B. Mitschang et al. (Hrsg.): BTW 2017 – Workshopband,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 99

Die Gratwanderung zwischen qualitativ hochwertigen und einfach zu erstellenden domänenspezifischen Textanalysen

Cornelia Kiefer[1]

Abstract: Die Textanalyse ist zu einem entscheidenden Werkzeug in verschiedenen Domänen wie den Geisteswissenschaften, Naturwissenschaften sowie auch in der Industrie geworden. Eine der größten Herausforderungen bei domänenspezifischen Textanalyseprojekten besteht darin, das Wissen aus den Bereichen IT und Text Mining mit dem Wissen aus der Domäne zusammenzubringen. Viele Textanalysetoolkits werden deshalb speziell für den Gebrauch durch Domänenexperten ohne oder mit wenig IT und Textanalysewissen vereinfacht. In diesem Beitrag diskutieren wir, inwiefern diese Vereinfachungen zu Qualitätsproblemen bei der Analyse von unsauberen Daten führen können.

Keywords: Textanalyse, Datenqualität, Analysequalität, überwachte maschinelle Lernverfahren, Textanalyse in den Geisteswissenschaften.

1 Einleitung

Viele Fragen in den Geisteswissenschaften, Naturwissenschaften und in der Industrie können beantwortet werden, indem Informationen aus großen Textkorpora extrahiert werden [FS07]. So werden zum Beispiel in den Sozialwissenschaften, in der Biologie, Linguistik und in der Automobilindustrie Textanalysen verwendet, um Fragen aus den jeweiligen Bereichen zu beantworten. All diese domänenspezifischen Projekte müssen sich derselben Herausforderung stellen: das Expertenwissen aus IT, Textanalyse und der Domäne muss zusammengebracht werden, um die Fragestellungen adäquat beantworten zu können. Eine Möglichkeit, die Verschmelzung der Kompetenzen zu ermöglichen, besteht darin, Menschen in allen dreien oder zumindest zwei der Kompetenzen auszubilden [Co]. Ein weiterer Ansatz ist es, das notwendige IT und Textanalysewissen stark zu reduzieren und vereinfachte Analyseumgebungen zur Verfügung zu stellen, die es dem Domänenexperten erlauben IT und Textanalyse für die Beantwortung seiner Forschungsfrage zu nutzen. Zum Beispiel wurde der Leipzig Corpus Miner (LCM) für Sozialwissenschaftler ohne IT-Hintergrund entwickelt [LW16]. Eine einfache Verwendbarkeit von Analysetools sollte jedoch nicht auf Kosten der Qualität gehen. In diesem Artikel stellen wir zwei Datenqualitätsindikatoren mit Blick auf die automatische Analyse von Textdaten in Textanalysepipelines vor.

[1] Universität Stuttgart, Graduate School of Excellence advanced Manufacturing Engineering (GSaME), Nobelstr. 12, 70569 Stuttgart, cornelia.kiefer@gsame.uni-stuttgart.de

2 Motivation und Anwendungsszenario

Um Geisteswissenschaftlern zu ermöglichen, selbst Textanalysen durchzuführen, werden Textanalysen vereinfacht. In diesen vereinfachten Analysetools werden zum Teil voreingestellte Trainingskorpora verwendet. Diese Vereinfachung kann jedoch zu qualitativ schlechten Textanalysen und zu falschen Forschungsergebnissen führen, wenn unsaubere Daten für Fragestellungen in der Domäne analysiert werden sollen. Auch wenn unterschiedliche Texttypen wie etwa Bewertungen, Forumsnachrichten und Tweets aus den sozialen Medien in einem Stream in Echtzeit analysiert werden sollen, kann die Verwendung von Analysetools, die nur für einen Datentyp trainiert wurden, zu schlechten Ergebnissen führen. Diese Probleme können durch entsprechende Datenqualitätsindikatoren angezeigt werden.

Im Folgenden beschreiben wir ein beispielhaftes **Anwendungsszenario**, das diese Probleme aufzeigen soll: Ein Linguist möchte die Jugendsprache in sozialen Netzwerken analysieren. Hierzu werden unsaubere Chat,- und Twitterdaten betrachtet. Im ersten Schritt werden alle englischen Beiträge herausgefiltert. Auf den herausgefilterten Beiträgen werden nachfolgend einige Vorverarbeitungsschritte durchgeführt, wie etwa, Sätze zu segmentieren (=Satzsegmentierung), Sätze in seine kleinsten bedeutungstragenden Elemente zu zerlegen (=Tokenisierung), und jedem Token die korrekte Wortart zuzuordnen (=Wortartentagging). Danach kann nach Adjektiven oder Nomen gefiltert werden um schließlich eine Häufigkeitsverteilung zu berechnen und die Ergebnisse in den letzten Schritten zu visualisieren und manuell auszuwerten.

3 Verwandte Arbeiten

In [Ki16] legen wir den Fokus auf die Messung der Qualität von unstrukturierten Daten und schlagen automatische Metriken für Textdaten vor. In dieser Arbeit illustrieren wir zwei dieser Datenqualitätsindikatoren anhand eines Anwendungsszenarios aus den Geisteswissenschaften.

Während es zur Qualität von strukturierten Daten sehr viel Forschung gibt, etwa zu Dimensionen [WS96], sowie Methoden zum Messen und Verbessern der Qualität von strukturierten Daten [Se13], gibt es bisher noch kaum Forschung zur Qualität von unstrukturierten Daten. Die Notwendigkeit, Methoden für die Messung und Verbesserung der Qualität von unstrukturierten Daten zu finden, wurde jedoch erkannt [Sc12]. In [So04] werden vier Kategorien für Datenqualitätsdimensionen für unstrukturierte Textdaten und konkretere Indikatoren, wie z.B. lexikalische Ambiguität und der Anteil an Rechtschreibfehlern aufgelistet. Wir illustrieren hingegen zwei Datenqualitätsprobleme in einem Anwendungsszenario aus den Geisteswissenschaften und zeigen Probleme bei der Analyse von unsauberen Daten an Beispielen mit echten Textkorpora auf.

4 Messen der Interpretierbarkeit von Textdaten

In dieser Arbeit fokussieren wir die Datenqualitätsdimension Interpretierbarkeit. Diese messen wir über die Ähnlichkeit zwischen den vom Datenkonsumenten erwarteten Daten zu den Eingabedaten (vgl. [Ki16]). Die Eingabedaten sind dabei die zu analysierenden Daten und die erwarteten Daten können im Falle eines überwachten Klassifikators konkret über die verwendeten Trainingsdaten dargestellt werden. Die *Ähnlichkeit von Trainingsdaten und Eingabedaten* kann in diesem Fall mit Textähnlichkeitsmetriken berechnet werden, die die semantische, strukturelle oder stilistische Ähnlichkeit von Texten messen (für Textähnlichkeitsmetriken siehe etwa [DTI13]). Weiterhin gibt der *Anteil an noisy data* Aufschluss darüber, wie gut ein Standardtool für das Natural Language Processing, das zumeist saubere Daten erwartet, diese Daten interpretieren kann. Dieser Anteil kann z.B. über den Anteil an Rechtschreibfehlern, unbekannten Wörtern und Abkürzungen bestimmt werden.

5 Identifikation von Problemen der Interpretierbarkeit im Anwendungsszenario

In diesem Abschnitt zeigen wir Probleme auf, die mit Hinblick auf die oben beschriebenen Indikatoren bei der Erstellung der Analysepipeline zu dem Anwendungsszenario aus Kapitel 2 auftreten können. Die im Anwendungsszenario skizzierte Analysepipeline fassen wir in Abbildung 1 zusammen. Für jeden Schritt in der Analysepipeline muss die Ähnlichkeit zwischen den erwarteten Daten und den Eingabedaten gemessen werden, um Datenqualitätsprobleme mit Blick auf die Interpretierbarkeit der Textdaten feststellen zu können. Wir diskutieren im Folgenden beispielhaft die automatische Erkennung der Sprache („Sprache“ in Abbildung 1) und die automatische Annotation von Wortarten („Wortarten“ in Abbildung 1). Die Tools zu beiden Vorverarbeitungsschritten basieren auf überwachten maschinellen Klassifikatoren, die auf großen Mengen manuell annotierter Daten trainiert wurden.

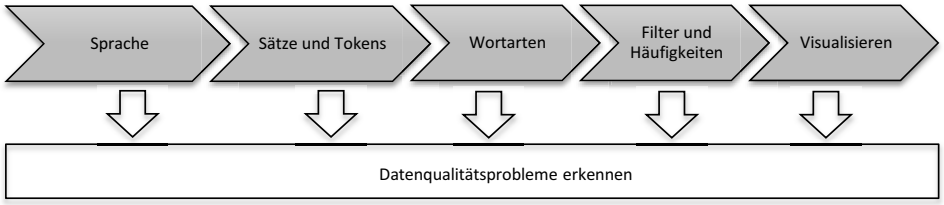

Abb. 1: Datenqualitätsprobleme können in jedem Analyseschritt in Textanalysepipelines auftreten

Im ersten betrachteten Vorverarbeitungsschritt, **„Sprache“**, soll für jeden Datensatz automatisch die korrekte Sprache erkannt werden. Da viele domänenspezifische Textanalysen unsaubere Daten verarbeiten müssen und die Domänenexperten die standardmäßig verwendeten Trainingsdaten zumeist nicht verändern, kann es zu einer

schlechten Erkennungsrate kommen. In Tabelle 1 stellen wir zur Illustration der Problematik die erreichte Genauigkeit (als prozentualen Anteil korrekt erkannter Sätze) für drei Spracherkenner, Apache *Tika*[2], den *language-detector*[3] und den *LanguageIdentifier*[4] für unterschiedliche Datentypen dar.

	Genauigkeit		
Datensatz	Tika	language-detector	Language Identifier
Zeitungstexte etc. (Penn Treebank[5], siehe [MMS93])	86	96	96
Prosa (Brown, siehe [FK79])	84	89	91
Tweets (Twitter Korpus, siehe [De13])	47	72	77
Chatnachrichten (NPS Chat, siehe [FLM])	20	22	33

Tab. 1: Genauigkeit der automatischen Filterung nach englischsprachigen Sätzen

Während Apache Tika ein Standardtool ist, das überwiegend auf sauberen Daten trainiert wurde, wurde der language-detector ebenfalls auf annotierten Tweets trainiert. Der LanguageIdentifier basiert auf [CT94], hier wurden Newsgroups als Trainingsdaten verwendet. Die Korpora sollten für dieses Beispiel jeweils Satzweise annotiert werden, wobei wir annehmen dass korrekterweise jeder Satz als englisch annotiert werden sollte[6]. Für saubere Daten, wie Zeitungstexte und Prosa liegt der Anteil an korrekt als Englisch erkannter Sätze bei allen drei Tools über 80%. Bei Tweets sinkt diese Rate erheblich für den Tika Spracherkenner, der saubere Daten erwartet (auf 47%) und weniger stark (auf 72% und 77%) für den language-detector und den LanguageIdentifier, die beide auch auf Tweets bzw. Newsgroups trainiert wurden. Für Chatnachrichten sinkt die Rate bei allen drei Erkennern erheblich, auf 20-33%. Bei den Chatnachrichten und Tweets sind insbesondere kurze Sätze mit einem hohen Grad an *noisy data* falsch klassifiziert worden. Zu den problematischen Sätzen zählen bspw. *„ah well“*, *„where did everyone gooo ?”, “RT @xochinadoll: I fel so blah today.”*. Die Messung des *Anteils an noisy data* sowie der *Ähnlichkeit zwischen Trainingsdaten und Eingabedaten* könnte diese Probleme indizieren.

Als nächsten beispielhaften Vorverarbeitungsschritt betrachten wir die automatische Annotation von **„Wortarten“**. In den Standardtools für die Wortartenannotation von Textdaten wird zumeist ein ausgewählter Trainingskorpus als Standard verwendet, der für die Analyse von sauberen Daten besonders geeignet ist. Wie schon bei der Komponente zur Spracherkennung kann eine Verwendung von Standardtools je nach Qualität der Daten zu Problemen führen. Zur Illustration der Problematik nennen wir in Tabelle 2 die erreichte Genauigkeit von verschiedenen Standardmodulen zur automatischen Bestimmung von Wortarten (als prozentualen Anteil der korrekt

[2] https://tika.apache.org/
[3] https://github.com/optimaize/language-detector
[4] Aus der Bibliothek DKPro Core: https://dkpro.github.io/dkpro-core/
[5] Verwendet wurde der in NLTK zur Verfügung gestellte Ausschnitt aus der Penn Treebank.
[6] Die Satzgrenzen wurden jeweils manuell bestimmt (bzw. es wurden entsprechende Goldannotationen für die Trennung der Korpora in Sätze verwendet).

zugewiesenen Wortarten von allen zugewiesenen Wortarten[7]). Wir haben drei Standardtools auf den unterschiedlichen Korpora getestet: Tool 1 ist der *NLTK* Maxent Treebank Tagger, Tool 2 ist der *Stanford* Tagger, und Tool 3, der *OpenNLP* Tagger. Im *Leipzig Corpus Miner* wird der OpenNLP Tagger für das Wortartentagging verwendet.

	Genauigkeit		
Datensatz	NLTK	Stanford	OpenNLP
Zeitungstexte etc. (Penn Treebank)	100	91	90
Prosa (Brown)	60	63	63
Tweets (Twitter Korpus)	65	67	70
Chatnachrichten (NPS Chat)	64	62	62

Tab. 2: Genauigkeit bei der automatischen Annotation von Wortarten

Sowohl die Goldannotationen als auch die vorhergesagten Wortarten basieren auf dem Penn Treebank Tagset [MMS93]. Tool 1 (NLTK) wurde auf der Penn Treebank trainiert, Tool 2 (Stanford) auf den Artikeln aus dem Wall Street Journal, die Teil der Penn Treebank sind und Tool 3 (OpenNLP) wurde auf eigenen Trainingsdaten von OpenNLP trainiert[8]. Bei sauberen Daten, wie etwa Zeitungstexten funktionieren die Standardtools einwandfrei, wohingegen bei unsauberen Daten wie Chatposts und Twitterdaten alle getesteten Tools versagen.

Wie in diesem Abschnitt anhand von zwei beispielhaften Vorverarbeitungsschritten aufgezeigt wurde, können bei jedem Analyseschritt in Textanalysepipelines Probleme auftauchen, die sich mit Vereinfachungen bzw. fehlenden Anpassungen durch Anwender ohne Textanalyseexpertise begründen lassen. Diese Problematik wird noch zusätzlich dadurch erschwert, dass die genannten Probleme in einer Analysepipeline propagieren und sich aufsummieren und letztlich die durch Textanalyse gewonnenen Antworten zur Forschungsfrage des Domänenexperten verfälschen können. Die illustrierten Probleme könnten durch eine Messung der Qualität der Textdaten, wie in Kapitel 4 vorgeschlagen, angezeigt werden.

6 Fazit und Ausblick

Um die Verschmelzung von Kompetenzen zu IT, Textanalyse und zu einer bestimmten Domäne zu ermöglichen, können vereinfachte Analyseumgebungen und Standardtools des Natural Language Processing eingesetzt werden. Die Vereinfachung von Textanalysetools sollte jedoch nicht auf Kosten der Qualität der Analysen gehen. In dieser Arbeit wurden hierzu zwei Datenqualitätsprobleme in einem konkreten Anwendungsszenario aus den Geisteswissenschaften illustriert. In der anschließenden Forschungsarbeit werden wir die entsprechenden Datenqualitätsindikatoren (*Ähnlichkeit*

[7] Es wurden jeweils die manuell annotierten Satz,- und Tokengrenzen verwendet.
[8] Zur Zusammenstellung letzterer Trainingsdaten gibt es keine näheren Angaben: https://opennlp.apache.org/documentation/manual/opennlp.html

von Trainingsdaten und Eingabedaten und *Anteil an noisy data*) und daraus abgeleitete Problemlösungen implementieren und den Ansatz in Experimenten validieren.

Literaturverzeichnis

[Co] Universität Jena: Computational and Data Science: Ein neuer Studiengang für eine neue Wissenschaftsdisziplin. http://www.cds.uni-jena.de/, 04.11.2016.

[CT94] Cavnar, W. B.; Trenkle, J. M.: N-gram-based text categorization. In Ann Arbor MI, 1994; S. 161–175.

[De13] Derczynski, L. et al.: Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data: Proceedings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics, 2013.

[DTI13] Daniel Bär; Torsten Zesch; Iryna Gurevych: DKPro Similarity: An Open Source Framework for Text Similarity: Proceedings of the Association for Computational Linguistics, Stroudsburg, USA, 2013.

[FK79] Francis, W. N.; Kučera, H.: Manual of Information to Accompany A Standard Corpus of Present-day Edited American English, for Use with Digital Computers. Brown University, Department of Lingustics, 1979.

[FLM] Forsyth, E.; Lin, J.; Martell, C.: The NPS Chat Corpus. http://faculty.nps.edu/cmartell/NPSChat.htm, 03.11.2016.

[FS07] Feldman, R.; Sanger, J.: The text mining handbook. Cambridge University Press, New York, 2007.

[Ki16] Kiefer, C.: Assessing the Quality of Unstructured Data: An Initial Overview. In (Ralf Krestel; Davide Mottin; Emmanuel Müller Hrsg.): CEUR Workshop Proceedings. Proceedings of the LWDA, Aachen, 2016; S. 62–73.

[LW16] Lemke, M.; Wiedemann, G.: Text Mining in den Sozialwissenschaften. Springer Fachmedien, Wiesbaden, 2016.

[MMS93] Marcus, M. P.; Marcinkiewicz, M. A.; Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank. In Comput. Linguist., 1993, 19; S. 313–330.

[Sc12] Schmidt, A.; Ireland, C.; Gonzales, E.; Del Pilar Angeles, M.; Burdescu, D. D.: On the Quality of Non-structured Data. http://www.iaria.org/conferences2012/filesDBKDA12/DBKDA_2012_PANEL.pdf, 03.11.2016.

[Se13] Sebastian-Coleman, L.: Measuring data quality for ongoing improvement. A data quality assessment framework. Elsevier Science, Burlington, 2013.

[So04] Sonntag, D.: Assessing the Quality of Natural Language Text Data: GI Jahrestagung, 2004; S. 259–263.

[WS96] Wang, R. Y.; Strong, D. M.: Beyond accuracy: what data quality means to data consumers. In J. Manage. Inf. Syst., 1996; S. 5–33.