

Protokoll zur Sitzung am 18. Mai 2015

Den ersten Vortrag des Tages hält [REDACTED], die von Herrn Dr. Zangenfeind betreut wird, über das Thema ***Lexikalische Funktionen für die maschinelle Übersetzung***. Unter Lexikalischen Funktionen versteht man im Allgemeinen syntagmatische und paradigmatische semantische Relationen zwischen Lexemen einer Sprache. Es gibt ungefähr siebzig Lexikalische Funktionen. Diese werden in zwei große Gruppen eingeteilt, nämlich Standard- und Nichtstandardfunktionen. Wichtig für die maschinelle Sprachverarbeitung sind vor Allem die Standardfunktionen, weil sie einen großen Argument- und Wertebereich haben und somit auch eine große Gültigkeit im Wortschatz besitzen. Nichtstandardfunktionen hingegen dienen zur Beschreibung von Kollokationen und haben einen relativ kleinen Gültigkeitsbereich, weil sie selten vorkommen. Es gibt zwei Arten von Lexikalischen Funktionen. Paradigmatische Funktionen nennt man auch lexikalische Substitute, das sind Ausdrücke, die durch andere Ausdrücke ersetzt werden. Ein Beispiel hierfür ist “Syn(Linguist)”, was man mit dem Begriff Sprachwissenschaftler übersetzen kann. Syntagmatische Funktionen sind lexikalische Parameter und beziehen sich auf die Kombinierbarkeit von Wörtern. Ein Beispiel ist “Magn(verletzt)”, was ‘schwer verletzt’ bedeutet. Eine wichtige Rolle spielen auch Lexikalische Funktionen für sogenannte Stützverbkonstruktionen, wie beispielsweise “Oper”, bei dem das Subjekt im Vordergrund steht, und ‘Func’, bei dem das Objekt im Vordergrund steht. Ein Modell ist für die Bachelorarbeit von zentraler Bedeutung, nämlich das sogenannte Bedeutung-Text-Modell. Gegründet wurde es unter anderem von Melcuk in den Sechziger Jahren. Es besteht im Wesentlichen aus zwei Komponenten, nämlich einer Grammatik und einem Wörterbuch. Die Grammatik dient als Translator, der Bedeutungen von formaler in natürliche Sprache und umgekehrt überträgt. Das Wörterbuch gibt Informationen über die Bedeutungen und Kombinierbarkeit von Lexemen. Am Schluss wird ETAP-3 vorgestellt, ein Übersetzungstool für die Russische Sprache.

Als nächstes ist [REDACTED] an der Reihe, der über das Thema ***WITTFIND: Semi-automatische Korrektur des Highlighting im Facsimile*** referiert und dessen Arbeit von Herrn Dr. Hadersbeck betreut wird. Es geht dabei um die Texterkennung in Bilddateien. Zum Einsatz kommen zwei Typen von Dokumenten aus dem Wittgenstein-Nachlass, nämlich “TS-213”, ein sogenanntes Typoskript, das heißt, gedruckter Text, und “MS-114”, ein Manuskript, das als Bilddatei vorliegt. Als Programmiersprachen dienen Python und Shell. Mithilfe von OCR (Optical Character Recognition), werden aus dem Typoskript Text und aus dem Manuskript Koordinaten extrahiert. Als Software wird hier Tesseract verwendet, die beste Open Source OCR-Software, die es zurzeit gibt. Bei der Extraktion geht man zeilenweise, und innerhalb der Seiten zeilenweise vor. Der Vergleich von der OCR-Ausgabe mit den Bilddaten geschieht folgendermaßen: Jede Zeile der Edition wird mit jeder Zeile aus der OCR verglichen, wobei top-down vorgegangen wird. Die Entscheidung, welcher Satz welcher Stelle im Bild entspricht, wird nach minimaler Levenstein-Distanz getroffen. Einige Funktionalitäten, die das Korrektursystem bereitstellen soll, ist die Aufteilung von zu bearbeitenden Seiten in Pakete, desweiteren eine Auswahlmöglichkeit des Bearbeitungsmodus, zum Beispiel “zeilenweise” oder “satzweise”.

Auch soll das System für den Benutzer übersichtlich gestaltet sein. Abschließend lässt sich sagen, dass sich Probleme hinsichtlich der Tatsache ergeben, dass es manchmal Unstimmigkeiten zwischen Bild und Edition gibt, oder auch, dass oft unbrauchbarer Output erzeugt wird. Für die restliche Bearbeitungszeit steht noch die Aufgabe an, einen besseren Algorithmus zur Aligrierung zu finden.

Den letzten Vortrag des heutigen Tages hält [REDACTED] über ihre Bachelorarbeit mit dem Thema ***Slot Filling based on Open Relation Extraction*** unter Leitung von Frau Heike Adel. Das Slot Filling ist ein Task wie auch Semeval und erfordert ein System zur Informationsextraktion aus großen Dokumenten, mit dessen Hilfe man Antworten auf Anfragen mit bestimmten Slots erhält. Es gibt insgesamt einundvierzig Slots, fünfundzwanzig davon haben mit Personen zu tun und sechzehn mit Organisationen. Als Slots gelten beispielsweise die Beziehung "Organisation - Tochtergesellschaften", die "Eltern - Kind-Beziehung" oder die Staatsangehörigkeit. Im Gegensatz zu vielen anderen Systemen, die Maschine Learning verwenden, wird in dieser Arbeit aus Pattern Matching zurückgegriffen. Die zu bearbeitenden Daten sind in Trainings-, Development- und Testdaten aufgeteilt. Das System, das bei der Arbeit zum Einsatz kommt, ist "OPEN IE" (Open Information Extraction). Es ist im Jahr 2007 bekannt geworden und zeichnet sich dadurch aus, dass es mit sehr großen Korpora arbeiten kann. Außerdem bildet es eine optimale Basis für das Pattern Matching.

Formal:

- * **übersichtliche Aufteilung**
- * **BA- Bertreuer, Name d. Referenten und Thema (durch Hervorhebung) im Text**
- * **Rechtschreibung, Kommasetzung**

Inhaltlich:

- * **ergebnisorientiert und nachvollziehbar**
- * **ausreichende Erläuterungen und Beispiele**
- * **Diskussion aufs Wesentliche beschränkt**