Vortrag von Tobias Eder, BA-Betreuer(in) Alexander Fraser, Fabienne Braune
Thema: Exploiting Bilingual Word Embeddings to Establish Translational Equivalence

In dieser Arbeit geht es darum bilinguale Word Embeddings zu finden um eine Übersetzung ohne externes Wörterbuch durchführen zu können. Hier liegt das Augenmerk vor allem auf Fachbereiche für die es keine bis wenig Ressourcen gibt, z.B. Medizin. Zunächst gibt der Vortragende einen Überblick über das Thema. Der Vortrag gliedert sich wie folgt:

1) Word Embeddings
2) Vektorraummodelle
3) lineare Abbildungen
4) Korpora und Experimentaufbau
5) Weitere Schritte/Regularisierung

Professor Schulz warf noch ein, das Evaluierung auch ein wichtiger Punkt sei.

zu 1)
Word Embedding ist eine Technik zum erlernen von Merkmalen in NLP. Wörter oder Phrasen werden dabei mit Vektoren repräsentiert. Hier dient das Word Embedding hauptsächlich dafür das Datenknappheitsproblem (engl. sparseness) zu verhindern. Professor Schulz ergänzt dabei noch, das das Datenknappheitsproblem nie vollständig behoben werden kann, da Wörter atomare Einheiten sind. Word2Vec ist ein Modell um Word Embeddings zu erstellen. Dabei werden stetige Wortsackmodelle (CBOW) und stetige Skipgramme verwendet. Ein anderes Beispiel ist FastText. Hier enthält ein Wort eine Subwort Information, z.B. Buchstaben n-Gramme.

zu 2)
In Vektorraummodellen werden syntaktische und semantische Beziehungen von Wörtern dargestellt. Je näher zwei Wörter beieinander liegen umso „ähnlicher“ sind sie sich syntaktisch und semantisch.

zu 3)
Der Vortragende erwähnt zwei Arten von Regression: Die lineare Regression und die „ridge“ Regression. Erstere ist ein statistisches Verfahren wodurch in diesem Fall ein Wort mithilfe von anderen Wörtern repräsentiert wird. „ridge“ Regression, auch Tikhonov Regularisierung genannt, bezeichnet hier das verringern zu hoher Gewichte.

zu 4)
Vier verschiedene Korpora werden zum trainieren und testen eingesetzt: ein „medical“ Korpus, ein „general“ Korpus, der EMEA Korpus (aus dem medizinischen Fachbereich) und die TED Talks (gesprochene Sprache). Des Weiteren wird noch ein kleiner Paralleltext-Korpus, für die Englisch - Deutsch Übersetzung verwendet. Für unterschiedliche Word Embeddings wird CBOW verwendet.

zu 5)
Der nächste Schritt in der Arbeit ist es herauszufinden wie man am besten mit niedrig frequentieren Wörtern umgeht. Welche Regualriesierungsmethoden sind dafür besonders geeignet? Wie geht man mit dem OOV (Out of Vocabulary) Problem um? Welche Evaluierungsmethode ist geeignet?

Am Ende wurde noch auf ein Paper von Tomas Mikolov zum Thema Word Embeddings hingewiesen.