

Computerlinguistisches Arbeiten

Protokoll zur Sitzung am 22.6.15: Kurzvorstellung von drei Bachelorarbeit-Themen

BA-Betreuer: Dr. Vu

Building data resources for sentiment analysis in Twitter for Russian Language

Dies ist eine Fortsetzung des Vortrags von IG vom 15.6.15. IG fasste noch einmal kurz die wichtigsten Begriffe wie Polaritätsanalyse, Datenakquisition und mögliche Bausteine eines Tweets zusammen.

IG verwendet die SVM-Methodik (Support Vector Machine), die bereits 1963 ihren Ursprung hat und die populärste Art ist, um Instanzen zu klassifizieren. IG setzt Multi-class-Klassifizierung ein, um ihre Twitterdaten in "positiv, negativ und neutral" zu trennen. Sie nutzt scikit-learn, ein Open Source Machine Learning Paket in Python. Dabei liegt als Idee zugrunde, dass ein ML-Algorithmus bessere Ergebnisse liefert, wenn er Daten, für die er trainieren soll, auswählen kann. Dieses "active learning" bereitet IG so vor, dass Tweets mit weniger als 5 Wörtern weggelassen werden und ebenso Tweets, bei denen nur weniger als ein Viertel seiner Wörter im Wörterbuch stehen.

Als bestes Resultat nannte IG eine 83%-ige Trefferquote. Nach fünf Iterationen über Datenpakete von jeweils 300 Tweets stellte sich eine Verbesserung auf 88% ein. Als weitere Ansätze zur Verbesserung zählte IG auf, die Trainingsdaten zu vergrößern und somit auch ihr Sentiment-Lexikon zu erweitern, ähnliche Tweets auszusortieren, mehr Features für negative Twittersätze zu finden, auf den Kontext einzugehen und n-Gramme einzubinden sowie auch Suffixe und Präfixe zu analysieren.

BA-Betreuer: Dr. Müller

Multilinguale Eigennamenerkennung

MS charakterisierte die multilinguale Eigennamenerkennung (Named Entity Recognition) als wichtigen vorausgehenden Schritt von NLP-Anwendungen, z.B. der Informationsextraktion, und den Zusatz multilingual als relevant für Anwendungen wie Business Intelligence.

Zu Anfang hat MS textuelle Daten in den Sprachen Englisch, Deutsch, Niederländisch und Spanisch gesammelt und verglichen. Ihre Features - etwa 20 - sind für alle Sprachen gleich, während die Modelle auf Spracheigenheiten eingehen.

Zur Historie der Eigennamenerkennung, beispielsweise in der Kategorisierung "Person, Ort, Organisation und Anderes", nannte MS die 6. Message Understanding Conference MUC 1995 für Englisch sowie CoNLL (Computational Natural Language Learning) 2002 und 2003 für Englisch, Deutsch, Niederländisch und Spanisch. Früher wurden eher regelbasierte Techniken für NER angewendet, heutzutage sind es eher statistische: Machine Learning, semi-supervised Techniken, Hidden Markow, SVM, Maximum Entropie und - als besonders gut bewährt - CRT (Conditional Random Field, ein probabilistisches, graphisches Modell für die Wahrscheinlichkeit $P(\text{Eingabesatz}|\text{Label})$). Letzteres verwendet MS auch in ihrer Arbeit bzw. den CIS-eigenen, morphologischen MarMoT-Tagger, der als generisches Conditional Random Field Framework dient

Die Trainingsdaten von MS stammen von CoNLL 2002 und 2003. Sie sind nach dem BIO-System markiert. Als Ansätze, die Ergebnisse zu verbessern, nannte MS die Erhöhung der Markow-Ordnung bzw. der Kontext-Tiefe.

Nun zählte MS ihre Features auf. Features auf Wortebene sind: das Token selbst und ein Fenster von +/- 2 Token. Dies erbrachte ein Ergebnis von 78% Treffern im Englischen und 35% im Deutschen. Durch Hinzunahme von Groß- und Kleinschreibung, Worttyp-Informationen, Prefix- und Suffix-Werten sowie POS- und Lemma-Angaben verbesserte sich das Ergebnis auf 88,76% bzw. 64,31%. Eine weitere Steigerung auf 91,46% und 71,92% brachten der Einsatz von Clustern (für die Generierung wurde das MarLiN Tool von Martin, Lierman and Ney verwendet) und Lookup-Wörterbüchern. Außerdem zog MS noch Wikipedia als multilinguale Resource für millionenfache Beiträge in allen Sprachen der Welt hinzu. Hier positionierte sie ihre Features auf die Info- und Kategorienboxen. Bei den Kategoriearten nimmt sie - bezogen auf deren Häufigkeitsverteilung - nur den Mittelteil her, nicht die zu seltenen oder die zu häufigen. Zusätzlich sucht sie in den erststehenden Erläuterungssätzen nach dem Kontext von "to be"-Konstruktionen. Dies brachte für die Trefferrate im Englischen eine Steigerung auf 91,40%. Problematisch ist dabei allerdings der Zeitfaktor für die Durchsuchung. Außerdem hat sich herausgestellt, dass für das Niederländische IOBES-Tags statt nur BIO günstiger wären.

Als allgemeine Probleme nannte MS noch Inkonsistenzen in den Daten oder deren Annotationen, Ambiguitäten und sprachspezifische Eigenheiten.

Zur Optimierung der Parameter ihres CRF-orientierten Modells zählte MS auf: das Korrigieren von zu hohen Gewichtungen, die aus Overfitting resultieren, das Durchprobieren der Markow-Ordnungen 1 bis 4 und weiterer Stellgrößen. Hierbei stellte MS fest, dass höhere Markow-Ordnungen die Ergebnisse für Englisch und Spanisch verschlechterten und dass sich aber bei Deutsch eine Verbesserung einstellte.

MS lieferte als Schlussfolgerungen, dass sich bereits mit einfachen Features akzeptable Ergebnisse erzielen lassen und dass Wikipedia und seine Kategorien viel zu einem guten Ergebnis beitragen, vor allem, wenn man Disambiguitäten beseitigt und noch mehr auf die Verlinkungsstruktur eingeht.

Dr. Schulz bemerkte, dass ein F-Score um die 80% zwar als eine hohe Zahl erscheint, in der Praxis aber mit einer Trefferrate von 4-Gute:1-Schlechter "keinen Käufer" finden würde. Er würde empfehlen, statt der etwa 60.000 in Wikipedia beschriebenen Personen und Orte lieber auf Vornamen- und Nachnamen- sowie Orte-Lexika zurückzugreifen, die 300.000 Einträge aufweisen. Auch Lexika für Firmen- und Menschenbezeichner wie "Herr, Direktor" oder Wortendungen wären hilfreich.

MS hatte in ihrer vorausgegangenen BA-Kurzzvorstellung jedoch bemerkt, dass nur Wikipedia für Eigennamen, die aus sehr vielen Token bestehen wie z.B. Film- oder Buchtitel, so gut geeignet sei und Multilingualität aus sich selbst heraus unterstütze.

 , BA-Betreuer: Dr. Frankwein

Lexikalische Funktionen bei der Übersetzung vom Russischen ins Deutsche

TS begann mit der Erklärung von Begriffen zu ihrer BA. Das ETAP3-Bedeutung-Text-Modell ist ein Sprachmodell mit einer vollständigen Sprachbeschreibung aus Grammatik und erklärend-kombinatorischem Wörterbuch, welches alle relevanten Informationen zu sprachlichen Eigenheiten und den Verknüpfungen eines Wortes zu anderen enthält. Es wurde um 1960 von Igor Melcuk entwickelt mit dem Ziel, alle konventionalisierten lexikalischen Relationen und Funktionen zu versammeln, um ein mühsames Zusammensuchen von Einzelinformationen zu vermeiden. ETAP3 wird vor allem für maschinelles Übersetzen, Parsen und Paraphrasieren verwendet.

TS zeigte ein ausführliches Schaubild für das generelle Schema des Übersetzen eines Eingabesatzes in einen Zielsatz. Dabei sind drei Zwischensprachen und mehrere Hilfsverzeichnisse für die morphologische und syntaktische Analyse, das Parsen, die Normalisierung (z.B. Entfernen von Artikeln), den Transfer, die Expansion und syntaktische und morphologische Synthese erforderlich.

Als Definition der lexikalischen Funktion zeigte TS die Formel $f(L)=\{L'\}$, mit L' seien Lexeme, Kollokationen oder Idiome gemeint. Zusätzlich gelten die Bedingungen der semantischen Homogenität von $f(L)$, z.B. "schließen:öffnen"="hindern:helfen", und der Maximalität von $f(L)$, z.B. "Bon(entscheiden)=mutig, überlegt, klug, richtig" und der phraseologische Charakter von $f(L)$.

TS erstellte unter diesen Vorbedingungen aus einer Liste von 68 lexikalischen Funktionen für zwanzig vorgegebene deutsche Verben deren ETAP3-Einträge. Das Heraussuchen und Prüfen der Kandidaten für ihre Einträge erfolgte anhand eines Lexikons und durch das Untersuchen von Beispielsätzen mit einer typischen Verwendung des betrachteten Verbs. Schwierig war oft das Finden von Gegenteilen.

Als Beispiel nannte TS das einfache Verb "öffnen". Sie fand als Belegung für die Nominalisierung "Öffnung", als Gegenteil "zumachen", als Verstärkung "weit, vollständig", als Gegenteil-Verstärkung "einen Spalt". Ziel ist es letztendlich, für die Übersetzung vom Russischen ins Deutsche die bestmöglichen Wortkombinationen zu finden.

Formal:

- * **übersichtliche Aufteilung**
- * **Name d. Referenten, BA- Betreuer, Thema**
- * **Rechtschreibung, Kommasetzung, Darstellung von Formeln**

Inhaltlich:

- * **ergebnisorientiert und nachvollziehbar**
- * **ausreichende Erläuterungen und Beispiele**
- * **Diskussion auf wesentliche Informationen beschränkt**