The Automatic Translation of Discourse Structures

Daniel Marcu
Information Sciences Institute and
Department of Computer Science
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
marcu@isi.edu

Lynn Carlson
U.S. Department of Defense
Ft. Meade, MD 20755
lmcarls@afterlife.ncsc.mil

Maki Watanabe
Department of Linguistics
University of Southern California
Los Angeles, CA 90089
mwatanab@usc.edu

Abstract

We empirically show that there are significant differences between the discourse structure of Japanese texts and the discourse structure of their corresponding English translations. To improve translation quality, we propose a computational model for rewriting discourse structures. When we train our model on a parallel corpus of manually built Japanese and English discourse structure trees, we learn to rewrite Japanese trees as trees that are closer to the natural English rendering than the original ones.

1 Motivation

Almost all current MT systems process text one sentence at a time. Because of this limited focus, MT systems cannot re-group and re-order the clauses and sentences of an input text to achieve the most natural rendering in a target language. Yet, even between languages as close as English and French, there is a 10% mismatch in number of sentences — what is said in two sentences in one language is said in only one, or in three, in the other (Gale and Church, 1993). For distant language pairs, such as Japanese and English, the differences are more significant.

Consider, for example, Japanese sentence (1), a word-by-word "gloss" of it (2), and a two-sentence translation of it that was produced by a professional translator (3).

> [厚生省が昨年公表した^1] [人口の将来推計では、^{2}] [将来、一・四九九人を最低に、^{3}] [その後は上昇に転ずると^4] [推計していたが、^{5}] [早くも予想がはずれる^6] [見通しとなった。^{7}] (1)

> [The Ministry of Health and Welfare last year revealed^1] [population of future estimate according to^2] [in future 1.499 persons as the lowest^3] [that after *SAB* rising to turn that^4] [*they* estimated but^5] [already the estimate misses a point^6] [prediction became.^{7}] (2)

> [In its future population estimates^1] [made public last year,^{2}] [the Ministry of Health and Welfare predicted that the SAB would drop to a new low of 1.499 in the future,^{3}] [but would make a comeback after that,^{4}] [increasing once again.^{5}] [However, it looks as if that prediction will be quickly shattered.^{6}] (3)

The labeled spans of text represent elementary discourse units (*edus*), i.e., minimal text spans that have an unambiguous discourse function (Mann and Thompson, 1988). If we analyze the text fragments closely, we will notice that in translating sentence (1), a professional translator chose to realize the information in Japanese unit 2 first (unit 2 in text (1) corresponds roughly to unit 1 in text (3)); to realize then some of the information in Japanese unit 1 (part of unit 1 in text (1) corresponds to unit 2 in text (3)); to fuse then information given in units 1, 3, and 5 in text (1) and realize it in English as unit 3; and so on. Also, the translator chose to repackage the information in the original Japanese sentence into two English sentences.

At the elementary unit level, the correspondence between Japanese sentence (1) and its English translation (3) can be represented as in (4), where $j \subset e$ denotes the fact that the semantic content of unit j is realized fully in unit e; $j \supset e$ denotes the fact that the semantic content of unit e is realized fully in unit j; $j = e$ denotes the fact that units j and e are semantically equivalent; and $j \cong e$ denotes the fact that there is a semantic overlap between units j and e, but neither proper inclusion nor proper equivalence.

$$\begin{array}{l} j_1 \supset e_2; j_1 \cong e_3; \\ j_2 = e_1; \\ j_3 \subset e_3; \\ j_4 \cong e_4; j_4 \cong e_5; \\ j_5 \cong e_3; \\ j_6 \subset e_6; \\ j_7 \subset e_6 \end{array} \tag{4}$$

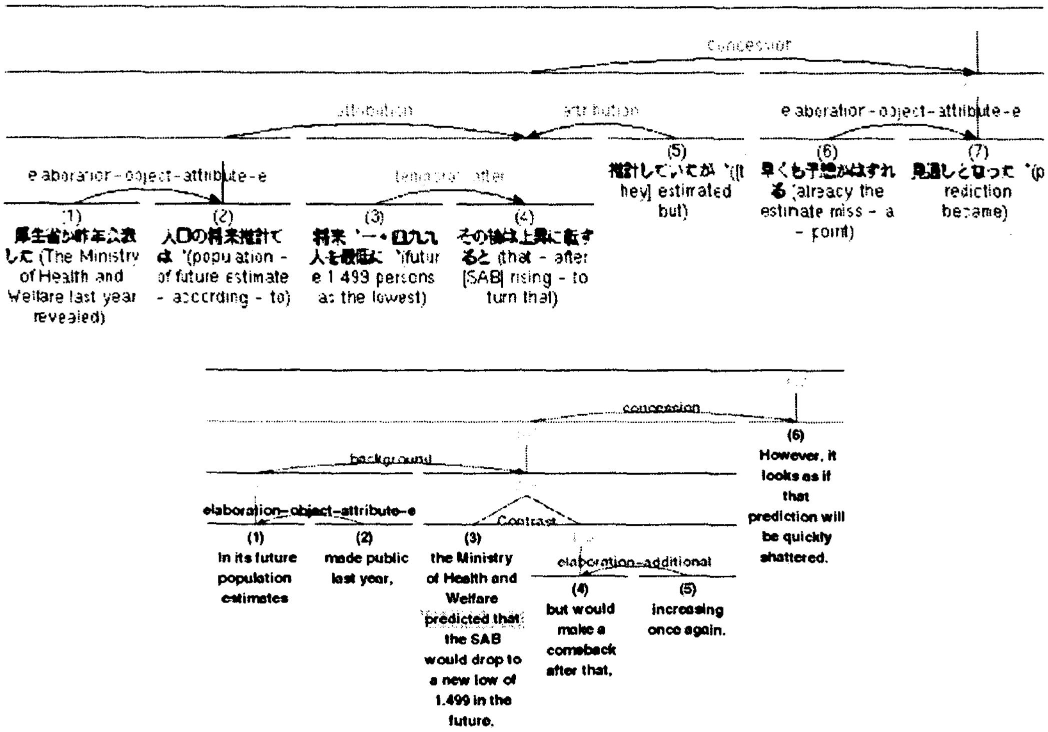
concession
attribution
attribution
elaboration-object-attribute-e
elaboration-object-attribute-e
temporal-after
(1)
厚生省が昨年公表した (The Ministry of Health and Welfare last year revealed)
(2)
人口の将来推計では '(population - of future estimate - according - to)
(3)
将来 '一・四九九人を最低に '(future 1 499 persons as the lowest)
(4)
その後は上昇に転ずると (that - after [SAB] rising - to turn that)
(5)
推計していたが '([they] estimated but)
(6)
早くも予想がはずれる (already the estimate miss - a - point)
(7)
見通しとなった '(prediction became)
concession
background
elaboration-object-attribute-e
Contrast
elaboration-additional
(1)
In its future population estimates
(2)
made public last year,
(3)
the Ministry of Health and Welfare predicted that the SAB would drop to a new low of 1.499 in the future,
(4)
but would make a comeback after that,
(5)
increasing once again.
(6)
However, it looks as if that prediction will be quickly shattered.

Figure 1: The discourse structures of texts (1) and (3).

Hence, the mappings in (4) provide an explicit representation of the way information is re-ordered and re-packaged when translated from Japanese into English. However, when translating text, it is also the case that the rhetorical rendering changes. What is realized in Japanese using an CONTRAST relation can be realized in English using, for example, a COMPARISON or a CONCESSION relation.

Figure 1 presents in the style of Mann and Thompson (1988) the discourse structures of text fragments (1) and (3). Each discourse structure is a tree whose leaves correspond to the *edus* and whose internal nodes correspond to contiguous text spans. Each node is characterized by a *status* (NUCLEUS or SATELLITE) and a *rhetorical relation*, which is a relation that holds between two non-overlapping text spans. The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's intention than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa. When spans are equally important, the relation is multinuclear: for example, the CONTRAST relation that holds between unit [3] and span [4,5] in the rhetorical structure of the English text in figure 1 is multinuclear. Rhetorical relations that end in the suffix "-e" denote relations that correspond to embedded syntactic constituents. For example, the ELABORATION-OBJECT-ATTRIBUTE-E relation that holds between units 2 and 1 in the English discourse structure corresponds to a restrictive relative.

If one knows the mappings at the *edu* level, one can determine the mappings at the span (discourse constituent) level as well. For example, using the elementary mappings in (4), one can determine that Japanese span [1,2] corresponds to English span [1,2], Japanese unit [4] to English span [4,5], Japanese span [6,7] to English unit [6], Japanese span [1,5] to English span [1,5], and so on. As Figure 1 shows, the CONCESSION relation that holds between spans [1,5] and [6,7] in the Japanese tree corresponds to a similar relation that holds between span [1,5] and unit [6] in the English tree (modulo the fact that, in Japanese, the relation holds between sentence fragments, while in English it holds between full sentences). However, the TEMPORAL-AFTER relation that holds between units [3] and [4] in the Japanese tree is realized as a CONTRAST relation between unit [3] and span [4,5] in the English tree. And because Japanese units [6] and [7] are fused into unit [6] in English, the relation ELABORATION-OBJECT-ATTRIBUTE-E is no longer made explicit in the English text.

Some of the differences between the two discourse trees in Figure 1 have been traditionally addressed

Corpus	k_u (#)	k_s (#)	k_n (#)	k_r (#)
Japanese	0.856 (80)	0.785 (3377)	0.724 (3377)	0.650 (3377)
English	0.925 (60)	0.866 (1826)	0.839 (1826)	0.748 (1826)

Table 1: Tagging reliability

in MT systems at the syntactic level. For example, the re-ordering of units 1 and 2, can be dealt with using only syntactic models. However, as we will see in Section 2, there are significant differences between Japanese and English with respect to the way information is packaged and organized rhetorically not only at the sentence level, but also, at the paragraph and text levels. More specifically, as humans translate Japanese into English, they *re-order* the clauses, sentences, and paragraphs of Japanese texts, they *re-package* the information into clauses, sentences, and paragraphs that *are not* a one-to-one mapping of the original Japanese units, and they *rhetorically re-organize* the structure of the translated text so as to reflect rhetorical constraints specific to English. If a translation system is to produce text that is not only grammatical but also coherent, it will have to ensure that the discourse structure of the target text reflects the natural renderings of the target language, and not that of the source language.

In Section 2, we empirically show that there are significant differences between the rhetorical structure of Japanese texts and their corresponding English translations. These differences justify our investigation into developing computational models for discourse structure rewriting. In Section 3, we present such a rewriting model, which re-orders the *edus* of the original text, determines English-specific clause, sentence, and paragraph boundaries, and rebuilds the Japanese discourse structure of a text using English-specific rhetorical renderings. In Section 4, we evaluate the performance of an implementation of this model. We end with a discussion.

2 Experiment

In order to assess the role of discourse structure in MT, we built manually a corpus of discourse trees for 40 Japanese texts and their corresponding translations. The texts were selected randomly from the ARPA corpus (White and O'Connell, 1994). On average, each text had about 460 words. The Japanese texts had a total of 335 paragraphs and 773 sentences. The English texts had a total of 337 paragraphs and 827 sentences.

We developed a discourse annotation protocol for Japanese and English along the lines followed by Marcu et al. (1999). We used Marcu's discourse annotation tool (1999) in order to manually construct the discourse structure of all Japanese and English texts in the corpus. 10% of the Japanese and English texts were rhetorically labeled by two of us. The tool and the annotation protocol are available at *http://www.isi.edu/~marcu/software/*. The annotation procedure yielded over the entire corpus 2641 Japanese *edus* and 2363 English *edus*.

We computed the reliability of the annotation using Marcu et al. (1999)'s method for computing kappa statistics (Siegel and Castellan, 1988) over hierarchical structures. Table 1 displays average kappa statistics that reflect the reliability of the annotation of elementary discourse units, k_u, hierarchical discourse spans, k_s, hierarchical nuclearity assignments, k_n, and hierarchical rhetorical relation assignments, k_r. Kappa figures higher than 0.8 correspond to good agreement; kappa figures higher than 0.6 correspond to acceptable agreement. All kappa statistics were statistically significant at levels higher than $\alpha = 0.01$. In addition to the kappa statistics, table 1 also displays in parentheses the average number of data points per document, over which the kappa statistics were computed.

For each pair of Japanese-English discourse structures, we also built manually an alignment file, which specified in the notation discussed on page 1 the correspondence between the *edus* of the Japanese text and the *edus* of its English translation.

We computed the similarity between English and Japanese discourse trees using labeled recall and precision figures that reflected the resemblance of the Japanese and English discourse structures with respect to their assignment of *edu* boundaries, hierarchical spans, nuclearity, and rhetorical relations.

Because the trees we compared differ from one language to the other in the number of elementary units, the order of these units, and the way the units are grouped recursively into discourse spans, we computed two types of recall and precision figures. In computing *Position-Dependent* (P-D) recall and precision figures, a Japanese span was considered to match an English span when the Japanese span contained all the Japanese *edus* that corresponded to the *edus* in the English span, and when the Japanese and English spans appeared in the same position with respect to the overall structure. For example, the English tree in figure 1 is characterized by 10 subsentential spans: [1], [2], [3], [4], [5], [6], [1,2], [4,5], [3,5], and [1,5]. (Span [1,6] subsumes 2 sentences, so it is not sub-sentential.) The Japanese discourse tree has only 4 spans that could be matched in the same positions with English spans, namely spans [1,2], [4], [5], and [1,5]. Hence the similarity between the Japanese tree and the English tree with respect

Level	Units		Spans		Status/Nuclearity		Relations	
	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P
Sentence	29.1	25.0	27.2	22.7	21.3	17.7	14.9	12.4
Paragraph	53.9	53.4	46.8	47.3	38.6	39.0	31.9	32.3
Text	41.3	42.6	31.5	32.6	28.8	29.9	26.1	27.1
Weighted Average	36.0	32.5	31.8	28.4	26.0	23.1	20.1	17.9
All	8.2	7.4	5.9	5.3	4.4	3.9	3.3	3.0
	P-I R	P-I P	P-I R	P-I P	P-I R	P-I P	P-I R	P-I P
Sentence	71.0	61.0	56.0	46.6	44.3	36.9	30.5	25.4
Paragraph	62.1	61.6	53.2	53.8	43.3	43.8	35.1	35.5
Text	74.1	76.5	54.4	56.5	48.5	50.4	41.1	42.7
Weighted Average	69.6	63.0	55.2	49.2	44.8	39.9	33.1	29.5
All	74.5	66.8	50.6	45.8	39.4	35.7	26.8	24.3

Table 2: Similarity of the Japanese and English discourse structures

to their discourse structure below the sentence level has a recall of 4/10 and a precision of 4/11 (in Figure 1, there are 11 sub-sentential Japanese spans).

In computing *Position-Independent* (P-I) recall and precision figures, even when a Japanese span "floated" during the translation to a position in the English tree that was different from the position in the initial tree, the P-I recall and precision figures were not affected. The Position-Independent figures reflect the intuition that if two trees t_1 and t_2 both have a subtree t, t_1 and t_2 are more similar than if they were if they didn't share any tree. At the sentence level, we hence assume that if, for example, the syntactic structure of a relative clause is translated appropriately (even though it is not appropriately attached), this is better than translating wrongly that clause. The Position-Independent figures offer a more optimistic metric for comparing discourse trees. They span a wider range of values than the Position-Dependent figures, which enable a finer grained comparison, which in turn enables a better characterization of the differences between Japanese and English discourse structures. When one takes an optimistic stance, for the spans at the sub-sentential level in the trees in Table 1 the recall is 6/10 and the precision is 6/11 because in addition to spans [1,2], [4], [5], and [1,5], one can also match Japanese span [1] to English span [2] and Japanese span [2] to Japanese span [1].

In order to provide a better estimate of how close two discourse trees were, we computed Position-Dependent and -Independent recall and precision figures for the sentential level (where units are given by *edus* and spans are given by sets of *edus* or single sentences); paragraph level (where units are given by sentences and spans are given by sets of sentences or single paragraphs); and text level (where units are given by paragraphs and spans are given by sets of paragraphs). These figures offer a detailed picture of how discourse structures and relations are mapped from one language to the other across all discourse levels, from sentence to text. The differences at the sentence level can be explained by differences between the syntactic structures of Japanese and English. The differences at the paragraph and text levels have a purely rhetorical explanation.

As expected, when we computed the recall and precision figures with respect to the nuclearity and relation assignments, we also factored in the statuses and the rhetorical relations that labeled each pair of spans.

Table 2 summarizes the results (P-D and P-I (R)ecall and (P)recision figures) for each level (Sentence, Paragraph, and Text). The numbers in the "Weighted Average" line report averages of the Sentence-, Paragraph-, and Text-specific figures, weighted according to the number of units at each level. The numbers in the "All" line reflect recall and precision figures computed across the entire trees, with no attention paid to sentence and paragraph boundaries.

Given the significantly different syntactic structures of Japanese and English, we were not surprised by the low recall and precision results that reflect the similarity between discourse trees built below the sentence level. However, as Table 2 shows, there are significant differences between discourse trees at the paragraph and text levels as well. For example, the Position-Independent figures show that only about 62% of the sentences and only about 53% of the hierarchical spans built across sentences could be matched between the two corpora. When one looks at the status and rhetorical relations associated with the spans built across sentences at the paragraph level, the P-I recall and precision figures drop to about 43% and 35% respectively.

The differences in recall and precision are explained both by differences in the way information is packaged into paragraphs in the two languages and the way it is structured rhetorically both within and above the paragraph level.

These results strongly suggest that if one attempts to translate Japanese into English on a sentence-by-sentence basis, it is likely that the resulting text will be unnatural from a discourse perspective. For example, if some information rendered using a CONTRAST relation in Japanese is rendered using an ELABORATION relation in English, it would be inappropriate to use a discourse marker like "but" in the English translation, although that would be consistent with the Japanese discourse structure.

An inspection of the rhetorical mappings between Japanese and English revealed that some Japanese rhetorical renderings are consistently mapped into one or a few preferred renderings in English. For example, 34 of 115 CONTRAST relations in the Japanese texts are mapped into CONTRAST relations in English; 27 become nuclei of relations such as ANTITHESIS and CONCESSION, 14 are translated as COMPARISON relations, 6 as satellites of CONCESSION relations, 5 as LIST relations, etc. Our goal is to learn these systematic discourse mapping rules and exploit them in a machine translation context.

3 Towards a discourse-based machine translation system

3.1 Overall architecture

We are currently working towards building the modules of a Discourse-Based Machine Translation system that works along the following lines.

1. A discourse parser, such as those described by Sumita et al. (1992), Kurohashi (1994), and Marcu (1999), initially derives the discourse structure of the text given as input.
2. A discourse-structure transfer module rewrites the discourse structure of the input text so as to reflect a discourse rendering that is natural to the target language.
3. A statistical module maps the input text into the target language using translation and language models that incorporate discourse-specific features, which are extracted from the outputs of the discourse parser and discourse transfer modules.

In this paper, we focus only on the discourse-structure transfer module. That is, we investigate the feasibility of building such a module.

3.2 The discourse-based transfer model

In order to learn to rewrite discourse structure trees, we first address a related problem, which we define below:

Definition 3.1 *Given two trees T_s and T_t and a correspondence Table C defined between T_s and T_t at the leaf level in terms of $=$, $\subset$, $\supset$, and $\cong$ relations, find a sequence of actions that rewrites the tree T_s into T_t.*

If for any tuple $\langle T_s, T_t, C\rangle$ such a sequence of actions can be derived, it is then possible to use a corpus of $\langle T_s, T_t, C\rangle$ tuples in order to automatically learn to derive from an unseen tree T_{s_i}, which has the same structural properties as the trees T_s, a tree T_{t_j}, which has structural properties similar to those of the trees T_t.

In order to solve the problem in definition 3.1, we extend the shift-reduce parsing paradigm applied by Magerman (1995), Hermjakob and Mooney (1997), and Marcu (1999). In this extended paradigm, the transfer process starts with an empty Stack and an Input List that contains a sequence of elementary discourse trees *edts*, one *edt* for each *edu* in the tree T_s given as input. The status and rhetorical relation associated with each *edt* is undefined. At each step, the transfer module applies an operation that is aimed at building from the units in T_s the discourse tree T_t. In the context of our discourse-transfer module, we need 7 types of operations:

- SHIFT operations transfer the first *edt* from the input list into the stack;
- REDUCE operations pop the two discourse trees located at the top of the stack; combine them into a new tree updating the statuses and rhetorical relation names of the trees involved in the operation; and push the new tree on the top of the stack. These operations are used to build the structure of the discourse tree in the target language.
- BREAK operations are used in order to break the *edt* at the beginning of the input list into a predetermined number of units. These operations are used to ensure that the resulting tree has the same number of *edts* as T_t. A BREAK operation is necessary whenever a Japanese *edu* is mapped into multiple English units.
- CREATE-NEXT operations are used in order to create English discourse constituents that have no correspondent in the Japanese tree.
- FUSE operations are used in order to fuse the *edt* at the top of the stack into the tree that immediately precedes it. These operations are used whenever multiple Japanese *edus* are mapped into one English *edu*.
- SWAP operations swap the *edt* at the beginning of the input list with an *edt* found one or more positions to the right. These operations are necessary for re-ordering discourse constituents.
- ASSIGNTYPE operations assign one or more of the following types to the tree at the top of the stack: Unit, MultiUnit, Sentence, Paragraph, MultiParagraph, and Text. These op-

erations are necessary in order to ensure sentence and paragraph boundaries that are specific to the target language.

For example, the first sentence of the English tree in Figure 1 can be obtained from the original Japanese sequence by following the sequence of actions (5), whose effects are shown in Figure 2. For the purpose of compactness, the figure does not illustrate the effect of ASSIGNTYPE actions. For the same purpose, some lines correspond to more than one action.

BREAK 2; SWAP 2; SHIFT; ASSIGNTYPE UNIT; SHIFT; REDUCE-NS-ELABORATION-OBJECT-ATTRIBUTE-E; ASSIGNTYPE MULTIUNIT; SHIFT; ASSIGNTYPE UNIT; SHIFT; ASSIGNTYPE UNIT; FUSE; ASSIGNTYPE UNIT; SWAP 2; SHIFT; ASSIGNTYPE UNIT; FUSE; BREAK 2; SHIFT; ASSIGNTYPE UNIT; SHIFT; ASSIGNTYPE UNIT; REDUCE-NS-ELABORATION-ADDITIONAL; ASSIGNTYPE MULTIUNIT; REDUCE-NS-CONTRAST; ASSIGNTYPE MULTIUNIT; REDUCE-SN-BACKGROUND; ASSIGNTYPE SENTENCE. (5)

For our corpus, in order to enable a discourse-based transfer module to derive any English discourse tree starting from any Japanese discourse tree, it is sufficient to implement:

- one SHIFT operation;
- 3 × 2 × 85 REDUCE operations; (For each of the three possible pairs of nuclearity assignments NUCLEUS-SATELLITE (NS), SATELLITE-NUCLEUS (SN), AND NUCLEUS-NUCLEUS (NN), there are two possible ways to reduce two adjacent trees (one results in a binary tree, the other in a non-binary tree (Marcu, 1999)), and 85 relation names.)
- three types of BREAK operations; (In our corpus, a Japanese unit is broken into two, three, or at most four units.)
- one type of CREATE-NEXT operation;
- one type of FUSE operation;
- eleven types of SWAP operations; (In our corpus, Japanese units are at most 11 positions away from their location in an English-specific rendering.)
- seven types of ASSIGNTYPE operations: Unit, MultiUnit, Sentence, MultiSentence, Paragraph, MultiParagraph, and Text.

These actions are sufficient for rewriting any tree T_s into any tree T_t, where T_t may have a different number of *edus*, where the *edus* of T_t may have a different ordering than the *edus* of T_s, and where the hierarchical structures of the two trees may be different as well.

3.3 Learning the parameters of the discourse-transfer model

We associate with each configuration of our transfer model a learning case. The cases were generated by a program that automatically derived the sequence of actions that mapped the Japanese trees in our corpus into the sibling English trees, using the correspondences at the elementary unit level that were constructed manually. Overall, the 40 pairs of Japanese and English discourse trees yielded 14108 cases.

To each learning example, we associated a set of features from the following classes:

Operational and discourse features reflect the number of trees in the stack, the input list, and the types of the last five operations. They encode information pertaining to the types of the partial trees built up to a certain time and the rhetorical relations that hold between these trees.

Correspondence-based features reflect the nuclearity, rhetorical relations, and types of the Japanese trees that correspond to the English-like partial trees derived up to a given time.

Lexical features specify whether the Japanese spans that correspond to the structures derived up to a given time use potential discourse markers, such as *dakara (because)* and *no ni (although)*.

The discourse transfer module uses the C4.5 program (Quinlan, 1993) in order to learn decision trees and rules that specify how Japanese discourse trees should be mapped into English-like trees. A ten-fold cross-validation evaluation of the classifier yielded an accuracy of 70.2% (± 0.21).

In order to better understand the strengths and weaknesses of the classifier, we also attempted to break the problem into smaller components. Hence, instead of learning all actions at once, we attempted to learn first whether the rewriting procedure should choose a SHIFT, REDUCE, BREAK, FUSE, SWAP, or ASSIGNTYPE operation (the "Main Action Type" classifier in table 3), and only then to refine this decision by determining what type of reduce operation to perform, how many units to break a Japanese units into, how big the distance to the SWAP-ed unit should be, and what type of ASSIGNTYPE operation one should perform. Table 3 shows the sizes of each

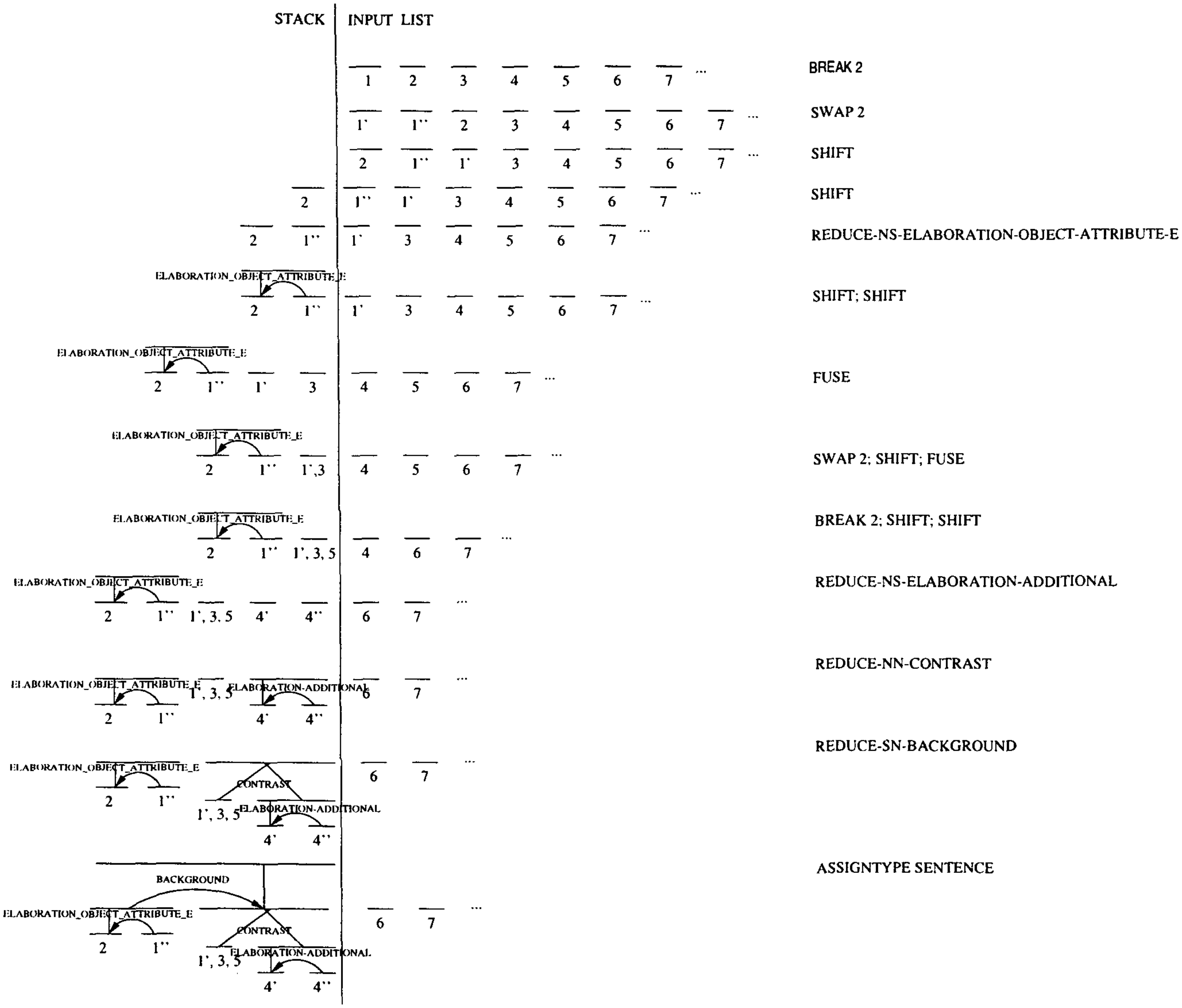

Figure 2: Example of incremental tree reconstruction.

data set and the performance of each of these classifiers, as determined using a ten-fold cross-validation procedure. For the purpose of comparison, each classifier is paired with a majority baseline.

The results in Table 3 show that the most difficult subtasks to learn are that of determining the number of units a Japanese unit should be broken into and that of determining the distance to the unit that is to be swapped. The features we used are not able to refine the baseline classifiers for these action types. The confusion matrix for the "Main Action Type" classifier (see Table 5) shows that the system has trouble mostly identifying BREAK and CREATE-NEXT actions. The system has difficulty learning what type of nuclearity ordering to prefer (the "Nuclearity-Reduce" classifier) and what relation to choose for the English-like structure (the "Relation-Reduce" classifier).

Figure 3 shows a typical learning curve, the one that corresponds to the "Reduce Relation" classifier. Our learning curves suggest that more training data may improve performance. However, they also suggest that better features may be needed in order to improve performance significantly.

Table 4 displays some learned rules. The first rule accounts for rhetorical mappings in which the order of the nucleus and satellite of an ATTRIBUTION relation is changed when translated from Japanese into English. The second rule was learned in order to map EXAMPLE Japanese satellites into EVIDENCE English satellites.

Classifier	# cases	Accuracy (10-fold cross validation)	Majority baseline accuracy
General (Learns all classes at once)	14108	70.20% (±0.21)	22.05% (on ASSIGNTYPE UNIT)
Main Action Type	14108	82.53% (±0.25)	45.47% (on ASSIGNTYPE)
AssignType	6416	90.46% (±0.39)	57.30% (on ASSIGNTYPE Unit)
Break	394	82.91% (±1.40)	82.91% (on BREAK 2)
Nuclearity-Reduce	2388	67.43% (±1.03)	50.92% (on NS)
Relation-Reduce	2388	48.20% (±1.01)	17.18% (on ELABORATION-OBJECT-ATTRIBUTE-E)
Swap	842	62.98% (±1.62)	62.98% (on SWAP 1)

Table 3: Performance of the classifiers

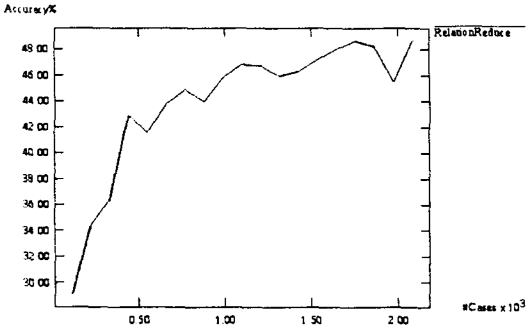

Figure 3: Learning curve for the Relation-Reduce classifier.

if rhetRelOfStack-1InJapTree = ATTRIBUTION
  then rhetRelOfTopStackInEngTree ← ATTRIBUTION

if rhetRelOfTopStackInJapTree = EXAMPLE ∧
  isSentenceTheLastUnitInJapTreeOfTopStack = false
  then rhetRelOfTopStackInEngTree ← EVIDENCE

Table 4: Rule examples for the Relation-Reduce classifier.

4 Evaluation of the discourse-based transfer module

By applying the General classifier or the other six classifiers successively, one can map any Japanese discourse tree into a tree whose structure comes closer to the natural rendering of English. To evaluate the discourse-based transfer module, we carried out a ten-fold cross-validation experiment. That is, we trained the classifiers on 36 pairs of manually built and aligned discourse structures, and we then used the learned classifiers in order to map 4 unseen Japanese discourse trees into English-like trees. We measured the similarity of the derived trees with the English trees built manually, using the metrics discussed in Section 2. We repeated the procedure ten times, each time training and testing on different subsets of tree pairs.

Action	(a)	(b)	(c)	(d)	(e)	(f)	(g)
ASSIGNTYPE (a)	660						
BREAK (b)		1			2	28	1
CREATE-NEXT (c)					1	8	
FUSE (d)				69	8	3	
REDUCE (e)		4		18	193	30	3
SHIFT (f)	1	4		15	44	243	25
SWAP (g)		3		4	14	43	25

Table 5: Confusion matrix for the Main Action Type classifier.

We take the results reported in Table 2 as a baseline for our model. The baseline corresponds to applying no knowledge of discourse. Table 6 displays the absolute improvement (in percentage points) in recall and precision figures obtained when the General classifier was used to map Japanese trees into English-looking trees. The General classifier yielded the best results. The results in Table 6 are averaged over a ten-fold cross-validation experiment.

The results in Table 6 show that our model outperforms the baseline with respect to building English-like discourse structures for sentences, but it under-performs the baseline with respect to building English-like structures at the paragraph and text levels. The main shortcoming of our model seems to come from its low performance in assigning paragraph boundaries. Because our classifier does not learn correctly which spans to consider paragraphs and which spans not, the recall and precision results at the paragraph and text levels are negatively affected. The poorer results at the paragraph and text levels can be also explained by errors whose effect cumulates during the step-by-step tree-reconstruction procedure; and by the fact that, for these levels, there is less data to learn from.

However, if one ignores the sentence and paragraph boundaries and evaluates the discourse structures overall, one can see that our model outperforms the baseline on all accounts according to the Position-Dependent evaluation; outperforms the baseline with respect to the assignment of elementary units, hierarchical spans, and nuclearity statuses according to the Position-Independent evaluation and under-performs the baseline only slightly

Level	Units		Spans		Status/Nuclearity		Relations	
	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P
Sentence	+9.1	+25.5	+2.0	+19.9	+0.4	+13.4	−0.01	+8.4
Paragraph	−14.7	+1.4	−12.5	−1.7	−11.0	−2.4	−9.9	−3.3
Text	−9.6	−13.5	−7.1	−11.1	−6.3	−10.0	−5.2	−8.8
Weighted Average	+1.5	+14.1	−2.1	+9.9	−3.1	+6.4	−3.0	+3.9
All	−1.2	+2.5	−0.1	+2.9	+0.6	+3.5	+0.7	+2.6
	P-I R	P-I P	P-I R	P-I P	P-I R	P-I P	P-I R	P-I P
Sentence	+13.4	+30.4	+3.1	+36.1	−6.3	+18.6	−10.1	+3.9
Paragraph	−15.6	+0.6	−13.5	−0.8	−11.7	−1.8	−10.3	−2.8
Text	−15.4	−23.3	−13.0	−20.4	−13.2	−19.5	−11.5	−17.0
Weighted Average	+3.6	+15.5	−2.7	+17.1	−8.5	+7.3	−10.5	-0.4
All	+12.7	+29.6	+2.0	+28.8	−5.1	+13.0	−7.9	+2.2

Table 6: Relative evaluation of the discourse-based transfer module with respect to the figures in Table 2.

with respect to the rhetorical relation assignment according to the Position-Independent evaluation. More sophisticated discourse features, such as those discussed by Maynard (1998), for example, and a tighter integration with the lexicogrammar of the two languages may yield better cues for learning discourse-based translation models.

5 Conclusion

We presented a systematic empirical study of the role of discourse structure in MT. Our study strongly supports the need for enriching MT systems with a discourse module, capable of re-ordering and re-packaging the information in a source text in a way that is consistent with the discourse rendering of a target language. We presented an extended shift-reduce parsing model that can be used to map discourse trees specific to a source language into discourse trees specific to a target language. Our model outperforms a baseline with respect to its ability to predict the discourse structure of sentences. Our model also outperforms the baseline with respect to its ability to derive discourse structures that are closer to the natural, rhetorical rendering in a target language than the original discourse structures in the source language. Our model is still unable to determine correctly how to re-package sentences into paragraphs; a better understanding of the notion of "paragraph" is required in order to improve this.

References

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Ulf Hermjakob and Raymond J. Mooney. 1997. Learning parse and translation decisions from examples with rich context. In *Proc. of ACL'97*, pages 482–489, Madrid, Spain. .

Sadao Kurohashi and Makoto Nagao. 1994. Automatic detection of discourse structure by checking surface information in sentences. In *Proc. of COLING'94*, volume 2, pages 1123–1127, Kyoto, Japan.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proc. of ACL'95*, pages 276–283, Cambridge, Massachusetts.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1999. A decision-based approach to rhetorical parsing. In *Proc. of ACL'99*, pages 365–372, Maryland.

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proc. of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, Maryland.

Senko K. Maynard. 1998. *Principles of Japanese Discourse: A Handbook*. Cambridge Univ. Press.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Sidney Siegel and N.J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Second edition.

Kazuo Sumita, Kenji Ono, T. Chino, Teruhiko Ukita, and Shin'ya Amano. 1992. A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, v 2, pages 1133–1140.

J. White and T. O'Connell. 1994. Evaluation in the ARPA machine-translation program: 1993 methodology. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 135–140, Washington, D.C.