048
Johannsen	0.033	0.048
Krüger	0.036	0.054
Lorenz	0.004	0.009
Roch	0.010	0.020
Torberg	0.036	0.058
MT System A	0.015	0.027
MT System B	0.012	0.023

Table 3: Standard deviation for Figures 4 and 5 (Sawyer with 4 and 2 reference translations)

In her translation, Jacobi turns the ‘jews-harp’ into a harmonica, the ‘tin soldier’ into a lead soldier, the ‘tadpoles’ into a piece of string, and the ‘knife handle’ into a knife blade. She also adds modifiers like *stone* marbles, *fairly damaged* harmonica, *half* a spool cannon, *old* key, *half broken* glass stopper, and *old* brass door-knob.

Roch turns the ‘key’ into a knife, the ‘kitten’ into a rabbit, and the ‘tadpoles’ into the head of a frog. He also leaves out ‘of a decanter’. In fact, five of the six human translators make these kinds of changes in this passage, which are not due to constraints of the target language.

4 Discussion

The n-gram metrics assume that a good translation will be similar to other good translations. Our assumption is that all of the human translations we have considered are “good” translations (at least certainly fluent). Our results show that these two assumptions are not entirely compatible. In particular, we saw that a clearly nonfluent machine translation could score better than a completely fluent human translation, in either genre.

There are two kinds of moral one can draw from this story. The first kind concerns potential pitfalls in using these metrics and how to avoid them. One potential problem is that a low n-gram score is *not* necessarily indicative of a poor translation, although a high n-gram score (where what counts as high depends on the number of reference translations and other factors) is probably indicative of a good translation.

A second point is that it is critical to control the type of translation represented by the reference translations. For example, the bulk of our human Bible translations were fairly “strict” (though not necessarily literal) translations. The human translations that scored poorly were generally “freer” translations. While Luther and its older language always scores somewhat lower than the group of more modern strict translations, the much less strict Hoffnung translation gets substantially lower scores than any of the more strict translations. In other words, the pool of translations favors the stricter translations over the freer one. In order to make the best use of an n-gram metric, the reference translations should be roughly the same style as the translation to be judged. That probably means fairly strict translations for the time being.

This is also suggested by an informal preliminary study we did with translations of a news story. In this study the translation by the only professional translator got worse scores than the translations of all seven non-professionals, even though it was superior in many respects. This is because the non-professional translations tended to be fairly literal and stayed as close to the source text as possible, and the professional translation was an exception in the otherwise fairly homogenous set of translations. Having relatively literal reference translations also resulted in better scores for the machine translation systems.

A third potential problem is that with only two reference translations, it is harder to distinguish good and not so good translations. With four reference translations, the contrast between good and not so good translations is much clearer, as can be seen when comparing Figure 4 and Figure 5.[5]

The second kind of moral one can draw concerns general properties of these metrics. One point to make in this regard is that absolute scores are not necessarily comparable. Our results confirm that a greater number of reference translations does give rise to higher scores, across the board. In addition, scores may vary by genre and type of translation. We can note that the BLEU scores for some human Bible translations were roughly double the highest scores for human translations reported by Papineni et al. 2002.[6] And of course we have just commented on how scores can vary according to the type of reference translations.

Furthermore, we would like the scores of a text compared to a completely unrelated set of references to be substantially lower than the scores of a true translation corresponding to that set of references. Figure 7 shows the NIST scores for all of the texts, Luke and Tom Sawyer, with four translations of Luke as the reference sets. While the scores for the actual translations of Luke are higher than those of the Sawyer translations, the differences are not always as great as we might like. For example, there is a real Luke translation (Hoffnung) whose score is closer to a false translation (Lorenz) than to another true translation (Luther).

A final important point is a reminder that the n-gram metrics *are* really document similarity measures rather than true translation quality measures. This point is emphasized by our result that relative scores of segmented and unsegmented translations were the same, even though the absolute scores of the unsegmented translations were slightly higher.

In this context it is interesting to evaluate texts that are not translations but that are fluent texts from the same genre about the same topic. To this end, we compared the Schlachter translations of the other three gospels, Matthew, Mark, and John, to the three lowest scoring translations of Luke (Hoffnung and the two MT systems), using the human translations of Luke as the reference translations.

[5] Note too that the NIST metric does not make as sharp a contrast as does BLEU.

[6] We speculate that a combination of tradition, scholarship, and close adherence to the original lead to more uniformity of the strict Bible translations, leading to higher n-gram scores.

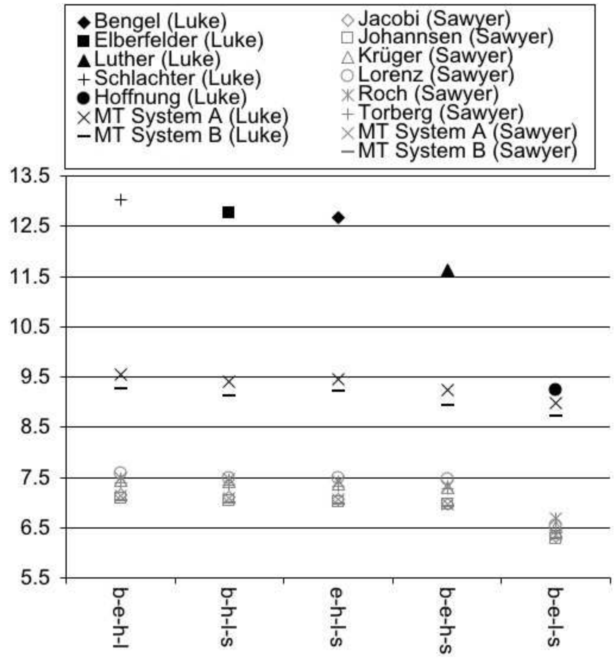

Figure 7: NIST scores for translations of Luke and Sawyer with Luke as the reference translations

The results are in Figure 8, where the solid black symbols are the non-Luke texts. What we see is that the Schlachter translations of Matthew, and especially Mark, are comparable to the MT translations of Luke. John, the gospel most different from the rest, is substantially lower. These results show that BLEU scores of fluent texts from the same genre about the same topic may be indistinguishable from the scores of actual translations by MT systems.

Of course, these n-gram metrics are only one type of document similarity metric. Perhaps other similarity metrics, either alone or in combination with these n-gram metrics, would overcome some of the problems we have seen here.

5 Conclusion

As with many things in life, we should be neither too pessimistic about the use of n-gram metrics to evaluate MT (a low score isn't necessarily bad) nor too optimistic (two reference translations may not be sufficient to differentiate among translations). Despite their problems, n-gram metrics have a very useful role to play in MT. It will be extremely useful to use them during the development cycle to iterate translation quality against a fixed set of reference translations. Indeed, this was the primary use that the BLEU team saw for BLEU. However, we should be much more cautious about using n-gram metrics as the basis for a formal evaluation.

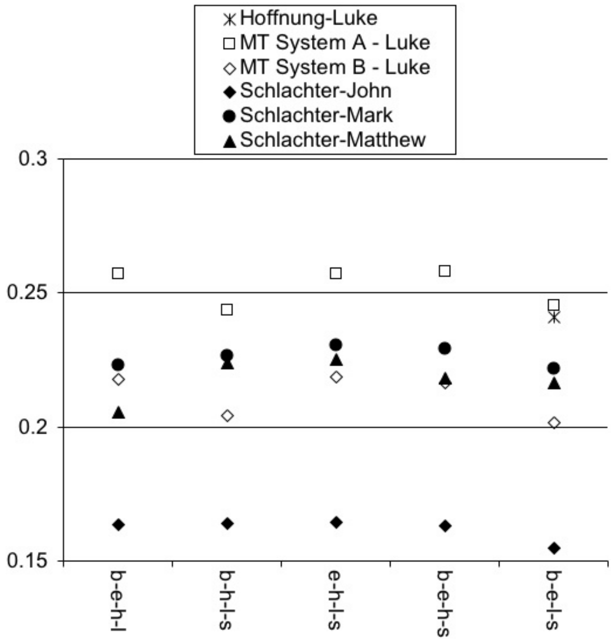

Figure 8: BLEU scores of the other gospels with Luke translations as the reference

Acknowledgments

Part of this work was done while the first author was at SRI International. We would like to thank Kristin Precoda, Ascander Dost, Dave Kale, Anand Venkataraman, and the audience of our Stanford NLP Group presentation for helpful discussion. This work was supported under SPAWAR contract N66001-99-D-8504 with funding from DARPA.

References

–. 1974. *Bengel Bibel*. Hänssler Verlag.

–. 1855. *Elberfelder Bibel*. (revised 1985). Brockhaus Verlag.

–. 1545. *Luther Bibel*.

–. 1951. *Schlachter Bibel*. Genfer Bibelgesellschaft.

–. 1999. *Hoffnung für Alle*. IBS.

–. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. NIST Report http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

Eduard Hovy, Maghi King, Andrei Popescu-Belis. 2002. An Introduction to MT Evaluation. *Machine Translation Evaluation: Human Evaluators Meet Automated Metrics*. Workshop at the LREC 2002 Conference.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 2000. The Bible as a parallel corpus: annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1-2):129-153.

Mark Twain, translated by Margarete Jacobi. 1930. *Tom Sawyers Abenteuer und Streiche*. Hesse & Becker, Leipzig.

Mark Twain, translated by Ulrich Johannsen. 1900. *Die Abenteuer des Tom Sawyer*. Maschler, Berlin.

Mark Twain, translated by Lore Krüger. 1985. *Tom Sawyers Abenteuer*. Diogenes Taschenbuch.

Mark Twain, translated by Hertha Lorenz. 1994. *Tom Sawyers Abenteuer*. Neuer Kaiser Verlag, Klagenfurt.

Mark Twain, translated by Andreas Roch. 2001. *Die Abenteuer von Tom Sawyer*. http://www.andiroch.de/twain/tomsawy.htm

Mark Twain, translated by Peter Torberg. 1996. *Die Abenteuer von Tom Sawyer*. Fischer Taschenbuchverlag, Frankfurt.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*