Joint Morphological-Lexical Language Modeling (JMLLM) for Arabic

Ruhi Sarikaya, Mohamed Afify, and Yuqing Gao

IBM T.J. Watson Research Center, Yorktown Heights NY 10598

{sarikaya,afify,yuqing}@us.ibm.com

Abstract

Language modeling for inflected languages such as Arabic poses new challenges for speech recognition due to rich morphology. The rich morphology results in large increases in perplexity and out-of-vocabulary (OOV) rate. In this study, we present a new language modeling method that takes advantage of Arabic morphology by combining morphological segments with the underlying lexical items and additional available information sources with regards to morphological segments and lexical items within a single joint model. Joint representation and modeling of morphological and lexical items reduces the OOV rate and provides smooth probability estimates. Preliminary experiments detailed in this paper show satisfactory improvements over word and morpheme based trigram language models and their interpolations.

Index Terms: Language Modeling, Maximum Entropy Modeling, Morphological Analysis, Joint Modeling.

1. Introduction

Arabic is a highly inflected language where affixes are appended to the beginning or end of a stem to generate new words that indicate case, gender, tense, number, etc. associated with the stem. Hence, it is natural that this leads to rapid vocabulary growth which is accompanied by worse language model (LM) probability estimation due to data sparsity and a higher OOV rate. For example, a parallel corpora (pairwise sentence translations) of 337K utterances between English and dialectal Arabic has about 24K and 80K unique words for the English and dialectal Arabic, respectively. A standard *n*-gram language model (LM) computes the probability of a word sequence, $W = \{w_1,....,w_K\}$

$$P(W) \approx \prod_{i=1}^{K} P(w_i|w_{i-1},...w_{i-n+1})$$

as a product of the conditional probabilities of each word given its history, which is typically approximated by *n* most recent words. There is an inverse relationship between the predictive power and the robust estimation of *n*-gram parameters. As such, as *n* increases the predictive power increases, however due to data sparsity the LM parameters may not be robustly estimated. Therefore, setting *n* to 2 or 3 appears to be a reasonable compromise between these competing goals. Robust parameter estimation problem is however more pronounced for Arabic due to its rich morphology. One would suspect that words may not be the best lexical units in this case and, perhaps, morphological units would be a better choice. Recently, there have been a number of new methods aiming at addressing robust parameter estimation and rapid vocabulary growth problems by using the morphological units to represent lexical items [1, 2, 3, 4]. Factored Language Models (FLMs) [5] share the same idea to some extent but here words are decomposed into a number of features and the resulting representation is used in a generalized back-off scheme to improve the robustness of probability estimates for rarely observed word *n*-grams.

In this work, we propose a tree structure called Morphological-Lexical Parse Tree (MLPT), to combine the information provided by a morphological analyzer with the lexical information within a single Joint Morphological-Lexical Language Model (JMLLM). The MLPT allows us to include other available information sources about the lexical items (i.e. POS), group of words (i.e. syntactic/semantic information), morphological segments[1] (i.e. prefix/stem/suffix), or the sentence (i.e. dialog state). This model enhances the language model parameter estimation and limits the rapid OOV growth. The model statistically estimates the joint probability of a sentence and its most likely morphological analysis. In this respect the model can also be used to guide the recognition for selecting high probability morphological sentence segmentations.

The rest of the paper is organized as follows. Section 2 provides a description of the morphological segmentation method. A short overview of Maximum Entropy modeling is given in Section 3. The proposed JMLLM is presented in Section 4. Section 5 describes the experimental results followed by the conclusions and future research directions in Section 7.

2. Morphological Analysis

In this study we use a rule-based morphological segmentation algorithm for Iraqi-Arabic [4], which is the language of choice for our experiments. This algorithm analyzes a given surface word, returning one of the four segmentations: {*stem, prefix+stem, suffix+stem, prefix+stem+suffix*}. Here, *stem* includes those words that do not have any affixes. Even though on the average the set of potential segmentations may be on the order of a dozen due to composite prefixes(suffixes), we use the longest prefixes(suffixes). Using finer affixes resulted in poorer speech recognition performance as compared to using the longest affixes. We attribute this to the fact that having too many small affixes reduces the *n*-gram language model span and leads to large insertion rate in the resulting decoded output. Therefore, we predefine a set of prefixes and suffixes and perform blind word segmentation. The difficulty about blind segmentation is that sometimes the beginning(ending) part of a word agrees with a prefix(suffix), and hence blind segmentation will lead to illegitimate Arabic stems. For example, the word AlqY[2] (threw in English) has its initial part agreeing with the popular prefix Al, and thus blind segmentation will lead to the

[1] We use "Morphological Segment" and Morpheme interchangeably.

[2] Using Buckwalter Arabic transliteration.

segmentation Al-qY and hence to the invalid stem qY. In order to avoid this situation we employ the following segmentation algorithm. Using the given set of prefixes and suffixes, a word is first blindly chopped to one of the four segmentations mentioned above. This segmentation is accepted if the following three rules apply:

• The resulting stem is longer than two characters in length.
• The resulting stem is accepted by the Buckwalter morphological analyzer [6].
• The resulting stem exists in the original dictionary.

The first rule eliminates many of the illegitimate segmentations. The second rule ensures that the word is a valid Arabic stem, given that the Buckwalter morphological analyzer covers all words in the Arabic language. Unfortunately, the fact that the stem is a valid Arabic stem does not always imply that the segmentation is valid. This is especially true for unvowelized text. For example, for the word AlgyA ("both canceled") the segmentation Al-gyA is not valid but the stem will be accepted by the Buckwalter morphological analyzer. The third rule, while still not offering such guarantee, simply prefers keeping the word intact if its stem does not occur in the lexicon. In our implementation we used a set of prefixes and suffixes for dialectal Iraqi. This list is given below:

• Prefix list: {chAl, bhAl, lhAl, whAl, wbAl, wAl, bAl, hAl, EAl, fAl, Al, cd, ll, b, f, c, d, w}.
• Suffix list: {thmA, tynA, hmA, thA, thm, tkm, tnA, tny ,whA, whm, wkm, wnA, wny, An, hA, hm, hn, km, kn, nA, ny, tm, wA, wh, wk, wn, yn, tk, th, h, k, t, y}.

These affixes are selected based on our knowledge of their adequacy for dialectal Iraqi Arabic. These lists differ from those for MSA [1, 2] because they have prefixes and suffixes that are particular to Iraqi Arabic. In addition, we found in preliminary experiments that keeping the top-N frequent decomposable words intact led to better performance. A value of N=5000 was experimentally found to work well in practice.

Using this segmentation method will produce prefixes and suffixes on the ASR output that should be glued to the following or previous word to form meaningful words. To facilitate such gluing we marked each prefix and suffix with a -, e.g. we have prefix Al- or suffix -yn. We used two gluing schemes. The first is very simple and just sticks any word that starts(ends) with a - to the previous(following) word. The second tries to apply some constraints to prevent sequences of affixes and to ensure that these affixes are not attached to words that start(end) with a prefix(suffix). No noticeable difference was seen between the two approaches. Next, we give a brief description of the Maximum Entropy (MaxEnt) modeling.

3. Maximum Entropy Modeling

The Maximum entropy (MaxEnt) method is an effective method to combine multiple information sources (features) in statistical modeling. The MaxEnt method is a flexible statistical modeling framework that has been used widely in many areas of natural language processing [9]. Maximum entropy modeling produces a probability model that is as uniform as possible while matching empirical feature expectations exactly. This can be interpreted as making as few assumptions as possible in the model. The MaxEnt modeling combines multiple overlapping information sources. The information sources are combined as follows:

$$P(o \mid h) = \frac{e^{\sum_i \lambda_i f_i(o,h)}}{\sum_{o'} e^{\sum_j \lambda_j f_j(o',h)}}$$

which describes the probability of a particular outcome (e.g. one of the morphemes) given the history or context. Notice that the denominator includes a sum over all possible outcomes, o', which is essentially a normalization factor for probabilities to sum to 1. The indicator functions, f_i or features are "activated" when certain outcomes are generated for certain context.

$$f_i(o \mid h) = \begin{cases} 1, & \text{if } o=o_i \text{ and } q_i(h)=1 \\ 0, & \text{otherwise} \end{cases}$$

where o_i is the outcome associated with feature f_i and $q_i(h)$ is an indicator function on histories. The MaxEnt models are trained using the improved iterative scaling algorithm [9].

For example, a bigram feature f_i representing the word sequence "ARABIC LANGUAGE" in the MaxEnt modeling would have o_i = "LANGUAGE" and $q_i(h)$ would be the question "Does the context *h* contain the word "ARABIC" as the previous word of the current word ?". Next, we present the MaxEnt based Joint Morphological-Lexical Language Modeling (JMLLM) method.

4. Joint Morphological-Lexical Language Modeling (JMLLM)

The purpose of morphological analysis is to split a word into its constituting segments. Hence, a set of segments can form a meaningful lexical unit such as a word. There may be additional information for words or group of words, such as part-of-speech (POS) tags, syntactic and semantic information (parse tree), or morpheme and word attributes. For example, in Arabic and to a certain extent in French, some words can be masculine/feminine or singular/plural. All of these information sources can be represented using a -what we call- Morphological-Lexical Parse Tree (MLPT). MLPT is a tree structured joint representation of lexical, morphological, attribute, syntactic and semantic content of the sentence. An example of a MLPT for an Arabic sentence is shown in Fig. 1. The leaves of the tree are morphological segments (morphemes) that are predicted by the language model. Each morphological segment has one of the three attributes: {*prefix, stem, suffix*} as generated by the morphological analysis mentioned in Sec. 2. Each word can take three sets of attributes: *{type, gender, number}*. Word *type* can be considered as POS, but here we consider only nouns (N), verbs (V) and rest is labeled as "other" (O). *Gender* can be masculine (M) or feminine (F). *Number* can be singular (S), plural (P) or double (D) (this is specific to Arabic). For example, NMP label for the first[1] word, شباب, shows that this word is a noun (N), male (M), plural (P). Using the information represented in MLPT for Arabic language modeling provides a back-off for smooth probability estimation even for those words that are not seen in the training data.

[1] In Arabic text is written (read) from right-to-left.

The MaxEnt models have been used in language modeling before, in the context of *n*-gram models, whole sentence models and syntactic structural language models [7] and semantic structured language models [8]. We present a new language modeling technique called Joint Morphological-Lexical Language Modeling (JMLLM) for Arabic which incorporates the local morpheme and word *n*-grams, morphological dependencies and attribute information associated with morphological segments and words. We also use the MaxEnt modeling to incorporate all the information contained in MLPT for language modeling. The dependencies or constraints represented in MLPT are integrated using the MaxEnt modeling.

We hypothesize that as we increase the amount of information represented in MLPT and the tightness of integration, JMLLM performance should improve. We can construct a single probability model that models the joint probability of all of the available information sources in the MLPT. To compute the joint probability of the morpheme sequence and its MLPT, we use features extracted from MLPT. Even though the framework is generic to jointly represent the information sources in the MLPT, in this study we limit ourselves to using only lexical and morphological content of the sentence, along with the morphological attributes simply because the lexical attributes are not available yet and we are in the process of labeling them. Therefore, the information we used from MLPT in Fig. 1 uses everything but the second row that contains lexical attributes (NFS, VFP, NFS, and NMP). Applying the morphological segmentation to data improves the coverage and reduces the OOV rate. For example, splitting the word, بالقهوة as بال (prefix) and قهوة (stem) as in Fig. 1, allows us to decode other combinations of this stem with the prefix and suffix list provided in Sec.2. These additional combinations certainly cover those words that are not seen in the unsegmented training data. The first step in building the MaxEnt model is to represent a MLPT as a sequence of morphological segments, morphological attributes, words, and word attributes using a bracket notation [8]. Converting the MLPT into a text sequence allows us to group the semantically related morphological segments and their attributes. In this notation, each morphological segment is associated (this association is denoted by "=") with an attribute (i.e. prefix/stem/suffix) and the lexical items are represented by opening and closing tokens, [WORD and WORD] respectively. The parse tree given in Fig. 1 can be converted into a token sequence in text format as follows:

[!S! [NMP شباب=stem NMP] [NFS [المنطقة ال=prefix منطقة=stem المنطقة] NFS] [VFP [يقعدون ي=prefix قعد=stem ون=suffix يقعدون] VFP] [NFS [بالقهوة بال=prefix بالقهوة=stem بالقهوة] NFS] !S!]

This representation uniquely defines the MLPT given in Fig. 1. Given the bracket notation of the text, JMLLM can be trained in two ways with varying degrees of "*tightness of integration*". A relatively "*loose integration"* involves using only the leaves of the MLPT as the model output and estimating *P(M|MLPT),* where *M* is the morpheme sequence. In this case JMLLM predicts only morphemes. A *tight integration* method would require every token in the bracket representation to be an outcome of the joint model. A tight integration can be achieved by building a joint probability

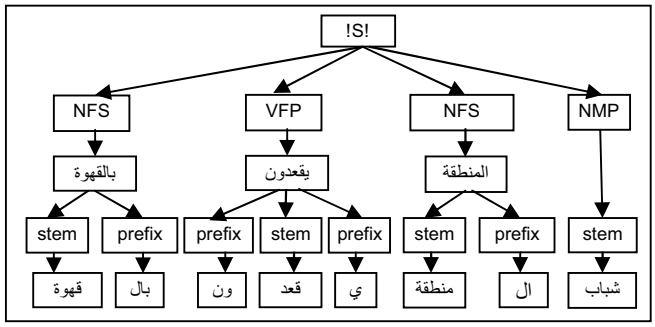

Fig 1. Morphological-Lexical Parse Tree.

model of a morphological sequence, *M*, and its MLPT, *P(M,MLPT)* which can be estimated using the tokens in the bracket notation:

$$P(M, MLPT) = \sum_{i=1}^{T} P(t_i | t_1, ... t_{i-1})$$

where t_i is a token in the bracket notation and T is the total number of tokens. The main benefit of *tight integration* using joint modeling becomes apparent when a set of alternatives are generated for a sentence rather than just a single surface form. For example, we may have more than one MLPT for a given sentence because of alternative morphological analysis, tagging or semantic/syntactic parses. Then, *tight integration* with joint modeling allows us not only to get the best morpheme sequence but also the best morphological analysis and/or tagging and/or semantic/syntactic parses of a sentence.

We note that the feature set stays the same independent of the "tightness of integration". In our preliminary experiments we chose the *loose integration* method, simply because the model training time was significantly faster than that for the *tight integration*.

The JMLLM can employ any type of questions one can derive from MLPT for predicting the next morphological segment. In addition to regular trigram questions about previous morphological segments, questions about the attributes of the previous morphological segments, parent lexical item and attributes of the parent lexical item can be used. Obviously joint questions combining these information sources are also used. These questions include (1) previous morpheme (m_{i-1}) and current active parent word, (w_i) (2) m_{i-1}, w_i and previous morpheme attribute, (ma_{i-1}), 3) ma_{i-1}, ma_{i-2}, w_i *lexical attribute* (wa_i) *and* m_{i-1}, m_{i-2}. The history given in $P(o \mid h)$ consists of answers to these questions. Clearly, there numerous questions one can ask from MLPT. The "best" feature set depends on the task, information sources and the amount of data. In our experiments, we have not exhaustively searched for the best feature set but rather used a small subset of these features which we believed to be helpful in predicting the next morpheme.

The language model score for a given morpheme using JMLLM is conditioned not only on the previous morphemes but also on their attributes, the lexical items and their morphological and lexical attributes. As such, the language model scores are smoother compared to *n*-gram models especially for unseen lexical items.

5. Experiments

We conducted our experiments on the Iraqi-Arabic speech recognition task. The Iraqi-Arabic acoustic training data consists of about 200 hours of speech collected in the context of our speech-to-speech (S2S) project [10], which covers the military and medical domains. The speech data is sampled at 16kHz and the feature vectors are computed every 10ms. The 24 dimensional MFCC features are then mean normalized, and 9 vectors are stacked leading to a 216-dimensional parameter space. The feature space is finally reduced to 40 dimensions using a combination of linear discriminant analysis (LDA), and maximum likelihood linear transformation (MLLT). There are 33 graphemes representing the speech and silence. Each grapheme is modeled with a 3-state left-to-right HMM. Building the decision tree for the Iraqi-Arabic data results in about 2K leaves and 75K Gaussians.

The language model training data has 2.8M words with 98K unique lexical items. The morphologically analyzed training data has 2.6M words with 58K unique vocabulary items. A statistical trigram language model using Modified Knesser-Ney smoothing [11] has been built for both the unsegmented data and the morphologically analyzed data. The test data consists of 2719 utterances spoken by 19 speakers. It has 3522 unsegmented lexical items, and morphological analysis reduces this figure to 3315. The OOV rate for the unsegmented test data is 3.3%, the corresponding number for the morphologically analyzed data is 2.7%. Hence, morphological segmentation reduces the OOV rate by 0.6%.

In order to evaluate the performance of the JMLLM, a lattice with low oracle error rate was generated by a Viterbi decoder using the word trigram model (Word-3gr) model. From the lattice at most 200 (N=200) sentences are extracted for each utterance to form an N-best list. These utterances are rescored using JMLLM and the trigram Morphological language model (Morph-3gr) that is built on the morphologically analyzed data. The results are presented in Table 1. The first entry (18.4%) is the oracle error rate of the N-best list. The Morph-3gr error rate is 0.9% better than the Word-3gr. Log-linear interpolation of these language models provides a small improvement (0.3%) over Morph-3gr. JMLLM obtains 30.5%, which is 1.7% and 0.8% better than Word-3gr and Morph-3gr, respectively. Interpolating JMLLM with Word-3gr improves the WER to 29.8%, which is 1.2% better than that for the interpolation of Word-3gr and Morph-3gr. The interpolation weights are set equally to 0.5 for each LM. Adding the Morph-3gr in a three way interpolation did not provide further improvement.

LANGUAGE MODELS	WER
N-best Oracle	18.4
Word-3gr	32.2
Morph-3gr	31.3
Word-3gr + Morph-3gr	31.0
JMLLM	**30.5**
JMLLM + Word-3gr	**29.8**

Table 1. Language Model Rescoring Experiments

6. Conclusions

We presented a new language modeling technique called Joint Lexical-Morphological Language Modeling (JMLLM) for Arabic. JMLLM allows joint modeling of lexical, morphological and additional information sources about morphological segments, lexical items and sentence. The results demonstrate that joint modeling provides encouraging improvements over the baseline word and morpheme based language models. Our future work will be directed towards several areas including 1) integration of the lexical attributes in JMLLM, 2) tight integration of all available information sources represented in MLPT by predicting the whole MLPT (including the internal nodes) rather than leaves of the tree, and 3) using smoothing for the MaxEnt training.

7. References

[1] A. Ghaoui, F. Yvon, C. Mokbel, and G. Chollet, "On the use of morphological constraints in N-gram statistical language model", *Eurospeech'05*, Lisbon, Portugal, 2005.

[2] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription", *ICASSP'06*, Toulouse, France, 2006.

[3] G. Choueiter, D. Povey, S.F. Chen, and G. Zweig, "Morpheme-based language modeling for Arabic LVCSR", *ICASSP'06*, Toulouse, France, 2006.

[4] M. Afify**,** R. Sarikaya, H-K J. Kuo, L. Besacier, and Y. Gao, "On the Use of Morphological Analysis for Dialectal Arabic Speech Recognition", *Interspeech'06*, Pittsburgh, PA 2006.

[5] D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition", *ICSLP '04*, Jeju Island, Korea, Oct. 2004.

[6] T. Buckwalter. "Buckwalter Arabic morphological analyzer version 1.0", *LDC2002L49 and ISBN 1-58563-257-0*, 2002.

[7] J. Wu and S. Khudanpur, "Combining nonlocal syntactic and n-gram dependencies in language modeling", *Eurospeech'99*, Budapest, Hungary, 1999.

[8] H. Erdogan, R. Sarikaya, S.F. Chen, Y. Gao and M. Picheny, "Using Semantic Analysis to Improve Speech Recognition Performance," *Computer Speech & Language Journal*, vol. 19(3), pp: 321-343, July 2005.

[9] S. F. Chen, R. Rosenfeld. "A survey of smooth-ing techniques for ME models", *IEEE Trans. Speech and Audio Process*. 8 (1), 37–50, 2000.

[10] Y. Gao, L. Gu, B. Zhou, R. Sarikaya, H.-K. Kuo. A.-V.I. Rosti, M. Afify, W. Zhu, "IBM MASTOR: Multilingual Automatic Speech-to-Speech Translator", *ICASSP'06*, Toulouse, France, 2006.

[11] S. Chen, J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling", *ACL-96*, Santa Cruz, CA, 1996.