Evaluation of a Method of Creating New Valency Entries

Francis Bond and Sanae Fujita

NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation
2-4 Hikari-dai Seika-cho, Kyoto, Japan 619-0237
{bond, sanae}@cslab.kecl.ntt.co.jp

Abstract

Information on subcategorization and selectional restrictions is important for natural language processing tasks such as deep parsing, rule-based machine translation and automatic summarization. In this paper we present a method of adding detailed entries to a bilingual dictionary, based on information in an existing valency dictionary. The method is based on two assumptions: words with similar meaning have similar subcategorization frames and selectional restrictions; and words with the same translations have similar meanings. Based on these assumptions, new valency entries are constructed from words in a plain bilingual dictionary, using entries with similar source-language meaning and the same target-language translations. We evaluate the effects of various measures of similarity in increasing accuracy.

1 Introduction

Although great progress has been made in learning statistical models from annotated corpora, most machine translation systems rely on detailed information compiled in lexicons. These are typically hand-built (Dorr, 1997). However, adding this detailed information to dictionaries is both time consuming and costly. Several automatic and semi-automatic methods have been proposed to construct lexicons. A common method is to attempt to learn information from corpora (Manning, 1993; Li & Abe, 1998; Kawahara & Kurohashi, 2001). Other work has attempted to extract knowledge from heterogeneous sources, such as existing lexicons (Fujita & Bond, 2002a; Dorr et al., 2002).

Our work differs from corpus-based work such as Manning (1993) or Kawahara & Kurohashi (2001) in that we are using existing lexical resources rather than a corpus. Our method is applicable to rare words, so long as we can find them in a bilingual dictionary, and know the English translation.

In order to demonstrate the utility of the valency information, we give an example of a sentence translated with the system default information (basically a choice between transitive and intransitive), and the full valency information in (1).[1] The verb is 頼む *tanomu* "ask" [NP-*ga* NP-*ni* Cl-*to* V], which takes a clause complement. Without the valency information the translation is incomprehensible: the clause complement is misinterpreted, the zero-pronoun is not resolved and the English to-inifinitive is not produced.

(1) 太郎 は 友達 に 話さ ない
Tarō wa tomodachi ni hanasa nai,
Tarou TOP friend DAT talk not
ように 頼んだ
yōni tanonda
QUOT asked
"Tarou asked his friend not to talk."
with: `Taro asked his friend not to talk.`
without: `As Taro did not talk to his friend, * asked.`

In general, translation tends to simplify text, because the target language will not be able to represent exactly the same shades of meaning as the source text: there is some semantic loss. Therefore, in many cases, a single target language entry is the translation of multiple similar source patterns. For example, there are 23 Japanese predicates linked to the English entry *report* in the valency dictionary used by the Japanese-to-English machine translation system **ALT-J/E**.

In this paper, we extend and re-evaluate the approach proposed by Fujita & Bond (2002a). New en-

[1] We use the following abbreviations: TOP: topic postposition; ACC: accusative postposition; DAT: dative postposition; QUOT: quotative postposition; REC: reciprocal postposition; NP: noun phrase; Cl: clause; V: verb. The sentence is translated using **ALT-J/E** Ikehara et al. (1991).

tries are based on existing entries, so have the same amount of detailed information. The method bootstraps from an initial hand-built lexicon, and allows new entries to be added cheaply and effectively. Although we will use Japanese and English as examples, the algorithm is not tied to any particular language pair or dictionary. The core idea is to add new entries to the valency dictionary by using Japanese-English pairs from a plain bilingual dictionary (without detailed information about valency or selectional restrictions), and build new entries for them based on existing entries.

Fujita & Bond (2002a) showed the approach allowed new patterns to be built at a cost of less than 7 minutes per pattern. An evaluation of 6,893 new patterns showed that adding them to a Japanese-to-English machine translation system improved the translation for 37.5% of sentences using these verbs, and degraded it for 12.6%, a substantial improvement in quality. However, they were unable to fully evaluate the effects of various filters on improving the output quality, such as paraphrasing and using a concept base (Kasahara et al., 1997), because the translation-based evaluation was too indirect.

In this paper, we directly evaluate the approach, by evaluating the created patterns directly. Overall, we are able to confirm the earlier results: high-quality entries can be created cheaply. We also show that evaluating the quality of patterns using paraphrase tests is difficult for non-experts, and ultimately not efficient. Further, we suggest and implement two refinements: creating multiple patterns simultaneously, using information about alternations, and merging similar entries.

The ultimate aim of this research is to identify what kinds of information are most effective in the creation of lexical entries. In particular we wish to discover what is the minimal amount of information necessary to reliably create new entries. Dillinger (2001) criticized previous research presented on lexical construction as paying "more attention to theoretical issues than to establishing effective processes for dictionary development". We try to address both issues here through rigorous evaluation of various methods, with an emphasis on producing usable entries as the final result.

2 The Method of Making New Patterns

The approach is based on that of Fujita & Bond (2002a). It crucially relies on the observation that verbs with similar meanings typically have similar valency structures (Levin, 1993). Given an unknown verb (J_U) which doesn't appear in the valency seed dictionary, if we can find its translation E in a bilingual dictionary and a verb with the same translation exists in the seed dictionary (the known verb J_K), then we assume J_U and J_K are similar in meaning. In this case we can copy the valency information of J_K for J_U. Because the method creates new patterns by copying from the existing pattens, so it's simple and robust.

The method used to determine similarity is translation equivalence: if two verbs have the same translation then they have similar meanings. This has some fundamental problems. Firstly, the set of verbs for which we can create valency patterns is limited. We can only make new entries for words in the bilingual dictionary whose English translations can be found in the valency dictionary. Secondly, it massively overgenerates: one sense of a verb may overlap, but not all will. Further, verbs with similar meanings may have different subcategorization (subcat) and selectional restrictions (SR).

In this work we extend our earlier work in two ways. Firstly we increase the cover by using data about verbal alternations (Levin, 1993). If the known verb, J_K participates in a known alternation then we create new entries based on both alternatives. Secondly, to filter the overgeneration, we investigate merging similar patterns. This is in addition to the existing filters suggested by Fujita & Bond (2002a): a simple human check (pre-filter), paraphrasing and association scores.

Section 2.1 recaps the basic method of constructing methods given in Fujita & Bond (2002a). Section 2.1.1 introduces the use of alternations to create more patterns, while Section 2.1.2 presents the use of merging to reduce redundant patterns.

2.1 Constructing Candidates

As a seed dictionary we use the verbs from **ALT-J/E**'s valency dictionary (Ikehara et al., 1991), ignoring all idiomatic and adjectival entries — this gave 5,062 verbs and 11,214 valency patterns (2.2 patterns/verb). Each **pattern** consists of source (Japanese) and target (English) language subcategorization information and selectional restrictions on the source side. Each argument on the Japanese side consists of head-word, a case-role, a list of postpositions and a list of selectional restrictions. There is also other information about aspectual class, ver-

bal semantic attributes and so on, which we will not discuss here, although it is included in the entries we create. Selectional restrictions are given as either nodes in the GoiTaikei thesaurus (3,710 semantic classes; Ikehara et al. (1997)) or strings. It takes an expert lexicographer an average of 30 minutes to create one entry from scratch.

To find translation equivalences, we used a plain bilingual dictionary which contains word pairs without valency information. This was made from **ALT-J/E**'s Japanese-English word transfer dictionary and an enhanced version of EDICT (Breen, 1995) where Japanese verbal-nouns were expanded into verbs (.e.g, 共同 *kyōdō* "cooperation" was expanded into 共同する *kyōdō-suru* "cooperate").

To create a candidate J_U, an Unknown word for which we have no valency information, we find all words where E, the English translation (or translations) is linked to one or more valency patterns J_K in the valency dictionary. Figure 1 shows the overall flow of creating new patterns. The only step which is not fully automatic is the pre-filter, which is done by an analyst (§ 2.2.1).

For each entry in the plain J-E dictionary

- If no entries with the same Japanese (J_U) exist in the valency dictionary
  - For each valency entry (J_K) with the same English (E)
    - Create a candidate pattern consisting of J_K replaced by J_U (§ 2.1)

For each candidate pattern J_U-E (from J_K-E)

1. If J_U is obviously different to J_K **`reject`** (§ 2.2.1) [human judgment]
2. If J_K-E has an alternation J_A-E_A also create candidate J_U-E_A (§ 2.1.1)
3. Merge very similar candidate patterns (§ 2.1.2)

Figure 1: Creating New Patterns

2.1.1 Adding Alternative Patterns

If the entry in the seed valency dictionary participates in a diathesis alternation (such as *I broke the cup* ⇔ *The cup broke*), then we create candidates for both alternatives at once.

For example, the unknown verb 着火する *chakka-suru* "ignite" matches 引火する *inka-suru* "ignite" which has two alternatives in the seed dictionary linked by the Causative/Inchoative Alternation. We make patterns for both of them, allowing us to match both (2) and (3).

(2) 導火線 が 着火した。
doukasen ga chakka-shita.
fuse ACC ignited
The fuse ignited.

(3) 彼 は 導火線 に 着火した。
kare wa doukasen ni chakka-shita.
He TOP fuse DAT ignited.
He ignited the fuse.

This can only be done if the seed dictionary contains information about alternations, but currently much research is being done to identify them and add them to lexicons, both by linguists (Furumaki & Tanaka, 2003) and computational linguists (Bond et al., 2002; McCarthy, 2000).

2.1.2 Method of Merging Patterns

Merging similar candidates is an important problem for corpus-based approaches, which normally have 10s to 1000s of candidates to merge (Li & Abe, 1998; McCarthy, 2000). In our case we have fewer candidates, and they have more information. Although the existence of very similar patterns does not effect the translation quality, the redundancy creates spurious ambiguity, which slows the system down and makes debugging harder.

We reduce the number of redundant patterns by merging similar entries. First, if two patterns were identical, we merge them. We then merge candidates that only differ in their postpositions and selectional restrictions. That is, they have the same Japanese head-word, the same English head-word, the same English subcat, the same number of arguments, and the same case-roles. If the entries have different postpositions, the merged entry is given the union of the two sets (for example if the argument of J_{U1} has {に} *ni* "to", and the argument of J_{U2} has {に,へ} *ni,e* "to", then the merged entry will have {に, へ} *ni,e* "to" as its case markers. However, if one of the similar entries is from a domain-specific

dictionary, it is rejected in favor of the entry from the general dictionary, rather than being merged.

We tested two strategies for merging selectional restrictions: **parent** and **child**. All pairs of selection restrictions from the two patterns are compared. In **parent**, if one restriction subsumes the other the least restrictive (the parent) is used. In **child**, the most restrictive (the child) is used. If neither restriction subsumes the other, then both are used. Multiple patterns can be merged, not only pairs of similar patterns.

2.2 Filtering Candidates

In order to filter out bad candidates, we investigate several other methods of judging similarity.

2.2.1 Pre-filter

The simplest method is to use human judgment. In the pre-filter, an analyst examines the two source language words (J_U, J_K) linked by an English translation, and rejects them if they do not have a similar meaning. Many words that are obviously dissimilar are linked due to the polysemy of the English pivot. Rejecting them is a very fast process. It only becomes slow if the analyst does not recognize one of the verbs and therefore has to look it up. The strength of this method is its accuracy, the weakness is that it requires human intervention, and is thus expensive.

2.2.2 Paraphrasing

The aim of filtering using paraphrases, first proposed by Fujita & Bond (2002b), is to eliminate candidate patterns with incorrect subcats. The method we use is described in detail in Fujita & Bond (2002a).

The filtering is done by an analyst, but it is claimed that they do not have to be an expert, just a native speaker. The analyst judges whether sentences with the candidate verb (J_U) replaced by the seed verb (J_K) (and vice-versa) are grammatical or not. Ideally, words with the same subcat will produce grammatical a paraphrase, while those with different subcats will not.

For example, both 結婚する *kekkon-suru* "marry" (J_K) and 嫁ぐ *totsugu* "marry into"(J_U) have similar meanings. But 結婚する *kekkon-suru* "marry" is a reciprocal verb: "a man and a woman marry", 嫁ぐ *totsugu* "marry into" on the other hand is directional, "a woman marries a man/into a family" and thus the subcat is different. This can be seen in (4) and (5), where 結婚する *kekkon-suru* "marry" is replaced with 嫁ぐ*totsugu* "marry into", but (5) is ungrammatical.

(4) 彼女 は 彼 と 結婚する。
kanojo wa kare to kekkon-suru
she TOP him REC marry
"She'll marry (with) him."

(5) * 彼女 は 彼 と 嫁ぐ。
kanojyo wa kare to totsugu

Two judgments are made for each paraphrase pair: is the paraphrase grammatical, and if it is grammatical, are the meanings similar? The three grammaticality classes are: **`grammatical`**, **`ungrammatical`**, **`grammatical in some context`**. Semantic similarity was divided into three major classes. In **`same or close`** the paraphrased sentence has almost the same meaning as the original; in **`different nuance`** the meaning is significantly broader or narrower, or only the same in certain contexts; in **`different`** the meaning changes in the paraphrased examples.

The strength of this method is that non-experts can make the judgments and there is supporting data for them. The weaknesses are that it requires example sentences and is labor intensive.

2.2.3 Association Scores

We also tested the use of association scores to measure similarity. This measure is designed to simulate human word association. We used the concept base built by Kasahara et al. (1997) where scores are calculated using word-vector distances taken from word-definitions and corpora. The measure itself is hard to use directly, but it can be used to rank words in order of similarity. We investigated only creating patterns for the most similar word, and for words that were within the top 10, 100, 1,000 and 10,000 most similar words. The strength of this method is that it is fully automatic. The weakness is that highly associated words are not necessarily syntactically or semantically similar (for example 結婚する *kekkon-suru* "marry" and 嫁ぐ *totsugu* "marry into").

2.2.4 Translation Link

We also evaluated the quality of the translation link used to create the candidates. If J_U has English translations $E(J_U)$, and they link through the valency dictionary to a Japanese word J_V with translations $E(J_K)$, then we calculate the strength of the

link using Dice's coefficient:

$$\text{link strength} = \frac{2 \times (|E(J_U) \cap E(J_K)|)}{|E(J_U)| + |E(J_K)|} \quad (1)$$

The strength of this method is that it is fully automatic. The weakness is that it depends entirely on the quality of the bilingual lexicon.

3 Experiment

There were 4,129 verbs in the bilingual dictionary where the Japanese had no entry in the valency dictionary, but the English did. We were able to find examples for 3,753 of these (90.9%), taken from a corpus of nine years of newspaper text. In order to test the paraphrase filter, candidate patterns were only made for those verbs for which we could find examples.

For the 3,753 target verbs, we did the check using the pre-filter and paraphrasing. The original number of candidates was enormous: 108,733 pairs of J_U and J_K. Most of these were removed in the prefiltering stage, leaving 2,570 unknown verbs matching 6,888 verbs in the valency dictionary (in fact, as the pre-filter check doesn't need the valency patterns, they can be made after this stage). When these were expanded into patterns, they made a total of 8,553 candidate patterns (3.3 patterns/verb).

An additional 175 patterns were made using alternations. Six were subsequently merged.

We were able to merge 2,934 similar patterns into 1,188, leaving 6,305 candidate patterns. The maximum number of patterns merged into one was nine (勘違いする *kanchigai-suru* "mistake"). Half the mergers used the **parent** method and half used the **child** method (§ 2.1.2). In the merging, 50% of the patterns had postpositions merged. After merging, there were 2.5 patterns/verb, a much closer ratio to that of the seed lexicon.

The created patterns were then checked using paraphrases as described in 2.2.2, which took the analysts about 7 minutes per verb. The data was split between three analysts, one a linguist and two people with no special training.

As a final evaluation, all the candidates which passed the pre-filter were checked directly by lexicographers who were familiar with the seed lexicon (not the authors). This took around 5 minutes per verb. Each pattern was marked as: **`acceptable`**, **`fixable`** or **`useless`**: **`acceptable`** patterns could be used as is; **`fixable`** patterns could be used with minor revisions; **`useless`** patterns were so poor that it would be easier to create an entry from scratch.

4 Results and Evaluation

In this section we present the results of the direct evaluation, and use them to see the worth of the various filters. Finally we compare our results to the translation-based evaluation of Fujita & Bond (2002a).

4.1 Direct Evaluation

The results of the direct analysis are given in Table 1. Separate columns are shown for patterns made using alternations, patterns that underwent merging using the parent method, patterns that underwent merging using the child method, the remainder of patterns and all patterns. The final column shows the results with the scores for merged patterns split among the original patterns. These results are used to evaluate the similarity filters.

The majority of patterns that passed the prefilter were usable as is (51.8%). A further 36.3% were usable with minor revisions, giving 88.1% potentially useful patterns. These are encouraging results.

Patterns made using the alternations were worse overall, while those made by merging were substantially better. One of the reasons for the poor quality of the alternations is that they added another transformation to the original. If we consider only alternations of acceptable patterns, then they are acceptable 44% of the time. Therefore, it is better to make patterns using alternations after all other filters have been applied.

Fewer fixes were necessary for the patterns merged with more general restrictions (**parent:child** — 62.1%:56.7%), although both were better than the remainder. Examining the kinds of changes needed by the merged patterns showed the child set needed their selectional restrictions corrected more often. This shows clearly that merging to the least restrictive values (the parent strategy) is the best.

4.2 Evaluation using Translation

Fujita & Bond (2002a) reported a translation-based evaluation of the effect on translation quality of new patterns, without merging, that had at least one paraphrase that was **`grammatical`**. There were 6,893 new patterns, for 2,305 kinds of verbs (3.0 patterns/verb). For each verb (J_U) they picked two shortish sentences (average length 81.8 characters

Table 1: Evaluations

	Alternation		Merge-Parent		Merge-Child		Remainder		Total		Expanded	
Result	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Acceptable	53	31.4	369	62.1	337	56.7	2505	50.6	3264	51.8	4273	53.4
Fixable	61	36.1	196	33.0	231	38.9	1803	36.4	2291	36.3	2899	36.2
Useless	55	32.5	29	4.9	26	4.4	640	12.9	750	11.9	829	10.4
Total	169	100	594	100	594	100	4948	100	6305	100	8008	100.0

(40 words)/sentence) from a corpus of newspaper text which had not been used in the paraphrasing stage.[2] This gave a total of 4,367 test sentences.

Translations were compared with and without the valency patterns. There were two set-ups. In the first, each pattern was added to the valency dictionary individually, to get a score for each pattern. Thus verbs with more than one pattern would be tested multiple times. In the second, all the patterns were added together, and the system selected the most appropriate pattern using the valency information and selectional restrictions (the normal way to use the lexicon). Translations were judged to be either: **`no change`**, **`improved`**, **`equivalent`** or **`degraded`**. The results of their evaluation are given in Table 2.

Table 2: Evaluation of New Valency Entries

	Each pattern		All patterns	
Judgment	No.	%	No.	%
`improved`	4,536	34.5	1,636	37.5
`no change`	3,238	24.6	1,063	24.3
`equivalent`	3,465	26.4	1,115	25.5
`degraded`	1,901	14.5	552	12.6
Total	13,140	100.0	4,366	100.0

Most sentences **`improved`** translation quality, followed by **`equivalent`** or **`no change`**. Few translations were **`degraded`**. Using only the pre-filter and the grammaticality judgments, 37.5% of translations improved and only 12.6% degraded, an overall improvement of 24.9%.

Fujita & Bond (2002a) also reported that using only the pre-filter gave an improvement of 32% versus a degradation of 16% (tested on each pattern individually), which should improve further if all patterns are tested together.

[2]Only one sentence could be found for some verbs.

4.3 Evaluation of the Filters

In this section we evaluate the effectiveness of the filters, using the check by the expert lexicographers as our gold standard. The results of several methods are given in Table 3. Thresholds were chosen after examining the data over a wide range of values, although we don't show all the results here.

Precision is the percentage of entries that passed the filter and were rated acceptable. Recall is the percentage of acceptable entries that passed the filter (from a total of 4,273: these scores are calculated using the expanded (unmerged) set of patterns). The baseline is to use all patterns that passed the pre-filter: this gives a precision of 53.4% and 100% recall.

The highest precision (72.3%) came from only using entries where the unknown verb (J_U) was the most similar to the known verb (J_K). The recall, however is a disappointing 3%. Using the paraphrase tests based on sentences where the unknown verb replaced the known verb, gave almost as high a precision with a higher recall (71.8% and 23.7% respectively).

We also attempted using a learner (C5.0) on the results of the various filters. It gave almost no improvement, separating patterns into acceptable vs the rest with an accuracy of only 52.8% (average over ten-fold cross-validation). However, when tested separately on the data from the three different paraphrase evaluators, the linguist's results were significantly better, leading to a discrimination accuracy of 59.1%.

4.4 Relationship between Two Evaluations

We compare the results of the direct evaluation with the translation-based evaluation of Fujita & Bond (2002a) in Table 4. Here **good** are patterns where the translation either improved or stayed the same. **equal** is used for patterns where one translation im-

Table 3: Filter Effectiveness

Filter	Cutoff	Precision (%)	Recall (%)	F-score (%)
Association Score	1st ranked	72.3	3.0	5.8
Translation Link	score ≥ 0.9	59.6	7.1	12.7
$J_U \Rightarrow J_K$: grammatical	$\% \geq 90$	57.1	61.0	59.0
$J_U \Rightarrow J_K$: ungrammatical	$\% = 0$	57.1	61.3	59.1
$J_U \Rightarrow J_K$: same or close	$\% \geq 90$	70.2	22.0	33.5
$J_K \Rightarrow J_U$: grammatical	$\% \geq 90$	61.7	55.1	58.2
$J_K \Rightarrow J_U$: ungrammatical	$\% = 0$	61.7	55.7	58.5
$J_K \Rightarrow J_U$: same or close	$\% \geq 90$	71.8	23.7	35.6
Pre-filter only		53.4%	100.0%	53.5%

proved and one degraded and **bad** where one degraded and one stayed the same or both degraded.

Table 4: Relationship between Direct Evaluation and Translation-based Evaluation

Direct	Translation-based Evaluation						
Evaluation	Good	%	Equal	%	Bad	%	Total
Acceptable	3174	76	507	12	496	12	4177
Fixable	1906	69	492	18	368	13	2766
Useless	500	65	152	20	115	15	767
Total	5624	72	1173	15	991	13	7788

There is general agreement, but it is not exact. One reason for this is that some bad patterns provided a translation where the system without the entry had none. This generally improves the translation quality, even if the subcat is wrong. The direct evaluation gave a clearer indication of the utility of the features than the translation-based one.

5 Discussion

The evaluation shows two things. The first is the utility of merging similar patterns: the resulting patterns are of high quality, and the dictionary becomes more compact. When merging, the best strategy is to create new patterns with less restrictive selectional restrictions. The second is that evaluation by paraphrasing is no better than using expert lexicographers. Although using paraphrase data does improve the quality of the dictionary, it is quicker (5 minutes vs 7 minutes) and more accurate to use lexicographers directly. Further, paraphrase judgments are hard to make for untrained analysts: linguists made paraphrase judgments with higher accuracy. This falsifies the claim that paraphrase judgments can be done cheaply with untrained analysts, and removes another incentive to use paraphrasing as a filter.

From a practical point of view the results are encouraging: we can produce useful new patterns with only a simple monolingual judgment as pre-filter: "are these verbs similar in meaning?", and it has been shown that these patterns improve the quality of translation in 32% of sentences versus degradations in only 16%.

The quality can further be improved by the candidates being checked by lexicographers. This is expensive, at an additional 5 minutes per pattern, but is still cheaper than creating patterns from scratch. Preliminary investigation shows that even correcting the fixable entries takes less than 10 additional minutes per entry on average, for a total of 15 minutes per entry.

Overall, our results show that hand-compilation is still necessary for building high quality lexicons. However, semi-automatic acquisition of candidates, and merging the acquired candidates can increase efficiency considerably.

We therefore propose a method of building information-rich lexicons that proceeds as follows: (1) build a seed lexicon by hand; (2) extend it semi-automatically using bilingual lexicons and a simple pre-filter check; (3) merge any similar entries, making the selectional restrictions broader rather than narrower; (4) revise the new entries as far as possible.

This method is also applicable to work in new language pairs. It will always be the case that simple bilingual lexicons are larger than information-rich lexicons — therefore it will be worthwhile using the former to extend the latter.

Our work is similar to that of Dorr et al. (2002), who link two information-rich resources (one English and one Chinese) using a bilingual dictionary. They then use the bilingual dictionary to fill in gaps, effectively using a simpler resource to increase the size of the information-rich lexicons.

6 Conclusion

In this paper we present a method of assigning valency information and selectional restrictions to entries in a bilingual dictionary. The method exploits existing dictionaries and is based on two basic assumptions: words with similar meaning have similar subcategorization frames and selectional restrictions; and words with the same translations have similar meanings.

A prototype system allowed new patterns to be built, with some human intervention, at a cost of around 6 minutes per pattern. Our evaluation showed that fully automatic creation of high quality patterns is still beyond our reach.

Acknowledgments

The authors would like to thank the other members of the NTT Machine Translation Research Group, Satoshi Shirai and Timothy Baldwin. This research was supported by the research collaboration between the NTT Communication Science Labs and CSLI, Stanford University.

References

Bond, Francis, Timothy Baldwin & Sanae Fujita: 2002, 'Detecting alternation instances in a valency dictionary', in *8th Annual Meeting of the Association for Natural Language Processing*, The Association for Natural Language Processing, pp. 519–522.

Breen, Jim: 1995, 'Building an electronic Japanese-English dictionary', Japanese Studies Association of Australia Conference (`http://www.csse.monash.edu.au/~jwb/jsaa_paper/hpaper.html`).

Dillinger, Mike: 2001, 'Dictionary development workflow for MT: Design and management', in *MT Summit VIII*, Santiago de Compostela, pp. 83–88.

Dorr, Bonnie J.: 1997, 'Large-scale dictionary construction for foreign language tutoring and interlingual machine translation', *Machine Translation*, **12**(4): 271–322.

Dorr, Bonnie J., Gina-Anne Levow & Dekang Lin: 2002, 'Construction of a Chinese-English verb lexicon for machine translation', *Machine Translation*, **17**(1–2), (in press).

Fujita, Sanae & Francis Bond: 2002a, 'Extending the coverage of a valency dictionary', in *COLING-2002 workshop on Machine Translation in Asia*, Taipei, pp. 67–73.

Fujita, Sanae & Francis Bond: 2002b, 'A method of adding new entries to a valency dictionary by exploiting existing lexical resources', in *Ninth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2002*, Keihanna, Japan, pp. 42–52.

Furumaki, Hisanori & Hozumi Tanaka: 2003, 'The consideration of <n-suru> for construction of the dynamic lexicon', in *9th Annual Meeting of The Association for Natural Language Processing*, pp. 298–301, (in Japanese).

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

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa: 1991, 'Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–', in *Third Machine Translation Summit: MT Summit III*, Washington DC, pp. 101–106, (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

Kasahara, Kaname, Kazumitsu Matsuzawa & Tsutomu Ishikawa: 1997, 'A method for judgment of semantic similarity between daily-used words by using machine readable dictionaries', *Transactions of IPSJ*, **38**(7): 1272–1283, (in Japanese).

Kawahara, Daisuke & Sadao Kurohashi: 2001, 'Japanese case frame construction by coupling the verb and its closest case component', in *Proceedings of First International Conference on Human Language Technology Research (HLT 2001)*, San Diego, pp. 204–210.

Levin, Beth: 1993, *English Verb Classes and Alternations*, Chicago, London: University of Chicago Press.

Li, Hang & Naoki Abe: 1998, 'Generalizing case frames using a thesaurus and the MDL principle', *Computational Linguistics*, **24**(2): 217–244.

Manning, Christopher D.: 1993, 'Automatic acquisition of a large subcategorization dictionary from corpora', in *31st Annual Meeting of the Association for Computational Linguistics: ACL-93*, pp. 235–242.

McCarthy, Diana: 2000, 'Using semantic preferences to identify verbal participation in role switching alternations', in *Proceedings of the first Conference of the North American Chapter of the Association for Computational Linguistics. (NAACL)*, Seattle, WA.