

Corpus-based Extraction of Modals in Consecutive Sentences

Robert Chartrand^a, Shunsuke Nakamoto^a, Hidenobu Kunichika^b, Akira Takeuchi^a

^a*Department of Artificial Intelligence,*

^b*Department of Creation Informatics,*

Faculty of Computer Science, Kyushu Institute of Technology, Japan

robert@minnie.ai.kyutech.ac.jp

Abstract: This work presents a method of learning modal auxiliaries. In our method, two consecutive uses of modals are given as example sentences. These are beneficial for learners to recognize the flow of thought in context. We also present a method of extracting example sentences from a corpus. Here sentences in the corpus are parsed and analyzed to extract only the main clauses, which are essential parts to convey the flow of ideas. Extracted sentences are simplified and examples can be retrieved in many different situations, therefore, examples can be useful for both teachers and learners alike.

Keywords: Corpus linguistics, modal auxiliary, computer-assister language learning

Introduction

In this paper, we aim to demonstrate a method to extract useful examples from two consecutive sentences using modals to use as a learning tool for English students or teachers. We conducted our analysis by using the British National Corpus (BNC) 2007 XML edition [1].

1. Consecutive use of modals

Modal auxiliaries are among the most difficult structures to teach to students of English as a second or foreign language (ESL/EFL) [2]. Because other languages often use different structures to convey the ideas expressed in English by modals, learners of English frequently make mistakes with modals [3]. Although there are a large number of grammar books that explain the form and meaning of the English modals, understanding the meaning of modals is difficult when only the explanation is offered. One solution is to apply corpus linguistics methods and show learners examples of authentic usage of modals [4]. In order to highlight the meaning/usage of modals, we focused on the consecutive use of modals. Useful example sentences can be extracted from a corpus by using computational linguistics techniques for parsing and simplifying sentences for learners to study the use of modal auxiliaries. In Example 1, “would, should” are extracted from two consecutive sentences:

IT REALLY WOULD WORK AFTER ALL
WE SHOULD KNOW SOON ENOUGH

Example 1: Consecutive modals extracted with our method

Extracting two consecutive sentences using modals is an important feature of our method. The combination of the two sentences makes the flow of thought more intuitive and offers a greater chance of comprehension to the learner by using deductive reasoning. The use of context and vocabulary allow for the learner to relate these two modals consecutively and appropriately. A single sentence will not have as much intrinsic impact and will offer no suggestions as to how to employ another modal in the vicinity of the first modal.

2. Extracting example sentences

While the BNC contains a large number of useful sentences containing modals, the process of automatically extracting consecutive use of modals involves certain difficulties. We considered only the main clauses for our analysis. We determined that by removing the subordinate clauses, this made the connotation of the sentence easier to understand for English language learners, and when presenting two consecutive sentences, the meaning becomes more evident. There is also a possibility that a single sentence from the BNC includes multiple main clauses. In this case, all main clauses are extracted as separate sentences.

To achieve this task, it was necessary to parse the sentences automatically, thus we used the Charniak Parser [5,6]. After parsing the sentences, we processed the output in order to recognize the information. The sentences were modified to remove the unnecessary subordinate clauses and the t-scores were calculated to determine the most likely occurrence of the modals (data not shown.) A higher t-score signifies greater confidence that there is a viable association between these two words. We can thus use this information for selecting practical example sentences. A learner could make use of our system to search for examples of the pairs of modals to see how these expressions are used in the English language.

3. Conclusion

The retrieval of information from a corpus and the processing of the result are important to find practical examples of modals in consecutive sentences. Therefore, a method of dealing with vast amounts of information is necessary to achieve this purpose. In our method, we utilized our knowledge of the English language to ascertain some attributes of modals in order to achieve the best results. Learners and educators could make use of these results to improve the quality of writing, a better understanding of modals and improving the process of learning a difficult aspect of the English language.

References

- [1] Research Technologies Service at Oxford University Computing Services. (2007). Retrieved April 27, 2008 from <http://www.natcorp.ox.ac.uk/XMLedition/URG/>
- [2] Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course (2nd Edition)*. New York: Heinle and Heinle.
- [3] Coelho, E. (2004). *Adding English: A guide to teaching in multilingual classrooms*. Don Mills: Pippin.
- [4] Meyer, C. F. (2002). *English Corpus Linguistics*. Cambridge: Cambridge University Press.
- [5] Charniak, E. (2000). A Maximum-entropy-inspired Parser. *Proceedings of NAACL*. 132-139.
- [6] Charniak, E. & Johnson, M. (2005). Coarse-to-file n-best parsing and MaxEnt discriminative reranking, *Proceedings of ACL*. 173-180.