Neural and Symbolic Arabic Paraphrasing with Automatic Evaluation

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Weijian Lin
Carnegie Mellon University
School of Computer Science
5000 Forbes Avenue, Pittsburgh, PA 15213
wlin1@cs.cmu.edu

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

2 Related Work

3 Obtaining Parallel Monolingual Data

4 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.

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

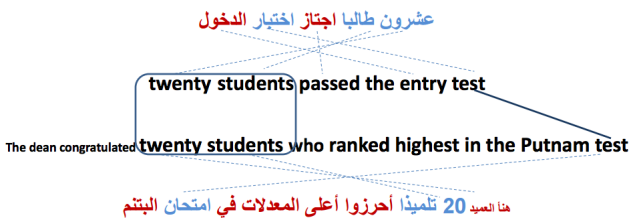

Figure 1: An example paraphrased sentence produced using the pivot method.

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. 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 then ran GIZA++ to obtain these alignments (Och and Ney, 2003). Once we have a database of paraphrase mappings, we can then substitute phrases for 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

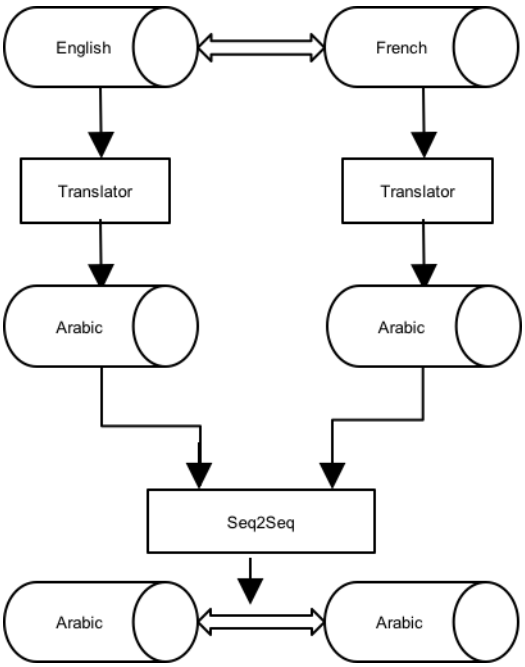

Figure 2: Overview of the process completed to obtain two parallel Arabic corpora generated from different source languages

5 Generating Paraphased Sentences

To address the shortage problem for Arabic language data, here we propose a novel method for generating Arabic parallel language sentences using other two languages 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 paraphraser model, the generation step for Arabic paraphrases is easily scalable. The second advantage is that 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 use europarl-v7.fr-en data which contains two millions sentence pairs, and then use Google translate API to generate French-Arabic and English-Arabic sentence pairs correspondingly. Then we

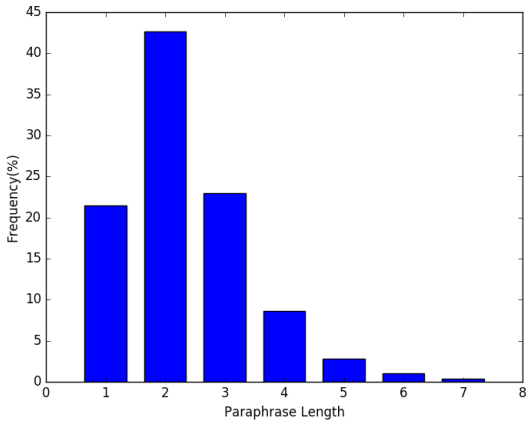

Figure 3: Distribution of phrase lengths in our paraphrases obtained through pivoting

use them as training data to train a single layer sequence to sequence attentional encoder-decoder model, with embedding size and hidden size set to be 512 while attention size set to be 128. The whole process is demonstrated in Figure 2

6 Results

We report results on the bilingual pivot and sequence to sequence approach detailing above.

6.1 Phrase Substitution Method

For the pivot method we report a few results. First

6.2 Neural Seq-to-Seq Method

7 An Automatic Evaluation Metric

7.1 Semantic Similarity

7.2 Surface Variation

8 Analysis

9 Evaluation

10 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 models required a relatively long time to run which limited the testing of other architectures and model options. We plan to conduct more experiments on different architectures and compare results.

Acknowledgments

References

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.

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

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