Adequacy–Fluency Metrics: Evaluating MT in the Continuous Space Model Framework

Rafael E. Banchs, *Member, IEEE*, Luis F. D'Haro, *Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—This work extends and evaluates a two-dimensional automatic evaluation metric for machine translation, which is designed to operate at the sentence level. The metric is based on the concepts of adequacy and fluency, aiming at decoupling both semantic and syntactic components of the translation process to provide a more balanced view on translation quality. These two elements are independently evaluated by using continuous space and n-gram language modeling frameworks, respectively. Two different implementations are evaluated: a monolingual version that fully operates on the target language side, and a cross-language version that has the main advantage of not requiring reference translations. Both implementations are evaluated by comparing their performance with state-of-the-art automatic metrics over a dataset involving five different European languages.**

***Index Terms*—Machine translation, natural language processing, system evaluation.**

I. Introduction

THE mathematical metaphor offered by the geometric concept of distance in continuous space with respect to semantics and meaning has been proven to be useful in many monolingual and cross-language natural language processing applications such as document classification [1], information retrieval [2], word similarity [3], and phrase modeling in machine translation [4], [5], among others [6] and [7].

The evaluation of Machine Translation (MT) results has always been one of the major issues concerning researchers in this area, as both human and automatic evaluation methods exhibit very important limitations.

Human evaluation, although highly reliable, suffers from inconsistency problems due to both inter- and intra-annotator agreement issues [8]. In addition, this type of evaluation is very expensive and time consuming. On the other hand, although more consistent, cheaper and much faster, automatic evaluation methods have the major disadvantage of requiring human-generated translation references. This makes automatic evaluation not reliable in the sense that a good translation hypothesis that does not match the available reference will be actually scored as a poor or bad translation.

Different from automatic evaluation metrics, which heavily rely on comparing translation results with a set of references, humans can produce absolute evaluations based on more meaningful metrics such as adequacy and fluency [9]. While adequacy focuses on the problem of measuring how much of the meaning is preserved during translation, fluency focuses on the problem of measuring the quality of the target language construction. In this sense, the adequacy and fluency scores can be considered as proxies to the semantic and syntactic appropriateness of a translation result, respectively.

The main objective of this work is to extend and evaluate our recently proposed automatic evaluation metric AM-FM (Adequacy Metric - Fluency Metric) [10], which was designed to access translation quality by independently addressing the semantic and syntactic dimensions of the translation process. In this way, AM-FM constitutes a two-dimensional metric that accounts for the two aforementioned dimensions of translation quality in an independent manner: a continuous space model framework is used for assessing adequacy, and an n-gram language model framework is used for assessing fluency. These two components are evaluated at the sentence level, making AM-FM a sentence-based metric by design, while still maintaining a reasonable[1] correlation with human-generated quality assessments.

In this work, two alternative versions of AM-FM are implemented and evaluated. In the first version, a monolingual implementation that operates over the target language alone is described. In this case, a monolingual continuous space model is used to compare translation outputs against translation references for the adequacy-oriented component of the metric. This provides a fairer basis for comparing the proposed metric with other currently existing metrics, which extensively use translation references. In the second version, a cross-language continuous space model is used for assessing adequacy by directly comparing translation outputs with their corresponding source inputs, which are used as evaluation reference.

Different from our original presentation of the metric [10], where only the cross-language setting was considered, in this work we focus our attention on the properties of AM-FM in the

Manuscript received July 15, 2014; revised December 10, 2014; accepted February 17, 2015. Date of current version February 26, 2015. This work was supported by the Human Language Technology Department of the Institute for Infocomm Research (I²R). The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Bowen Zhou.

R. E. Banchs and L. F. D'Haro are with the Human Language Technology Department, I²R, Singapore 138632 (e-mail: rembanchs@i2r.a-star.edu.sg; luisdhe@i2r.a-star.edu.sg).

H. Li is with the Human Language Technology Department, I²R, Singapore 138632, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: hli@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2405751

[1]By 'reasonable' we mean better than or, at least, similar to those correlations observed between human evaluations and currently available automatic evaluation metrics.

monolingual setting, in which translation references are available. Additionally, different from the weighted harmonic mean used in [10], we explore two alternative strategies for combining the AM and FM components into a single metric. More specifically, we consider the weighted mean and the weighted $L2$-norm as alternative combination strategies.

The rest of the paper is organized as follows. In Section II, background work on MT automatic evaluation metrics is presented along with the specific dataset to be used. In Section III, details about AM-FM and its computational implementation are provided. In Section IV, comparative studies on how AM-FM correlates with human evaluations are presented; followed by an analysis, in Section V, on how sensitive the metric and its meta-parameters are to the training data variations. Finally, in Section VI, conclusions and open questions for future research are presented.

II. Related Work and Dataset

For several years BLEU [11] has been the most commonly used metric for machine translation evaluation. Other metrics such as WER, PER [12], NIST [13] and, more recently, Meteor [14], TER [15] and MEANT [16], [17] have been commonly used too. Indeed, a large amount of good work is available on the problem of machine translation evaluation [18], [19], and [20]. This group of metrics assumes availability of human translated references and attempts to compare the translation outputs against the references in different ways.

Some recent work has also been focused on the problem of MT confidence estimation and the specific idea of evaluating translation outputs without translation references. In this sense, several different approaches have been proposed and evaluated such as round-trip translation [21] and [22], cross-language semantic-frame base evaluation XMEANT [23], as well as some regression- and classification-based approaches [24], [25], [26], [27], [28] and [29]. Instead of using human translated references, this group of metrics leverages on knowledge captured by other natural language processing or machine translation systems, therefore, their performance depends on how good those systems are.

Recently, there is also a common tendency to use either semi-automated evaluation procedures by incorporating humans in the loop, such as in the case of metrics like HTER [15] and HMEANT [30], [31], or human-based comparative evaluations [32], which are easier than absolute evaluations and exhibit better intra- and inter-annotator agreements.

Table I summarizes the main characteristics of the most commonly and currently used evaluation metrics, along with both, the monolingual (mAM-FM) and cross-language (xAM-FM) implementations of the AM-FM evaluation framework.

As part of the efforts on machine translation evaluation, two workshops have been organizing shared-tasks and evaluation campaigns over the last few years: the NIST Metrics for Machine Translation Challenge[2] (MetricsMATR) and the Workshop on Statistical Machine Translation[3] (WMT); which were held as one single event for the first time in 2010.

The dataset used in the experimental part of this work corresponds to the dataset of the 2007's edition of WMT. The reasons for using such a dataset instead of a more recent one are basically two: human evaluation data is not freely available for the case of MetricsMATR, and no human judgments on adequacy and fluency have been conducted in WMT after year 2007.[4]

TABLE I
Summary of Main Currently Available Evaluation Metrics

Assessment Level	Needs for References	Cross-Language	Humans in the Loop
Words	WER, PER	–	–
Word *n*-grams	BLEU, NIST	–	–
Stems & Synonyms	METEOR	–	–
Edit Distances	TER	–	HTER
Semantic Roles	MEANT	XMEANT	HMEANT
Continuous Space	*m*AM-FM	*x*AM-FM	–

The dataset used in this work includes translation outputs, training data and reference translations for fourteen tasks involving five different European languages and two different domains. The languages are English (EN), German (DE), Czech (CZ), French (FR) and Spanish (ES); and the domains are News Commentaries (News) and European Parliament Plenary Sessions (EPPS). A complete description on the WMT-07 evaluation campaign and dataset is available in [19]. Here, we will only provide details on those specific aspects of the dataset that are relevant to our experimental work.

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

For removing individual voting patterns, as well as averaging votes (in those cases where more than one human evaluation was conducted for the same sentence translation), we used the vote standardization procedure described in [8]. The process was applied to all adequacy and fluency scores. Table II provides information on the corresponding domain, source language and target language 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.

III. The AM-FM Evaluation Metric

As already mentioned in the introduction, the AM-FM evaluation framework is intended to access translation quality along two different dimensions: semantics and syntax. The two-component evaluation metric is based on the concepts of adequacy and fluency [9]. While adequacy accounts for the amount of source information (meaning) preserved in the translation, fluency accounts for the quality of the target language constructions used in the translation. As adequacy and fluency are directly related to semantics and syntax, respectively, we can say

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

[3]http://www.statmt.org/wmt14/

[4]As the AM-FM evaluation framework is based on the concepts of adequacy and fluency, we need to evaluate its performance against human judgments on these two concepts. Then, the only freely available dataset that is suitable for our purposes is WMT-07 (http://www.statmt.org/wmt07/results.html)

TABLE II
THE FOURTEEN TRANSLATION TASKS INCLUDED IN THE WMT-07 DATASE

Task	Domain	Source	Target	Systems	Sentences
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

Domain, source language, target language, system outputs and total amount of sentence translations (with both, adequacy and fluency, human assessments available).

TABLE III
FIVE-POINT SCALES AND DEFINITION GUIDELINES FOR HUMAN ASSESSMENTS ON ADEQUACY AND FLUENCY

Metric	Score	Definition
ADEQUACY	1	None of the meaning is preserved
	2	Little of the meaning is preserved
	3	Much of the meaning is preserved
	4	Most of the meaning is preserved
	5	All the meaning is preserved
FLUENCY	1	Incomprehensible target language
	2	Disfluent target language
	3	Non-native kind of target language
	4	Good quality target language
	5	Flawless target language

that the proposed AM-FM metric aims at evaluating both components independently. However, in practice, these two components are strongly correlated.

Human evaluators typically assess adequacy and fluency by using a five-point scale, which is depicted in Table III. However, after the vote standardization procedure (see Section 5.4 of [8]), both metrics are rescaled to the [0, 1] interval.

The AM-FM metric was originally proposed as a means to assess translation quality in such scenarios in which translation references were not available [10], while maintaining consistency with human quality assessments. Different from round-trip and regression-based evaluations, AM-FM relies on a continuous space model to define a metric able to assess semantic similarities at the sentence level, within both the monolingual and the cross-language settings.

In the next sub-sections, we describe in detail the implementation of both, the monolingual and the cross-language versions of the AM-FM evaluation metric.

A. Metric Definition

For implementing the adequacy-oriented component of the metric (referred to as the AM component), a Latent Semantic Indexing [33] approach is used. This approach allows for both a monolingual implementation, in which reference translations are required, and a cross-language implementation [34], in which the source sentences originating the translation are used as the evaluation reference.

In both cases, adequacy is measured by means of a distance metric in a low-dimensional continuous space, in which sentences are modeled as bag-of-words [35]. According to this kind of representation, the AM component can be regarded to be pure adequacy-oriented as no relevant information on word ordering is taken into account.

On the other hand, for implementing the fluency-oriented component of the proposed metric (referred to as the FM component), an n-gram based language model approach is used [36]. This component can be regarded to be pure fluency-oriented, as it is computed on the target language side in a manner that is totally independent from the source language.

For combining the two components into a single evaluation metric, three different schemes are considered and evaluated: a weighted harmonic mean, a weighted mean, and a weighted $L2$-norm,

$$AM\text{-}FM_{HM} = \frac{AM \times FM}{(\alpha AM + (1-\alpha)\ FM)} \tag{1}$$

$$AM\text{-}FM_{WM} = (1-\alpha)\, AM + \alpha FM \tag{2}$$

$$AM\text{-}FM_{L2} = \sqrt{(1-\alpha)\ AM^2 + \alpha FM^2} \tag{3}$$

where, in all three cases, α is a weighting factor which should be in the range from $\alpha = 0$ (pure AM component) to $\alpha = 1$ (pure FM component), and can be adjusted to maximize the correlation between the proposed metric and human evaluation scores.

B. Adequacy Metric (AM) Implementation Details

As already mentioned, the adequacy-oriented component of the metric AM is implemented by means of Latent Semantic Indexing [33], which is based on the well-known singular value decomposition of a rectangular matrix [37]. Basically, the Latent Semantic Indexing algorithm exploits the fact that a term-document matrix $\mathbf{X}$ [35] of dimensions $M \times N$, where M and N are the number of vocabulary terms and documents (sentences in our case), can be factorized as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{4}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices of dimensions $M \times M$ and $N \times N$, respectively, and $\mathbf{\Sigma}$ is an $M \times N$ diagonal matrix containing the singular values associated to the decomposition.

According to [33], a low-dimensional representation of a given document (sentence in our case) can be obtained from the singular value decomposition depicted in (4) as follows:

$$\mathbf{y}^T = \mathbf{x}^T \mathbf{U}_{M \times L} \tag{5}$$

where $\mathbf{y}$ represents the L-dimensional vector corresponding to the projection of a full-dimensional sentence vector $\mathbf{x}$, and $\mathbf{U}_{M \times L}$ is the projection matrix, which is composed of the L first column vectors of the unitary matrix $\mathbf{U}$ obtained in (4). This kind of rank reduction has been proven to preserve the most important semantic information in the collection of documents while reducing some of the noise present in the sparse full-rank representation.

The AM component of the monolingual version of the metric is then computed in the projected space by calculating the cosine

similarity between projected sentences. More specifically, similarities between translation outputs and their corresponding reference translations are considered. According to this, the monolingual AM score (referred to as mAM) is computed as follows:

$$mAM = \frac{\mathbf{r}^T\mathbf{U}_{M\times L}\left(\mathbf{t}^T\mathbf{U}_{M\times L}\right)^T}{|\mathbf{r}^T\mathbf{U}_{M\times L}|\,|\mathbf{t}^T\mathbf{U}_{M\times L}|} \tag{6}$$

where $\mathbf{U}_{M\times L}$ is the monolingual projection matrix, $\mathbf{r}$ and $\mathbf{t}$ are corresponding vector representations of the reference and translation output being compared, and $||$ is the $L2$-norm operator.

In a final implementation stage, the range of possible values for mAM is restricted to the interval [0, 1] by truncating those negative values resulting from vectors that are more than *90* degrees apart in the resulting low-dimensional embedding.

The same Latent Semantic Indexing methodology can be extended to the cross-language case [34], where the main difference is that the term-document matrix $\mathbf{X}$ is constructed with parallel documents (in our case sentences) in two languages. In the cross-language case, the singular value decomposition depicted in (4) can be reformulated as follows:

$$[\mathbf{X}_a;\mathbf{X}_b] = \mathbf{U}_{ab}\mathbf{\Sigma}_{ab}{\mathbf{V}_{ab}}^T \tag{7}$$

where $[\mathbf{X}_a;\mathrm{X}_b]$ is a bilingual term-document matrix obtained by concatenating two monolingual term-document matrices for a parallel document collection. In this case, the low-dimensional projections for two given sentence vectors $\mathbf{x}_a$ and $\mathbf{x}_b$, in languages a and b, can be obtained as follows:

$${\mathbf{y}_a}^T = [\mathbf{x}_a;\mathbf{0}]^T\mathbf{U}_{ab\ M\times L} \tag{8a}$$

$${\mathbf{y}_b}^T = [\mathbf{0};\mathbf{x}_b]^T\mathbf{U}_{ab\ M\times L} \tag{8b}$$

where $\mathbf{y}_a$ and $\mathbf{y}_b$ represent the L-dimensional vectors corresponding to the projections of the full-dimensional vectors $\mathbf{x}_a$ and $\mathbf{x}_b$, respectively, and $\mathbf{U}_{abM\times L}$ is the cross-language projection matrix composed of the L first column vectors of the unitary matrix $\mathbf{U}_{ab}$ obtained in (7).

Notice, from (8a) and (8b), that both sentence vectors $\mathbf{x}_a$ and $\mathbf{x}_b$ are padded with zeros, at each of the corresponding other-language-vocabulary locations, before performing the projections. As similar terms in different languages would have similar distributions, 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 resulting low-dimensional cross-language continuous space.

Similar to the monolingual version of the metric, the AM component of the cross-language version is computed in the projected space by means of the cosine similarity between projected sentences. However, in this case, projections of the translation outputs and the source sentences being translated are considered. According to this, the cross-language AM score (referred to as xAM) is computed as follows:

$$xAM = \frac{[\mathbf{s};\mathbf{0}]^T\mathbf{U}_{ab\ M\times L}\left([\mathbf{0};\mathbf{t}]^T\mathbf{U}_{ab\ M\times L}\right)^T}{\left|[\mathbf{s};\mathbf{0}]^T\mathbf{U}_{ab\ M\times L}\right|\left|[\mathbf{0};\mathbf{t}]^T\mathbf{U}_{ab\ M\times L}\right|} \tag{9}$$

where $\mathbf{U}_{abM\times L}$ is a truncated version of the cross-language projection matrix computed in (7), $[\mathbf{s};\mathbf{0}]$ and $[\mathbf{0};\mathbf{t}]$ are vector space representations of the source and translation sentences being compared (with their target and source vocabulary elements set to zero, respectively), and $||$ is the $L2$-norm operator. Again, the range of possible values for xAM is restricted to the interval [0, 1] by truncating all resulting negative cosine similarities.

For computing all required projection matrices $\mathbf{U}_{M\times L}$ and $\mathbf{U}_{\mathbf{ab}\,M\times L}$, *10,000* parallel sentences[5] were randomly drawn from the available training datasets and used for constructing the corresponding term document matrices $\mathbf{X}$ and $[\mathbf{X_a};\mathbf{X_b}]$. The only restriction we imposed to this sentence random selection process was that each of the extracted sentences should contain at least *10* words.

Fourteen projection matrices were constructed in total for the monolingual implementation of AM, one for each task; and seven projection matrices were constructed for the cross-language implementation, one for each combination of domain and language pair. All computations related to singular value decompositions, sentence projections and cosine similarities were conducted with Python's NumPy (www.numpy.org).

C. Fluency Metric (FM) Implementation Details

The fluency-oriented component FM, which is the same for both the monolingual and the cross-language versions of the metric, was implemented by means of an n-gram language model. In order to compensate for 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:

$$FM = \exp\left(\frac{1}{N}\sum\nolimits_{n=1:N}\log\left(p(w_n|w_{n-1},\ldots)\right)\right) \tag{10}$$

where $p(w_n|w_{n-1},\ldots)$ represents 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 components AM and FM range within the same interval. Fourteen language models were trained in total, one per task, by using all available training data for each task and language. The models were computed with the SRILM toolbox [38].

As seen from (6), (9) and (10), different from other conventional metrics that compute matches between translation outputs and references, in the AM-FM framework, a semantic continuous space embedding is used for assessing the similarities between outputs and inputs (9), or outputs and reference translations (6), and, independently, an n-gram model is used for evaluating output language quality (10).

According to this, the prior knowledge involved in the AM-FM evaluation framework is actually separated by the two components of the metric. While syntax level information is

[5]Although this represents a very small proportion of the datasets (20% of News and 1% of European Parliament), it allowed for maintaining computational requirements bounded while still providing good vocabulary coverage. Similar experiments were conducted with sets of *5 K* and *15 K* parallel sentences and the observed relative variations in the correlation coefficients computed in Section IV were smaller than 5% (see Section V-A for more details).

mainly captured by the fluency-oriented component FM by means of a target language model, semantic level information is mainly captured by the adequacy-oriented component AM by means of a semantic continuous space model. Furthermore, both components can be adaptively combined in several different ways, as already described in (1), (2) and (3), for improving its correlation with human-generated scores.

IV. Correlation with Human Evaluations

In order to evaluate the performance of both the mono-lingual mAM-FM and the cross-language xAM-FM metrics, a comparative analysis was conducted considering the fourteen tasks presented in Table II. For this comparative evaluation, BLEU, NIST, TER[6] [39] and Meteor were used as reference metrics. We selected these specific metrics, as they constitute the most commonly used automatic evaluation metrics in machine translation evaluation campaigns.

A. Parameter Selection

Three fundamental parameters should be adjusted for any of the implementations described in equations (1), (2) and (3). They are: the dimensionality of the reduced space for AM, the order of n-gram for FM, and the weighting parameter α.

In our conception of the AM-FM metric, these parameter values should be selected for maximizing the correlation between the metric and the observed human evaluations. According to this, we use the Pearson's correlation coefficient as objective function, which is computed over the subset of system outputs, for which translations were evaluated for both adequacy and fluency by human judges.

As optimizing all three parameters simultaneously can be troublesome, we will present here some results corresponding to the variation of only two parameters at a time. Indeed, we will optimize the reduced space dimensionality looking only at the behavior of the correlation coefficients when jointly varying the dimensionality and the weighting parameter α. The order of the language model, on the other hand, will be optimized independently from the weighting parameter α, and a more detailed evaluation on the effects of the weighting parameter α will be performed in the next sub-section.

Different from [19], where Spearman's correlations were used, we use the Pearson's correlation coefficient here. This is basically because, instead of focusing on ranking, we are much more interested in evaluating the significance and noisiness of the association, if any, between the proposed AM-FM metric and the human-generated scores. All correlations presented in this work are computed using system level scores (i.e. units of analysis are system outputs). According to this, each coefficient is computed over a data sample of *86* points, which corresponds to the *86* system outputs described in the fifth column of Table II. As human-generated and AM-FM scores are both sentence-based scores, average values are computed to obtain scores at the system level.

[6]More specifically, we use TER-Plus (TERp), a version known for better correlating with human judgements. This version of TER is available online at: http://www.umiacs.umd.edu/~snover/terp/

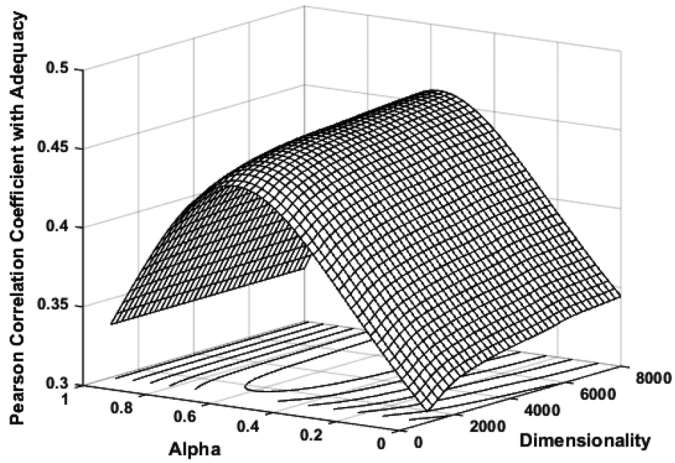

Fig. 1. Pearson's correlation coefficients between the weighted mean implementation of m AM-FM and human-generated adequacy scores for different values of the weighting parameter α and mAM dimensionality.

To see the combined effects of the weighting parameter α and the reduced-space dimensionality for the monolingual implementation of the AM-FM metric, Fig. 1 shows the resulting Pearson's correlation coefficients between mAM-FM and human-generated scores for adequacy. For the specific result illustrated in Fig. 1, the weighted mean version of the metric, as defined in equation (2), has been used.[7]For the FM component, a *5*-gram language model was used.

All the correlation coefficients depicted in Fig. 1 are statistically significant with $p < 0.01$, p being the probability of getting the same correlation coefficient, with a similar number of samples, by pure chance.

Notice from the figure that the best correlation coefficient values are consistently achieved for all dimensionalities when the weighting parameter α is around *0.6*. By looking at the correlation coefficient values across the dimensionality axis, it can be noticed that, different from other type of applications (such as text classification and information retrieval) in which optimal dimensionalities typically range from *400* to *800*, the values of the correlation coefficients continue to increase as the dimensionality value approaches its maximum possible value of *10,000* (full-rank projections).

However, it can be confirmed that these values actually reach a plateau for dimensionalities around *7,500* and above. According to this, we selected as the optimal dimensionality value for computing the mAM component of the monolingual version of the metric the value of *7,600* dimensions.

Finally, although statistically significant, the resulting correlation values are low. This is basically due to the very noisy association between the AM metric and the human-generated scores. This noisy association mainly results from two facts: first, human judgments are intrinsically inconsistent as they exhibit low rates of both intra- and inter-annotator agreement [19]; second, human judgments do not actually reflect a pure adequacy assessment due to intrinsic human's difficulty to

[7]Correlation surfaces for the weighted harmonic mean and weighted $L2$-norm implementations look similar, with the only difference that optimal α values vary significantly from one implementation to another. This will be explored in more detail in the next sub-section.

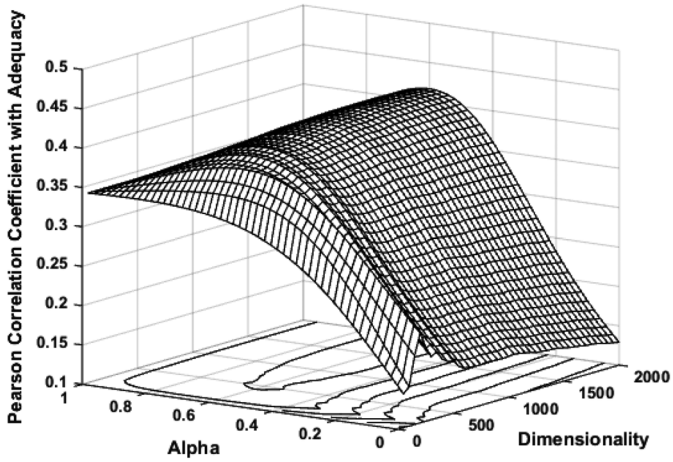

Fig. 2. Pearson's correlation coefficients between the weighted mean implementation of xAM-FM and human-generated adequacy scores for different values of the weighting parameter α and x AM dimensionality.

effectively separate adequacy from fluency (more on this in Section IV-C).

To see the effects of reduced-space dimensionality selection in the cross-language implementation of the AM-FM metric, Fig. 2 shows the resulting Pearson's correlation coefficients between xAM-FM and human-generated scores for adequacy. For the specific result illustrated in Fig. 2, the weighted mean version of the metric, as defined in equation (2), has been used[8] and, again, a *5*-gram language model was used for the FM component implementation.

Notice from the figure how, in this case, the general behavior of the correlation coefficient across the parameter space is similar to the monolingual case. However, two important differences can be observed. First, the starting values of the correlation values for the xAM component ($\alpha = 0$) are much lower than in the monolingual case. Although the optimal combination remains at the region around $\alpha = 0.6$, in this case the overall correlation response seems to be mostly driven by the FM component.

Second, if we pay detailed attention to the correlation values for the pure AM component ($\alpha = 0$), different from the monolingual setting, the correlation coefficients exhibit a maximum at a very low dimensionality value, which is around *200* dimensions. However, if we look at the behavior of the correlation coefficients across this fixed dimensionality value, we can observe that its maximum value is not precisely the best correlation coefficient in the parameter space.

Indeed, similar to the monolingual case, it can be confirmed that the correlation of the xAM-FM combination reaches a plateau at larger dimensionality values, which in this case are around *1,500*. According to this, we selected *2,000* as the optimal dimensionality for computing the xAM component of the cross-language version of the metric.

[8]Correlation surfaces for the weighted harmonic mean and weighted $L2$-norm implementations look similar, with the only difference that optimal α values vary significantly from one implementation to another. This will be explored in more detail in the next sub-section.

TABLE IV
PEARSON'S CORRELATION COEFFICIENTS BETWEEN THE FM COMPONENT AND HUMAN-GENERATED SCORES FOR DIFFERENT ORDERS OF THE n-GRAM LANGUAGE MODEL

Order	Adequacy	Fluency
2-gram	0.3346	0.4225
3-gram	0.3383	0.4245
4-gram	0.3393	0.4251
5-gram	0.3408	**0.4267**
6-gram	**0.3409**	0.4266

All coefficients are significant with $p < 0.01$

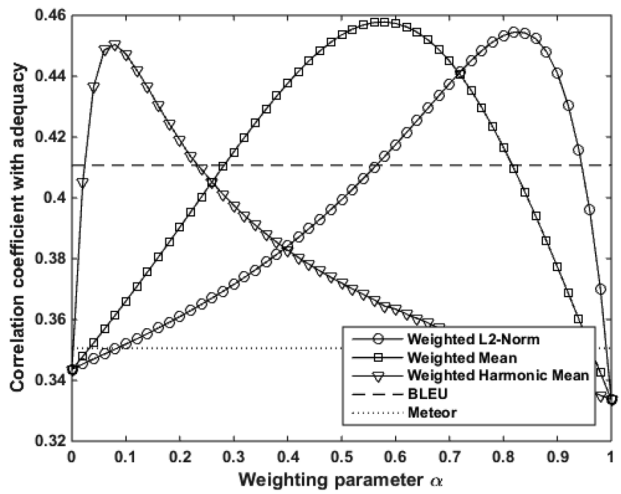

Fig. 3. Pearson's correlation coefficients between the three monolingual implementations of mAM-FM and human-generated scores for adequacy.

To evaluate the effects of n-gram model size, Pearson's correlation coefficients between the FM component alone, as defined in equation (10), and the human-generated scores for both adequacy and fluency were computed for different orders of n-gram models. More specifically, language model orders from *2* to *6* were considered. The obtained results are summarized in Table IV.

As seen from the table, the best correlations are achieved for $n = 5$ and $n = 6$. As no significant improvement is actually observed when increasing the model order from *5* to *6*, we selected *5* as the optimal value for the FM component. Finally, as expected, the FM component correlates better with human-generated fluency than with human-generated adequacy.

B. Comparative Evaluations

For the comparative evaluations, as already mentioned, we selected BLEU, NIST, TER and Meteor, as they are the automatic evaluation metrics that are most commonly used in machine translation evaluation campaigns.

First, we evaluated the performance of the monolingual version of the AM-FM metric using all three implementations defined in (1), (2) and (3). For this analysis, optimal values of *7,600* for dimensionality reduction and *5* for n-gram model size were used, while the effect of varying the weighting parameter α was evaluated in more detail.

Figs. 3 and 4 illustrate the resulting Pearson's correlation coefficients between the three implementations of mAM-FM with

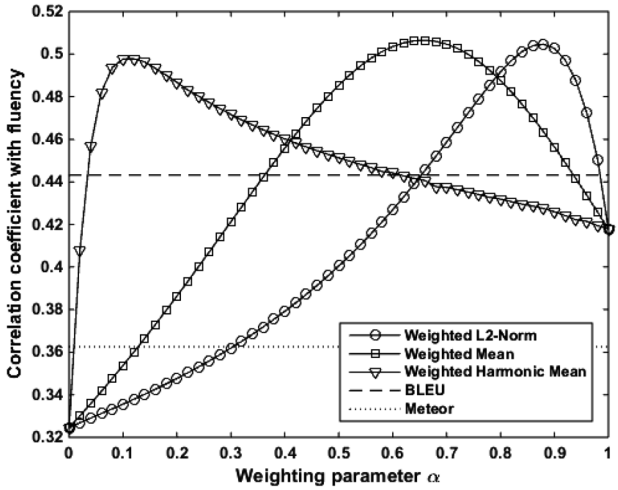

Fig. 4. Pearson's correlation coefficients between the three monolingual implementations of mAM-FM and human-generated scores for fluency.

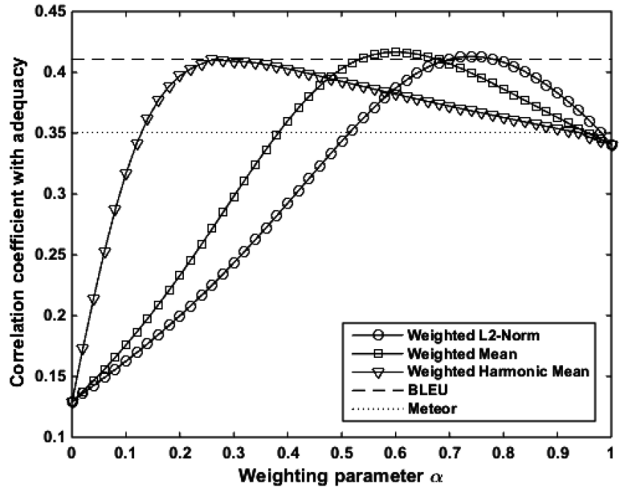

Fig. 5. Pearson's correlation coefficients between the three cross-language implementations of xAM-FM and human-generated scores for adequacy.

TABLE V
PEARSON'S CORRELATION COEFFICIENTS BETWEEN AUTOMATIC METRICS AND HUMAN-GENERATED SCORES FOR ADEQUACY AND FLUENCY

Metric	α	Adequacy	Fluency
BLEU	–	0.4107	0.4432
Meteor	–	0.3505	0.3626
NIST	–	0.3226	0.3444
TER-Plus	–	0.3068	0.3170
mAM	–	0.3435	0.3245
xAM	–	0.1291*	0.0330*
FM	–	0.3408	0.4267
mAM-FM $_{HM}$	0.10	0.4473	0.4977
mAM-FM $_{WM}$	0.60	**0.4574**	0.5036
mAM-FM $_{L2}$	0.86	0.4523	**0.5040**
xAM-FM $_{HM}$	0.30	0.4091	0.4503
xAM-FM $_{WM}$	0.60	0.4167	0.4442
xAM-FM $_{L2}$	0.80	0.4084	0.4493

All coefficients (except those marked with '$*$') are significant with $p < 0.01$

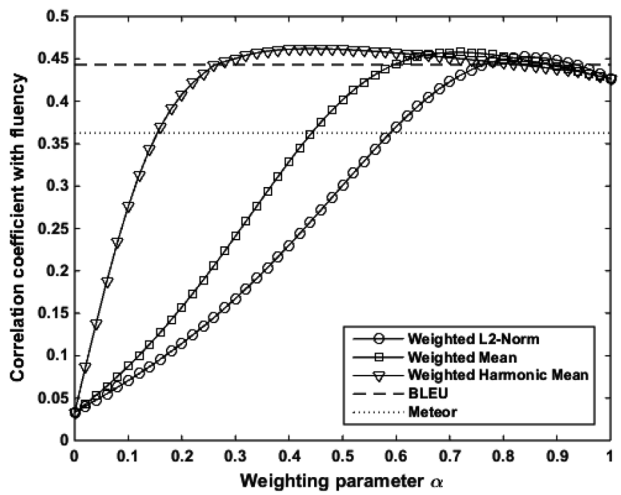

Fig. 6. Pearson's correlation coefficients between the three cross-language implementations of xAM-FM and human-generated scores for fluency.

human-generated scores for adequacy and fluency, respectively. The corresponding correlation coefficient values for BLEU and Meteor are also presented in the figures.[9]

Next we evaluated the performance of the cross-language version of the AM-FM metric. Figs. 5 and 6 illustrate the correlation coefficients between the three implementations of xAM-FM and human-generated adequacy and fluency, respectively. In these cases, optimal values of *2,000* and *5* for dimensionality and n-gram order were used, respectively.

Notice, from Figs. 3 and 4, how for some intervals of the weighting parameter α, the monolingual implementations of AM-FM correlates better with both human-generated scores than state-of-the-art evaluation metrics. However, as seen from Figs. 5 and 6, the correlation coefficients for the cross-language implementations are slightly lower. As both versions of AM-FM use the same FM component, it can be concluded that this performance drop is basically due to the AM component of the metric, which is actually worse in the case of the cross-language setting. Indeed, this correlation drop represents the price paid for not using reference translations and reflects the increased difficulty of assessing semantic similarity within a cross-language continuous space model with respect to the case of a monolingual one.

Also, as seen from all figures, the effect of the weighting parameter α is different for each of the three considered AM-FM combination strategies. While the weighted harmonic mean attains its maxima for low values of α, the weighted mean and $L2$-norm attain their maxima at intermediated and large values of α. It can also be noticed that the weighted mean and the weighted $L2$-norm strategies seem to be performing slightly better than the weighted harmonic mean in most of the considered scenarios.

Moreover, comparing Figs. 3 and 4, as well as Figs. 5 and 6, it can be seen that the optimal value for the combination parameter α are slightly different for the cases of adequacy and fluency. This means that a trade-off exists when the correlation coefficients are to be optimized simultaneously for both adequacy and fluency. However, it can also be seen that such dif-

[9]Only BLEU and Meteor are reported in the figures, as these are the best performing reference metrics. Correlation coefficient values for NIST and TER are reported in Table V.

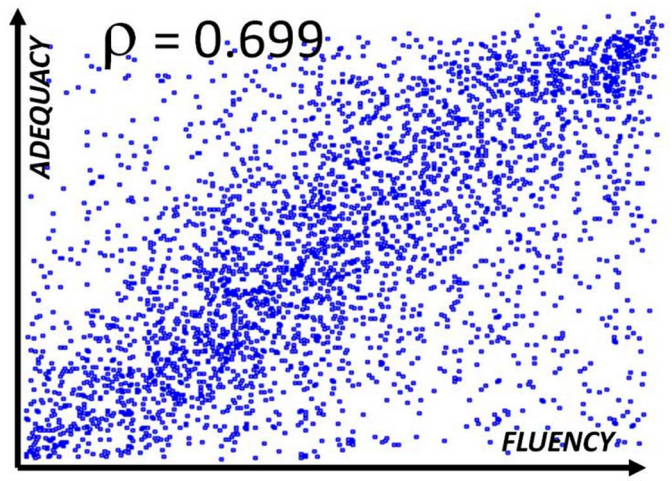

Fig. 7. Cross-plot of human-generated accuracy and fluency for the *10,754* translations that were rated for both adequacy and fluency by human judges. (The resulting Pearson's correlation coefficient is *0.699*).

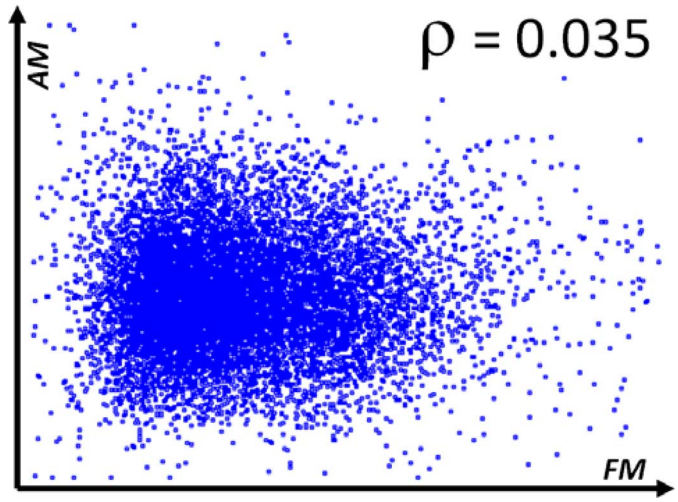

Fig. 8. Cross-plot of AM and FM metrics computed for the *10,754* translations that were rated for both adequacy and fluency by human judges. (The resulting Pearson's correlation coefficient is *0.035*).

ferences are relative small and a good suboptimal value for α can be easily set.

Table V summarizes some of the obtained correlation coefficients between the considered automatic metrics and the human-generated scores for adequacy and fluency. Some selected suboptimal values for the weighting parameter α are considered in the cases of AM-FM metrics.

C. Further Analysis and Discussion

As already mentioned before, one of the main problems affecting human-based judgments on adequacy and fluency is that they do not actually reflect a purely independent adequacy or fluency assessments. This is basically due to the intrinsic human's difficulty to effectively separate information related to adequacy from information related to fluency when performing assessments. As a result, human-generated adequacy and fluency are highly correlated.

To better illustrate this point, Fig. 7 depicts a cross-plot between adequacy and fluency judgments for the *10,754* translation outputs that were annotated by human judges. As seen from the figure, a clear trend of dependence between adequacy judgments and fluency judgments is observed.

The resulting correlation coefficient between the human-generated adequacy and fluency scores depicted in Fig. 7 is *0.699*, which can be considered to reflect a high association.

On the other hand, Fig. 8 presents the cross-plot between the AM and the FM metric components computed over the same set of *10,754* translation outputs that were annotated by humans. As seen from the figure, the dependence between these two metric components is practically inexistent. Indeed, the resulting correlation coefficient for this case is *0.035*, which is evidence of very low or inexistent association.

V. Sensitiveness to Training Data Variations

In this section, we explore in more detail how sensitive the proposed AM-FM metric is to variations of the training data. First, we evaluate the differences in performance for both metric components, in terms of their correlations with human-generated scores, when different data subsets are used for training purposes. Second, we evaluate the dependence of the AM-FM meta-parameter optimal values on different training data conditions.

TABLE VI
Pearson's Correlation Coefficients Between AM and Human-Generated Adequacy Scores for Varying Sizes of Training Data, and Their Relative Variations with Respect to the *10 K* Base Case

	1K	*5K*	*10K*	*15K*
*m*AM	0.3219	0.3291	0.3409	0.3423
Variation	−5.57%	−3.46%	–	+0.41%

A. Component Performance Sensitiveness

Regarding the sensitiveness of the AM and FM components with respect to training data variations, we focus our attention on three scenarios: the size of the dataset used to compute the AM projection matrix, the random subset selection for AM projection matrix, and the in-domain versus out-of-domain data proportion used for training the FM's n-gram model.

In the first scenario, we compare the performance of the AM component for four different sizes of the training dataset: 1 K,5 K,10 K and 15 K sentences. Table VI depicts the obtained correlation coefficients, between the pure AM component and human-generated adequacy, as well as their relative variations with respect to the 10 K case. For each result in the table, the dimensionality was set to be *70*% of the full-rank one. The analysis was conducted for the monolingual case (mAM) only.

As seen from the table, while a *5.5*% reduction of the correlation coefficient is observed for the 1 K case, the observed differences are smaller for the 5 K and 15 K cases. Indeed, as previously mentioned in Sub-Section III-B, observed differences for the 5 K and 15 K cases are smaller than 5%.

In the second scenario, we compare the performance of the AM component across different subsets of training data. More specifically, we consider five 10 K-sentence data folds, which have been randomly extracted from the available WMT-07 training data. Table VII presents the resulting correlations,

TABLE VII
PEARSON'S CORRELATION COEFFICIENTS BETWEEN AM AND HUMAN-GENERATED SCORES FOR ADEQUACY FOR DIFFERENT DATA FOLDS OF *10 K* SENTENCES, ALONG WITH MEAN AND STANDARD DEVIATION VALUES

	Fold-1	*Fold-2*	*Fold-3*	*Fold-4*	*Fold-5*	*Mean*	*Stdev*
*m*AM	0.3435	0.3390	0.3498	0.3404	0.3427	0.3431	0.0041
*x*AM	0.1291	0.1267	0.1205	0.1346	0.1153	0.1252	0.0075

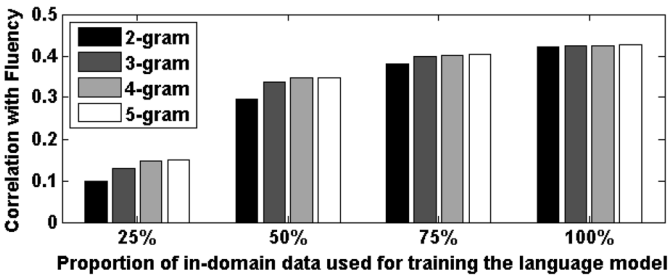

Fig. 9. Pearson's correlation coefficients between the FM metric component and human-generated scores for fluency, when different proportions of in-domain and out-of-domain data are used to train the n-gram language models. Four different models (*2*-gram, *3*-gram, *4*-gram and *5*-gram) are considered.

between the pure AM components and human-generated adequacy, for each of the five data folds under consideration, along with the corresponding average and standard deviation values (the first data fold corresponds to the one used in all experiments described in Section IV). The dimensionality values used for computing these results are *7,600* and *2,000* for mAM and xAM, respectively.

As seen from the table, different random selections of the training datasets used to compute the AM projection matrix yield very similar results. Indeed, the observed standard deviations are as small as *1.2*% of the mean value for the case of mAM and *6.0*% for xAM.

In the last scenario considered in this sub-section, we evaluate the performance of FM, in terms of its correlation with human-generated fluency, when varying the proportion of in-domain versus out-of-domain data used for training the n-gram probabilities. Fig. 9 depicts the observed variations of the correlation coefficients for the cases of *2*-gram, *3*-gram, *4*-gram and *5*-gram models when different proportions of in-domain training data are used. As seen from the figure, the FM performance barely changes when *75*% of in-domain data is preserved, while degradation becomes significant when *50*% of out-of-domain data is introduced and very severe when *75*% of the data is out-of-domain.

B. Meta-parameter Sensitiveness

Regarding the sensitiveness of AM-FM meta-parameters to training data variations, we focus our attention on the weighting parameter α, as it is the only one significantly affected by varying data conditions. Indeed, as seen in Fig. 9, regardless of the proportion of out-of-domain data, the *5*-gram model consistently exhibits the best performance. Similarly, detailed exploration of AM dimensionalities result on wide ranges of optimal regions, such as those observed in Figs. 1 and 2.

To see how α is affected by data variations, we search for optimal values of it across three different scenarios: a sweep of *50* different dimensionalities around its optimal value and, the five 10 K-sentence data folds and four different proportions of in-domain versus out-of-domain data described in the previous sub-section. Table VIII presents the observed standard deviations (and their percentages) for the optimal values of α across the considered scenarios. Only results corresponding to the weighted mean implementation of AM-FM are presented. As seen from the table, α is mainly sensitive to variations of in-domain data proportions used for training FM. Also, more sensitiveness to dimensionality variations is observed for the cross-language version than for the monolingual one.

TABLE VIII
OBSERVED STANDARD DEVIATIONS FOR OPTIMAL VALUES OF α (AND THEIR PERCENTAGE VALUES) ACROSS THREE DIFFERENT DATA VARIATION CONDITIONS. BOTH ADEQUACY AND FLUENCY ARE CONSIDERED

		Dimensionality	*Folds*	*In- vs out-domain*
mAM-FM$_{\text{WM}}$	Adequacy	0.0078 (1.3%)	0.0089 (1.5%)	0.0443 (7.4%)
	Fluency	0.0085 (1.4%)	0.0089 (1.5%)	0.0365 (6.1%)
xAM-FM$_{\text{WM}}$	Adequacy	0.0129 (2.2%)	0.0089 (1.5%)	0.0443 (7.4%)
	Fluency	0.0225 (3.7%)	0.0089 (1.5%)	0.0365 (6.1%)

VI. CONCLUSIONS AND FUTURE WORK

This work extended and evaluated the AM-FM framework, a two-dimensional automatic metric for machine translation evaluation, which is designed to operate at the sentence level. Two different versions were evaluated: monolingual (mAM-FM) and cross-language (xAM-FM). Comparative evaluations were conducted to study how the metric correlates with human-generated scores for adequacy and fluency.

Obtained results show that mAM-FM can be tuned to achieve better correlations with human evaluations for both adequacy and fluency than other conventional metrics. On the other hand, although the xAM-FM version allows for conducting quality assessments without the need for a set of reference translations, its performance is significantly below than the monolingual version, but still comparable to the performance of other state-of-the-art automatic evaluation metrics.

As future research, we plan to study the effects of vector space model parameters, as well as different techniques for constructing the semantic continuous space embeddings, with the objective of improving the performance of the AM metric component, in both the monolingual and the cross-language scenarios. More specifically, we plan to explore the use of neural network architectures, such as deep-autoencoders [1] or recurrent neural networks [41], as an alternative to construct the continuous space representations.

Additionally, we also want to study fusion techniques for combining all three metric components, mAM, xAM and FM into a single score for machine translation evaluation, which we expect should combine the advantages of assessing adequacy by means of semantic similarity in both monolingual and cross-language continuous spaces.

Finally, we plan to explore other possible uses of the AM-FM framework in other statistical machine translation sub-tasks such as, for instance, its use as objective function for MERT optimization, as a model for phrase-table pruning, and as a translation feature during decoding, as in [4].

REFERENCES

[1] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[2] R. E. Banchs and M. R. Costa-jussà, "Cross-language document retrieval by using nonlinear semantic mapping," *Appl. Artif. Intell.*, vol. 27, no. 9, pp. 1–22, Oct. 2013.

[3] T. Mikolov *et al.*, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, vol. 26, pp. 3111–3119.

[4] J. Gao *et al.*, "Learning continuous phrase representations for translation modelling," in *Proc. 52nd Annu. Meeting Assoc. Comput. l Linguist.*, Baltimore, MD, USA, Jun. 2014, vol. 1, pp. 699–709.

[5] J. Zhang *et al.*, "Bilingually-constrained phrase embeddings for machine translation," in *Proc. 52nd Annu. Meeting Assoc. Computat. Linguist.*, Baltimore, MD, USA, Jun. 2014, vol. 1, pp. 111–121.

[6] R. Socher *et al.*, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proc. Conf. Empir. Meth. Nat. l Lang. Process.*, Edinburgh, U.K., Jul. 2011, pp. 151–161.

[7] T. Mikolov *et al.*, "Recurrent neural network based language model," in *Proc. Interspeech*, Florence, Italy, Aug. 2010, pp. 1045–1048.

[8] J. Blatz *et al.*, "Confidence estimation for machine translation," Johns Hopkins Univ., Baltimore, MD, USA, Final Report, 2003, WS2003 CLSP Summer Workshop.

[9] J. S. White, T. O'Cornell, and F. O'Nava, "The ARPA MT evaluation methodologies: Evolution, lessons and future approaches," in *Proc. Assoc. Mach. Translat. Amer.*, Oct. 1994, pp. 193–205.

[10] R. E. Banchs and H. Li, "AM-FM: A semantic framework for translation quality assessment," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguist.*, Portland, OR, USA, Jun. 2011, pp. 153–158.

[11] K. Papineni *et al.*, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318.

[12] C. Tillmann *et al.*, "Accelerated DP based search for statistical translation," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, Rhodos, Greece, Sep. 1997, pp. 2667–2670.

[13] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proc. Human Lang. Tech. Conf.*, San Diego, CA, USA, Mar. 2002.

[14] A. Lavie and M. J. Denkowski, "The meteor metric for automatic evaluation of machine translation," *Mach. Translat.*, vol. 23, pp. 105–115, May 2009.

[15] M. Snover *et al.*, "Study of translation edit rate with targeted human annotation," in *Proc. 7th Biennial Conf. Assoc. Mach. Translat. Amer.*, Cambridge, MA, USA, Aug. 2006.

[16] C. K. Lo and D. Wu, "MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles," in *Proc. 49th Annu. Meeting Assoc. for Comput. Linguist.*, Portland, OR, USA, Jun. 2011, pp. 220–229.

[17] C. K. Lo and D. Wu, "MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric," in *Proc. 8th Workshop Statist. Mach. Translat.*, Sofia, Bulgaria, Aug. 2013, pp. 422–428.

[18] E. Hovy, M. King, and A. Popescu-Belis, "Principles of context-based machine translation evaluation," *Mach. Translat.*, vol. 17, no. 1, pp. 43–75, Mar. 2002.

[19] C. Callison-Burch *et al.*, "(Meta-) evaluation of machine translation," in *Proc. 2nd Workshop Statist. Mach. Translat.*, Prague, Czech Republic, Jun. 2007, pp. 136–158.

[20] A. Lopez, "Statistical machine translation," *ACM Comput. Surveys*, vol. 40, no. 3, pp. 8–49, Aug. 2008.

[21] H. Somers, "Round-trip translation: What is it good for?," in *Proc. Workshop Australasian Lang. Technol. Assoc*, Sydney, Australia, Dec. 2005, pp. 127–133.

[22] R. Rapp, "The back-translation score: Automatic MT evaluation at the sentences level without reference translations," in *Proc. 47th Annu. Meeting Assoc. Comput. Linguist.*, Singapore, Aug. 2009, pp. 133–136.

[23] C. K. Lo *et al.*, "XMEANT: Better semantic MT evaluation without reference translations," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguist.*, Baltimore, MD, USA, Jun. 2014, pp. 765–771.

[24] C. B. Quirk, "Training a sentence-level machine translation confidence measure," in *Proc. 4th Int. Conf. Lang. Resources Eval.*, Lisbon, Portugal, May 2004, pp. 825–828.

[25] M. Gamon, A. Aue, and M. Smets, "Sentence-level MT evaluation without reference translations: Beyond language modeling," in *Proc. 10th Annu. Conf. Eur. Assoc. Mach. Translat.*, Budapest, Hungary, May 2005, pp. 103–111.

[26] J. S. Albrecht and R. Hwa, "Regression for sentence-level MT evaluation with pseudo references," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguist.*, Prague, Czech Republic, Jun. 2007, pp. 296–303.

[27] L. Specia *et al.*, "Improving the confidence of machine translation quality estimates," in *Proc. MT Summit XII*, Ottawa, ON, Canada, Aug. 2009.

[28] A. L. F. Han *et al.*, "LEPOR: A robust evaluation metric for machine translation with augmented factors," in *Proc. 24th Int. Conf. Comput. Linguist.*, Mumbai, India, Dec. 2012.

[29] M. Macháček and O. Bojar, "Results of the WMT13 metrics shared task," in *Proc. 8th Workshop Statist. Mach. Translat.*, Sofia, Bulgaria, Aug. 2013, pp. 45–51.

[30] C. K. Lo and D. Wu, "On the reliability and inter-annotator agreement of human semantic MT evaluation via HMEANT," in *Proc. 14th Int. Conf. Lang. Resources Eval.*, Reykjavik, Iceland, May 2014, pp. 602–607.

[31] A. Birch *et al.*, "The feasibility of HMEANT as a human MT evaluation metric," in *Proc. 8th Workshop Statist. Mach. Translat., Assoc. for Comput. Linguist.*, Sofia, Bulgaria, Aug. 2013, pp. 52–61.

[32] D. Vilar *et al.*, "Human evaluation of machine translation through binary system comparisons," in *Proc. 2nd Workshop Statist. Mach. Translat.*, Prague, Czech Republic, Jun. 2007, pp. 96–103.

[33] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to latent semantic analysis," *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[34] S. Dumais, T. K. Landauer, and M. L. Littman, "Automatic cross-linguistic information retrieval using latent semantic indexing," in *Proc. SIGIR Workshop Cross-Lingual Inf. Retrieval*, Philadelphia, PA, USA, Jul. 1997, pp. 16–23.

[35] G. M. Salton, A. K. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[36] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press, 1999, ch. 6.

[37] G. H. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *J. Soc. Ind. Appl. Math.: Numer. Anal.*, vol. 2, no. 2, pp. 205–224, 1965.

[38] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, Denver, CO, USA, Sep. 2002.

[39] M. Snover *et al.*, "Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric," in *Proc. 4th Workshop Statist. Mach. Translat.*, Athens, Greece, Mar. 2009.

[40] F. A. Gers and J. Schmidhuber, "LSTM recurrent networks learn simple context free and context sensitive languages," *IEEE Trans. Neural Netw.*, vol. 12, no. 6, pp. 1333–1340, Nov. 2001.

Rafael E. Banchs (M'14) is currently a Research Scientist at the Institute for Infocomm Research in Singapore. He received his Ph.D. in electrical engineering from the University of Texas at Austin in 1998. He was awarded a Ramon y Cajal fellowship from the Spanish Ministry of Education and Science from 2004 to 2009. His recent areas of research include Machine Translation, Information Retrieval, Cross-language Information Retrieval and Dialogue Systems. More specifically, he has been working on the application of vector space models along with linear and non-linear projection techniques to improve the quality of statistical machine translation and cross-language information retrieval systems.

He has served as co-organizer of the 2nd TC-STAR Workshop on Speech to Speech Translation 2006; the First International Workshop on Content Analysis in the Web2.0 (CAW2) at WWW 2009, the ESIRMT-HyTra Joint Workshop at EACL'12, the CREDISLAS workshop at LREC'12, the Special Session "Rediscovering 50 Years of Discoveries" at ACL'12, and HyTra-2 and HyTra-3 workshops at ACL'13 and EACL'14. He has also served as area chair for IJCNLP'11, general co-chair of AIRS'13 and PC chair of IALP'14.

Luis Fernando D'Haro (M'00) earned his degree as electronics engineer in 2000, from Universidad Autónoma de Occidente in Cali, Colombia, and his Ph.D. (with highest honors) from the Technical University of Madrid in 2009. During his Ph.D. studies he was a Visiting Researcher at the I6 Human Language Technology and Pattern Recognition Group in Aachen, Germany (2005) and AT&T Research labs (2006). Later he made a postdoctoral research stay at the Speech Processing Group in Brno, Czech Republic (2011). He is currently a Research Scientist at the Institute for Infocomm Research in Singapore. His current research is mainly focused on Spoken Dialogue Systems, Information Retrieval and Machine Translation. Previously, he has worked on Speaker and Language Recognition.

Haizhou Li (M'91–SM'01–F'14) received the B.Sc., M.Sc., and Ph.D. degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. Dr, Li is currently the Principal Scientist, Department Head of Human Language Technology in the Institute for Infocomm Research (I^2R), Singapore. He is also an Adjunct Professor at the National University of Singapore and a Conjoint Professor at the University of New South Wales, Australia. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing.

Prior to joining I^2R, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor at CRIN in France (1994–1995), a Research Manager at the Apple-ISS Research Centre (1996–1998), a Research Director in Lernout & Hauspie Asia Pacific (1999–2001), and the Vice President in InfoTalk Corp. Ltd. (2001–2003).

Dr. Li is currently the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2015–2017). He has served in the Editorial Board of *Computer Speech and Language* (2012–2014). He is an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the Vice President of the International Speech Communication Association (2013–2014), the President of Asia Pacific Signal and Information Processing Association (2015–2016). He was the General Chair of ACL 2012 and INTERSPEECH 2014.

Dr. Li is a Fellow of the IEEE. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one the two Nokia Visiting Professors in 2009 by the Nokia Foundation.