AM-FM: A Semantic Framework for Translation Quality Assessment

Rafael E. Banchs
Human Language Technology Department
Institute for Infocomm Research
1 Fusionopolis Way, Singapore 138632
rembanchs@i2r.a-star.edu.sg

Haizhou Li
Human Language Technology Department
Institute for Infocomm Research
1 Fusionopolis Way, Singapore 138632
hli@i2r.a-star.edu.sg

Abstract

This work introduces AM-FM, a semantic framework for machine translation evaluation. Based upon this framework, a new evaluation metric, which is able to operate without the need for reference translations, is implemented and evaluated. The metric is based on the concepts of adequacy and fluency, which are independently assessed by using a cross-language latent semantic indexing approach and an n-gram based language model approach, respectively. Comparative analyses with conventional evaluation metrics are conducted on two different evaluation tasks (overall quality assessment and comparative ranking) over a large collection of human evaluations involving five European languages. Finally, the main pros and cons of the proposed framework are discussed along with future research directions.

1 Introduction

Evaluation has always been one of the major issues in Machine Translation research, as both human and automatic evaluation methods exhibit very important limitations. On the one hand, although highly reliable, in addition to being expensive and time consuming, human evaluation suffers from inconsistency problems due to inter- and intra-annotator agreement issues. On the other hand, while being consistent, fast and cheap, automatic evaluation has the major disadvantage of requiring reference translations. This makes automatic evaluation not reliable in the sense that good translations not matching the available references are evaluated as poor or bad translations.

The main objective of this work is to propose and evaluate AM-FM, a semantic framework for assessing translation quality without the need for reference translations. The proposed framework is theoretically grounded on the classical concepts of adequacy and fluency, and it is designed to account for these two components of translation quality in an independent manner. First, a cross-language latent semantic indexing model is used for assessing the adequacy component by directly comparing the output translation with the input sentence it was generated from. Second, an n-gram based language model of the target language is used for assessing the fluency component.

Both components of the metric are evaluated at the sentence level, providing the means for defining and implementing a sentence-based evaluation metric. Finally, the two components are combined into a single measure by implementing a weighted harmonic mean, for which the weighting factor can be adjusted for optimizing the metric performance.

The rest of the paper is organized as follows. Section 2, presents some background work and the specific dataset that has been used in the experimental work. Section 3, provides details on the proposed AM-FM framework and the specific metric implementation. Section 4 presents the results of the conducted comparative evaluations. Finally, section 5 presents the main conclusions and relevant issues to be dealt with in future research.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 153–158,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

2 Related Work and Dataset

Although BLEU (Papineni *et al.*, 2002) has become a *de facto* standard for machine translation evaluation, other metrics such as NIST (Doddington, 2002) and, more recently, Meteor (Banerjee and Lavie, 2005), are commonly used too. Regarding the specific idea of evaluating machine translation without using reference translations, several works have proposed and evaluated different approaches, including round-trip translation (Somers, 2005; Rapp, 2009), as well as other regression- and classification-based approaches (Quirk, 2004; Gamon *et al.*, 2005; Albrecht and Hwa, 2007; Specia *et al.*, 2009).

As part of the recent efforts on machine translation evaluation, two workshops have been organizing shared-tasks and evaluation campaigns over the last four years: the NIST Metrics for Machine Translation Challenge[1] (MetricsMATR) and the Workshop on Statistical Machine Translation[2] (WMT); which were actually held as one single event in their most recent edition in 2010.

The dataset used in this work corresponds to WMT-07. This dataset is used, instead of a more recent one, because no human judgments on adequacy and fluency have been conducted in WMT after year 2007, and human evaluation data is not freely available from MetricsMATR.

In this dataset, translation outputs are available for fourteen tasks involving five European languages: English (EN), Spanish (ES), German (DE), French (FR) and Czech (CZ); and two domains: News Commentaries (News) and European Parliament Debates (EPPS). A complete description on WMT-07 evaluation campaign and dataset is available in Callison-Burch *et al.* (2007).

System outputs for fourteen of the fifteen systems that participated in the evaluation are available. This accounts for 86 independent system outputs with a total of 172,315 individual sentence translations, from which only 10,754 were rated for both adequacy and fluency by human judges.

The specific vote standardization procedure described in section 5.4 of Blatz *et al.* (2003) was applied to all adequacy and fluency scores for removing individual voting patterns and averaging votes. Table 1 provides information on the corresponding domain, and source and target languages for each of the fourteen translation tasks, along with their corresponding number of system outputs and the amount of sentence translations for which human evaluations are available.

Task	Domain	Src.	Tgt.	Syst.	Sent.
T1	News	CZ	EN	3	727
T2	News	EN	CZ	2	806
T3	EPPS	EN	FR	7	577
T4	News	EN	FR	8	561
T5	EPPS	EN	DE	6	924
T6	News	EN	DE	6	892
T7	EPPS	EN	ES	6	703
T8	News	EN	ES	7	832
T9	EPPS	FR	EN	7	624
T10	News	FR	EN	7	740
T11	EPPS	DE	EN	7	949
T12	News	DE	EN	5	939
T13	EPPS	ES	EN	8	812
T14	News	ES	EN	7	668

Table 1: Domain, source language, target language, system outputs and total amount of sentence translations (with both adequacy and fluency human assessments) included in the WMT-07 dataset

3 Semantic Evaluation Framework

The framework proposed in this work (AM-FM) aims at assessing translation quality without the need for reference translations, while maintaining consistency with human quality assessments. Different from other approaches not using reference translations, we rely on a cross-language version of latent semantic indexing (Dumais *et al.*, 1997) for creating a semantic space where translation outputs and inputs can be directly compared.

A two-component evaluation metric, based on the concepts of adequacy and fluency (White *et al.*, 1994) is defined. While adequacy accounts for the amount of source meaning being preserved by the translation (5:all, 4:most, 3:much, 2:little, 1:none), fluency accounts for the quality of the target language in the translation (5:flawless, 4:good, 3:non-native, 2:disfluent, 1:incomprehensible).

3.1 Metric Definition

For implementing the adequacy-oriented component (AM) of the metric, the cross-language latent semantic indexing approach is used (Dumais *et al.*, 1997), in which the source sentence originating the translation is used as evaluation reference. Accord-

[1] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/

[2] http://www.statmt.org/wmt10/

ing to this, the AM component can be regarded to be mainly adequacy-oriented as it is computed on a cross-language semantic space.

For implementing the fluency-oriented component (FM) of the proposed metric, an n-gram based language model approach is used (Manning and Schutze, 1999). This component can be regarded to be mainly fluency-oriented as it is computed on the target language side in a manner that is totally independent from the source language.

For combining both components into a single metric, a weighted harmonic mean is proposed:

$$\boldsymbol{AM\text{-}FM = AM\ FM / (\alpha\, AM + (1\text{-}\alpha)\ FM)} \qquad (1)$$

where α is a weighting factor ranging from $\alpha=0$ (pure AM component) to $\alpha=1$ (pure FM component), which can be adjusted for maximizing the correlation between the proposed metric AM-FM and human evaluation scores.

3.2 Implementation Details

The adequacy-oriented component of the metric (AM) was implemented by following the procedure proposed by Dumais *et al.* (1997), where a bilingual collection of data is used to generate a cross-language projection matrix for a vector-space representation of texts (Salton *et al.*, 1975) by using singular value decomposition: SVD (Golub and Kahan, 1965).

According to this formulation, a bilingual term-document matrix $\boldsymbol{X}_{ab}$ of dimensions $M*N$, where $M=(M_a+M_b)$ are vocabulary terms in languages *a* and *b*, and *N* are documents (sentences in our case), can be decomposed as follows:

$$\boldsymbol{X}_{ab} = [\boldsymbol{X}_a;\boldsymbol{X}_b] = \boldsymbol{U_{ab}\, \Sigma_{ab}\, V_{ab}}^T \qquad (2)$$

where $[\boldsymbol{X}_a;\boldsymbol{X}_b]$ is the concatenation of the two monolingual term-document matrices $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ (of dimensions M_a*N and M_b*N) corresponding to the available parallel training collection, $\boldsymbol{U_{ab}}$ and $\boldsymbol{V_{ab}}$ are unitary matrices of dimensions $M*M$ and $N*N$, respectively, and $\boldsymbol{\Sigma}$ is an $M*N$ diagonal matrix containing the singular values associated to the decomposition.

From the singular value decomposition depicted in (2), a low-dimensional representation for any sentence vector $\boldsymbol{x_a}$ or $\boldsymbol{x_b}$, in language *a* or *b*, can be computed as follows:

$$\boldsymbol{y_a}^T = [\boldsymbol{x_a};\boldsymbol{0}]^T\ \boldsymbol{U}_{\boldsymbol{ab}M*L} \qquad (3.a)$$

$$\boldsymbol{y_b}^T = [\boldsymbol{0};\boldsymbol{x_b}]^T\ \boldsymbol{U}_{\boldsymbol{ab}M*L} \qquad (3.b)$$

where $\boldsymbol{y_a}$ and $\boldsymbol{y_b}$ represent the *L*-dimensional vectors corresponding to the projections of the full-dimensional sentence vectors $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$, respectively; and $\boldsymbol{U}_{\boldsymbol{ab}M*L}$ is a cross-language projection matrix composed of the first *L* column vectors of the unitary matrix $\boldsymbol{U_{ab}}$ obtained in (2).

Notice, from (3a) and (3b), how both sentence vectors $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ are padded with zeros at each corresponding other-language vocabulary locations for performing the cross-language projections. As similar terms in different languages would have similar occurrence patterns, theoretically, a close representation in the cross-language reduced space should be obtained for terms and sentences that are semantically related. Therefore, sentences can be compared across languages in the reduced space.

The AM component of the metric is finally computed in the projected space by using the cosine similarity between the source and target sentences:

$$\boldsymbol{AM = [s;0]^T P\ ([0;t]^T P)^T /\ |[s;0]^T P|\ /\ |[0;t]^T P|} \qquad (4)$$

where $\boldsymbol{P}$ is the projection matrix $\boldsymbol{U}_{\boldsymbol{ab}M*L}$ described in (3a) and (3b), $\boldsymbol{[s;0]}$ and $\boldsymbol{[0;t]}$ are vector space representations of the source and target sentences being compared (with their target and source vocabulary elements set to zero, respectively), and | | is the *L2-norm* operator. In a final implementation stage, the range of AM is restricted to the interval [0,1] by truncating negative results.

For computing the projection matrices, random sets of 10,000 parallel sentences[3] were drawn from the available training datasets. The only restriction we imposed to the extracted sentences was that each should contain at least 10 words. Seven projection matrices were constructed in total, one for each different combination of domain and language pair. TF-IDF weighting was applied to the constructed term-document matrices while maintaining all words in the vocabularies (i.e. no stop-words were removed). All computations related to SVD, sentence projections and cosine similarities were conducted with MATLAB.

[3] Although this accounts for a small proportion of the datasets (20% of News and 1% of European Parliament), it allowed for maintaining computational requirements under control while still providing a good vocabulary coverage.

The fluency-oriented component FM is implemented by using an n-gram language model. In order to avoid possible effects derived from differences in sentence lengths, a compensation factor is introduced in log-probability space. According to this, the FM component is computed as follows:

$$\boldsymbol{FM} = exp(\Sigma_{n=1:N} \log(p(w_n|w_{n-1},\ldots))/N) \quad (5)$$

where $p(w_n|w_{n-1},\ldots)$ represent the target language n-gram probabilities and N is the total number of words in the target sentence being evaluated.

By construction, the values of FM are also restricted to the interval [0,1]; so, both component values range within the same interval.

Fourteen language models were trained in total, one per task, by using the available training datasets. The models were computed with the SRILM toolbox (Stolcke, 2002).

As seen from (4) and (5), different from conventional metrics that compute matches between translation outputs and references, in the AM-FM framework, a semantic embedding is used for assessing the similarities between outputs and inputs (4) and, independently, an n-gram model is used for evaluating output language quality (5).

4 Comparative Evaluations

In order to evaluate the AM-FM framework, two comparative evaluations with standard metrics were conducted. More specifically, BLEU, NIST and Meteor were considered, as they are the metrics most frequently used in machine translation evaluation campaigns.

4.1 Correlation with Human Scores

In this first evaluation, AM-FM is compared with standard evaluation metrics in terms of their correlations with human-generated scores. Different from Callison-Burch *et al.* (2007), where Spearman's correlation coefficients were used, we use here Pearson's coefficients as, instead of focusing on ranking; this first evaluation exercise focuses on evaluating the significance and noisiness of the association, if any, between the automatic metrics and human-generated scores.

Three parameters should be adjusted for the AM-FM implementation described in (1): the dimensionality of the reduced space for AM, the order of n-gram model for FM, and the harmonic mean weighting parameter α. Such parameters can be adjusted for maximizing the correlation coefficient between the AM-FM metric and human-generated scores.[4] After exploring the solution space, the following values were selected, dimensionality for AM: *1,000*; order of n-gram model for FM: *3*; and, weighting parameter α: *0.30*

In the comparative evaluation presented here, correlation coefficients between the automatic metrics and human-generated scores were computed at the system level (i.e. the units of analysis were system outputs), by considering all 86 available system outputs (see Table 1). For computing human scores and AM-FM at the system level, average values of sentence-based scores for each system output were considered.

Table 2 presents the Pearson's correlation coefficients computed between the automatic metrics (BLEU, NIST, Meteor and our proposed AM-FM) and the human-generated scores (adequacy, fluency and the harmonic mean of both; i.e. $2af/(a+f)$). All correlation coefficients presented in the table are statistically significant with $p<0.01$ (where p is the probability of getting the same correlation coefficient, with a similar number of 86 samples, by chance).

Metric	Adequacy	Fluency	H Mean
BLEU	**0.4232**	**0.4670**	**0.4516**
NIST	0.3178	0.3490	0.3396
Meteor	0.4048	0.3920	0.4065
AM-FM	0.3719	0.4558	0.4170

Table 2: Pearson's correlation coefficients (computed at the system level) between automatic metrics and human-generated scores

As seen from the table, BLEU is the metric exhibiting the largest correlation coefficients with human-generated scores, followed by Meteor and AM-FM, while NIST exhibits the lowest correlation coefficient values. Recall that our proposed AM-FM metric is not using reference translations for assessing translation quality, while the other three metrics are.

In a similar exercise, the correlation coefficients were also computed at the sentence level (i.e. the units of analysis were sentences). These results are summarized in Table 3. As metrics are computed

[4] As no development dataset was available for this particular task, a subset of the same evaluation dataset had to be used.

at the sentence level, smoothed-bleu (Lin and Och, 2004) was used in this case. Again, all correlation coefficients presented in the table are statistically significant with $p<0.01$.

Metric	Adequacy	Fluency	H Mean
sBLEU	0.3089	**0.3361**	**0.3486**
NIST	0.1208	0.0834	0.1201
Meteor	**0.3220**	0.3065	0.3405
AM-FM	0.2142	0.2256	0.2406

Table 3: Pearson's correlation coefficients (computed at the sentence level) between automatic metrics and human-generated scores

As seen from the table, in this case, BLEU and Meteor are the metrics exhibiting the largest correlation coefficients, followed by AM-FM and NIST.

4.2 Reproducing Rankings

In addition to adequacy and fluency, the WMT-07 dataset includes rankings of sentence translations. To evaluate the usefulness of AM-FM and its components in a different evaluation setting, we also conducted a comparative evaluation on their capacity for predicting human-generated rankings.

As ranking evaluations allowed for ties among sentence translations, we restricted our analysis to evaluate whether automatic metrics were able to predict the best, the worst and both sentence translations for each of the 4,060 available rankings[5]. The number of items per ranking varies from 2 to 5, with an average of 4.11 items per ranking. Table 4 presents the results of the comparative evaluation on predicting rankings.

As seen from the table, Meteor is the automatic metric exhibiting the largest ranking prediction capability, followed by BLEU and NIST, while our proposed AM-FM metric exhibits the lowest ranking prediction capability. However, it still performs well above random chance predictions, which, for the given average of 4 items per ranking, is about 25% for best and worst ranking predictions, and about 8.33% for both. Again, recall that the AM-FM metric is not using reference translations, while the other three metrics are. Also, it is worth mentioning that human rankings were conducted by looking at the reference translations and not the source. See Callison-Burch *et al.* (2007) for details on the human evaluation task.

Metric	Best	Worst	Both
sBLEU	51.08%	54.90%	37.86%
NIST	49.56%	54.98%	37.36%
Meteor	**52.83%**	**58.03%**	**39.85%**
AM-FM	35.25%	41.11%	25.20%
AM	37.19%	46.92%	28.47%
FM	34.01%	39.01%	24.11%

Table 4: Percentage of cases in which each automatic metric is able to predict the best, the worst, and both ranked sentence translations

Additionally, results for the individual components, AM and FM, are also presented in the table. Notice how the AM component exhibits a better ranking capability than the FM component.

5 Conclusions and Future Work

This work presented AM-FM, a semantic framework for translation quality assessment. Two comparative evaluations with standard metrics have been conducted over a large collection of human-generated scores involving different languages. Although the obtained performance is below standard metrics, the proposed method has the main advantage of not requiring reference translations.

Notice that a monolingual version of AM-FM is also possible by using monolingual latent semantic indexing (Landauer *et al.*, 1998) along with a set of reference translations. A detailed evaluation of a monolingual implementation of AM-FM can be found in Banchs and Li (2011).

As future research, we plan to study the impact of different dataset sizes and vector space model parameters for improving the performance of the AM component of the metric. This will include the study of learning curves based on the amount of training data used, and the evaluation of different vector model construction strategies, such as removing stop-words and considering bigrams and word categories in addition to individual words.

Finally, we also plan to study alternative uses of AM-FM within the context of statistical machine translation as, for example, a metric for MERT optimization, or using the AM component alone as an additional feature for decoding, rescoring and/or confidence estimation.

[5] We discarded those rankings involving the translation system for which translation outputs were not available that, consequently, only had one translation output left.

References

Joshua S. Albrecht and Rebeca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 296-303.

Rafael E. Banchs and Haizhou Li. 2011. Monolingual AM-FM: a two-dimensional machine translation evaluation method. Submitted to the Conference on Empirical Methods in Natural Language Processing.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 65-72.

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 WS2003 CLSP Summer Workshop, Johns Hopkins University

Chris Callison-Burch, Cameron Fordyce,Philipp Koehn, Christof Monz and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In Proceedings of Statistical Machine Translation Workshop, 136-158.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the Human Language Technology Conference.

Susan Dumais, Thomas K. Landauer and Michael L. Littman. 1997. Automatic cross-linguistic information retrieval using latent semantic indexing. In Proceedings of the SIGIR Workshop on Cross-Lingual Information Retrieval, 16-23.

Michael Gamon, Anthony Aue and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: beyond language modeling. In Proceedings of the 10th Annual Conference of the European Association for Machine Translation, 103-111.

G. H. Golub and W. Kahan. 1965. Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics: Numerical Analysis, 2(2):205-224.

Thomas K. Landauer, Peter W. Foltz and Darrell Laham. 1998. Introduction to Latent Semantic Analysis. Discourse Processes, 25:259-284.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In Proceedings of the 20th international conference on Computational Linguistics, pp 501, Morristown, NJ.

Christopher D. Manning and Hinrich Schutze. 1999. Foundations of Statistical Natural Language Processing (Chapter 6). Cambridge, MA: The MIT Press.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jung Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the Association for Computational Linguistics, 311-318.

Christopher B. Quirk. 2004. Training a sentence-level machine translation confidence measure. In Proceedings of the 4th International Conference on Language Resources and Evaluation, 825-828.

Reinhard Rapp. 2009. The back-translation score: automatic MT evaluation at the sentences level without reference translations. In Proceedings of the ACL-IJCNLP, 133-136.

Gerard M. Salton, Andrew K. Wong and C. S. Yang. 1975. A vector space model for automatic indexing. Communications of the ACM, 18(11):613-620.

Harold Somers. 2005. Round-trip translation: what is it good for? In proceedings of the Australasian Language Technology Workshop, 127-133.

Lucia Specia, Craig Saunders, Marco Turchi, Zhuoran Wang and John Shawe-Taylor. 2009. Improving the confidence of machine translation quality estimates. In Proceedings of MT Summit XII. Ottawa, Canada.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing.

John S. White, Theresa O'Cornell and Francis O'Nava. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. In Proceedings of the Association for Machine Translation in the Americas, 193-205.