Providing sufficient parallel corpora essentially boosts the quality of machine translation system. Parallel texts – as the most important resource in statistical machine translation (SMT) – appear to be limited in quantity, genre and language coverage. Recent research work has focused on exploring comparable corpora, which contain bilingual information. This information compensates the existing parallel texts with additional vocabularies and phrase translation candidates. Therefore, our goal is to find a new method that exploits comparable corpora for collecting parallel data.

Munteanu and Marcu (2005, 2006) have developed two systems for mining parallel fragments and sentences from comparable corpora. However, they left several issues unsolved: 1) in the work of Munteanu and Marcu (2006), they cannot measure the correlation of extracted fragments due to a lack of metrics that could determine whether the pair is equivalent translation; 2) each of their presented solutions is restricted to just one of the two relevant levels of extraction: sentential and sub-sentential fragments. To address these problems, we propose a modified IBM Model 1 for fragment detection, and use two-level classifiers for further verifying both sentences and sub-sentential fragments; in this two-level classification step, more features are investigated and utilized for improving the accuracy of the results.

The evaluation is conducted in similar-domain and out-domain translation test corpora of the German-English language pair. We compare the proposed method with the re-implemented system of Munteanu and Marcu (2006). The results show that our framework achieves BLEU score improvements of up to 0.98 %. Moreover, our experiments on different domains and training corpus sizes show the potential of future enhancement.