Evaluation without references: IBM1 scores as evaluation metrics

Maja Popović, David Vilar, Eleftherios Avramidis, Aljoscha Burchardt
German Research Center for Artificial Intelligence (DFKI)
Language Technology (LT), Berlin, Germany
name.surname@dfki.de

Abstract

Current metrics for evaluating machine translation quality have the huge drawback that they require human-quality reference translations. We propose a truly automatic evaluation metric based on IBM1 lexicon probabilities which does not need any reference translations. Several variants of IBM1 scores are systematically explored in order to find the most promising directions. Correlations between the new metrics and human judgments are calculated on the data of the third, fourth and fifth shared tasks of the Statistical Machine Translation Workshop. Five different European languages are taken into account: English, Spanish, French, German and Czech. The results show that the IBM1 scores are competitive with the classic evaluation metrics, the most promising being IBM1 scores calculated on morphemes and POS-4grams.

1 Introduction

Currently used evaluation metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), etc. are based on the comparison between human reference translations and the automatically generated hypotheses in the target language to be evaluated. While this scenario helps in the design of machine translation systems, it has two major drawbacks. The first one is the practical criticism that using reference translations is inefficient and expensive: in real-life situations, the quality of machine translation must be evaluated without having to pay humans for producing reference translations first. The second criticism is methodological: in using reference translation, the problem of evaluating translation quality (e.g., completeness, ordering, domain fit, etc.) is transformed into a kind of paraphrase evaluation in the target language, which is a very difficult problem itself. In addition, the set of selected references always represents only a small subset of all good translations. To remedy these drawbacks, we propose a truly automatic evaluation metric which is based on the IBM1 lexicon scores (Brown et al., 1993).

The inclusion of IBM1 scores in translation systems has shown experimentally to improve translation quality (Och et al., 2003). They also have been used for confidence estimation for machine translation (Blatz et al., 2003). To the best of our knowledge, these scores have not yet been used as an evaluation metric.

We carry out a systematic comparison between several variants of IBM1 scores. The Spearman's rank correlation coefficients on the document (system) level between the IBM1 metrics and the human ranking are computed on the English, French, Spanish, German and Czech texts generated by various translation systems in the framework of the third (Callison-Burch et al., 2008), fourth (Callison-Burch et al., 2009) and fifth (Callison-Burch et al., 2010) shared translation tasks.

2 IBM1 scores

The IBM1 model is a bag-of-word translation model which gives the sum of all possible alignment probabilities between the words in the source sentence and the words in the target sentence. Brown et al. (1993) defined the IBM1 probability score for a translation

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 99–103,
Edinburgh, Scotland, UK, July 30–31, 2011.

pair f_1^J and e_1^I in the following way:

$$P(f_1^J|e_1^I) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(f_j|e_i) \quad (1)$$

where f_1^J is the source language sentence of length J and e_1^I is the target language sentence of length I.

As it is a conditional probability distribution, we investigated both directions as evaluation metrics. In order to avoid frequent confusions about what is the source and what the target language, we defined our scores in the following way:

- source-to-hypothesis (sh) IBM1 score:

$$\text{IBM1}_{sh} = \frac{1}{(H+1)^S} \prod_{j=1}^{S} \sum_{i=0}^{H} p(s_j|h_i) \quad (2)$$

- hypothesis-to-source (hs) IBM1 score:

$$\text{IBM1}_{hs} = \frac{1}{(S+1)^H} \prod_{i=1}^{H} \sum_{j=0}^{S} p(h_i|s_j) \quad (3)$$

where s_j are the words of the original source language sentence, S is the length of this sentence, h_i are the words of the target language hypothesis, and H is the length of this hypothesis.

In addition to the standard IBM1 scores calculated on words, we also investigated:

- MIBM1 scores – IBM1 scores of word morphemes in each direction;
- PnIBM1 scores – IBM1 scores of POS n-grams in each direction.

A parallel bilingual corpus for the desired language pair and a tool for training the IBM1 model are required in order to obtain IBM1 probabilities $p(f_j|e_i)$. For the POS n-gram scores, appropriate POS taggers for each of the languages are necessary. The POS tags cannot be only basic but must have all details (e.g. verb tenses, cases, number, gender, etc.). For the morpheme scores, a tool for splitting words into morphemes is necessary.

3 Experiments on WMT 2008, WMT 2009 and WMT 2010 test data

3.1 Experimental set-up

The IBM1 probabilities necessary for the IBM1 scores are learnt using the WMT 2010 News Commentary bilingual corpora consisting of the Spanish-English, French-English, German-English and Czech-English parallel texts. Spanish, French, German and English POS tags were produced using the TreeTagger[1], and the Czech texts are tagged using the COMPOST tagger (Spoustová et al., 2009). The morphemes for all languages are obtained using the Morfessor tool (Creutz and Lagus, 2005). The tool is corpus-based and language-independent: it takes a text as input and produces a segmentation of the word forms observed in the text. The obtained results are not strictly linguistic, however they often resemble a linguistic morpheme segmentation. Once a morpheme segmentation has been learnt from some text, it can be used for segmenting new texts. In our experiments, the splitting are learnt from the training corpus used for the IBM1 lexicon probabilities. The obtained segmentation is then used for splitting the corresponding source texts and hypotheses. Detailed corpus statistics are shown in Table 1.

Using the obtained IBM1 probabilities of words, morphemes and POS n-grams, the scores described in Section 2 are calculated for the Spanish-English, French-English, German-English and Czech-English translation outputs from each translation direction. For each of the IBM1 scores, the system level Spearman correlation coefficients ρ with the human ranking are calculated for each document. In total, 32 correlation coefficients are obtained for each score – four English outputs from the WMT 2010 task, four from the WMT 2009 and eight from the WMT 2008 task, together with sixteen outputs in other four target languages. The obtained correlation results were then summarised into the following three values:

- *mean*
  a correlation coefficient averaged over all translation outputs;

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

	Spanish	English	French	English	German	English	Czech	English
sentences	97122		83967		100222		94693	
running words	2661344	2338495	2395141	2042085	2475359	2398780	2061422	2249365
vocabulary:								
words	69620	53527	56295	50082	107278	54270	125614	52081
morphemes	14178	13449	12004	12485	22211	13499	18789	12961
POS tags	69	44	33	44	54	44	611	44
POS-2grams	2459	1443	826	1443	1611	1454	27835	1457
POS-3grams	27350	20474	10409	19838	19928	20769	209481	20522
POS-4grams	135166	121182	62177	114555	114314	123550	637337	120646

Table 1: Statistics of the corpora for training IBM1 lexicon models.

- *rank*$>$
  percentage of documents where the particular score has better correlation than the other IBM1 scores;

- *rank*$\geq$
  percentage of documents where the particular score has better or equal correlation than the other IBM1 scores.

3.2 Comparison of IBM1 scores

The first step towards deciding which IBM1 score to submit to the WMT 2011 evaluation task was a comparison of the average correlations i.e. *mean* values. These values for each of the IBM1 scores are presented in Table 2. The left column shows average correlations of the source-hypothesis (sh) scores, and the right one of the hypothesis-source (hs) scores.

mean	IBM1$_{sh}$	IBM1$_{hs}$
words	0.066	0.308
morphemes	0.227	0.445
POS tags	0.006	0.337
POS-2grams	0.058	0.337
POS-3grams	0.172	0.376
POS-4grams	0.196	0.442

Table 2: Average correlations of source-hypothesis (left column) and hypothesis-source (right column) IBM1 scores.

It can be seen that the morpheme, POS-3gram and POS-4gram scores have the best correlations in both directions. Apart from that, it can be observed that all the hs scores have better correlations than sh scores. Therefore, all the further experiments will deal only with the hs scores, and the subscript hs is omitted.

In the next step, all the hs scores are sorted according to each of the three values described in Section 3.1, i.e. average correlation *mean*, *rank*$>$ and *rank*$\geq$, and the results are shown in Table 3. The most promising scores according to each of the three values are morpheme score MIBM1, POS-3gram score P3IBM1 and POS-4gram score P4IBM1.

3.2.1 Combined IBM1 scores

The last experiment was to combine the most promising IBM1 scores in order to see if the correlation with human rankings can be further improved. In general, a combined IBM1 score is defined as arithmetic mean of various individual IBM1$_{hs}$ scores described in Section 2:

$$\text{COMBIBM1} = \sum_{k=1}^{K} w_k \cdot \text{IBM1}_k \qquad (4)$$

The following combinations were investigated:

- P1234IBM1
  combination of all POS n-gram scores;

- MP1234IBM1
  combination of all POS n-gram scores and the morpheme score;

- MP34IBM1
  combination of the most promising individual scores, i.e. POS-3gram, POS-4gram and morpheme scores;

mean		*rank*$>$		*rank*$\geq$	
0.445	morphemes	60.6	POS-4grams	71.3	POS-4grams
0.442	POS-4grams	54.4	morphemes	61.3	POS-3grams
0.376	POS-3grams	50.6	POS-3grams	56.3	morphemes
0.337	POS-2grams	39.4	POS tags	48.1	POS tags
0.337	POS tags	36.3	words	43.7	POS-2grams
0.308	words	35.6	POS-2grams	42.5	words

Table 3: IBM1$_{hs}$ scores sorted by average correlation (column 1), *rank*$>$ value (column 2) and *rank*$\geq$ value (column 3). The most promising scores are those calculated on morphemes (MIBM1), POS-3grams (P3IBM1) and POS-4grams (P4IBM1).

- MP4IBM1
  combination of the two most promising individual scores, i.e. POS-4gram score and morpheme score.

For each of the scores, two variants were investigated, with and without (i.e. with uniform) weights w_k. The weigths were choosen proportionally to the average correlation of each individual score. Table 4 contains average correlations for all combined scores, together with the weight values.

combined score	*mean*
P1234IBM1	0.403
+weights (0.15, 0.15, 0.3, 0.4)	0.414
MP1234IBM1	0.466
+weights (0.2, 0.05, 0.05, 0.2, 0.5)	0.486
MP34IBM1	0.480
+weights (0.25, 0.25, 0.5)	**0.498**
MP4IBM1	0.494
+weights (0.4, 0.6)	**0.496**

Table 4: Average correlations of the investigated IBM1$_{hs}$ combinations. The weight values are choosen according to the average correlation of the particular individual IBM1 score.

The POS n-gram combination alone does not yield any improvement over the best individual scores. Introduction of the morpheme score increases the average correlation, especially when only the best n-gram scores are chosen. Apart from that, introducing weights improves the average correlation for each of the combined scores.

The final step in our experiments consists of ranking the weighted combined scores. The *rank*$>$ and *rank*$\geq$ values for these scores are presented in Table 5. According to the *rank*$>$ values, the MP4IBM1 score clearly outperforms all other scores. This score also has the highest *mean* value together with the MP34IBM1 score. As for *rank*$\geq$ values, all morpheme-POS scores have similar values significantly outperforming the P1234IBM1 score.

combined score	*rank*$>$	*rank*$\geq$
P1234IBM1	25.0	36.4
MP1234IBM1	44.8	68.7
MP34IBM1	39.6	64.6
MP4IBM1	**55.2**	65.7

Table 5: *rank*$>$ (column 1) and *rank*$\geq$ (column 2) values of the weighted IBM1$_{hs}$ combinations.

Following all these observations, we decided to submit the MP4IBM1 score to the WMT 2011 evaluation task.

4 Conclusions and outlook

The results presented in this article show that the IBM1 scores have the potential to be used as replacement of current evaluation metrics based on reference translations. Especially the scores abstracting away from word surface particularities (i.e. vocabulary, domain) based on morphemes, POS-3grams and 4grams show a high average correlation of about 0.5 (the average correlation of the BLEU score on the same data is 0.566).

An important point for future optimisation is to investigate effects of the selection of training data for the IBM1 models (and its similarity to the training data of the involved statistical translation systems). Furthermore, investigation of how to assign the weights for combining the corresponding indi-

vidual scores, as well as of the possible impact of different morpheme splittings should be carried out. Other direction for future work is combination with other features (i.e. POS language models).

This method is currently being tested and further developed in the framework of the TARAXÜ project[2]. In this project, three industry and one research partners develop a hybrid machine translation architecture that satisfies current industry needs, which includes a number of large-scale evaluation rounds involving various languages: English, French, German, Czech, Spanish, Russian, Chinese and Japanese. By the time of writing this article, the first human evaluation round in TARAXÜ on a pilot set of about 7000 sentences is running. The metrics proposed in this paper will be tested on the TARAXÜ data as soon as they are available. First results will be reported in the presentation of this paper.

Acknowledgments

This work has been partly developed within the TARAXÜ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of the ACL 05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, June.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Final report, JHU/CLSP Summer Workshop.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the 3rd ACL 08 Workshop on Statistical Machine Translation (WMT 08)*, pages 70–106, Columbus, Ohio, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (WMT 10)*, pages 17–53, Uppsala, Sweden, July.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report Report A81, Computer and Information Science, Helsinki University of Technology, Helsinki, Finland, March.

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. Syntax for statistical machine translation. Technical report, Johns Hopkins University 2003 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, USA, August.

Kishore Papineni, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, July.

Drahomíra "Johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March.

[2] http://taraxu.dfki.de/