

Robust Parsing of Japanese Technical Text: An Initial Study

Richard F. E. Sutcliffe[†] and Nao Nashimoto*

Department of Computer Science
and Information Systems[†]
Department of Languages
and Cultural Studies*
University of Limerick
Limerick, Ireland

+353 61 202706 (Direct)
+353 61 202734 (Fax)

Richard.Sutcliffe@ul.ie (Email)
www.csis.ul.ie/staff/richard.sutcliffe (URL)

Abstract

A grammar and robust parsing system for the analysis of technical Japanese taken from a software instruction manual has been developed and evaluated. Initial results suggest a performance of 93% on simple noun phrases and 88% on verb phrases.

1 Introduction

Many language engineering tasks require accurate grammatical analysis of an input in order to facilitate further processing. In the domain of help systems relating to desktop software for example, it is possible to experiment with more sophisticated retrieval strategies if predicate-argument data can be extracted from the help text. For this reason we have previously carried out studies into the analysis of English text extracted from various software manuals (Sutcliffe, Koch and McElligott, 1996). These in turn led to the development of a simple multiple-pass robust parser. The objective of the present work was to investigate the extent to which the approach could be applied to Japanese working in the same application domain.

Much work has already been carried out on the syntactic analysis of Japanese. For example Kurohashi and Nagao (1998) describe the well-known Juman tokenisation and part-of-speech tagging system as well as the KNP dependency-based parser which uses Juman output. Mitsuishi, Torisawa and Tsujii (1998) have developed a grammar for Japanese designed for reasonable

accuracy and wide coverage, based on HPSG. This has been evaluated on the EDR Corpus¹ with good results.

Shirai, Inui, Tokunaga and Tanaka (1998) describe the application of a statistical language model combining lexical association statistics with syntactic preference for the disambiguation of dependency structures in parsing. Uchimoto, Sekine and Isahara (1998) use a maximum entropy model to predict dependency relations between phrasal units in a sentence. Finally, Haruno, Shirai and Ooyama (1998) describe a method for computing a matrix which describes the probability that one phrase modifies another in a sentence for each pair of phrases, based on the use of decision trees. This is effectively a technique for resolving structural ambiguity.

However, all the above work is being carried out in terms of either the EDR Corpus or the Kyoto Corpus² which are treebanks for magazine and newspaper texts, and newspaper texts respectively. No studies so far appear to have focussed on software instruction manuals. The key objectives were thus as follows:

- To study the syntactic characteristics of technical Japanese,
- To design a grammar for recognising certain constructs in this domain,
- To build and evaluate a robust parser using that grammar.

2 Method

2.1 Application Domain

The study was carried out on text derived from the Ami Pro User's Guide Release 3.1J (Ami Pro, 1994). Text within this guide can be characterised grammatically as follows:

- Many compound nouns to denote word processing and computing concepts, e.g. アンドウ レベル [undo level]³;
- Frequent and complicated and/or coordination at many levels, e.g. ドラッグ & ドロップ で 文字 を 移動 、 コピー する 手順 [drag (and) drop by-means-of text (obj) movement (comma-meaning-or) copy do procedure];

¹ www.ijjnet.or.jp/edr

² www-nagao.kuee.kyoto-u.ac.jp/index-e.html

³ Japanese examples are shown artificially tokenised for clarity and each is followed by a word-for-word translation.

- Pseudo constructs to denote specific information (e.g. key combinations):
〈編集 (E) コピー (C) 〉 を選択します。 [(left-angle-bracket) editing (left-round-bracket) E (right-round-bracket) copying (left-round-bracket) C (right-round-bracket) (right-angle-bracket) (obj) selection do verb suffix polite (full-stop)];
- Frequent use of Katakana and Roma-ji to denote computer terms, key strokes, command lines etc, as illustrated in the previous example;
- Many utterances which are not complete sentences but merely noun phrases, verb phrases etc. These are used for headings, bullet points, examples and so on;
- In other respects a limited, regular syntax.

2.2 Selection of Sentences

Sentences for analysis were selected from the manual as follows. The starting point was the set of 200 English sentences in Ami Pro (1993) which had been used in a previous study. These had been selected to illustrate the range of constructions and utterance types which are typical of the manual. For each of these the closest matching sentence or sentence pair was found in the Japanese version of the same manual. These were then sorted by sentence type and tagged using Juman 3.6⁴. The first 102 sentences which were correctly tokenised and correctly tagged by Juman were then selected.

2.3 Grammar Developed

The grammar is designed to recognise the following constructs (see Table 1):

- The Noun (N), incorporating prefixes and suffixes, the adjective (ADJ), including the demonstrative adjective and adjectival noun suffix, the Verb (V), which in Japanese can incorporate a Coordinated Noun Group plus a form of the verb する [do], and the particle (P),
- The noun group (NG) and coordinated noun group (CNG), in which noun groups are coordinated with ‘and’, comma or ‘or’,
- The Simple Noun Phrase (SNP), incorporating adjectives and the Coordinated Noun Group, and the Simple Noun Phrase 2/3, (SNP2/3) which joins SNPs with the particle の [’s],
- The particle phrase (PP), comprising an SNP3 followed by a Particle,
- The verb group (VG), Adverb Group ADVG, and Total Verb Group (TVG), including an

⁴ www-nagao.kuee.kyoto-u.ac.jp/index-e.html

ADVG and a possible Adjectival Predicate Suffix (APS),

- The Verb Phrase (VP), grouping particle phrases and the total verb group together.

Construct	Example Parse Tree of Construct
CNG	cng(and/[ng([n([ドラッグ]))], ng([n([ドロップ]))])) drag (and) drop
VG	vg([or/[ng([n([コピー]))], ng([n([移動]))], v(する)])] copying (or) movement do
TVG	tvg([adv(adv([adv(すでに)])], vg(vg([v(開いて)], v(いる)])), aps(aps([])))] already opening (verbal-suffix)
SNP2	snp2(no(snp([cng(and/[ng([n([今日]))])]), snp([cng(and/[ng([n([日付]))])])]))] today 's date
PP	pp([object, p(を), snp3(snp2(snp([cng(and/[ng([n([文字]))])])])))] text (object)
VP	vp([pp([object, p(を), snp3(snp2(snp([cng(and/[ng([n([文字]))])])])))]], tvg([adv(adv([])], vg(vg([and/[ng([n([削除]))], v(する)])), aps(aps([])))]])] text (object) deletion do

Table 1: Examples of Parse Trees Produced by Grammar Constructs

2.4 Parsing Algorithm

Parsing starts with a text which has been tokenised and tagged for part-of-speech using the Juman 3.6 system. This was used to replace the Brill Tagger employed for the English parser (Brill, 1992). Multiple passes are made over the input. During each pass an attempt is made to recognise a particular non-terminal in the grammar. If found, it is replaced by the parse tree for that non-terminal. For each analysis pass there is effectively a separate context-free grammar expressed in terms of terminal symbols or non-terminals recognised by the grammars corresponding to previous passes. From the perspective of parsing, a terminal symbol comprises a tuple containing the word, the general part-of-speech and the specific part-of-speech, as provided by Juman.

2.5 Evaluation

A standard method of evaluating parser performance is to use a parsed corpus and to make measurements based on a comparison of constituent boundaries in the reference parse with those of the candidate parse (Carroll, Briscoe and Sanfilippo, 1998). The Parseval scheme is a well-known form of this method (Black *et al.* 1991) while the method of Gaizauskas, Hepple and Huyck (1998) is a refinement of it.

Kurohashi and Nagao (1998) discuss their corpus of 20,000 sentences parsed with Juman and KNP. However this is based on newspaper texts taken from *Mainichi*. The EDR Corpus comprises sentences from magazine and newspaper articles. To our knowledge there is no corpus of Japanese

technical instruction manuals which could be used to evaluate our parser. For this reason we have resorted to a grammar-dependent method carried out in the following way:

- Output from the parser was inspected by an expert linguist and native speaker of Japanese;
- The following constructions were evaluated: TVG, SNP2, PP and VP;
- Each instance of a construct of one of these types was judged either totally correct or incorrect;
- Counts were then made of the total number of each type of construct, and the number of correct instances of each type. The results are summarised in Table 2.

	Number Correct	Total Number	Percent Correct
TVG	160	162	99%
SNP2	217	234	93%
PP	178	186	96%
VP	143	163	88%

Table 2: Results of Evaluation

3 Discussion and Conclusions

Considering the simplicity of the grammar, a large number of constructions can be handled accurately.

However, parsing is only as accurate as the part-of-speech tagging. While Juman is an excellent system, our initial trials in this domain suggest that 31% of utterances may contain an error either of tokenisation or tagging. We have not yet carried out a detailed analysis of the types of error and neither have we yet attempted to adapt the system to see if performance can be improved.

We can not analyse certain constructs due to deficiencies in the grammar. The main ones we have observed are:

- Relative clauses. We can in fact analyse relative clauses in simple nominal sentences such as 文書間でコピー、移動する手順 [document gap (location) copy (comma) movement do procedure] but these are not included in the current evaluation. More complicated examples in which a noun phrase including a relative clause is a constituent of another construct can not at present be handled, e.g. 文字を挿入する場所に挿入点を置きます。 [text (obj) insertion do place in insertion point (obj) place verb-suffix-polite (full stop)].
- One verb phrase in *te*-form modifying another. In this example *putting-together* is in *te*-form: 文字をまとめて削除する場合 [text (obj) putting-together deletion do case].

- Adverbial uses of number and counter noun suffix combinations such as: 文字メニューを1回選択すると、 [text menu (object) one counter-noun-suffix selection do when (comma) ...].
- Ranges of numbers expressed as ‘from ... to ...’, for example: 1 から 4 まで [...one from 4 until...].
- Syntactic uses of punctuation other than in comma coordination of noun groups. One example is the processing of quoted expressions. These can form constituents of other structures as in: 「操作の取り消し」を参照してください。 [(open-quote) operation 's cancellation (close-quote) (obj) referring doing please-give (full-stop)].

The structural ambiguity inherent in の [’s] processing as discussed in (Wu, de Paiva Alves and Furugori, 1998) can also not be handled. However, we have not yet encountered an instance of this in the software manual domain.

Acknowledgements

We are grateful to Prof. Makoto Nagao of University of Kyoto for making the Juman system publicly available. We are also indebted to Dr. Sadao Kurohashi for supplying several papers about Juman. As always, Denis Hickey and Redmond O’Brien provided invaluable systems support.

References

- Ami Pro (1993). *Lotus Ami Pro Word Processor for Windows User’s Guide Release 3*. Atlanta, GA: Lotus Development Corporation, Word Processing Division.
- Ami Pro (1994). *Lotus Ami Pro Word Processor for Windows User’s Guide Release 3.1J* (Japanese Version). Cambridge MA: Lotus Development Corporation.
- Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., & Strzalkowski, T. (1991). A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. *Proceedings of the DARPA Speech and Natural Language Workshop*, 306-311.
- Brill, E. (1992). A Simple Rule-Based Part-of-Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing, ANLP, Trento, Italy, 1992*.
- Carroll, J., Briscoe, T., Sanfilippo, A. (1998). Parser Evaluation: A Survey and a New Proposal. *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 28-30 May 1998*, 447-454.

- Gaizauskas, R., Hepple, M., & Huyck, C. (1998). A Scheme for the Comparative Evaluation of Diverse Parsing Systems. *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 28-30 May 1998*, 143-149.
- Haruno, M., Shirai, S., & Ooyama, Y. (1998). Using Decision Trees to Construct a Practical Parser. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Coling-ACL'98, Université de Montréal, Canada, 10-14 August 1998*, 505-511.
- Kurohashi, S., & Nagao, M. (1998). Building a Japanese Parsed Corpus while Improving the Parsing System. *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 28-30 May 1998*, 719-724.
- Mitsuishi, Y., Torisawa, K., & Tsujii, J. (1998). HPSG-Style Underspecified Japanese Grammar with Wide Coverage. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Coling-ACL'98, Université de Montréal, Canada, 10-14 August 1998*, 876-880.
- Shirai, K., Inui, K., Tokunaga, T., & Tanaka, H. (1998). An Empirical Evaluation on Statistical Parsing of Japanese Sentences using Lexical Association Statistics. *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, 80-87.
- Sutcliffe, R. F. E., Koch, H.-D., McElligott, A. (Eds.) (1996). *Industrial Parsing of Software Manuals*. Amsterdam, The Netherlands: Rodopi.
- Uchimoto, K., Sekine, S., & Isahara, H. (1998). Japanese Dependency Structure Analysis based on Maximum Entropy Models. *Proceedings of the 9th Conference of EACL*, 196-203.
- Wu, H., de Paiva Alves, E., & Furugori, T. (1998). Structural Disambiguation Based on Reliable Estimation of Strength Association. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Coling-ACL'98, Université de Montréal, Canada, 10-14 August 1998*, 1416-1422.