Three English Learner Assistance Systems Using Automatic Paraphrasing Techniques

Masaki Murata and Hitoshi Isahara
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{murata,isahara}@nict.go.jp

Abstract

We developed three systems based on automatic paraphrasing techniques to help English learners and English-language beginners. One system extracts personal error patterns in the user's English usage. The second transforms English sentences containing the letters "l" and "r" into sentences containing fewer instances of these letters, which Japanese people have trouble pronouncing properly in English. This system could be used, for example, to transform a draft of a presentation that a Japanese speaker was to present to an audience. The third is an annotation system that provides definition sentences of difficult English words, making them easier to understand. We believe that these systems will be useful both for learners of English and in studies on second-language acquisition.

1 Introduction

Studies on paraphrasing (IWPT, 2001; Murata and Isahara, 2001) are relevant to a range of research topics including sentence generation, summarization, and question-answering in natural language processing (Katoh and Uratani, 1999; Takahashi et al., 2003). Several techniques have been constructed for paraphrasing. We used automatic paraphrasing techniques based on natural language processing to develop three learner-assistance systems for English language learners and beginners. One system extracts an individual's personal error patterns in using English. The second transforms English sentences containing the letters "l" and "r" into sentences with fewer instances of these letters, which Japanese people have trouble pronouncing properly in English. The third system transforms difficult English words into easier ones by providing definitions of the difficult words that clarify their meaning. These systems and studies using them may also have applications in studies on second language acquisition (Dulay et al., 1982; Larsen-Freeman and Long, 1991; Granger, 1998).

2 Extraction of personal error patterns in using English

We investigated methods of automatically extracting differences from pairs of texts that correspond to each other within the domain of paraphrasing studies. One study examined the extraction of differences between spoken and written language and transformation rules for automatic translation from written to spoken language by matching presentations at academic conferences and the papers corresponding to them (Murata and Isahara, 2002a). Another study focused on extracting rewriting rules and synonyms in patents by matching the patent claim and its embodiment, both of which have the same meaning (Murata and Isahara, 2002c). In this study, we extracted an individual's personal English error patterns by matching a paper before proofreading with the corresponding paper after proofreading. (The proofreading was performed by native speakers.) We used 80 papers by a first author to extract the author's

Table 1: Examples of personal English errors

Before proofreading		After proofreading	Frequency
“the”	⇒	ϕ	393
ϕ	⇒	“the”	357
“a”	⇒	“the”	146
“a”	⇒	ϕ	122
ϕ	⇒	“a”	120
“the”	⇒	“a”	116
“is”	⇒	“was”	68
ϕ	⇒	“of”	56
“of”	⇒	“for”	49
“in”	⇒	“of”	36
“are”	⇒	“were”	36
“of”	⇒	ϕ	34
“of”	⇒	“in”	32
“in”	⇒	“for”	32
“an”	⇒	“the”	32
“which”	⇒	“that”	31
“is”	⇒	“are”	28
“a”	⇒	“an”	28
ϕ	⇒	“by”	28
ϕ	⇒	“an”	27
“having”	⇒	“with”	25
ϕ	⇒	“that”	25
ϕ	⇒	“in”	24
“the”	⇒	“an”	23
ϕ	⇒	“used”	23
“result”	⇒	“results”	22
“and”	⇒	“or”	22
ϕ	⇒	“also”	22
ϕ	⇒	“to”	21
ϕ	⇒	“thus”	21
ϕ	⇒	“only”	20
ϕ	⇒	“and”	20
“we”	⇒	ϕ	19
“that”	⇒	ϕ	19
ϕ	⇒	“as”	18
“metonymy”	⇒	“metonymic”	17
“cases”	⇒	“case particles”	17
“as”	⇒	ϕ	17
“are”	⇒	“is”	17
ϕ	⇒	“have”	17
“use”	⇒	“used”	16
“short term”	⇒	“short-term”	16
“only”	⇒	ϕ	16
“in”	⇒	“at”	16
“have”	⇒	ϕ	16
“by”	⇒	ϕ	15
“When a”	⇒	“A”	15
ϕ	⇒	“method”	15

English error patterns. The error patterns were extracted using a Unix Diff command as the differences between a paper before proofreading and the corresponding paper after proofreading. (For a detailed description of the method used to extract differences, see (Murata and Isahara, 2002a; Murata and Isahara, 2002c; Murata, 2002; Murata and Isahara, 2002b).) We then counted the frequencies of the extracted English error patterns. The results are shown in Table 1. ϕ indicates a void: "$\phi \Rightarrow$" and "$\Rightarrow \phi$" indicate insertion and deletion, respectively.

The results showed that most of the errors made by the first author related to usage of "the", with "a" being the next most frequent source of error. Correct usage of articles ("a" and "the") is very difficult for Japanese people (Murata and Nagao, 1993) and in this case, the author was Japanese. The next most frequent source of error was tenses such as "is" ⇒ "was" and "are" ⇒ "were". Interestingly, the errors relating to articles were symmetrical, while those relating to tenses were not. The frequency of "is" ⇒ "was" was high, but that of "was" ⇒ "is" was low.

The reason for this lack of symmetry may be that "is" and "are" are default forms and they are often used.

The next most common errors related to prepositions such as $\phi \Rightarrow$ "of", "of" ⇒ "for", and "in" ⇒ "of". We also noticed errors relating to "which" ⇒ "that" and "having" ⇒ "with". We found that our system worked well in determining the author's personal English errors.

A personalized education system designed to meet the needs of specific users would be extremely useful. (This was also described in the call for papers for this workshop.) Our system of extracting personal English errors would be useful in developing such a system. It could also be used to gather the errors made by a specific class of English language students or by students taught by a particular teacher, which would be useful for English teachers.

3 Transformation of English sentences containing the letters "l" and "r" into sentences with fewer instances of these letters

We have carried out several studies on automatically paraphrasing sentences for various purposes including automatic translation from written to spoken language (Murata and Isahara, 2002a); polishing sentences; compressing sentences without changing their meaning, as in summarization; and transforming sentences to sentences that have the same meaning and are similar to input questions in answering questions. These systems were constructed using one unique model (Murata and Isahara, 2001). In this section, we describe a system for transforming English sentences containing the letters "l" and "r" into sentences containing fewer instances of these letters. The system was constructed using paraphrasing techniques. Japanese people have difficulty in pronouncing "l" and "r" and this system could be useful, for example, for transforming a draft of a presentation to be presented by a Japanese speaker into sentences containing fewer "l" and "r" letters.

The paraphrasing technique consists of two modules: a transformation module and an evaluation module. The sentence to be transformed is input into the system. Several potential transformation types are generated in the transformation module and then tested in the evaluation module, where the most appropriate one is selected. This one is used for the transformation and the result is output.

We used synonyms in WordNet 2.0 (Princeton University, 2003) for the transformation module. The following conditions were used for the evaluation module.

- A word containing fewer instances of "l" and "r" preceding vowels is more appropriate for transformation. ("l" and "r" preceding vowels are particular difficult for Japanese people to pronounce.)

- For words with the same number of instances of "l" and "r" preceding vowels, a word

Table 2: Examples of transforming English sentences into sentences containing fewer instances of “l” and “r” letters preceding vowels

Correct transformation
We think a good | approach way | is to construct it using “X *no* Y”.
The criteria used to | select determine | the most appropriate transformation type must be predefined.
This figure shows the | structure composition | of the thesaurus.
length d is the | length size | of a document d.
Incorrect transformation
This is the | title name | of the query.
P of d and t is the | location determination | of the first occurrence of a term *t* in the document *d*.
This term is for weighting terms which are | followed used | by the Japanese-language particle “nado”.
We think that this | problem question | can be overcome by enlarging the corpus from which the examples are extracted.

that occurs more frequently in an English corpus with contexts is more appropriate for transformation.

- When a word does not occur in an English corpus with contexts, it is not used for transformation.

We used the two words preceding the transformed expressions, the transformed expressions, and the two words following the transformed expressions as the contexts. We used the British National Corpus (BNC), which contains over 100 million words (Oxford University Computing Services, 1995), for the English corpus.

We assessed the system experimentally using drafts of our presentations at academic conferences as input. Examples of the output of the system are shown in Table 2. A part enclosed by vertical lines "|" indicates a transformed expression. The upper expression was transformed to the lower expression. Each expression was transformed to an expression containing fewer instances of "l" and "r" preceding vowels, e.g. "approach" and "length" were transformed into "size" and "way", respectively. However, in the current system, this transformation sometimes changed the meaning of a sentence.

For this reason, it is better to use the system semi-automatically; i.e., the system outputs candidates for transformation with fewer instances of "l" and "r" preceding vowels and the frequency of their occurrence in an English corpus with contexts. The user then selects the appropriate expression from these candidates.

The current system focuses only on the letters "l" and "r". In future, we would like to add other letters that are difficult for English beginners to pronounce such as "f" or "v" and we would also like to examine which letters are most difficult for English beginners to pronounce.

4 Providing definition sentences for difficult words to clarify their meaning

Kaji and Kurohashi constructed a dictionary of definition sentences that could be used to transform sentences into sentences that were easier to understand (Kaji and Kurohashi, 2004). Definition sentences explain the meaning of words, so they can be used to transform a difficult sentence into one that is easier to understand. We developed a system to make reading easier by providing difficult words with definition sentences.

Our system first detects a difficult word and then gives it a definition sentence. The current system used two methods for detecting difficult words. One method extracts words that the user has not yet written as difficult words, and the other extracts words with low frequency in an English corpus. (Word frequency in an English corpus is often used in language education to extract easy or difficult words (Granger, 1998).)

Examples of the outputs of our system are shown in Table 3. The difficult words that were extracted were words with a frequency of less than 1000 in the BNC corpus. We used the WordNet 2.0 manual as input and definition sentences from the EDR dictionary (EDR, 1993). We also used WordNet 2.0 for stemming words. The part "[Notes: ...]" indicates a definition sentence provided by our system. The current system outputs all the definition sentences for a particular word, and the symbol "|" is used to separate definition sentences with different meanings. Although we used English definition sentences in Table 3, we can also use Japanese definition sentences for Japanese users. "[Caution!]" is assigned to words that are judged to be difficult but that are not included in the EDR dictionary. "[Above!]" indicates that the definition sentence has been given previously. "Corpus", "consortium", and "lexicography" were appropriately extracted as difficult words and given definition sentences.

Users have their own professional domains and our system can use individual user's characteristics to provide definition sentences only for words that they are unlikely to know. For example,

Table 3: Examples of giving difficult words definition sentences to make reading easier

<table>
<tr><td>The British National Corpus [Notes: a collection of all the works of a special type, on a special subject|a gland of an insect, called corpus allatum|a total assemblage of law in a country|an extract of corpus luteum of a pig or a cow|the glassy fluid from the eye|a set of material or data for use during study, especially for linguistic analysis] is a very large corpus [Above!] of modern English, containing over 100 million words from both spoken and written English texts.</td></tr>
<tr><td>The project was carried out and is managed by an industrial/academic consortium [Notes: an association of creditor nations] led by Oxford University Press, ...</td></tr>
<tr><td>The spoken part (10%) includes a large amount of unscripted [Notes: (of a speech or discussion that is broadcast) spoken naturally or without previous arrangement] informal conversation, recordeded [Caution!] by volunteers selected from different age, region and social classes in a demographically [Notes: in a demographical way] balanced way, ...</td></tr>
<tr><td>... will be useful for a very wide variety of research purposes, in fields as distinct as lexicography, [Notes: the creating and printing of dictionaries] artificial intelligence, speech recognition and synthesis, literary studies, and all varieties of linguistics.</td></tr>
<tr><td>The corpus [Above!] is encoded [Caution!] according to the Guidelines ...</td></tr>
<tr><td>..., the equivalent of more than a thousand high capacity floppy diskettes. [Notes: a small plastic disk coated with magnetic material on which computer data can be stored called a floppy disk]</td></tr>
<tr><td>To put these numbers into perspective, the thickness of the average paperback [Notes: of the form of a publication enveloped by a slender cardboard cover—a small book bound with a thin cardboard cover] book is about 250 pages per centimetre [Caution!];</td></tr>
</table>

when we used the method to extract difficult words but to avoid extracting words occurring in our papers, “corpus”, which we know very well, was not extracted.

“encoded” is given “[Caution!]”, so the user realizes that “encoded” is difficult. This helps English learners. “[Caution!]” has another interesting function. “recorededed”, which is a typo, is also given “[Caution!]”.

In future work, we would like to develop a system that stores all the sentences that a user has written and read and show him/her which word he/she first encountered when a new sentence is presented. (We have already investigated a system for highlighting expressions that appear first in documents (Murata and Isahara, 2002c).) We also consider that it may be interesting to highlight the words that appear first in English language school textbooks. Although there are systems that provide translations or meaning glosses to assist readers (Poznanski et al., 1998), the ideas of user-dependent processing and highlighting expressions that appear first are novel and in future, we would like to further examine these concepts.

5 Conclusion

We used automatic paraphrasing techniques based on natural language processing to develop three systems for helping English learners and beginners. They included a system for extracting personal error patterns in the user’s English usage; a system for transforming English sentences containing the letters “l” and “r”, which Japanese people have trouble pronouncing, into sentences containing fewer instances of these letters; and an annotation system that provides definition phrases for difficult words. We believe that these systems will be useful for English learners and in studies on second language acquisition (Dulay et al., 1982; Larsen-Freeman and Long, 1991; Granger, 1998).

Acknowledgments

We are grateful to Ms. Izumi and Dr. Kotani of the National Institute of Information and Communications Technology for their help and valuable comments on this study.

References

Dulay, Heidi, Marina Burt, and Stephen Krashen. 1982. *Language Two.* Oxford University Press.

EDR, 1993. *EDR Electronic Dictionary Technical Guide.* EDR (Japan Electronic Dictionary Research Institute, Ltd.).

Granger, Sylviane. 1998. *Learner English on Computer.* Longman.

IWPT. 2001. *NLPRS’2001 Workshop on Automatic Paraphrasing: Theories and Applications.*

Kaji, Nobuhiro and Sadao Kurohashi. 2004. Recognition and paraphrasing of periphrastic and overlapping verb phrases. In *LERC 2004.*

Katoh, Naoto and Noriyoshi Uratani. 1999. A new approach to acquiring linguistic knowledge for locally summarizing Japanese news sentences. *Journal of Natural Language Processing*, 6(7). (in Japanese).

Larsen-Freeman, Diane and Michael H. Long. 1991. *An Introduction to second language acquisition research.* Longman.

Murata, Masaki. 2002. NLP using DIFF — use of convenient tool for detecting differences, MDIFF —. *Journal of Natural Language Processing*, 9(2). (in Japanese).

Murata, Masaki and Hitoshi Isahara. 2001. Universal model for paraphrasing — using transformation based on a defined criteria —. In *NLPRS’2001 Workshop on Automatic Paraphrasing: Theories and Applications.*

Murata, Masaki and Hitoshi Isahara. 2002a. Automatic extraction of differences between spoken and written languages, and automatic translation from the written to the spoken language. In *LERC 2002*.

Murata, Masaki and Hitoshi Isahara. 2002b. Using the diff command for natural language processing. http://arxiv.org/abs/cs.CL/0208020.

Murata, Masaki and Hitoshi Isahara. 2002c. Using the diff command in patent documents. *Proceedings of the Third NTCIR Workshop (PATENT)*.

Murata, Masaki and Makoto Nagao. 1993. Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *Proceedings of the 5th TMI*, pages 218–225.

Oxford University Computing Services. 1995. British national corpus.

Poznanski, Victor, Pete Whitelock, Jan IJdens, and Steffan Corley. 1998. Practical glossing by prioritised tiling. *COLING-ACL '98*, pages 1060–1066.

Princeton University. 2003. Wordnet 2.0.

Takahashi, Tetsuro, Kozo Nawata, Kentaro Inui, and Yuji Matsumoto. 2003. Effect of structural matching and paraphrasing in question answering. *IEICE Transactions on Information and Systems*, E86–D(9):1677–1685.