# A Study on the Segmentation for Dialogue

Kanayo Ogura

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
September 2001

**Keywords:** Discourse segmentation, Cue words , Communicative phrases , Texttiling algorithm, Maximum entropy method,

In recent years new technologies like the Internet and cellular phones have greatly changed our communication environment, and created new communication concepts like email and chat, which have the characteristics of anonymity and/or asynchronism. Currently this change has effects mainly on younger generation, but will spread or is now spreading across all generations. These technologies affect not only our every day situation, but also education and bussiness environments. With this new technologies, as the role of communication is getting more important, supporting effective andefficient communication is recognized to be necessary. However the communication technologies have changed, linguistic exchanges are fundamental to communication. In every day conversation, we achieve efficient communication without much effort through these linguistic exchanges. There might exist differences among communication patterns by the media such as face-to-face, computer-mediated and through cellular phones and the characteristics of the partner --- human-to-human vs. human-to-machine ---, if the knowledge on linguistic exchanges could be utilized for supporting communication, it would contribute to achieving effective and efficient communication.

The purpose of the research is to examine our knowledge on segmenting dialogues with topic boundary, which can be used for building computer dialogue systems and automatic summarization of dialogues.Clue words such as discourse markers, fillers, conjunctives, interjectives and communicative phrases and the word distribution proposed in the text tiling algorithm are examined to be the factors or accurate segmentation.

50 task-oriented dialogues, consisting of 36 scheduling dialogues and 14 dialogues of various tasks were used for evaluation.The amount of dialogues are not sufficient, but a variety of the tasks enables us to examine useful suggestions.

In 14 dialogues, conjunctives, discourse markers and fillers are manually annotated. Using these discourse markers can achieve 85.1% precision. In 36 dialogues, instead of using manually annotated information, using the results of morphological analysis, conjuncives can achieve 71.7% precision and 32.7% recall (72.0% precision and 28.0% recall in 14 dialogues). Communicative phrases extracted from 14 dialogues by hand achieve 82.8% precision and 60.6% recall for the same dialogues, while achieved 57.1% precision and 44.9% recall for 36
dialogues. Using word distribution --- term frequency and the number of new words --- can achieve only 30 -- 50% precision and 60 —80% recall. The maximum entropy method is used for selecting and combining clue words and word distribution and achieved 59.4% precision and 78.8% recall. The features of conjunctives, communicative phrases, interjectives, word distribution of term frequency and the number of new words are found to contribute to accurate segmentation.