Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

A Comparative Study of Dictionaries and Corpora as Methods for Language Resource Addition

Shinsuke Mori · Graham Neubig

Received: date / Accepted: date

Abstract In this paper, we investigate the relative effect of two strategies for language resource addition for Japanese morphological analysis, a joint task of word segmentation and part-of-speech tagging. 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 addition of annotated sentences to the training corpus is better than the addition of entries to the dictionary. In particular, adding annotated sentences is especially efficient when we add new words with contexts of several real occurrences as partially annotated sentences, *i.e.* sentences in which only some words are annotated with word boundary information.

According to this knowledge, we performed real annotation experiments on invention disclosure texts and observed word segmentation accuracy.

Finally we investigated various language resource addition cases and introduced the notion of *non-maleficence*, *asymmetricity*, and *additivity* of language resources for a task. In the WS case, we found that language resource addition is non-maleficent (adding new resources causes no harm in other domains) and sometimes additive (adding new resources helps other domains). We conclude that it is reasonable for us, NLP tool providers, to distribute only one general-domain model trained from all the language resources we have.

Keywords Partial annotation · Domain adaptation · Dictionary · Word segmentation · POS tagging · Non-maleficence of language resources

The current paper describes and extends the language resource creation activities, experimental results, and findings that have previously appeared as an LREC paper (Mori and Neubig, 2014).

Shinsuke Mori
Academic Center for Computing and Media Studies, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, Japan
Tel.: +81-75-753-7486 Fax: +81-75-753-7475 E-mail: forest@i.kyoto-u.ac.jp

Graham Neubig
Nara Institute of Science and Technology
8916-5 Takayamacho, Ikoma, Nara, Japan
E-mail: neubig@is.naist.jp

1 Introduction

The importance of language resources continues to increase in the era of natural language processing (NLP) based on machine learning (ML) techniques. Well defined annotation standards and rich language resources have enabled us to build very accurate ML-based analyzers in the general domain. Representative tasks include word segmentation for languages without word boundaries such as Japanese and Chinese, part-of-speech (POS) tagging, and many others. These are the first step in NLP for many languages and have a great impact on subsequent processes. However, while the accuracies are more than 97% for texts from the same domain as the training corpus, large drops in accuracy are seen when using these models in a different domain. To cope with the problem of adapting to new domains, 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 still highly effective and not unbearably costly. In fact, according to authors' experience with annotation for Japanese word segmentation, it only took 7 hours $\times$ 10 days to annotate 5,000 sentences (about 40 words per sentence) with word-boundary information including two check processes.[1] As shown in the subsequent parts of this paper, 5,000 annotated sentences are often enough to achieve large gains in domain adaptation for sequence labeling.

The 5,000 sentences mentioned above were so-called fully annotated sentences, where all the positions between two characters are annotated with word-boundary information. Within the context of sequence labeling, however, a variety of resources can be used, including partially annotated sentences and dictionaries. In contrast to fully annotated sentences, partially annotated sentences lack labels at some points. These annotated sentences give us information about word use in context, without requiring the annotator to annotate the full sentence. On the other hand, dictionaries lack context information but are often available at large scale.

In this paper, we first investigate the relative effect of dictionary expansion and annotated corpus addition (full annotation and partial annotation) for the Japanese morphological analysis (MA) problem (a joint task of word segmentation and POS tagging) and the word segmentation problem. Then we present some real adaptation cases, including invention disclosures and recipe texts. Finally we introduce a notion of *non-maleficence* of language resources, which means that the addition of language resources in a certain domain causes no harm to segmentation accuracy for another domain.

2 Related Work

The task we have chosen is Japanese MA, a joint task of word segmentation and POS tagging. Although in Japanese most of the ambiguity in MA lies in

[1] In the first check process the annotator focused on words appearing only in the newly annotated 5,000 sentences. In the second process we divide annotated sentences into several parts and the annotator checked the differences between the manual annotations of each part and the machine decisions by a model trained on the corpus including the other parts, similarly to cross validation.

word segmentation,[2] in the first half of the experimental evaluation we address the full MA task for comparison between two standard methods in the field: the joint sequence-based method (Kudo et al, 2004) and the 2-step pointwise method (Neubig et al, 2011). In the second half, we focus on word segmentation.

After a long history of rule-based MA methods, the first statistical method was proposed by Nagata (1994). It was based on a hidden Markov model whose states correspond to POS tags. Then Mori and Kurata (2005) extended it by lexicalizing the states like many works in that era, grouping the word-POS pairs into clusters inspired by the class-based n-gram model (Brown et al, 1992), and making the history length variable like it has been done for a POS tagger in English (Ron et al, 1996). In parallel, Kudo et al (2004) applied conditional random fields (CRFs) (Lafferty et al, 2001) to this task and showed that it achieved better performance than a POS-based Markov model. This CRF-based method does not have an unknown word model and large drops in accuracy are seen when it is applied to a different domain from the training data. For unknown words Nakagawa (2004) and Kruengkrai et al (2009) proposed a word-based model (not CRF-based) equipped with an unknown word model represented by position-of-character tags and reported comparable accuracies to the CRF-based method.

Along with the model evolution, the NLP community has been becoming more and more aware of the importance of language resources. In addition we have observed a drastic degradation in word segmentation accuracy in domain adaptation cases. Because most unknown words are nouns and the rest fall into other content word categories such as verbs, adjectives, it is less difficult to estimate the POS given the correct segmentation than word segmentation of sentences including unknown words. Thus the Japanese MA research focus has been shifted to word segmentation. Given this background Mori and Oda (2009) proposed a method for referring to the words in a dictionary prepared for humans, not computers. The entries in these dictionaries tend to be compound words but not words defined according to the annotation standard.[3] Another important extension to enlarge language resource availability is CRFs trainable from partially annotated sentences (Tsuboi et al, 2008). This is applied to Chinese word segmentation capitalizing on so-called natural annotations such as tags in hyper-texts (Yang and Vozila, 2014). Because the training time of sequence-based methods tends to be long, a simple method based on pointwise classification has been shown to be comparable to sequence-based methods (Neubig et al, 2011). Since the pointwise method decides whether there is a word boundary or not between two characters without referring to the decisions on the other points, we can train the model from partially annotated sentences in a straightforward way.[4]

[2] We had run some experiments. BCCWJ consists of six domains. We split each of them into the test and train. Then we built a model from five training data and tested it on the rest of the data in the other domain. When we use Yahoo! QA as test, WS and MA accuracies are 98.64 and 97.78, respectively. The WS errors are 61.3% of those of MA. When the test is Yahoo! blogs, the most difficult domain among six, the accuracies are 96.98 and 95.77, so the WS errors are 71.4% of those of MA.

[3] For example an entry “フランス語” (French language) is a combination of “フランス” (France) and “語” (language).

[4] Note that it is also possible to learn sequence-based models from partial annotations (Tsuboi et al, 2008; Yang and Vozila, 2014), which may provide an increase of accuracy at the cost of an increase in training time (the total time for training CRFs on partially annotated data scales in the number of words in sentences with at least one annotation, in contrast

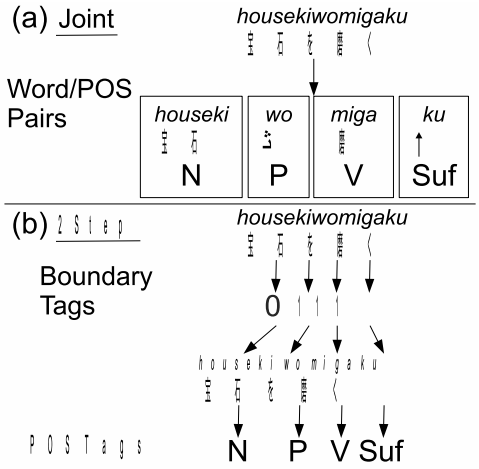

Fig. 1 Joint MA (a) performs maximization over the entire sequence, while two-step MA (b) maximizes the 4 boundaries and 4 POS tags independently.

Type	Feature strings
1-gram	t_j, $t_j w_j$, $c(w_j)$, $t_j c(w_j)$
2-gram	$t_{j-1}t_j$, $t_{j-1}t_j w_{j-1}$, $t_{j-1}t_j w_j$, $t_{j-1}t_j w_{j-1}w_j$

Table 1 Features for the joint model using tags t and words w. $c(\cdot)$ is a mapping function onto the sequence of one of the six character types mentioned in the body text.

3 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 POS (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). In this section we explain these approaches briefly, and contrast their various characteristics.

3.1 Joint Sequence-Based MA

Japanese MA has traditionally used sequence-based models, finding the highest-scoring POS sequence for entire sentences as in **Fig. 1** (a). The CRF-based method presented by Kudo et al (2004) is a widely used baseline 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.

to the pointwise approach, which scales in the number of annotated words). A comparison between these two methods is orthogonal to our present goal of comparing dictionary and corpus addition, and thus we use pointwise predictors in our experiments.

Type	Feature strings
Character n-gram	x_l, x_r, $x_{l-1}x_l$, x_lx_r, x_rx_{r+1}, $x_{l-1}x_lx_r$, $x_lx_rx_{r+1}$
Character type n-gram	$c(x_l)$, $c(x_r)$ $c(x_{l-1})c(x_l)$, $c(x_l)c(x_r)$, $c(x_r)c(x_{r+1})$ $c(x_{l-2})c(x_{l-1})c(x_l)$, $c(x_{l-1})c(x_l)c(x_r)$ $c(x_l)c(x_r)c(x_{r+1})$, $c(x_r)c(x_{r+1})c(x_{r+2})$
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. For example, when estimating word boundary b_i, $l = i - 1$ and $r = i$. l_s, r_s, and i_s represent the features expressing whether a word in the dictionary exists to the left, right, or spanning the current boundary, while d_{jk} indicates that tag k exists in the dictionary for word j.

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.[5] 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.

We follow (Kudo et al, 2004) in defining our feature set, as summarized in **Table 1**[6]. Lexical features were trained for the top 5,000 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 POS.

3.2 Two-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 POS $\boldsymbol{t}_1^J$. This approach is shown in **Fig. 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 CRFs.

[5] 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).

[6] 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.

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 (Neubig et al, 2011). The character n-gram features fire binary features for each character n-grams of length 1-3 in the neighborhood of the current segmentation boundary. The character type n-gram features similarly are binary features, but with each character generalized from its surface form to one of six different types of characters widely used in the Japanese language: *kanji*, *hiragana*, *katakana*, *arabic number*, *alphabet*, and *symbol*. *kanji* is ideograms, while *hiragana* and *katakana* are phonograms mainly used for function words and imported words, respectively. Dictionary features include l_s and r_s which are active if a string of s characters 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 s characters. 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 Subsection 3.1 (Kudo et al, 2004), 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.

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, 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[7]. This allows for training from sentences that are partially annotated practically.

4 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.

4.1 Experimental Setting

The task we use as our test bed is the domain adaptation of Japanese MA. We use the Core part of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa, 2008) as the data for our experiments. The BCCWJ Core

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

Corpus		
Domain	#words	
General	784,000	
General + Web	898,000	
Web for test	13,000	
Dictionary		
Domain	#words	Coverage (word-POS pairs)
General	29,700	96.3%
General + Web	32,500	97.9%

Table 3 Language Resource Specifications.

data is divided into six 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 general domain we use a joint corpus built by putting together the other five domains of BCCWJ Core data. **Table 3** shows the specifications of the corpus and dictionary.

As morphological analyzers, we use the following two publicly available tools.[8]

1. `MeCab`: CRF-based joint method (Kudo et al, 2004)
2. `KyTea`: two-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 weights, we set the weight of all new words to the same value as infrequent words of the same POS tag, following the instructions on the `MeCab` Web page[9] (`MeCab` only [10]).
- 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: Combine the training corpora from the general and Web domains, and train the parameters on the result.

4.2 Evaluation Criterion

As an evaluation criterion we follow Nagata (1994) and use precision and recall based on word-POS pairs. 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:

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

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

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

[10] `KyTea` requires re-training.

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 4 Word Segmentation Accuracy (F-measure).

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

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

4.3 Results and Discussion

Table 4 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 the MeCab case[11]) the accuracies of MeCab and KyTea increase by 1.39% and 1.21% respectively. However, by actually adding annotated sentences to the training corpus we can further increase by 0.26% and 0.40% respectively. That is to say, 75∼84% of the accuracy increase can be achieved through dictionary expansion and the remaining 16∼25% can be realized only by adding the context information included in the corpus.

The following are the examples of increases realized only by the corpus addition for MeCab.

- な(na) / ん(n) ⇒ 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.” Note that every Chinese character (ideogram) can be written by *hiragana*, (phonogram). Thus the morphological analyzer divides the string into the auxiliary verb “な”(become, get) 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 “^” are a single 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(kan) / じ(ji) ⇒ 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 two-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.

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

	#sent.	#r-NEs	#words	#char.
Training	1,760	13,197	33,088	50,002
Test	724	–	13,147	19,975

Table 5 Specifications of the recipe corpus.

各	/ホット ドッグ/F	に	/チリ/F	、	/チーズ/F	、	/オニオン/F	
(each)	(hot dog)	(*cmi*)	(chili)	,	(cheese)	,	(onion)	
を	/ふりかけ/Ac	る						
(*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 the case marker for complement, direct object, and indirect object, respectively. *infl.* stands for inflectional ending. F and Ac are type tags and mean food and action by chef, respectively.

Fig. 2 Example sentences in the r-FG corpus.

5 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 POS.

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 as we mentioned in Section 2, 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.

5.1 Recipe Domain

The first case we examine is the adaptation to cooking recipe texts. A cooking recipe text consists of sentences describing procedures. Because they do not contain difficult language phenomena for NLP such as tense, aspect, viewpoint, etc., they can be thought to be relatively easy for computers to understand. Thus they are suitable as a benchmark for the language understanding research. In addition, there is a large demand for more accurate NLP in recipe domains for intelligent recipe processing (Wang et al, 2008; Yamakata et al, 2013).

In the following experiment we used the recipe flow graph corpus (r-FG corpus) (Mori et al, 2014). **Table 5** shows the specifications of the r-FG corpus relating to the word segmentation experiment. In the corpus word sequences important for cooking are annotated with types (recipe named entities; r-NEs) and they are correctly segmented into words. **Fig. 2** shows two example sentences.

Adaptation strategy	#occurrences		#words	WS F-measure	
	Maximum (n)	Average		BCCWJ	Recipe
No adaptation	–	–	0	99.01%	95.56%
Dictionary	–	–	1,999	99.00%	95.78%
Partial annotation	1	1.00	1,999	99.01%	95.80%
	2	1.60	3,191	99.01%	96.06%
	3	2.02	4,046	99.01%	96.15%
	4	2.36	4,727	99.01%	96.27%
	8	3.26	6,523	99.01%	96.28%
	16	4.26	8,512	99.01%	96.33%
	32	5.10	10,203	99.01%	96.36%
	64	5.77	11,542	99.02%	96.38%
	∞	6.60	13,197	99.01%	96.43%

Table 6 Word segmentation accuracy in the partial annotation case.

5.1.1 Experimental Setting

As the adaptation strategies, we used the following two methods in addition to "No adaptation." The examples are taken from **Fig. 2**. F and Ac are type tags and mean food and action by the chef, respectively.

Dictionary: Use the training data as a dictionary.

1. Extract r-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 r-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.

1. Extract n occurrences at maximum of the r-NEs from the training data (see **Fig. 2**, where the r-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 r-NEs and the inside of the r-NEs are annotated with word-boundary information. If the r-NE in focus is ホットドッグ composed of two words, then the partially segmented sentences are
   ex.) 各|ホ-ッ-ト|ド-ッ-グ|に␣チ␣リ␣、 …,
   ex.) |ホ-ッ-ト|ド-ッ-グ|を␣ア␣ル␣ミ…,
   where the symbols "|," "-," and "␣" mean word boundary, no word boundary, and no information, respectively.
3. Use the partially annotated data as an additional language resource to train the model.

5.1.2 Result and Discussion

Table 6 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 vary depending on the parameter n (the maximum number of occur-

rences). The table shows these results with the real average number of occurrences in the partially segmented sentences.

From the result we can note several things. The BCCWJ results show that adding language resources in the recipe domain has no effect on the general domain accuracy. Regarding the results of the first recipe, 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 4**. 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. We describe a concrete way of doing so for real unannotated data in the following section.

5.2 Invention Disclosure Domain

Finally we report the result of a real adaptation experiment that we performed. The target domain is invention disclosure texts from patents, which are an important domain for NLP, especially information extraction (Nanba et al, 2011, inter alia) and machine translation (Goto et al, 2011, inter alia).

5.2.1 Setting

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

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.[12]

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 (average: 2.36) and 8 (average: 3.26) in **Table 6**, because the average number of the occurrences is expected to be three.

In practice, we first assign each word a default annotation, then ask an annotator to check unknown word candidates with three different contexts in the raw corpus and correct the word boundary information if the default is incorrect. **Fig. 3** shows two example word candidates with their three occurrences. The default segmentation assumes that the candidate word is really a word. In the first example case, "御部" is assumed to be a word, but it is a concatenation of a word fragment "御" in "制御" (control) and a suffix "部" (part) at all three occurrences.

[12] 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 the distribution of the word candidate and that of the words. See (Mori and Nagao, 1996) for more detail.

	#sent.	#words	#char.
Test	500	20,658	32,139

Table 7 Specifications of the invention disclosure corpus.

Freq. = 200, candidate = “御部(gyobu)”
Default annotation by the system
␣こ␣の␣場␣合␣、␣第␣２␣の␣制|御-部|２␣２␣は␣初␣期␣動␣作␣ル␣ー␣チ␣ン␣の␣み␣を␣行␣う␣。
␣ま␣た␣、␣第␣１␣の␣制|御-部|２␣１␣は␣、␣Ｓ␣８␣０␣で␣、␣ホ␣ス␣ト␣か␣ら␣記␣録␣再␣生␣…
␣プ␣リ␣ン␣タ␣９␣５␣Ｂ␣の␣プ␣リ␣ン␣タ␣制|御-部|か␣ら␣開␣始␣信␣号␣が␣帰␣っ␣て␣く␣る␣…
After annotation work
␣こ␣の␣場␣合␣、␣第␣２␣の␣制-御|部|２␣２␣は␣初␣期␣動␣作␣ル␣ー␣チ␣ン␣の␣み␣を␣行␣う␣。
␣ま␣た␣、␣第␣１␣の␣制-御|部|２␣１␣は␣、␣Ｓ␣８␣０␣で␣、␣ホ␣ス␣ト␣か␣ら␣記␣録␣再␣生␣…
␣プ␣リ␣ン␣タ␣９␣５␣Ｂ␣の␣プ␣リ␣ン␣タ␣制-御|部|か␣ら␣開␣始␣信␣号␣が␣帰␣っ␣て␣く␣る␣…

Freq. = 80, candidate = “積層(sekiso)”
Default annotation by the system
…␣ｐ␣型␣半␣導␣体␣基␣板␣上␣に␣電␣荷␣蓄|積-層|（␣浮␣遊␣ゲ␣ー␣ト␣）␣と␣制␣御␣ゲ␣ー␣…
…␣バ␣イ␣モ␣ル␣フ␣形␣の␣も␣の␣の␣他␣、|積-層|形␣の␣も␣の␣を␣採␣用␣し␣て␣も␣良␣い␣。
…␣モ␣リ␣ブ␣デ␣ン␣等␣の␣金␣属␣を␣１␣０␣０␣０␣Å␣の␣膜␣厚␣に|積-層|し␣て␣、␣金␣属␣…
After annotation work
…␣ｐ␣型␣半␣導␣体␣基␣板␣上␣に␣電␣荷␣蓄-積|層|（␣浮␣遊␣ゲ␣ー␣ト␣）␣と␣制␣御␣ゲ␣ー␣…
…␣バ␣イ␣モ␣ル␣フ␣形␣の␣も␣の␣の␣他␣、|積-層|形␣の␣も␣の␣を␣採␣用␣し␣て␣も␣良␣い␣。
…␣モ␣リ␣ブ␣デ␣ン␣等␣の␣金␣属␣を␣１␣０␣０␣０␣Å␣の␣膜␣厚␣に|積-層|し␣て␣、␣金␣属␣…

Fig. 3 KWIC (key word in context) of unknown word candidates.

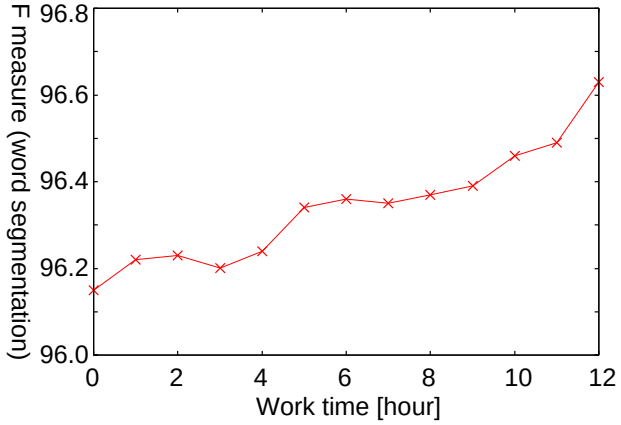

Fig. 4 Accuracy increase.

So the annotator changed the word-boundary information as shown in “After annotation work.” In the second example case, “積層(sekiso)” in the first line (context) is a concatenation of a word fragment “積(seki)” in “蓄積(chikuseki)” (accumulate) and a suffix “層(sou)” (layer). So the annotator changed the word-boundary information as shown in “After annotation work.” In the second and third line (context), however, “積層(sekiso)” (lamination) is a word and the annotator leaves it as the default without change.

Table 8 WS F-measure of the adapted models

Test domain	General	Recipe	Patent	Twitter
#sentences	3,680	724	500	542
Adaptation method	–	r-NE	KWIC	KWIC
		$n = 8$	12 hours	47 hours
No adapatation	99.01	95.56	96.15	85.90
Adaptation to				
recipe	99.01	96.28	96.56	85.99
patent	99.02	95.54	96.63	85.94
twitter	99.01	95.73	96.22	87.63
all	99.01	96.35	97.05	88.35

5.2.2 Result and Discussion

The learning curve is shown in **Fig. 4**. The leftmost point corresponds to the "No adaptation" case. The accuracy in this case is high compared with the recipe domain (**Table 6**) because the invention disclosure domain is not as stylistically different from the general domain containing newspaper articles etc. The most important thing to note 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 12% 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). This improved word segmenter model 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 work based on the partial annotation strategy.

6 Non-maleficence of Language Resources

In this section we introduce a notion of *non-maleficence*[13] of language resources. Briefly this means that the effects of a language resource in a certain domain does no harm in another domain.

Our experiments indicate that non-maleficence holds for Japanese word segmentation. In order to show this empirically, we provide results for an experiment in which we executed the adaptation of word segmenter `KyTea` to Twitter (micro blog) for 47 hours in addition to two domains which we described in Section 5. The adaptation method is exactly the same as the invention disclosure domain (patent).

In the experiment we made five models of `KyTea` in total. The first one is the default model without adaptation. The next three are models adapted to the recipe domain, invention disclosure (patent) domain, and Twitter. They are trained from the language resources in each domain in addition to the default language resources. In the recipe case we set $n = 8$ because it is realistic as we discussed above. In the patent case we used the maximum size of partially

[13] We borrow this terminology from medicine, where non-maleficence indicates the property of "doing no harm."

annotated sentences, i.e the 12-hour work result corresponding to the rightmost point in **Fig. 4**. The final `KyTea` model is trained from the language resources in all three adaptation domains in addition to the default language resources. Then we measured the accuracy of each model on the test data in the general domain, recipe domain, patent domain, and Twitter.

Table 8 shows the results. From these results we first see that the corpus addition to each domain improves the performance in that domain. In addition it does not degrade the performance in other domains.[14] We call this characteristic of the pair of a task and language resources non-maleficence. Its conditions are as follows:

- Language resource addition in a target domain is efficient for the target domain
- Language resource addition in other domains is not harmful for the target domain

We also see that in some cases language resource addition in other domains has a positive effect on the target domain. We term this *additivity* of language resources. A clear example is the model trained from the recipe tested on the patents. The reason may be that the recipe texts covers how-to expressions and the many frequent technical terms have already been covered by the general texts. On the contrary, language resource addition in the patent domain does not improve the performance on the recipe domain. Thus, we can say that the language resource addition is not *symmetric*. The model adapted to all the three domains performs almost always the best.[15] In this case as well, language resource addition is non-maleficent. The users of the NLP tool can always use the model trained from the maximum language resources, but not the model trained only from the language resource in the target domain.[16]

7 Conclusion

In this paper, we first reported to what extent two strategies of language resource addition contributed to improvements in word segmentation and POS tagging accuracy 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 allows for achievement of better accuracy than the dictionary addition strategy in the Japanese morphological analysis task.

We then introduced the partial annotation strategy, in which only important points are annotated with word-boundary information, and reported the real cases focusing on 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 a word segmenter.

[14] A very slight degradation is observed in case of recipe WS by the model trained from patent texts (from 95.56% to 95.54%). This is not statistically significant.

[15] The only exception is that the model adapted to the patent tested on the general domain is better than the others (from 99.01% to 99.02%). The change is, however, not significant.

[16] ML technologies have a possibility to adapt the model to an unexpected input automatically.

Finally we investigated various language resource addition cases and introduced the notion of non-maleficence, additivity, and asymmetricity of language resources for a task. Briefly non-maleficence means that language resource addition in a certain domain does no harm in another domain, additivity means that language resource addition in other domains has a positive effect on the target domain, and asymmetricity means that the effect of language resource addition is not symmetric. In the WS case, language addition is non-maleficent and sometimes additive. So it is enough for us, NLP tool providers, to distribute the only one model trained from all the language resources we have.

Acknowledgements This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 26280084 and 26540190, and NTT agreement dated 05/23/2013.

References

Brown PF, Pietra VJD, deSouza PV, Lai JC, Mercer RL (1992) Class-based n-gram models of natural language. Computational Linguistics 18(4):467–479

Goto I, Lu B, Chow KP, Sumita E, Tsou BK (2011) Overview of the patent machine translation task at the NTCIR-9 workshop. In: Proceedings of NTCIR-9 Workshop Meeting, pp 559–578

Kaji N, Kitsuregawa M (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, Nagoya, Japan, pp 153–161

Kruengkrai C, Uchimoto K, Kazama J, Wang Y, Torisawa K, Isahara H (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, pp 513–521

Kudo T, Yamamoto K, Matsumoto Y (2004) Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp 230–237

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth ICML, pp 282–289

Liang P, Daumé III H, Klein D (2008) Structure compilation: trading structure for features. In: Proceedings of the 25th ICML, pp 592–599

Maekawa K (2008) Balanced corpus of contemporary written Japanese. In: Proceedings of the 6th Workshop on Asian Language Resources, pp 101–102

Mori S, Kurata G (2005) Class-based variable memory length Markov model. In: Proceedings of the InterSpeech2005, pp 13–16

Mori S, Nagao M (1996) Word extraction from corpora and its part-of-speech estimation using distributional analysis. In: Proceedings of the 16th International Conference on Computational Linguistics, pp 1119–1122

Mori S, Neubig G (2014) Language resource addition: Dictionary or corpus? In: Proceedings of the Nineth International Conference on Language Resources and Evaluation, pp 1631–1636

Mori S, Oda H (2009) Automatic word segmentation using three types of dictionaries. In: Proceedings of the Eighth International Conference Pacific Association for Computational Linguistics

Mori S, Maeta H, Yamakata Y, Sasada T (2014) Flow graph corpus from recipe texts. In: Proceedings of the Nineth International Conference on Language Resources and Evaluation, pp 2370–2377

Nagata M (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, pp 201–207

Nakagawa T (2004) Chinese and Japanese word segmentation using word-level and character-level information. In: Proceedings of the 20th International Conference on Computational Linguistics, pp 466–472

Nanba H, Fujii A, Iwayama M, Hashimoto T (2011) Overview of the patent mining task at the ntcir-8 workshop. In: Proceedings of NTCIR-8 Workshop Meeting, pp 293–302

Neubig G, Mori S (2010) Word-based partial annotation for efficient corpus construction. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, pp 2723–2727

Neubig G, Nakata Y, Mori S (2011) Pointwise prediction for robust, adaptable Japanese morphological analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp 529–533

Ng HT, Low JK (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, pp 277–284

Peng F, Feng F, McCallum A (2004) Chinese segmentation and new word detection using conditional random fields. In: Proceedings of the 20th International Conference on Computational Linguistics, pp 562–568

Ron D, Singer Y, Tishby N (1996) The power of amnesia: Learning probabilistic automata with variable memory length. Machine Learning 25:117–149

Sassano M (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, pp 505–512

Settles B, Craven M, Friedland L (2008) Active learning with real annotation costs. In: NIPS Workshop on Cost-Sensitive Learning

Tomanek K, Hahn U (2009) Semi-supervised active learning for sequence labeling. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, pp 1039–1047

Tsuboi Y, Kashima H, Mori S, Oda H, Matsumoto Y (2008) Training conditional random fields using incomplete annotations. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp 897–904

Wang L, Li Q, Li N, Dong G, Yang Y (2008) Substructure similarity measurement in Chinese recipes. In: Proceedings of the 17th International Conference on World Wide Web, pp 978–988

Yamakata Y, Imahori S, Sugiyama Y, Mori S, Tanaka K (2013) Feature extraction and summarization of recipes using flow graph. In: Proceedings of the 5th International Conference on Social Informatics, LNCS 8238, pp 241–254

Yang F, Vozila P (2014) Semi-supervised Chinese word segmentation using partial-label learning with conditional random fields. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp 90–98