Language Resource Addition: Dictionary or Corpus?

Shinsuke Mori,[1] Graham Neubig[2]

[1]Academic Center for Computing and Media Studies, Kyoto University
[2]Nara Institute of Science and Technology
[1]Yoshidahonmachi, Sakyo-ku, Kyoto, Japan
[2]8916-5 Takayamacho, Ikoma, Nara, Japan
[1]forest@i.kyoto-u.ac.jp, [2]neubig@is.naist.jp

Abstract

In this paper, we investigate the relative effect of two strategies of language resource additions to the word segmentation problem and part-of-speech tagging problem in Japanese. The first strategy is adding entries to the dictionary and the second is adding annotated sentences to the training corpus. The experimental results showed that the annotated sentence addition to the training corpus is better than the entries addition to the dictionary. And the annotated sentence addition is efficient especially when we add new words with contexts of three real occurrences as partially annotated sentences. According to this knowledge, we executed annotation on the invention disclosure texts and observed word segmentation accuracy.

Keywords: Partial annotation, Dictionary, Word segmentation, POS tagging

1. Introduction

The importance of language resources continues to increase in the era of natural language processing (NLP) based on machine learning techniques. For mature NLP applied to real problems, such as word segmentation, part-of-speech (POS) tagging, etc., relatively high accuracies are achieved on general-domain data, and much of the problem lies in adaptation to new domains. To cope with this problem, there are many attempts at semi-supervised training and active learning (Tomanek and Hahn, 2009; Settles et al., 2008; Sassano, 2002). However, the simple strategies of corpus annotation or dictionary expansion are highly effective and not so costly. In fact, according to authors' experiences it only took 7 hours $\times$ 10 days to annotate 5,000 sentences precisely with word boundary information, enough to achieve large gains in a domain adaptation setting.

Within the context of sequence labeling, a variety of resources can be used, including annotated training data, which gives us information about word use in context, and dictionaries, which lack context information but are often available at large scale. In this paper, we investigate the relative effect of dictionary expansion and annotated corpus addition (full annotation and partial annotation) to the Japanese morphological analysis problem (MA; a joint task of word segmentation and POS tagging) and word segmentation problem.

2. 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; Neubig et al., 2011), or as a single joint process of finding a morpheme/POS string from unsegmented text (Nagata, 1994; Mori and Kurata, 2005; Kudo et al., 2004; Nakagawa, 2004; Kruengkrai et al., 2009).

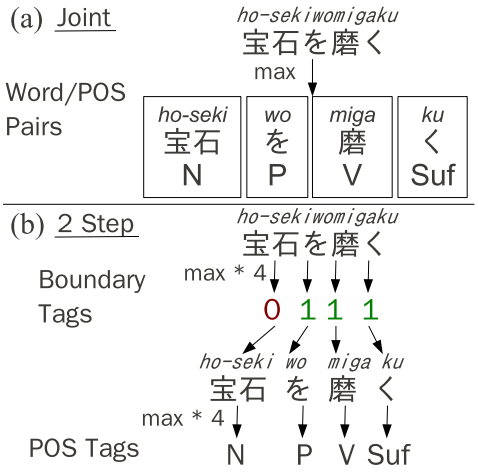

Figure 1: Joint MA (a) performs maximization over the entire sequence, while two-step MA (b) maximizes the 4 boundary and 4 POS tags independently.

2.1. Joint Sequence-Based MA

Japanese MA has traditionally used sequence based models, finding a maximal POS sequence for entire 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.

The main feature of this approach in the context of the current paper is that it relies heavily on a complete and accurate dictionary. In general when building the lattice of candidates from which to choose, it is common to consider only candidates that are in a pre-defined dictionary, only adding character sequences that are not in the dictionary

when there are no in-vocabulary candidates.[1] Thus, if the dictionary contains all of the words present in the sentences we want to analyze, these methods will obtain relatively high accuracy, but any words not included in the dictionary will almost certainly be given a mistaken analysis.

2.2. 2-Step Pointwise MA

In the two-step approach (Neubig et al., 2011), on the other hand, we first segment character sequence $\boldsymbol{x}_1^I$ into the word sequence $\boldsymbol{w}_1^J$ with the highest probability, then tag each word with parts of speech $\boldsymbol{t}_1^J$. This approach is shown in Figure 1 (b).

Word segmentation is formulated 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).

As features for these classification problems, it is common to use information about the surrounding characters (character and character-type n-grams), as well as the presence or absence of words in the dictionary. The details of the features can be found in Neubig et al. (2011), but as dictionary features are particularly important in the context of this paper we explain them shortly here. Dictionary features for word segmentation can include, for example, l_s and r_s which 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 which is active if the present word boundary is included in a dictionary word of length s. Dictionary feature d_{jk} for POS estimation can indicate whether the current word w_j occurs as a dictionary entry with tag t_k.

Compared to the joint sequence-based method described in the previous section, the two-step approach is a dictionary-light method. In fact, given a corpus of segmented and POS-tagged sentences, it is possible to perform analysis without the dictionary features, relying entirely on the information about the surrounding n-grams learned from the corpus. However, as large-coverage dictionaries often exist in many domains for consumption by either computer or human, having the possibility to use these as additional features is expected to give a gain in accuracy, which we verify experimentally in the following section.

3. Experimental Evaluation

To observe the difference between the addition of annotated sentences to the training corpus, and addition of entries to the dictionary, we conducted the experiments described below.

[1] It should be noted that there has been a recently proposed method to loosen this restriction, although this adds some complexity to the decoding process and reduces speed somewhat (Kaji and Kitsuregawa, 2013).

Corpus		
Domain	#words	
General	784k	
General + Web	898k	
Web for test	13.0k	
Dictionary		
Domain	#words	Coverage (word/POS)
General	29.7k	96.3%
General + Web	32.5k	97.9%

Table 1: Language Resource Specification.

Adaptation strategy	MeCab	KyTea
No adaptation	95.20%	95.54%
Dict. addition (no re-training)	96.59%	-
Dict. addition (re-training)	96.55%	96.75%
Corpus addition	96.85%	97.15%

Table 2: Word Segmentation Accuracy (F-measure).

3.1. Experimental Setting

The task we use as our test bed is the domain adaptation of Japanese morphological analysis. We use the Balanced Corpus of Contemporary Written Japanese (BCCWJ) as the testbed for our experiments (Maekawa, 2008). BCCWJ is divided into several sections, each from a different source, so this is ideal for domain adaptation experiments.

As our target domain, we use data from the Web (Yahoo! *Chiebukuro* in BCCWJ) and as the source domain we use the other five domains of BCCWJ Core data. Table 1 shows the specification of the corpus and dictionary.

As morphological analyzers, we use the following two publicly available tools[2].

1. MeCab: CRF-based joint method (Kudo et al., 2004)
2. KyTea: 2-step pointwise method (Neubig et al., 2011)

We compare the following adaptation strategies for the two morphological analyzers.

- No adaptation: Use the corpus and the dictionary in the general domain.
- Dictionary addition (no re-training): Add words appearing in the Web training corpus to the dictionary. As the dictionary includes costs, we set the cost of all new words to the same value as infrequent words of the same POS tag, following the instructions on the MeCab Web page[3] (MeCab only).
- Dictionary addition (re-training): Add words appearing in the Web corpus to the dictionary and estimate the weights of the model on the general domain training data again.
- Corpus addition: Create a dictionary from both the general and Web domains, and train the parameters on the same corpus from both domains.

[2] We did not precisely tune the parameters, so there still may be room for further improvement.

[3] `http://mecab.sourceforge.net/dic.html`

3.2. Evaluation Criterion

As an evaluation criterion we follow (Nagata, 1994) and use precision and recall based on word-POS pair. First the longest common subsequence (LCS) is found between the correct answer and system output. Then let N_{REF} be the number of word-POS pairs in the correct sentence, N_{SYS} be that in the output in a system, and N_{LCS} be that in the LCS of the correct sentence and the output of the system, so the recall R and precision P are defined as follows:

$$R = \frac{N_{LCS}}{N_{REF}}, \quad P = \frac{N_{LCS}}{N_{SYS}}.$$

Finally we calculate F-measure defined as the harmonic mean of the recall and the precision:

$$F = \left\{ \frac{1}{2}(R^{-1} + P^{-1}) \right\}^{-1} = \frac{2N_{LCS}}{N_{REF} + N_{SYS}}.$$

3.3. Result and Discussion

Table 2 shows the experimental result. From this table, we can see that just adding entries to the dictionary has a large positive effect on the accuracy. By adding entries to the dictionary (no re-training in MeCab case[4]) the accuracies of MeCab and KyTea increase by 1.35% and 1.21% respectively. However, by actually adding annotated sentences to the training corpus we can further increase by 0.30% and 0.40% respectively. That is to say, 75∼80% of accuracy increase can be achieved through dictionary expansion and the remaining 20∼25% can realized only by adding the context information included in the corpus.
The followings are the examples of increases realized only by the corpus addition for MeCab.

- な / ん ⇒ なん (freq.=4)
  In books and newspaper articles “なん”(what) is written in the Chinese character “何” instead of the *hiragana* “なん.” Thus the morphological analyzer divides the string into the auxiliary verb “な” and its inflectional ending “ん” which appear many times in these domains.

- ^ / ^ ⇒ ^^ (freq.=3)
  Smiley faces are rare in the general domain but often used in Web domain. And characters including “^” make a word in many cases. Thus we need to add a Web domain training corpus to estimate that the smiley face is sufficiently common as a single word and should not be divided.

- 感 / じ ⇒ 感じ (freq.=2)
  “感じ”(feeling) as a noun does not appear in the general domain corpus and is segmented into a verb “感” and inflectional endings “じ”, but using this word as a noun is common in the Web domain.

Another remark is that the accuracy gain is almost the same in CRF-based joint method (MeCab) and 2-step pointwise method (KyTea) contrary to our expectation that MeCab depends more on the dictionary than KyTea. Thus both morphological analyzers are making good use of dictionary information, but also can be improved with the context provided by the corpus.

[4] As we can see in Table 2, renewing CRF parameters decreased the accuracy.

	#Sent.	#NEs	#Words	#Char.
Training	1,760	13,197	33,088	50,002
Test	724	–	13,147	19,975

Table 3: Specifications of the recipe corpus.

4. Realistic Cases

The experimental results that we described in the previous section are somewhat artificial or *in-vitro*. In the corpus addition case, it is assumed that the sentences are entirely annotated with word boundary information and all the words are annotated with their POSs.
In this section, we report results under two other adaptation methods used in real or *in-vivo* adaptation scenarios. In both cases, the language resources to be added are partially annotated corpora (Neubig and Mori, 2010). Because MeCab is not capable of training a model from such corpora, we only report the result of KyTea.
As the problem, we focus on word segmentation, because in Japanese most ambiguity in MA lies in word segmentation, especially in the domain adaptation situation where most of unknown words are nouns and the rest fall into other content word categories such as verbs, adjectives, etc.
Figure 2

4.1. Recipe Domain

The first case is the adaptation to cooking recipe texts. We used recipe flow graph corpus (r-FG corpus) (Mori et al., 2014) in which word sequences important for cooking are annotated with types (recipe named entities; recipe NEs). They are also correctly segmented into words (see Figure 2).

4.1.1. Experimental Setting

Table 3 shows the specifications of the r-FG corpus relating to the word segmentation experiment. As the adaptation strategies, we used the following two methods in addition to “No adaptation.” The examples are taken from Figure 2.

Dictionary Use the training data as a dictionary.

1. Extract NEs from the training data,
   ex.) /ホットドッグ/F, /チリ/F, /チーズ/F, /オニオン/F, /ふりかけ/Ac, /ホットドッグ/F, /アルミホイル/F, /覆/Ac
2. Make a dictionary containing the words in these NEs,
   ex.) ホット, ドッグ, チリ, チーズ, オニオン, ふりかけ, アルミ, ホイル, 覆
3. Use the dictionary as the additional language resource to train the model.

Partial annotation Use the training data as partially annotated data.

各 /ホットドッグ/F に /チリ/F 、 /チーズ/F 、 /オニオン/F を /ふりかけ/Ac る
(each) (hot dog) (*cmi*) (chili) , (cheese) , (onion) (*cmd*) (sprinkle) (*infl.*)

/ホットドッグ/F を /アルミホイル/F で /覆/Ac う
(hot dog) (*cmd*) (aluminum foil) (*cmc*) (cover) (*infl.*)

English is added for explanation only. *cmc*, *cmd*, and *cmi* stand for case marker for complement, direct object, and indirect object, respectively. *infl.* stands for inflectional ending.

Figure 2: Example sentences in the r-FG corpus.

Adaptation strategy	#occurrences		#words	WS F-measure	
	maximum (n)	average		BCCWJ	Recipe
No adaptation	–	–	0	98.87%	94.35%
Dictionary	–	–	1,999	98.90%	94.54%
Partial annotation	1	1.00	1,999	98.89%	95.56%
	2	1.60	3,191	98.89%	95.81%
	3	2.02	4,046	98.89%	95.94%
	4	2.36	4,727	98.89%	96.01%
	8	3.26	6,523	98.89%	96.07%
	16	4.26	8,512	98.89%	96.14%
	32	5.10	10,203	98.89%	96.21%
	64	5.77	11,542	98.89%	96.28%
	∞	6.60	13,197	98.89%	96.29%

Table 4: Word Segmentation Accuracy in Partial Annotation Case.

1. Extract n occurrences at maximum of the NEs from the training data (see Figure 2, where the NE in focus is ホットドッグ and $n = 2$),
2. Convert them into partially segmented sentences in which only both edges of the NEs and the inside of the NEs are annotated with word boundary information.

   ex.) If the NE in focus is ホットドッグ, then
   各 | ホ – ッ – ト | ド – ッ – グ | に␣チ␣リ␣、…,
   | ホ – ッ – ト | ド – ッ – グ | を␣ア␣ル␣ミ…,

   where the symbols "|," "–," and "␣" mean word boundary, not a word boundary, and no information, respectively.
3. Use the partially annotated data as the additional language resource to train the model.

4.1.2. Result and Discussion

Table 4 shows the word segmentation accuracies (WS F-measure) of "No adaptation" and the strategies that we explained above. The results of the partial annotation strategy varies depending on the parameter n (the maximum occurrences). The table shows these results with the real average occurrences in the partially segmented sentences.

From the result we can note the following. First, the addition of new words as the dictionary to the training data improves the word segmenter. This is consistent with the results shown in Table 2. Second, the partial annotation strategy with one occurrence ($n = 1$) is as good as the dictionary addition strategy. And as we increase the number of occurrences (n), the segmenter improves. The degree of improvement, however, shrinks as n increases. In a real situation, we have to prepare such partially annotated data and the annotation cost is proportional to the number of occurrences to be annotated. Therefore it is good to start annotating new words in descending order of frequency, selecting a threshold based on the number of occurrences.

4.2. Invention Disclosure Domain

Finally we report the result of a real adaptation that performed. The target domain is the invention disclosure texts, which are one of the important domains for NLP, especially information extraction and machine translation.

4.2.1. Setting

Based on the knowledge we described above, we adopted the partial annotation strategy. Concretely, we performed the following procedures.

1. Extract unknown word candidates based on the distributional similarity from a large raw corpus in the target domain (Mori and Nagao, 1996),
2. Annotate three occurrences with word boundary information to make partially segmented sentences for each unknown word candidate in the descending order of the expected frequencies[5].

For frequent word candidates, i.e. in the beginning of the annotation work, the three-occurrence annotation corresponds to the case of those with the maximum occurrence count of 4 and 8 in Table 4, because the average number of the occurrences is expected to be three.

In the practice, we asked an annotator to check unknown word candidates with three different contexts in the raw

[5]The expected frequency of a word candidate is the frequency as a string in the raw corpus multiplied by the word likelihood estimated by the comparison between its distribution and that of the words. See (Mori and Nagao, 1996) for more detail.

	#Sent.	#Words	#Char.
Test	500	20,658	32,139

Table 5: Specifications of the invention disclosure corpus.

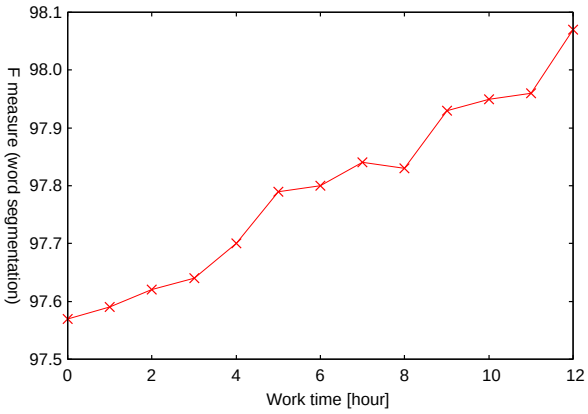

Figure 3: Accuracy increase.

corpus and correct the word boundary information if the default[6] is incorrect.

Every time the annotator finished one hour work, we built a word segmenter by adding the partially annotated sentences and measured the accuracy (WS F-measure) on the test set shown in Table 5.

4.2.2. Result and Discussion

The learning curve shown in Figure 3. The left most point corresponds to the "No adaptation" case. The accuracy in this case is high compared with the recipe domain (Table 4). The reason is that the invention disclosure domain is not much different from the general domain containing newspaper articles etc. The most important remark is that the accuracy gets higher as we add more unknown word candidates to the training data as partially annotated sentences. After 12 hours of annotation work, we succeeded to eliminate 20% of the errors. The absolute F-measure is almost the same as that of the state-of-the-art word segmenter on the test set in the same domain as the training data (Neubig et al., 2011). Thus the word segmenter model itself is capable of contributing to various NLP applications in the invention disclosure domain in Japanese. In addition the accuracy does not seem to be saturating, thus we can improve more by only more annotator's work based on the partial annotation strategy.

5. Conclusion

In this paper, we reported to what extent two strategies of language resource additions contribute to improvement in the word segmentation problem and POS tagging problem in Japanese. The first strategy is adding entries to the dictionary and the second is adding annotated sentences to the training corpus.

In the experimental evaluations, we first showed that the corpus addition strategy is better than the dictionary addition strategy in the Japanese morphological task. Then we introduced the partial annotation strategy, in which only important points are annotated with word boundary information, and reported the real cases focusing on the word segmentation in Japanese. The experiment showed that adding word candidates to the training data as partially annotated data with about three different contexts is efficient to improve the word segmenter.

6. Acknowledgments

This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 23500177 and NTT agreement dated 05/23/2013.

7. References

Nobuhiro Kaji and Masaru Kitsuregawa. 2013. Efficient word lattice generation for joint word segmentation and pos tagging in japanese. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 153–161, Nagoya, Japan.

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.

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

Shinsuke Mori and Gakuto Kurata. 2005. Class-based variable memory length markov model. In *Proceedings of the InterSpeech2005*, pages 13–16.

Shinsuke Mori and Makoto Nagao. 1996. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Proceedings of the 16th International Conference on Computational Linguistics*.

Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. 2014. Flow graph corpus from recipe texts. In *Proceedings of the Nineth International Conference on Language Resources and Evaluation*.

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 Seventh International Conference on Language Resources and Evaluation*.

[6]The default segmentation assumes that the candidate word is a word. That is to say, there are word boundaries on the both edges and no word boundary inside the string.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

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*.

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, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *NIPS Workshop on Cost-Sensitive Learning*.

Katrin Tomanek and Udo Hahn. 2009. Semi-supervised active learning for sequence labeling. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1047.