A MULTI-PURPOSE JAPANESE SPOKEN LANGUAGE CORPUS

Shiho Ogino, Shinsuke Mori and Nobuyasu Itoh

Tokyo Research Laboratory, IBM Research
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan
{shiho, e29176, iton}@jp.ibm.com

ABSTRACT

Corpus data is essential to the improvement of natural language processing by statistical methods. In Japan, some on-line text corpora which are annotated with word boundary, part-of-speech and some other syntactic information have been created, however, most of them are based on written Japanese. The corpora of spoken Japanese previously described are dialog-based, and their domains are restricted to such limited areas as reservations for travelers. Written Japanese corpora and spoken Japanese corpora are usually created by different people and for different reasons, and therefore an annotation system which can be applied to annotating both written Japanese and spoken Japanese is not defined yet.

We are creating a Japanese tagged corpus of spoken Japanese with the aim of making a multi-purpose corpus that can be used for developing various systems. such as transcription systems, machine translation systems, and so on. When we started to make a Japanese written language corpus, we carefully defined the corpus format and annotation rules so that they could be applied to annotating a Japanese spoken language corpus with minimal expansion. Now we have created a spoken language corpus and have begun to annotate it with syntactic information such as word boundaries, parts-of-speech, and word-to-word dependency relationships, applying the same annotation rules used in annotating our written language corpus. We focused on a collection of spoken Japanese that is used in presentations and lectures (monologues) and selected classes broadcast by the University of the Air.

We parsed both our written language corpus and spoken language corpus by using a parser based on a stochastic language model. We found there is no critical difference in the accuracy of parsing the written language corpus versus the spoken language corpus. This result shows our corpus annotation method is applicable for both a written and a spoken language corpora.

The authors are grateful to The Mainichi Newspaper Co. Ltd (CD-ROM Mainichi 91-95), Nippon Keizai Shinbun, Inc., and the University of the Air.

1. INTRODUCTION

Corpus data is essential for use in stochastic processing of natural language. An on-line corpus annotated at least with word boundary information is necessary for word based and morphological based statistical processing of Japanese.

Recently, some organizations have provided Japanese tagged corpora that are morphologically parsed and annotated with some syntactic information such as dependency relationships [1, 2, 3]. However, most of them are written language corpora. In most languages, spontaneous speech is substantially different from speech read from written sources. Research for English by using the Switchboard corpus [4] shows clear statistical differences. We believe a spontaneous-speech corpus is important for any natural language processing of Japanese such as recognition, machine translation [5], or information retrieval using speech communication [6].

There has been some research reported on collecting spontaneous Japanese speech corpora [7, 8, 9]. However, since their foci are dialogues on limited topics such as hotel reservations and queries on travel information, the sizes and topical range of the corpora collected from Japanese spontaneous speech are unsatisfactory. In addition, these corpora were created without consideration of tags suitable for a written Japanese corpus, it is not clear that the annotating strategies of these corpora are applicable to annotating a written language corpus. That is, these annotating strategies may not be directly applicable to written language corpora and multi-purpose corpora.

We think the requirements for a Japanese spoken-language corpus are as follows:

- Not limited to specific topics
- The annotating rules should not be different from

those of written-text corpora, but an extension of them

- General rules which can be applied to many languages are preferred over uniquely Japanese rules

We therefore selected 78 subjects from broadcast lectures (The University of the Air) and transcribed 148 classes, which include about 83 hours of speech. Filled-pauses and word fragments are written phonologically with tags [10].

When we started to make a Japanese written language corpus [13], we carefully defined the corpus format and annotation rules so they could be applied to annotating a Japanese spoken language corpus with minimal expansion. This time, in annotating the spoken language corpus, we applied the same annotation rules used in annotating the written text corpus. This paper describes our efforts in annotating the spoken language corpus with the aim of building a multi-purpose annotated corpus.

In Section 2, we describe our corpus format and the annotation rules. We report on our effort to annotate the transcribed text of the University of the Air in Section 3. The results of our experiment in which we parsed both our written language corpus and a spoken language corpus and compared the accuracy is shown in Section 4 in order to examine whether or not our annotating strategy is appropriate for building a multi-purpose annotated corpus.

2. CORPUS FORMAT AND ANNOTATING STRATEGY

2.1. Corpus format

In annotating a multi-purpose corpus, the information to be given to the corpus tends to be detailed because each system that uses the corpus needs different information in its processing. However, annotation that is too detailed increases the work load of the corpus annotators, and tends to lead to a domain-dependent or application-dependent annotated corpus. For our corpus, we decided to use so simple information that is independent of a particular domain or application.

In most Japanese corpora post-positional phrases (called "bunsetsu"), which consist of a content word and zero or more subsequent function words, are used as a syntactic unit and showing dependencies between them is a common annotation. However, the definition of the phrases is sometimes ambiguous and inconsistent between annotators. This is why we did not attempt to define dependencies between the phrases, but rather dependencies between words.

attribute	tag	attribute value
sentence head	-	*BOS* BOS
sentence end	-	*EOS* EOS
sentence ID	ID	1, 2, ..., n
part-of-speech	pos	noun, proper noun, verb, adjective, adjective noun, auxiliary, adverb, conjunction, digit, adnominal,prefix, suffix, special characters, particle, inflection, interjection, unknown word, others
modifiee	mod	1, 2, 3, ..., m
relation	mlbl	MOD, WRD, QUT, FLT, UND

Table 1: Annotated Information

value	relation
MOD	common word-to-word dependency
WRD	relationships between a word and its inflection
QUT	relationships between quotation marks
FLT	words that has no clear modifiee, such as an interjection, are annotated as FLT
UND	undefined relationships

Table 2: Dependency Types

Table 1 shows the information annotated in our corpus. The detail of the dependency relation codes are described in Table 2. The value of the modifiee of a word indicates the ID number of the modifiee of that word. We allow right-to-left modification in our annotation method in addition to left-to-right modification. The modifiee of a word for which the dependency relation is FLT is tentatively set to the subsequent word.

The type of dependency is not necessarily based on a syntactic structure. For example, a demonstrative word occasionally does not correspond to a specific word but to a whole sentence [11]. We generalized the definition of dependency and made strict rules on the scope of dependencies, which should be a useful annotation for multiple purposes.

Figure 1 shows an example of our annotated corpus from the transcribed lectures of the University of the Air.

2.2. Work flow

Currently we have graduate students of linguistics correct the output of the automatic morphological analysis of the corpus, and annotate the word-to-word dependency information. A researcher of our group super-

```
*BOS* BOS ID=54
de(and)          pos=interjection         mod=2  mlbl=MOD
,                pos=special-character    mod=8  mlbl=MOD
ashi(foot)       pos=noun                 mod=4  mlbl=MOD
ni(on)           pos=particle             mod=5  mlbl=MOD
ha(topic.)       pos=particle             mod=8  mlbl=MOD
kutsu(shoes)     pos=noun                 mod=7  mlbl=MOD
wo(obj.)         pos=particle             mod=8  mlbl=MOD
ha(wear)         pos=verb                 mod=9  mlbl=WRD
i(infl.)         pos=inflection           mod=10 mlbl=MOD
te(particle)     pos=particle             mod=11 mlbl=MOD
i(be-ing)        pos=auxiliary            mod=12 mlbl=WRD
ru(infl.)        pos=inflection           mod=13 mlbl=MOD
.                pos=special-character
EOS* EOS
```

Figure 1: Example of Annotated Corpus

vises all of the corpus annotation works, and several researchers who represent each of the application areas for which the annotated corpus is used support that supervising researcher.

Before beginning to annotate the spoken language corpus, we had annotators make reports on problems and their solutions in annotating the existing written language corpus, and defined annotation rules according to these solutions. We then applied these annotation rules to the spoken language corpus annotation. Each annotator makes a list of any new problems. The annotators work to improve their consistency by discussing annotation problems and their solutions. The supervising researcher and the other researchers discuss the new problems regularly, and decide how to solve them. The annotation rules are expanded according to these new solutions.

Currently, there are about 10,000 sentences in the written language corpus and 3,000 sentences in the spoken language corpus have been annotated.

3. ANNOTATING THE SPOKEN-LANGUAGE CORPUS

3.1. Lectures of the University of the Air

The lectures of the University of the Air that we selected as the spoken language corpus domain have the characteristics of monologues rather than dialogs, because a lecturer makes a speech on a certain topic to unspecified people in each lecture. However, the speech is natural because it is not a simulated dialogue, and includes many spoken language expressions such as disfluencies and rephrasings. We selected lectures from various domains, and therefore the words which appear in the corpus can be expected to be sufficiently well balanced.

Filled-pauses and word fragments are written phonologically with tags [10]. We tokenized all texts into units of two types; one is a word-based recognition unit, the other is the component morphemes. The corpus consists of about 1,099K words and about 1,263K morphemes respectively. About 8.75% of all words are filled-pauses and 0.35% word fragments, which suggests that most of the utterances are spontaneous.

3.2. Problems in annotating spoken-language corpus

About six hundred new problems were found by the annotators while annotating the spoken language corpus. About ten percent of them are about colloquial expressions, and ninety percent of them are general problems in annotating both written and spoken Japanese, such as compound word segmentation, part-of-speech definitions, and so on. This might be partly due to the characteristics of our spoken language corpus as based on monologues. In addition, lecturers might to be expected to use less colloquial expressions compared to casual dialogs. Normal dialog data is likely to include more special colloquial expressions.

The following list shows typical problems in annotating our spoken language corpus and their current solutions.

- Very long sentences
  In spoken language, a sentence could be longer than in the written language [1], because the speaker corrects the contents of the utterance not by deleting information, but by adding information [12]. We prepared a written language sentence set and a spoken language sentence set that include almost the same number of sentences and counted the number of words included each sentence set. Table 3 shows the results. The spoken language sentence set includes about 1.6 times as many characters as the written language sentence set does. Annotators often wavered over setting dependency relationships between sentential units that were included in the same sentence but which do not have a direct relationship both syntactically and semantically with each other.

  Currently we keep the transcribed sentence boundaries as they were originally determined. However, how we decide on sentence boundaries and how we annotate sentence-to-sentence relationships in comparison with word-to-word relationships are still left as unsolved problems.

[1] Definition of a sentence in transcription is described in [10].

- Rephrasing
  Rephrasing is one of the causes of the lengthy sentences. Currently, we annotate dependency relations between rephrased expressions and original expressions when they satisfy the conditions defined by the annotators in order to identify rephrasing. However, we should think carefully whether we can consider the dependency between rephrased expressions as being at the same level as the common word-to-word dependencies such as subcategorizations when we use this information in various application systems.

- Pronouns without obvious referents
  There are more pronouns that refer not to a specific word but to a whole sentence or preliminary discourse than in written language corpus. As we mentioned above, a demonstrative word occasionally does not correspond to a specific word but to a whole sentence.

We expanded the annotation rules for annotating the spoken language corpus according to the problem list made by the annotators.

The next section describes our experimental results investigating the quality of the corpus that is annotated by applying the expanded the annotation rules.

4. EXPERIMENT

Using our experimental natural language parser, we reparsed both our written language corpus and spoken language corpus and compared the accuracy of the results in order to determine whether our annotation rules are appropriate for not only written language but for spoken language. If there are major differences in accuracy between them, we should carefully investigate whether the rules are applicable to only one of these corpora. Our parser is based on a stochastic language model [13].

Table 3 shows details of the data used in the experiment. In order to make the training conditions comparable, the two training data sets have almost the same size.

Each training set was divided into nine chunks. Words that appear in only one chunk among the nine were considered as unknown words. As we can see in Table 3, the rate of the unknown words is lower in the spoken language corpus than in the written language corpus. That is, the number of distinct words used was smaller for the spoken language corpus, showing that speakers reuse the same words more often, while writers use a wider range of vocabulary.

Spoken Language Corpus

	sent.	morph.	char.	word-base coverage
training	990	42243	61812	96.29%
test	110	4455	6561	94.01%

Written Language Corpus

	sent.	morph.	char.	word-base coverage
training	1656	40450	61709	91.93%
test	434	10210	15920	82.83%

Table 3: Data Set for Experiment

Word-based Accuracy

corpus	precision	recall
spoken language	94.90%	94.86%
written language	88.69%	88.70%

Sentence-based Accuracy

corpus	accuracy
spoken language	29.09%
written language	22.81%

Table 4: Result of Morphological Analysis

Table 4 shows the results of the morphological analysis. The accuracy of analyzing spoken language morphologically is higher than analyzing written language. The unknown word rate might have an effect on this difference in the accuracies of the morphological analysis.

Table 5 shows that there is no major difference between the accuracy of the dependency analysis for each corpus. In order to observe whether the dependency annotation rules are appropriate or not, it is better that the accuracy of the dependency analyses are not affected by the accuracy of the morphological analyses. Therefore the input of the dependency analysis was the manually corrected results of the morphological analysis.

Intuitively, we expect lower accuracy in analyzing the spoken language corpus where the average sentence length is longer than for the written language corpus. Indeed, the sentence-based accuracy of the dependency

corpus	word-based	sentence-based
spoken language	87.54%	9.09%
written language	87.06%	23.27%

Table 5: Result of Dependency Analysis

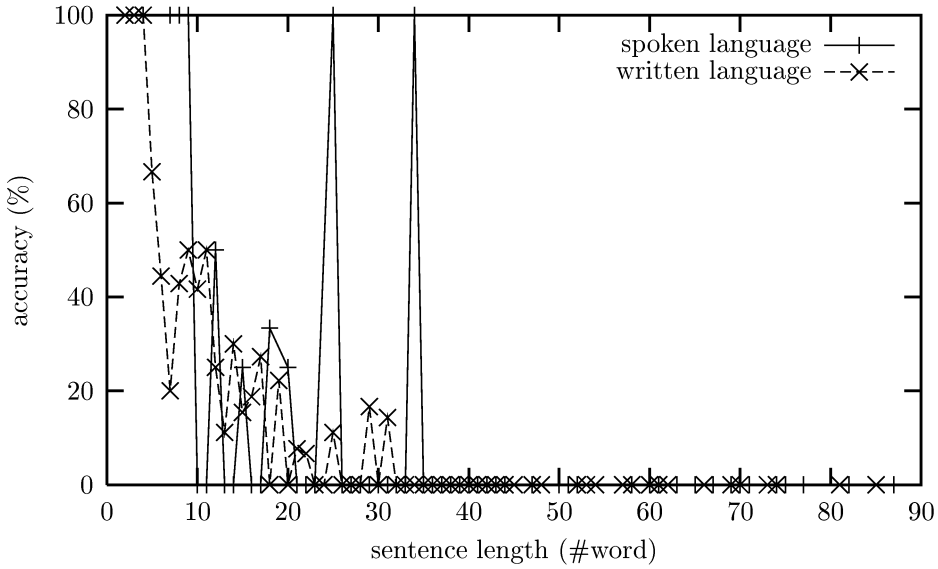

Figure 2: Sentence-based Accuracy and Sentence Length

analysis of the spoken language corpus is lower than for the written language corpus; on the other hand, the difference in average sentence length does not make any serious difference for word-based the accuracy of the dependency analysis.

The sentence length has effects on the sentence-based accuracy of the dependency analysis not only the spoken language corpus but also for the written language corpus. Figure 2 shows the relationship between sentence-based accuracy of the dependency analysis and the sentence length. Accuracy decreases as the sentences become longer for both corpora. The pattern of decrease in the accuracy of both corpora is quite similar.

These experimental results show that there is no fundamental difference between the accuracy of analyzing a written language corpus and a spoken language corpus. According to these results, it can be said that our annotation strategy is appropriate both for annotating a written language corpus and a spoken language corpus.

5. CONCLUDING REMARKS

We applied the same annotation rules in annotating both a written language corpus and a spoken language corpus. The experimental results of parsing both corpora show that we can get annotated corpora that have similar quality by applying the same annotation strategy.

We will continue annotating our spoken language corpus and expanding our annotation rules for other domains, such as dialogs, with the aim of building a multi-purpose corpus. We are also planning to define clear sentence boundaries for spoken language and sentence-to-sentence relations for annotating richer information into our corpus. These new annotations will be useful for various applications.

ACKNOWLEDGMENTS

We would like to thank Dr. Masafumi Nishimura and Dr. Hideo Watanabe for their support of this research.

REFERENCES

[1] http://www.itl.atr.co.jp/Japanese/overview/index.html

[2] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

[3] http://www.rwcp.or.jp/wswg/rwcdb/text/index.html

[4] Godfrey, J. J. et al., "Switchboard: Telephone speech corpus for research and development," *Proc. of ICASSP 1992*, pp. 517–520, 1992.

[5] Watanabe, H. et al., "Improving Natural Language Processing by Linguistic Document Annotation," *Proc. of COLING 2000 Workshop for Semantic Annotation and Intelligent Content*, pp. 20–27, 2000.

[6] Furui, S., "Spoken Language Resources," *GSK symposium*, 1999.

[7] http://www.itl.atr.co.jp/Japanese/overview/index.html

[8] Yamamoto, M., "Current status of construction of spoken dialog database", *Journal of the Acoustical Society of Japan*, Vol. 54, No. 11, pp. 797–802, 1998, in Japanese.

[9] Japan Electronic Industry Development Association, *Annual report on trends in natural language processing systems*, pp. 162–178, 1997, in Japanese.

[10] Itoh, N. et al., "A Method for Style Adaptation to Spontaneous Speech by Using a Semi-linear Interpolation Technique," *Proc. of ICSLP 2000*, in press.

[11] Hobbs, J., "Resolving Pronoun References," *Lingua 44*, pp. 311–338, 1978.

[12] National Language Research Institute of Japan, "Sentence patterns of spoken language (1)," *National Language Research Institute Report 18*, 1960, in Japanese.

[13] Mori, S. et al., "A Stochastic Parser Based on a Structural Word Prediction Model," *Proc. of COLING 2000*, pp. 558–564, 2000.