

# EINLADUNG

Zeit: 10. Juli 2008, 13:45 Uhr  
Ort: AH V, Ahornstr. 55  
Referent: Shahram Khadivi, M. Sc. Comp. Eng.  
Titel: Statistical Computer-Assisted Translation

## Abstract:

In recent years, significant improvements have been achieved in statistical machine translation (MT), but still even the best machine translation technology is far from replacing or even competing with human translators. However, an MT system helps to increase the productivity of human translators. Usually, human translators edit the MT system output to correct the errors, or they may edit the source text to limit vocabulary. A way of increasing the productivity of the whole translation process (MT plus human work) is to incorporate the human correction activities in the translation process, thereby shifting the MT paradigm to that of computer-assisted translation (CAT). In a CAT system, the human translator begins to type the translation of a given source text; by typing each character the MT system interactively offers and enhances the completion of the translation. Human translator may continue typing or accept the whole completion or part of it. Here, we will use a fully fledged translation system, phrase-based MT, to develop computer-assisted translation systems. An important factor in a CAT system is the response time of the MT system. We will describe an efficient search space representation using word hypotheses graphs, so as to guarantee a fast response time. The experiments will be done on a small and a large standard task.

Skilled human translators are faster in dictating than typing the translations, therefore a desired feature of a CAT system is the integration of human speech into the CAT system. In a CAT system with integrated speech, two sources of information are available to recognize the speech input: the target language speech and the given source language text. The target language speech is a human-produced translation of the source language text. The main challenge in the integration of the automatic speech recognition (ASR) and the MT models in a CAT system, is the search. The search in the MT and in the ASR systems are already very complex, therefore a full single search to combine the ASR and the MT models will considerably increase the complexity. In addition, a full single search becomes more complex since there is not any specific model nor any appropriate training data. In this work, we study different methods to integrate the ASR and the MT models. We propose several new integration methods based on N-best list and word graph rescoring strategies. We study the integration of both single-word based MT and phrase-based MT with ASR models. The experiments are performed on a standard large task, namely the European parliament plenary sessions.

A CAT system might be equipped with a memory-based module that does not actually translate, but find the translation from a large database of exact or similar matches from sentences or phrases that are already known. Such a database, known as bilingual corpora are also essential in training the statistical machine translation models. Therefore, having a larger database means a more accurate and faster translation system. In this thesis, we will also investigate the efficient ways to compile bilingual sentence-aligned corpora from the Internet. We propose two new methods for sentence alignment. The first one is a typical extension of the existing methods in the field of sentence alignment for parallel texts. We will show how we can employ sentence-length based models, word-to-word translation models, cognates, bilingual lexica, and any other features in an efficient way. In the second method, we propose a new method for aligning sentences based on bipartite graph matching. We show that this new algorithm has a competitive performance with other methods for parallel corpora, and at the same time its very useful in handling different order of sentences in a source text and its corresponding translation text. Further, we propose an efficient way to recognize and filter out wrong sentence pairs from the bilingual corpora.

## Formlos überreicht mit der Bitte um Kenntnisnahme und Aushang

Prof. Dr. B. Vöcking, Informatik 1  
Dr. M. Westermann, DFG-Nachwuchsgruppenleiter, Informatik 1  
Prof. Dr. P. Rossmaniith, Informatik 1  
Prof. Dr. Ir. J.-P. Katoen, Informatik 2  
Prof. (em.) Dr. K. Indermark, Informatik 2  
Prof. Dr. J. Giesl, Informatik 2  
Prof. Dr. M. Nagl, Informatik 3  
Prof. Dr. H. Lichter, Informatik 3  
Prof. Dr. O. Spaniol, Informatik 4  
Prof. Dr. K. Wehrle, Informatik 4  
Prof. Dr. M. Jarke, Informatik 5  
Prof. G. Lakemeyer, Informatik 5  
Prof. Dr. H. Ney, Informatik 6  
Prof. Dr. W. Thomas, Informatik 7  
Prof. (em.) Dr. W. Oberschelp, Informatik 7  
Prof. Dr. E. Grädel, Informatik 7  
Prof. Dr. L. Kobbelt, Informatik 8  
Prof. Dr. Th. Seidl, Informatik 9  
Prof. Dr. U. Schroeder, Informatik 9  
Prof. Dr. J. Borchers, Informatik 10  
Prof. Dr. J. Kowalewski, Informatik 11  
Prof. C. Bischof, Ph.D, Informatik 12  
Prof. Dr. U. Naumann, Informatik 12  
Prof. Dr. J. Roßmann, Lehrstuhl f. Mensch-Maschine-Interaktion  
Prof. Dr. F. Wolf, Helmholtz-Nachwuchsgruppe  
Prof. Dr. M. Wiegner, Lehrstuhl I f. Mathematik  
Prof. Dr. E. Triesch, Lehrstuhl II f. Mathematik für Ingenieure  
Prof. Dr. W. Dahmen, Lehrstuhl f. Mathematik und Institut f. Geometrie und Praktische Mathematik  
Prof. Dr. A. Reusken, Lehrstuhl f. Numerische Mathematik  
Prof. Dr. A. Krieg, Lehrstuhl A f. Mathematik  
Prof. Dr. W. Plesken, Lehrstuhl B f. Mathematik  
Prof. Dr. H. Th. Jongen, Lehrstuhl C f. Mathematik  
Prof. Dr. G. Hiß, Lehrstuhl D f. Mathematik  
Prof. Dr. G. Nebe, Lehr- und Forschungsgebiet Mathematik (Algebra)  
Prof. Dr. E. Zerz, Lehr- und Forschungsgebiet Mathematik  
Prof. Dr. M. Herty, Lehr- und Forschungsgebiet Mathematik  
Prof. Dr. U. Kamps, Lehrstuhl f. Statistik  
Prof. Dr. H.-H. Bock, Lehr- und Forschungsgebiet Angewandte Statistik  
Prof. Dr. A. Steland, Lehrstuhl f. Stochastik  
Prof. Dr. E. Cramer, Lehr- und Forschungsgebiet Angewandte Stochastik  
Prof. Dr. D. Bothe, MathCCES, Pauwelsstraße  
Prof. Dr. R. Mathar, Lehrstuhl f. theoretische Informationstechnik  
Prof. Dr. T. Noll, Lehrstuhl f. Allgemeine Elektrotechnik und Datenverarbeitungssysteme  
Prof. Dr. Th. Bemmerl, Lehrstuhl f. Betriebssysteme  
Prof. Dr. J.-R. Ohm, Lehrstuhl f. Elektrische Nachrichtentechnik  
Prof. Dr. P. Vary, Lehrstuhl u. Institut f. Nachrichtengeräte u. Datenverarbeitung  
Prof. Dr. B. Rembold, Lehrstuhl u. Institut f. Hochfrequenztechnik  
Prof. Dr. T. Aach, Lehrstuhl f. Bildverarbeitung  
Prof. Dr. G. Ascheid, Lehrstuhl f. Integrierte Systeme d. Signalverarbeitung  
Prof. Dr. B. Walke, Lehrstuhl f. Kommunikationsnetze  
Prof. Dr. W. Michaeli, Institut f. Kunststoffverarbeitung  
Prof. Dr. D. Abel, Institut f. Regelungstechnik  
Prof. Dr. G. Hirt, Institut f. Bildsame Formgebung  
Prof. Dr. M. Bastian, Lehrstuhl f. Wirtschaftsinformatik