In-memory Translation Spotting Using Big Data Analysis

Jan Scherbaum

Abstract

Localisation service providers look for ways to minimise the time needed for translations. Automated machine translation techniques do not provide quality of translations, which could be used in professional services. Translation memories are databases of previous translations which are collected in an attempt to reuse previously professionally translated segments of text. Translation memory systems however pose a number of usability issues. Translation spotting is a technique of identifying a translation of a given segment of text from an aligned sentence, solving some of the usability issues of translation memory systems. Previously explored translation spotting techniques were using statistical translation models. With the emerging trend of in-memory databases, big data analysis becomes possible in a very performant fashion.

This dissertation explores the extent to which in-memory databases could be used to perform translation spotting with big data analysis. An in-memory translation memory framework is designed and implemented. A number of algorithms are proposed and implemented, which analyse the co-occurrence of patterns in order to identify translations of queries.

Both the translation memory framework and the translation spotting algorithms are then evaluated in terms of extensibility, effectiveness and efficiency. It is found, that the proposed single word algorithms yield up to 94% of effectiveness, scale very well and are able to perform in real-time. The proposed multi-word algorithms yield up to 76% effectiveness, even though their performance varies greatly based on the query and size of the analysed data set.

The limitations of the implemented system are identified and a number of improvements are presented as avenues for potential future research.