VERBAL CASE FRAME ACQUISITION FROM A BILINGUAL CORPUS: GRADUAL KNOWLEDGE ACQUISITION

Hideki Tanaka†
NHK Science and Technical Research Laboratories
†tanakah@strl.nhk.or.jp

Abstract

This paper describes acquisition of English surface case frames from a corpus, based on a gradual knowledge acquisition approach. To acquire and unambiguously accumulate precise knowledge, the process is divided into three steps which are assigned to the most appropriate processor: either a human or a computer. The data is prepared by human workers and the knowledge is acquired and accumulated by a leaning program. By using this method, inconsistent human judgement is minimized. The acquired case frames basically duplicate human work, but are more precise and intelligible.

1 Gradual Knowledge Acquisition

We have been developing an English-to-Japanese machine translation (MT) system for news reports in English (Aizawa T., 1990) (Tanaka H., 1991) and have so far studied the translation selection problem for common English verbs (Tanaka H., 1992). Recently, we examined the problem of multiple translations for common English verbs (Tanaka H., 1993). Our MT system uses surface verbal case frames (simply written as case frames) to select a Japanese translation for an English verb. The need to acquire and accumulate case frames leads directly to three problems.

(1) How to obtain detailed case frames which are accurate enough to translate highly polysemous verbs?

(2) How to accumulate a number of case frames in an unambiguous way.

(3) Manual case frame acquisition tends to yield inconsistent results since human judgements are changeable. How can we maintain consistency?

We need to devise a clear methodology for acquiring sufficient case frames and accumulating them in a way that is unambiguous and consistent.

In this paper, we propose a gradually building up a knowledge base from a bilingual corpus to cope with these three problems. The knowledge base is a collection of case frames. Fig. 1 shows an overall view of our approach.

The process is divided into three steps which are assigned to the most appropriate processor: a human or a computer. Using this method, detailed knowledge is obtained from the target domain texts, unstable human judgement is confined, and case frames are accumulated unambiguously by using a learning algorithm.

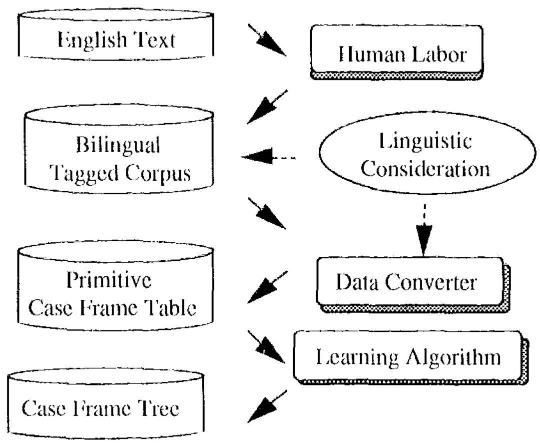

Fig. 1: Case-Frame Tree Acquisition from a Bilingual Corpus

We begin by preparing a tagged bilingual corpus seeking detailed knowledge in target domain texts. The annotation described in the corpus is the syntactic information of the texts and the translation. They are assigned manually since human translators can do such jobs as syntactic tagging and translation with far more consistency than writing case frames directly.

Next, the corpus is converted into an intermediate data form called the primitive case-frame table (PCFT). Finally a statistical learning algorithm is used to extract the case frames from the PCFT and accumulate them in a clear-cut fashion.

While this approach let us avoid writing case frames directly using linguistic contemplation, human activity plays an important role in designing and constructing the corpus and converting it into the PCFT (Fig. 1).

The case frames are represented in a discrimination tree, which has several attractive features for word-sense selection (Okumura M., 1990). The biggest attraction of the learning algorithm, we think, is its intelligibility; compared with the algorithms for neural networks, for example, it produces highly intelligible results if the input is appropriate.

Knowledge acquisition by machine learning from a corpus has recently been getting more attention than ever in some natural language processing fields. Cardie(1992, 1993) applied this approach to predict the antecedent of relative pronouns and attributes of unknown words. Utsuro(1993) introduced a methodology for automatically acquiring the verbal case frames from bilingual corpora in a different way than our methodology.

2 Case Frames for Translation

Our machine translation system uses case frames for the translation of English verbs. Fig. 2 shows illustrative case frames for the word *take*.

SN [man] take ON [boy] 選ぶ (select)
SN [I] take ON [him] PN[to] PNc[BUILD] 連れていく (escort)
SN [HUMAN] take ON [CON] PN[to] PNc[BUILD] 持っていく (bring)

Fig. 2: Example of Case Frames for *take*

We write case categories (SN (subject noun) and PN(preposition) here) and specify their restrictions. The restriction can be a semantic category like HUMAN or a word form itself like *boy* . There may be several hundred case frames for the most common English verbs.

The translation selection is performed after the parser produces a syntactic structure for the input sentence. The system compares the syntactic structure with the case frames and selects the translation from the best-matching case frame. Translation selection is performed without considering the context. Our new case frames are designed to follow the same protocol.

There are three factors to consider at this point.

(1) How many and what kinds of case categories should be used?
(2) In which order should the system compare the syntactic structure and the case categories in a case frame?
(3) What kind of restriction should we use?

In this paper, we will deal mainly with the first two factors. Our solution is to use a discrimination tree for the case-frame representation and a statistical algorithm for learning. The necessary case categories are selected and stacked in a tree form, one by one, according to their contribution to the translation selection . We call the obtained tree the case-frame tree. Fig. 3a is an example of a case-frame tree for *take*.

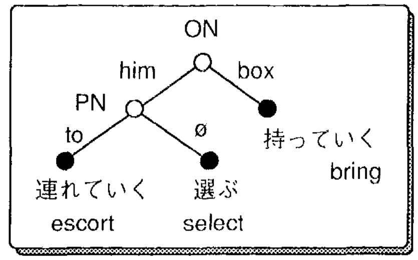

Fig. 3a: Example of a Case-Frame Tree

ON[box]	持っていく
ON[him]	選ぶ
ON[him] PN[to]	連れていく

Fig. 3b: Linear Case Frames for Fig. 3a

Comparison with the syntactic structure is made from the root node to the leaf nodes of the case frame-tree and no backtracking is allowed. The comparison is executed deterministically. If we read the tree from the root to the leafs, it can be expanded into a linear case frame, as shown in Fig. 3b. This increases the intelligibility of the case-frame tree enabling a human lexicographer to evaluate it from a linguistic viewpoint.

3 Learning from the PCFT

A case-frame tree can be regarded as a decision tree. Decision-tree learning has a long research history and many algorithms have been developed. Among them, the ID3 group (Quinlan J., 1993) of programs and its descendants satisfy our solution in Sec. 2. We apply the latest program, C4.5 (Quinlan J., 1993), to our problem. This algorithm learns a decision tree from an attribute-value and class table. An example of such a table is shown in Table 1.

Table 1: Example of a Primitive Case Frame Table

SN	V	ON	PN	PNc	translation
I	take	him	to	theater	連れていく (escort)
you	take	him	to	school	連れていく (escort)
you	take	him	to	park	連れていく (escort)
you	take	box	to	theater	持っていく (bring)
you	take	box	to	park	持っていく (bring)
I	take	box	to	school	持っていく (bring)
I	take	him	ø	ø	選ぶ (select)
you	take	him	ø	ø	選ぶ (select)

The first row of the table represents the attributes or the case categories. The values of the attributes are the restrictions of the case categories. Word forms are used in this. Since the algorithm produces a case-frame tree from this table, we term the table a "Primitive Case-frame Table (PCFT)."

The C4.5 first puts all translations listed in the PCFT under a root node then recursively selects one case category and partitions the translations according to the word forms of the selected category. For the case category selection, a criteria based on the entropy reduction of translations gained by the partitioning is used. See (Quinlan J., 1993) for more details. In a word, this algorithm places case categories from the root node to the leaf nodes according to the category's ability for translation discrimination. The case-frame tree in Fig. 3a was produced from Table 1. It does not have a node corresponding to a subject. This simply means the subject information is redundant in selecting the translation of *take* in Table 1.

4 Data Preparation

4.1 Construction of the Bilingual Corpus

As mentioned in Sec. 1, the data for machine learning is prepared in two steps: construction of a bilingual corpus and its conversion into a PCFT. Following are the factors considered and the steps taken to put together our corpus.

• Source

Since we could not find a readily available bilingual corpus from the news domain, we decided to make one ourselves by using the Associated Press (AP) wire service news text and adding a Japanese translation to it.

• Target

We selected 15 verbs known to be problematic verbs for machine translation: *come, get, give, go, make, take, run, call, cut, fall, keep, look, put, stand,* and *turn.*

Since case frames correspond to simple sentences, we did not deal with long sentences. The maximum sentence length was set at 15 words.

• Quantity of Data

To estimate the necessary amount of data, we investigated the monthly frequency of each verb appearing over six months. The frequency showed a fixed tendency over the measurement periods, suggesting that the data for one month is a good starting point. We decided to use two months, January 1990 and January 1991, for the English sentence extraction.

• Construction

(1) Preparing the English text

Sentences up to 15 words long which contain one or more of the 15 target verbs were automatically extracted from the two-month AP source text.

(2) Identifying the range governed by the verb

The range which the target verb directly governs in the English text was manually identified. The two lines starting with ENG in Fig. 4 are an example.

(3) Constructing the English case data

The a priori-defined category labels for each part of the ENG data were manually marked and the head word and functional word in each category were identified. The lines starting with CASE in Fig. 4 correspond to this data.

We had defined 34 category labels beforehand. Twelve of them (*sentence category labels*) were assigned to verbs to identify the sentence category from which the verb was extracted. Example categories are: V (declarative sentence), PVQ (polar question), IMV (imperative sentence), PASV (passive sentence), and IV (to-infinitive clause). Twenty-two of the category labels (*case category labels*) identify the surface cases or the syntactic categories of other components in the sentence. Examples are: SN (subject noun clause), SIN (subject to-infinitive clause), and PN (prepositional phrase [modifying the target verb]).

(4) Constructing the Japanese data

Japanese translations were assigned to each of the English head words and functional words. When translation was not possible simply reading the English sentence, its context was given to the translators. The two lines starting with JAP in Fig. 4 show the translations.

The complete corpus took about 12 man-months of labor to construct. Table 2 shows the corpus statistics for seven verbs. Row (2) shows the percentage of sentences that required the context for translation. This figure indicates the limitations of manual translation without context. Most of these sentences had pronouns like *it* and the translators needed the context to clarify the referents.

19 : " I just know I'm going to take those rubles and build another restaurant, " he said .
ENG : I'm going to take those rubles
CASE : SN<|I|> AX<|be going to|> V<|take|> ON<those |ruble|>
JAP : SN<|私|{は}> AX<|BE GOING TO|> V<|受け取る|> ON<|ルーブル|{を}>

20 : " I take everybody seriously , " Graf said .
ENG : I take everybody seriously
CASE : SN<|I|> V<|take|> ON<|everybody|> DD<|seriously|>
JAP : SN<|私|{は}> V<|受け止める|> ON<|すべての人|{を}> DD<|真剣に|>

<> category label, | | head word, { } functional word

Fig. 4: Part of a Tagged Bilingual Corpus

4.2 Conversion into a PCFT

The bilingual corpus must be converted into a PCFT be-

Table 2: Corpus Statistics [1]

	come	get	give	go	make
(1)	795	867	635	1204	1024
(2)	3.4%	5.2%	4.1%	3.7%	6.6%
(3)	782	849	637	941	1020

	run	take
(1)	440	1062
(2)	6.0%	4.0%
(3)	303	1067

(1) Number of English sentences
(2) Percentage requiring context to translate
(3)Number of obtained quadruplets

fore a case frame can be learned. We can now directly control the information used for learning. We followed the principals below.

- Develop one case-frame tree from each sentence category
  This was intended to observe how the sentence category affects the appearance of case-frame trees.
- Use all case categories in the corpus as attributes
  This was to select effective case categories without any bias.
- Use head words and functional words as values for case categories
  These words are the primary elements representing each case category so it is reasonable to use them as the value.

5. Case-frame Tree Learning Experiments

Several learning experiments were conducted on the PCFT obtained from each sentence category of the target verbs. Complete results from the experiments are not presented here due to space limitations. Table 3 shows the statistical results for seven verbs.

Table 3: Statistics of Case-Frame Trees (from declarative sentences)

(1)	come	get	give	go	make
(2)	398	274	292	225	367
(3)	30	28	31	20	33
(4)	10	9	9	8	8
(5)	6	5	5	6	6
(6)	10.1%	5.5%	13.0%	10.2%	6.2%

(1)	run	take
(2)	68	285
(3)	15	21
(4)	3	10
(5)	3	5
(6)	0.0%	6.0%

(1)Verbs (2) Number of training data
(3)Number of case categories appearing in the PCFT (attribute size)
(4) Number of translations (class size)
(5) Number of case categories appearing in the case-frame tree
(6) Error rate when the tree was used to re-classify the training data

[1] We are now increasing the corpus for *give, make,* and *take* by 4,000 sets.

Translations occurring less than ten times were not included in the PCFT for this experiment. The overall error rate in Table 3 was quite low. Part of the *take* tree is shown in Fig. 5. The figures at the end of each line show the result of the reclassification of the training PCFT by the learned tree: (number of data items which fell on this leaf / number of errors, if any). As is shown, the case-frame tree is highly intelligible.

D<> = over: 引き継ぐ (12.0)
D<> = up: 取る (3.0/1.0)
D<> = 0: *need time*
¦ ON<> = 0: かかる (5.0/1.0) ← A
¦ ON<> = action: とる (8.0)
¦ ON<> = bronze: 獲得する (9.0)
¦ ON<> = hour: かかる (11.0/3.0)
¦ ON<> = measures: とる (10.0) *join*
¦ ON<> = part: 参加する (33.0/1.0) ← B
¦ ON<> = while: かかる (6.0)
¦ ON<> = place:
¦ ¦ SN<> = Sergei Shupletsov: 獲得する (1.0)
¦ ¦ SN<> = attack: 行われる (4.0) *win*
¦ ON<> = time: *happen* C
¦ ¦ AX<> = 0: かかる (4.0/2.0)
¦ ¦ AX<> = may: 必要とする (1.0)
¦ ¦ AX<> = could: かかる (1.0)

Fig. 5: Part of Case-Frame Tree for *Take*

- **Similarity**

The number of case categories actually used in the case-frame tree was drastically smaller than the number used in the PCFT, (row (3) vs. row (5) of Table 3). In the case-frame tree for *take*, for example, the following case categories were used: AX (adverb equivalents), D (adverbial particles), ON (object noun clause), SIN (subject to-infinitive clause), and SN (subject noun clause). The top node, i.e. the most important node, became D, the adverbial particle, following the description in an ordinary dictionary. Most of these syntactical categories are usually used to describe the verb patterns in ordinary dictionaries. The case-frame tree basically duplicates the verb patterns found in an ordinary dictionary.

- **Precision**

From the line marked A in Fig. 5, the translation became kakaru (need time) under the condition of (ON=0) though *take* is usually used as a transitive verb, so the lack of an object noun looks unnatural; this part of the tree, however, corresponds to time expressions like "take long" and "take awhile" which do not have object nouns. This is reasonable learning.

From the line marked B, the idiomatic expression "take part in" was learned as "take part." The word *in* was judged to be redundant and thus an ineffective element. While our corpus did reveal one example that did not have *in* it still had the same translation: "sanka suru." This learning is more precise than the description in an ordinary dictionary.

• Complementary learning

The lines marked C in Fig. 5 show an example of what we call complementary learning. The case-frame tree surprisingly distinguished "kakutoku suru" (win) from "okonawa reru" (happen). The former was learned from "take third place." The latter corresponds to an idiomatic expression, "take place". The way the algorithm learns is unique. The key to discrimination was found in SN, the subject noun, which sounds reasonable. Discrimination is done in terms of the subject's nature: person vs. action noun. However, this could also be distinguished by the existence of the modifier to *place*, since in the idiomatic sense, no modification is allowed between *take* and *place*. In our PCFT, modifiers were not included and the system found complementary knowledge to distinguish the translations. The same phenomenon was found in many parts of the trees. The learning algorithm does its best to sub-categorize the translations within the given case categories. While this can yield linguistically-skewed case frames, they are still effective, at least in the corpus.

• Differences among sentence categories

The results from other sentence categories had a much different appearance. Trees for *make* and *take* which were obtained from the PCFT for the to-infinitive clause contained only one case category, ON (object noun clause). The case categories effective in the declarative sentence, like the adverbial particle, were not effective for this sentence category. This strongly suggests that translations should be selected by using the case frames for the sentence type.

6 Conclusion

We proposed the idea of gradual knowledge acquisition from a bilingual corpus. The knowledge addressed in this paper was the surface verbal case frames for the Japanese translation of English verbs. The process consists of three steps: corpus construction, data conversion, and machine learning.

The case-frame trees we obtained were highly intelligible: they can be interpreted from the linguistic viewpoint. They basically matched linguistic intuition and more precise knowledge was sometimes acquired. Tree analysis showed that in some cases complementary learning occurred even when necessary knowledge was not available.

The trees successfully distinguished the translations of the training data.

Our approach basically fulfills our primary goal: acquiring detailed knowledge and accumulating it in a way that is consistent and unambiguous.

There are several areas for future work. The work in this paper used the word forms as the restrictions for the case categories, resulting in case-frame trees with limited translation power for open data. To increase the translation power, we are generalizing the corpus by using semantic codes and plan to produce case-frame trees with them.

Acknowledgements

I would like to thank Prof. Makoto Nagao of Kyoto University and Prof. Hozumi Tanaka of the Tokyo Institute of Technology for their valuable suggestions. I would also like to thank my supervisors Dr. Yuichi Ninomiya, Dr. Teruaki Aizawa, and Dr. Terumasa Ehara, and my colleagues whose discussions helped clarify this work. The anonymous reviewers made very constructive comments which gave us valuable pointers for our future work.

References

Aizawa, T., Ehara, Uratani and Tanaka (1990). A Machine Translation System for Foreign News in Satellite Broadcasting. *Proc. of Coling-90, Vol. 3*, pp. 308-310.

Cardie, C. (1992). Learning to Disambiguate Relative Pronouns. *Proc. of AAAI-92*, pp. 38-43.

Cardie, C. (1993). A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis. *Proc. of AAAI-93*, pp. 798-803.

Okumura, M. and Tanaka (1990). Towards Incremental Disambiguation with a Generalized Discrimination Network. *Proc. of AAAI-90, Vol. 2*, pp. 990-995.

Quinlan, J, R. (1993). C4.5 programs for machine learning, Morgan Kaufmann.

Tanaka, H. (1991). The MT User Experience. *Proc. of MT Summit III*, pp. 123-125.

Tanaka, H., Aizawa, Kim and Hatada. (1992). A Method of Translating English Delexical Structures into Japanese. *Proc. of Coling-92, Vol. 2*, pp. 567-573.

Tanaka, H. and Ehara (1993). Automatic Verbal Case Frame Acquisition from Bilingual Corpora (in Japanese). *Proc. 47th Annual Convention IPS Japan, Vol. 3*, pp. 195-196.

Utsuro, T., Matsumoto and Nagao.(1993). Verbal Case Frame Acquisition from Bilingual Corpora. *Proc. of the IJCAI-93, Vol. 2*, pp. 1150-1156.