Lexical Choice via Topic Adaptation for Paraphrasing Written Language to Spoken Language

Nobuhiro Kaji[1] and Sadao Kurohashi[2]

[1] Institute of Industrial Science, The University of Tokyo,
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
kaji@tkl.iis.u-tokyo.ac.jp
[2] Graduate School of Information Science and Technology,
The University of Tokyo, 7-3-1 Hongo,
Bunkyo-ku, Tokyo 113-8656, Japan
kuro@kc.t.u-tokyo.ac.jp

Abstract. Our research aims at developing a system that paraphrases written language text to spoken language style. In such a system, it is important to distinguish between appropriate and inappropriate words in an input text for spoken language. We call this task lexical choice for paraphrasing. In this paper, we describe a method of lexical choice that considers the topic. Basically, our method is based on the word probabilities in written and spoken language corpora. The novelty of our method is topic adaptation. In our framework, the corpora are classified into topic categories, and the probability is estimated using such corpora that have the same topic as input text. The result of evaluation showed the effectiveness of topic adaptation.

1 Introduction

Written language is different from spoken language. That difference has various aspects. For example, spoken language is often ungrammatical, or uses simplified words rather than difficult ones etc. Among these aspects this paper examines difficulty. Difficult words are characteristic of written language and are not appropriate for spoken language.

Our research aims at developing a system that paraphrases written language text into spoken language style. It helps text-to-speech generating natural voice when the input is in written language. In order to create such a system, the following procedure is required: (1) the system has to detect inappropriate words in the input text for spoken language, (2) generate paraphrases of inappropriate words, and (3) confirm that the generated paraphrases are appropriate. This paper examines step (1) and (3), which we call lexical choice for paraphrasing written language to spoken language.

Broadly speaking, lexical choice can be defined as binary classification task: the input is a word and a system outputs whether it is appropriate for spoken

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 981–992, 2005.

language or not. This definition is valid if we can assume that the word difficulty is independent of such factors as context or listeners. However, we think such assumption is not always true. One example is business jargon (or technical term). Generally speaking, business jargon is difficult and inappropriate for spoken language. Notwithstanding, it is often used in business talk. This example implies that the word difficulty is dependent on the topic of text/talk.

In this paper, we define the input of lexical choice as a word and text where it occurs (= the topic). Such definition makes it possible for a system to consider the topic. We think the topic plays an important role in lexical choice, when dealing with such words that are specific to a certain topic, e.g., business jargon. Hereafter, those words are called *topical words*, and others are called non-topical words. Of course, in addition to the topic, we have to consider other factors such as listeners and so on. But, the study of such factors lies outside the scope of this paper.

Based on the above discussion, we describe a method of lexical choice that considers the topic. Basically, our method is based on the word probabilities in written and spoken language corpora. It is reasonable to assume that these two probabilities reflect whether the word is appropriate or not. The novelty of the method is topic adaptation. In order to adapt to the topic of the input text, the corpora are classified into topic categories, and the probability is estimated using such corpora that have the same topic category as the input text. This process enables us to estimate topic-adapted probability. Our method was evaluated by human judges. Experimental results demonstrated that our method can accurately deal with topical words.

This paper is organized as follows. Section 2 represents method overview. Section 3 and Section 4 describe the corpora construction. Section 5 represents learning lexical choice. Section 6 reports experimental results. Section 7 describes related works. We conclude this paper in Section 8.

2 Method Overview

Our method uses written and spoken language corpora classified into topic categories. They are automatically constructed from the WWW. The construction procedure consists of the following two processes (Figure 1).

1. **Style Classification**
   Web pages are downloaded from the WWW, and are classified into written and spoken language style. Those pages classified as written/spoken language are referred as written/spoken language corpus. In this process, we discarded ambiguous pages that are difficult to classify.
2. **Topic Classification**
   The written and spoken language corpora are classified into 14 topic categories, such as arts, computers and so on.

Both classification methods are represented in Section 3 and Section 4.

Given an input word and a text where it occurs, it is decided as follows whether the input word is appropriate or inappropriate for spoken language.

1. The topic category of the input text is decided by the same method as the one used to classify Web pages into topic categories.
2. We estimate the probabilities of the input word in the written and spoken language corpora. We use such corpora that have the same topic as the input text.
3. Using the two probabilities, we decide whether the input word is appropriate or not. Section 5 describes this method.

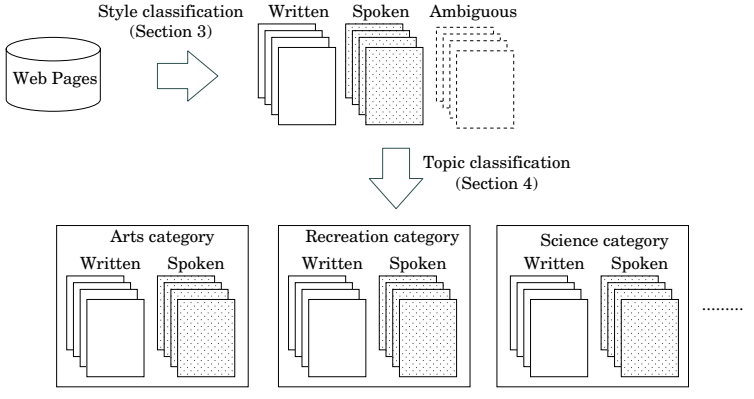

Fig. 1. Written and spoken language corpora construction

3 Style Classification

In order to construct written and spoken language corpora classified into topic categories, first of all, Web pages are classified into written and spoken language pages (Figure 1). Note that what is called spoken language here is not real utterance but chat like texts. Although it is not real spoken language, it works as a good substitute, as some researchers pointed out [2,11].

We follow a method proposed by Kaji et al (2004). Their method classifies Web pages into three types: (1) written language page, (2) spoken language page, and (3) ambiguous page. Then, Web pages classified into type (1) or (2) are used. Ambiguous pages are discarded because classification precision decreases if such pages are used. This Section summarizes their method. See [11] for detail. Note that for this method the target language is Japanese, and its procedure is dependent on Japanese characteristics.

3.1 Basic Idea

Web pages are classified based on interpersonal expressions, which imply an attitude of a speaker toward listeners, such as familiarity, politeness, honor or contempt etc. Interpersonal expressions are often used in spoken language, although

not frequently used in written language. For example, when spoken language is used, one of the most basic situations is face-to-face communication. On the other hand, such situation hardly happens when written language is used.

Therefore, Web pages containing many interpersonal expressions are classified as spoken language, and vice versa. Among interpersonal expressions, such expressions that represent familiarity or politeness are used, because:

- Those two kinds of interpersonal expressions frequently appear in spoken language,
- They are represented by postpositional particle in Japanese and, therefore, are easily recognized as such.

Hereafter, interpersonal expression that represents familiarity/politeness is called familiarity/politeness expression.

3.2 Style Classification Procedure

Web pages are classified into the three types based on the following two ratios:

- Familiarity ratio (F-ratio): '# of sentences including familiarity expressions' divided by '# of all the sentences in the page'.
- Politeness ratio (P-ratio): '# of sentences including politeness expressions' divided by '# of all the sentences in the page'.

The procedure is as follows. First, Web pages are processed by Japanese morphological analyzer JUMAN[3]. And then, in order to calculate F-ratio and P-ratio, sentences which include familiarity or politeness expressions are recognized in the following manner. A sentence is considered to include the familiarity expression, if it has one of the following six postpositional particles: *ne*, *yo*, *wa*, *sa*, *ze*, *na*. A sentence is considered to include the politeness expression, if it has one of the following four postpositional particles: *desu*, *masu*, *kudasai*, *gozaimasu*.

After calculating the two ratios, the page is classified according to the rules illustrated in Figure 2. If F-ratio and P-ratio are equal to 0, the page is classified as written language page. If F-ratio is more than 0.2, or if F-ratio is more than 0.1 and P-ratio is more than 0.2, the page is classified as spoken language page. The other pages are regarded as ambiguous and are discarded.

3.3 The Result

Table 1 shows the number of pages and words (noun, verb, and adjective) in the corpora constructed from the WWW. About 8,680k pages were downloaded from the WWW, and 994k/1,338k were classified as written/spoken language. The rest were classified as ambiguous page and they were discarded. The precision of this method was reported by Kaji et al (2004). According to their experiment, the precision was 94%.

[3] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman-e.html

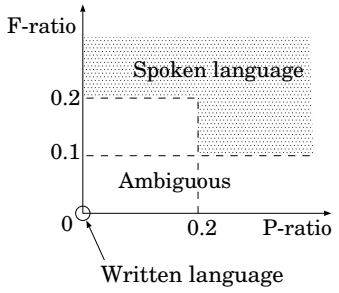

Fig. 2. Style classification rule

Table 1. The size of the written and spoken language corpora

	# of pages	# of words
Written language	989k	432M
Spoken language	1,337k	907M

Table 2. The size of the training and test data

Topic category	Training	Test
Arts	2,834	150
Business & Economy	5,475	289
Computers & Internet	6,156	325
Education	2,943	155
Entertainment	6,221	328
Government	3,131	165
Health	1,800	95
News	2,888	152
Recreation	4,352	230
Reference	1,099	58
Regional	4,423	233
Science	3,868	204
Social Science	5,410	285
Society & Culture	5,208	275

4 Topic Classification

The written and spoken language corpora are classified into 14 topic categories (Figure 1). This task is what is called text categorization. We used Support Vector Machine because it is reported to achieve high performance in this task.The training data was automatically built from Yahoo! Japan[4].

The category provided by Yahoo! Japan have hierarchy structure. For example, there are Arts and Music categories, and Music is one of the subcategories of Arts. We used 14 categories located at the top level of the hierarchy. We downloaded Web pages categorized in one of the 14 categories. Note that we did not use Web pages assigned more than one categories. And then, the Web pages were divided them into 20 segments. One of them was used as the test data, and the others were used as the training data (Table 2). In the Table, the

[4] http://www.yahoo.co.jp/

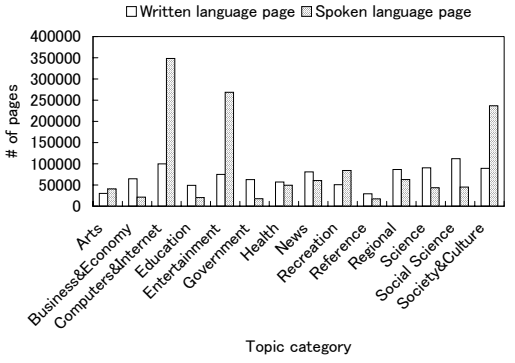

Fig. 3. The size of written and spoken language corpora in each topic category

first column shows the name of the 14 topic categories. The second/third column shows the number of pages in the training/test data.

SVM was trained using the training data. In order to build multi-class classifier, we used One-VS-Rest method. Features of SVM are probabilities of nouns in a page. Kernel function was linear. After the training, it was applied to the test data. The macro-averaged accuracy was 86%.

The written and spoken language corpora constructed from the WWW were classified into 14 categories by SVM. Figure 3 depicts the number of pages in each category.

5 Learning Lexical Choice

We can now construct the written and spoken language corpora classified into topic categories. The next step is discrimination between inappropriate and appropriate words for spoken language using the probabilities in written and spoken language corpora (Section 2). This paper proposes two methods: one is based on Decision Tree (DT), and the other is based on SVM. This Section first describes the creation of gold standard data, which is used for both training and evaluation. Then, we describe the features given to DT and SVM.

5.1 Creation of Gold Standard Data

We prepared data consisting of pairs of a word and binary tag. The tag represents whether that word is inappropriate or appropriate for spoken language. This data is referred as gold standard data. Note that the gold standard is created for each topic category.

Gold standard data of topic T is created as follows.

1. Web pages in topic T are downloaded from Yahoo! Japan, and we sampled words (verbs, nouns, and adjectives) from those pages at random.
2. Three human judges individually mark each word as INAPPROPRIATE, APPROPRIATE or NEUTRAL. NEUTRAL tag is used when a judge cannot mark a word as INAPPROPRIATE or APPROPRIATE with certainty.

3. The three annotations are combined, and single gold standard data is created. A word is marked as INAPPROPRIATE/APPROPRIATE in the gold standard, if
   - All judges agree that it is INAPPROPRIATE/APPROPRIATE, or
   - Two judges agree that it is INAPPROPRIATE/APPROPRIATE and the other marked it as NEUTRAL.

   The other words are not used in the gold standard data.

5.2 The Features

Both DT and SVM use the same three features: the word probability in written language corpus, the word probability in spoken language corpus, and the ratio of the word probability in spoken language corpus to that in written language. Note that when DT and SVM are trained on the gold standard of topic T, the probability is estimated using the corpus in topic T.

6 Evaluation

This Section first reports the gold standard creation. Then, we show that DT and SVM can successfully classify INAPPROPRIATE and APPROPRIATE words in the gold standard. Finally, the effect of topic adaptation is represented.

6.1 The Gold Standard Data

The annotation was performed by three human judges (Judge1, Judge2 and Judge3) on 410 words sampled from Business category, and 445 words sampled from Health category. Then, we created the gold standard data in each category (Table 3). The average Kappa value [3] between the judges was 0.60, which corresponds to substantial agreement.

Table 3. Gold standard data

	Business	Health
INAPPROPRIATE	49	38
APPROPRIATE	267	340
Total	316	378

Table 4. # of words in Business and Health categories corpora

	Business	Health
Written language	29,891k	30,778k
Spoken language	9,018k	32,235k

6.2 Lexical Choice Evaluation

DT and SVM were trained and tested on the gold standard data using Leave-One-Out (LOO) cross validation. DT and SVM were implemented using C4.5[5] and TinySVM[6] packages. The kernel function of SVM was Gaussian RBF. Table

[5] http://www.rulequest.com/Personal/
[6] http://chasen.org/ taku/software/TinySVM/

Table 5. The result of LOO cross validation

Topic	Method	Accuracy	# of correct answers	Precision	Recall
Business	DT	.915 (289/316)	31 + 258 = 289	.775	.660
	SVM	.889 (281/316)	21 + 260 = 281	.750	.429
	MCB	.845 (267/316)	0 + 267 = 267	—	.000
Health	DT	.918 (347/378)	21 + 326 = 347	.600	.552
	SVM	.918 (347/378)	13 + 334 = 347	.684	.342
	MCB	.899 (340/378)	0 + 340 = 340	—	.000

4 shows the number of words in Business and Health categories corpora. Three features described in Section 5 were used.

The result is summarized in Table 5. For example, in Business category, the accuracy of DT was 91.5%. 289 out of 316 words were classified successfully, and the 289 consists of 31 INAPPROPRIATE and 258 APPROPRIATE words. The last two columns show the precision and recall of INAPPROPRIATE words. MCB is Majority Class Baseline, which marks every word as APPROPRIATE.

Judging from the accuracy in Health category, one may think that our method shows only a little improvement over MCB. However, considering other evaluation measures such as recall of INAPPROPRIATE words, it is obvious that the proposed method overwhelms MCB. We would like to emphasize the fact that MCB is not at all practical lexical choice method. If MCB is used, all words in the input text are regarded as appropriate for spoken language and the input is never paraphrased.

One problem of our method is that the recall of INAPPROPRIATE words is low. We think that the reason is as follows. The number of INAPPROPRIATE words in the gold standard is much smaller than that of APPROPRIATE words. Hence, we think a system that is biased to classify words as APPROPRIATE often achieves high accuracy. It is one of future works to improve the recall while keeping high accuracy.

We examined discrimination rules learned by DT. Figure 4 depicts the rules learned by DT when the whole gold standard data of Business category is used as a training data. In the Figure, the horizontal/vertical axis corresponds to the probability in the written/spoken language corpus. Words in the gold standard can be mapped into this two dimension space. INAPPROPRIATE/ APPROPRIATE words are represented by a cross/square. The line represents discrimination rules. Words below the line are classified as INAPPROPRIATE, and the others are classified as APPROPRIATE.

6.3 Effect of Topic Adaptation

Finally, we investigated the effect of topic adaptation by comparing our method to a baseline method that does not consider topic.

Our method consists of two steps: (1) mapping from a word to features, and (2) applying discrimination rules to the features. In the step (1), the probability is estimated using the written and spoken language corpora in a certain topic

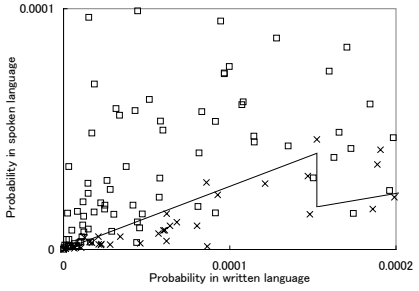

Fig. 4. Decision tree rules in Business category

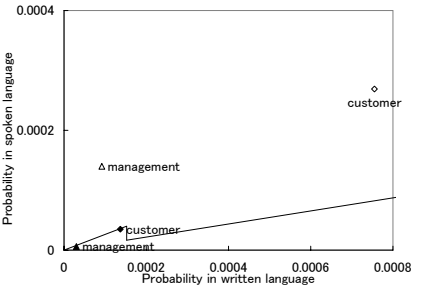

Fig. 5. Examples in Business category

T. In the step (2), discrimination rules are learned by DT using the whole gold standard data of topic T. We used DT rather than SVM because rules are easy for humans to understand. On the other hand, the baseline uses the same discrimination rules as our method, but uses the whole written and spoken language corpora to map a word to features. Hereafter, the two methods are referred as PROPOSED and BASELINE. Both methods use the same discrimination rules but map a word to features in a different way. Therefore, there are such words that are classified as INAPPROPRIATE by PROPOSED and are classified as APPROPRIATE by BASELINE, and vice versa. In the evaluation, we compared the classification results of such words.

We evaluated the results of topical words and non-topical words separately. This is because we think PROPOSED is good at dealing with topical words and hence we can clearly confirm the effectiveness of topic adaptation. Here, a word is regarded as topical word in topic T, if its probabilities in the written and spoken language corpora assigned topic category T are larger than those in the whole corpora with statistical significance (the 5% level). Otherwise it is regarded as non-topical word in topic T. As a statistical test log-likelihood ratio [4] was used. The evaluation procedure is as follows.

1. Web pages in Business category were downloaded from Yahoo! Japan, and words in those pages were classified by the two methods. If the results of the two methods disagree, such words were stocked.
2. From the stocked words, we randomly sampled 50 topical words in Business and 50 non-topical words. Note that we did not use such words that are contained in the gold standard.
3. Using Web pages in Health category, we also sampled 50 topical words in Health and 50 non-topical words in the same manner.
4. As a result, 100 topical words and 100 non-topical words were prepared. For each word, two judges (Judge-A and Judge-B) individually assessed which method successfully classified the word. Some classification results were difficult even for human judges to assess. In such cases, the results of the both methods were regarded as correct.

Table 6 represents the classification accuracy of the 100 topical words. For example, according to assessment by Judge-A, 75 out of 100 words were classified successfully by PROPOSED. Similarly, Table 7 represents the accuracy of the 100 non-topical words. The overall agreement between the two judges according to the Kappa value was 0.56. We compared the result of the two methods using McNemar's test [8], and we found statistically significant difference (the 5% level) in the results. There was no significant difference in the result of non-topical words assessed by the Judge-A.

Table 6. Accuracy of topical words classification

Judge	Method	Accuracy
Judge-A	PROPOSED	75% (75/100)
	BASELINE	52% (52/100)
Judge-B	PROPOSED	72% (72/100)
	BASELINE	53% (53/100)

Table 7. Accuracy of non-topical words classification

Judge	Method	Accuracy
Judge-A	PROPOSED	48% (48/100)
	BASELINE	66% (66/100)
Judge-B	PROPOSED	38% (38/100)
	BASELINE	78% (78/100)

6.4 Discussion and Future Work

PROPOSED outperformed BASELINE in topical words classification. This result indicates that the difficulty of topical words depends on the topic and we have to consider the topic. On the other hand, the result of PROPOSED was not good when applied to non-topical words. We think this result is caused by two reasons: (1) the difficulty of non-topical words is independent of the topic, and (2) BASELINE uses larger corpora than PROPOSED (see Table 1 and Table 4). Therefore, we think this result does not deny the effectiveness of topic adaptation. These results mean that PROPOSED and BASELINE are complementary to each other, and it is effective to combine the two methods: PROPOSED/BASELINE is applied to topical/non-topical words. It is obvious from the experimental results that such combination is effective.

We found that BASELINE is prone to classify topical words as inappropriate and such bias decreases the accuracy. Figure 5 depicts typical examples sampled from topical words in Business. Both judges regarded 'management' and 'customer'[7] as appropriate for spoken language in Business topic. The white triangle and diamond in the Figure represent their features when the probability is estimated using the corpora in Business category. They are located above the line, which corresponds to discrimination rules, and are successfully classified as appropriate by PROPOSED. However, if the probability is estimated using the whole corpora, the features shift to the black triangle and diamond, and BASELINE wrongly classified the two as inappropriate. In Health category, we could observe similar examples such as 'lung cancer' or 'metastasis'.

[7] Our target language is Japanese. Examples illustrated here are translation of the original Japanese words.

These examples can be explained in the following way. Consider topical words in Business. When the probability is estimated using the whole corpora, it is influenced by the topic but Business, where topical words in Business are often inappropriate for spoken language. Therefore, we think that BASELINE is biased to classify topical words as inappropriate.

Besides the lexical choice method addressed in this paper, we proposed lexical paraphrase generation method [10]. Our future direction is to apply these methods to written language texts and evaluate the output of text-to-speech. So far, the methods were tested on a small set of reports.

Although the main focus of this paper is lexical paraphrases, we think that it is also important to deal with structural paraphrases. So far, we implemented a system that paraphrases compound nouns into nominal phrases. It is our future work to build a system that generates other kinds of structural paraphrases.

7 Related Work

Lexical choice has been widely discussed in both paraphrasing and natural language generation (NLG). However, to the best of our knowledge, no researches address topic adaptation. Previous approaches are topic-independent or specific to only certain topic.

Lexical choice has been one of the central issues in NLG. However, the main focus is mapping from concepts to words, (e.g., [1]). In NLG, a work by Edmonds and Hirst is related to our research [5]. They proposed a computational model that represents the connotation of words.

Some paraphrasing researches focus on lexical choice. Murata and Isahara addressed paraphrasing written language to spoken language. They used only probability in spoken language corpus [12]. Kaji et al. also discussed paraphrasing written language to spoken language, and they used the probabilities in written and spoken language corpora [11]. On the other hand, Inkpen et al. examined paraphrasing positive and negative text [9]. They used the computational model proposed by Edmonds and Hirst [5].

The proposed method is based on the probability, which can be considered as a simple language model. In language model works, many researchers have discussed topic adaptation in order to precisely estimate the probability of topical words [6,7,13]. Our work can be regarded as one application of such language model technique.

8 Conclusion

This paper proposed lexical choice method that considers the topic. The method utilizes written and spoken language corpora classified into topic categories, and estimate the word probability that is adapted to the topic of the input text. From the experimental result we could confirm the effectiveness of topic adaptation.

References

1. Berzilay, R., Lee, L.: Bootstrapping Lexical Choice via Multiple-Sequence Alignment. Proceedings of EMNLP. (2002) 50–57
2. Bulyko, I., Ostendorf, M., and Stolcke, A.: Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures. Proceedings of HLT-NAACL (2003) 7–9
3. Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics. **22** (2). (1996) 249–255
4. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics. **19** (1). (1993) 61–74
5. Edmonds, P., Hirst, G.: Near-Synonymy and Lexical Choice. Computational Linguistics. **28** (2). (2002) 105–144
6. Florian, R., Yarowsky, D.: Dynamic Nonlocal Language Modeling via Hierarchical Topic-Based Adaptation: Proceedings of ACL. (1999) 167–174
7. Gildea, D., Hofmann, T.; TOPIC-BASED LANGUAGE MODELS USING EM. Proceedings of EUROSPEECH. (1999) 2167–2170
8. Gillick, L., Cox, S.: Some Statistical Issues in the Comparison of Speech Recognition Algorithms. Proceedings of ICASSP. (1989) 532–535
9. Inkpen, D., Feiguina, O., and Hirst, G.: Generating more-positive and more-negative text. Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text. (2004)
10. Kaji, N., Kawahara, D., Kurohashi, S., and Satoshi, S. : Verb Paraphrase based on Case Frame Alignment. Proceedings of ACL. (2002) 215–222
11. Kaji, N., Okamoto, M., and Kurohasih, S.: Paraphrasing Predicates from Written Language to Spoken Language Using the Web. Proceedings of HLT-NAACL. (2004) 241–248
12. Murata, M., Isahara, H.: Automatic Extraction of Differences Between Spoken and Written Languages, and Automatic Translation from the Written to the Spoken Language. Proceedings of LREC. (2002)
13. Wu, J., Khudanpur, S.: BUILDING A TOPIC-DEPENDENT MAXIMUM ENTROPY MODEL FOR VERY LARGE CORPORA. Proceedings of ICASSP. (2002) 777–780