MT @HanBaldwin: Fightin OOVs in German #twitter

Uladzimir Sidarenka, Tatjana Scheffler und Manfred Stede

Universität Potsdam

Postanschrift: Universität Potsdam
Karl-Liebknecht Str. 24-25, 2.32
14476 Potsdam

E-Mail: {uladzimir.sidarenka|tatjana.scheffler|manfred.stede}@uni-potsdam.de

Mit 58 Millionen versendeten Kurznachrichten pro Tag stellt Twitter eine unschätzbare Ressource an Textinformationen dar, deren qualitätvolle automatische Verarbeitung aber erst durch eine vorausgehende Normalisierung möglich wird. Und obwohl es schon etliche Verfahren für Vorverarbeitung von englischen "noisy" Texten gibt (vgl. Aw 2006, Kaufman 2010, Clark 2011, Han 2011), möchten wir in diesem Artikel ein neues Normalisierungsverfahren vorstellen, das explizit auf deutsche Tweets ausgerichtet ist und den störenden Effekt von Faktoren der internetbasierten Kommunikation (IBK) in diesem Textgenre verringert.

Um herauszufinden, welche Faktoren es eigentlich sind, analysierten wir 10,000 Tweets mit `TreeTagger` (Schmid, 1994) und `GNU Hunspell` und berechneten dabei, wie viele Tokens des Eingabetexts von diesen Programmen als "out-of-vocabulary" (OOV) erkannt wurden. Durch eine Auszählung und detaillierte manuelle Klassifizierung der OOV-Wörter wurde festgestellt, dass OOV-Tokens im Durchschnitt mehr als 20 % aller Wörter im Twitter-Diskurs ausmachen und von diesen mehr als die Hälfte auf Twitter-spezifisches Markup, IBK-Phänomene oder auch Rechtschreibfehler zurückzuführen ist.

Um die Prozentrate der OOV-Tokens zu senken, wendeten wir unser Set von Heuristiken an. Im ersten Verarbeitungsschritt dieses Sets wurden Hashtag-Zeichen und Retweet-Angaben entfernt sowie Twitter-Erwähnungen durch ein spezielles Wort ersetzt, das zum Lexikon der jeweiligen Tools auch hinzugefügt wurde. In nachfolgenden Etappen wurden dann weitere häufig beobachtete OOV-Phänomene in Twitter adressiert vor allem durch: a) Wiederherstellung von Umlauten (aus "ae", "ue" usw.), b) Rückführung von vervielfachten Buchstaben (z.B. "Haaallllooo") auf die Grundform, c) Normalisierung von Slangausdrücken und Kontraktionen ("isser") und d) spezielle Satzgrenzenerkennung und Tokenisierung.

Durch die Anwendung dieser Verfahren wurde eine Reduzierung der durchschnittlichen OOV-Rate um 6 bis 9 % erreicht. Eine nachfolgende Analyse der verbleibenden OOV-Tokens zeigte auch, dass der Rest der OOVs eher von der Unvollständigkeit der maschinenlesbaren Lexika zeugt oder durch Produktivität der Sprache selbst zu erklären ist, was aber auch in anderen Textgenres häufig vorkommt.

Literatur

Aw, A., Zhang, M., Xiao ,J., Su, J.: Lexical Normalization of Short Text Messages: Makn Sens a #twitter. A Phrase-based Statistical Model for SMS Text Normalization. In COLING/ACL Main Conference Poster Sessions, Seiten 33–40. (2011)

Han, B., Baldwin T.: Lexical Normalization of Short Text Messages: Makn Sens a #twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT). Seiten 368–378. (2011)

Clark E., Araki, K.: Text Normalization in Social Media: Progress, Problems and Applications for a Pre-processing System of Casual English. In Pacific Association for Computational Linguistics (PACLING) (2011)

Kaufman M.: Syntactic normalization of twitter messages. International Conference on Natural Language Processing (2010)

Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. International Conference on New Methods in Language Processing. (1994)