A Semi-Supervised Learning Approach for Morpheme Segmentation for An Arabic Dialect

Mei Yang[1†], *Jing Zheng*[2], *Andreas Kathol*[2]

[1]Dept. of Electrical Engineering, University of Washington, Seattle, WA, USA
[2]SRI International, Menlo Park, CA, USA
yangmei@u.washington.edu, {zj,kathol}@speech.sri.com

Abstract

We present a semi-supervised learning approach which utilizes a heuristic model for learning morpheme segmentation for Arabic dialects. We evaluate our approach by applying morpheme segmentation to the training data of a statistical machine translation (SMT) system. Experiments show that our approach is less sensitive to the availability of annotated stems than a previous rule-based approach and learns 12% more segmentations on our Iraqi Arabic data. When applied in an SMT system, our approach yields a 8% relative reduction in the training vocabulary size and a 0.8% relative reduction in the out-of-vocabulary (OOV) rate on the test set, again as compared to the rule-based approach. Finally, our approach also results in a modest increase in BLEU scores.

Index Terms: Iraqi Arabic, morpheme segmentation

1. Introduction

Previous studies ([1], [2] and [3]) have shown that the data sparseness problem in translating morphologically complex languages in SMT can be effectively addressed using morphological knowledge. However, morphological analysis, such as tagging, stemming, and morpheme segmentation, is a difficult problem in itself, especially for languages with complex morphology but limited resources. In this paper, we present a semi-supervised learning approach for morpheme segmentation for Arabic dialects, which utilizes a heuristic model with an annotated lexicon. Our goal is to increase the translation accuracy of SMT systems by applying the morpheme segmentation to the training data in order to reduce the number of word forms in the training vocabulary and the OOV rate on the test set.

Arabic has rich morphology resulting in a large number of word forms. An exhaustive morphological analysis for Arabic usually requires a considerable manual effort, as for instance in the case of the Buckwalter morphological analyzer for Modern Standard Arabic (MSA) [8]. In contrast to MSA, Arabic dialects are spoken languages that rarely appear in written materials. Data can be obtained only by manually transcribing audio recordings, a costly effort that prevents large-scale data collections. Because of the paucity of training resources, the computational analysis of Arabic dialectal morphology has not been developed very extensively and hence derived tools such as lexicons, tokenizers, and taggers are rare or non-existent.

† The first author performed the work described while doing an internship at SRI International. We thank Kristin Precoda for helpful comments on an earlier draft of this paper.

2. Related work

Some work has recently been done on the morphological analysis of Arabic dialects. [4] present a morphological analyzer and generator for Arabic dialects ("MAGEAD"), which utilizes a unified processing architecture to generalize morphology to all variants of Arabic. However, since MAGEAD has been developed mainly on the basis of linguistic analysis, adapting MAGEAD for a new Arabic dialect requires a native speaker or trained linguist.

On the other hand, [5] and [6] report that a shallow analyzer of morpheme segmentation for Arabic dialects has proven effective in increasing the accuracy of SMT systems. [5] applies a rule-based approach (see section 3.1 for details) to perform morpheme segmentation for Iraqi Arabic with a pre-compiled list of affixes and stems. In subsequent work [6], a more sophisticated trie model is employed to strip affixes from words. The trie model consists of trie-based classifiers, which are trained for each affix on a small amount of annotated data and are constructed in a cascading manner. The model parameters are chosen to maximize the accuracy of classifiers on a held-out set with an exhaustive search in the parameter space. Applied to two SMT systems for Iraqi-English and Levantine-English, the trie model outperforms the rule-based model even with little training data, and moreover, further improvement is observed when a hybrid model is used which first uses the rule-based model for segmentation and backs off to the trie model when no analysis can be made. The disadvantage of the trie model lies in the fact that the model parameters might be biased toward the training data. In addition, in contrast to the rule-based model it is difficult to incorporate linguistic knowledge into the trie model without increasing the number of parameters and the complexity of the model.

Another promising recent approach to developing morphological analyzers for Arabic dialects involves data-sharing techniques. [7] investigate part-of-speech (POS) tagging for Egyptian Arabic using cross-dialectal data sharing. The authors exploit the commonalities among similar dialects and train a contextual model jointly on Egyptian and Levantine data. Their experiments show that the data sharing approach yields the best performance of all methods explored.

3. Morpheme segmentation for Arabic dialects

In this paper, we present a semi-supervised learning approach for morpheme segmentation for Arabic dialects. Our approach adapts the rule-based segmentation model in [5] into a heuristic architecture where various types of resources are incorpo-

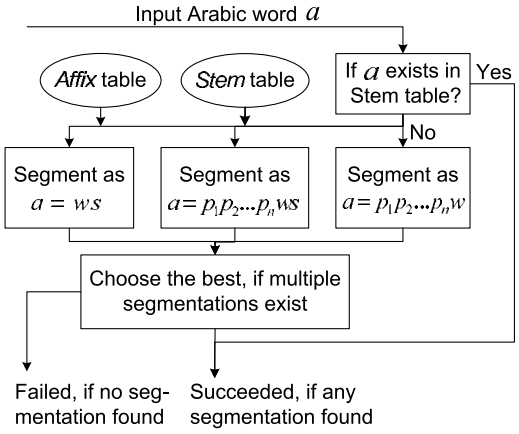

Figure 1: Rule-based model for morpheme segmentation, where w, p_i, and s represent the stem, prefix(es), and suffix respectively.

rated, including linguistic rules, a small annotated lexicon, a pre-compiled list of affixes and stems, and word forms found in unannotated data.

3.1. Rule-based segmentation model

The rule-based model for morpheme segmentation is illustrated in Figure 1 where the *Affix* table contains all the affixes considered for segmentation and the *Stem* table is a pre-compiled list of Arabic stems with their frequencies. The rule-based model first checks whether an input word exists in the *Stem* table or not. If it does, no affixes need to be segmented, and the word will be output as is. Otherwise, if the input word is not a known stem, the model tries to analyze the word according to three possible segmentations: prefix(es)-stem, stem-suffix, and prefix(es)-stem-suffix. If multiple segmentations are found, the frequencies of their stems are compared and the one with the highest frequency is chosen as the output. If no segmentation is found, the word cannot be segmented and the form is left unaltered.

3.2. Heuristic model for morpheme segmentation

Figure 2 shows our proposed heuristic model for morpheme segmentation. In addition to the *Affix* and *Stem* tables, the heuristic model records all segmentations learned so far in the *Seed* and *Word* tables, and uses them for segmenting unseen words. These two tables can be initialized with an annotated lexicon so that the *Seed* table contains all words that have already received segmentations, while the *Word* table contains words that may be considered to be stems when no segmentations can be found by applying the rule-based model with the *Stem* table. Note that words in the *Word* table are treated differently from those in the *Stem* table and have different priority as stems for segmentation.

Given an input word a to be segmented, the heuristic model starts by checking whether a has received a segmentation or not. If it has, we output the corresponding segmentation. Otherwise, we attempt a rule-based segmentation using the *Stem* table. If no segmentation can be found, another rule-based segmentation is tried, this time using the *Word* table. When a segmentation is found for a using a word w from the *Word* table, we try to

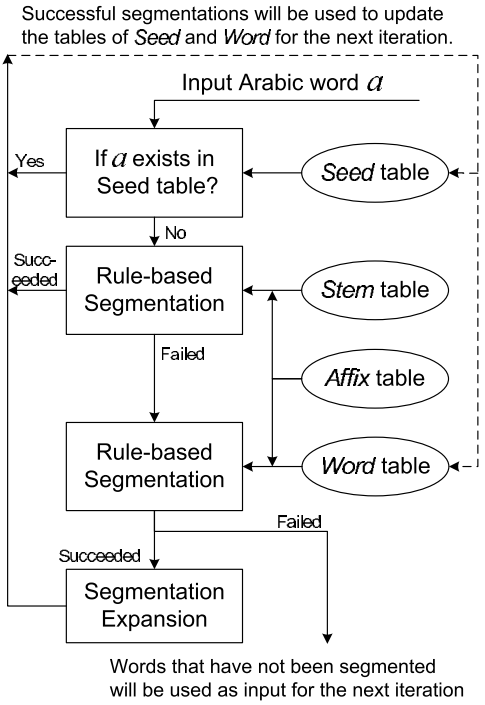

Figure 2: Heuristic model for morpheme segmentation.

expand it with the segmentation learned for w. Assume that a can be segmented as

$$a = p_1p_2...p_nws_1s_2...s_m \quad (1)$$

If w has already been segmented as

$$w = q_1q_2...q_kvt_1t_2...t_l \quad (2)$$

we combine (1) and (2), and output the final segmentation for a

$$a = p_1p_2...p_nq_1q_2...q_kvt_1t_2...t_ls_1s_2...s_m \quad (3)$$

where p_i, q_i are prefixes, s_i, t_i are suffixes, and v is the stem. If w cannot be further segmented, (1) is the final segmentation for a. If neither application of the rule-based segmentation succeeds, we conclude that the word a cannot be segmented for the time being and proceed to the next word.

In each learning iteration, each word that has not been segmented is processed once and the *Seed* and *Word* tables are updated at the end of that iteration. For simplicity, we currently update the two tables by adding in all the words that have been successfully segmented in that iteration. Words that cannot be segmented remain as input for the next iteration. The learning procedure exits when no more words can be segmented.

Note that additional linguistic constraints could be incorporated to prevent potential over-segmentation in the learning procedure, for instance, allowing only one suffix in Arabic. However, our phrase-based translation scheme (see section 4.3) is not sensitive to over-segmentation, as long as the segmentation is consistent.

4. Experiments

In our experiments we used the rule-based model as the baseline model and the proposed heuristic model to perform morpheme segmentation on the Iraqi Arabic side of parallel training

data for an English-Iraqi Arabic SMT system developed at SRI. We examined the reduction in the number of word forms in the training vocabulary as well as the OOV rate on the test set and evaluated the performance with the translation accuracy of the SMT system.

4.1. Data

The training corpus for the English-Iraqi Arabic SMT system consists of 40 hours of transcribed speech audio from DARPA's Transtac program. Table 1 summarizes the statistics of the data. The Iraqi Arabic side is more sparse than the English side due to its larger vocabulary size and lower word token frequencies. Furthermore, the test set on Iraqi Arabic side has a higher OOV rate as shown in Table 2.

The component tables in the heuristic model were initialized with the affixes shown in Table 3 and a manually compiled list of Iraqi Arabic stems which contains 28,955 unique stems as well as their frequencies. For the *Seed* and *Word* tables, we used an annotated lexicon compiled by the Linguistic Data Consortium containing root-and-pattern analysis for 12,934 unique Iraqi Arabic words. Since we are only interested in stem-and-affix analysis, the more fine-grained morphological analysis of that lexicon were semi-automatically mapped onto stem-and-affix segmentation before the lexicon was used to initialize the heuristic model.

	Iraqi Arabic	English
Word Types	51,500	17,494
Word Tokens	950,310	1,373,108
Token Freq.	18	78

Table 1: Training data for the English-Iraqi Arabic SMT system. .

Iraqi Arabic		English	
Dev	Test	Dev	Test
9.93	8.86	5.28	4.32

Table 2: Type OOV rates (%) of the dev and test sets .

Prefix		Gloss	Suffix		Gloss
و	w	and+	ي	y	+1-sg-pron
ف	f	so/then+	ني	ny	+1-sg-pron (verbal)
ب	b	to/in+	ك	k	+2-sg-pron
ل	l	for+	ه	h	+3-sg-masc-pron
ش	$	what+	ها	hA	+3-sg-fem-pron
عال	EAl	on+the+	نا	nA	+1-pl-pron
هال	hAl	this+	كم	km	+2-pl-masc-pron
د	d	prog+	كن	kn	+2-pl-fem-pron
ال	Al	the+	هم	hm	+3-pl-masc-pron
لا	lA	neg+	هن	hn	+3-pl-fem-pron
ما	mA	neg+			
مو	mw	neg+			
لل	ll	for+the+			

Table 3: Iraqi Arabic affixes considered for morpheme segmentation. Each affix is shown in Arabic orthography followed by its Buckwalter transliteration and gloss.

4.2. Reduction of vocabulary and OOV rate

Table 4 shows the reduction in the number of word forms in the vocabulary for rule-based vs. heuristic segmentation models. Table 5 shows the reduction in OOV rates. Compared to the rule-based model, our heuristic model achieves a relative reduction of 8% of the training vocabulary size and of 1% and 0.8% of the OOV rates for the dev and test sets, respectively.

	Train	Dev	Test
No segmentation	51,500	2,589	3,982
Rule-based model	31,301	1,956	2,812
Heuristic model	27,253	1,900	2,728

Table 4: Number of word forms in the training, dev and test sets for Iraqi Arabic. .

	Dev	Test
No segmentation	9.93	8.86
Rule-based model	7.21	6.69
Heuristic model	6.16	5.90

Table 5: Type OOV rates (%) of the dev and test sets for Iraqi Arabic. .

To further analyze the effect of the component tables in the heuristic model, we applied segmentation to the Iraqi Arabic side of parallel training data with four different configurations. Firstly, we disabled usage of the *Seed* and *Word* tables, using only the *Affix* and *Stem* tables. This is equivalent to the baseline rule-based model. Then we added the *Word* table, the *Seed* table, and finally both tables. Figure 3 shows the learning curves for the four configurations as the number of stems in the *Stem* table varies from 100 to 20,000. In all cases the heuristic model yields the largest number of segmentations, while segmentation count is lowest for the rule-based model. The *Word* table helps to increase the number of segmentations, especially when the *Seed* table is enabled, which bootstraps the number of learned segmentations initially. Moreover, we can see that even with 100 stems in the *Stem* table, the heuristic model still learns segmentations for 46.7% of word forms in the training vocabulary while the rule-based model only does so for 2.3%. This suggests that the heuristic model is less sensitive to the availability of annotated stems.

4.3. Application to SMT

For translation experiments, we used SRI's SRInterp phrase-based SMT system (cf. also [9]), which was used in NIST's MT 2006 evaluations and obtained state-of-the-art results. Word alignment was trained using GIZA++ [10] with IBM-4 model and phrases were extracted using Och's method [9]. A four-gram language model was implemented using the SRILM toolkit [11]. Decoding used a log-linear model consisting of seven scores: phrase and lexicon scores in both directions, a four-gram language model score, a distortion score and a word penalty score. The score weights were optimized using the algorithm described in [12].

Our baseline system was trained on the original Iraqi Arabic data, while the other two systems were trained on the pre-processed data using segmentations from the rule-based and heuristic models respectively. For translating from Iraqi Ara-

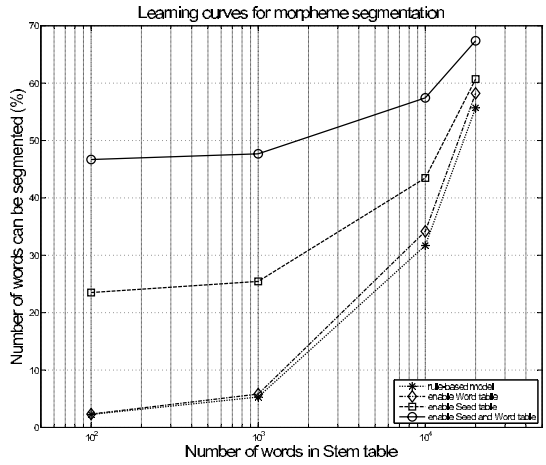

Figure 3: Learning curves for the training vocabulary with different model configurations: rule-based, rule-based with the *Word* table, rule-based with the *Seed* table, and the full heuristic model. The number of stems in the *Stem* table varied from 100 to 20,000.

bic to English, the segmented word forms were used for training word alignment and extracting phrases. For translating from English to Iraqi Arabic stemmed word forms (segmented word forms with all affixes removed) were used for training word alignment, specifically to extract phrases from the original Iraqi Arabic data.

Table 6 shows the BLEU scores for the English-Iraqi Arabic SMT systems. We can see that both the rule-based and heuristic models help to increase the translation accuracy and furthermore that our heuristic model yields the best BLEU scores in all cases except for the test set for translating from English to Iraqi Arabic.

	Iraqi-to-English		English-to-Iraqi	
	Dev	Test	Dev	Test
No segmentation	26.33	34.96	18.88	24.66
Rule-based model	27.92	36.60	19.65	**25.07**
Heuristic model	**28.16**	**37.69**	**19.95**	24.97

Table 6: BLEU scores for English-Iraqi Arabic SMT systems trained on the original and pre-processed Iraqi Arabic data. .

5. Conclusion and future work

We have presented a semi-supervised learning approach which utilizes a heuristic model for learning morpheme segmentations for Arabic dialects and which outperforms the previous rule-based segmentation approach. We have examined the learning curves using different amounts of training resources and showed that our proposed heuristic model produces more segmentations than the rule-based model even when the training resources are sparse. Although it is difficult to verify the accuracy of morpheme segmentations, we can evaluate their effectiveness with the increase of accuracy in SMT systems. Our experiments have shown that when applied to SMT, these two segmentation approaches can address the data sparseness problem for translating morphologically complex languages, and our proposed heuristic model yields better BLEU scores than the rule-based model in most cases.

One thing to mention is that the segmentation produced in our approach are not guaranteed to be correct in the linguistic sense, but linguistic correctness usually does not have much impact on phrase-based SMT when the segmentation is consistent. In future work, we will continue to investigate additional knowledge sources for producing linguistically sound segmentations. Moreover, we will explore the use of MSA resources for the morphological analysis of Arabic dialects. We expect to use data-sharing techniques to incorporate a large quantity of MSA resources into our heuristic architecture.

6. Acknowledgements

This material is based upon work supported by SRI International internal funding and in part supported by the Defense Advanced Research Projects Agency (DARPA) and the Department of Interior-National Business Center (DOI-NBC) under contract number NBCHD040058.

7. References

[1] S. Goldwater and D. McCloskey, "Improving statistical MT through morphological analysis", Proceedings of HLT/EMNLP, 2005

[2] M. Yang and K. Kirchhoff, "Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages", Proceedings of EACL, 2006

[3] F. Sadat and N. Habash, "Combination of Arabic Preprocessing Schemes for Statistical Machine Translation", Proceedings of COLING/ACL, 2006

[4] N. Habash and O. Rambow, "MAGEAD: a morphological analyzer and generator for the Arabic dialects", Proceedings of COLING/ACL, 2006

[5] J. Riesa, B. Mohit, K. Knight, D. Marcu, "Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources", Proceedings of Interspeech, 2006

[6] J. Riesa and D. Yarowsky, "Minimally Supervised Morphological Segmentation with Applications to Machine Translation", Proceedings of AMTA, 2006.

[7] K. Duh and K. Kirchhoff, "POS Tagging of Dialectal Arabic: A Minimally Supervised Approach", Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, 2005

[8] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 2.0", Linguistic Data Consortium, catalog number LDC2004L02, 2004

[9] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation", Computational Linguistics, Volume 30, Number 4, p.417-449, 2004.

[10] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51, 2003.

[11] A. Stolcke, "Srilm - An Extensible Language Modeling Toolkit", Proceedings of ICSLP, 2002

[12] F. J. Och, "Minimum error rate training in statistical machine translation", Proceedings of ACL, 2003