Pointwise Prediction and Sequence-based Reranking for Adaptable Part-of-Speech Tagging

Shinsuke Mori
Academic Center for Computing and Media Studies,
Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501, Japan
forest@i.kyoto-u.ac.jp

Yosuke Nakata[*]
NTT Communications
1-1-6 Uchisachicho, Chiyoda-ku, Tokyo, 100-8019, Japan
ruberukuraku@gmail.com

Graham Neubig
Nara Institute of Science and Technology,
8916-5 Takayamacho, Ikoma, Nara, 630-0192, Japan
neubig@is.naist.jp

Tetsuro Sasada
Academic Center for Computing and Media Studies,
Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501, Japan
sasada@ar.media.kyoto-u.ac.jp

***Abstract*—This paper proposes an accurate method for part-of-speech (POS) tagging that is highly domain-adaptable. The method is based on an assumption that the POS transition tendencies do not depend on domains, and has the following three characteristics: 1) it is trainable from partially annotated data, 2) it uses efficiently trainable pointwise POS taggers to allow for active learning, and 3) is more accurate than the pointwise or sequence-based POS taggers. The proposed method estimates POS tags by stacking pointwise and sequence-based predictors.**

In the experiments we deal with the joint problem of word segmentation and POS tagging in Japanese. We show that our proposed stacking process improves over pointwise and sequence-based methods (hidden Markov models and conditional random fields) both in the general domain and the target domain. In addition we show the learning curve in a domain adaptation scenario. The result shows that our method outperforms state-of-the-art methods in the same domain as the training data and is better than them in domain adaptation situations as well.

***Keywords*-Active learning; Reranking; Word segmentation; Part-of-speech tagging; Pointwise prediction;**

I. Introduction

Part-of-speech (POS) tagging [1], [2, and many others] is a fundamental step of natural language processing (NLP) in many languages, and many NLP applications use POS tagging results. Thus POS tagging accuracy has a great impact on these NLP applications. With large annotated corpora [3, *inter alia*] and methods based on machine learning techniques, the NLP community achieved a high accuracy around 97% or more in various languages. However, with the diversification of the domains to which NLP is applied, such as medical texts or texts in user generated contents (blog, twitter, etc.), we sometimes observe a severe degradation in POS tagging accuracy in text domains different from that of the training data . Therefore we can say that there is still a large demand for improving POS tagging accuracy, especially in domain adaptation situations.

In addition to the machine learning techniques, the NLP community is increasingly aware of importance of language resources, as the easiest way to improve an NLP based on a particular machine learning technique for a certain domain is to just add annotated texts in that domain to the training data. This strategy does require time and money, however, so we are interested in reducing annotation work by allowing annotators to focus on informative points that will provide a good performance/cost trade-off. Thus, there is a large amount of research on training NLP systems from partially annotated data or incomplete data, in which only some points are annotated with labels [4]. In this setting, some points lack the correct labels, and some may have multiple labels.

One of the major tasks to which these methods have been applied is word segmentation (WS) for languages without obvious word boundaries. A method for training conditional random fields (CRFs) from partially annotated data has been proposed and tested in Japanese WS [4]. This CRF extension is used to improve Chinese WS by referring to so-called natural annotations, such as partially segmented sentences converted from Wikipedia assuming that HTML tags are word boundaries [6]. A word segmenter based on a binary classifier [7] is another implementation trainable from partially annotated data. In this method the WS system decides whether there is a word boundary or not at each point between two characters without referring to the estimated labels surrounding the decision point. This is called a pointwise prediction method (or simply a pointwise method). Compared with sequence prediction methods like Markov models or CRFs, pointwise prediction requires less time to estimate model parameters even from partially annotated data, and is thus suitable for active learning, which

[*] This work was done when he was at Kyoto University.

is performed by alternating rounds of selecting uncertain points for annotation, performing annotation, and retraining the classifier [8].

Given this background, in this paper we propose a POS tagger equipped with following three characteristics:

1) It is trainable from partially annotated data.
2) Training is as fast as pointwise POS taggers to allow active learning.
3) It is more accurate than the pointwise and sequence-based POS taggers.

Our method performs POS estimation by stacking pointwise and sequence-based predictors, using pointwise prediction followed by reranking using sequence-based predictors [9]. The first module, pointwise POS estimation, is trainable from partially annotated sentences. The second module, sequence-based POS reranking, is efficiently trainable only from fully annotated sentences. Thus the training data of the second module is a subset of the first module. However, sequence-based predictors can use POS sequence information, and thus there may be room for improvement by referring to the combination of label candidates. We assume that these POS transition tendencies do not depend on the domain, and thus even if the sequence-based labeler is trained only on general domain data, it might be able to contribute to improve POS tagging accuracy even on texts in the adaptation target domain where only partially annotated data is available.

In the experiment we deal with the joint problem of WS and POS tagging in Japanese, traditionally called morphological analysis (MA).[1] We show that our stacking process improves over pointwise MA [8] and sequence-based MA [10] [11] [12] both in the general domain and in the target domain. In addition we show the learning curve in a domain adaptation scenario, which finds that the proposed method is as domain adaptable as purely pointwise approaches.

II. Joint Problem Solution Baseline

Our method, which we propose in this paper, uses MA by pointwise classifiers [8] as the first step. MA by pointwise classifiers (PW-MA) solves the problem step-by-step. First, it segments an input sentence into a word sequence. Then, it estimates the POS of each word like an English POS tagger. At each step PW-MA refers only to the input but not to any estimation results (or dynamic information) as features. In [8] linear support vector machines (SVMs) [13] are used because of their classification accuracy and speed. In this section we describe PW-MA in detail.

A. Word Segmentation by Pointwise Classification

The two-step approach [8] segments character sequence $\boldsymbol{x} = x_1 x_2 \cdots x_k$ into the word sequence $\boldsymbol{w}$. Word segmentation is formulated as a series of binary classification problems, estimating boundary tags $b_1,\ b_2,\ \ldots,\ b_{k-1}$. 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.

As features it uses 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 are as follows:

1) Character n-grams: substrings surrounding the decision point i. There are two parameters: the window width m and the length n. Features are all the substrings of the length up to n in the $2m$ long substring $x_{i-m+1}, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_{i+m}$. Figure 1 shows an example.
2) Character type n-grams: the same as the character n-grams but the characters in the substring are converted into the character type. The character types are Chinese character (K), *katakana* (k), *hiragana* (H), Roman alphabet (R), Arabic number (N), or other (O). Figure 1 shows an example.
3) Dictionary: three flags indicating that the word starting at i, ending at i, or containing i are included in the dictionary, and the length of that word.

As the above explanation indicates, PW-MA is trained from only the annotated points between two characters and it does not require any modification to estimate its parameters from partially annotated data. Thus it is both simple and fast enough to make active learning realistic.

B. POS Tagging by Pointwise Classification

POS tagging by pointwise classification performs one of the following four processes depending on the target word.

1) If the word appears as more than one POS in the training corpus, estimate the POS by a classifier,
2) If the word appears as only one POS in the training corpus, return its POS,
3) If the word does not appear in the training corpus but in the dictionary, return the POS of the first entry,
4) Otherwise, return noun.

In the first case, POS estimation is formulated as a multi-class classification problem, where we choose one tag t_j for each word w_j. The input is a word sequence but the classifier regards it as the target word and the character sequences preceding it ($\boldsymbol{x}_-$) and following it ($\boldsymbol{x}_+$). The POS of w_j is estimated from $\boldsymbol{x}_-$, w_j, and $\boldsymbol{x}_+$. When the window width is m', then the information referred to is $x_{-m'} \cdots x_{-2} x_{-1}, w_j, x_1 x_2 \cdots x_{m'}$. Putting it in another way, it only refers to the fact that there are word boundaries on both sides of w_j and that there is no word boundary inside w_j, and two character sequences $\boldsymbol{x}_-$ and $\boldsymbol{x}_+$.

The features for POS estimation are as follows (see Figure 2):

1) Word in focus,

[1] MA often also performs recovery of word base forms, but we do not handle this element in the present work.

(vaccinate a healthy child with this medicine)

x_{i-2} x_{i-1} x_i x_{i+1} x_{i+2} x_{i+3}

Text: 健 康 児 に 本 剤 を 接 種 し

↑

t_i: Decision point

Character (type) 1-gram: -3/児 (K), -2/に (H), -1/本 (K), 1/剤 (K), 2/を (K), 3/接 (K)
Character (type) 2-gram: -3/児に (KH), -2/に本 (HK), -1/本剤 (KK), 1/剤を (KH), 2/を接 (HK)
Character (type) 3-gram: -3/児に本 (KHK), -2/に本剤 (HKK), -1/本剤を (KKH), 1/剤を接 (KHK)

Figure 1. Features referred to in word segmentation (window width m = 3, n = 1, 2, 3).

(vaccinate a healthy child with this medicine)

x_{-3} x_{i-2} x_{-1} w x_{i+1} x_{i+2} x_{i+3}

Text: 健 康 児 に 本剤 を 接 種 し

↑

Word in focus

Character (type) 1-gram: -3/康 (K), -2/児 (K), -1/に (H), 1/を (H), 2/接 (K) 3/種 (K),
Character (type) 2-gram: -3/康児 (KK), -2/児に (KH), -1/にを (HH), 1/を接 (HK), 2/接種 (KK)
Character (type) 3-gram: -3/康児に (KKH), -2/児にを (KHH), -1/にを接 (HHK), 1/を接種 (HKK)

Figure 2. Features referred to in POS tagging (window width m' = 3, n = 1, 2, 3).

2) Character n-grams included in $\boldsymbol{x}_{-}\boldsymbol{x}_{+}$,
3) Character type n-grams included in $\boldsymbol{x}_{-}\boldsymbol{x}_{+}$.

Similar to PW-MA, POS tagging based on pointwise prediction is trained from only the words annotated with their POS and it does not require any modification to estimate its parameters from partially annotated data and is enough fast to make active learning realistic.

C. Flexible Language Resource Usage by Pointwise Prediction

WS or POS tagging based on the pointwise prediction allows us to use the following new types of language resources, making it possible to more efficiently adapt the tagger to new domains.

1) Partially annotated corpora: Only some points between two characters in a sentence are annotated with word boundary information or only some words are annotated with POSs. For MA a corpus annotated only with word boundaries is also a partial annotation corpus. Partially segmented or partially POS-annotated corpora also fall in this category.
2) Word dictionary: A list of words without POSs. This type of dictionary is often available in many domains. We can use this for automatic WS.

Of course the pointwise prediction can use a fully annotated corpus in which all the sentences are completely segmented into words and all the words are annotated with their POSs, and a list of words with POSs. These fully annotated corpora and dictionaries are sometimes difficult to prepare in a target domain, but partial annotations are relatively easy to prepare. Thus, MA based on the pointwise prediction makes it easier to adapt to new domains by making it possible to retrieve information even from these various language resources.

III. 2-Step POS Estimation

The PW-MA described in the previous section can not use the POS sequence information in the training corpus. This information may be, however, important for POS estimation. In this paper we assume that the domain dependency of POS transition tendencies is low and propose a new method for POS estimation based on this assumption. In this method, we use stacking to combine pointwise and sequence-based predictors, with the domain-specific pointwise predictor capturing domain knowledge, and the domain independent sequence-based predictor reranking the POS estimation result of PW-MA.

A. Overview of the Proposed Method

The proposed method combines the following three processes in a cascade:

1) word boundary estimation by a pointwise prediction,
2) POS-confidence pair estimation by pointwise prediction, and
3) POS reranking by sequence prediction.

Given an input sentence, first we segment it into a word sequence by word boundary estimation based on a pointwise prediction. This process is completely the same as the one described in Section II-A. Then we estimate a POS for each word in the word sequence. This process is similar to the one described in Section II-B, but we enumerate all the possible POSs with confidences. Finally we rerank the POS sequences based on sequence-based prediction referring to the confidences.

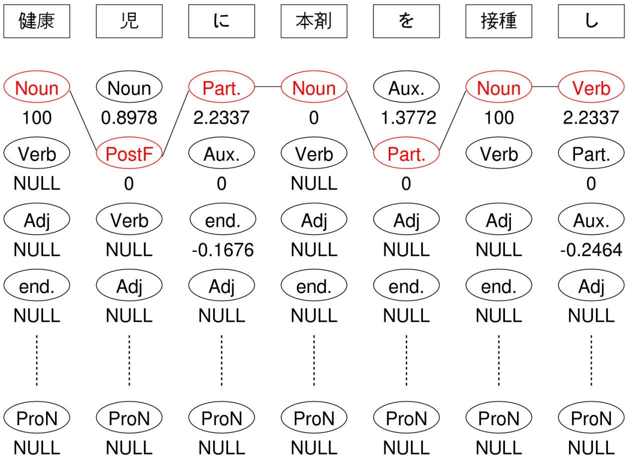

Figure 3. POS reranking by sequence labeling.

B. *POS Estimation with Confidence by the Pointwise Prediction*

The pointwise POS estimation described in Subsection II-B outputs only one POS for each word. In the proposed method, however, we calculate the confidences for all the possible POSs for a word and we use these confidences in the reranking process (see Figure 3).

The confidence of a POS for a word is defined as follows. First let d_r be the distance (margin) from the separation hyper-plane of the r-th ($r \geq 1$) POS candidate. And we define the confidence of the r-th POS candidate as $c_r = d_r - d_2$. As a result, the confidence of the first candidate is a positive value (in almost all cases $c_r \ll 100$ because of L2 regularization), that of the second candidate is 0, and those of the other candidates are negative values. If there is no POS candidate (the case 4 in Section II-B), this process returns a noun with confidence 0. And if there is only one POS candidate (the case 2 or 3 in Section II-B), this process returns that POS with confidence 100 (a special value). Figure 3 shows an example of POS candidates and their confidences.

C. *POS Reranking by a Sequential Prediction*

We have a word sequence and all the possible POSs with confidences as the output of the process above. Then we search for the best POS sequence among all the possible POS sequences by referring to the POS-confidence pair estimation result and the POS sequence statistics taken in the training corpus. Note that the word boundaries estimated by the pointwise word segmentation are not changed, because we do not rerank the word boundaries.

As a sequential prediction method we use CRFs [14], a standard method for sequence labeling problems because of their flexible feature design and high classification accuracy. The correct labels in the training data are the POS sequences in the full annotation corpus. The features are divided into two types: context features and confidence features, which we describe in detail in the subsequent section. In the prediction step the CRFs output the most likely POS sequence taking the output of the pointwise prediction results with confidence as the input. In the example shown in Figure 3, the CRFs output the POS sequence connected by the solid line, where the POS of the word “児” (child) has been changed into prefix from noun.

D. *Features*

As we mentioned, the CRFs for POS reranking refer to context features and confidence features. The confidence features are the followings calculated from the POS-confidence pairs output by the pointwise prediction.

Rule 1:
If the word has multiple POS candidates, the t-th feature ($1 \leq t \leq T$) is the confidence of the t-th POS.

Rule 2:
If the t-th POS is not a candidate, the $(T+t)$-th feature ($T+1 \leq T+t \leq 2T$) is set be 1.

Rule 3:
If the t-th POS is the only candidate, the $(2T+t)$-th feature ($2T+1 \leq 2T+t \leq 3T$) is set be 1.

When the condition of each rule is not satisfied, the feature value is set to be “NULL” (i.e. many features are NULL). The rationale of the rule 2 is to provide information about the POSs not in the candidate list. That of the rule 3 is to indicate POSs with high confidence according to the pointwise prediction, that may not need to be changed.

The other feature set is the context. We list them as follows:

1) word n-grams in the window width m'' including the word in focus at the center.
2) character type set n-grams of the words in the window width m'' including the word in focus at the center.

The character type set is a set of character types included in the spelling of a word. We set 6 character types, which are the same as those used in the word boundary prediction (Subsection II-A). Thus the character type set has $2^6 - 1$ combinations. The character type set n-grams are sequences of the character types for a word sequence.

E. *Training Data Creation*

As the training data of the CRFs for POS reranking, we need the correct POS tag sequence and those estimated by the pointwise prediction for a word sequence for feature creation. The estimated POS tag sequence should be similar to that given at the runtime. Thus the confidence estimation target has to be different from the training data of the pointwise prediction for the POS-confidence pair estimation. So we propose the following procedure similar to deleted interpolation [15].

<table>
<tr><th colspan="3">General domain (G)</th></tr>
<tr><td rowspan="2">Word boundary (W)</td><td>Full (F):</td><td>|文–化|交–流|使|事–業|を|</td></tr>
<tr><td>Partial (P):</td><td>␣文␣化|交–流|使␣事␣業␣を␣</td></tr>
<tr><td rowspan="2">Word boundary /POS (T)</td><td>Full (F):</td><td>|文–化/Noun|交–流/Noun|使/PostF|事–業/Noun|を/PP|</td></tr>
<tr><td>Partial (P):</td><td>␣文␣化|交–流/Noun|使␣事␣業␣を␣</td></tr>
<tr><th colspan="3">Target domain (A)</th></tr>
<tr><td rowspan="2">Word boundary (W)</td><td>Full (F):</td><td>|血|小–板|の|減–少|が|</td></tr>
<tr><td>Partial (P):</td><td>␣血|小–板|の␣減␣少␣が␣</td></tr>
<tr><td rowspan="2">Word boundary /POS (T)</td><td>Full (F):</td><td>|血/PreF|小–板/Noun|の/PP|減–少/Noun|が/PP|</td></tr>
<tr><td>Partial (P):</td><td>␣血|小–板/Noun|の␣減␣少␣が␣</td></tr>
</table>

Figure 6. Examples of various types of corpora.

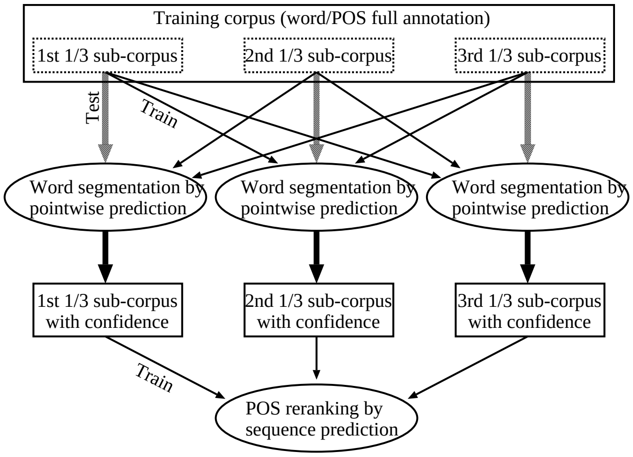
Training corpus (word/POS full annotation)
1st 1/3 sub-corpus
2nd 1/3 sub-corpus
3rd 1/3 sub-corpus
Test
Train
Word segmentation by pointwise prediction
Word segmentation by pointwise prediction
Word segmentation by pointwise prediction
1st 1/3 sub-corpus with confidence
2nd 1/3 sub-corpus with confidence
3rd 1/3 sub-corpus with confidence
Train
POS reranking by sequence prediction

Figure 4. Procedure for generating the training corpora for POS reranking by sequence labeling ($k = 3$).

1) Divide the training corpus C into k subsets $C_1,\ C_2,\ \ldots,\ C_k$.
2) For each $i \in \{1, 2, ..., k\}$
   a) Train the i-th pointwise MA from $k - 1$ subsets except for C_i
   b) Estimate POS-confidence pairs on C_i by the i-th pointwise MA with the model obtained by the step a)

Figure 4 illustrates the above procedure in the case of $k = 3$. The above procedure produces the subsets annotated with POS candidates and their confidences $C'_1,\ C'_2,\ \ldots,\ C'_k$. By adding the correct POS tag sequence in C to them, we have the training data of our CRFs for POS reranking.

F. Proposed Method and Language Resource

At the end of this section, we discuss the relationship between the proposed method and the corpus types. In the general domain, many fully annotated corpora (GTF in Figure 5), in which the sentences are divided into words completely and all the words are annotated with POSs, are available. Almost all annotated corpora produced through corpus annotation research [16], [17] fall in this category. The annotation work for the target (adaptation) domain corpus requires, however, domain knowledge in addition to the linguistic knowledge of the annotation standard. Thus the full annotation corpus in the target domain is costly. But a partial annotation corpus, in which some words are identified and some are annotated with POSs, is relatively easy to prepare. Figure 6 shows examples. In Figure 6, we use a notation called the extended 3-valued notation, which is our extension of the following 3-valued notation [18].

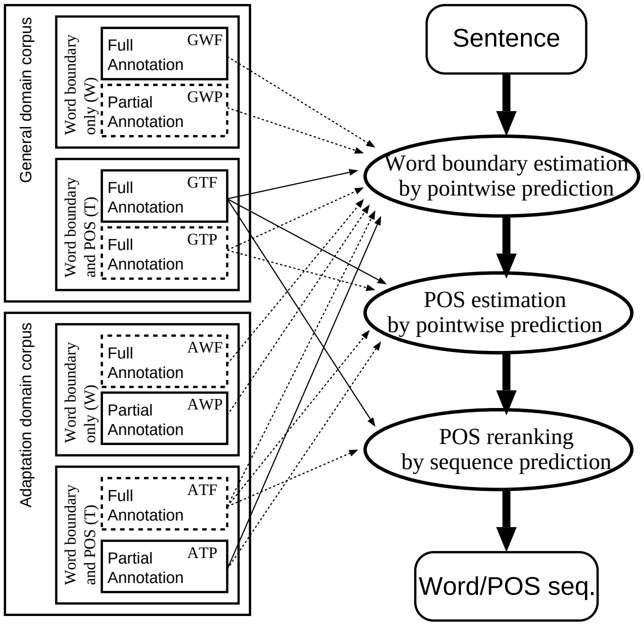

Figure 5. Relationship between the proposed method and various types of corpora. Theoretically we can use language resources connected by both dotted and solid lines for each process indicated by the ovals. Practically we use language resources connected by solid lines.

| : There is a word boundary.
– : There is no word boundary.
␣ : There is no information.

As an extension we add “/” to denote the POS of the word after it like |s–p–e–l–l–i–n–g/POS| if annotated. The

Table I
CORPUS SPECIFICATION.

Name	Source	Usage	#Sent.	#Words	#Char.
BCCWJ	White paper, Book, Newspaper	Training	27,338	782,584	1,131,317
	(General domain as GTF)	Test	3,038	87,458	126,154
	Yahoo!QA	Training	5,800	114,265	158,000
	(Target as ATF or ATP)	Test	645	13,018	17,980

following are the information that we can extract from various types of corpora.

- GWF, AWF: Word sequence and surrounding characters of word boundaries
- GWP, AWP: Surrounding characters of word boundaries
- GTF, ATF: Word sequence, surrounding characters of word boundaries, POS sequence, word-POS pair sequence, and surrounding characters of word-POS pairs
- GTP, ATP: Word sequence, surrounding characters of word boundaries, and surrounding characters of word-POS pairs

Theoretically speaking the WS based on a pointwise prediction can use any type of corpora containing one or more word boundaries with the characters surrounding them. So the pointwise WS can be trained from all types of corpora. The POS tagging based on a pointwise prediction can use any type of corpora containing one or more words annotated with their POSs with the characters surrounding them. In the above list, this applies to GWP, GWF, AWP, and AWF. POS tagging based on a sequence prediction can use corpora in which sentences are divided into a word sequence and the words are annotated with their POSs without any missing elements. In the above list, this applies to GTF and ATF.

In practical domain adaptation situations the available training data are a large GTF and AWP or ATP which are relatively easy to build. Full annotation corpora (AWF and ATF) are costly because it requires both linguistic and domain knowledge to build them. Figure 5 summarizes these remarks. In this figure the corpora connected by solid lines are usable by the processes listed on the right hand side: the WS or POS tagging based on pointwise prediction and the POS reranking based on sequence prediction.

IV. EVALUATION

In order to test the effectiveness of the proposed method, we conducted two experiments. One is a comparison on the general domain among existing methods and the proposed method. The other is a comparison among the major methods in a domain adaptation situation. We set the parameters n in n-gram to 2 and the window width m, m', and m'' to 5 in all cases based on the results of preliminary experiments. We divided the training corpus into 9 parts in the training data creation for the POS reranking (see Section III-E). For the sequence labeling we used CRFsuite [19].

A. Corpus

The corpus we used is the core part of Balanced Corpus of Contemporary Written Japanese (BCCWJ) [17] The sentences are divided into words and each word is annotated with a POS. We only used 21 coarse grained POS tags. The sources are white papers, books, newspapers, and Yahoo!QA. As [17] states, Yahoo!QA is different from the others. Thus we regard Yahoo!QA as the target domain and the others as the general domain. Table I shows the corpus specifications.

B. Evaluation Criterion

As an evaluation criterion we follow [10] 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:

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

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

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

C. Evaluation 1: Comparison with Existing Methods

First we compared our method with popular existing methods in the general domain. The methods are based on POS 2-grams model[2] [10], word-POS pair n-grams (n = 2,3) [11], CRFs (MeCab) [12], or pointwise prediction (KyTea) [8]. In this experiment, we assumed that only the full annotation corpus (GTF in Figure 5 and 6) is available to compare our method with existing ones trained from the same language resources.

To train the CRFs for reranking in the proposed method, we used the corpus generated from the general domain corpus produced by the procedure described in Subsection III-E. We tested the methods on the corpora in general domain and in the target domain. First we performed MA using pointwise prediction and then reranked the resulted POSs using sequence-based prediction.

[2] [10] reports POS 3-gram model but POS 3-gram model is less accurate than word-POS pair 3-gram model.

Table II
ACCURACIES OF WS AND THE JOINT PROBLEM ON THE GENERAL DOMAIN.

	Word boundary estimation			Joint		
Method	Precision [%]	Recall [%]	F-measure	Precision [%]	Recall [%]	F-measure
POS 2-gram model (HMM)	96.32	96.84	96.58	93.77	94.27	94.02
Pair 2-gram model	97.44	98.52	97.98	96.58	97.65	97.11
Pair 3-gram model	97.49	98.53	98.00	96.70	97.73	97.21
CRFs (MeCab)	97.19	98.30	97.74	96.72	97.84	97.28
Pointwise (KyTea)	98.73	98.71	98.72	98.07	98.06	98.06
Pointwise + Reranking	98.73	98.71	98.72	**98.38**	**98.37**	**98.38**

Table III
ACCURACIES OF WS AND THE JOINT PROBLEM ON THE TARGET DOMAIN (YAHOO!QA).

	Word boundary estimation			Joint		
Method	Precision [%]	Recall [%]	F-measure	Precision [%]	Recall [%]	F-measure
POS 2-gram model (HMM)	93.17	94.44	93.80	86.78	87.96	87.36
Pair 2-gram model	94.52	96.65	95.57	92.01	94.09	93.04
Pair 3-gram model	94.52	96.71	95.60	92.10	94.24	93.16
CRFs (MeCab)	94.89	96.87	95.87	93.69	95.65	94.66
Pointwise (KyTea)	96.93	97.26	97.09	95.19	95.51	95.35
Pointwise + reranking	96.93	97.26	97.09	**95.86**	**96.18**	**96.02**

Table II and III show the accuracies in the general domain and the target domain respectively. In these tables, "pointwise" means the results of "pointwise prediction," the second oval from the top in Figure 5. "pointwise + reranking" means the results of the POS reranking by the proposed method, that is the third oval in Figure 5. Since we do not rerank the WS results, word boundary estimation accuracies of these two methods are the same. From the tables we can say that the proposed method improves the joint problem accuracy both in the general domain and the target domain. The improvement is larger in the target domain. From these results, our assumption that the POS transition tendencies does not depend on the domain (see Section III) is plausible and we can improve PW-MA based on this assumption without losing the flexibility in choosing language resources. From the above observations, we can say that the proposed method is effective.

D. Evaluation 2: Adaptation Case

Second, we evaluate our method in a domain adaptation scenario. The existing method that is the most flexible in this scenario is pointwise MA, as it is trainable from partial annotations. In the experiment, we emulated active learning by adding partially annotated sentences. Along with the proposed method we tested pointwise MA and sequence-based MA. We started with the training corpus in the general domain and added partially annotated sentences gradually.

The concrete procedure is as follows (see Figure 7).

1) Train the pointwise MA from the training corpus in the general domain (GTF in Figure 5 and 6),
2) Estimate confidences of the training corpus in the target domain by the above obtained model without referring to the correct tags.

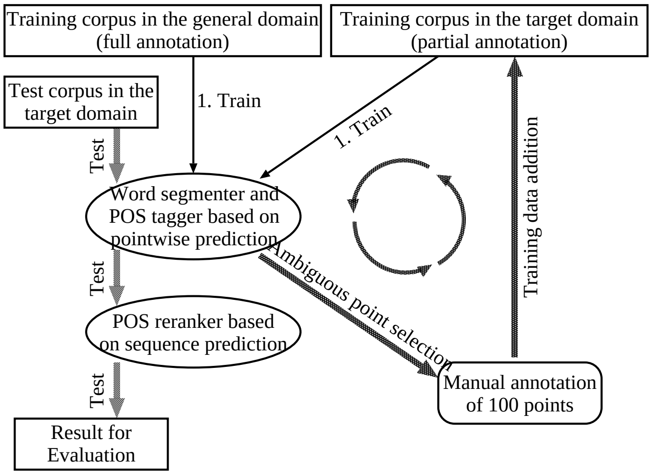

Figure 7. Domain adaptation scheme based on active learning using partial annotation.

3) Annotate 100 points of low confidence in the corpus in the target domain with word boundary or POS producing a partial annotation corpus in the target domain (ATP in Figure 5 and 6), and
4) Add the above partial annotation corpus to training corpus and train the model again and go to 2).

We repeated this procedure for 200 iterations. Each time we measured the accuracies on the target domain. The baselines are the pointwise MA (pointwise:part) trained from the same corpus as the proposed method (pointwise+CRFsuite:part) and the CRFs with new words appearing in the partially annotation corpus added to the dictionary (CRF:part).

The result is shown in Figure 8. From this graph we see that the proposed method outperforms the pointwise

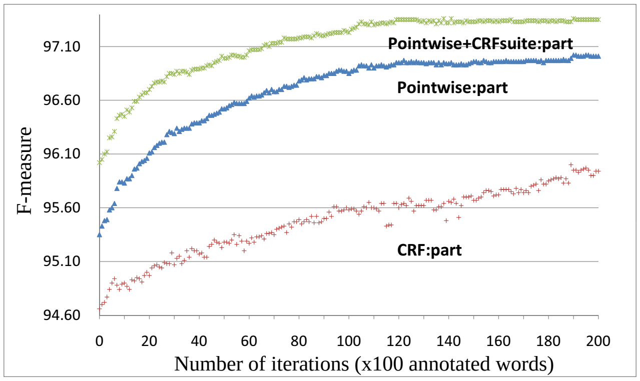

Figure 8. Learning curve in the case of domain adaptation.

MA consistently. In addition the proposed method improves the accuracy in the domain adaptation case. Putting it in other words, the proposed method successfully increased the accuracy without losing the domain adaptability of pointwise MA. Therefore we can say that the proposed method is superior to existing ones in this case as well.

V. Conclusion

In this paper we have proposed a POS tagging method allowing flexible usage of language resources. The method is based on pointwise prediction and reranking by sequence-based prediction combined in the cascaded manner. The experimental results showed that the accuracy in the resource-rich domain is higher than existing methods. In a domain adaptation scenario where we add partially annotated corpora, the proposed method outperformed the existing pointwise method constantly. These results showed that the proposed method is capable of providing high domain adaptability while keeping high accuracy in the general domain.

Interesting research directions include testing POS tagging in other languages and the application of our reranking technique in various sequence labeling problems in NLP or other fields.

Acknowledgments

This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 26280084, 26540190, 24240030, and 26280039.

References

[1] S. J. DeRose, “Grammatical category disambiguation by statistical optimization,” *Computational Linguistics*, vol. 14, no. 1, pp. 31–39, 1988.

[2] B. Merialdo, “Tagging English text with a probabilistic model,” *Computational Linguistics*, vol. 20, no. 2, pp. 155–171, 1994.

[3] M. P. Marcus and B. Santorini, “Building a large annotated corpus of English: The Penn treebank,” *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[4] Y. Tsuboi, H. Kashima, S. Mori, H. Oda, and Y. Matsumoto, “Training conditional random fields using incomplete annotations,” in *Proceedings of the 22nd International Conference on Computational Linguistics*, 2008.

[5] F. Pereira and Y. Schabes, “Inside-outside reestimation from partially bracketed corpora,” in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 1992, pp. 128–135.

[6] Y. Liu, Y. Zhang, W. Che, T. Liu, and F. Wu, “Domain adaptation for CRF-based Chinese word segmentation using free annotations,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 864–874.

[7] G. Neubig and S. Mori, “Word-based partial annotation for efficient corpus construction,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010.

[8] G. Neubig, Y. Nakata, and S. Mori, “Pointwise prediction for robust, adaptable Japanese morphological analysis,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 529–533.

[9] E. Charniak and M. Johnson, “Coarse-to-fine n-best parsing and maxent discriminative reranking,” in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005, pp. 173–180.

[10] M. Nagata, “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*, 1994, pp. 201–207.

[11] S. Mori and G. Kurata, “Class-based variable memory length Markov model,” in *Proceedings of the InterSpeech*, 2005, pp. 13–16.

[12] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to Japanese morphological analysis,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 230–237.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[14] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of the Eighteenth ICML*, 2001.

[15] F. Jelinek, “Self-organized language modeling for speech recognition,” IBM T. J. Watson Research Center, Tech. Rep., 1985.

[16] *EDR Electronic Dictionary Technical Guide*, Japan Electronic Dictionary Research Institute, Ltd., 1993.

[17] K. Maekawa, M. Yamazaki, T. Maruyama, M. Yamaguchi, H. Ogura, W. Kashino, T. Ogiso, H. Koiso, and Y. Den, “Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010.

[18] S. Mori and H. Oda, “Automatic word segmentation using three types of dictionaries,” in *Proceedings of the Eighth International Conference Pacific Association for Computational Linguistics*, 2009.

[19] N. Okazaki, “Crfsuite: a fast implementation of conditional random fields,” 2007. [Online]. Available: http://www.chokkan.org/software/crfsuite/