Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model

Hajime Morita[1,2] **Daisuke Kawahara**[1] **Sadao Kurohashi**[1,2]
[1] Kyoto University [2] CREST, Japan Science and Technology Agency
{hmorita, dk, kuro}@i.kyoto-u.ac.jp

Abstract

We present a new morphological analysis model that considers semantic plausibility of word sequences by using a recurrent neural network language model (RNNLM). In unsegmented languages, since language models are learned from automatically segmented texts and inevitably contain errors, it is not apparent that conventional language models contribute to morphological analysis. To solve this problem, we do not use language models based on raw word sequences but use a semantically generalized language model, RNNLM, in morphological analysis. In our experiments on two Japanese corpora, our proposed model significantly outperformed baseline models. This result indicates the effectiveness of RNNLM in morphological analysis.

1 Introduction

In contrast to space-delimited languages like English, word segmentation is the first and most crucial step for natural language processing (NLP) in unsegmented languages like Japanese, Chinese, and Thai (Kudo et al., 2004; Kaji and Kitsuregawa, 2014; Shen et al., 2014; Kruengkrai et al., 2006). Word segmentation is usually performed jointly with related analysis: POS tagging for Chinese, and POS tagging and lemmatization (analysis of inflected words) for Japanese. Morphological analysis including word segmentation has been widely and actively studied, and for example, Japanese word segmentation accuracy is in the high 90s. However, we often observe that strange outputs of downstream NLP applications such as machine translation and question answering come from incorrect word segmentations.

For example, the state-of-the-art and popular Japanese morphological analyzers, JUMAN (Kurohashi and Kawahara, 2009) and MeCab (Kudo et al., 2004) both analyze "外国人参政権 (foreigner's right to vote)" not into the correct segmentation of (1a), but into the incorrect and awkward segmentation of (1b).

(1) a. 外国 / 人 / 参政 / 権
foreigner right to vote
b. 外国 / 人参 / 政権
foreign carrot regime

JUMAN is a rule-based morphological analyzer, defining word-to-word (including inflection) connectivities and their scores. MeCab is a supervised morphological analyzer, learning the probabilities of word/POS/inflection sequence from an annotated corpus of tens of thousands of sentences. Both systems, however, cannot realize semantically appropriate analysis, and often produce totally strange outputs like the above.

This paper proposes a semantically appropriate morphological analysis method for unsegmented languages using a language model. For unsegmented languages, morphological analysis and language modeling form a chicken-and-egg problem. That is, if high-quality morphological analysis is available, we can learn a high-quality language model from a morphologically analyzed large corpus. On the other hand, if a high-quality language model is available, we can achieve high-quality morphological analysis by looking for a segmented word sequence with a large language model score. However, even if we learn a language model from a corpus analyzed by a certain level of morphological analyzer, the language model is affected by the analysis errors of the morphological analyzer and it is no practical use for the improvement of the morphological analyzer. A language model trained by incorrectly segmented "外国 (foreign)/人参 (carrot)/政権 (regime)" just supports that incorrect segmentation.

The point of the paper is that we have tackled the chicken-and-egg problem, not by using a lan-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2292–2297,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

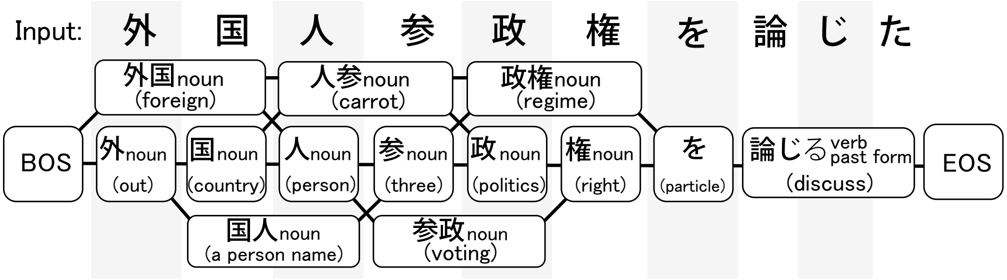

Figure 1: An example of a word lattice.

guage model of raw word sequences, but by using a semantically generalized language model based on word embeddings, RNNLM (Recurrent Neural Network Language Model) (Mikolov et al., 2010; Mikolov et al., 2011). The RNNLM is trained on an automatically analyzed corpus of ten million sentences, which possibly includes incorrect segmentations such as "外国 (foreign)/人参 (carrot)/政権 (regime)." However, on semantically generalized level, it is an unnatural semantic sequence like *nation vegetable politics*. Since the state-of-the-art morphological analyzer achieves the high accuracy, it does not often produce incorrect analyses which support such a semantically strange sequence. This would prefer analysis toward semantically appropriate word sequences. When a morphological analyzer utilizes such a generalized and reasonable language model, it can penalize strange segmentations like "外国 (foreign)/人参 (carrot)/政権 (regime)," leading to better accuracy.

We furthermore retrain RNNLM using an annotated corpus of manually segmented 45k sentences, which further improves morphological analysis.

2 Related Work

There have been several studies that have integrated language models into morphological analysis. Wang et al. (2011) improved Chinese word segmentation and POS tagging by using N-gram features learned from an automatically segmented corpus. However, since the auto-segmented corpus inevitably contains segmentation errors, frequent N-grams are not always correct and thus this problem might affect the performance of morphological analysis. They also divided N-gram frequencies into three binned features: high-frequency, middle-frequency and low-frequency. Such coarse features cannot express slight differences in the likelihood of language models.

Kaji and Kitsuregawa (2014) used a bigram language model feature for Japanese word segmentation and POS tagging. Their objective of using a language model is to normalize informally spelled words in microblogs. Therefore, their objective is different from ours.

Some studies have used character-based language models for Chinese word segmentation and POS tagging (Zheng et al., 2013; Liu et al., 2014). Although their approaches have no drawbacks of learning incorrect segmentations, they only capture more local information than word-based language models.

Word embeddings have been also used for morphological analysis. Neural network based models have been proposed for Chinese word segmentation and POS tagging (Pei et al., 2014) or word segmentation (Mansur et al., 2013). These methods acquire word embeddings from a corpus, and then use them as the input of the neural networks. Our proposed model learns word embeddings via RNNLM, and these embeddings are used for scoring word transitions in morphological analysis. Our usage of word embeddings is different from the previous studies.

3 Proposed Method

We propose a new morphological analysis model that considers semantic plausibility of word sequences by using RNNLM. We integrate RNNLM into morphological analysis (Figure 2). We train the RNNLM using both an automatically analyzed corpus and a manually labeled corpus.

3.1 Recurrent Neural Network Language Model

RNNLM is a recurrent neural network language model (Mikolov et al., 2010), which outputs a probability distribution of the next word, given the embedding of the last word and its context. We

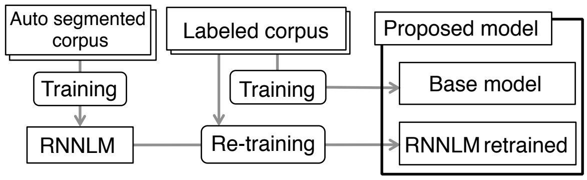

Figure 2: Workflow for training RNNLM and base model.

employ the RNNME language model[1] proposed by (Mikolov et al., 2011; Mikolov, 2012) as the implementation of RNNLM. The RNNME language model has direct connections from the input layer of the recurrent neural network to the output layer, which act as a maximum entropy model and avoid to waste a lot of parameters to describe simple patterns. Hereafter, we refer to the RNNME language model simply as RNNLM.

To train RNNLM, we use a raw corpus of 10 million sentences from the web corpus (Kawahara and Kurohashi, 2006). These sentences are automatically segmented by JUMAN (Kurohashi and Kawahara, 2009). The training of RNNLM is based on lemmatized word sequences without POS tags.

The trained model contains errors caused by an automatically analyzed corpus. We retrain RNNLM using a manually labeled corpus after training RNNLM using the automatically analyzed corpus as shown in Figure 2. The retraining aims to cope with errors related to function word sequences.

3.2 Base Model

For our base model, we adopt a model for supervised morphological analysis, which performs segmentation, lemmatization and POS tagging jointly. We train this model using a tagged corpus of tens of thousands of sentences that contain gold segmentations, lemmas, inflection forms and POS tags. To predict the most probable sequence of words with lemmas and POS tags given an input sentence, we execute the following procedure:

1. Look up the string of the input sentence using a dictionary.
2. Make a word lattice.
3. Search for the path with the highest score from the lattice.

Figure 1 illustrates the constructed lattice during the procedure. At the dictionary lookup step, we use the basic dictionary of JUMAN and an additional dictionary comprising 0.8 million words, both of which have lemma, POS and inflection information. The additional dictionary mainly consists of itemizations in articles and article titles in Japanese Wikipedia.

We define the scoring function as follows:

$$\text{score}_B(\mathbf{y}) = \Phi(\mathbf{y}) \cdot \vec{w}, \tag{1}$$

where $\mathbf{y}$ is a tagged word sequence, $\Phi(\mathbf{y})$ is a feature vector for $\mathbf{y}$, and $\vec{w}$ is a weight vector. Each element in $\vec{w}$ gives a weight to its corresponding feature in $\Phi(\mathbf{y})$. We use the unigram and the bigram features composed from word base form, POS and inflection described in Kudo et al. (2004). We also use additional lexical features such as character type, and trigram features used in Zhang and Clark (2008). To learn the weight vector, we adopt exact soft confidence-weighted learning (Wang et al., 2012).

To consider out-of-vocabulary (OOV) words that are not found in the dictionary, we automatically generate words at the lookup step by segmenting the input string by character types[2]. For training, we regard words that are not found in the dictionary but found in the training corpus as OOV words to learn their weights.

3.3 RNNLM Integrated Model

Based on retrained RNNLM, we calculate an RNNLM score ($\text{score}_R(\mathbf{y})$) to be integrated into the base model. The RNNLM score is defined as the log probability of the next word given its context (path). Here, the score for an OOV word is given by the following formula:

$$-C_p - L_p \cdot \text{length}(n), \tag{2}$$

where C_p is a constant penalty for OOV words, L_p is a factor for the character length penalty, and $\mathrm{length}(n)$ returns the character length of the next word n. This formula is defined to penalize longer words, which are likely to produce segmentation errors.

We then integrate the RNNLM score into the base model using the following equation:

$$\text{score}_I(\mathbf{y}) = (1-\alpha)\text{score}_B(\mathbf{y}) + \alpha\ \text{score}_R(\mathbf{y}), \tag{3}$$

[1] RNNME is the abbreviation of Recurrent Neural Network trained jointly with Maximum Entropy model.

[2] Japanese has three types of characters: Kanji, Hiragana and Katakana.

where α is an interpolation parameter that is tuned on development data.

For decoding, we employ beam search as used in Zhang and Clark (2008). Since the possible context (paths in the word lattice) considered in RNNLM falls into combinatorial explosion in morphological analysis, we keep only probable context candidates inside the beam. That is, each node keeps candidates inside the beam width. Each candidate has a vector representing context, and two words of history. The recurrent model makes decoding harder than non-recurrent neural network language models. However, we use RNNLM because the model outperforms other NNLMs (Mikolov, 2012) and the result suggests that the model is more likely to capture semantic plausibility. Since a sentence rarely contains ambiguous and semantically appropriate word sequences, we think that beam search with enough beam size is able to keep the ambiguous candidates of word sequences. In the case of non-recurrent NNLMs and the base model, which uses trigram features, we can conduct exact decoding using the second-order Viterbi algorithm (Thede and Harper, 1999).

4 Experiments

4.1 Experimental Settings

In our experiments, we used the Kyoto University Text Corpus (Kawahara et al., 2002) and Kyoto University Web Document Leads Corpus (Hangyo et al., 2012) as manually tagged corpora. We randomly chose 2,000 sentences from each corpus for test data, and 500 sentences for development data. We used the remaining part of the corpora as training data to train our base model and retrain RNNLM. In total, we used 45,000 sentences for training.

For comparative purposes, we used the following four baselines: the Japanese morphological analyzer JUMAN, the supervised morphological analyzer MeCab, the base model, and a model using a conventional language model. For this language model, we built a trigram language model with Kneser-Ney smoothing using SRILM (Stolcke, 2002) from the same automatically segmented corpus. The language model is modified to have an interpolation parameter α and length penalty for OOV, L_p.

We set the beam width to 5 by preliminary experiments. We also set a constant penalty for OOV words (C_p) as 5, which is the default value in the implementation of Mikolov et al. (2011). We tuned the parameters of our proposed model and the baseline model (α and L_p) and the parameters of language models using grid search on the development data. We set α = 0.3, L_p =1.5 for the proposed model (" Base + RNNLM$_{\mathrm{retrain}}$").[3]

We measured the performance of the baseline models and the proposed model by F-value of word segmentation and F-value of joint evaluation of word segmentation and POS tagging. We calculated F-value for the two corpora (news and web) and the merged corpus (all).

We used the bootstrapping method (Zhang et al., 2004) to test statistical significance between proposed models and other models. Suppose we have a test set T that includes N sentences. The method repeatedly creates M new test sets by resampling N sentences with replacement from T. We calculate the F-value of each model on $M + 1$ test sets including T, and then we have $M + 1$ score differences. From the scores, we calculate the 95% confidence interval. If the interval does not overlap with zero, the two models are considered as statistically significantly different. In our evaluation, M is set to 2,000.

4.2 Results and Discussions

Table 1 lists the results of our proposed model and the baseline models. Our proposed model ("Base + RNNLM$_{\mathrm{retrain}}$") significantly outperforms all the baseline models and "Base + RNNLM," which does not use retraining. In particular, we achieved a large improvement for segmentation. This can be attributed to the use of RNNLM that was learned based on lemmatized word sequence without POS tags.

"Base + SRILM" segmented the example described in Section 1 ("外国人参政権") into the incorrect segmentation "外国/人参/政権" in the same way as JUMAN. This segmentation error was caused by errors in the automatically segmented corpus that was used to train the language model. Our proposed model can correctly segment this example if a proper context is available by semantically capturing word transitions using RNNLM.

The base model, JUMAN and "Base + SRILM" incorrectly segmented "健康 (healthy)/など (etc.)/

[3] We set α = 0.1, L_p = 2.0 for "Base + RNNLM", and α = 0.3, L_p = 0.5 for "Base + SRILM."

	Segmentation (news)	Seg + POS (news)	Segmentation (web)	Seg + POS (web)	Segmentation (all)	Seg + POS (all)
JUMAN	98.92	98.47	98.20	97.64	98.64	98.14
MeCab	99.07	98.58	98.22	97.51	98.74	98.16
Base model	98.94	98.46	97.71	96.90	98.46	97.85
Base + SRILM	98.94	98.40	98.13	97.33	98.62	97.98
Base + RNNLM	99.06	98.59	98.17	97.45	98.71	98.14
Base + $\text{RNNLM}_{\text{retrain}}$	$\mathbf{99.15}^*$	$\mathbf{98.70}^*$	$\mathbf{98.37}^*$	$\mathbf{97.68}^*$	$\mathbf{98.84}^*$	$\mathbf{98.30}^*$

Table 1: Results for test datasets. $*$ means the score of "Base + $\text{RNNLM}_{\text{retrain}}$" is significantly improved from that of all other models.

の (of)/点 (point)/で (in)/······" (in terms of health and so on) into "健康 な(healthy)/どの(any)/点 (point)/で (in)/······." Although this segmentation can be grammatically accepted, it is difficult to semantically interpret this word sequence. Our proposed model can correctly segment this example because RNNLM learns semantically plausible word sequences.

5 Conclusion

In this paper, we proposed a new model for morphological analysis that is integrated with RNNLM. We trained RNNLM on an automatically segmented corpus and tuned on a manually tagged corpus. The proposed model was able to significantly reduce errors in the base model by capturing semantic plausibility of word sequences using RNNLM. In the future, we will design features derived from RNNLM models, and integrate them into a unified learning framework. We also intend to apply our method to unsegmented languages other than Japanese, such as Chinese and Thai.

References

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 535–544.

Nobuhiro Kaji and Masaru Kitsuregawa. 2014. Accurate word segmentation and POS tagging for japanese microblogs: Corpus annotation and joint modeling with lexical normalization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 99–109, Doha, Qatar. Association for Computational Linguistics.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1344–1347.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L02-1302.

Canasai Kruengkrai, Virach Sornlertlamvanich, and Hitoshi Isahara. 2006. A conditional random field framework for Thai morphological analysis. In *Proceedings of LREC*, pages 2419–2424.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the Conference on Emprical Methods in Natural Language Processing (EMNLP 2004)*, volume 2004.

Sadao Kurohashi and Daisuke Kawahara, 2009. *Japanese Morphological Analysis System JUMAN 6.0 Users Manual*. `http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN`.

Xiaodong Liu, Kevin Duh, Yuji Matsumoto, and Tomoya Iwakura. 2014. Learning character representations for Chinese word segmentation. In *NIPS 2014 Workshop on Modern Machine Learning and Natural Language Processing*.

Mairgup Mansur, Wenzhe Pei, and Baobao Chang. 2013. Feature-based neural language model and Chinese word segmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1271–1277, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Tomas Mikolov, Anoop Deoras, Dan Povey, Lukar Burget, and Jan Honza Cernocky. 2011. Strategies for training large scale neural network language models. In *Proceedings of ASRU 2011*, pages 196–201. IEEE Automatic Speech Recognition and Understanding Workshop.

Tomas Mikolov. 2012. *Statistical language models based on neural networks*. Ph.D. thesis, Brno university of technology.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for Chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 293–303.

Mo Shen, Hongxiao Liu, Daisuke Kawahara, and Sadao Kurohashi. 2014. Chinese Morphological Analysis with Character-level POS Tagging. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 253–258, Baltimore, Maryland. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In John H L Hansen and Bryan L Pellom, editors, *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, pages 901–904. ISCA.

Scott M. Thede and Mary P. Harper. 1999. A second-order Hidden Markov Model for part-of-speech tagging. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 175–182, Morristown, NJ, USA, June. Association for Computational Linguistics.

Yiou Wang, Jun'ichi Kazama, Wenliang Chen, Yujie Zhang, Kentaro Torisawa, and Yoshimasa Tsuruoka. 2011. Improving chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP-2011)*, pages 309–317, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Jialei Wang, Peilin Zhao, and Steven C.H. Hoi. 2012. Exact soft confidence-weighted learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 121–128.

Yue Zhang and Stephen Clark. 2008. Joint word segmentation and pos tagging using a single perceptron. In *Proceedings of ACL-08: HLT*, pages 888–896, Columbus, Ohio, June. Association for Computational Linguistics.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L04-1489.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep learning for Chinese word segmentation and POS tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 647–657.