CIS, LMU München
Fach: Kolloquium Computerlinguistisches Arbeiten SS2017
Studentin: Faridis Alberteris Azar

Protokoll zur Sitzung am 15.05.2017

Thema: Exploiting Bilingual Word Embeddings to Establish Translational Equivalence

Student: Tobias Eder
Betreuer/in: Alexander Fraser und Fabienne Braune

Wenn man Dokumente über spezifische Themen übersetzen muss, stoßt man gegen das Problem, dass es nicht immer die entsprechenden Wörterbücher existieren oder wenn es doch gibt, sind manchmal manche Begriffe nicht vorhanden. Diese Tatsache hat den Studenten Tobias Eder motiviert, im Rahmen seiner Bachelorarbeit eine Übersetzung ohne Wörterbuch zu schaffen.

Nachdem Tobias Eder seine Motivation am Anfang der Präsentation ausgedruckt hat, hat er das *Toolkit* erklärt, die er benutzt, um sein Ziel zu erreichen:

1) Wordembedding
Das *Wordembedding* ist die Repräsentation von Wörtern in einem Vektor. Dieser Vektor ist hochdimensional und die Wörter mit den ähnlichen Bedeutungen sind näher von einander. Also im Vektorraum lassen sich die Wörter *clustern*. An dieser Stelle der Präsentation wurde über einen großen Nachteil des *Wordembeddinds* diskutiert, und zwar über das *Sparse*- Data-Problem.Das bedeutet, das es wenige Trainingsdaten zur Verfügung stehen. Wenn die Wörter nicht im Trainingsdaten vorkommen, bieten *Wordembeddings* und insgesamt probabilistische Modelle keine Lösung an.

2) Word2Vec
Tobias Eder hat erklärt, dass Word2vec eine Gruppe von Modellen ist, die um die Erzeugung von *Wordembeddings* verwendet wird. Es wurde von einem Forscher- Team unter der Leitung von Tomas Mikolov im Rahmen eines Google- Projektes erfunden und im Jahr 2013 veröffentlicht. Das Ziel dieser Modellen ist, vorhersagen zu können, welches Wort gemäß des Kontextes an der Stelle kommt.

3) FastText
FastText, laut Tobias Eder, ist Teil einer Facebook Recherche, die im Jahr 2016 stattgefunden hat. Es handelt sich um das *Learning* von Wortrepräsentationen (engl.: *word representations*). Jedes Wort wird in a *bag of n-grams* repräsentiert. Diese Methode ist vorteilhaft, denn sie ermöglicht ein schnelles Training für ein großes Korpus. Dieses *Tool* kann auch für die Textklassifikation verwendet werden. Aber wie Tobias Eder klargestellt hat, wird diese Applikation nicht im Rahmen seiner Bachelorarbeit angewendet.

4) Lineare Abbildungen
Auf der Basis dieses mathematischen Verfahrens werden, so Eder, zwei Vektorräume verglichen, in denen vektorielle Repräsentationen von Wörtern aus jeweils 2 verschiedenen Domains bzw. Sprachen enthalten sind. Der Student zieht die Variante der linearen Regression bevor, denn es

geeigneter und einfacher für seine Zwecke ist. Eine Nachteil ist es in diesem Fall, dass die Resultaten mit diesem Verfahren sehr nah an die Trainingsdaten drankommen können. Um das *Overfitting* zu vermeiden, wendet der Student Tobias Eder die L-2- Regularisierung an.

Für diese Bachelorarbeit werden unterschiedliche parallele Korpora verwendet, wie zum Beispiel Medical Big(ca. 50M Tokens), EMEA (ca. 4M Tokens) oder TED Talks (ca. 2M Tokens)

Der Vortragende hat geschildert, dass er unterschiedliche *Embeddings* von unterschiedlichen Modellen nimmt. Danach baut er ein kleines Englisch- Deutsch- Korpus mit ca. 5000 Wörter, die bereits übersetzt wurden. Aus diesem Korpus wählt er die 1000 Hochfrequenten Wörter und testet dann für sie, wie gut die Übersetzung stattgefunden hat. Edler benutzt ein domainspezifisches Testset, mit denen er die Differenzen bei den Performances der unterschiedlichen Modelle überprüft (wo haben sie am besten funktioniert)

Kurz vor dem Ende des Vortrages hat er weitere Schritte erwähnt:

- Es wird überprüft, wie die Modelle bei niedrigfrequenten Wörtern funktionieren.
- Wie bessere Abbildungen erhalten werden können.
- Es wird nach anderen Regularisierungsmethoden geschaut (evtl. werden extra Faktoren zur Regularisierung hinzugefügt)
- Evaluation auf OOV- Wörter

Am Ende der Präsentation hat Tobias Eder seine Bibliographie gezeigt und einige interessante Artikel empfohlen, unter denen das Paper von Tomas Mokolov (oben erwähnt) unter dem Titel *Linguistic Regularities in Continous Space Word Representation*, wo in das Thema Word2Vec vertieft werden kann.