Statusbericht: Sentimentanalyse auf Amazon-Reviews

Designing Experiments for Machine Learning Tasks

bei Éva Mújdricza-Maydt

Institut für Computerlinguistik

Ruprecht-Karls-Universität Heidelberg

Caroline Berg Simon Will

16. Dezember 2015

1 Motivation

Ziel unseres Projektes ist es, mit Methoden des Maschinellen Lernens aus den Texten von Amazon-Reviews, die zu einer bestimmten Produktklasse gehören, die jeweilige zum Review gehörige Bewertung (1–5 Sterne) vorherzusagen. Dabei möchten wir feststellen, ob die Käufer ihre Bewertungen so eindeutig formulieren, dass es mithilfe von Maschinellem Lernen möglich ist, sie entsprechend ihrer angegebenen Bewertung zu klassifizieren. Besonderes Augenmerk wollen wir darauf legen, herauszufinden, welche Attribute am hilfreichsten beim Lernen sind.

2 Ressourcen

2.1 Amazon-Reviews

Datengrundlage unseres Projekts sind Amazon-Bewertungen von Smartphones. Um diese Texte herunterzuladen, verwenden wir ein Python-Skript (s. 3.1), welches die entsprechende Überblicksseite auf Amazon aufruft und die Review-Texte zusammen mit Meta-Informationen wie der Bewertung abspeichert.

2.2 SentiWS

SentiWS ist eine öffentlich zugängliche deutschsprachige Datenbank, für *sentiment analysis* und *opinion mining*, wobei die dort aufgeführten Wörter im Intervall $[-1; 1]$ je nach ihrem *sentiment* gewichtet und mit einem POS-Tag versehen sind. (Remus, Quasthoff und Heyer 2010, vgl. S. 1168)

3 Module

3.1 Amazon-Scraper

Das Modul `vilperg_amazonreview.py` ermöglicht den Zugriff auf Amazon-Produkte und Amazon-Reviews. Das Skript `write_amazon_reviews.py` schaut auf einer Überlicksseite von Amazon[1] alle angezeigten Produkte und deren Reviews an und lädt Information über sie herunter. Dann wird die darauf folgende Überlicksseite aufgerufen und Informationen über die angezeigten Produkte und Reviews heruntergeladen, usw.

3.2 `arff`-Ersteller und Attribut-Ersteller

Es soll ein Skript geschrieben werden, das eine `arff`-Datei erstellt. Die Attribute für die `arff`-Datei sollen dabei beim Aufrufen des Skripts angegeben werden können. So können dann dynamisch verschiedene `arff`-Dateien erstellt werden.

4 Attribute

4.1 overall_sentiment

Das Attribut overall_sentiment errechnet sich aus den *sentiments* der Wörter in jedem einzelnen Review. Die jeweiligen *sentiments* beziehen wir aus dem Datensatz SentiWS. Wörter, die nicht in SentiWS aufgeführt werden, werden ignoriert.

4.2 Sentiments der Wortarten

Die *sentiments* werden noch einmal für verschiedene Wortarten einzeln erfasst: noun_sentiment, verb_sentiment und adjective_sentiment. Hierbei wird mithilfe der POS-Tags für die jeweilige Wortart ein sentiment-Wert errechnet, analog zu overall_sentiment.

4.3 Sentiments zu Schlüsselwörter

Es sollen für Domäne der Reviews (z. B. Smartphones) die Schlüsselwörter gefunden werden, indem die Texte aus der Domäne mit einem größeren Textaussschnitt allgemeinerer Texte verglichen werdenlüsselwörter gefunden werden, indem die Texte aus der Domäne mit einem größeren Textaussschnitt allgemeinerer Texte verglichen werden. Für jedes Schlüsselwort soll dann ein Attribut erstellt werden, dessen Wert sich aus den *sentiments* der Wörter in der Umgebung des jeweiligen Schlüsselwortes errechnet.

4.4 Länge des Textes

Das Attribut beschreibt die Länge des Textes in Wörtern.

[1] Z. B. `http://www.amazon.de/Handys-Telefone/b/ref=sv_hv_1?ie=UTF8&node=3468301`

4.5 Satzzeichen

Weitere interessante Attribute stellen besondere Vorkommen von Satzzeichen wie z. B. '!' oder '. . . ' dar.

5 Stand der Dinge und weiterer Ausblick

5.1 Stand der Dinge

Wir haben bereits mithilfe des Amazon-Scrapers (s. 3.1) erfolgreich Reviews heruntergeladen. Jetzt soll die Tokenisierung folgen, die wir für die POS-Tags benötigen und es sollen die einfacheren Attribute extrahiert werden.

5.2 Probleme

Bisher haben wir noch nicht entschieden, wie wir mit informell geschriebenen Kommentaren umgehen (falsch geschriebene Wörter, Kleinschreibung, abweichende Zeichensetzung). Ein weiteres Problem ist, dass die Reviews sehr selten mit 2 oder 3 Sternen bewertet werden. Daher müssen wir noch überlegen, inwiefern wir dies gewichten und in unsere Reviewauswahl mit einfließen lassen.

5.3 Ausblick

Als nächstes geht es darum die heruntergeladenen Reviews für die Weiterverarbeitung mit WeKa zu formatieren, bzw. die oben genannten Features aus den Rohdaten zu extrahieren. Ein besonderes Augenmerk werden wir darauf legen, die Extraktion der verschiedenen Attribute modular zu gestalten, sodass `arff`-Dateien leicht mit beliebiger Attribut-Kombination erstellt werden können. So können wir später eine gute Evaluationen der Nützlichkeit der Attribute durchführen.

6 Einteilung

6.1 Caroline Berg

- Zwischenbericht
- Attribut-Extraktion

6.2 Simon Will

- Struktur der Programme
- Amazon-Scraper
- Keyword-Extraktion

Literatur

[1] Robert Remus, Uwe Quasthoff und Gerhard Heyer. “SentiWS – a Publicly Available German-language Resource for Sentiment Analysis”. In: *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*. 2010, S. 1168–1171.