Joint Morphological-Lexical Language Modeling for Machine Translation

Ruhi Sarikaya **Yonggang Deng**

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
sarikaya@us.ibm.com ydeng@us.ibm.com

Abstract

We present a joint morphological-lexical language model (JMLLM) for use in statistical machine translation (SMT) of language pairs where one or both of the languages are morphologically rich. The proposed JMLLM takes advantage of the rich morphology to reduce the Out-Of-Vocabulary (OOV) rate, while keeping the predictive power of the whole words. It also allows incorporation of additional available semantic, syntactic and linguistic information about the morphemes and words into the language model. Preliminary experiments with an English to Dialectal-Arabic SMT system demonstrate improved translation performance over trigram based baseline language model.

1 Introduction

Statistical machine translation (SMT) methods have evolved from using the simple word based models (Brown et al., 1993) to phrase based models (Marcu and Wong, 2002; Koehn et al., 2004; Och and Ney, 2004). More recently, there is a significant effort focusing on integrating richer knowledge, such as syntactic parse trees (Huang and Knight, 2006) within the translation process to overcome the limitations of the phrase based models. The SMT has been formulated as a noisy channel model in which the target language sentence, *e* is seen as distorted by the channel into the foreign language *f* :

$$\hat{e} = \underset{e}{\operatorname{argmax}} \; P(e \mid f) = \underset{e}{\operatorname{argmax}} \; P(f \mid e)P(e)$$

where *P(f | e)* is the translation model and *P(e)* is language model of the target language. The overwhelming proportion of the SMT research has been focusing on improving the translation model. Despite several new studies (Kirchhoff and Yang, 2004; Schwenk et al., 2006), language modeling for SMT has not been receiving much attention. Currently, the state-of-the-art SMT systems have been using the standard word n-gram models. Since n-gram models learn from given data, a severe drop in performance may be observed if the target domain is not adequately covered in the training data. The coverage problem is aggravated for morphologically rich languages. Arabic is such a language where affixes are appended to the beginning or end of a stem to generate new words that indicate case, gender, tense etc. associated with the stem. Hence, it is natural that this leads to rapid vocabulary growth, which is accompanied by worse language model probability estimation due to data sparsity and high Out-Of-Vocabulary (OOV) rate.

Due to rich morphology, one would suspect that words may not be the best lexical units for Arabic, and perhaps morphological units would be a better choice. Recently, there have been a number of new methods using the morphological units to represent lexical items (Ghaoui et al., 2005; Xiang et al., 2006; Choueiter et al., 2006). Factored Language Models (FLMs) (Kirchhoff and Yang, 2004) 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 study 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 available syntactic and semantic information about the morphological segments[1] (i.e. prefix/stem/suffix), words or group of words. The JMLLM 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 introduces the SMT system and Section 6 describes the experimental results followed by the conclusions in Section 7.

2 Morphological Segmentation

Applying the morphological segmentation to data improves the coverage and reduces the OOV rate. In

[1] We use “Morphological Segment” and “Morpheme” interchangeably.

Proceedings of NAACL HLT 2007, Companion Volume, pages 145–148,
Rochester, NY, April 2007.

this study we use a rule-based morphological segmentation algorithm for Iraqi-Arabic (Afify et. al., 2006). This algorithm analyzes a given surface word, and generates one of the four possible segmentations: {*stem, prefix+stem, suffix+stem, prefix+stem+suffix*}. Here, *stem* includes those words that do not have any affixes. We use the longest prefixes (suffixes). Using finer affixes reduces the n-gram language model span, and leads to poor performance for a fixed n-gram size. Therefore, we predefine a set of prefixes and suffixes and perform blind word segmentation. In order to minimize the illegitimate segmentations we employ the following 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:

(1) The resulting stem has more than two characters.
(2) The resulting stem is accepted by the Buckwalter morphological analyzer (Buckwalter, 2002).
(3) 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. 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 the following set of prefixes and suffixes for dialectal Iraqi:

- 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. 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 SMT output that are glued to the following or previous word to form meaningful words.

3 Maximum Entropy Modeling

The Maximum Entropy (MaxEnt) method is an effective method to combine multiple information sources (features) in statistical modeling and has been used widely in many areas of natural language processing (Berger et al.,, 2000). The MaxEnt modeling produces a probability model that is as uniform as possible while matching empirical feature expectations exactly:

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

which describes the probability of a particular outcome (e.g. one of the morphemes) given the history (*h*) 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. The MaxEnt model is trained using the Improved Iterative Scaling algorithm.

4 Joint Morphological-Lexical Language Modeling

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 (from parse tree) and semantic information, 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 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 the rest are 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[2] 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 before.

The JMLLM integrates the local morpheme and word *n*-grams, morphological dependencies and attribute information associated with morphological segments and words, which are all represented in the MLPT using the MaxEnt framework. We trained JMLLM for Iraqi-

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

Arabic speech recognition task (Sarikaya et al., 2007), and obtained significant improvements over word and morpheme based trigram language models.

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).

Using the morphological segmentation improves the coverage, 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. 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. 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*. segment. 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. 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} .

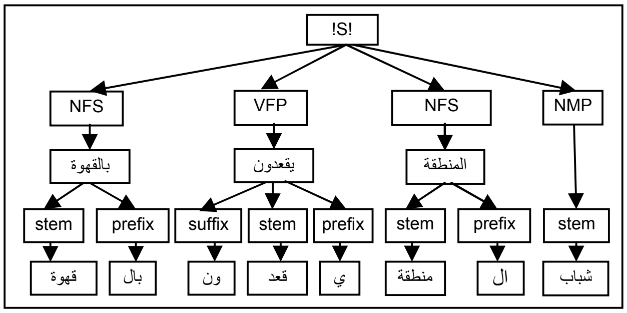

Fig 1. Morphological-Lexical Parse Tree.

The history given in $P(o \mid h)$ consists of answers to these questions. 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, and the lexical items and their attributes. As such, the language model scores are smoother compared to *n*-gram models especially for unseen lexical items.

5 Statistical Machine Translation System

Starting from a collection of parallel sentences, we trained word alignment models in two translation directions, from English to Iraqi Arabic and from Iraqi Arabic to English, and derived two sets of Viterbi alignments. By combining word alignments in two directions using heuristics (Och and Ney, 2003), a single set of static word alignments was then formed. All phrase pairs which respect to the word alignment boundary constraint were identified and pooled together to build phrase translation tables with the Maximum Likelihood criterion. The maximum number of words in Arabic phrases was set to 5.

Our decoder is the phrase-based multi-stack implementation of log-linear models similar to Pharaoh (Koehn et al, 2004). Like most other MaxEnt-based decoders, active features in our decoder include translation models in two directions, lexicon weights in two directions, language model, distortion model, and sentence length penalty.

6 Experiments

The parallel corpus has 459K utterance pairs with 90K words (50K morphemes). The Iraqi-Arabic language model training data is slightly larger than the Iraqi-Arabic side of the parallel corpus and it 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 has been built for the morphologically segmented data. The test data consists of 2242 utterances (3474 unique words). The OOV rate for the unsegmented test data is 8.7%, the corresponding number for the morphologically analyzed data is 7.4%. Hence, morphological segmentation reduces the OOV rate by 1.3% (15% relative), which is not as large reduction as compared to training data (about 40% relative reduction). We believe this would limit the potential improvement we could get from JMLLM, since JMLLM is expected to be more effective compared to word n-gram models, when the OOV rate is significantly reduced after segmentation.

We measure translation performance by the BLEU score (Papineni *et al*, 2002) with one reference for each hypothesis. In order to evaluate the performance of the JMLLM, a translation N-best list (N=10) is generated using the baseline Morpheme-trigram language model. First, on a heldout development data all feature weights including the language model weight are optimized to maximize the BLEU score using the downhill simplex method (Och and Hey, 2002). These weights are fixed when the language models are used on the test data. The translation BLEU (%) scores are given in Table 1. The first entry (37.59) is the oracle BLEU score for the N-best list. The baseline morpheme-trigram achieved 29.63, word-trigram rescoring improved the BLEU score to 29.91. The JMLLM achieved 30.20 and log-linear interpolation with the morpheme-trigram improved the BLEU score to 30.41.

Table 1. SMT N-best list rescoring.

LANGUAGE MODELS	BLEU (%)
N-best Oracle	37.59
Morpheme-trigram	29.63
Word-trigram	29.91
JMLLM	**30.20**
JMLLM + Morpheme-Trigram	**30.41**

7 Conclusions

We presented a new language modeling technique called Joint Morphological-Lexical Language Modeling (JMLLM) for use in SMT. JMLLM allows joint modeling of lexical, morphological and additional information sources about morphological segments, lexical items and sentence. The translation results demonstrate that joint modeling provides encouraging improvement over morpheme based language model. Our future work will be directed towards tight integration of all available information by predicting the entire MLPT (besides leaves).

References

P. Brown er al.,. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

A. Berger, S. Della Pietra and V. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," Computational Linguistics, vol. 22, no. 1, March 1996

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

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

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

B. Huang and K. Knight. 2006. Relabeling Syntax Trees to Improve Syntax-Based Machine Translation Quality. In HLT/NAACL.

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

K. Kirchhoff and M. Yang. 2005. Improved language modeling for statistical machine translation. In *ACL'05 workshop on Building and Using Parallel Text*, pages 125–128.

P. Koehn, F. J. Och, and D. Marcu. 2004. Pharaoh: A beam search decoder for phrase based statistical machine translation models. In *Proc. of 6th Conf. of AMTA*.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*, pages 295–302, University of Pennsylvania.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. Comp. Linguistics, 29(1):9--51.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of machine translation. In ACL 02, pages 311–318

H. Schwenk, D. D´echelotte and J-L. Gauvain. 2006. Continuous space language models for statistical machine translation. In ACL/*COLING*, pages 723–730.

M. Afify, R. Sarikaya, H-K. J. Kuo, L. Besacier and Y. Gao. 2006. On the Use of Morphological Analysis for Dialectal Arabic Speech Recognition, In Interspeech-2006, Pittsburgh PA.

R. Sarikaya, M .Afify and Y. Gao. 2007. Joint Morphological-Lexical Modeling (JMLLM) for Arabic. ICASSP 2007, Honolulu Hawaii.