Abstract

This thesis presents scientific contributions to the field of multimodal interactive structured prediction (MISP). The aim of MISP is to reduce the human effort required to supervise an automatic output, in an efficient and ergonomic way. Hence, this thesis focuses on the two aspects of MISP systems. The first aspect, which refers to the interactive part of MISP, is the study of strategies for efficient human–computer collaboration to produce error-free outputs. Multimodality, the second aspect, deals with other more ergonomic modalities of communication with the computer rather than keyboard and mouse.

To begin with, in sequential interaction the user is assumed to supervise the output from left-to-right so that errors are corrected in sequential order. We study the problem under the decision theory framework and define an optimum decoding algorithm. The optimum algorithm is compared to the usually applied, standard approach. Experimental results on several tasks suggests that the optimum algorithm is slightly better than the standard algorithm.

In contrast to sequential interaction, in active interaction it is the system that decides what should be given to the user for supervision. On the one hand, user supervision can be reduced if the user is required to supervise only the outputs that the system expects to be erroneous. In this respect, we define a strategy that retrieves first the outputs with highest expected error first. Moreover, we prove that this strategy is optimum under certain conditions, which is validated by experimental results. On the other hand, if the goal is to reduce the number of corrections, active interaction works by selecting elements, one by one, e.g., words of a given output to be supervised by the user. For this case, several strategies are compared. Unlike the previous case, the strategy that performs better is to choose the element with highest confidence, which coincides with the findings of the optimum algorithm for sequential interaction. However, this also suggests that minimizing effort and supervision are contradictory goals.

With respect to the multimodality aspect, this thesis delves into techniques to make multimodal systems more robust. To achieve that, multimodal systems are improved by providing contextual information of the application at hand. First, we study how to integrate e-pen interaction in a machine translation task. We contribute to the state-of-the-art by leveraging the information from

the source sentence. Several strategies are compared basically grouped into two approaches: inspired by word-based translation models and n-grams generated from a phrase-based system. The experiments show that the former outperforms the latter for this task. Furthermore, the results present remarkable improvements against not using contextual information. Second, similar experiments are conducted on a speech-enabled interface for interactive machine translation. The improvements over the baseline are also noticeable. However, in this case, phrase-based models perform much better than word-based models. We attribute that to the fact that acoustic models are poorer estimations than morphologic models and, thus, they benefit more from the language model. Finally, similar techniques are proposed for dictation of handwritten documents. The results show that speech and handwritten recognition can be combined in an effective way.

Finally, an evaluation with real users is carried out to compare an interactive machine translation prototype with a post-editing prototype. The results of the study reveal that users are very sensitive to the usability aspects of the user interface. Therefore, usability is a crucial aspect to consider in an human evaluation that can hinder the real benefits of the technology being evaluated. Hopefully, once usability problems are fixed, the evaluation indicates that users are more favorable to work with the interactive machine translation system than to the post-editing system.