Towards a Malay Derivational Lexicon: Learning Affixes Using Expectation Maximization

Suriani Sulaiman, Michael Gasser, Sandra Kübler
Indiana University
{ss23,gasser,skuebler}@indiana.edu

Abstract

We propose an unsupervised training method to guide the learning of Malay derivational morphology from a set of morphological segmentations produced by a naïve morphological analyzer. Using a morphology-based language model, we first estimate the probability of a given segmentation. We train the model with EM to find the segmentation that maximizes the probability of each morpheme. We extract the set of affix patterns produced by our algorithm and evaluate them against two references: a list of affix patterns extracted from our hand-segmented derivational wordlist and a derivational history produced by a stemmer.

1 Introduction

For languages with complex morphology, morphological analysis is a crucial step. In most languages, morphological analyzers built with comprehensive morpho-phonological rules are used to predict properties of words such as part-of-speech (POS) or morpho-syntactic features on the basis of affixes. Designing a morphological analyzer capable of producing a complete analysis requires extensive human effort and there is therefore considerable interest in machine learning of morphology.

In languages where words are not separated by spaces, such as Chinese and Japanese, statistical language modeling and unsupervised learning are the preferred methods of learning segmentation of sentences into words (Ge et al., 1999; Peng and Schuurmans, 2001; Kit et al., 2003). For morphological segmentation, unsupervised methods include the use of minimum description length (Goldsmith, 2001; Creutz and Lagus, 2005), the learning of suffixation operations and derivational rules from an inflectional lexicon (Gaussier, 1999), the application of minimum edit distance and mutual information (Baroni et al., 2002), and the mutation of virtual morphs (Kohonen et al., 2008). Most of these studies focus on well-resourced languages with mostly inflectional morphology such as English, German, and French that usually take no more than one prefix or suffix; the techniques have not been proven to work on an under-resourced language like Malay. The only effort to learn Malay morphology through a corpus based approach that we are aware of is the work of Knowles and Mohd Don (2006) who discovered Malay word classes using a stemmer. Unfortunately, their work lacks a technical discussion of the learning approach, and the origin of the stemmer remains unclear.

In this paper, we adopt a modified version of the unsupervised technique from Chinese word segmentation (Ge et al., 1999; Peng and Schuurmans, 2001; Kit et al., 2003) to learn the derivational morphology of Malay, a language with hardly any inflectional morphology, by manipulating the output of a naïve morphological analyzer. Given a Malay word, the analyzer guesses all its possible morphological segmentations, producing a list of potential hypotheses. We then use the EM algorithm to find the segmentation that maximizes the probability of each morpheme. Finally, we extract the set of all possible affix patterns from the best segmentations and evaluate them against our gold standard. Our task is not to evaluate the performance of the analyzer per se but to collect as many reliable affix patterns as possible with the help of language modeling and EM in an effort to build a Malay derivational morphological lexicon.

The remainder of the paper is organized as follows: Sec. 2 describes the basics of Malay derivational morphology. Sec. 3 presents an overview of the unsupervised learning of morphological segmentation. Sec. 4 discusses results and evaluation and Sec. 5 concludes.

Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, , pages 30–34,
Chiang Mai, Thailand, November 8, 2011.

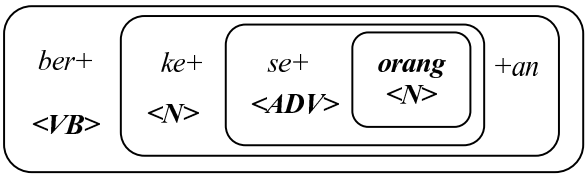

Figure 1: Nested structure of Malay morphology

English:	Malay:
use-**ful-ness**	**per-se-***faham*-**an**
help*-ness-ful**	**se-per-***juang*-**an**

Figure 2: English versus Malay morphotactics

2 Malay Derivational Morphology

Malay is an Austronesian language with rich concatenative word structure and productive derivational morphology. A Malay word can be divided into discrete morphemes with clearly defined boundaries, including roots, prefixes, suffixes, infixes, and circumfixes (Knowles and Mohd Don, 2006). In Malay morphology, affixes can be nested, as shown in Figure 1.

The loose restriction on word formation and the productive nature of certain affixes in Malay results in a large number of possible affix patterns, and the nested structures impose complex constraints on how affixes are combined. Unlike in English, some affixes in Malay can be combined in different orders, depending on the roots, to produce derived words with distinct parts-of-speech (Figure 2).

Malay derivational morphology also makes use of reduplication, which is the only non-concatenative feature in Malay for which morpheme boundaries are difficult to handle (Beesley and Karttunen, 2003). In this experiment, we exclude reduplication for the sake of simplicity.

3 Unsupervised Learning of Derivational Morphology

We first extract unique word types from our training corpora and feed them into the Malay morphological analyzer. We then build an n-gram model from the output of the analyzer. For each derived word type, the analyzer provides a list of possible morphological segmentations. However, these are unreliable because of the limitations of the analyzer (see next section). In order to get a better estimate of the probability of each morpheme, we train the n-gram model with EM on a new list of pre-segmented derived word types produced by the same analyzer using larger corpora from a different domain. Finally, the best segmentations are chosen, and unique affix patterns are extracted as initial steps in developing a derivational lexicon.

Malay word:	*diketahui* (Eng.: "*know*")
Hypothesis :	{*di-ketahu-i, di-ketahui, di-ke-tahu-i, diketahui, di-ke-tahui, diketahu-i*}

Figure 3: Sample analysis from Malay analyzer

3.1 MorfoMelayu

We use a finite-state Malay morphological analyzer, MorfoMelayu,[1] provided with an undifferentiated list of about 5000 Malay roots, a list of prefixes, and a list of suffixes. The analyzer is naïve in the sense that it knows no constraints on the order or co-occurrence of affixes. Given an input Malay word, it produces all possible segmentations of the word based on its limited knowledge of the language (Figure 3).

Although this list should include the correct segmentation, it will normally also include an average of five incorrect ones for every word analyzed. It is the task of our machine learning algorithm to learn the precise morphotactics of Malay derivational morphology.

3.2 Morphology-based Language Model

n-gram models are widely used in statistical language modeling to estimate the probability of a character or word sequence. They can be utilized to find the most probable segmentation of a word or sentence. In morphology-based language modeling, morphemes are treated as the modeling unit (Tachbelie, 2010) instead of characters or words. Since Malay morphology is mostly concatenative, it is reasonable to use morphemes as n-gram units. Given a Malay word $w = m_1 m_2 \dots m_k$, where k represents the number of morphemes, its most likely segmentation into a morpheme sequence can be determined according to maximum likelihood estimation (MLE) as:

$$s(w) = argmax \prod_{i}^{k} p_{ML}(m \mid m_{i-n+1}^{i-1}) \quad (1)$$

where m_{i-n+1}^{i-1} is the context of morpheme m_i and n the order of the n-gram model. We choose

[1] MorfoMelayu can be downloaded from `https://www.cs.indiana.edu/~gasser/Research/software.html`.

a bigram model for this experiment because it is less likely for a sequence of morphemes than for a single morpheme to coincide with a root. As an example, the Malay prefix sequence *meN-teR* is very likely to be part of a derived word, e.g., *meN-teR-**tawa*** (laugh), while the prefix *teR* alone can easily be part of the root, e.g., *terbang* (fly) or *terjun* (jump). Given a list of pre-segmented Malay derived words from the output of the Malay morphological analyzer, which we refer to as *L-model-news*, we collect the frequency counts of bigram morphemes from each word and estimate their probability:

$$p_{ML}(m_i \mid m_{i-1}) = \frac{f(m_{i-1}, m_i)}{f(m_{i-1})} \quad (2)$$

For smoothing, we apply Jelinek-Mercer linear interpolation, which has been shown to perform well on smaller training sets (Chen and Goodman, 1998) on our n-gram model. We reserve a section of the training corpus for heldout data, *L-heldout-news*, containing 1,303 pre-segmented words containing 2,347 unique bigrams. The bigrams are partitioned into 4 different buckets according to their frequencies and independently trained with the parameter value λ, tuned between 0.1 and 0.9. We linearly interpolate the bigram and unigram model:

$$p_{itp}(m_i \mid m_{i-1}) = \lambda p_{ML}(m_i \mid m_{i-1}) + (1-\lambda) p_{ML}(m_i) \quad (3)$$

where λ is set to 0.1 for low frequency bigrams (0-2 counts), 0.5 for high frequency bigrams (>10 counts) and 0.9 for bigrams of intermediate frequency (3-10 counts). Given that the output of the Malay morphological analyzer is only partially reliable to begin with, we train the bigram model with EM on a different pre-segmented wordlist *L-train-lit* produced by the same analyzer. This step ensures a more reliable $p_{ML}(m_i)$ by minimizing the bias towards the performance of the language model, forcing EM to learn to generalize from the model.

3.3 EM Training

EM is favored mainly due to its guaranteed convergence to a good probability model that locally maximizes the likelihood or posterior probability of the training data (Dempster et al., 1977). In this experiment, given a set of hypotheses for all possible segmentations of a particular word w, $s(w) = \{w'_1, w'_2, \ldots, w'_j\}$, we use EM to find the most probable segmentation that maximizes $s(w)$. Instead of initializing with uniform distribution across the training data, we use the initial probability estimation from the bigram model to boost the slow convergence of EM and perform 10 iterations to produce a more reliable $f(m)$ for estimating $p(m)$ using (4):

$$f^{t+1} = \sum_{w \in L-tr} \sum_{w' \in S(w)} \frac{p^t(w')}{\alpha} f^t(m \in w') \quad (4)$$

where m now represents a sequence of two morphemes, t the current iteration and $f^t(m \in w')$ the number of times a morpheme sequence m occurs in segmentation w'. Since maximum likelihood training is known to penalize longer sequences, we add the normalization factor α in (4), which is the sum of the probabilities of all possible segmentations for a particular word w. We assume a uniform distribution for each unique morpheme in the training list *L-train-lit* and assign$f^0(m)$ a frequency of 1. We adjust (2) as (5) for simplicity, where $f(m)$ is the sum of frequency of all bigrams in *L-model-news*. We derive $p^0(m)$ and its subsequent values from (5).

$$p(m_i) = \frac{f(m_i)}{\sum_{w \in L-model} f(m)} \quad (5)$$

We update the count of each morpheme through (4) for an optimum value of $p(m_i)$. The updated value of $p(m_i)$ is then used to re-calculate $s(w)$ through (1) at the end of each iteration. Note that this differs slightly from the normal implementation of EM in which $s(w)$ is re-estimated at each step. We find that this method speeds up the convergence process and improves the overall performance of EM for our tasks.

3.4 Derivational Lexicon of Affix Patterns

Based on the best segmentations produced by our EM algorithm, we extract all unique affix patterns by combining over possible roots. We then construct a lexicon consisting of unique affix patterns (e.g., *meN-X-kan*, *ber-ke-X-an*, where *X* represents a possible root) for Malay derivational morphology. We evaluate the validity of the affix patterns produced by our algorithm by comparing them with a list of affix patterns extracted from a hand-segmented list of derived words produced by a native speaker of Malay and an automatically derived list produced by a stemmer (Knowles and Mohd Don, 2006).

	Hand Segmented		Stemmer
	L_H-eval-news	L_H-eval-lit	L_S-eval-lit
Precision	33.17	27.14	40.7
Recall	61.11	58.06	36.16
F-Score	42.99	36.99	38.29
Lex. size	108	93	224
Pat. not recov.	42	39	143

Table 1: Experimental results

Error type	Analyzer Output	Hand-segment	Pattern error
Root-Pref.	meN-teR-nak	meN-ternak	meN-teR-X
Root-Suf.	beR-nila-i	beR-nilai	beR-X-i
Suffix Recursion	peN-tah-an-an	peN-tahan-an	X-an-an
All affix	peN-di-di-kan	peN-didik-an	peN-di-di-kan
OOV	beR-se-belah-an	-	ber-se-X-an

Table 2: Typical errors of affix patterns

3.5 Datasets

Four different corpora are used for training and evaluation. The first training corpus, used to build the morphology-based bigram model, consists of 14,869 word types compiled from Malay news articles. The pre-segmented list, *L-model-news*, contains 8,563 derived words (13,514 unique bigrams). The second corpus, used for EM training, consists of 18,438 word types collected from Malay literature. After post-processing, the pre-segmented list, *L-train-lit*, contains 15,916 derived words producing 215 unique affix patterns. For evaluation, two separate corpora are collected from Malay news articles and literature. The news articles contain 5,797 word types with 2,584 derived words (L_H*-eval-news*), producing 108 unique affix patterns, while the literature has 2,832 word types with 1,439 derived words (L_H*-eval-lit*), producing 93 unique affix patterns. Finally, we use a reference list of derivational history (L_S*-eval-lit*) collected by Knowles and Mohd Don (2006) from 4 Malay texts (119,471 words) and generated by a stemmer (224 affix patterns).

4 Results and Evaluation

To evaluate the lexicon we extracted from the training data, we compared the affix patterns extracted from the evaluation corpora, by hand or using the stemmer, with the patterns in the lexicon. The results are shown in Table 1.

There are a few observations to be made from these results. Firstly, our implementation of EM is still biased towards shorter morpheme sequences despite the added normalizing factor α, failing to choose correct segmentations with longer sequences. Secondly, a large amount of data is crucial to extract as many unique affix patterns as possible (an average of 4 unique affix patterns per 100 derived words). The limited amount of hand-segmented data used as the gold standard and the tendency of our algorithm to choose words with fewer morphemes represent major weaknesses in our evaluation, resulting in very low precision values (33.17% and 27.14%). Thirdly, the use of different domains for evaluation does not seem to affect the results, suggesting that domain is not a critical factor in collecting diversified affix patterns. Finally, we find that most affix patterns not recovered from the training corpus are either out of the vocabulary or result from ambiguous affixes that also exist as parts of roots (affix-like syllables). These ambiguous affixes occur so often that our algorithm fails to tell them apart. Table 2 shows typical errors produced by the analyzer.

5 Conclusion and Future Work

We have explored the feasibility of using a naïve morphological analyzer, a morphology-based language model, and EM training for learning the derivational morphology of an under-resourced language like Malay. As far as we know, this is the first attempt to combine these three methods in the learning of morphology. Our low precision and F-score indicate that our algorithm suffers from over-segmentation, which we believe is due to the small reference sets used for evaluation. Despite the discouraging overall results, our promising recall values (61.11% and 58.06%) show that most of the frequent affix patterns from our gold standard are recognized from the analysis. Eventually, the error analysis can serve as a guideline to improve the performance of the Malay morphological analyzer. In future, we will compare the performance of our algorithm with Morfessor 1.0 for unsupervised morphology learning (Creutz and Lagus, 2005). Our ultimate goal is to construct a hierarchical lexicon for Malay derivational morphology by clustering affixes based on their positions, precedence and lexical classes with the help of the improved analyzer.

Acknowledgments

The first author is funded by the Ministry of Higher Education of Malaysia.

References

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and seman-tic similarity. In *Proceedings of the ACL Workshop on Morphological and Phonological Learning*, pages 48–57, Philadelphia, PA.

Kenneth Beesley and Lauri Karttunen. 2003. *Finite-State Morphology*. CSLI Publications.

Stanley Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki, Finland.

Arthur Dempster, Nan Laird, and Donald Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B(34):1–38.

Eric Gaussier. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 24–30, College Park, MD.

Xianping Ge, Wanda Prat, and Padhraic Smyth. 1999. Discovering Chinese words from unsegmented text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–272, Berkeley, CA.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Chunyu Kit, Zhiming Xu, and Jonathan Webster. 2003. Integrating $n$gram model and case-based learning for Chinese word segmentation. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 160–163, Sapporo, Japan.

Gerald Knowles and Zuraidah Mohd Don. 2006. *Word Class in Malay: A Corpus Based Approach*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia.

Oskar Kohonen, Sami Virpioja, and Mikaela Klami. 2008. Allomorfessor: Towards unsupervised morpheme analysis. In *Proceedings of the 9th Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*, pages 975–982, Aarhus, Denmark.

Fuchun Peng and Dale Schuurmans. 2001. Self-supervised Chinese word segmentation. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, pages 238–247, Cascais, Portugal.

Martha Y. Tachbelie. 2010. *Morphology-Based Language Modeling for Amharic*. Ph.D. thesis, University of Hamburg, Hamburg, Germany.