Machine Translation Evaluation for Arabic using Morphologically-Enriched Embeddings

Francisco Guzmán, Houda Bouamor*, Ramy Baly† and Nizar Habash‡
Qatar Computing Research Institute
*Carnegie Mellon University in Qatar
†American University of Beirut
‡New York University Abu Dhabi
fguzman@qf.org.qa, hbouamor@cmu.edu,
rgb15@mail.aub.edu, nizar.habash@nyu.edu

Abstract

Evaluation of machine translation (MT) into morphologically rich languages (MRL) has not been well studied despite posing many challenges. In this paper, we explore the use of embeddings obtained from different levels of lexical and morpho-syntactic linguistic analysis and show that they improve MT evaluation into an MRL. Specifically we report on Arabic, a language with complex and rich morphology. Our results show that using a neural-network model with different input representations produces results that clearly outperform the state-of-the-art for MT evaluation into Arabic, by almost over 75% increase in correlation with human judgments on pairwise MT evaluation quality task. More importantly, we demonstrate the usefulness of morpho-syntactic representations to model sentence similarity for MT evaluation and address complex linguistic phenomena of Arabic.

1 Introduction

Statistical machine translation (SMT) into morphologically rich languages (MRL) faces many challenges: from handling a complex and rich vocabulary, to designing adequate MT metrics that take morphology into account. While the first problem has widely explored (e.g. by using morphological analysis tools to reduce sparsity), the evaluation part has only been partly addressed. This is problematic since traditional MT metrics struggle to distinguish between (*i*) incorrect lexical choices; (*ii*) valid alternative lexical or syntactic variations; and (*iii*) differences in morphological inflection that are the result of incorrect case assignment or morphological agreement. While metrics like METEOR (Denkowski and Lavie, 2011) have made it possible to distinguish between (*i*) and (*ii*) by using paraphrases, (*iii*) is still an open problem. As a result, progress in SMT for MRL is hindered by the lack of adequate evaluation metrics. Since SMT metrics are used not only for evaluation but also for tuning system parameters, it is crucial that the MT metrics correctly handle morphology.

Most recently, deep learning models have been used more heavily in different parts of the natural language processing (NLP) community, including MT and MT evaluation. One of the main advantages of such models is the use of distributed word representations (embeddings). It has been shown that word embeddings are able to capture to certain semantic and syntactic aspects of words (Mikolov et al., 2013). Further refinements allow the inclusion of morphological information into distributed representations (Cotterell and Schütze, 2015). Word embeddings have been shown to help with modeling textual similarity well in the context of MT evaluation for MT into English (Guzmán et al., 2015), and community Question Answering (Guzmán et al., 2016). Nonetheless little exploration has been done on the use of embeddings for MT into MRL.

In this paper, we investigate how embeddings obtained from different levels of lexical and morpho-syntactic linguistic analysis can improve MT evaluation into a MRL. Specifically we report on Arabic, a language with complex and rich morphology paired with a high degree of ambiguity (Habash, 2010). Our results show that using a pairwise neural-network over different representations produces results

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 1398–1408, Osaka, Japan, December 11-17 2016.

that clearly outperform the state-of-the-art for MT evaluation into Arabic, by almost over 75% increase in correlation with human judgments on pairwise MT evaluation quality task. More importantly, we demonstrate that the use of embeddings based on morpho-syntactic representations in conjunction with the non-linear modeling capabilities of a neural-network help to capture the preferences of human judges.

Next, we present related work in Section 2. We describe our approach in detail in Section 3; and we evaluate in Section 4. We present a discussion of our findings in Section 5.

2 Related Work

Despite its well-known shortcomings (Callison-Burch et al., 2006), BLEU continues to be the de-facto MT evaluation metric. Several studies have attempted to improve upon it by taking into account different aspects of linguistic structures including: (*i*) synonym dictionaries or paraphrase tables (Denkowski and Lavie, 2011; Snover et al., 2010); (*ii*) syntactic information (Liu and Gildea, 2005; Giménez and Màrquez, 2007; Liu et al., 2010; Chen and Kuhn, 2011); (*iii*) morphology (Tantug et al., 2008); (*iv*) semantics (Dahlmeier et al., 2011; Lo et al., 2012) and (*v*) discourse (Guzmán et al., 2014b; Joty et al., 2014). Generally, these metrics have been focused on translation into English. However, there has been little attention into their direct applicability to languages with rich morphology.

Our work focuses on automatic evaluation of translation into morphologically rich languages, Arabic more specifically. In that sense, our work is related to AL-BLEU (Bouamor et al., 2014) which is an adaptation of BLEU that gives partial credits for stem and morphological matchings of hypothesis and reference words. Here, in addition to using lexical information captured by n-gram metrics, we show that using morpho-syntactic representations can significantly improve the correlation with human judgments. Furthermore, we use a neural-network, which uses non-linearities to improve modeling.

Over the past few years, neural network models have dramatically improved the state-of-the-art of different NLP applications (Goldberg, 2015). For instance, in SMT we have observed an increased use of neural nets for language modeling (Bengio et al., 2003; Mikolov et al., 2010), for improving answer ranking in community Question Answering (Guzmán et al., 2016), for improving the translation modeling (Devlin et al., 2014; Bahdanau et al., 2014; Cho et al., 2014) and for machine translation evaluation (Guzmán et al., 2015; Gupta et al., 2015). Our work is related to Guzmán et al. (2015), in several levels of lexical, syntactic and semantic are combined in a compact fashion using a pairwise neural framework. There are several differences between that work and ours: (*i*) we do not use syntactic embedding representations, (*ii*) we include additional pairwise features, namely the pairwise cosine similarity between embeddings; and (*ii*) we focus on an MRL language. While use of syntactic representations has proven a useful component to evaluate English, it relies heavily on an syntactic neural parser (Socher et al., 2013), which increases the complexity of the evaluation setup, and is not readily available for every language. Here, we instead use morpho-syntactic representations which capture both syntactic and morphological aspects of language. In our experiments, these simple representations are powerful enough to provide state-of-the-art performance.

In this work, we use neural network models to improve MT evaluation into Arabic using representations that capture morphology. Morphological structure has been shown to improve the quality of word clusters (Clark, 2003), word vector representations (Cotterell and Schütze, 2015) and neural language models (Botha and Blunsom, 2014). The novelty of our work resides in the way we integrate lexical and morpho-syntactic distributed representations into a neural-network. We demonstrate that combining several sources of complementary information is useful to capture sentence similarity in a translation evaluation scenario. And arguably, capture complex phenomena like morphological agreement.

3 Approach

We use a pairwise approach to translation evaluation (Guzmán et al., 2014a) using neural networks. We use neural networks for two reasons. First, to take advantage of their ability to model complex non-linear relationships efficiently. Second, to have a framework that allows for easy incorporation of distributed representations captured by lexical and morpho-syntactic embeddings. In this section, we describe the neural-network model and the distributed representations we use as features in this work.

3.1 Learning framework

Neural Network Model Our full neural network model for pairwise evaluation is depicted in Figure 1. It is a direct adaptation of the feed-forward NN proposed for English MTE (Guzmán et al., 2015). Technically, we have a binary classification task with input $x = (r, t_1, t_2)$, which outputs 1 if t_1 is a *better* translation than t_2 in the context of the reference r, or 0 otherwise. The network computes a sigmoid function $f(r, t_1, t_2) = \mathrm{sig}(\mathbf{w_v^T}\phi(r, t_1, t_2) + b_v)$, where $\phi(x)$ transforms the input x through the hidden layer, $\mathbf{w_v}$ are the weights from the hidden layer to the output layer, and b_v is a bias term.

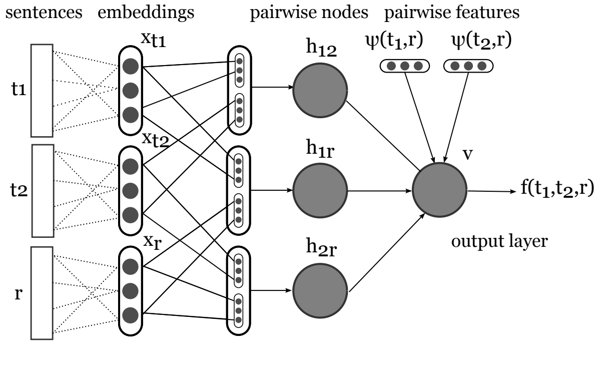

Figure 1: Overall architecture of the neural network.

To decide which hypothesis is *better* given the tuple (r, t_1, t_2) as input, we map the hypotheses and the reference to a fixed-length vector $[\mathbf{x}_r, \mathbf{x}_{t_1}, \mathbf{x}_{t_2}]$, using embeddings based on different lexical and morpho-syntactic representations.

We model three types of interactions (between t_1, t_2 and r) using different groups of nodes in the hidden layer h_{12}, h_{1r}, h_{2r}. The input to each of these groups is the concatenation of the vector representations of the two interacting components i.e., $\mathbf{x_{1r}} = [\mathbf{x}_{t_1}, \mathbf{x}_r]$, $\mathbf{x_{2r}} = [\mathbf{x}_{t_2}, \mathbf{x}_r]$, $\mathbf{x_{12}} = [\mathbf{x}_{t_1}, \mathbf{x}_{t_2}]$.

In summary, the transformation $\phi(t_1, t_2, r) = [\mathbf{h_{12}}, \mathbf{h_{1r}}, \mathbf{h_{2r}}]$ can be written as :

$$\mathbf{h_{ij}} = \tanh(\mathbf{W_{ij}}\mathbf{x_{ij}} + \mathbf{b_{ij}})$$

where $ij \in \{12, 1r, 2r\}$, $\tanh(.)$ is a non-linear component-wise activation function, $\mathbf{W} \in \mathbb{R}^{H \times N}$ are the associated weights between the input layer and the hidden layer, and $\mathbf{b}$ are the bias terms.

The model further allows to incorporate external sources of information in the form of *skip arcs* ψ that go directly from the input to the output layer. These arcs represent pairwise *similarity* between each translation and the reference (we denote them as $\psi_{1r} = \psi(t_1, r)$ and $\psi_{2r} = \psi(t_2, r)$). We use these feature vectors to encode two basic elements: (*i*) the pairwise cosine similarity between the embeddings from each translation and the reference; and (*ii*) N-gram based MT evaluation measures (e.g., AL-BLEU, METEOR, and NIST). We provide more detail about pairwise features in the next section.

Pairwise Network Training The negative log-likelihood of the training data for the model parameters $\theta = (\mathbf{W_{12}}, \mathbf{W_{1r}}, \mathbf{W_{2r}}, \mathbf{w_v}, \mathbf{b_{12}}, \mathbf{b_{1r}}, \mathbf{b_{2r}}, b_v)$ can be written as follows:

$$J_\theta = -\sum_n y_n \log \hat{y}_{n\theta} + (1 - y_n) \log (1 - \hat{y}_{n\theta}) + \lambda \sum \theta^2 \tag{1}$$

where $\hat{y}_{n\theta} = f_n(t_1, t_2, r)$ is the activation at the output layer for the n-th data instance, and λ is the L_2 regularization penalty. The network is trained with stochastic gradient descent (SGD), mini-batches and adagrad updates (Duchi et al., 2011), using Theano (Bergstra et al., 2010).

Evaluating a single translation Most of the MT evaluation metrics are not designed to do pairwise, but absolute evaluation. In other words, we are interested in generating a score for a single translation t_1 given a reference r. To achieve this with our pairwise network, we compute the *goodness* margin, which tells us how good translation t_1 is better than any other translation $t_\varnothing$ given the reference r. In this case, $t_\varnothing$ is the average representation for all translations observed during training.

To compute the margin, we subtract the scores for the direct and reverse network predictions, and generate the final score for the sentence: $score(t, r) = 1 + (f(r, t, t_\varnothing) - f(r, t_\varnothing, t))/2$.

Note that we use both direct and reverse predictions as our network is not exactly symmetric. We also shift the score to range between $[0, 1]$.

3.2 Features

In this work, we compare different sets of features representing different levels lexical and morpho-syntactic information. As a baseline, we also used several MT metrics that are based on n-gram matches.

Lexical units A distinguishing characteristic of Arabic morphology is the presence of concatenative morphemes, where words are formed via concatenations of stems, affixes and clitics. To allow our system to model how morphemes interact at a finer level, we split the morphemes. We used MADAMIRA (Pasha et al., 2014), the state-of-the-art morphological analyzer and disambiguator, to perform morphological tokenization following ATB scheme (Habash and Sadat, 2006). We extracted two forms of lexical features: NORM and TOKEN, which are tokens with and without Alef/Yaa normalization, respectively. We also extract the LEMMA feature; a morphological abstraction that represents words related by inflectional morphology.

Morpho-Syntactic units We extracted part-of-speech (POS) tags according to different POS tagsets including: (*i*) CATIBPOS(Habash et al., 2009), (*ii*) KULICKPOS[1] (Kulick et al., 2006), (*iii*) BUCKWALTERPOS (Buckwalter, 2004) and (*iv*) STANFORDPOS tagsets. These tagsets differ in their richness and complexity they capture. CATIBPOS is the simplest with only 6 base tags[2], BUCKWALTERPOS is the richest with 485 base tags, and KULICKPOS and STANFORDPOS come in-between with 43 and 32 base tags, respectively. These tags were extracted using MADAMIRA, except for the Stanford tags, for which we used Stanford CoreNLP (Manning et al., 2014). Table 1 illustrates an example sentence with its lexical and morpho-syntactic features.[3]

Sentence:	*ςAd AlmSrywn Alðyn AxtTfwA Ǎlý bldhm* – عاد المصريون الذين اختطفوا إلى بلدهم						
TOKEN	+هم *+hm*	بلد *bld*	إلى *Ǎlý*	اختطفوا *AxtTfwA*	الذين *Alðyn*	المصريون *AlmSrywn*	عاد *ςAd*
NORM	+هم *+hm*	بلد *bld*	الي *Āly*	اختطفوا *AxtTfwA*	الذين *Alðyn*	المصريون *AlmSrywn*	عاد *ςAd*
LEMMA	+هُم *+hum*	بَلَد *balad*	إِلَى *Ǎilaý*	اِخْتَطَف *Aix.taTaf*	الَّذِي *Al∼aðiy*	مِصرِيّ *miS.riy∼*	عَاد *ςaAd*
CATIBPOS	+NOM	NOM	PRT	VRB-PASS	NOM	NOM	VRB
KULICKPOS	+PRP$	NN	IN	VBN	WP	DT+NNS	VBD
STANFORDPOS	PRP$	NN	IN	VBN	WP	DTNNS	VBD
BUCKWALTERPOS	+POSS _PRON _3MP	NOUN +CASE _DEF_GEN	PREP	PV_PASS +PVSUFF _SUBJ:3MP	REL_PRON	DET+NOUN +NSUFF _MASC_PL_NOM	PV+ PVSUFF _SUBJ:3MS
GLOSS	their	country	to	were abducted	which	the Egyptians	returned
English	The Egyptians who were abducted returned to their country.						

Table 1: Illustration of the lexical and morpho-syntactic feature representations extracted for an example sentence. The sentence is presented from right to left following the directionality of writing Arabic.

Distributed representations To obtain embeddings based on the lexical and morpho-syntactic units described above, we annotated the fifth edition of the Gigaword corpus (LDC2011T11) with each of the representations. Then, we trained word embeddings using word2vec (Mikolov et al., 2013). To

[1] (Pasha et al., 2014) refers to Kulick tagset as the Penn ATB tagset, while it is Buckwalter tagset that is used in Penn ATB.

[2] The size of each tagset is expected to increase since Arabic morphemes can function as clitics, and their corresponding POS tags are then assigned a clitic marker (+). The final sizes of the extracted POS tags are shown in Table 4.

[3] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007) (in alphabetical order): أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي *Â b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y*, and the additional symbols: ء ’, أ Â, إ Ǎ, آ Ā, ؤ ŵ, ئ ŷ, ة ħ, ى ý. Diacritics are represented as: ـَ a, ـُ u, ـِ i, ـْ ., ـً ã, ـٌ ũ, ـٍ ĩ, and ـّ ∼.

obtain sentence-level representations for each of the translations and the references, we used additive composition (Mitchell and Lapata, 2010) with dropping unknown words.

N-gram MT metrics We used the different n-gram based metrics to serve as a benchmark, and as additional features that capture lexical similarity. We used: BLEU+1 (Nakov et al., 2012),NIST (Doddington, 2002); METEOR (Denkowski and Lavie, 2011), 1-TER (Snover et al., 2006), and AL-BLEU (Bouamor et al., 2014), to compute scores at the sentence-level. For consistency with previous work, we report scores over words, and not over morphemes.

4 Experimental Setup

In this section, we describe the experimental settings we used through our study. First, we introduce our evaluation criteria, then we elaborate on the dataset and various settings we used for our experiments.

4.1 Performance evaluation

Automatic evaluation metrics are evaluated based on their correlation with human-performed evaluations (Soricut and Brill, 2004). In this work, we use Kendall's τ, a coefficient that measures the agreement between rankings produced by human judgments and rankings produced by an automatic metric, at the sentence-level. We use the WMT'12 (workshop of machine translation) definition of Kendall's τ that ignores ties, and is calculated as follows: $[\tau = (\text{\# concordant pairs} - \text{\# discordant pairs})\,/\text{total pairs}]$, where the *# concordant pairs* is the number of times the human judgment and the automatic metric agree in the ranking of any two translations that belong to the same source sentence. The *# discordant pairs* is the opposite. The value of τ ranges from -1 (all pairs are discordant) to $+1$ (all pairs are concordant).

4.2 Data

To evaluate the performance of the neural-network, we evaluated its Kendall's τ given different input representations. We used a medium-scale corpus of human judgments for Arabic MT outputs covering different topics in the news domain (Bouamor et al., 2014). The corpus is composed of 1,383 sentences selected from two datasets: (*i*) the standard English-Arabic NIST 2005 corpus, commonly used for MT evaluations and composed of political news stories; and (*ii*) a small dataset of translated Wikipedia articles. This corpus contains the source and target text along with the automatic translations produced by five English-to-Arabic MT systems: three research-oriented phrase-based systems with various morphological and syntactic features and two commercial, off-the-shelf systems. The corpus contains annotations that assess the quality of the five systems, by ranking their translation candidates from best to worst for each source sentence in the corpus. The annotation was conducted by mirroring the WMT evaluation campaigns (Callison-Burch et al., 2011), but with few key differences: (*i*) the full corpus was annotated with no random sampling, and (*ii*) the task was performed by two independent native speakers of Arabic. The total number of annotated pairs in this corpus is 33,192 (each sentence has two annotations, each yielding 12 rankings). The agreement between the annotators in terms of Kendall's τ is 49.20 (which roughly translates to agreement in 75% of all rankings). We used random partitions of 783 sentences for training (TRAIN), 300 sentences for tuning (DEV) and 300 sentences for testing (TEST).

4.3 Network Settings

We train our model on TRAIN with hidden layers of size 10 for 30 epochs with mini batches of size 30, L_2 regularization of 0.0001, and a learning rate of 0.01. We normalize the input feature values to the $[-1; 1]$ interval using minmax, and we initialize the network weights by sampling from a uniform distribution as in (Bengio and Glorot, 2010).

We evaluate the model on DEV after each epoch, and keep the one that achieves the highest accuracy. We selected the above parameter values on the DEV dataset using the full model, and we use them for all experiments described in Section 5, where we evaluate on the official TEST dataset.

Note that, we train the pairwise neural-network using all pairwise rankings in the TRAIN set. At test time, we compute the scores of the translations in TEST using the absolute scoring previously described.

5 Results

In this section we present the main results from our experiments. Since the dataset is new, we first present the Kendall's τ scores obtained from five n-gram based metrics that are popular in the community. Then, we present the results of using embeddings obtained from different representations as input to train the neural-network. Finally, we present combination of representations, that shed light on the capabilities of the neural-network to learn how to exploit different levels of lexical and morpho-syntactic information.

A. MT Metrics			**B. Embeddings**		
		Kendall's τ			**Kendall's τ**
			Lexical		
1	NIST	17.94	7	TOKEN	**24.35**
2	METEOR	17.90	8	NORM	23.22
3	AL-BLEU	17.02	9	LEMMA	21.17
4	BLEU	16.11	*Morpho-syntactic*		
5	1-TER	4.97	10	BUCKWALTERPOS	**25.49**
			11	KULICKPOS	16.25
6	5METRICS	**18.12**	12	STANFORDPOS	10.90
			13	CATiBPOS	5.41

Table 2: Kendall's τ on the TEST set for traditional n-gram based metrics as well as metrics built using different lexical and morpho-syntactic embeddings as input to the neural-network.

5.1 MT Metrics

In Table 2.A, we present four of the most popular MT metrics: BLEU, NIST, METEOR and 1-TER, along with AL-BLEU; a precision-based metric that is designed to handle Arabic morphology. These metrics were calculated over tokenized text. From the results, we observe that NIST and METEOR obtain very similar performances for this task, with 17.94 and 17.90, respectively.[4] This suggests that both the paraphrasing that METEOR uses, and the precision weighting by n-gram *importance* that NIST does, yield results that are more in line with human judgments than other metrics. Additionally, the role of morphology is important, as AL-BLEU presents an improvement over BLEU (17.02 vs 16.11).

Next, we combine the five metrics in a logistic regression model. We observe that the 5METRICS combination yields only minor improvements over the best single metrics, suggesting limited complementarity. We use this 5METRICS combination as a baseline to compare to next experiments.

5.2 Embeddings

In Table 2.B, we present the results of using the neural-network with embeddings for different representations. Using the embeddings for any lexical representations (TOKEN, NORM, LEMMA) produces significant improvements (+3%) over any of the MT metrics and their combination, yielding state-of-the-art results. The relative increase over 5METRICS ranges from 17% (+3.05% absolute with LEMMA) to 34% (+6.23% absolute with TOKEN).

Using the embeddings obtained for morpho-syntactic representations results in a wide-range of results. Surprisingly, the BUCKWALTERPOS representation obtains very competitive scores, even surpassing the lexical representations and yielding the highest score yet with a 41% relative increase over 5METRICS (+7.37% absolute). However, none of the other morpho-syntactic representations improve on 5METRICS. There is a clear correlation between the size of the POS tag-set and the performance of the metric. We explore this relationship more in depth in Section. 5.4.

5.3 Combination of Representations

Given the complimentary information embedded in the different representations, it is natural to combine them to obtain a stronger metric. To combine different embedding representations, we simply concatenate the different embedding representations before feeding them to the network. Below, we present

[4] Note that the Kendall's τ results for METEOR are in the range of the results for translation from English into other two morphologically rich languages (German: 18.0 and Czech 16.0) reported in WMT 2012 (Callison-Burch et al., 2012)

		Kendall's τ		
Combinations		**result**	**prev. best**	**delta**
C. Embeddings and N-gram based metrics				
Lexical				
14	5METRICS+TOKEN	23.62	24.35	(-0.73)
15	5METRICS+NORM	**24.17**	23.22	(+0.95)
16	5METRICS+LEMMA	23.51	21.17	(+2.34)
Morpho-syntactic				
17	5METRICS+BUCKWALTERPOS	**29.81**	25.49	(+4.32)
18	5METRICS+KULICKPOS	23.58	18.12	(+5.46)
19	5METRICS+STANFORDPOS	21.79	18.12	(+3.67)
20	5METRICS+CATIBPOS	18.93	18.12	(+0.81)
D. Embedding mixtures				
Lexical + Lexical				
21	TOKEN +NORM	25.12	24.35	(+0.77)
22	NORM + LEMMA	**25.42**	23.22	(+2.20)
23	TOKEN+LEMMA	24.90	24.35	(+0.55)
24	TOKEN+NORM+LEMMA	25.34	25.42	(-0.08)
Lexical + Morpho-syntactic				
25	BUCKWALTERPOS+TOKEN	* **31.87**	25.49	(+6.38)
26	BUCKWALTERPOS+NORM	30.69	25.49	(+5.19)
27	BUCKWALTERPOS+LEMMA	30.69	25.49	(+5.19)
E. Embedding mixtures + Ngram-based metrics				
Lexical + Lexical				
28	5METRICS+TOKEN+NORM	25.56	25.12	(+0.44)
29	5METRICS+NORM+LEMMA	25.42	25.42	(+0.00)
30	5METRICS+TOKEN+LEMMA	25.45	24.90	(+0.55)
31	5METRICS+TOKEN+NORM+LEMMA	**28.35**	25.42	(+2.93)
Lexical + Morpho-syntactic				
32	5METRICS+BUCKWALTERPOS+TOKEN	29.78	31.87	(-2.09)
33	5METRICS+BUCKWALTERPOS+NORM	**30.73**	30.69	(+0.04)
34	5METRICS+BUCKWALTERPOS+TOKEN+LEMMA+NORM	30.44	31.87	(-1.43)

Table 3: Kendall's τ on the TEST set for metrics built using combination of lexical and morpho-syntactic embeddings in addition to the 5METRICS. For comparison, we compare the result of the combination to the best result of any of the components in the combination. The best result overall is marked with *.

three sets of results involving different types of combinations in Table 3. In addition to the Kendall's τ of a particular combination x, we indicate the best previous result (i.e., result of a sub-combination that was presented already, *underlined*), and its delta from combination x.

Embeddings and MT Metrics In Table 3.C, we present the results of adding the 5METRICS to the different embeddings in Table 2.B as *skip-arc* features. For lexical embeddings, we observe slight improvements, except for the TOKEN representation, which has a small decrease. The combination of each of the morpho-syntactic embeddings with 5METRICS improves over 5METRICS. For the best performer so far, BUCKWALTERPOS, we get even more improvements, reaching an 11.69 absolute increase over 5METRICS (65% relative). This showcases that the neural-network is able to successfully make use of the complementarity between n-gram-based MT metrics and other sources of morpho-syntactic information.

Embedding Mixtures Table 3.D presents the results of combining lexical and morpho-syntactic embeddings. Every pairwise combination of lexical embeddings improves over either embedding combined. However, putting all lexical embeddings together is not as good as NORM+LEMMA. Combining the best morpho-syntactic performer (BUCKWALTERPOS) with each lexical embedding produces large improvements. This highlights the ability of the neural-network to exploit the interactions between different representations to provide a more robust metric. With BUCKWALTERPOS+TOKEN, we reach our best results, improving over the 5METRICS baseline by 13.75% (or 75.9% relative improvement).

Embedding Mixtures and MT Metrics Finally, we present in Table 3.E the results of combining 5METRICS with different combinations of embeddings. The addition of 5METRICS to pairs of lexical embeddings does not hurt and only slightly improves the scores. However, putting all of the lexical embeddings together with 5METRICS makes a nice increase, although still not competitive with our best result so far. Finally, combining 5METRICS+BUCKWALTERPOS with different lexical embeddings seems to make little improvement if any.

5.4 Discussion

One of the interesting results from Table 2.B is the wide range of scores for the different embeddings. Surprisingly, the BUCKWALTERPOS representation does remarkably well. Here we investigate some possibilities for this behavior.

Feature expressiveness In Table 4, we consider different aspects of the different embeddings: their vocabulary size, their token out-of-vocabulary rate, the standard deviation of the cosine similarity scores they assign to each sentence in TEST, as well as their respective Kendall's τ scores. The larger the variance, the more expressive and informative the representation is, as it gives more diverse values for cosine similarity between translation and references.

	Vocab Size	Token OOV Rate (%)	StDev Cos Sim	Kendall's τ
TOKEN	98,923	1.32	$9.57\cdot10^{-2}$	24.35
NORM	97,870	1.29	$9.51\cdot10^{-2}$	23.22
LEMMA	51,493	1.25	$9.12\cdot10^{-2}$	21.17
BUCKWALTERPOS	714	0.00	$9.87\cdot10^{-2}$	25.49
KULICKPOS	53	0.00	$4.97\cdot10^{-2}$	16.25
STANFORDPOS	32	0.00	$4.16\cdot10^{-2}$	10.90
CATIBPOS	11	0.00	$3.84\cdot10^{-2}$	5.41

Table 4: Statistics for the different embedding representations and their impact on TEST: vocabulary size of the embedding representations, token OOV rate of the embedding vocabulary on the TEST set, the standard deviation (StDev) of cosine similarity values for all translation–reference pairs in the TEST set and the Kendall's τon the TEST.

Overall, we observe that the variance (or the standard deviation) of the values in the cosine similarity and the Kendall's τ scores correlate very well at $\rho = 0.94$. One way to interpret this is that the more expressive the representation is, the better it performs at MT evaluation. The size of the vocabulary generally adds expressiveness to a feature, and correlates within the lexical subset and the morpho-syntactic subset of the representations (but not across or overall). However, lexical representations lose expressiveness because OOVs.

Agreement The BUCKWALTERPOS representation is particularly rich and captures the morphological complexity of Arabic. Thus, it can be used to represent patterns of grammatical agreement across words, e.g., verb-subject and noun-adjective. While the lexical embedding may capture that the two verbs يكتب *yktb* 'he writes' and تكتب *tktb* 'she writes' may occur in similar contexts, the BUCKWALTERPOS representation models for agreement with the context they appear in. Since Arabic uses agreement heavily across verbs and noun phrases, we expect that the simple additive combination used with word embeddings is able to capture impressions of gender or number, just enough to allow the model to distinguish between sequences that are closer to the reference from those that are not. For example, the sentence يريد الطفل الصغير أن يكتب *yryd AlTfl AlSγyr Ân yktb* 'the little boy wants to write' encodes the masculine singular gender in four of its words. Simpler POS tags do not mark this information and thus are unable to distinguish between sentences that use the correct and incorrect gender information. In our experience, human judges pay careful attention to agreement errors as they reduce the fluency of the text even as the basic "accuracy" of it is kept.

Morpho-semantics The BUCKWALTERPOS representation contains some semantic features related to specific POS tags. To give a concrete example, in our data set, a future verb was translated correctly using the particle سوف *swf* 'will' (BUCKWALTERPOS: FUT_PART) by system A, and incorrectly as the particle لن *ln* 'will not' (BUCKWALTERPOS: NEG_PART) by system B. However, the reference had no future particle. While both translations were penalized equally (in terms of cosine similarity) in the TOKEN representation, the sentence with the negation POS NEG_PART was penalized more in the BUCKWALTERPOS representation. There are a number of negative particles in Arabic and all get the same BUCKWALTERPOS tag. This, we believe allows the neural-network to abstract and model them correctly.

Task specific Another reason that using morpho-syntactic helps MT evaluation even in isolation is that by definition, the setup of the MT evaluation forces translation and references to be somewhat close. Here, capturing agreement and POS semantics seem to correlate well with human judgments. We don't expect that using the morpho-syntactic to capture other type of sentence similarity (e.g mining sentence pairs in comparable corpora) will work as well.

6 Conclusions and Future Work

In this paper, we explored the use of different lexical and morpho-syntactic representations to model similarity in the context of MT evaluation for a morphologically rich language such as Arabic. Our results show that using the neural-network with the input embeddings obtained from different representations makes impressive gains individually and, especially, in combination. This confirms the neural-network ability to model complex interactions between the feature sets. The fact that the best performers in each of these subsets when combined (BUCKWALTERPOS+TOKEN) show very high gains; suggests that they are modeling very different aspects of the text. The lexical embeddings we posit model semantic similarity between test and reference; while the morpho-syntactic embeddings model syntactic similarity, and complex phenomena like morphological agreement.

Furthermore, when paired to morpho-syntactic representations, the distributed lexical information seems to be a good alternative to n-gram metrics to obtain state-of-the-art results. In the future, we would like to use rich morpho-syntactic representations for evaluation into other MRL languages to validate our observations for Arabic. For instance, this framework can be easily ported to European languages, such as German, Russian or Czech, where: (*i*) there the existence of morphological analyzers makes it feasible to develop morpho-syntactic embeddings, and (*ii*) there are datasets available (from WMT Evaluation Campaigns) to develop and evaluate such metrics. Finally, we want to test the use of morpho-syntactic representation in other tasks such as source language modeling for Neural Machine Translation.

Acknowledgment

We would like to thank the annotators, Fatma Tlili and Aisha Mohamed, for providing the machine translation quality human judgments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio and Xavier Glorot. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of AISTATS 2010*, volume 9, pages 249–256.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference*, SciPy '10, Austin, Texas.

Jan A Botha and Phil Blunsom. 2014. Compositional Morphology for Word Representations and Language Modeling. In *ICML*, pages 1899–1907.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A Human Judgement Corpus and a Metric for Arabic MT Evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Doha, Qatar.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.: LDC2004l02. Technical report, ISBN 1-58563-324-0.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Boxing Chen and Roland Kuhn. 2011. AMBER: A Modified BLEU, Enhanced Ranking Metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 71–77, Edinburgh, Scotland.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN encoder-decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*.

Alexander Clark. 2003. Combining Distributional and Morphological Information for Part of Speech Induction. In *Proceedings of the Tenth Conference on European chapter of the Association for Computational Linguistics*, pages 59–66.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological Word-Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. Tesla at WMT 2011: Translation Evaluation and Tunable Metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 78–84.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, UK.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1370–1380, Baltimore, Maryland.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality using n-gram Co-occurrence Statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264.

Yoav Goldberg. 2015. A Primer on Neural Network Models for Natural Language Processing. *CoRR*, abs/1510.00726.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014a. Learning to differentiate better from worse translations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Doha, Qatar, October. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014b. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, Maryland.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 805–814, Beijing, China.

Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2016. Machine translation evaluation meets community question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 460–466, Berlin, Germany.

Nizar Habash and Fatiha Sadat. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 49–52. Association for Computational Linguistics.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Reem Faraj, and Ryan Roth. 2009. Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of MEDAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt*.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using Discourse Structure for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14, pages 402–408, Baltimore, Maryland, USA.

Seth Kulick, Ryan Gabbard, and Mitchell Marcus. 2006. Parsing the Arabic Treebank: Analysis and Improvements. In *Proceedings of the Treebanks and Linguistic Theories Conference*, pages 31–42. Citeseer.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. TESLA: Translation Evaluation of Sentences with Linear-Programming-Based Analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 354–359.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully Automatic Semantic MT Evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. In *Interspeech*, volume 2, page 3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1439.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In *Proceedings of COLING 2012*, pages 1979–1994, Mumbai, India.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *LREC*, volume 14, pages 1094–1101.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, AMTA '06, Cambridge, Massachusetts, USA.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2010. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 455–465, Sofia, Bulgaria.

Radu Soricut and Eric Brill. 2004. A Unified Framework For Automatic Evaluation Using 4-Gram Co-occurrence Statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, pages 613–620, Barcelona, Spain.

Cüneyd Tantug, Kemal Oflazer, and Ilknur Durgar El-Kahlout. 2008. BLEU+: a Tool for Fine-Grained BLEU Computation. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, Marrakech, Morocco.