INTRAFIND

# Opponent review of the Master Thesis

Title: „Analysis of Word Embeddings"

Author: Matthias Thurnbauer

Opponent: Christoph Goller, PhD, Director Research, IntraFind Software AG, Email: christoph.goller@intrafind.com

The goal of the theses was to investigate whether word embeddings can be used to improve traditional search engines (information retrieval systems) especially in the context of Enterprise Search. The intended application implies the necessity of domain-specific training with customer data on standard hardware for on premise installations. Three questions had to be answered: (1) How much computation time / power is needed to train standard word embeddings? (2) Can they be used to solve the vocabulary mismatch problem (words used in queries differ from those used in documents)? (3) Can they be used to identify ambiguous single-word queries and provide additional information for query refinement to the user?

For the provided master thesis extensive literature research was performed, which is reflected by the thoroughly chosen publications cited in the work. The thesis is written in a very precise way with a clear structure.

Extensive performance tests were conducted and documented in chapter 5. They show that training word embeddings for a mid-size corpus like the German Wikipedia is feasible on standard server hardware. The actual neural network training is much faster than the tokenization preprocessing step. Chapter 6 presents results for solving the vocabulary mismatch problem. The author demonstrates a thorough understanding of the underlying neural network architectures and analyses their influence on word similarity. With the knowledge and deep understanding of the network architectures acquired in chapter 6 the author tackles the challenging problem of identifying word ambiguity in chapter 7. The interesting result is that he is not only able to identify and describe (label) different meanings of ambiguous words. By using a density-based clustering algorithm, he also presents a method to decide whether a word is ambiguous or not.

Overall, I recommend the theses as an excellent work with the highest possible rating.

Dr. Christoph Goller