密级：公开

机器翻译分词研究

中国科学技术信息研究所

2011 年 6 月

中国科学技术信息研究所

中国科学技术信息研究所

机器翻译分词研究

Research on Chinese Word Segmentation in Machine Translation

课题组长：王惠临

课题成员：石崇德 何彦青 张均胜 刘丹

中国科学技术信息研究所

中国科学技术信息研究所

中国科学技术信息研究所

摘 要

本研究对中文分词研究的历史、现状、难点以及主要切分方法进行了广泛调研，并对机器翻译中使用断字、机械分词和字标注分词等不同分词方法进行了对比实验，通过分析和统计不同切分方法对机器翻译的影响，总结和归纳了包括总分词数、分词词表数和分词错误等影响机器翻译的主要因素，以作为机器翻译中分词优化方向的指导。

中国科学技术信息研究所

中国科学技术信息研究所

Abstract

This research focused on how different Chinese Word Segmentation (CWS) influences machine translation. We firstly broadly examined the history, current state, difficulties, and mostly used paradigms of CWS. Then we compared different CWS methods like char-based, dictionary-based and char-labeling-based segmentation and how they worked in statistical machine translation systems. According to different impact of these methods in machine translation, we conclude that three factors in different segmentation play the most important part in machine translation: total number of segmented words, total number of words in vocabulary and segment error. These three factors is our direction of optimize CWS in machine translation.

中国科学技术信息研究所

中国科学技术信息研究所

目　录

中国科学技术信息研究所

中国科学技术信息研究所