Neural and Symbolic Arabic Paraphrasing with Automatic Evaluation

Fatima Al-Raisi
Carnegie Mellon University
School of Computer Science

Abdelwahab Bourai
Carnegie Mellon University
School of Computer Science

Weijian Lin
Carnegie Mellon University
School of Computer Science

Abstract

We present a tool for Arabic paraphrasing that yields good paraphrasing accuracy. We present and compare several methods for paraphrasing and obtaining monolingual parallel data. We also present first results on Arabic paraphrasing using neural methods. Additionally, we propose a new evaluation metric for paraphrasing that is shown to correlate highly with human judgement.

1 Introduction

Paraphrasing and paraphrase detection are two important problems in natural language processing. Paraphrasing-based applications include text simplification and text generation form structured knowledge (Pavlick and Callison-Burch, 2016). Other paraphrastic models include machine translation and sentence summarization (Koehn et al., 2003; Zhou et al., 2006; Callison-Burch et al., 2006). Paraphrases are useful not only in generation tasks but also in analysis tasks such as information retrieval and question answering (Zukerman et al., 2002; Rinaldi et al., 2003; Dept and Wallis, 1993).

We present and compare two different approaches for sentence paraphrasing in Arabic: a phrase-based method and a neural method. To our knowledge, this is the first work on sentence paraphrasing for modern standard Arabic.

We also present a novel approach for obtaining parallel monolingual data and use the acquired data to train our neural sequence-to-sequence model.

As a by-product of this exercise, we contribute a large parallel monolingual corpus for Arabic containing two million sentence pairs. We also build a phrase database for Arabic containing over 88K phrase pairs of various lengths.

Another contribution of our work is devising and testing a new evaluation metric for paraphrasing. We present encouraging initial results in this paper.

The remainder of this paper is structured as follows: we contextualize our work within paraphrasing research in Section 2, we present the phrase-based and neural approaches for paraphrasing sentences and building phrase dictionaries in sections 3 and 4. We present details and discuss experiments on the evaluation metric in Section 6. We conclude with plans for future extensions of the work.

2 Related Work

The paraphrase database project PPDB has paraphrase resources for multiple languages (Bannard and Callison-Burch, 2005), including Arabic. The paraphrases are obtained using parallel bilingual corpora by applying the pivot method where one language is used as a bridge or intermediate meaning representation (Bannard and Callison-Burch, 2005). Paraphrases from dialectal Arabic to standard Arabic have been used in (Salloum and Habash, 2011) to improve Arabic-English statistical machine translation. Turker assisted paraphrasing has been used in (Denkowski et al., 2010) to improve English-Arabic MT. A comparison between various paraphrase acquisition techniques on sentential paraphrasing is given in (Bouamor et al., 2010) but does not include experiments on Arabic sentential paraphrasing.

3 Extracting Paraphrases from Bilingual Data

Our first approach to Arabic paraphrasing is the pivot method proposed by Bannard and Callison-Burch (Bannard and Callison-Burch, 2005). A key benefit is that it is language-agnostic and is based on the idea that any two source strings e_1 and e_2 that both translate to a reference string f_1 have similar meaning. Bannard and Callison-Burch used English as the reference string f, but in our study we will instead pivot into English to obtain paraphrase pairs (Bannard and Callison-Burch, 2005). We obtain the final paraphrase probability by marginalizing over the English translation probabilities with e and Arabic phrases a_1 and a_2. A mathematical formulation of the approach can be found in Equation 1.

$$p(a_2|a_1) = \sum_e p(a_2|e)p(e|a_1) \quad (1)$$

In order to extract paraphrases, we first obtained a parallel bilingual corpus through English and Arabic versions of the EUROPARL dataset (Koehn, 2005). We pruned the corpus to only contain sentences with less than 80 words and tokenized using the StanfordNLP Arabic Tokenizer (Manning et al., 2014). This gave us a final corpus size of 241,902 sentences.

Additionally, to calculate conditional probabilities for our paraphrase equation, we need alignments. Thus we ran GIZA++ to obtain these alignments (Och and Ney, 2003). Once we have a database of paraphrase mappings, we can then substitute phrases with their corresponding paraphrases by selecting the phrase with the highest probability. This substitution approach was used by Bannard and Callison-Burch in their study as well (Bannard and Callison-Burch, 2005) The way we extract the paraphrase is summarized in Equation 2.

$$\hat{a_2} = \text{argmax}_{a_2 \neq a_1} p(a_2|a_1) \quad (2)$$

An example of this process can be seen in Figure 1

3.1 Improving Coverage of Phrase Database

In our initial experiments, we noticed that generated phrase pairs did not necessarily match in some grammatical features such as definiteness and number. We post-processed the phrase dictionary to add

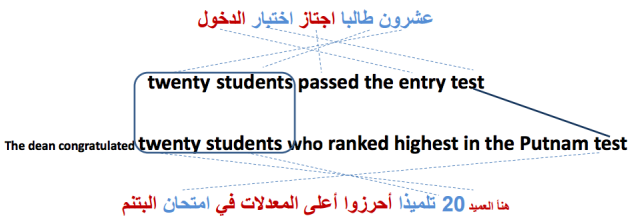

Figure 1: An example paraphrased sentence produced using the pivot method.

entries with other variants of the phenomenon for completion. For example, for a word pair that appears in the phrase table where one word is definite and the other is not, we add two entries where both are definite and both are indefinite. We did not adjust the scores to reflect this but ordered the entries according to observed frequency of the word/phrase. For definiteness, we limited the addition to the clear definite marker "al-" in Arabic. We applied the same for simple cases of number matching where the morphology is concatenative or easily processed. We note that this modifications did not include all possible mismatches since they were based on simple heuristics. However, this may have contributed to better grammaticality as discussed in Section 5.

We also noticed cases like the following in the generated phrase table:

x ||| y ||| score
x ||| z ||| score

We included, for improved coverage, the following entry:

y ||| z ||| score

Again, this was limited in scope since we relied on simple string match to identify such entries.

3.2 Phrase-substitution

We randomly sampled 100 sentences from the datasets we have (Xiaoyi et al., 2004; Graff and Walker, 2001) and performed phrase substitution. Figure 2 shows a sample of paraphrased sentences acquired using phrase substitution. We note that in the last example the output sentence differs from the original in only one word (last word) but the meaning is entirely altered. We discuss results and experiments on the quality of paraphrased sentences in Section 5.

4 Monolingual Parallel Data for Seq-to-Seq Paraphrasing

We need parallel monolingual data to train our sequence-to-sequence paraphrasing model. To address the lack of parallel monolingual data for Arabic, we propose a novel method for generating Arabic parallel language sentences using two other language pairs as resource. The idea is to use paired sentences data from two other languages, translate them into Arabic correspondingly, then use them as the training data for sequence-to-sequence machine translation model to train a translator to generate Arabic to Arabic paraphrases.

The first advantage of this approach is its scalability. After preparing enough training data for the paraphrase model, the generation step for Arabic paraphrases is easily scalable. The second advantage is that the seq-to-seq paraphraser model may contain valuable insight for building Arabic paraphrases database at words and phrases level, since state of the art neural machine translation techniques are capable of capturing word and phrase level similarity by projecting word embeddings into vector space.

We used europarl-v7 fr-en (Koehn, 2005) data which contains two million sentence pairs, and then used Google translate API to generate French-Arabic and English-Arabic sentence pairs correspondingly. Then we paired the output to construct parallel monolingual data and used it as training data for a Bi-LSTM with embedding size and hidden size set to be 512 and attention size set to be 128. The whole process is demonstrated in Figure 3. Training the Bi-LSTM took about 7 days on this dataset and we obtained a corpus of monolingual Arabic containing two million parallel sentences.

5 Results

We report results on the bilingual pivot and sequenc-to-sequence approaches detailed above.

5.1 Phrase-based Method

Using the pivot method, we obtained over 88K phrase pairs. We report a few results. First, Figure 4 shows the length distribution for phrases in the database.

	ICC(single)	ICC(average)
Grammaticality	0.375	0.545
Meaning	0.707	0.547

Table 1: Inter-annotator agreement

No. of changes	5	4	3	2	1
Frequency (Sent.)	3	4	5	16	26

Table 2: Number of changes in paraphrased sentences with high meaning preservation score

We obtained human evaluations from two native speakers on the grammaticality and meaning preservation aspects of the paraphrased sentences. For each criterion, we asked the annotator to judge the quality of the output on a scale from 1 to 5. We chose this scale to capture variation in the level of grammaticality (since there are minor and more serious grammatical mistakes) and in the extent to which the paraphrased sentence preserved the meaning of the original sentence. The agreement between annotators calculated in terms of IntraClass Correlation, preferred for ordinal data, is summarized in Table 1. It was not expected to find higher agreement on meaning preservation since it is more subjective than grammaticality. It is possible that phrases substituted were of similar meaning yet possibly resulted in unusual sentence structure which made the meaning preservation judgement straightforward while grammaticality harder to judge. We analyzed changes made to reference sentences that received the highest score in meaning preservation after paraphrase substitution (score = 5) . A change is a word replacement or deletion. Table 2 summarizes these changes. When n consecutive words are replaced by n or more words, we consider those to be n changes rather than 1 phrasal change. Table 3 summarizes the evaluation of paraphrased sentences obtained using phrase substitution.

As for the neural model, the produced corpus of sentence pairs was large (2M). By the time the output was produced, we had exhausted our human evaluation resources, so we qualitatively evaluated it on a small subset consisting of the first 10 sentence pairs. Based on this small sample, the monolingual parallel sentences obtained were grammatical or near grammatical with minor mistakes and highly

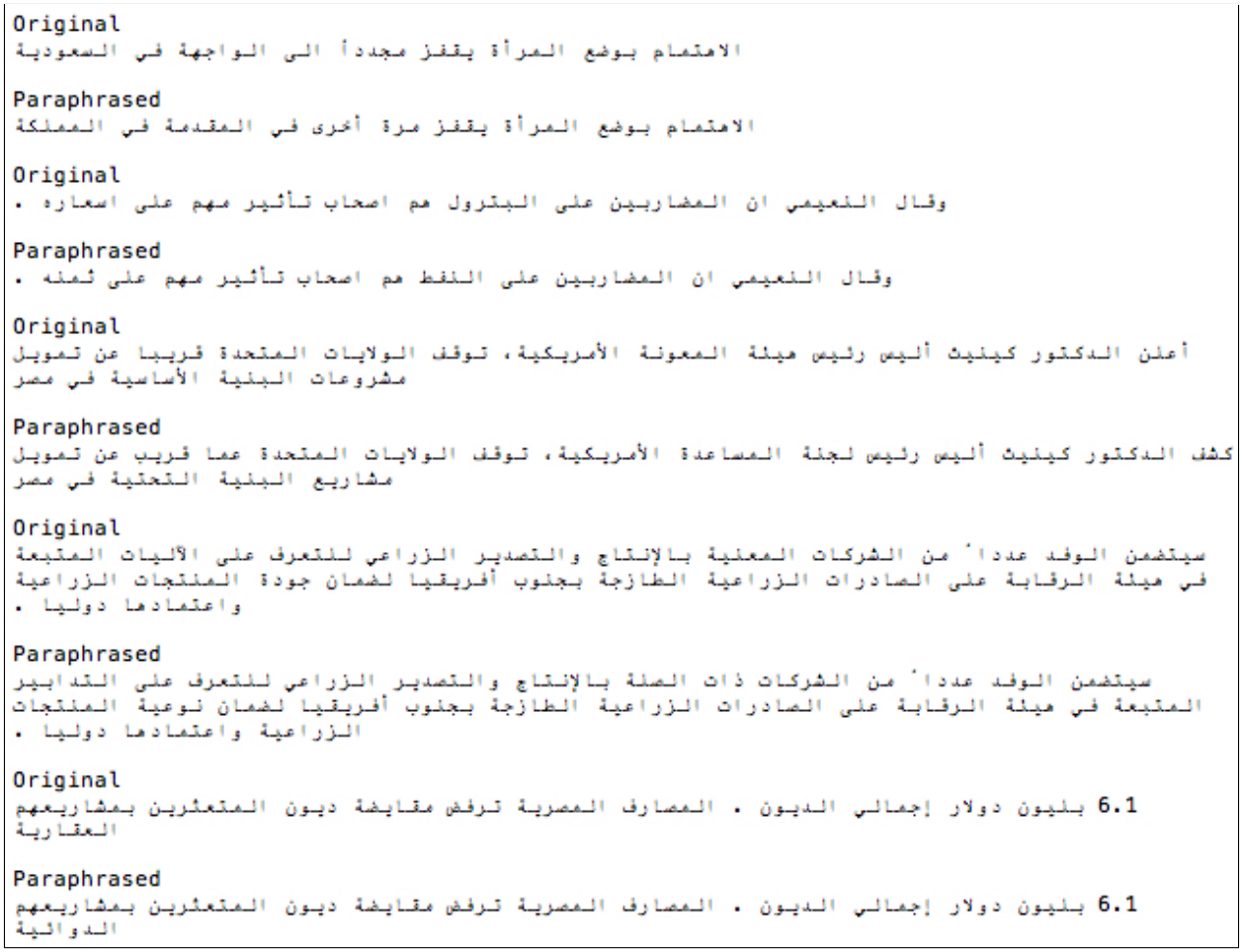

Figure 2: Sample output produced using phrase substitution

scale	Grammaticality %	Meaning %	Both %
5	62	62	50
4	21	16	7
3	10	9	3
2	7	12	4
1	0	1	0

Table 3: Evaluation of paraphrased sentences (averaged and rounded from two annotator ratings)

similar in meaning. Also, they were sufficiently different in surface form which encouraged their use as input to a seq-to-seq paraphrasing model. Initially, we were concerned of the possibility that the en→ar and fr→ar translation tools we used were trained on the same Arabic dataset with similar model architecture and parameters, in which case the output would not be diverse enough to be useful as input to the paraphrasing system. We were also concerned of the possibility that the French source was translated to English and then to Arabic. However, judging by the clear diversity in the output of these models, we were further encouraged to use the resultant parallel monolingual corpus to train the neural paraphrasing model. We plan to quantitatively compare the surface diversity in the output of the fr-ar and en-ar MT models by computing word overlap and hamming distance between corresponding sentences. Due to time constraints we leave this for future work.

The paraphrased output from the neural system was far from grammatical and meaning was incomplete due to early sentence truncation, especially for long sentences. Better output was observed for shorter sentences. However, we make the following observations about the output:

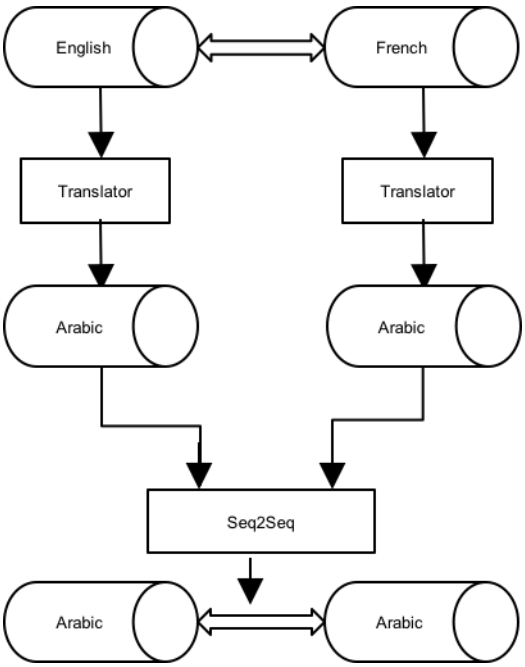

Figure 3: Overview of the process completed to obtain two parallel Arabic corpora generated from different source languages

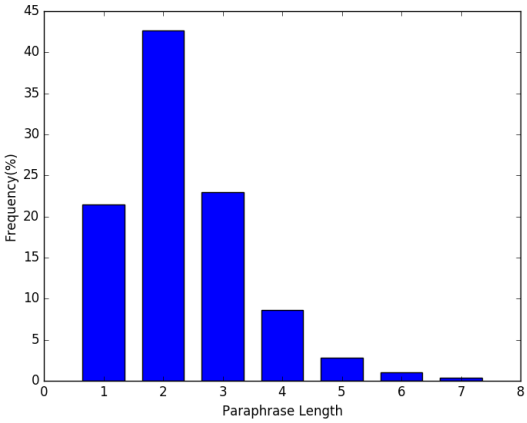

Figure 4: Distribution of phrase lengths in our paraphrases obtained through pivoting

- The model learns what phrases to use at the beginning of the sentence. It uses things like "Moreoever" and "As you know," (translated from Arabic), exactly at the beginning of the sentence.
- The model seem to lean and include central parts of the sentence in the output such as the subject or the location of the event.
- The model learns correspondences between the main parts of the sentence; e.g., the byline vs. remainder of the sentence and quoted text vs. part before the quotation.
- The model often fails in producing output with the correct word order.
- As observed with neural language models, it tends to repeat words.

Figure 5 shows a sample output from the neural model.

6 An Automatic Evaluation Metric

Since human evaluation is time-consuming, subjective and not always available, we propose an automatic evaluation metric for paraphrasing. We propose criteria for judging paraphrase quality and operationalize those criteria using well-defined functions. A good paraphrase has the following two properties:

1. maintains the meaning of the original text, yet
2. expresses the same meaning using different surface realizations.

To evaluate the semantic similarity and surface variation in a paraphrase we employ well-defined metrics discussed next.

6.1 Semantic Similarity

Several methods exist for capturing the semantic similarity of text (Ferreira et al., 2014; Corley and Mihalcea, 2005; Agirre et al., 2009). One simple approach uses the distributional properties of words in the sentence and embeds a sentence $x = \langle x_1, x_2, \ldots, x_n \rangle$ by averaging the embedding vectors of its words. We choose this method for its simplicity and efficiency. The sentence vector is thus given by:

$$w_x = \frac{1}{|x|} \sum_{1}^{n} w_{x_i} \quad (3)$$

where n is the number of words in the sentence. Although this method is simple and does not consider word-order or syntactic structure, it performs surprisingly well when compared to more advanced

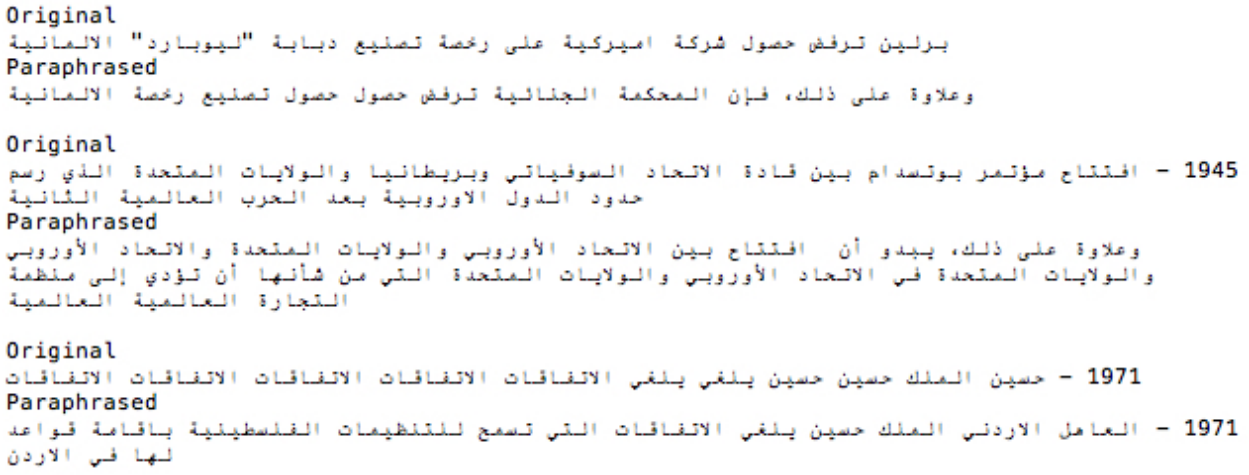

Figure 5: Sample output from neural parpahrasing model

neural- based methods designed to capture sentential semantics (Wieting et al., 2015). It also correlates highly with human judgement (Wieting et al., 2015). Also, being agnostic to word order is actually a desired property in the paraphrase case since valid paraphrases may only differ in the order of the words or the construction used. For example, in Arabic the SVO word order can almost always be changed to VSO without changing the meaning of the sentence[1] and without introducing any other particle. Another example is English active voice and the corresponding passive voice sentence (+ *by* Subj) of the original sentence.

To compute the semantic similarity between two sentences, the original sentence and the paraphrase, the cosine similarity between the two sentence vectors is computed. In our experiments, we use word embeddings with 300 dimensions trained on Arabic text from (Bojanowski et al., 2016).

6.2 Surface Variation

To capture surface variation we first map each sentence into a common vocabulary space and compute the hamming distance between the sentence vectors in that space. This also limits sentence length bias where short sentences will naturally have less surface overlap. In our experiments, we map sentences into a vocabulary space of 610977 words (Bojanowski et al., 2016). We present experiments and results in section 6.4.

[1]except for emphasis

6.3 Combining Criteria for Meaning and Form

Minimal change in surface form can result in maximal preservation of original meaning. However it will score low on surface variation. Similarly, if surface form is significantly changed, we may risk altering the meaning of the original sentence. Since semantic similarity and surface variation are two competing criteria, we combine them using (balanced) harmonic mean. The final score of the paraphrase is given by:

$$s = 2\,\frac{\mathit{SemanticSimilarity} \cdot \mathit{LexicalDistance}}{\mathit{SemanticSimilarity} + \mathit{LexicalDistance}} \quad (4)$$

6.4 Results

We evaluate a set of 198 Arabic sentence pairs sampled from newswire data (Graff and Walker, 2001) as follows. These include headlines on similar topics and for each headline one or two sentences detailing the event in the headline or reiterating it. Pairs including the headline and the following sentence were reasonable paraphrase pairs whereas the other pairs varied in paraphrasing potential from moderate (some overlap in meaning) to poor (unrelated, contradicting or little overlap). We created 576 sentence pairs from the dataset but obtained annotations for only 198 of them. This evaluation of the metric was conducted before obtaining paraphrasing results from our phrase-based and neural models. Therefore, we created sentence and paraphrase pairs following this approach. Since sentences were

obtained from newswire data, we assumed they are grammatical and did not obtain grammatical judgement from annotators. On paraphrastic quality, human evaluations were obtained from three annotators who are native speakers of Arabic. Each annotator was asked to judge the quality of the paraphrase, on whether it preserved meaning and was expressed differently, on a ordinal scale from 1 to 5 where 1 indicates poor quality. We used R to measure inter-annotator (absolute) agreement using IntraClass Correlation (ICC) and agreement was measured at 0.714 ICC which is considered "very good." The biserial correlation between the binarized human evaluations and the evaluation metric scores was 0.813.

6.5 Analysis

Observing high correlation between human evaluation and the proposed evaluation metric, we examined the dataset to see if results were biased by sampling or data peculiarities. For sentence pairs including the headline and the following sentence, both human and evaluation metric scores were high. For most of the other sentence pairs, the paraphrase was judged as weak or poor. In both of these cases, the judgement was "easy" and straightforward and this perhaps lead to the surprisingly good results. Perhaps sentence pairs with finer and more subtle semantic phenomena such as polysemy and synonymy would have been harder to score accurately by the metric. We need to conduct more experiments to verify this.

We also explored the effect of the embedding dimension on the correlation between human evaluation and the metric we proposed. We experimented with lower dimensions: 50, 100, 150, 200, 250 and obtained slightly lower values of biserial correlation as we decreased the embedding dimension. With 50 dimensions, the absolute difference between the previous biserial correlation value and the new one was 0.03. It is worth mentioning that when only using overlap as a measure of semantic similarity, the correlation between human judgement and the evaluation metric is 0.47. We clearly gain by using word embeddings but even something as simple as word overlap can capture semantic distance to some extent and explain a good amount of variation in human judgement.

We initially planned to experiment with the size of the vocabulary space in which the lexical distance between the sentence and a paraphrase is computed but due to time constraints we leave that as future work. We initially set out to compute the surface distance between two sentences using a measure that only depends on the two sentences as input; such as token/character overlap or minimum edit distance. However, we decided to compute the hamming distance in a canonical space since the first approach can have a length bias.

We also prefer using a common canonical space for doing the various computations as these spaces comprise a reference against which candidates are evaluated. When using the metric to evaluate outputs from different systems, the reference can be decided at test time to avoid "gaming" the metric. It is desirable to have a metric that does not require a reference sentence when evaluating a candidate paraphrase since 1. results and rankings can be sensitive to the choice of the reference sentence which is often subjective and 2. the notion of a "reference paraphrase" is problematic here since paraphrasing is essentially based on divergence from a given surface form while preserving the underlying meaning and hence is loosely constrained. However, we do recognize the problem with not having a reference for an evaluation metric: it can make it susceptible to "gaming" by players competing to optimize the metric score. Therefore, we propose to decide the vocabulary space in which surface distance function is computed and the semantics space in which semantic similarity is computed at test time.

Also, since this metric combines competing objectives, the parameter controlling the relative strength of each component can also be decided at test time to improve the metric robustness. While it has been shown that it is possible to find a Pareto optimal hypothesis that aims to jointly optimize several different objectives (Duh et al., 2012), we argue that those objectives are not exactly *competing* or orthogonal. They may be weakly correlated but they are still positively correlated since they compare surface distance against the same reference (as in BLUE and TER), which is not the case in our proposed metric setting.

7 Conclusion

We presented and compared two different approaches for sentence paraphrasing in Arabic. The phrase-based approach yielded very good results when used to create paraphrase sentences. The neural method output is still far from practical applicability but the model has learned interesting linguistic constructs like phrases used for sentence opening. We also presented a novel approach for obtaining parallel monolingual data and contributed a dataset of two million parallel sentence pairs in Arabic using this approach. We applied the pivoting method to construct a large coverage paraphrase database for Arabic that includes over 88K phrase pairs. When used to create new sentences, the paraphrase dictionary gave very good results on grammaticality and meaning preservation. We proposed a new automatic evaluation metric for paraphrasing that does not require the use of a reference sentence while evaluating candidate hypotheses. We showed encouraging preliminary results in terms of correlation with human judgement.

8 Future Work

We plan to explore other options for obtaining monolingual parallel data. One possible approach is to retrieve headlines of news articles from different agencies covering the same event. We expect headlines describing the same event to have some degree of semantic similarity yet different surface realizations.

The sequence-to-sequence model required a relatively long time to run which limited the testing of other architectures and model parameters. We plan to conduct more experiments on different architectures and compare results. More specifically, whether we get better results with a uni-directional model since sequences from the same language will tend to have similar word order. We also intend to incorporate the concept of "coverage" (Mi et al., 2016) to address issues with fluency of the neural model output.

We obtained encouraging results from the evaluation metric experiments but we need to verify its usefulness in settings with subtle semantic phenomena such as polysemy and synonymy. We also plan to use it at word subunits such as morphemes and even character level especially for surface distance comparison in morphology rich languages like Arabic.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Stroudsburg, PA, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Houda Bouamor, Aurélien Max, and Anne Vilnat, 2010. *Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases*, pages 67–78. Springer Berlin Heidelberg, Berlin, Heidelberg.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Denkowski, Hassan Al-Haj, and Alon Lavie. 2010. Turker-assisted paraphrasing for english-arabic machine translation. https://www.cs.cmu.edu/ mdenkows/pdf/paraphrase-mturk-2010.pdf.

Peter Wallis Dept and Peter Wallis. 1993. Information retrieval based on paraphrase. In *In Proceedings of PACLING Conference*.

Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada, and Masaaki Nagata. 2012. Learning to translate with multiple objectives. In *Proceedings of*

the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rafael Ferreira, Rafael Dueire Lins, Fred Freitas, Steven J. Simske, and Marcelo Riss. 2014. A new sentence similarity assessment measure based on a three-layer sentence representation. In *Proceedings of the 2014 ACM Symposium on Document Engineering*, DocEng '14, pages 25–34, New York, NY, USA. ACM.

David Graff and Kevin Walker. 2001. Arabic newswire part 1. .

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. http://www.statmt.org/europarl/.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. A coverage embedding model for neural machine translation. *CoRR*, abs/1605.03148.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Fabio Rinaldi, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. 2003. Exploiting paraphrases in a question answering system. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, DIALECTS '11, pages 10–21, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *CoRR*, abs/1511.08198.

Ma Xiaoyi, Dalal Zakhary, and Moussa Bamba. 2004. Arabic news translation text part 1 ldc2004t17. .

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 447–454, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ingrid Zukerman, Bhavani Raskutti, and Yingying Wen, 2002. *Experiments in Query Paraphrasing for Information Retrieval*, pages 24–35. Springer Berlin Heidelberg, Berlin, Heidelberg.