Evaluating Contextual Dependency of Paraphrases using a Latent Variable Model

Kiyonroi OHTAKE

Spoken Language Communication Research Laboratories
Advanced Telecommunications Research Institute International
Kyoto 619-0288 Japan
`kiyonori.ohtake + @atr.jp`

Abstract

This paper presents an evaluation method employing a latent variable model for paraphrases with their contexts. We assume that the context of a sentence is indicated by a latent variable of the model as a topic and that the likelihood of each variable can be inferred. A paraphrase is evaluated for whether its sentences are used in the same context. Experimental results showed that the proposed method achieves almost 60% accuracy and that there is not a large performance difference between the two models. The results also revealed an upper bound of accuracy of 77% with the method when using only topic information.

1 Introduction

This paper proposes a method to evaluate whether a paraphrasing pair is contextually independent. Evaluating a paraphrasing pair is important when we extract paraphrases from a corpus or apply a paraphrase to a sentence, since we must guarantee that the paraphrase carries almost the same meaning. However, the meaning carried by a sentence is affected by its context. Thus, we focus on the contextual dependency of paraphrases.

A thing can be expressed by various expressions, and a single idea can be paraphrased in many ways to enrich its expression or to increase understanding. Paraphrasing plays a very important role in natural language expressions. However, it is very hard for machines to handle different expressions that carry the same meaning.

The importance of paraphrasing has been widely acknowledged, and many paraphrasing studies have been carried out. Using only surface similarity is insufficient for evaluating paraphrases because there are not only surface differences but many other kinds of differences between paraphrased sentences. Thus, it is not easy to evaluate whether two sentences carry almost the same meaning.

Some studies have constructed and evaluated hand-made rules (Takahashi et al., 2001; Ohtake and Yamamoto, 2001). Others have tried to extract paraphrases from corpora (Barzilay and McKeown, 2001; Lin and Pantel, 2001), which are very useful because they enable us to construct paraphrasing rules. In addition, we can construct an example-based or a Statistical Machine Translation (SMT)-like paraphrasing system that utilizes paraphrasing examples. Thus, collecting paraphrased examples must be continued to achieve high-performance paraphrasing systems.

Several methods of acquiring paraphrases have been proposed (Barzilay and McKeown, 2001; Shimohata and Sumita, 2002; Yamamoto, 2002). Some use parallel corpora as resources to obtain paraphrases, which seems a promising way to extract high-quality paraphrases.

However, unlike translation, there is no obvious paraphrasing direction. Given paraphrasing pair E_1:E_2, we have to know the paraphrasing direction to paraphrase from E_1 to E_2 and vice versa. When extracting paraphrasing pairs from corpora, whether the paraphrasing pairs are con-

textually dependent paraphrases is a serious problem, and thus there is a specific paraphrase direction for each pair. In addition, it is also important to evaluate a paraphrasing pair not only when extracting but also when applying a paraphrase.

Consider this example, automatically extracted from a corpus: *Can I pay by traveler's check? / Do you take traveler's checks?* This example seems contextually independent. On the other hand, here is another example: *I want to buy a pair of sandals. / I'm looking for sandals.* This example seems to be contextually dependent, because we don't know whether the speaker is only looking for a single pair of sandals. In some contexts, the latter sentence means that the speaker is seeking or searching for sandals. In other words, the former sentence carries specific meaning, but the latter carries generic meaning. Thus, the paraphrasing sentences are contextually dependent, and although the paraphrasing direction from specific to generic might be acceptable, the opposite direction may not be.

We can solve part of this problem by inferring the contexts of the paraphrasing sentences. A text model with latent variables can be used to infer the topic of a text, since latent variables correspond to the topics indicated by texts. We assume that a topic indicated by a latent variable of a text model can be used as an approximation of context. Needless to say, however, such an approximation is very rough, and a more complex model or more powerful approach must be developed to achieve performances that match human judgement in evaluating paraphrases.

The final goal of this study is the evaluation of paraphrasing pairs based on the following two factors: contextual dependency and paraphrasing direction. In this paper, however, as a first step to evaluate paraphrasing pairs, we focus on the evaluation of contextual dependency by using probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as text models with latent variables.

2   Latent Variable Models and Topic Inference

In this section, we introduce two latent variable models, pLSI and LDA, and also explain how to infer a topic with the models.

In addition to pLSI and LDA, there are other latent variable models such as mixture of unigrams. We used pLSI and LDA because Blei et al. have already demonstrated that LDA outperforms mixture of unigrams and pLSI (Blei et al., 2003), and a toolkit has been developed for each model.

From a practical viewpoint, we want to determine how much performance difference exists between pLSI and LDA through evaluations of contextual paraphrase dependency. The time complexity required to infer a topic by LDA is larger than that by pLSI, and thus it is valuable to know the performance difference.

2.1   Probabilistic LSI

PLSI is a latent variable model for general co-occurrence data that associates an unobserved topic variable $z \in \mathcal{Z} = \{z_1, \cdots, z_K\}$ with each observation, i.e., with each occurrence of word $w \in \mathcal{W} = \{w_1, \cdots, w_M\}$ in document $d \in \mathcal{D} = \{d_1, \cdots d_N\}$.

PLSI gives joint probability for a word and a document as follows:

$$P(d, w) = P(d)P(w|d), \qquad (1)$$

where

$$P(w|d) = \sum_{z \in \mathcal{Z}} P(w|z)P(z|d). \qquad (2)$$

However, to infer a topic indicated by a document, we have to obtain $P(z|d)$. From (Hofmann, 1999), we can derive the following formulas:

$$P(z|d, w) \propto P(z)P(d|z)P(w|z) \qquad (3)$$

and

$$P(d|z) \propto \sum_w n(d, w)P(z|d, w), \qquad (4)$$

where $n(d, w)$ denotes term frequency, which is the number of times w occurs in d. Assuming that $P(d|z) = \prod_{w \in d} P(w|z)$, the probability of a topic under document ($P(z|d)$) is proportional to the following formula:

$$P(z)^2 \prod_{w \in d} P(w|z) \sum_w n(d, w)P(w|z). \qquad (5)$$

After a pLSI model is constructed with a learning corpus, we can infer topic $z \in \mathcal{Z}$ indicated

by given document $d = w_1, \cdots, w_{M(d)}$ with Formula 5. A topic z that maximizes Formula 5 is inferred as the topic of document d.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA gives us the marginal distribution of a document $(p(d|\alpha, \beta), d = (w_1, w_2, \cdots w_N))$ by the following formula:

$$\int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta) \right) d\theta, \quad (6)$$

where α parameterizes Dirichlet random variable θ and β parameterizes the word probabilities, and z_n indicates a topic variable $z_n \in Z = \{z_1, z_2, \cdots, z_N\}$. To obtain the probability of a corpus, we take the product of the marginal probabilities of single documents.

Here, we omit the details of parameter estimation and the inference of LDA due to space limitations. However, the important point is that the Dirichlet parameters used to infer the probability of a document can be seen as providing a representation of the document in the topic simplex. In other words, these parameters indicate a point in the topic simplex. Thus, in this paper, we use the largest elements of the parameters to infer the topic (as an approximation of context) to which a given text belongs.

3 Evaluating Paraphrases with Latent Variable Models

To evaluate a paraphrasing pair of sentences, we must prepare a learning corpus for constructing latent variable models. It must be organized so that it consist of documents, and each document must be implicated in a specific context.

Both latent variable models pLSI and LDA require vector format data for their learning. In this paper, we follow the bag-of-words approach and prepare vector data that consist of words and their frequency for each document in the learning corpus.

After constructing the pLSI and LDA models, we can infer a topic by using the models with vector data that correspond to a target sentence. The vector data for the target sentence are constructed by using the target sentence and the sentences that surround it. From these sentences, the vector data that correspond to the target sentence are constructed. We call the number of sentences used to construct vector data "window size."

Evaluating a paraphrasing pair $(P1{:}P2)$ is simple. Construct vector data ($vec(P1)$ and $vec(P2)$) and infer contexts ($T(P1)$ and $T(P2)$) by using a latent variable model. Using pLSI, the topic that indicates the highest probability is used as the inferred result, and using LDA, the largest parameter that corresponds to the topic is used as the inferred result. If topics $T(P1)$ and $T(P2)$ are different, the sentences might be used in different contexts, and the paraphrasing pair would be contextually dependent; otherwise, the paraphrasing pair would be contextually independent.

4 Experiments

We carried out several experiments that automatically evaluated extracted paraphrases with pLSI and LDA. To carry out these experiments, we used plsi-0.03[1] by Kudo for pLSI and lda-c[2] toolkit by Blei (Blei et al., 2003) for LDA.

4.1 Data set

We used a bilingual corpus of travel conversation containing Japanese sentences and corresponding English translations (Takezawa et al., 2002). Since the translations were made sentence by sentence, this corpus was sentence-aligned from its origin and consisted of 162,000 sentence pairs.

The corpus was manually and roughly annotated with topics. Each topic had a two-level hierarchical structure whose first level consisted of 19 topics. Each first-level topic had several subtopics. The second level consisted of 218 topics, after expanding all subtopics of each topic in the first level. A rough annotation example is shown in Table 1; the hierarchical structure of this topic seems unorganized. For example, in the first-level topic, there are topics labeled *basic* and *communication*, which seem to overlap.

[1] http://chasen.org/˜taku/software/plsi/
[2] http://www.cs.berkeley.edu/˜blei/lda-c/

Table 1: Examples of manually annotated topics

sentence	1st topic	2nd topic
Where is the nearest department store?	shopping	buying something
That's too flashy for me.	shopping	choosing something
There seems to be a mistake on my bill.	staying	checkout
There seems to be a mistake on my bill.	staying	complaining

In the corpus, however, there is an obvious textual cohesion such that sentences of the same topic are locally gathered. Each series of sentences can be used as a document for a text model. Under the assumption that each series of sentences is a document, the average number of sentences included in a document is 18.7, and the average number of words included in a document is 44.9.

4.2 Extracting paraphrases

A large collection of parallel texts contains many sentences in one language that correspond to the same expression in the other language for translation. For example, if Japanese sentences $J_{i1}, ..., J_{im}$ correspond to English sentence E_i, then these Japanese sentences would be paraphrases.

We utilized a very simple method to extract Japanese paraphrases from the corpus. First, we extracted duplicate English sentences by exact matching. From the learning set, 18,505 sentences were extracted. Second, we collected Japanese sentences that correspond to each extracted English sentence. Next, we obtained sets of Japanese sentences collected by using English sentences as pivots. In the corpus, one English sentence averaged almost 4.5 Japanese sentences, but this number included duplicate sentences. If duplicate sentences are excluded, the average number of Japanese sentences corresponding to an English sentence becomes 2.4. Finally, we obtained 944,547 Japanese paraphrasing pairs by combining sentences in each group of Japanese sentences.

4.3 Comparing human judgement and inference by latent variable models

In this section, we determine the difference between manually annotated topics and inference results using pLSI and LDA. We originally considered evaluating each paraphrase as a binary classification problem that determines whether both sentences of the paraphrase are used in the same context. We evaluated the inferred results by comparison with the manually annotated topics, and thus accuracy could be calculated when the manually annotated topics were correct. However, accuracy is inappropriate for evaluating results inferred by a latent variable model, since the topics were roughly annotated by humans as mentioned in Section 4.1. Accordingly, we employed Kappa statistics as a rough guide for the correctness of the inferred results by latent variable models.

Tables 2 and 3 show the comparison results, where the window size is 11 (the target sentence + the previous five and the following five sentences). When constructing pLSI models, the parameter for tempered EM (TEM) is set to 0.9 (we use this value in all of the experiments in this paper), because it showed the best performance in preliminary experiments. We performed the experiments on several topics.

Table 2: Comparing results of first-level topic (19)

# of topics	κ by pLSI	κ by LDA
10	0.4812	0.4798
20	0.5085	0.5185
30	0.5087	0.5094
40	0.5392	0.5245
50	0.5185	0.4897

window size = 11

As mentioned in Sections 2.1 and 2.2, we can treat inference results as vector data. Thus, we can use a metric to classify the two vectors that correspond to the inferred results of any two given sentences. We use cosine as a metric and con-

Table 3: Comparing results of second-level topic (218)

# of topicss	κ by pLSI	κ by LDA
30	0.3523	0.3883
40	0.3663	0.4093
50	0.4122	0.4111
60	0.4184	0.4186
70	0.4196	0.4133
80	0.3665	0.3702
90	0.3437	0.3596
100	0.3076	0.3526

window size = 11

ducted comparison experiments for the first- and second-level topics, as shown in Tables 4 and 5. The threshold values used to judge whether topics are the same are indicated in the parentheses.

Table 4: Comparing results of first-level topic (19) with cosine metric

# of topics	κ by pLSI	κ by LDA
10	0.4873(0.5)	0.5042(0.5)
20	0.5230(10^{-6})	0.5841(0.5)
30	0.5502(10^{-6})	0.5672(0.5)
40	0.5808(10^{-6})	0.5871(0.5)
50	0.5611(10^{-6})	0.5573(0.5)

window size = 11

Table 5: Comparing results of second-level topic (218) with cosine metric

# of topics	κ by pLSI	κ by LDA
30	0.3536(0.5)	0.3726(0.5)
40	0.3679(0.5)	0.4006(0.5)
50	0.4127(0.5)	0.4085(0.5)
60	0.4186(0.5)	0.4218(0.5)
70	0.4202(0.5)	0.4202(0.5)
80	0.3733(0.5)	$5.2 * 10^{-7}$(0.5)

window size = 11

We also performed an experiment to confirm the relationship between Kappa statistics and window-size context. Experiments were done under the following conditions: the number of topics was 20 for both pLSI and LDA, Kappa statistics were calculated for the first-level topic, and window sizes were 5, 11, 15, 21, 25, and 31. Table 6

Table 6: Window size and Kappa statistics for first-level annotation

window size	pLSI (20 topics)	LDA (20 topics)
5	0.4580	0.2527
11	0.5085	0.5185
15	0.5165	0.5440
21	0.4613	0.5396
25	0.3286	0.5286
31	0.1730	0.5157

shows the experimental results.

The actual computing time needed to evaluate 944,547 paraphrases with a Pentium M 1.4-GHz, 1-GB memory computer is shown in Table 7. It is important to note that the inference program for pLSI was written in Perl, but for LDA it was written in C.

Table 7: Computing time to evaluate paraphrases

# of topics	pLSI	LDA
20	665 sec.	996 sec.
60	1411 sec.	2223 sec.

window size = 15

4.4 Experiments from paraphrasing perspectives

To investigate the upper bound of our method, we carried out several experiments. So far in this paper, we have discussed topic information as an approximation of contextual information by comparing topics annotated by humans and automatically inferred by pLSI and LDA. However, since our goal is to evaluate paraphrases, we need to determine whether latent variable models detect a difference of topics for sentences of paraphrases.

First, we randomly selected 1% of the English seed sentences. Each sentence corresponds to several Japanese sentences, so we could produce Japanese paraphrasing pairs. The number of selected English sentences was 185.

Second, we generated 9,091 Japanese paraphrasing pairs from the English seed sentences. However, identical sentences existed in some generated paraphrasing pairs. In other words, these sentences were simply collected from different

places in the corpus. From a paraphrasing perspective, such pairs are useless. Thus we removed them and randomly selected one pair from one English seed sentence.

Finally, we sampled 117 paraphrasing pairs and evaluated them based on a paraphrasing perspective: whether a paraphrase is contextually independent. There were 71 contextually independent paraphrases and 37 contextually dependent paraphrases. Nine paraphrases had problems, all of which were caused by translation errors. The phrase "contextually independent paraphrases" means that the paraphrases can be used in any context and can be applied as two-way paraphrases. On the other hand, "contextually dependent paraphrases" means that the paraphrases are one-way, and so we have to give consideration to the direction of each paraphrase.

Table 8: Evaluation with manually annotated labels

	independent		dependent	
	same	diff.	same	diff.
1st level	46	25	18	19
2nd level	25	46	11	26

We removed the nine problematic paraphrasing pairs and evaluated the remaining samples with manually annotated topic labels, as shown in Table 8. According to the basic idea of this method, a contextually independent paraphrasing pair should be judged as having the same topic, and a contextually dependent pair should be judged as having a different topic. Thus, we introduced a criterion to evaluate labeling results in terms of an error rate, defined as follows:

$$Error\ rate = \frac{|D_{indep}| + |S_{dep}|}{\#\ of\ judged\ pairs}, \quad (7)$$

where D_{indep} denotes a set that consists of paraphrasing pairs that are judged as having different topics but are contextually independent. On the other hand, S_{dep} denotes a set that consists of paraphrasing pairs that are judged as having the same topic, but are contextually dependent.

For example, from the results in Table 8, the error rate of the results for the first-level topic is 0.398 ((25 + 18)/108), and that for the second-level topic is 0.528 ((46 + 11)/108).

To estimate the upper bound of this method, we also investigated potentially unavoidable errors. Several paraphrasing pairs are used for the exact same topic, but they seem contextually dependent because several words are different. On the other hand, some paraphrasing pairs seem to be used in obviously different topics but are contextually independent. Table 9 shows the investigation results; at least ten paraphrasing pairs seem contextually independent but are actually used in different topics. In addition, there are at least 15 paraphrasing pairs whose topic is obviously the same, but several differences of words make them contextually dependent. Moreover, in this case, the error rate is 0.231 ((15+10)/108), meaning that it is difficult to judge all of the paraphrasing pairs correctly by using only topic (contextual) information. Thus, this method's upper bound of accuracy when using only topic information is estimated to be around 77%.

Table 9: Potential upper bound of this method

human judgement from paraphrasing perspective	human judgement based on topic	
	same	different
independent	61	10
dependent	15	22

We prepared several latent variable models to investigate the performance of the proposed method and applied it to the sampled paraphrasing sentences mentioned above. Table 10 shows the evaluation results.

5 Discussion

First, there is no major performance difference between pLSI and LDA in paraphrasing evaluation. On average, LDA is slightly better than pLSI. Blei et al. showed that LDA outperforms pLSI in (Blei et al., 2003); however, in some of the cases shown in Tables 2 and 3, pLSI outperforms LDA. On the contrary, using a cosine metric, LDA has a significant problem: it loses its distinguishing ability when the number of topics (latent variables) becomes large. With such a large number of topics, LDA always infers a point near the gravity point of the topic simplex. In addition, using a cosine metric also requires a threshold to

Table 10: Evaluating contextual dependency of paraphrases by latent variable models

model (threshold)	window size	independent same	diff.	dependent same	diff.	err. rate	corrected err. rate
pLSI20	11	43	28	14	23	0.3889	0.2048
pLSI20	15	39	32	14	23	0.4259	0.2530
pLSI40	11	33	38	12	25	0.4630	0.3012
pLSI40	15	34	37	16	21	0.4907	0.3373
pLSI20cos(10^{-6})	11	45	26	17	20	0.3981	0.2169
pLSI20cos(10^{-6})	15	31	40	15	22	0.5093	0.3614
pLSI40cos(10^{-6})	11	43	28	17	20	0.4167	0.2410
pLSI40cos(10^{-6})	15	29	42	13	24	0.5093	0.3614
LDA20	11	39	32	19	18	0.4722	0.3133
LDA20	15	42	29	16	21	0.4167	0.2410
LDA40	11	40	31	14	23	0.4167	0.2410
LDA40	15	35	36	15	22	0.4722	0.3133
LDA20cos(0.5)	11	49	22	23	14	0.4167	0.2410
LDA20cos(0.5)	15	51	20	21	16	0.3796	0.1928
LDA40cos(0.5)	11	47	24	18	19	0.3889	0.2048
LDA40cos(0.5)	15	43	28	17	20	0.4167	0.2410
1st-level topic	–	46	25	18	19	0.3981	0.2169

judge a pair of paraphrasing sentences.

From Table 6, LDA seems robust against the inclusion of noisy sentences with a large window, but it is easily affected by a small window. On the other hand, pLSI seems robust against information shortages due to a small window, but it is not effective with a large window. The best performances were shown at window size 15 for both pLSI and LDA, since the average number of sentences in a document (segment) is 18.7, as shown in Section 4.1.

Table 7 shows that in spite of the difference in programing language, pLSI is faster than LDA in practice. In addition, Table 8 reveals that judging the contextual dependency of paraphrasing pairs does not require fine-grained topics.

From the results shown in Table 10, we can conclude that topic inference by latent variable models resembles context judgement by humans as recorded in error rate. However, we note that the error rate was not weighted for contextually independent or dependent results. Error rate is simply a relative index. For example, if there is a result in which all of the inferences reflect the same topic, then the error rate becomes 0.3426. Thus it is important to detect a contextually de-

pendent paraphrase. Considering these points, pLSI20 with window size 11 shows very good results in Table 10.

In Section 4.4, we showed the potential upper bound of this method. The smallest error rate is 0.231, and we can estimate a corrected error by the following formula:

$$\frac{|D_{indep}| + |S_{dep}| - C}{\# \ of \ judged \ pairs - C}, \qquad (8)$$

where C denotes the correction value that corresponds to the number of paraphrasing pairs judged incorrectly with only contextual information. In our experiments, from the results shown in Table 9, C is set to 25. From the results shown in Table 10, we can conclude that the performance of our method is almost the same as that by the manually annotated topics, and the accuracy of our method is almost 80% for paraphrasing pairs that can be judged by contextual information.

There are several possibilities for improving accuracy. One is using a $fixed$ window to obtain contextual information. Irrelevant sentences are sometimes included in fixed windows, and latent variable models fail on inference. If we could infer a boundary of topics with high accuracy,

we would be able to dynamically detect a precise window using some other reliable text models specialized to text segmentation.

So far, we have mainly discussed the contextual dependency of paraphrasing pairs. However, when a paraphrasing pair is contextually dependent, it is also important to infer its specific paraphrasing direction. Unfortunately, we conclude that inferring the paraphrasing direction with contextual information is difficult. In the experimental results, however, there were several examples whose direction could be inferred from their contextual information. Thus, contextual information may benefit the inference of paraphrasing direction. Actually, in the experiments, 11 of 37 contextual dependent pairs had obvious paraphrasing directions. In most of the paraphrasing pairs, different words were used or inserted, or some words were deleted. Thus, to infer a paraphrasing direction, we need more specific information for words or sentences; for example what words carry specific or generic meaning and so on.

One might consider a supervised learning method, such as Support Vector Machine, to infer topics (e.g., (Lane et al., 2004)). However, we cannot know the best number of topics for an application in advance. Thus, a supervised learning method is promising only if we already know the best number of topics for which we can prepare an appropriate learning set.

6 Conclusion

We proposed an evaluation method for the contextual dependency of paraphrasing pairs using two latent variable models, pLSI and LDA. To evaluate a paraphrasing pair, we used sentences surrounding the given sentence as contextual information and approximated context by topics that correspond to a latent variable of a text model. The experimental results with paraphrases automatically extracted from a corpus showed that the proposed method achieved almost 60% accuracy. In addition, there is no major performance difference between pLSI and LDA. However, they have slightly different characteristics: LDA is robust against noisy sentences with long context, while pLSI is robust against information shortage due to short context. The results also revealed that any method's upper bound of accuracy using only contextual information is almost 77%.

References

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the ACL*, pages 50–57.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57.

Ian R. Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2004. Topic classification and verification modeling for out-of-domain utterance detection. In *Proceedings of ICSLP*, pages 2197–2200.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rule for question-answering. *Natural Language Engineering*, 7(4):343–360.

Kiyonori Ohtake and Kazuhide Yamamoto. 2001. Paraphrasing honorifics. In *Workshop Proceedings of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Post-Conference Workshop)*, pages 13–20.

Mitsuo Shimohata and Eiichiro Sumita. 2002. Automatic paraphrasing based on parallel corpus for normalization. In *Proceedings of LREC 2002*, pages 453–457.

Tetsuro Takahashi, Tomoya Iwakura, Ryu Iida, Atsushi Fujita, and Kentaro Inui. 2001. KURA: A transfer-based lexico-structural paraphrasing engine. In *Proceedings of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Workshop)*, pages 37–46.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, pages 147–152.

Kazuhide Yamamoto. 2002. Acquisition of lexical paraphrases from texts. In *Proceedings of the 2nd International Workshop on Computational Terminology (Computerm 2002, in conjunction with Coling 2002)*, pages 22–28.