Carnegie Mellon University Qatar

Unsupervised Arabic Word Segmentation and Statistical Machine Translation

Senior Thesis

School of Computer Science

Hanan Alshikhabobakr

halshikh@qatar.cmu.edu

Advisor: Kemal Oflazer

ko@cs.cmu.edu

Co-advisor: Mohit Behrang

behrang@cmu.edu

May 2013

ABSTRACT

Word segmentation is a necessary step for Natural Language Processing (NLP) for morphologically rich languages, such as Arabic. In this thesis, we experiment with unsupervised word segmentation systems proposed in the literature, to perform segmentation on Arabic, and couple word segmentation with Statistical Machine Translation (SMT). Our results indicate that unsupervised segmentation systems turn out to be inaccurate and do not help with improving SMT quality. Although minimal automatic post-processing improves the translation accuracy, word baseline accuracy turn out to be better. We conclude that semi-supervised word segmentation systems have more potential to improve Arabic to English translation in SMT.

Acknowledgements

I sincerely thank my advisors Prof. Kemal Oflazer and Dr. Mohit Behrang for their constant support and guidance throughout this research. Although it has been my first time to get exposed to NLP, my advisors were extremely helpful and patient in my learning process. I would also like to forward my sincere gratitude to Prof. Mark Stehlik for his constant motivation and support to me throughout the year.

I definitely owe a lot to my friends and family who were by my side whenever I needed them and for their support to me during all the hard times I faced.

Contents

1. Introduction

Word segmentation plays an important role for morphologically rich languages in many NLP applications. Arabic is a morphologically rich language, so we use it in this research as the target language for segmentation. Although there are accurate word segmentation systems for Arabic, such as MADA (Habash, 2007), they are manually-built systems that incorporate rules of the Arabic language and their exceptions. In this work, we look at unsupervised word segmentation systems to see how well they perform word segmentation, without relying on any linguistic information about the language. Hence the methodology of this research can be applied to many other morphologically-complex languages. We focus on three leading unsupervised word segmentation systems in the literature: Morfessor (Creutz and Lagus, 2002), ParaMor (Monson, 2007), and Demberg's system (Demberg, 2007). For each of the three systems, we train segmentation models from the same training set and test accuracy on a test set. We then apply the word segmentation model in an NLP application, statistical machine translation (SMT). As a result we observe that Morfessor works best with SMT, and when we apply minimal post-processing on its segmentations, it gets closer to the baseline, as it improves translation by a factor of 3 from the original result obtained from Morfessor.

Based on our observation we conclude that 1) unsupervised segmentation models does not seem to improve MT output quality, 2) unsupervised segmentation accuracy does not predict SMT output quality, and 3) some additional post-processing could help.

2. Literature Review

2.1 Word Segmentation

Word segmentation break words into grammatically meaningful segments, which we refer to as morphemes. For example, "meaningless" could be segmented into "mean+ing+less", where each segment (or morpheme) has a grammatical meaning/function. Figure 1 illustrates a word segmentation example for the word "talking" and for its Arabic equivalent in meaning:

In this work we investigate three unsupervised word segmentation systems and one manually-built system.

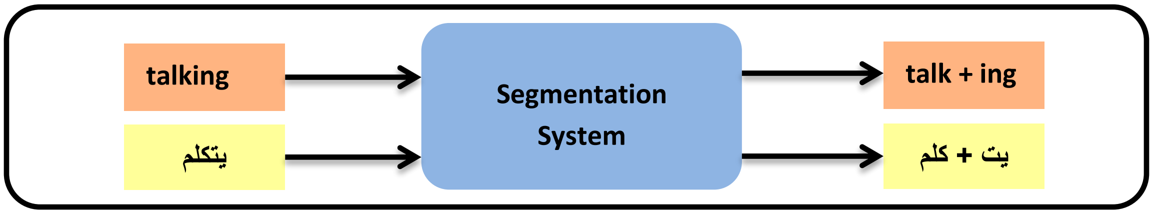

Figure 1: Examples of word segmentation for English and Arabic

2.2 Unsupervised Word Segmentation Systems

An unsupervised word segmentation system is one which learns the segmentation from a list of words that are not annotated or pre-processed in any way that helps the system to predict the correct segmentation. The main task of an unsupervised system is to create a segmentation model that then can take new words and output their segmentation.

We study the word segmentation performance of three unsupervised systems: Morfessor (Creutz and Lagus, 2002), ParaMor (Monson, 2007), and Demberg's system (Demberg, 2007). We briefly describe each of the systems below. We also experiment with a manually-built system for Arabic words Segmentation, MADA (Habash et al., 2008), and use it as a standard for some of our evaluations.

Morfessor

Morfessor tries to discover the most compact description of the data (that is, the set of words). It does that through finding substrings that appears frequently enough in several word forms, so that it can propose them as morphemes. This is called the Minimal Description Length (MDL) principle: Morfessor tries to minimize the total description length of unique morphemes to account for the training data.

Demberg's word segmentation model

Demberg's segmentation model is based on RePortS (Keshava and Pitler, 2006) but adds some extensions to it. RePortS uses words that appear as substring of other words and transition probabilities between letters in a word, to detect morpheme boundaries. RePortS assumes that root words do appear in the corpus, which may not be the case for all languages. Demberg's model adds to RePortS algorithm, an extension to fix this assumption by having an intermediate step which creates a candidate list of root words.

ParaMor

Segmentation in ParaMor is carried out by identifying the morpheme boundaries using letter transition probabilities, and then identifying morpheme-internal bigrams or trigrams. ParaMor then discovers the relationship between pairs of words. Finally, it uses an information-theoretic approach to minimize the number of letters in the morphemes of the language.

MADA

MADA (Morphological Analysis and Disambiguation for Arabic) (Habash, 2007) is the state-of-the-art manually-built morphological analysis system of the Arabic language. Along with word segmentation, MADA is an excellent word-in-context analyzer, and therefore provides accurate segmentation of a word in its context in a sentence. MADA has a high accuracy of usually over 94%. TOKAN, a component of MADA, allows a user to specify the tokenization (or segmentation) scheme. Each scheme has its own characteristics. This work uses two of the schemes: D1 and D2; D1 is a less aggressive in segmentation than D2, that is, D1 produces less overall segments than D2, on the average.

2.2 Statistical Machine Translation

Machine Translation is the task of automatically converting a text from one language to another. Statistical Machine Translation uses statistics from a parallel corpus to build a statistical model of translation.

An SMT model for Arabic and English is created through the following steps:

1. An Arabic-English parallel corpus (i.e., Arabic sentences and their aligned English translations) is given as input to the SMT learner which produces a corresponding SMT model.
2. The resulting SMT model is then used to translate Arabic into English with an SMT decoder.

Table 1 illustrates the matching alignment between Arabic and English sentences in the table below. Notice here that some English words correspond to only a morpheme (substring) in Arabic words. So we can see that word segmentation could be useful for Arabic to English translation.

English	The boy is playing with the ball	The boy is play+ing with the ball
Arabic	يلعب الولد بالكرة	يـ+لعب الـ+ولد بـ+الـ+كرة

Figure 1: Example of a sentence translated from Arabic to English. The matching substrings are highlighted with the same color.

In this research, we use the MOSES toolkit (Koehn et al., 2007), an SMT tool that allows a user to build an SMT system for any pair of languages using a parallel corpus.

3. Methodology

We now describe the method in which we perform the unsupervised segmentation learning task, the core of this research. We then describe how to carry out the machine translation task. Finally, we explain how we couple word segmentation task with SMT.

3.1 Data

In this work, we used two sets of data:

Set 1: A list of 1.7 million unique and punctuation-free words extracted from a corpus of 400 million words. These then were transliterated to Buckwalter transliteration for processing purposes (Buckwalter, 2004).

Set 2: An Arabic-English parallel corpus of 120,000 sentences, of which 119,000 were used for SMT training, and a 1,000 for SMT testing.

3.2 The Segmentation Task

For each of the unsupervised word segmentation systems, we have two phases:

1. **Training:** We input a list of unique Arabic words, each word on line without annotation, into the learner. We get a segmentation model after this step. (Figure 2, step 1)
2. **Testing:** We use the resulting segmentation model from the first phase and use it to segment a smaller Arabic word list, again each word in a line. (Figure 2, step 2)

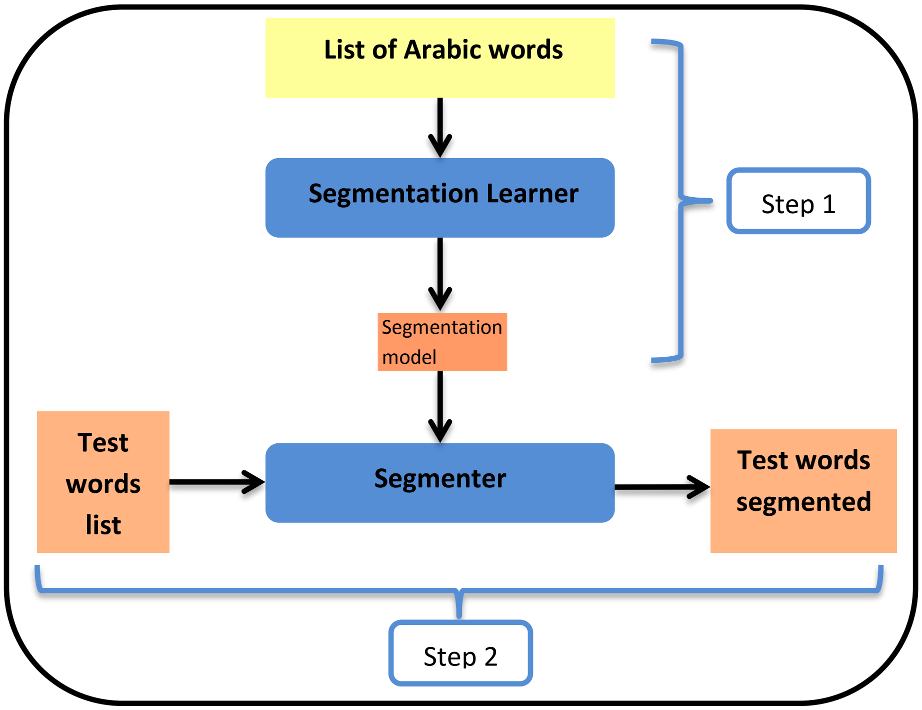

Figure 2: Unsupervised word segmentation

3.3 The Translation Task

Figure 3 shows the block diagram of the SMT data flow. We explain the diagram in three steps:

1. We run the Arabic side corpus through a segmenter and replace it with the original Arabic corpus, while keeping the English unsegmented, and input this modified parallel corpus into the SMT learner which produced an SMT model.
2. We run Arabic test corpus that we wish to translate through the same segmenter used in step-1. Now er run the segmented Arabic test set through the SMT decoder to get the English translation.
3. We compute the translation accuracy through running BLEU on translation comparing with gold-standard translations.

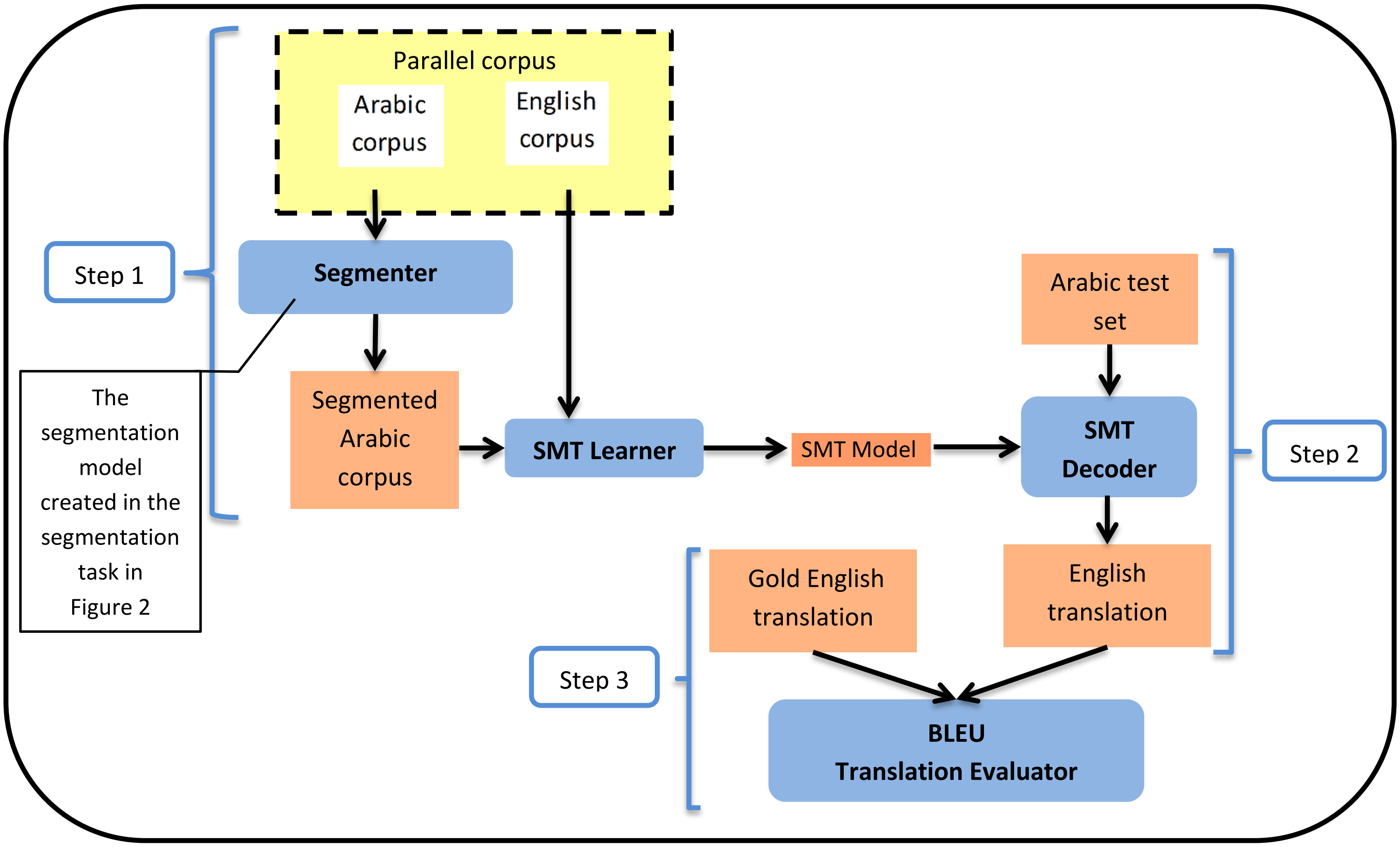

Figure 3: SMT methodology. Note that the "Segmentation Model" is created by the Segmentation task.

4 Evaluation

We evaluate both the accuracy of segmentation intrinsically and then evaluate the impact of different segmentation schemes on SMT.

4.1 Evaluation of Word Segmentation

The accuracy of a segmentation system is computed in the following way:

$$Accuracy = \frac{number\ of\ correctly\ segmented\ test\ words}{total\ number\ of\ test\ words}$$

where the number of the correctly segmented words is calculated either manually or by comparing it against MADA.

We run the following segmentation experiments:

1. **10-fold experiment**: We use a list of unique words of size 1,700,000 from which we create 10 experiments. In each experiment (or fold) the training set is 9 times the size of the test. We evaluate the correctness of segmentation by comparing it against MADA's segmentation.
2. **200 words test**: We compute the segmentation accuracy of 200 words output by each of the unsupervised systems and compare them against (1) MADA's segmentation and (2) manual segmentation.
3. **100 words test:** We take 100 words from the parallel corpus that is later to be translated and we evaluate the segmentation accuracy manually.

4.2 Evaluation of Statistical Machine Translation

One of the most common metrics to evaluate machine translation is through Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002). BLEU evaluates a translation by matching n-grams between a translation and a gold standard translation. Thus BLEU not only evaluates the accuracy of the words in the translation, but also evaluates the order of the words, quantifying the fluency of a translation. BLEU also allows for multiple human translation references as standard. In this research, we use four correct translation references to evaluate translation with BLEU.

5. Experiments and results

In Table 2, we present the results obtained for all the experiments. As we can see, Morfessor produces the best segmentation in two of the experiments, while ParaMor surpasses Morfessor in two of the experiments. Demberg's system overall has lower accuracy. Notice here that in the test of 200 words, once against MADA and once against manual segmentation, the accuracy does not match because although MADA is accurate, it does not cover all segmentation cases.

System	Morfessor	ParaMor	Demberg
10-fold vs. MADA	25.88%	**32.97%**	27.20%
200 words vs. MADA	**49.00%**	47.00%	31.00%
200 words vs. Gold	48.00%	**65.00%**	47.00%
100 words vs. Gold	**66.00%**	24.00%	37.00%

Table 2: Accuracy of the unsupervised segmentation systems for each experiment.

For the translation task, we use BLEU to evaluate the translation accuracy and fluency. In Table 3, we report the BLEU translation score for each system. Note that the baseline score refers to SMT model without using word segmentation. Also note that we have two scores for MADA: D1 and D2 due to using two different schemes for segmentation, where D2 is a more aggressive segmentation than D1.

	Baseline	MADA-D2	MADA-D1	Morfessor	ParaMor	Demberg	Morfessor+
BLEU	41.31%	36.87%	**43.78%**	38.29%	20.89%	36.73%	41.17%

Table 3: BLEU scores for the word baseline and for all the segmentation systems used.

We notice that amongst the three unsupervised systems, Morfessor is performing the best in translation. Although ParaMor performs better than Morfessor in word segmentation task, Morfessor outperforms ParaMor in translation. We claim that this is because although ParaMor has a better segmentation accuracy, it segments the words aggressively. As we can see from the Table 4, the number of unique segments that ParaMor produces is much higher than what Morfessor produces.

System	Morfessor	ParaMor	Demberg
Unique morphemes of words used in the translation evaluation for 7954 unique words	4,280	6,618	6,615

Table 4: Number of unique morphemes obtained by each segmentation system

As Morfessor is the best unsupervised segmentation system (Table 3), we now created a modified version, Morfessor+, a post-processing modification of Morfessor, where we try to make the segmentation less aggressive. We added three simple rules: attach “A” (Alef equivalent in Buckwalter) at the beginning of a word, attach “Al” (Alef-Lam equivalent in Buckwalter) at the beginning of a word, and remove segmentation from any two letter words. We see an improvement in translation from Morfessor to Morfessor+. But nevertheless, none of the systems proposed beat the baseline and MADA-D1.

6. Conclusions

We conclude that accurate manually-built word segmentation does improve translation (as the case for MADA-D1), especially while keeping word segmentation is balanced. However, even manually-built word segmentation may not improve translation, if segmentation was aggressive. As we see MADA-D2 has a lower BLEU compared to the baseline. The usefulness of balanced word segmentation in SMT also applies to the unsupervised systems. We have seen that even if segmentation is more accurate (in the case of ParaMor), it performs poorly when coupled with translation, and the more balanced the segmentation is (in the case of Morfessor), the better the translation score obtained. We also see that lowering the number of segmentation in Morfessor generates a better SMT (the case of Morfessor+).

We also see potential of unsupervised word segmentation to improve when post-processing is applied (as in the case form Morfessor to Morfessor+), and is very close to outperform the baseline. Therefore we propose that semi-supervised word segmentation has more potential to improve machine translation in SMT.

7. References

C. Mathias and K. Lagus. 2005b. Morfessor in the Morpho Challenge. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, Unsupervised segmentation of words into morphemes – Challenge 2005, pages 12–17, Helsinki University of Technology, Helsinki.

V. Demberg. 2007. A language independent unsupervised model for morphological segmentation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 920–927, Prague.

S. Keshava and E. Pitler. 2006. A simpler, intuitive approach to morpheme induction. In Proceedings of 2nd Pascal Challenges Workshop, pages 31–35, Venice, Italy.

C. Monson. 2009. ParaMor: From Paradigm Structure to Natural Language Morphology Induction. Ph.D. thesis, Carnegie Mellon University.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In Proceedings of Association for Computational Linguistics (ACL), Columbus, Ohio, 2008.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of ACL, pages 311–318, Philadelphia, PA.

T. Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium (LDC2004L02).