Augmenting a Semantic Verb Lexicon with a Large Scale Collection of Example Sentences

Kentaro Inui*, Toru Hirano†, Ryu Iida*, Atsushi Fujita‡ and Yuji Matsumoto*

* Nara Institute of Science and Technology
Takayama, Ikoma, Nara, 630-0192, Japan
{inui,ryu-i,matsu}@is.naist.jp
† Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
fujita@nuee.nagoya-u.ac.jp
‡ NTT Cyber Space Laboratories
Hikarinooka, Yokosuka, Kanagawa, 239-0847, Japan
hirano.tohru@lab.ntt.co.jp

Abstract

One of the crucial issues in semantic parsing is how to reduce costs of collecting a sufficiently large amount of labeled data. This paper presents a new approach to cost-saving annotation of example sentences with predicate-argument structure information, taking Japanese as a target language. In this scheme, a large collection of unlabeled examples are first clustered and selectively sampled, and for each sampled cluster, only one representative example is given a label by a human annotator. The advantages of this approach are empirically supported by the results of our preliminary experiments, where we use an existing similarity function and naive sampling strategy.

1. Introduction

The task of identifying the argument structure and its thematic role fillers for each predicate in a given input text is called predicate-argument structure analysis, or more simply semantic parsing. It has been attracting an increasing number of researchers because (a) it is considered as a crucially important technology in a wide range of NLP applications and (b) resources such as semantic lexicons of verbs (Dorr, 1997; Baker et al., 1998, etc.) and semantically annotated corpora (Palmer et al., 2005) are becoming increasingly available.

Like other NLP tasks, existing methods for semantic parsing can be classified as either the supervised approach (Gildea and Jurafsky, 2002; Thompson et al., 2003, etc.) or the unsupervised approach (Lapata and Brew, 2004, etc.). Supervised methods have so far tended to significantly outperform the unsupervised methods, while requiring considerable costs of supervision. If the supervised approach is taken, therefore, the issue of how to reduce costs of collecting labeled data needs to be addressed. Considering this background, in this paper, we present a new approach to cost-saving annotation of example sentences with argument structure information, taking Japanese as a target language.

2. Semantic role labeling in Japanese

In Japanese, a clause consists of a sequence of arguments and adverbial modifiers followed by a predicate as exemplified by sentence (1).

(1)

kare-wa	*shousai-o*	*keisatsu-ni*	*hanashi-ta*
he-TOP	details-ACC	police-DAT	tell-PAST
Nominative	Accusative	Dative	Predicate
Agent	Theme	Beneficiary	hanasu#1

He toled the details to the police.

For each clause in a given, syntactically parsed input sentence, semantic parsing is the task of resolving ambiguity at the following three levels:

(a) *Syntactic case*: The syntactic case (or grammatical role) of an argument is typically indicated by such a case marker as '*o* (ACC)' (e.g. '*shousai-o* (details-ACC)' in (1)). However, syntactic cases are not always linguistically indicated; for instance, the syntactic case of an argument is unspecified if it is marked by the topic-marking particle '*wa*', as in '*kare-wa* (he-TOP)' in (1). The syntactic case of the gap in a relative clause is also linguistically unspecified.

(b) *Verb sense*: the predicate may be polysemous, each sense associated with one or more predicate-argument structure frames. In (1), for instance, the verb '*hanasu* (to tell)' is used in the sense of hanasu#1, which has the frame ⟨Agent, Theme, Beneficiary⟩.

(c) *Semantic role*: The mapping from syntactic cases to semantic roles is not one-to-one. For instance, the Dative case, marked by '*ni*', maps to several semantic roles such as Goal, Beneficiary and Adjunct-Time.

It is also important to note that, in Japanese, even obligatory arguments of a predicate are often elided when they are inferable from the context, and the predicate-argument structure of such an incomplete clause may be undecidable unless semantic parsing is carried out interacting with ellipsis resolution. While it is an intriguing issue in itself, in this paper, we assume that semantic parsing is carried out independently before ellipsis resolution. If multiple semantic interpretations are possible for a given clause, the semantic parsing model is required to output all the interpretations while excluding wrong ones.

3. Proposal

The basic idea is the following. A large collection of unlabeled examples are first clustered and selectively sampled, and for each sampled cluster, only one representative example is labeled by a human annotator. More specifically, the overall supervision process is designed as follows:

1. *Initialization*: For each clause in a given syntactically parsed corpus, extract an example of the form of $\langle N_{i1}\text{-}C_{i1}, \ldots, N_{im}\text{-}C_{im}; V_i \rangle$, where N_{ij} and C_{ij} are the head noun of the case filler and the case marker of the *j*-th case-marked argument of verb V_i, deleting arguments that are not marked by a case marker (e.g. TOP-marked arguments).

2. *Clustering*: Automatically cluster examples using an inter-example similarity function.

3. *Selective sampling*: Automatically sample a cluster that is expected to be most useful for semantic parsing and choose one representative example from it, i.e. the medoid of the cluster.

4. *Annotation*: Manually annotate the chosen representative example with the predicate-argument structure. For instance, if sentence (1) is chosen, it is annotated with the predicate-argument structure: ⟨Agent:*kare*, Theme:*shousai*, Beneficiary:*keisatsu*; hanasu#1⟩.

5. *Label spreading*: Automatically label all the remaining examples in the cluster analogously to the manually annotated representative example.

6. *Termination*: Go to 3 unless a predefined termination condition is satisfied.

If examples are appropriately clustered and if clusters to annotate are carefully chosen, labeling only a small portion of a given example set is expected to achieve performance that is comparable to what would be achieved with all the examples manually labeled.

4. Clustering of examples

The goal of example clustering is to merge as many examples that share the same predicate-argument structure as possible while avoiding creating erroneous clusters. We achieve this by two sorts of bottom-up clustering methods: verb-wise clustering and cross-verb clustering.

4.1. Verb-wise clustering

In verb-wise clustering, examples that share the same verb are considered to be merged. For this job, in our preliminary experiments, we have so far examined the similarity function proposed by (Kawahara and Kurohashi, 2002) with a few minor modifications. Given a pair of examples, $\langle N_{11}\text{-}C_{11}, \ldots, N_{1m}\text{-}C_{1m}; V_1 \rangle$ and $\langle N_{21}\text{-}C_{21}, \ldots, N_{2n}\text{-}C_{2n}; V_2 \rangle$, where V_1 and V_2 are identical, the similarity between them is given by a similarity function that takes the following factors into account:

1. the similarity between the arguments immediately followed by the verb, N_{1m}-C_{1m} and N_{2n}-C_{2n},

2. the overlap between the sets of the belonging case markers, $\{C_{11}, \ldots, C_{1m}\}$ and $\{C_{21}, \ldots, C_{2n}\}$, and

3. the similarity between N_{1i}-C_{1i} and N_{2j}-C_{2j}, which is given by the semantic similarity between N_{1i} and N_{2j} if C_{1i} and C_{2j} are identical or zero otherwise.

The semantic similarity between two nouns is calculated by a function of the length of the path between them in a hierarchically organized thesaurus.
Merging two examples produces a cluster containing the union of their argument sets. For example, when merging (2a) and (2b), we obtain (2c).

(2) a. *shushou-ga* *gejun-ni* *houan-o* *happyou-suru*
PM-NOM end-DAT bill-ACC announce

b. *keikaku-o* *gatsu-ni* *happyou-suru*
plan-ACC (month)-DAT announce

c. {*shushou*}-*ga* { *gejun* / *gatsu* }-*ni* { *houan* / *keikaku* }-*o* *happyou-suru*

The similarity between clusters is defined analogously.

4.2. Cross-verb clustering

Suppose that a set of sampled examples for a certain verb, say '*kouhyou-suru* (to announce/disclose)', are already manually labeled. This set of labeled examples should be useful for clustering examples of other verbs of the same semantic group, such as '*happyou-suru* (to announce/publish)', because semantically similar verbs tend to have similar predicate-argument structures as demonstrated by many linguists (Levin, 1993, etc.).
Given a source verb, V_s, whose examples are already annotated, and a target verb, V_t, whose examples are to be clustered and annotated, cross-verb clustering is carried out through the following process.

1. Choose a representative example from each cluster of the examples of V_t,

2. For each representative example chosen in 1, find its most similar example from the V_s examples, and

3. If the V_s examples associated with two representative V_t examples have the same predicate-argument structure, merge the clusters of the two V_t examples.

5. Preliminary experiments

We report the present results on our preliminary experiments to show how effectively our example clustering method reduces the cost of manual annotation while maintaining the quality of annotation and the accuracy of semantic role labeling.

5.1. Data

We first chose four verbs, '*hanasu* (to talk)', '*hatsubai-suru* (to put on sale)', '*happyou-suru* (to announce/publish)' and '*fueru* (to increase)'. They are highly frequent in our newspaper corpus and each represents a distinct semantic group

verb	arg.str.	training	test
hanasu (talk)	5	1867	57
happyou-suru (announce)	3	2282	76
hatsubai-suru (put on sale)	1	635	10
fueru (increase)	2	3061	77

Table 1: Distribution of the training and test examples

of verbs. For each verb, we defined a set of predicate-argument structure frames based on the definition given by the IPA Lexicon of Basic Japanese Verbs (Information-Technology Promotion Agency, Japan, 1987). The number of the predicate-argument structure frames associated with each verb is shown in Table 1. One may suspect that the verb '*hatsubai-suru* (put on sale)' is not appropriate for evaluation because it is not polysemous. However, recall that the semantic parsing task we consider includes the interpretation of ambiguous case markers, such as '*ni* (Goal, Beneficiary, Adjunct-Time, etc)'; semantic parsing is thus not necessarily a trivial job even for such unambiguous verbs.

For the four verbs, we then collected a set of examples, which was used for training a semantic parsing model, through the following process:

1. For each verb, extract all the examples from the 13-year worth of the Mainichi Newspaper articles.
2. Discard passivized or causativized examples.
3. Delete arguments whose syntactic cases is not linguistically indicated.
4. Merge identical examples.

We finally annotate all the extracted examples (8385 types in total) with a sense tag and semantic role labels, in which ambiguous were labeled with all the interpretations. We call this training set the *primary* training set.

To simulate cross-verb clustering, we additionally created another set of training examples. We chose another four verbs, '*kataru* (to talk)', '*kouhyou-suru* (to announce/disclose)', '*uridasu* (to offer/launch)', and '*heru* (to decrease)', so that each corresponded to one of the former four verbs, and obtained 4551 labeled examples in total from the same corpus as above. We call this set the *secondary* training set.

For experiments on semantic parsing, we collected another distinct set of examples from a one-month worth of newspaper articles that were not overlapped with the corpus for training data collection. Manual annotation was done in an analogous manner to the above except that non-case-marked arguments, such as topic-marked arguments, were not deleted but were annotated with their syntactic case labels. As a result, we obtained 220 test examples.

5.2. Results

We evaluate the effects of verb-wise clustering and cross-verb clustering by comparing them with a baseline in which all the examples in the primary training set are individually labeled by a human annotator.

	(a) SV cost	(b) [%] SV err. rate	(c) [%] SP Acc.
Baseline	8385	0.00	99.3
VW clustering	2745	0.31	97.7
VW+CV clustering	1505	0.70	97.3

Table 2: Effects of verb-wise (VW) clustering and cross-verb (CV) clustering on (a) the cost of supervision (i.e. the number of clusters), (b) the error rate of the supervision of the primary training set, and (c) the accuracy of semantic parsing on the test data.

5.2.1. Clustering

Table 2(a) shows how much of the supervision cost was reduced by clustering. Since we assume the supervision scheme where a human annotator labels only one representative example for each cluster, we estimate the cost of supervision by the number of clusters. In the baseline method, all of the 8385 examples in the primary training set needed to be manually labeled. On the other hand, when verb-based clustering was applied, this number reduced to 2745, which further reduced almost by half when verb-based clustering was followed by cross-verb clustering.

More examples are merged in the clustering process, more examples in the training set are annotated automatically by label spreading, which may produce noise in the training set. Table 2(b) shows the error rates of the supervision by verb-wise and cross-verb clustering, which indicates that our way of clustering examples and spreading labels made only a very small number (0.70%) of errors in the training set.

5.2.2. Semantic parsing

We then used those labeled examples as training data in an experiment on semantic parsing. In this experiment, we applied a simple nearest neighbor-based classification algorithm to the aforementioned 220 test examples, where we used an inter-example similarity function analogous to the heuristic function devised by Kurohashi and Nagao (1994). The performance was measured by accuracy, where the model's output was considered correct only if the verb sense and semantic roles are all correctly chosen.

The results are shown in Table 2(c). First of all, the baseline model almost perfectly solved the task, achieving 99.3%, with the 8385 error-free training examples, which indicates that the training set was large enough for this problem setting. Compared with this, the proposed models gained a slightly lower accuracy. Given the amount of the reduction of annotation cost, however, this fall may be traded off.

Next we plotted the learning curve of the proposed model (VW+CV clustering) by altering the amount of the primary training data. The results are presented in Figure 1, where the x-axis denotes how many year worth of newspaper articles were used to collect the training data. The curve shows that the accuracy hit the ceiling in the middle and slightly fell down toward the end. This may be an indication that the size of the test set was not large enough for statistically reliable evaluation.

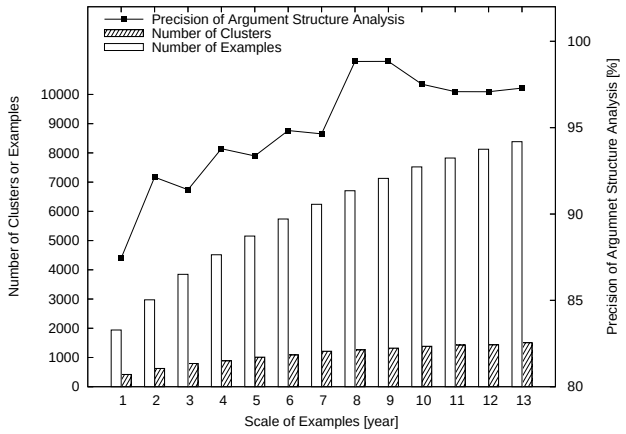

Figure 1: Learning curve in semantic parsing

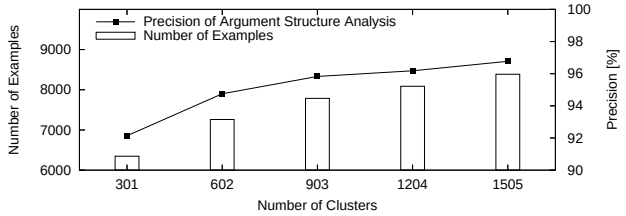

Figure 2: Learning curve of the proposed model with selective sampling

5.2.3. Selective sampling

Finally, we also examined a very simple strategy for selective sampling by simulating the situation where clusters were annotated in the descending order of the size, and plotted the learning curve by altering the number of the clusters to annotate as shown in Figure 2.

According to the results, the model achieved 92.1% in accuracy when only 301 of the 1505 clusters were annotated. In the experiment presented in 5.2.2, on the other hand, we obtained 420 clusters from one-year worth of newspaper articles and the model achieved only 87.4% accuracy when those clusters were all labeled, which was significantly lower then the above figure with selective sampling. This suggests that results achieved by labeling only a small portion of the set of clusters selectively sampled from a very large set of examples can be better than those achieved by labeling all the clusters obtained from a small example set.

6. Conclusion

In this paper we have presented a new approach to cost-saving annotation of example sentences with predicate-argument structure information, taking Japanese as a target language. The advantages of this approach are empirically supported by the results of our preliminary experiments although the scale of experiments clearly needs to be extended.

We are currently promoting a three-year project that aims at the development of a semantic lexicon of Japanese verbs based on the theoretical framework of Lexical Conceptual Structure (Takeuchi et al., 2005). We are planning to augment this lexicon with a large collection of semantically annotated examples by applying the method presented in this paper.

In the experiments, we have so far used manually tuned similarity metrics and threshold parameters for both example clustering and semantic role labeling. The issue of optimizing similarity metrics can be addressed in the framework of semi-supervised clustering, which scopes both metric learning and constrained clustering (Bilenko et al., 2004, etc.).

7. References

C. F. Baker, C. J. Filmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pages 86–90.

M. Bilenko, S. Basu, and R. J. Mooney. 2004. Integrating constrains and metric learning in semi-supervised clustering. In *Proceedings of the International Conference on Machine Learning*.

B. J. Dorr. 1997. Large-scale acquisition of lcs-based lexicons for foreign language tutoring. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 139–146.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Information-Technology Promotion Agency, Japan. 1987. *Japanese Verbs: A Guide to the IPA Lexicon of Basic Japanese Verbs*.

D. Kawahara and S. Kurohashi. 2002. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*.

S. Kurohashi and M. Nagao. 1994. A method of case structure analysis for Japanese sentences based on examples in case farme dictionary. *IEICE Transactions on Information and Systems*, E77-D(2):227–239.

M. Lapata and C. Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(2):45–73.

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago Press.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):266–274.

K. Takeuchi, K. Inui, and A. Fujita. 2005. Construction of compositional lexical database based on Lexical Concetual Structure for Japanese verbs (in Japanese). *Lexicon Forum*, 2.

C. A. Thompson, R. Levy, and C. D. Manning. 2003. A generative model for semantic role labeling. In *Proceedings of the 14th Europian Conference on Machine Learning (ECML)*, pages 397–408.