Homework 1.2

Fabian Otto – Matrikelnummer: 2792549
Foundations of Language Technology

27. Oktober 2017

1 Most challenging tasks in NLP

Gemäß der „Natural Language Processing Group“ von Google [2] ist das Ziel die Analyse, das Verständnis und die Generierung von Sprachen, die von Menschen verwendet werden, um eine Menschen-gleiche Kommunikation zwischen Mensch und Maschine zu ermöglichen. Im Rahmen dessen stellen sich der Gruppe Herausforderungen, wie die Mehrdeutigkeit der menschlichen Sprache. Dies schließt Konstrukte der Ironie bzw. des Sarkasmus oder Wortwitze ebenso ein wie auch die kontextbezogene Bedeutung von Wörtern oder Phrasen. Bezogen auf den Kontext ist es z.B. möglich, dass Kennedys berühmte Worte „Ich bin ein Berliner“ für eine Auswertung nicht zum gewünschten Ergebnis führen sondern Kennedy sich als ein Gebäckstück bezeichnet.

Auf einer detaillierter Ebene stellt sich diese Problematik deutlicher heraus, hier bietet z.B. eine Einführung in die Thematik aus dem „Journal of the American Medical Informatics Association“ [4] eine gute Zusammenfassung. In dieser werden unter anderem die Verarbeitungen von Abkürzungen, wie „Dr.“. „Prof.“, in Verbindung mit der Bildung von Tokens angeführt. Es ist bei der *Tokenization* schwer sicherzustellen, dass Tokens immer korrekt gebildet werden. Dies lässt sich neben den oben genannten Abkürzungen auch an einem weiteren Beispiel zeigen: Die Verwendung eines „/“ Zeichens kann im Fall von „Winter-/Sommerterm“ die Bildung von mehr als einem Token benötigen. Jedoch für den Fall $/h oder mg/Tag ist ausschließlich eine Bildung von einem Token sinnvoll. Dies lässt sich für weitere Fälle zeigen und stellt ein wesentliches Problem im NLP dar (Allgemein das Thema Information Retrieval auch von Manning [3] in einer Cambridge Publikation behandelt. Weiterhin finden sich auch im Rahmen von Part-of-Speech (*POS*) Tagging durch Sonderfälle, Mehrdeutigkeiten, etc. Herausforderungen, da Muster oder statistische Wahrscheinlichkeiten zu falschen Ergebnissen führen, z.B. „to book“ und „book“. Zudem finden sich auch Forschungsgebiete im Bereich der *Lemmatization*, d.h. das Finden eines gemeinsamen Grundbausteins (z.B. gerannt $\longrightarrow$ rennen). Zusätzlich stellt *Chunking*, d.h. die Erkennung von zusammengehörigen Phrasen, wie „das Haus“, einen weiteren Bereich mit Herausforderungen für NLP dar.

Von Fracis Bond [1] werden für einen Menschen einfach erkennbare Tatsachen, wie Rechtschreibfehler, Groß- und Kleinschreibung und Trennzeichen benannt. Dies fällt zum Teil auch in den schon oben beschriebenen Bereich der Tokenbildung. Hierbei seien auch Umgangssprache oder ähnliche Konzepte erwähnt, die eine korrektes POS Tagging oder eine sinnvolle Interpretation erschweren.

Zusammenfassend ist damit festzustellen, dass sich in NLP viele Herausforderungen finden. Beginnend mit der korrekten Bildung von Tokens/Trennung von Worten oder Phrasen zusammen mit dem korrekten POS Tagging, um anschließend zusätzlich die Problematik der inhaltlichen Erkennung mit den dazugehörigen Mehrdeutigkeiten und Interpretationsmöglichkeiten in bestimmten Kontexten zu betrachten. Die Probleme erstrecken sich dabei über die üblichen Schichten der linguistischen Analyse:

- Phonetics & Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics

2 Quellen

- Francis Bond, Overview of NLP, Main Issues, 2014. `http://compling.hss.ntu.edu.sg/courses/hg8003.2014/pdf/wk-01.pdf`, aufgerufen am 25.10.2017.
- Google, Natural Language Processing Group, 2017. `https://www.microsoft.com/en-us/research/group/natural-language-processing/`, aufgerufen am 25.10.2017.
- Christopher Manning, et. al., An Introduction to Information Retrieval, 2009.`https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf`, aufgerufen am 25.10.2017.
- Prakash Nadkarni, et. al., Natural language processing: an introduction, 2001. Journal of the American Medical Informatics Association, Volume 18, Issue 5, Pages 544–551, `https://doi.org/10.1136/amiajnl-2011-000464`, aufgerufen am 25.10.2017.