SuW GRADUIERTENPLATTFORM
Forschungsnetzwerk Sprache und Wissen

DIGITALE TAGUNG
01. BIS 02. OKTOBER 2020

Sprache und Wissen hin und zurück – iterative Annotation als linguistische Forschungsmethode

Thomas Jurczyk (Ruhr-Universität Bochum)

Projektvorstellung „ReligionML – Annotation religiöser Texte für Machine Learning“

Das in meinem Vortrag vorzustellende Projekt *ReligionML* wird derzeit als interne Arbeitsgruppe am Centrum für Religionswissenschaftliche Studien (CERES, Ruhr-Universität Bochum) durchgeführt. Die Arbeitsgruppe bestehend aus Religionswissenschaftler:innen verschiedener Schwerpunktbereiche hat es sich zum Ziel gesetzt, religionswissenschaftlich relevante Texte gemeinsam zu annotieren, um so mit der Zeit ein verlässlich annotiertes religionswissenschaftliches Corpus zu schaffen, das sowohl für die automatisierte als auch manuelle Bearbeitung religionswissenschaftlicher Fragen herangezogen werden kann.

Obwohl das finale Corpus für unterschiedliche Forschungsfragen nutzbar sein soll, steht im theoretischen Zentrum0F[1] der Gruppe derzeit die Frage, wie religiöse Semantik1F[2] in unterschiedlichen gesellschaftlichen Kontexten (Politik, Kunst, Wirtschaft, Religion etc.) verwendet wird. Das Corpus soll es dabei ermöglichen, diese Frage nicht nur punktuell, sondern möglichst umfangreich und repräsentativ bearbeiten zu können. Außerdem sollen Machine Learning Modelle, die auf Basis der annotierten Daten des Corpus trainiert wurden, dabei helfen, unbekannte Daten vorzufiltern und beispielsweise einzuordnen, ob es sich bei einem Text, der religiöse Semantik beinhaltet, um religiöse oder nicht-religiöse Kommunikation handelt2F[3] und aus welchem gesellschaftlichen Bereich diese stammt. Die Erstellung solcher automatisierter Klassifizierungsmodelle würde nicht nur die Filterung großer Datenmengen ermöglichen, um spezifischere Fragen zu bearbeiten,3F[4] sondern auch Rückschlüsse auf die Besonderheiten der religiösen bzw. nicht-religiösen Verwendung religiöser Semantik ermöglichen, die in den Klassifizierern zugrunde liegenden Entscheidungsparametern erkennbar sind.

Das Projekt *ReligionML* befindet sich noch in der Anfangsphase. Es basiert technisch auf einer von mir erstellten Webapplikation und konzentriert sich inhaltlich derzeit auf die Annotation von englischen Tweets, die Wörter wie „holy“ oder „religion“ enthalten, wobei das Corpus stetig erweitert werden soll und bereits wird. Es wurden bisher zwei Annotationsschritte implementiert:

[1] Frei nach dem ersten Schritt des MATTER Modells in (Pustejovsky, Bunt, and Zaenen 2017).
[2] Beispielsweise Heiligkeitssemantiken.
[3] Erste Tests mit simplen Machine Learning Modellen wie KNN und Logistic Regression wurden dabei bereits durchgeführt.
[4] Zum Beispiel, wenn die Frage im Zentrum steht, wie religiöse Semantiken in politischer Kommunikation verwendet werden.

Zum einen werden die Tweets als Ganzes von den Annotatoren:innen klassifiziert bzw. annotiert.4F[5] Zum anderen haben die Annotatoren:innen die Möglichkeit, einzelne Wörter aus den Tweets separat zu annotieren. Wir arbeiten dabei bewusst nicht mit einem Goldstandard, sondern die Annotationskategorien werden während regelmäßiger Treffen weiterentwickelt. Anfangs sind wir dabei von einer binären religiös/nicht-religiös Klassifizierung ausgegangen, haben dann allerdings schnell gemerkt, dass dies nicht ausreichend ist, und unsere Annotationen immer weiter differenziert bzw. ausgeweitet.

Während meines Vortrages möchte ich den bisherigen Stand unserer Diskussion sowie insbesondere unser iteratives Vorgehen vorstellen und diskutieren. In diesem Zusammenhang möchte ich als besonderes Merkmal unseres Vorhabens hervorheben, dass wir bewusst mit der Ambiguität der Texte bzw. der Annotationen umzugehen versuchen. So sehen wir beispielsweise divergierende Kategorisierungen durch die einzelnen Annotatoren:innen nicht als Problem an, sondern vielmehr als Teil der Phänomenbeschreibung, die es uns erlaubt, Einordnungswahrscheinlichkeiten von Texten prozentual wiederzugeben, anstatt eine eindeutige Zuordnung vorzunehmen, die so oftmals auch in den Texten schlicht nicht gegeben ist.

Besonders interessiert bin ich darüber hinaus an existierenden Annotationsschemata aus dem Bereich der Annotation religiöser Texte, die mir bisher nicht bekannt sind, sowie an einem allgemeinen Austausch und der Knüpfung von Kontakten, da wir uns noch in einer sehr frühen Projektphase befinden.

Literatur

Pustejovsky, James, Harry Bunt, and Annie Zaenen. 2017. “Designing Annotation Schemes: From Theory to Model.” In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 73–113. Dordrecht: Springer.

[5] Dabei sieht das derzeitige Annotationsschema der Tweets auf der Makroebene wie folgt aus: **a) Inner-religion** means that the text has a religious meaning and was stated from within religion. A typical example would be a Christian or Muslim saying something about his or her belief. **b) Religion-transcendence** means that although the text does not belong to an inner-religious sphere, the overall context is still referring to religion as a social system dealing with the immanence/transcendence distinction. A typical example is two non-religious persons talking about religion as a worldview and religious truth claims compared to, for instance, philosophical ones. **c) Religion-immanence** means that the text still uses religious semantics but more in the sense of a general distinction marker. A good example is "religion" as an ethnic/political category. **d) Metaphorical-use** means that the terminology is no longer directly referring to "religion" but uses the "religion" domain in another context. A typical example are sentences such as "Electronic music is my religion." **e) No-religion** means that the text has nothing to do with religion at all.