Technologievorrauschau OneTIPP

Autor: SE / v1 / 14.2.2016

Zielstellung:

- Time to Market verkürzen durch optimierten Technologieeinsatz
- SEMPRIA bietet interessante Technologie ->aber: Abhängigkeit von Technologieanbieter

Lösungsansätze:

1. Nutzung von Technologie ähnlich der von SEMPIRA (kurz: SEMP)
2. Ausbau von Technologie basierend auf DEEP Learning (kurz: DEEP)

Lösungsansatz „SEMP“ (Natural Language Generation):

1. **Prozessablaufschritt 1 „Text planning“** (Informationen aus Wissensdatenbanken abrufen)
    a. Aufbau eigene Datenbanken
        i. WikiPedia Inhalte (http://dumps.wikimedia.org/dewiki/latest/)
        ii. FreeBase Datenbank (https://www.freebase.com)
        iii. Eigene Fachwort und Synonymdatenbanken (s. Roadmap)
        iv. Eigene Themendatenbanken (s. Roadmap)
        v. Inhaltsdatenbanken mit Semantischen Verknüpfungen und Beziehungen (s. Roadmap)
        vi. Statistische Daten: https://www.govdata.de/ | http://www.data.gov/
    b. Einsatz von Software und Algorithmen
        i. Sentiment Analysis („Stimmungsanalyse“: Positive, neutrale oder negative Stimmung des Textes erkennen -> Teil des Autorenprofils)
        ii. Named Entity Recognition („Entity Extraktion“: Eigennamen, Orte, Zeiten oder ähnliches herausgezogen und klassifiziert werden)
        iii. Folgende Open Source Software für Named Entity Recognition ist vorhanden:
            1. https://en.wikipedia.org/wiki/General_Architecture_for_Text_Engineering
            2. https://en.wikipedia.org/wiki/OpenNLP
    c. Zusammenfassung:
        i. Abruf aller Semantischen Verknüpfungen, die zwischen dem Eingabetext und den Inhalten unserer Datenbanken stehen
        ii. Festlegen der Textstimmung und der Named Entities, die nicht verändert werden dürfen und im Zieltext wieder vorhanden sein müssen

2. **Prozessablaufschritt 2 „Sentence planning“**
    a. Word sense disambiguation („Sinnbestimmung eines Wortes“: Welches Wort passt in den Kontext des Textes -> wichtig für Synonymaustausch
    b. Einsatz von Software und Algorithmen für „Word sense disambiguation“:
        i. http://aclweb.org/aclwiki/index.php?title=Word_sense_disambiguation_resources
        ii. https://en.wikipedia.org/wiki/SemEval
        iii. http://www.sfs.uni-tuebingen.de/GermaNet/ und Wrapper https://github.com/wroberts/pygermanet
    c. Information Extraction („Schlüsselinformationen extrahieren“: automatisch strukturierte Informationen aus unstrukturierten maschinell lesbaren Dokumenten erstellen)
        i. Verwendung moderner Verfahren -> Sequence models (HMM, CMM, CRF) vgl. https://en.wikipedia.org/wiki/Information_extraction#Approaches)
        ii. Einsatz von Software und Algorithmen

1. https://github.com/mimno/Mallet
2. https://github.com/recski/HunTag
3. https://github.com/tpeng/python-crfsuite
4. https://wapiti.limsi.fr/
5. http://taku910.github.io/crfpp/
6. https://en.wikipedia.org/wiki/Conditional_random_field

3. **Prozessablaufschritt 3 „Text Realization“**
    a. Semantic Parser with Execution („Semantischer Parser“: Zerlegung der Eingabetextinhalte in semantische Repräsentationen -> Hintergrund: https://web.stanford.edu/class/cs224u/materials/cs224u-sempre-slides.pdf
    b. Semantisches Parsen muss die Named Entities und die Inhalte aus Information Extraction beinhalten
    c. Einsatz von Software und Algorithmen für Semantic Parsing:
        i. https://github.com/percyliang/sempre
        ii. https://github.com/opencog/link-grammar
        iii. https://github.com/opencog/opencog/tree/master/opencog/nlp/relex2logic
    d. Natural Language Generation („Spracherstellung“):
        i. http://wiki.opencog.org/wikihome/index.php/Natural_language_generation
    e. Einsatz von Software und Algorithmen
        i. https://launchpad.net/nlgen2
        ii. https://github.com/opencog/opencog/tree/master/opencog/nlp/sureal
        iii.

Lösungsansatz „DEEP“ (Deep Learning Natural Language Generation):

Frage: Können wir ein AI oder Maschine Learning System erstellen, welches z.b eine Art "Character-Aware Neural Language Models" sichert, bei dem man ein Seq2Seq Modell nachschaltet. Diesem Seq2Seq Modell müssen Parameter wie Person, Locality, Action, Organization, Nomephraseelements etc übergeben werden und das Seq2Seq Modell erstellt daraus neue Texte, mit den vorgegebenen Elementen`?

https://github.com/carpedm20/neural-summary-tensorflow
https://github.com/carpedm20/lstm-char-cnn-tensorflow

Dieses AI System könnte dann mit diesem Punkt 0 - bis Punkt 5. System trainiert werden und zusätzlich unser Autorenprofil enthalten.

http://alias-i.com/lingpipe/

https://gate.ac.uk/

https://github.com/opencog/opencog