Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis

Graham Neubig, Yosuke Nakata, Shinsuke Mori
Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, Japan

Abstract

We present a pointwise approach to Japanese morphological analysis (MA) that ignores structure information during learning and tagging. Despite the lack of structure, it is able to outperform the current state-of-the-art structured approach for Japanese MA, and achieves accuracy similar to that of structured predictors using the same feature set. We also find that the method is both robust to out-of-domain data, and can be easily adapted through the use of a combination of partial annotation and active learning.

1 Introduction

Japanese morphological analysis (MA) takes an unsegmented string of Japanese text as input, and outputs a string of morphemes annotated with parts of speech (POSs). As MA is the first step in Japanese NLP, its accuracy directly affects the accuracy of NLP systems as a whole. In addition, with the proliferation of text in various domains, there is increasing need for methods that are both robust and adaptable to out-of-domain data (Escudero et al., 2000).

Previous approaches have used structured predictors such as hidden Markov models (HMMs) or conditional random fields (CRFs), which consider the interactions between neighboring words and parts of speech (Nagata, 1994; Asahara and Matsumoto, 2000; Kudo et al., 2004). However, while structure does provide valuable information, Liang et al. (2008) have shown that gains provided by structured prediction can be largely recovered by using a richer feature set. This approach has also been called "pointwise" prediction, as it makes a single independent decision at each point (Neubig and Mori, 2010).

While Liang et al. (2008) focus on the speed benefits of pointwise prediction, we demonstrate that it also allows for more robust and adaptable MA. We find experimental evidence that pointwise MA can exceed the accuracy of a state-of-the-art structured approach (Kudo et al., 2004) on in-domain data, and is significantly more robust to out-of-domain data.

We also show that pointwise MA can be adapted to new domains with minimal effort through the combination of active learning and partial annotation (Tsuboi et al., 2008), where only informative parts of a particular sentence are annotated. In a realistic domain adaptation scenario, we find that a combination of pointwise prediction, partial annotation, and active learning allows for easy adaptation.

2 Japanese Morphological Analysis

Japanese MA takes an unsegmented string of characters $\boldsymbol{x}_1^I$ as input, segments it into morphemes $\boldsymbol{w}_1^J$, and annotates each morpheme with a part of speech $\boldsymbol{t}_1^J$. This can be formulated as a two-step process of first segmenting words, then estimating POSs (Ng and Low, 2004), or as a single joint process of finding a morpheme/POS string from unsegmented text (Kudo et al., 2004; Nakagawa, 2004; Kruengkrai et al., 2009). In this section we describe an existing joint sequence-based method for Japanese MA, as well as our proposed two-step pointwise method.

2.1 Joint Sequence-Based MA

Japanese MA has traditionally used sequence based models, finding a maximal POS sequence for en-

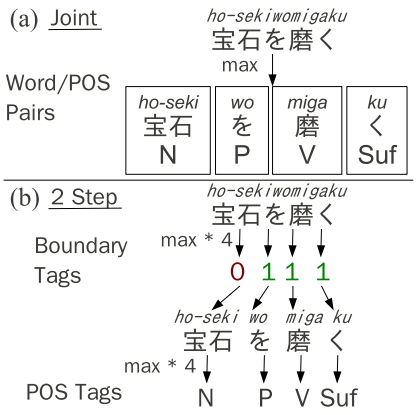

Figure 1: Joint MA (a) performs maximization over the entire sequence, while two-step MA (b) maximizes the 4 boundary and 4 POS tags independently.

Type	Feature Strings
Unigram	t_j, $t_j w_j$, $c(w_j)$, $t_j c(w_j)$
Bigram	$t_{j-1}t_j$, $t_{j-1}t_j w_{j-1}$, $t_{j-1}t_j w_j$, $t_{j-1}t_j w_{j-1}w_j$

Table 1: Features for the joint model using tags t and words w. $c(\cdot)$ is a mapping function onto character types (*kanji*, *katakana*, etc.).

tire sentences as in Figure 1 (a). The CRF-based method presented by Kudo et al. (2004) is generally accepted as the state-of-the-art in this paradigm. CRFs are trained over segmentation lattices, which allows for the handling of variable length sequences that occur due to multiple segmentations. The model is able to take into account arbitrary features, as well as the context between neighboring tags.

We follow Kudo et al. (2004) in defining our feature set, as summarized in Table 1[1]. Lexical features were trained for the top 5000 most frequent words in the corpus. It should be noted that these are word-based features, and information about transitions between POS tags is included. When creating training data, the use of word-based features indicates that word boundaries must be annotated, while the use of POS transition information further indicates that all of these words must be annotated with POSs.

[1]More fine-grained POS tags have provided small boosts in accuracy in previous research (Kudo et al., 2004), but these increase the annotation burden, which is contrary to our goal.

Type	Feature Strings
Character n-gram	x_l, x_r, $x_{l-1}x_l$, $x_l x_r$, $x_r x_{r+1}$, $x_{l-1}x_l x_r$, $x_l x_r x_{r+1}$
Char. Type n-gram	$c(x_l)$, $c(x_r)$, $c(x_{l-1}x_l)$, $c(x_l x_r)$, $c(x_r x_{r+1})$, $c(x_{l-2}x_{l-1}x_l)$, $c(x_{l-1}x_l x_r)$, $c(x_l x_r x_{r+1})$, $c(x_r x_{r+1}x_{r+2})$
WS Only	l_s, r_s, i_s
POS Only	w_j, $c(w_j)$, d_{jk}

Table 2: Features for the two-step model. x_l and x_r indicate the characters to the left and right of the word boundary or word w_j in question. l_s, r_s, and i_s represent the left, right, and inside dictionary features, while d_{jk} indicates that tag k exists in the dictionary for word j.

2.2 2-Step Pointwise MA

In our research, we take a two-step approach, first segmenting character sequence $\boldsymbol{x}_1^I$ into the word sequence $\boldsymbol{w}_1^J$ with the highest probability, then tagging each word with parts of speech $\boldsymbol{t}_1^J$. This approach is shown in Figure 1 (b).

We follow Sassano (2002) in formulating word segmentation as a binary classification problem, estimating boundary tags $\boldsymbol{b}_1^{I-1}$. Tag $b_i = 1$ indicates that a word boundary exists between characters x_i and x_{i+1}, while $b_i = 0$ indicates that a word boundary does not exist. POS estimation can also be formulated as a multi-class classification problem, where we choose one tag t_j for each word w_j. These two classification problems can be solved by tools in the standard machine learning toolbox such as logistic regression (LR), support vector machines (SVMs), or conditional random fields (CRFs).

We use information about the surrounding characters (character and character-type n-grams), as well as the presence or absence of words in the dictionary as features (Table 2). Specifically dictionary features for word segmentation l_s and r_s are active if a string of length s included in the dictionary is present directly to the left or right of the present word boundary, and i_s is active if the present word boundary is included in a dictionary word of length s. Dictionary feature d_{jk} for POS estimation indicates whether the current word w_j occurs as a dictionary entry with tag t_k.

Previous work using this two-stage approach has

used sequence-based prediction methods, such as maximum entropy Markov models (MEMMs) or CRFs (Ng and Low, 2004; Peng et al., 2004). However, as Liang et al. (2008) note, and we confirm, sequence-based predictors are often not necessary when an appropriately rich feature set is used. One important difference between our formulation and that of Liang et al. (2008) and all other previous methods is that we rely only on features that are directly calculable from the surface string, without using estimated information such as word boundaries or neighboring POS tags[2]. This allows for training from sentences that are partially annotated as described in the following section.

3 Domain Adaptation for Morphological Analysis

NLP is now being used in domains such as medical text and legal documents, and it is necessary that MA be easily adaptable to these areas. In a domain adaptation situation, we have at our disposal both annotated general domain data, and unannotated target domain data. We would like to annotate the target domain data efficiently to achieve a maximal gain in accuracy for a minimal amount of work.

Active learning has been used as a way to pick data that is useful to annotate in this scenario for several applications (Chan and Ng, 2007; Rai et al., 2010) so we adopt an active-learning-based approach here. When adapting sequence-based prediction methods, most active learning approaches have focused on picking full sentences that are valuable to annotate (Ringger et al., 2007; Settles and Craven, 2008). However, even within sentences, there are generally a few points of interest surrounded by large segments that are well covered by already annotated data.

Partial annotation provides a solution to this problem (Tsuboi et al., 2008; Sassano and Kurohashi, 2010). In partial annotation, data that will not contribute to the improvement of the classifier is left untagged. For example, if there is a single difficult word in a long sentence, only the word boundaries and POS of the difficult word will be tagged. "Difficult" words can be selected using active learning approaches, choosing words with the lowest classifier accuracy to annotate. In addition, corpora that are tagged with word boundaries but not POS tags are often available; this is another type of partial annotation.

[2]Dictionary features are active if the string exists, regardless of whether it is treated as a single word in $\boldsymbol{w}_1^J$, and thus can be calculated without the word segmentation result.

Type	Train	Test
General	782k	87.5k
Target	153k	17.3k

Table 3: General and target domain corpus sizes in words.

When using sequence-based prediction, learning on partially annotated data is not straightforward, as the data that must be used to train context-based transition probabilities may be left unannotated. In contrast, in the pointwise prediction framework, training using this data is both simple and efficient; unannotated points are simply ignored. A method for learning CRFs from partially annotated data has been presented by Tsuboi et al. (2008). However, when using partial annotation, CRFs' already slow training time becomes slower still, as they must be trained over every sequence that has at least one annotated point. Training time is important in an active learning situation, as an annotator must wait while the model is being re-trained.

4 Experiments

In order to test the effectiveness of pointwise MA, we did an experiment measuring accuracy both on in-domain data, and in a domain-adaptation situation. We used the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa, 2008), specifying the whitepaper, news, and books sections as our general domain corpus, and the web text section as our target domain corpus (Table 3).

As a representative of joint sequence-based MA described in 2.1, we used MeCab (Kudo, 2006), an open source implementation of Kudo et al. (2004)'s CRF-based method (we will call this JOINT). For the pointwise two-step method, we trained logistic regression models with the LIBLINEAR toolkit (Fan et al., 2008) using the features described in Section 2.2 (2-LR). In addition, we trained a CRF-based model with the CRFSuite toolkit (Okazaki, 2007) using the same features and set-up (for both word

Train	Test	JOINT	2-CRF	2-LR
GEN	GEN	97.31%	98.08%	98.03%
GEN	TAR	94.57%	95.39%	95.13%
GEN+TAR	TAR	96.45%	96.91%	96.82%

Table 4: Word/POS F-measure for each method when trained and tested on general (GEN) or target (TAR) domain corpora.

segmentation and POS tagging) to examine the contribution of context information (2-CRF).

To create the dictionary, we added all of the words in the corpus, but left out a small portion of singletons to prevent overfitting on the training data[3]. As an evaluation measure, we follow Nagata (1994) and Kudo et al. (2004) and use Word/POS tag pair F-measure, so that both word boundaries and POS tags must be correct for a word to be considered correct.

4.1 Analysis Results

In our first experiment we compared the accuracy of the three methods on both the in-domain and out-of-domain test sets (Table 4). It can be seen that 2-LR outperforms JOINT, and achieves similar but slightly inferior results to 2-CRF. The reason for accuracy gains over JOINT lies largely in the fact that while JOINT is more reliant on the dictionary, and thus tends to mis-segment unknown words, the two-step methods are significantly more robust. The small difference between 2-LR and 2-CRF indicates that given a significantly rich feature set, context-based features provide little advantage, although the advantage is larger on out-of-domain data. In addition, training of 2-LR is significantly faster than 2-CRF. 2-LR took 16m44s to train, while 2-CRF took 51m19s to train on a 3.33GHz Intel Xeon CPU.

4.2 Domain Adaptation

Our second experiment focused on the domain adaptability of each method. Using the target domain training corpus as a pool of unannotated data, we performed active learning-based domain adaptation using two techniques.

- Sentence-based annotation (SENT), where sentences with the lowest total POS and word boundary probabilities were annotated first.
- Word-based partial annotation (PART), where the word or word boundary with the smallest probability margin between the first and second candidates was chosen. This can only be used with the pointwise 2-LR approach[4] .

[3]For JOINT we removed singletons randomly until coverage was 99.99%, and for 2-LR and 2-CRF coverage was set to 99%, which gave the best results on held-out data.

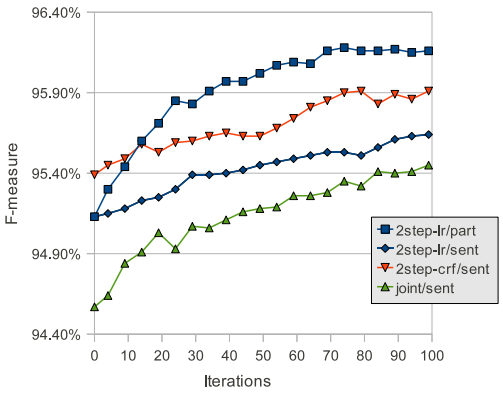

Figure 2: Domain adaptation results for three approaches and two annotation methods.

For both methods, 100 words (or for SENT until the end of the sentence in which the 100th word is reached) are annotated, then the classifier is retrained and new probability scores are generated. Each set of 100 words is a single iteration, and 100 iterations were performed for each method.

From the results in Figure 2, it can be seen that the combination of PART and 2-LR allows for significantly faster adaptation than other approaches, achieving accuracy gains in 15 iterations that are achieved in 100 iterations with SENT, and surpassing 2-CRF after 15 iterations. Finally, it can be seen that JOINT improves at a pace similar to PART, likely due to the fact that its pre-adaptation accuracy is lower than the other methods. It can be seen from Table 4 that even after adaptation with the full corpus, it will still lag behind the two-step methods.

5 Conclusion

This paper proposed a pointwise approach to Japanese morphological analysis. It showed that despite the lack of structure, it was able to achieve re-

[4]In order to prevent wasteful annotation, each unique word was only annotated once per iteration.

sults that meet or exceed structured prediction methods. We also demonstrated that it is both robust and adaptable to out-of-domain text through the use of partial annotation and active learning. Future work in this area will include examination of performance on other tasks and languages.

References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech tagger. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 21–27.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Taku Kudo. 2006. MeCab: yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net.

Percy Liang, Hal Daumé III, and Dan Klein. 2008. Structure compilation: trading structure for features. In *Proceedings of the 25th International Conference on Machine Learning*, pages 592–599.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*, pages 101–102.

Masaaki Nagata. 1994. A stochastic Japanese morphological analyzer using a forward-DP backward-A* N-best search algorithm. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 201–207.

Tetsuji Nakagawa. 2004. Chinese and Japanese word segmentation using word-level and character-level information. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Graham Neubig and Shinsuke Mori. 2010. Word-based partial annotation for efficient corpus construction. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. 2010. Domain Adaptation meets Active Learning. In *Workshop on Active Learning for Natural Language Processing (ALNLP-10)*.

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, pages 101–108.

Manabu. Sassano and Sadao Kurohashi. 2010. Using smaller constituents rather than sentences in active learning for Japanese dependency parsing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 356–365.

Manabu Sassano. 2002. An empirical study of active learning with support vector machines for Japanese word segmentation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 505–512.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Yuta Tsuboi, Hisashi Kashima, Hiroki Oda, Shinsuke Mori, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of the 22th International Conference on Computational Linguistics*, pages 897–904.