An Implemented Description of Japanese: The Lexeed Dictionary and the Hinoki Treebank

Sanae Fujita, Takaaki Tanaka, Francis Bond, Hiromi Nakaiwa
NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation
{sanae, takaaki, bond, nakaiwa}@cslab.kecl.ntt.co.jp

Abstract

In this paper we describe the current state of a new Japanese lexical resource: the Hinoki treebank. The treebank is built from dictionary definition sentences, and uses an HPSG based Japanese grammar to encode both syntactic and semantic information. It is combined with an ontology based on the definition sentences to give a detailed sense level description of the most familiar 28,000 words of Japanese.

1 Introduction

In this paper we describe the current state of a new lexical resource: the Hinoki treebank. The ultimate goal of our research is natural language understanding — we aim to create a system that can parse text into some useful semantic representation. This is an ambitious goal, and this presentation does not present a complete solution, but rather a road-map to the solution, with some progress along the way.

The first phase of the project, which we present here, is to construct a syntactically and semantically annotated corpus based on the machine readable dictionary Lexeed (Kasahara et al., 2004). This is a hand built self-contained lexicon: it consists of headwords and their definitions for the most familiar 28,000 words of Japanese. Each definition and example sentence has been parsed, and the most appropriate analysis selected. Each content word in the sentences has been marked with the appropriate Lexeed sense. The syntactic model is embodied in a grammar, while the semantic model is linked by an ontology. This makes it possible to test the use of similarity and/or semantic class based back-offs for parsing and generation with both symbolic grammars and statistical models.

In order to make the system self sustaining we base the first growth of our treebank on the dictionary definition sentences themselves. We then train a statistical model on the treebank and parse the entire lexicon. From this we induce a thesaurus. We are currently tagging other genres with the same information. We will then use this information and the thesaurus to build a parsing model that combines syntactic and semantic information. We will also produce a richer ontology — for example extracting selectional preferences. In the last phase, we will look at ways of extending our lexicon and ontology to less familiar words.

2 The Lexeed Semantic Database of Japanese

The Lexeed Semantic Database of Japanese consists of all Japanese words with a familiarity greater than or equal to five on a seven point scale (Kasahara et al., 2004). This gives 28,000 words in all, with 46,000 different senses. Definition sentences for these sentences were rewritten to use only the 28,000 familiar words (and some function words). The defining vocabulary is actually 16,900 different words (60% of all possible words). A simplified example entry for the last two senses of the word ドライバー *doraibā* "driver" is given in Figure 1, with English glosses added, but omitting the example sentences. Lexeed itself consists of just the definitions, familiarity and part of speech, all the underlined features are those added by the Hinoki project.

3 The Hinoki Treebank

The structure of our treebank is inspired by the Redwoods treebank of English (Oepen et al., 2002) in which utterances are parsed and the annotator selects the best parse from the full analyses derived by the grammar. We had four main reasons for selecting this approach. The first was that we wanted to develop a precise broad-coverage

Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 65–68,
Sydney, July 2006. ©2006 Association for Computational Linguistics

INDEX		ドライバー *doraibā*
POS		noun Lexical-Type `noun-lex`
FAMILIARITY		6.5 [1–7] (≥ 5) Frequency 37 Entropy 0.79
SENSE 1		…
SENSE 2	DEFINITION	自動車$_1$/を/運転$_1$/する/人$_1$/。 Someone who drives a car.
P(S$_2$) = 0.84	HYPERNYM	人$_1$ *hito* “person”
	SEM. CLASS	⟨`292:chauffeur/driver`⟩ (⊂ ⟨`5:person`⟩)
	WORDNET	*driver*$_1$
SENSE 3	DEFINITION	ゴルフ$_1$/で/、/遠距離$_1$/用/の/クラブ$_3$/。一番/ウッド/。 In golf, a long-distance club. A number one wood.
P(S$_2$) = 0.05	HYPERNYM	クラブ$_3$ *kurabu* “club”
	SEM. CLASS	⟨`921:leisure equipment`⟩ (⊂ 921)
	WORDNET	*driver*$_5$
	DOMAIN	ゴルフ$_1$ *gorufu* “golf”

Figure 1: Entry for the Word *doraibā* “driver” (with English glosses)

grammar in tandem with the treebank, as part of our research into natural language understanding. Treebanking the output of the parser allows us to immediately identify problems in the grammar, and improving the grammar directly improves the quality of the treebank in a mutually beneficial feedback loop.

The second reason is that we wanted to annotate to a high level of detail, marking not only dependency and constituent structure but also detailed semantic relations. By using a Japanese grammar (JACY: Siegel (2000)) based on a monostratal theory of grammar (Head Driven Phrase Structure Grammar) we could simultaneously annotate syntactic and semantic structure without overburdening the annotator. The treebank records the complete syntacto-semantic analysis provided by the HPSG grammar, along with an annotator's choice of the most appropriate parse. From this record, all kinds of information can be extracted at various levels of granularity: A simplified example of the labeled tree, minimal recursion semantics representation (MRS) and semantic dependency views for the definition of ドライバー$_2$ *doraibā* “driver” is given in Figure 2.

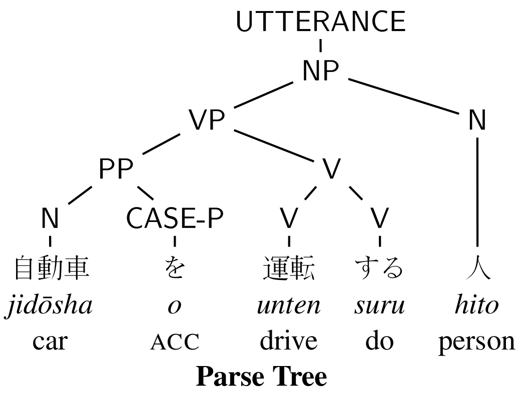
The third reason was that use of the grammar as a base enforces consistency — all sentences annotated are guaranteed to have well-formed parses. The last reason was the availability of a reasonably robust existing HPSG of Japanese (JACY), and a wide range of open source tools for developing the grammars. We made extensive use of tools from the the Deep Linguistic Processing with HPSG Initiative (DELPH-IN: `http://www.delph-in.net/`) These existing resources enabled us to rapidly develop and test our approach.

3.1 Syntactic Annotation

The construction of the treebank is a two stage process. First, the corpus is parsed (in our case using JACY), and then the annotator selects the correct analysis (or occasionally rejects all analyses). Selection is done through a choice of discriminants. The system selects features that distinguish between different parses, and the annotator selects or rejects the features until only one parse is left. The number of decisions for each sentence is proportional to $\log_2$ in the length of the sentence (Tanaka et al., 2005). Because the disambiguating choices made by the annotators are saved, it is possible to semi-automatically update the treebank when the grammar changes. Re-annotation is only necessary in cases where the parse has become more ambiguous or, more rarely, existing rules or lexical items have changed so much that the system cannot reconstruct the parse.

The Lexeed definition sentences were already POS tagged. We experimented with using the POS tags to mark trees as good or bad (Tanaka et al., 2005). This enabled us to reduce the number of annotator decisions by 20%.

One concern with Redwoods style treebanking is that it is only possible to annotate those trees that the grammar can parse. Sentences for which no analysis had been implemented in the grammar or which fail to parse due to processing constraints are left unannotated. This makes grammar cov-



$$\begin{array}{ll}
\langle h_0, x_1 \{ h_0 : \mathit{proposition\_m}(h_1) & \\
\quad h_1 : \mathit{hito\_n}(x_1) & \text{“person”} \\
\quad h_2 : \mathit{udef\_q}(x_1, h_1, h_6) & \\
\quad h_3 : \mathit{jidosha\_n}(x_2) & \text{“car”} \\
\quad h_4 : \mathit{udef\_q}(x_2, h_3, h_7) & \\
\quad h_5 : \mathit{unten\_s}(e_1, x_1, x_2) \} \rangle & \text{“drive”}
\end{array}$$

MRS

$$\begin{array}{l}
\{ x_1 : \\
\quad e_1 : \mathit{unten\_s}(\text{ARG}_1\ x_1 : \mathit{hito\_n}, \text{ARG}_2\ x_2 : \mathit{jidosha\_n}) \\
\quad r_1 : \mathit{proposition\_m}(\text{MARG}\ e_1 : \mathit{unten\_s}) \}
\end{array}$$

Semantic Dependency

Figure 2: Parse Tree, Simplified MRS and Dependency Views for ドライバー$_2$ *doraibā* “driver”

erage a significant issue. We extended JACY by adding the defining vocabulary, and added some new rules and lexical-types (more detail is given in Bond et al. (2004)). None of the rules are specific to the dictionary domain. The grammatical coverage over all sentences is now 86%. Around 12% of the parsed sentences were rejected by the treebankers due to an incomplete semantic representation. The total size of the treebank is currently 53,600 definition sentences and 36,000 example sentences: 89,600 sentences in total.

3.2 Sense Annotation

All open class words were annotated with their sense by five annotators. Inter-annotator agreement ranges from 0.79 to 0.83. For example, the word クラブ *kurabu* “club” is tagged as sense 3 in the definition sentence for *driver*$_3$, with the meaning “golf-club”. For each sense, we calculate the entropy and per sense probabilities over four corpora: the Lexeed definition and example sentences and Newspaper text from the Kyoto University and Senseval 2 corpora (Tanaka et al., 2006).

4 Applications

4.1 Stochastic Parse Ranking

Using the treebanked data, we built a stochastic parse ranking model. The ranker uses a maximum entropy learner to train a PCFG over the parse derivation trees, with the current node, two grandparents and several other conditioning features. A preliminary experiment showed the correct parse is ranked first 69% of the time (10-fold cross validation on 13,000 sentences; evaluated per sentence). We are now experimenting with extensions based on constituent weight, hypernym, semantic class and selectional preferences.

4.2 Ontology Acquisition

To extract hypernyms, we parse the first definition sentence for each sense (Nichols et al., 2005). The parser uses the stochastic parse ranking model learned from the Hinoki treebank, and returns the semantic representation (MRS) of the first ranked parse. In cases where JACY fails to return a parse, we use a dependency parser instead. The highest scoping real predicate is generally the hypernym. For example, for *doraibā*$_2$ the hypernym is 人 *hito* “person” and for *doraibā*$_3$ the hypernym is クラブ *kurabu* “club” (see Figure 1). We also extract other relationships, such as synonym and domain. Because the words are sense tags, we can specialize the relations to relations between senses, rather than just words: ⟨hypernym: doraibā$_3$, *kurabu*$_3$⟩.

Once we have synonym/hypernym relations, we can link the lexicon to other lexical resources. For example, for the manually constructed Japanese ontology **Goi-Taikei** (Ikehara et al., 1997) we link to its semantic classes by the following heuristic: look up the semantic classes C for both the headword (w_i) and hypernym(s) (w_g). If at least one of the index word’s classes is subsumed by at least one of the genus’ classes, then we consider the relationship confirmed. To link cross-linguistically, we look up the headwords and hypernym(s) in a translation lexicon and compare the set of translations $c_i \subset C(T(w_i))$ with WordNet (Fellbaum, 1998)). Although looking up the translation adds noise, the additional filter of the relationship triple effectively filters it out again.

Adding the ontology to the dictionary interface makes a far more flexible resource. For example, by clicking on the ⟨hypernym: doraibā$_3$, $goruf u_1$⟩ link, it is possible to see a list of all the senses related to golf, a link that is inaccessible in the paper dictionary.

4.3 Semi-Automatic Grammar Documentation

A detailed grammar is a fundamental component for **precise** natural language processing. It provides not only detailed syntactic and morphological information on linguistic expressions but also precise and usually language-independent semantic structures of them. To simplify grammar development, we take a snapshot of the grammar used to treebank in each development cycle. From this we extract information about lexical items and their types from both the grammar and treebank and convert it into an electronically accesible structured database (the lexical-type database: Hashimoto et al., 2005). This allows grammar developers and treebankers to see comprehensive up-to-date information about lexical types, including documentation, syntactic properties (super types, valence, category and so on), usage examples from the treebank and links to other dictionaries.

5 Further Work

We are currently concentrating on three tasks. The first is improving the coverage of the grammar, so that we can parse more sentences to a correct parse. The second is improving the knowledge acquisition, in particular learning other information from the parsed defining sentences — such as lexical-types, semantic association scores, meronyms, and antonyms. The third task is adding the knowledge of hypernyms into the stochastic model.

The Hinoki project is being extended in several ways. For Japanese, we are treebanking other genres, starting with Newspaper text, and increasing the vocabulary, initially by parsing other machine readable dictionaries. We are also extending the approach multilingually with other grammars in the DELPH-IN group. We have started with the English Resource Grammar and the Gnu Contemporary International Dictionary of English and are investigating Korean and Norwegian through cooperation with the Korean Research Grammar and NorSource.

6 Conclusion

In this paper we have described the current state of the Hinoki treebank. We have further showed how it is being used to develop a language-independent system for acquiring thesauruses from machine-readable dictionaries.

With the improved the grammar and ontology, we will use the knowledge learned to extend our model to words not in Lexeed, using definition sentences from machine-readable dictionaries or where they appear within normal text. In this way, we can grow an extensible lexicon and thesaurus from Lexeed.

Acknowledgements

We thank the treebankers, Takayuki Kuribayashi, Tomoko Hirata and Koji Yamashita, for their hard work and attention to detail.

References

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano. 2004. The Hinoki treebank: A treebank for text understanding. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*. Springer Verlag. (in press).

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Chikara Hashimoto, Francis Bond, Takaaki Tanaka, and Melanie Siegel. 2005. Integration of a lexical type database with a linguistically interpreted corpus. In *6th International Workshop on Linguistically Integrated Corpora (LINC-2005)*, pages 31–40. Cheju, Korea.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo. 5 volumes/CDROM.

Kaname Kasahara, Hiroshi Sato, Francis Bond, Takaaki Tanaka, Sanae Fujita, Tomoko Kanasugi, and Shigeaki Amano. 2004. Construction of a Japanese semantic lexicon: Lexeed. SIG NLC-159, IPSJ, Tokyo. (in Japanese).

Eric Nichols, Francis Bond, and Daniel Flickinger. 2005. Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, pages 1111–1116. Edinburgh.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christoper D. Manning, Dan Flickinger, and Thorsten Brant. 2002. The LinGO redwoods treebank: Motivation and preliminary applications. In *19th International Conference on Computational Linguistics: COLING-2002*, pages 1253–7. Taipei, Taiwan.

Melanie Siegel. 2000. HPSG analysis of Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 265–280. Springer, Berlin, Germany.

Takaaki Tanaka, Francis Bond, and Sanae Fujita. 2006. The Hinoki sensebank — a large-scale word sense tagged corpus of Japanese —. In *Frontiers in Linguistically Annotated Corpora 2006*. Sydney. (ACL Workshop).

Takaaki Tanaka, Francis Bond, Stephan Oepen, and Sanae Fujita. 2005. High precision treebanking – blazing useful trees using POS information. In *ACL-2005*, pages 330–337.