The Semi-Automatic Construction of Part-Of-Speech Taggers for Specific Languages by Statistical Methods

Tomohiro YAMASAKI **Hiromi WAKAKI** **Masaru SUZUKI**
Toshiba Corp., Corporate Research & Development Center, Knowledge Media Laboratory
1. Komukai Toshiba-cho, Saiwai-ku, Kawasaki, JAPAN
{tomohiro2.yamasaki, hiromi.wakaki, masaru1.suzuki}@toshiba.co.jp

Abstract

Economic activities now keep being globalized more and more. Thus we are driven to deal with not only the documents written in English but also those written in other languages. In order to enable us to develop processors of any language quickly, we have been making a framework based on statistical processing and machine learning. At present, we confirmed that part-of-speech (POS) taggers of some target languages can be built by using this framework and the information of source languages. In this paper, we describe the method of acquiring POS lexicons and that of generating supervisors of POS sequences, which are used to learn grammatical models of target languages. We also explain the experimental results of building POS taggers of Portuguese and Indonesian by using some source languages.

1 Introduction

The natural language processing, for example, part-of-speech (POS) tagging, syntactic parsing, and named entity extraction, is the fundamental technology for information extraction from text documents. This means that the preparation of processors of a specific language enables us to develop various applications for that language such as keyword extraction, document classification, and machine translation. However, most parts of the processors we have already built are dependent on the characteristics of each language since we have developed lexicons and grammars manually according to those of target languages such as Japanese and English. This means that we have to spend much time and effort when we try to prepare processors of a new language in the similar way before.

On the other hand, economic activities keep being globalized and thus we should provide people all over the world with appropriate services and products. In particular, the following needs are increasing:

- to estimate customers' concerns and intentions in order to provide the best service,
- to grasp customers' reputations and complaints in order to avoid troubles,
- and to analyze the documents written in local languages in order to achieve two above-mentioned statements.

We have mainly worked on processing of English until now, since many people tend to consider to be international as to use English much. After now, however, we must work on not only English but also other languages all over the world in order to be *truly* international.

Therefore, we have been working on the establishment of the framework that enables us to develop processors of any language quickly. Concretely speaking, we aim to build lexicons and grammatical models semi-automatically by using statistical processing. We also aim to achieve processors for POS tagging and more advanced language processing by using only the combination of surface and statistical information of documents given. However, we make it a condition that the documents written in target languages have many translations with source languages because it is difficult to build processors without any clue at all.

Roughly speaking, the technical points of our research are divided into the development of lexicons and that of grammatical models. In

Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, , pages 23–29,
Chiang Mai, Thailand, November 8, 2011.

this paper, we choose POS taggers as an example of processors and describe the method of the following processes:

- to acquire POS lexicons that are composed of [word, POS] pairs,
- to generate supervisors of POS sequences,
- and to learn grammatical models by using the above-mentioned lexicons and supervisors.

As a result of these processes, we can obtain the POS tagger of the target language semi-automatically. Finally, we do the experiment of building POS taggers by using some source languages and evaluate the accuracy of those taggers.

1.1 Related Work

Recently, it has been found that various problems of tasks in the natural language processing can often be solved easily by machine learning if we can prepare a large amount of tagged corpora. However, it is a large problem to prepare tagged corpora that can be used as supervisors of each task.

On the other hand, it is easy to obtain raw corpora from the Internet and so on. Therefore, there are some studies about the methods for building processors by using not tagged corpora but only raw ones. (Goldsmith, 2001) acquires the inflections of each word on the basis of Minimum Description Length (MDL) model. However, in order to use the method of (Goldsmith, 2001), we first have to generate probabilistic grammars manually, because this method is to distinguish the ones acceptable and the ones not acceptable. This means that we have to know the characteristics of the target language well to some degree, and that it is difficult to build processors of the language we hardly know by this method.

In addition, semi-supervised learning is receiving much attention as the method for solving the problem of preparing a large amount of tagged corpora in these days. This is a method aiming to obtain the same effect as the case where we prepare a large amount of tagged corpora by giving only a small amount of tagged data to a large amount of raw corpora. (Niu et al., 2003) learns the extraction rules from the seed words given first, generates the corpora of named entities by those rules, and finally builds a named entity extractor. As to semi-supervised learning, however, it is known that if tagged data include errors even a little, errors increase rapidly in the phase of automatic generation of supervisors and thus it is difficult to achieve enough accuracy. It is also difficult to give data with accurate tags when we hardly know the target language. Therefore, we have to do trial and error so as not to cause the error propagation.

1.2 Policy

When we use translations with some specific languages, the degree of difficulty of obtaining them has a big influence on us. Generally speaking, major news websites often deliver not only articles written in local languages but also those written in English. In other words, there is a large probability that the documents written in local languages have the English translations, which we can use as parallel corpora. However, we note that even if we can obtain the translations with languages X and Y, the sentences within the translations do not always have one-to-one relations. Generally speaking, it is difficult to associate the sentences of language X with the sentences of language Y with high accuracy when we hardly know the relations of words of both languages. Much less, it is almost impossible when we hardly know the target languages.

Therefore, we decided to use the translations of the Bible as our experimental corpora. The Bible is one of the most familiar documents that are read all over the world and the translations with many languages are open to the public on the Internet ((The Unbound Bible,)). In addition, the number of chapters and sections are the same in any language though each translation of the Bible is partitioned into many chapters and sections. This means that the sentences have almost one-to-one relations because each section has few sentences.

On the other hand, as we described above, we aim to achieve processors for advanced language processing by using only the combination of surface and statistical information of documents given. As the first approach, we decided not to target the languages as follows:

- the languages whose character system has not been digitalized yet,

- the languages whose words are not written with a space between them,
- and the languages whose orthographies do not distinguish common nouns and proper nouns.

Not only the languages that have very few users but also some of those that are used in India are known that their character systems have not been digitalized yet. We cannot disregard those Indian languages because they have many users, but we cannot perform the computer statistics if there is no digitalized corpora. Next, Thai, Cambodian, and Laotian languages are known that their words are not written with a space between them. These languages, similar to Japanese, have a large problem that it is very difficult for computers to divide a sentence into words. Then, Arabic, Hebrew, and Hindi languages have no case sensitivity. These languages, similar to German whose nouns always start with capital letters, have difficulties to extract the relations of words of other languages because it is not easy to determine proper nouns.

For these reasons, we mainly target the languages that use Latin characters. Particularly in this paper, we consider Portuguese and Indonesian as major targets. However, our method can be applied also to other languages like French and Italian.

2 Extracting the relations of words

Our method for acquiring POS lexicons is composed of two processes. One is a process of extracting useful words by using statistics of only one language. The other is a process of extracting the relations of words of two languages by using statistics of both languages. In this section, we describe both processes.

2.1 Extracting useful words on the basis of statistical information of a single language

Here, we describe the process of extracting the words whose surfaces are similar to one another (say sim-set), proper nouns, and word collocations on the basis of statistical information of a single language. The purpose of extracting sim-sets is to presume the inflections/derivations of each word at the next process.

As we described in Section 1.2, we consider Portuguese and Indonesian as major targets. This means that the words that always start with capital letters must be proper nouns, though we have to take into account the words that appear at the beginning of sentences. Therefore, we partition all sentences with spaces and symbols into words and extract each word w that satisfies the following conditions from them:

- $c_{small}(w)$, which is the count that w has only small letters, is equal to 0.
- $c_{capital}(w)$, which is the count that w starts with capital letters, is greater than or equal to 5.

The probability that a word that is not a proper noun satisfies the condition $c_{small}(w) = 0$ and $c_{capital}(w) \geq 5$, is less than $(1/2)^5 = 1/32$ even if we assume that the probability that it appears at the beginning of sentences is $1/2$. It follows that we can decide whether a word is a proper noun with significance level of 5%.

Next, C-value (Frantzi and Ananiadou, 1996) is known well as a method for extracting word collocations from the text documents. This method calculates the connectivity between the words, defined as $C - value(\mathbf{w}) = (l-1)(n - t/c)$, where $\mathbf{w}$ is a word collocation $w_1 \ldots w_l$, t and c are the total count and the distinct count of word collocations that include $\mathbf{w}$ and that are longer than $\mathbf{w}$.

When the connectivity between some words is strong, these words often appear composing a group and C-value tends to be large because t tend to be small in comparison with n. However, when the word collocations is short, C-value tends to be unreasonably large because c tends to be very large in comparison with n. Therefore, we use not only C-value but also C'-value (Yamasaki, 2008) in order to extract word collocations. In other words, we extract the word collocations whose C-value and C'-value are larger than a threshold given.

Here, Portuguese is classified into the inflectional language grammatically as well as other European languages. The inflectional languages have the property that the elements of grammatical functions are embedded in each word and thus each word changes its form according to the case, the gender, and the number. This means that we must have the means

Table 1: Example of french words extracted from the French Bible

Proper nouns	Word collocations	Sim-sets
Jubal	en paix	{répara,réparer,réparé,réparât,
Assyrie	le livre	réparent,réparèrent},
Jébusien	car vous	{sanctifie,sanctifie-la,sanctifier,sanctifié,
Guérar	nos pères	sanctifieras,sanctifiée,anctifiez-vous,
Nimrod	l'autel	sanctifierai,sanctifierez,sanctification,
Calakh	de guerre	sanctifiés,ssanctifièrent,sanctifiez-le,
Gaza	sa femme	sanctifiez,sanctifiaient,sanctifient,
Dikla	d'Égypte	sanctifiait,sanctifieront,sanctifiât}

by which we can determine inflection forms of each word. Indonesian is classified into the agglutinative language as well as Japanese. The agglutinative languages have the property that most words are formed with the joint of the elements of grammatical functions. This means that we must have the means by which we can determine the stem of derivation words.

In most languages, it is known that the beginning or the end of each word change its form, though the middle does in Arabic and Hebrew. Therefore, we formally define a sim-set as the words whose common affix is longer than a threshold given. Now, we partition all sentences with spaces and symbols into words and perform the following process for each pair of words (w_1, w_2):

- let L, l be max, min of $(|w_1|, |w_2|)$, respectively.
- define $w_1 \sim w_2$ if and only if $l \geq L/2$ and the length of common prefix $pre(w_1, w_2) \geq L/2$ or the length of common suffix $suf(w_1, w_2) \geq L/2$.
- partition all words into equivalence class based on $\sim^*$, which is defined as the reflexive transitive closure of $\sim$.

We note that the definition of $\sim^*$ does not depend on the definition of $\sim$. This means that if we define $\sim$ by using common subsequence instead of common affix, we may apply the same method to the languages where the middle of each word changes.

2.2 Extracting the relations of words on the basis of statistical information of two languages

Here, we describe the process of extracting the relations of words of two languages on the basis of statistical information of both languages.

We expect that when a word w^x of language X corresponds to a word w^y of language Y, the positions of w^x in corpora are related to those of w^y. Here, we note that it is not easy to decide whether the positions have any relations because the sentences within the translations do not always have one-to-one relations. However, it is easy to do it when we use the translations of the Bible because the sentences are almost parallel. Assume that an X–Y parallel corpus has n corresponding sentences and that the numbers of sentences where w^x and w^y appear are shown in Table 2. For example, both appear in a sentences, only w^x (w^y) in b (c), and neither in d.

For such a table, it is known that χ^2-value, defined as $\chi^2 = n(ad - bc)^2/efgh$, follows a χ^2 distribution. On the basis of this value, we can decide whether the words correspond to each other. In addition, we can also decide the relations of 2-grams and those of word collocations in the same way, because this test uses only the number of sentences and does not depend on the characteristics of languages and the length of each sentence. On the other hand, because this test does not use the information where the word appears in a sentence, we sometimes obtain two or more words that correspond to a word given. This does not matter so much if we can finally acquire POS lexicons composed of [word, POS] pairs. However, in order to extract one-to-one relations in any case, we make it a condition that we select the most similar one in the similarity of surfaces. This is because a proper noun is probably pronounced similarly in any language. In that sense, it is more general to calculate the similarity after we convert the surface into the pronunciation.

Now, we have described the method of extracting words and their relations by using not language dependent information but sta-

Table 2: The number of sentences where w^x and w^y appear

	w^y appears	w^y does not appear	sum
w^x appears	a	b	$e = a + b$
w^x does not appear	c	d	$f = c + d$
sum	$g = a + c$	$h = b + d$	$n = a + b + c + d$

Table 3: List of part-of-speeches

A	ADJECTIVE	P	PRONOUN
C	CONJUNCTION	R	ADVERB
D	DETERMINER	S	PREPOSITION
I	INTERJECTION	V	VERB
M	NUMERAL	0	DIGIT
N	NOUN	_	SYMBOL

tistical information. From here, on the assumption that we know language X well (= we have a POS tagger of language X), we describe the method of extracting the inflections/derivations of words of language Y we hardly know.

As we described in the previous section, a sim-set includes candidates of inflection/derivation forms of a word. Because we have a POS tagger of language X, we can decide whether some different words are in truth the same by restoring each word to its standard form. In other words, we can extract inflection/derivation forms of language Y that correspond to a standard form of language X by finding the subset that is contained in a sim-set of language Y and is the most relevant to the standard form of language X. Therefore, we perform the following processes:

- choose a standard form of language X $\overline{w}^x$ and a sim-set of language Y $sim^y = \{w^y_1, w^y_2, \ldots\}$.
- calculate χ^2-value for each subset $\overline{sim^y}$, which is contained in sim^y.
- find the subset whose χ^2-value is maximum.

3 Acquiring POS lexicons and generating supervisors of POS sequences

In the previous section, we explained the method of extracting the relations of words of languages X and Y on the basis of statistical information obtained from X–Y parallel corpora. In order to acquire POS lexicons of language Y finally, it is necessary to estimate the POS of each word w^y of language Y. Because we can know the POS of each word w^x of language X on the assumption that we have a POS tagger of language X, we consider the POS of w^x corresponding to w^y as that of w^y.

Here, we note that we may not be able to decide the unique POS of w^x. For example, it is known that many English words are used as a NOUN and a VERB. In other words, most of English words have two or more POSes. While the English word "name" can be used as a NOUN and a VERB, the Portuguese word "nome" is used as a NOUN only. Therefore, from the viewpoint of the relevance ratio, it is thought to be better that we estimate POSes on the basis of the context. However, in order to make our method simple, we consider all possible POSes of w^x as those of w^y.

It is known well that most of European languages belong to Indo-European languages and there are few differences in the fundamental grammars between them. Conversely speaking, this means that the difference of languages does not affect so much the POS sequences of the corresponding sentences. Though Indonesian does not belong to Indo-European languages, we generate the supervisors of POS sequences of language Y on the basis of POS sequences of language X by solving the Minimum Cost Matching Problem that has the following conditions:

- the POSes of D, P, S, 0 and _ can match the same POSes only, which is because these POSes are thought to be the same POSes for other languages,
- the skip cost is c_{skip},
- the match cost is 0 if $cand(w^y) = \emptyset$ or $pos(w^x) \in cand(w^y)$, otherwise c_{diff},

where $pos(w^x)$ is the POS of a word w^x of language X and $cand(w^y)$ is the POS candidates of a word w^y of language Y.

For example, Figure 3 shows that the French word "commencement" matches the English word " beginning" and thus is estimated to

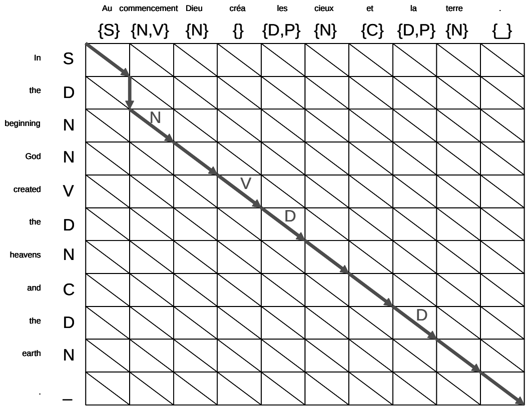

Figure 1: A solution of Minimum Cost Matching Problem solved by Dynamic Programing

be a NOUN. It also shows that "créa" matches "created" and thus is estimated to be a VERB. In order to make our method simple, we do not use the relations of words this time. However, we may make the condition that the match cost reflects the relations of words.

4 Experimental results

We have already built the POS taggers of English, Spanish and Esperanto manually. In this section, we explain the experimental results of building POS taggers of some target languages semi-automatically on the assumption that English, Spanish and Esperanto are used as the source languages. While there are some versions of the Bible by different translators in some languages, we used the following versions shown in Table 4 on this experiment.

First, we show the covering ratios in Figure 2. The total and distinct covering ratios are defined as the ratios of total and distinct words with one or more estimated POSes by using our method, respectively. Though there are a few differences, as you can see, the covering ratios in Figure 2 are almost the same degree even if the source language is English, Spanish or Esperanto.

This means that our method is stable and is independent of the characteristics of source languages. In addition, we confirmed that we acquired the POSes to almost all words by using statistical processing because the total covering ratio exceeds 0.8. However, the distinct covering ratio of Indonesian is about 0.25 and is lower than expected. There is still room for improvement.

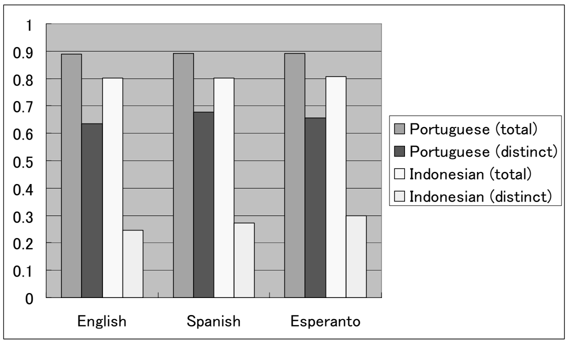

Figure 2: Total and distinct covering ratios

Next, we generated the supervisors of POS sequences based on the above-mentioned POS lexicons and performed the machine learning of grammatical models by using CRF (Laffert, 2001). After that, we obtained the POS taggers of the target languages semi-automatically. We show the accuracy ratio in Figure 3. The accuracy ratio is defined as the ratio of correct POSes that the taggers tagged onto words of sentences given. As you can see, POS information is not attached to the Bible. In order to evaluate the accuracy ratio, we extracted about 60 sentences (about 900 words) from the Bible and made the POS answers manually. Figure 3 shows that the Portuguese tagger achieved high accuracy of about 0.9 even though they are built semi-automatically. Figure 3 also shows that the accuracy of the Indonesian tagger is about 0.6. This is probably because the differences between Indonesian and source languages are large.

On the other hand, we analyzed failure cases and confirmed that one of the causes of incorrect POSes that the taggers tagged is to reflect grammatical features of source languages. For one example, the word "there" in English is ADVERB but is often expletive. For this rea-

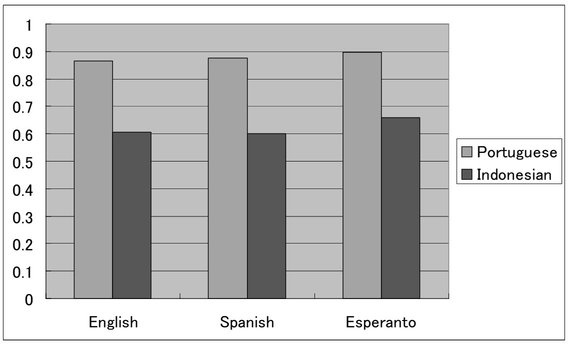

Figure 3: The accuracy ratios of POS taggers

Table 4: List of languages and versions of the Bible

Language	Version	Sections	Total words	Distinct words
English	American Standard	31103	918287	13256
Spanish	Reina-Valera	31103	824760	28874
Esperanto	British and Foreign Bible Society	31103	796700	30760
Portuguese	João Ferreira de Almeida	31103	828352	29306
Indonesian	Bahasa Indonesia Sehari-hari	31103	765810	47947

son, our taggers sometimes predicted by mistake some words as ADVERB, though those words should be NOUN in Portuguese and Indonesian. For another example, ADJECTIVE comes ahead of NOUN in English although ADJECTIVE comes behind NOUN in Portuguese and Indonesian. For this reason, at the sequences of words with the possibility of being ADJECTIVE and NOUN, our taggers sometimes predicted the previous word as ADJECTIVE as if the English tagger does.

Well, as you can easily see, many words that do not appear in the Bible appear in modern documents. This brings us a worry that the accuracy ratio might drop in proportion to the drop of the covering ratios, because as to the words that do not appear in the POS lexicons, our taggers must predict POSes from only peripheral words. Therefore, it will be important to develop the method of extracting modern words and estimating their POSes from large corpora such as Wikipedia documents, for example, by using grammatical knowledge of target languages given by hand at the minimum.

5 Conclusion

In this paper, we described our method that is composed of two following processes. One is the process of acquiring POS lexicons that are composed of [word, POS] pairs by using parallel corpora of source languages and target languages. The other is the process of generating supervisors that are used for machine learning of grammatical models. And we confirmed that Portuguese and Indonesian POS taggers are built semi-automatically by using the Bible as parallel corpora and by using English, Spanish and Esperanto as the source languages. In addition, we confirmed that the Portuguese tagger achieved high accuracy of about 0.9 while the accuracy of the Indonesian tagger is about 0.6.

Although we did not target the languages that use Cyrillic characters and Greek characters in this paper, we have a mind to expand the coverage of our method to such languages as Russian, Ukrainian and Greek in the future. On the other hand, a method (Mochihashi et al., 2009) has attracted a great deal of attention from many researchers in these years. This method partitions each sentence into words by using only statistical information of the documents given. We will work on word segmentation and will expand the coverage of our method to the languages which are not written with a space between words.

References

Biora University. 2005. *The Unbound Bible.* http://unbound.biola.edu/.

C. Niu, W. Li, J. Ding, R.K. Srihari. 2003. *A bootstrapping approach to named entity classification using successive learners.* Proc. of 41st Annual Meeting of ACL, pp.335–342, July 2003.

J. Goldsmith. 2001. *Unsupervised learning of the morphology of a natural language.* Journal of Computational Linguistics (2001), vol.27, no.2, pp.153–198.

K.T. Frantzi and S. Ananiadou. 1996. *Extracting nested collocations.* COLING-96, pp.41–46.

T. Yamasaki. 2008. *Topic extraction from Electronic Program Guides by using decomposition of the co-occurrence graph into strongly connected components.* Journal of Computational Linguistics (2001), vol.27, no.2, pp.153–198.

J. Laffert. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* Proc. of Machine Learning (2001), pp.282–289.

D. Mochihashi, T. Yamada, N. Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol.1, pp.100-108.