Technologievorrauschau OneTIPP

Autor: SE / v1.2 / 16.2.2016

Zielstellung:

- Time to Market verkürzen durch optimierten Technologieeinsatz
- SEMPRIA bietet interessante Technologie ->aber: Abhängigkeit von Technologieanbieter
    - Aufbau eigener Technologie, angelehnt an Sempria Technologie -> Aufbau v. Assets, Inhouse, Unabhängigkeit, Skalierbarkeit

Lösungsansätze:

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

These:

- Mit Lösungsansatz SEMP oder DEEP oder Kombination Beider kann die Funktionalität der Sempria GmbH und deren Umformulierungsdienst ganz oder teilweise abgebildet werden?
- Ansatz DEEP und/oder SEMP ist die Zukunftstechnologie von OneTIPP, sodass in Roadmap Phase B2B der 6 monatige Forschungsaufwand entfällt und die genannte Software nur noch für unsere Zwecke angepasst und trainiert werden muss -> sprich wir direkt ein Produkt entwickeln können, ohne langwierigen Forschungsaufwand
- Hintergrund:
    - Autorenprofil muss natürlich erforscht werden.
    - Jedoch sind wir mit diesem DEEP/SEMP Ansatz kein Forschungsstartup mehr und können gleich SEED Kapital einsammeln und wenige Zeit später ein eigenes Produkt aufzeigen.

Fragen an Gerhard:

- Bitte schau dir bei beiden Lösungsansätzen die grünen Elemente genau an, schau dich bitte auch auf den grün Website Links um und lies dich ein wenig ein.
- Wie schätzt du die aufgestellte These auf? Ist es realistisch mit dieser (grün hervorgehobenen Algorithmen bzw. Softwarelinks) Anleitung das richtige Software Grundgerüst für OneTIPP gefunden zu haben? Wo siehst du Probleme? Bitte um Prognose.
- Welchen Ansatz bevorzugst du, bzw. schein langfristig der wertvollere für uns zu sein (DEEP oder SEMP)? Wertvoll unter den Stichworten: „eigene Assets“, „weitere Finanzierungsrunden“, „Skalierbarkeit“.
- Ist ein anpassbares Seq2Seq Modell mit Parameterübergabe technisch zu realisieren? Bitte mit Zeitaufwand beim Einsatz von diesen fachlichen IT Experten: 1 bzw. 5 bzw. 10 bzw. 25 Entwickler?
- Kann nur das Seq2Seq Modell + Parameter schon allein für sich (eine mögliche) Lösung für OneTIPP sein?

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. Wörterbuch Deutsch: https://sourceforge.net/projects/germandict/files/?source=navbar
        vii. 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. https://github.com/opencog/opencog/tree/master/opencog/nlp
        1. Demo: microplanning
        2. Demo: sureal

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

Fragestellung:

1. Können wir „Character-Aware Neural Language Models“ sinnvoll als DEEP Learning Ansatz für OneTIPP verwenden?
    a. Vgl: http://arxiv.org/abs/1508.06615
    b. PDF: http://arxiv.org/pdf/1508.06615v4.pdf
    c. Quellcode: https://github.com/carpedm20/lstm-char-cnn-tensorflow
2. Können wir „Neural Attention Model for Abstractive Sentence Summarization“ sinnvoll als DEEP Learning Ansatz für OneTIPP verwenden?
    a. Vgl: http://arxiv.org/abs/1509.00685
    b. PDF: http://arxiv.org/pdf/1509.00685v2.pdf
    c. Quellcode: https://github.com/carpedm20/neural-summary-tensorflow
3. Können wir ein Sequence to Sequence als Deep Learning Modell erstellen, bei dem wir dem trainierten Seq2Seq Modell Parameter übergeben (z.B.: Ort, Person, Handlung, Organisation, Nomenphrasen etc) und unser Advanced Seq2Seq Modell dann sinnvolle Sätze erstellt, wobei die Parameter sinnvoll integriert werden

Interessante Software:

- http://alias-i.com/lingpipe/
- https://gate.ac.uk/
- https://github.com/opencog/opencog