Detection of Transfer Errors in Automatic Paraphrasing

Atsushi FUJITA
Computational Linguistics Lab.
atsush-f@is.aist-nara.ac.jp
http://cl.aist-nara.ac.jp/˜atsush-f/

1 Introduction

Paraphrasing into simpler and plainer texts is an effective way of assisting users on the Internet to access to the information contained there. Our goal is to automate lexical and syntactic paraphrasing as exhibited by the following examples[1] [2]:

(1) s. She *burst into tears*, and he tried to *console* her.
t. She *cried*, and he tried to *comfort* her.

(2) s1. This car *was sold to* Tom *by* John.
s2. Tom *bought* a car *from* John.
t. John *sold* a car *to* Tom.

2 Transfer errors

One of the major problems in automating lexical and syntactic paraphrasing is in the difficulty of specifying the applicability conditions of each paraphrasing pattern. Paraphrasing patterns with wrong applicability conditions would produce various types of errors in generalizing paraphrases from input, which we call *transfer errors*. We thus need to seek a robust method to detect and correct transfer errors in the post-transfer process.

Our preliminary investigation into transfer errors occurring in paraphrasing of Japanese revealed that the most dominant type of error is incorrect case assignment [1]. Motivated by this observation, we address the task of automatically detecting this type of error. Case assignments can be incorrect at two different levels. (i) Violation of syntactic constraints: in paraphrase (3t), the verb "*tsuranuku*" cannot take the "*ni*" case. (ii) Violation of semantic constraints: in paraphrase (4t), the verb "*katameru*" requires a concrete object for its "*o*" case, but the noun "*kontei*" does not satisfy this constraint.

(3) s. *team play-ni* *tessuru*. (He devotes himself to team play.)
(team play-DAT) (devote-PRES)
t. **team play-ni* *tsuranuku*. (He devotes himself to team play.)
(team play-DAT) (devote-PRES)

(4) s. *building-no* *kiban-o* *katameta*. (He strengthened the foundation of the building.)
(building-MOD) (foundation-DAT) (strengthen-PAST)
t. **building-no* *kontei-o* *katameta*. (*He strengthened the basis of the building.)
(building-MOD) (basis-DAT) (strengthen-PAST)

3 Automatic detection of incorrect case assignments

One may suspect that incorrect case assignments can be detected simply by using a handcrafted case frame dictionary, which describes allowable cases and semantic constraints for each verb. However, in existing case frame dictionaries, since semantic constraints are specified in terms of coarse-grained semantic classes, they are not adequate for the detection of incorrect case assignments. For example, though the difference between two nouns "*kiban* (foundation)" and "*kontei* (basis)" is crucial in the context of example (4), we cannot distinguish these two nouns because they tend to belong to the same semantic class "*basis*".

[1] For each example, s denotes an input and t denotes its paraphrase. Note that our target language is Japanese although example (1) and (2) in this manuscript are the examples in English.

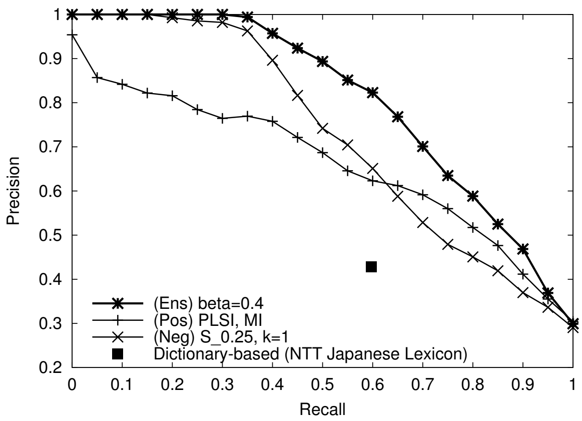

Figure 1: Recall-Precision curves of error detection

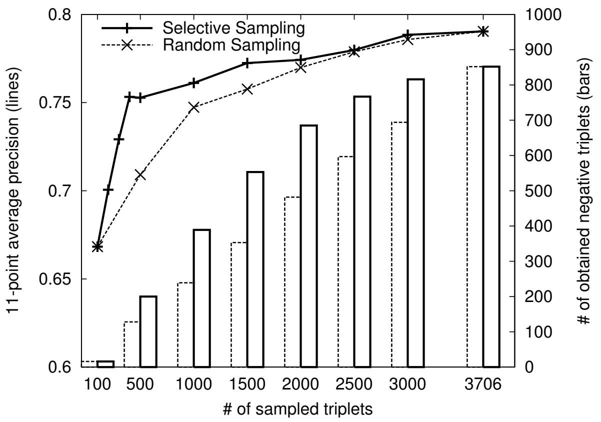

Figure 2: Effects of selective sampling

Instead of relying on semantic classes, we take words themselves into account. Let v, n and c be verb, noun and case markers. The task of detecting incorrect case assignments can be decomposed into the classification of triplet $\langle v, c, n\rangle$ into *correct* or *incorrect*. The correctness of a given paraphrased sentence is determined by combining the classification result for each $\langle v, c, n\rangle$ contained in it. To realize this task, we need to address the following two issues:

1. While positive examples can be collected from existing corpora, negative examples are generally not available. A challenging issue is therefore how to effectively use a limited number of manually collected negative examples in combination with a large number of positive examples.
2. Manual collection of negative examples is costly and time-consuming. Furthermore, any such collection is sparse in relation to the combinatorial space. Hence, we need an effective labeling scheme in order to collect negative examples that truly contribute to error detection.

In respect to these issues, we propose: (i) an empirical method to detect incorrect case assignments, where we make an ensemble of two error detection models Pos and Neg that are separately trained on a large collection of positive examples and a small collection of negative examples, respectively, and (ii) a selective sampling scheme which assigns preference values for each example based on the correctness output by Pos and the sparsity output by Neg.

4 Experimental results

Figure 1 shows the results of the 5-fold cross-validation over 3,169 hand-evaluated paraphrased sentences and 852 manually collected negative examples. Our error detection model (Ens) outperformed the dictionary based model and the two subcomponent models for all ranges of recall, achieving 79.4% 11-point average precision where the 11-points are $0.0, 0.1, \ldots, 1.0$ for recall; a 13.3 point improvement over Pos, the model trained only on positive examples.

Figure 2 shows the result of selective sampling. Here, the bars show the numbers of obtained negative examples, while the curves show the change in performance. Our preference function efficiently chose negative examples that effectively contributed to error detection. The improvement was remarkable particularly at the early stage of learning. The selective sampling achieved 75% 11-point precision with 400 samples, where the random sampling achieved the same precision with 1,500.

5 Conclusion

We focused on the issue of transfer errors in paraphrasing. Motivated by observation of the tendency of transfer errors, we addressed the automatic detection of incorrect case assignments. Our empirical method was justified through empirical experiments.

References

[1] A. Fujita and K. Inui. Exploring transfer errors in lexical and structural paraphrasing. *Journal of Information Processing Society of Japan*, Vol. 44, No. 11, Nov., 2003 (to appear, in Japanese).

[2] K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. Text simplification for reading assistance: a project note. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, pp.9-16, Jul., 2003.