MARSILIUS KOLLEG

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

WISSENSCHAFTLICHES SCHREIBEN UND SEINE WIRKUNG?

Eine computerlinguistische Analyse

Fellowbericht

Michael Strube

DOI: 10.11588/fmk.2021.0.78680

MARSILIUS-KOLLEG

2019/2020

MICHAEL STRUBE

WISSENSCHAFTLICHES SCHREIBEN UND SEINE WIRKUNG?

Eine computerlinguistische Analyse

Hängt die wissenschaftliche Wirkung einer Veröffentlichung nur vom wissenschaftlichen Inhalt ab, oder hat die Qualität des wissenschaftlichen Schreibens einen maßgeblichen Einfluss? Dieser Frage gehen Vera Nünning (Anglistik), Frauke Gräter (Biophysik) und ich nach. Meine Perspektive ist dabei die der Computerlinguistik: Ich versuche die Qualität des wissenschaftlichen Schreibens automatisch zu bestimmen und mit der wissenschaftlichen Wirkung statistisch zu korrelieren.

PLOS BIOLOGY-KORPUS

Als Gegenstand der Untersuchung wählten wir die Zeitschrift *PLOS Biology*[1] aus, eine Zeitschrift mit hoher wissenschaftlicher Reputation aus der *Public Library of Science*. Die Zeitschrift ist online verfügbar und *Open Access*. Sie wird unter der *Creative Commons „Attribution" Licence (CC BY)* veröffentlicht. Das heißt, man kann den Inhalt herunterladen, weiterverarbeiten und weitergeben, wenn man den Urheber des Inhalts nennt. *PLOS Biology* wird seit Oktober 2003 veröffentlicht und enthält jetzt etwa 3000 Forschungsartikel (und zusätzlich Editorials, Essays, Kurzberichte, etc.). Alle Artikel können von der *PLOS Biology* Webseite nicht nur im HTML- und PDF-Format heruntergeladen werden, sondern auch im XML-Format, was die automatische Weiterverarbeitung außerordentlich erleichtert. Des Weiteren ist jeder Artikel mit aktuellen Download- und Zitationsmetriken versehen. Diese Metriken betrachten wir als Näherung der wissenschaftlichen Wirkung eines Artikels. Frauke Gräter als Domänenspezialistin überprüft, ob die Anzahl der Zitationen mit ihrer Einschätzung der wissenschaftlichen Qualität eines Artikels übereinstimmt.

Um den Datensatz (Korpus) zu erstellen, schreibe ich im ersten Schritt einen *Webcrawler,* der die Webseite von *PLOS Biology* systematisch Jahrgang für Jahrgang und Monat für Monat durchgeht. Der *Webcrawler* lädt die Artikel im XML-Format herunter, streicht Nicht-Textinhalte (Referenzen, Formeln, Abbildungen, Tabellen, etc.) und speichert die Texte in einem vereinfachten XML-Format, das die Information über die Kapitelstruktur erhält. Die DOI, Autorennamen, Affiliationen, der Titel des Artikels, von *PLOS Biology* vergebene Stichwörter und Zitationen speichere ich in einer Datenbank. Das XML-Format einiger weniger Artikel ist nicht regelkonform. Diese Artikel werden nicht heruntergeladen, so dass am Ende 2764 Artikel bleiben. Wird ein Text unmittelbar nach der Publikation zitiert, kann dies von anderen Faktoren als der wissenschaftlichen Qualität abhängen (etwa Öffentlichkeitsarbeit der Institution). Deshalb berücksichtige ich nur Artikel, die bis spätestens Ende 2015 publiziert wurden. Dies entspricht 2105 Artikeln, die ich in 1265 für das Training, 420 für die Entwicklung und 421 für das Testen einteile. Alle Ergebnisse berichte ich auf dem Entwicklungsdatensatz. Der Testdatensatz bleibt vorerst unberührt.

PLOS Biology berichtet die Gesamtzahl der Zitationen. Da alte Artikel mehr Zeit hatten, zitiert zu werden, als neue, implementiert Frauke Gräters Doktorand Florian Franz eine Funktion zur Normalisierung der Zitationen über die Zeit. Damit können alte und neue Publikationen gleichermaßen betrachtet werden.

METHODEN

In der Computerlinguistik gibt es eine lange Tradition, die Lesbarkeit von Texten automatisch zu bewerten. Der *Flesch-Kincaid*-Lesbarkeitstest[2] etwa beruht auf der Annahme, dass kurze Sätze und kurze Wörter besser lesbar sind, und kombiniert durchschnittliche Satz- und Wortlänge in einer Formel:

Flesch Reading Ease:
206.835 - 1.015(total words/total sentences) - 84.6*(total syllables/total words)*

Davon ausgehend testen wir folgende Merkmale als *Baselines* (Vergleichsmaßstab): Anzahl Sätze, Anzahl Wörter, Anzahl Buchstaben, Anzahl Silben, durchschnittliche Satzlänge in Wörtern, Buchstaben und Silben, durchschnittliche Wortlänge in Buchstaben und Silben und *Flesch Reading Ease.*

Aus den Diskussionen in der Fellow-Arbeitsgruppe mit Vera Nünning (qualitative Analyse) ergibt sich die Notwendigkeit, linguistisch anspruchsvollere Merkmale zu untersuchen, zu implementieren und zu analysieren. Die Empfehlung, in wissenschaftlichen Texten das Passiv zu vermeiden, implementiere ich durch eine Heuristik, die festzustellen versucht, ob ein Satz im aktiven oder passiven Modus verfasst ist. Das Verhältnis von Aktiv zu Passiv halte ich in der *Active/Passive Ratio* fest. Die Größe des Wortschatzes hat ebenfalls Einfluss auf die Lesbarkeit und kann einfach mit Hilfe der *Type/Token Ratio* gemessen werden. Wenn die *Type/Token Ratio* niedrig ist, ist der Text redundant und enthält wenig Information. Ist sie hoch, enthält der Text viel Information, ist aber möglicherweise nicht gut lesbar.

Bestimmte Inhalts- und Funktionswörter können Einfluss auf die Lesbarkeit haben. In der Computerlinguistik berechnet man *n-Grams*[3] – Abfolgen von *n* Wörtern, wobei $n \geq 1$. Dokumente können dann durch die Anzahl bestimmter *n-Grams* charakterisiert werden. Ich tokenisiere die Texte, d.h., segmentiere in Einzelwörter, reduziere auf den Wortstamm und zähle. Das Ergebnis wird in einer weitgehend leeren Matrix *(Sparse Matrix)* festgehalten, die das Gesamtvokabular eines Korpus umfasst. Um von Einzelwörtern zu abstrahieren, führe ich die gleiche Zählung auf *Part-of-Speech n-Grams* durch. *Part-of-Speech* ist die Wortart eines Wortes im

Kontext, etwa Artikel, Nomen oder Verb. Dafür tokenisiere ich das Dokument, bestimme automatisch die Wortart eines jeden *Tokens (Part-of-Speech Tagging)* und zähle. Während n-Grams den Inhalt eines Textes kategorisieren (mit Ausnahme einiger Funktionswörter und spezieller Ausdrücke, die auf Stil hinweisen), charakterisieren *Part-of-Speech n-Grams* den Stil eines Textes.

Neben diesen einfachen, zählbasierten Methoden stellt die Computerlinguistik Verfahren zur Verfügung, die den emotionalen Gehalt von Sprache analysieren. Stellen die Autoren das Ergebnis ihrer Arbeit positiv oder negativ dar oder verwenden sie weitgehend neutrale Sprache? Zur Beantwortung dieser Frage verwende ich die Kategorien *Valence, Arousal* und *Dominance,* die ich mit Hilfe des von Saif Mohammad entwickelten Lexikons bestimme.[4] *Valence* beschreibt die Dimension positive-negative Emotion, *Arousal* die Dimension erregt-ruhig, *Dominance* die Dimension Macht-Schwäche. Um Wörter mit diesen Kategorien zu versehen, tokenisiere ich die Texte, reduziere die Wörter auf ihre Grundform (Lemmatisierung), reduziere die Einträge im Lexikon auf ihre Grundform und bilde die Kategorien dann auf die Lemmata ab. Schließlich berechne ich die durchschnittlichen Werte für *Valence, Arousal* und *Dominance* für jeden Text. Ein verwandtes Verfahren ist die *Sentiment*-Analyse. Ich verwende die Ressource *SentiWordNet*,[5] die für jeden *WordNet Synset*[6] einen der drei

Werte positiv, negativ oder objektiv vergibt. Ich tokenisiere die Texte, disambiguiere die Wortbedeutungen und bilde die *SentiWordNet*-Werte auf die Wörter ab. Schließlich berechne ich die durchschnittliche Positivität und Negativität eines Textes.

EXPERIMENTE

Die aktuelle Forschung in der Computerlinguistik verwendet in der Regel neuronale Netzwerke, die ohne linguistische Vorverarbeitung Prädiktionen auf Texten vornehmen und häufig exzellente Ergebnisse aufweisen. Ich dagegen bin daran interessiert, welche textuellen Eigenschaften für wissenschaftliche Wirkung ausschlaggebend sind. Deshalb verwende ich die oben beschriebenen Merkmale, um ein einfaches maschinelles Lernverfahren – lineare Regression – zu trainieren. Das erwünschte Ziel ist ein Wert für die statistische Korrelation zwischen den Merkmalen und der Zitationshäufigkeit als Näherung für wissenschaftliche Wirkung. Lineare Regression bestimmt den Zusammenhang zwischen einer abhängigen Variable (Zitationshäufigkeit) und einer unabhängigen Variable (den oben beschriebenen Merkmalen). Mich interessiert, welche Merkmale die größte Korrelation mit der Zitationshäufigkeit haben. Dafür wende ich einen Merkmalsauswahlalgorithmus an:

1. selected features = []
2. add each single feature separately to selected features -- run regression
3. determine feature with best performance when added to selected features
4. move this feature from features to selected features
5. if performance greater than previous highest performance, set this as highest performance, remember selected features
6. if less than 50 features selected, go to 2, else go to 7
7. report selected features at highest performance

ERGEBNISSE UND DISKUSSION

Die ausgewählten Merkmale und die Zitationshäufigkeit haben in der Tat eine statistisch signifikante, moderat positive Korrelation von rho=0.55 *(Spearman's rho).* Die ausgewählten Merkmale entstammen vielen der oben beschriebenen Merkmalsgruppen, wobei *Part-of-Speech n-Grams* und *n-Grams* die höchste Korrelation mit der Zielvariablen aufweisen, nahe gefolgt von *SentiWordNet.* Das zeigt, dass Merkmale, die Emotionen und Bewertungen ausdrücken, einen Einfluss auf die

Zitationshäufigkeit haben. Stark negativ korreliert mit der Zielvariable ist das *n-Gram "the drosophila".* Dieser Standardorganismus der Mikrobiologen kann keine große Begeisterung mehr wecken. Positiv korreliert ist dagegen das *n-Gram "microbiome communities",* ein Forschungsgebiet, das in der näheren Vergangenheit in Mode kam, und deshalb häufig zitiert wird. Neben Inhaltswörtern spielen aber auch Funktionswörter und -konstruktionen eine wichtige Rolle. *"Data from"* zeigt, dass datengetriebene Forschung wichtig ist. *"Proper", "rich and", "abundant", "normal", "complex with",* und *"that many"* werden eher in der Interpretation eingesetzt und drücken häufig eine positive Einstellung zu den Ergebnissen aus. Dies bestätigt die Auswahl der Merkmale, die auf *SentiWordNet* beruhen, da diese Merkmale ebenfalls eine positive oder negative Einstellung erfassen. Die Merkmale *Valence* und *Arousal* werden ebenfalls ausgewählt, sind aber verhältnismäßig schwach. Die *Baseline*-Merkmale scheinen dagegen keinen Einfluss auf die Zitationshäufigkeit zu haben. *Active/Passive-* und *Type/Token-Ratio* sind ebenfalls nicht wichtig.

Als vorläufiges Ergebnis bleibt festzuhalten, dass traditionelle Merkmale, die die Qualität des Schreibens quantifizieren, keine oder nur eine sehr schwache Korrelation mit der Zitationshäufigkeit eines wissenschaftlichen Artikels haben. Sprachliche Merkmale, die eine positive Einstellung zu den wissenschaftlichen Ergebnissen ausdrücken, weisen dagegen eine moderat positive Korrelation auf. Dies kann entweder damit zusammenhängen, dass die Ergebnisse der wissenschaftlichen Arbeit in der Tat positiv sind, oder aber, dass der Autor die Ergebnisse in ein positives Licht zu rücken versucht.

1 Vgl. https://journals.plos.org/plosbiology/

2 Vgl. Rudolph Flesch: *A New Readability Yardstick,* in: *Journal of Applied Psychology 32* (1948), S. 221-233, https://doi.org/10.1037/h0057532; J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers und Brad S. Chissom: *Derivation of New Readability Formulas for Navy Enlisted Personnel,* in: *NTC, Naval Air Station Memphis-Millington, Tenn: Technical Report* 8-75 (1975); Sarah E. Schwarm und Mari Ostendorf: *Reading Level Assessment Using Support Vector Machines, in: Proceedings of the 43rd Annual Meeting of the ACL* (2005), S. 523-530, https://doi.org/10.3115/1219840.1219905.

3 N-Gramme sind selbstbestimmte Einheiten, in die ein Text zerlegt werden kann. Geht man beispielsweise nach Wörtern vor, würde ein n-Gramm eine Einheit aus n Wörtern bilden.

4 Vgl. Saif M. Mohammad: *Obtaining reliable human ratings for valence, arousal, and dominance for 20,000 English words,* in: *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics 1 (2018), S. 174-184, https://doi.org/10.18653/v1/P18-1017; http://saifmohammad.com/WebPages/nrc-vad.html

[5] Vgl. https://github.com/aesuli/SentiWordNet

[6] "WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations." direkt zitiert nach und vgl. https://wordnet.princeton.edu/