Computerlinguistische Verfahren zur Aufdeckung struktureller Ähnlichkeiten in Narrativen

Einführung

In diesem Beitrag stellen wir eine Methode zur automatischen Erkennung von strukturellen Ähnlichkeiten narrativer Texte auf der Handlungsebene vor. Dafür operationalisieren wir strukturelle Ähnlichkeiten als (intertextuelle) Verbindungen (*Alignments*) zwischen Ereignissen. Die verwendeten Alignierungsalgorithmen bauen auf automatisch erzeugten linguistischen Analysen der Texte auf und verwenden als Kriterien Eigenschaften verschiedener linguistischer Ebenen. Ziel unseres Ansatzes ist es, materiell in Texten vorliegende Ähnlichkeiten auffindbar zu machen und hervorzuheben, so dass sie von Wissenschaftlerinnen und Wissenschaftlern zielgerichtet analysiert und interpretiert werden können.

Anwendungsszenarien

Die Untersuchung struktureller Ähnlichkeiten zwischen Narrativen spielt in vielen geisteswissenschaftlichen Disziplinen eine Rolle. Als Beispielszenarien verwenden wir die Märchen- und Ritualforschung.

Ähnlichkeiten zwischen **Märchen** sind auf verschiedenen Granularitätsebenen untersucht worden. Propp (1958) veröffentlichte eine Analyse, in der in russischen Märchen prototypische Handlungen und Charaktere identifiziert werden. Regelmäßigkeiten im Auftreten von Handlungen und Charakteren werden in einer sog. „Morphology of the Folktale“ erfasst. Damit sollen typische Handlungsmuster (Ereignis X folgt auf Ereignis Y) beschrieben werden. Am anderen Ende der Granularitätsskala existieren Sammlungen wie der ATU-Index (Uther, 2014), in dem Märchen mit gleichen Handlungselementen (Aussetzen von Kindern) oder Charakteren (Lebkuchenhaus) in Klassen zusammengefasst werden.

Im Bereich der **Ritualforschung** werden Rituale aus diversen religiösen, kulturellen oder politischen Kontexten untersucht. Unter dem Stichwort „Ritualgrammatik“ (vgl. Hellwig und Michaels, 2013) wird diskutiert, dass in verschiedenen Ritualen ähnliche Handlungen vorkommen und Teilnehmer ähnliche Rollen übernehmen. Verschiedene Forscher vertreten die Auffassung, dass die Zusammensetzung wiederkehrender Ereignisse zu Ritualen Regeln folgt. Existierende Überlegungen zur Ritualgrammatik sind nicht formalisiert und daher für eine automatische Analyse nur begrenzt nutzbar.

Um unsere Methode entwickeln und testen zu können, haben wir für diese beiden Szenarien ein englischsprachiges Korpus zusammengestellt, das mehrere Beschreibungen des gleichen Typs enthält (ATU-Märchenklasse bzw. Ritualtyp).

Computerlinguistische Verarbeitung

Wir wenden die gleichen computerlinguistischen Komponenten auf beide Korpora an. Damit werden linguistische Repräsentationen für Wortarten, (syntaktische) Dependenzrelationen, semantische Rollen, Wortbedeutungen und Koreferenzketten erstellt. Verknüpft ergeben diese Annotationen eine Diskursrepräsentation, die als Basis für die Alignierungsverfahren verwendet wird. Da Ritualbeschreibungen untypische linguistische Phänomene enthalten, wurden sämtliche Komponenten auf die Domäne angepasst (*Domain Adaptation*). Dadurch konnten deutliche Qualitätssteigerungen der computerlinguistischen Analyse erreicht werden.

Alignierungsexperimente

Drei Alignierungsalgorithmen mit unterschiedlicher Mächtigkeit wurden verglichen: *Sequence alignment* (Needleman-Wunsch, 1970) ist der einfachste Algorithmus, der ausschließlich paarweise und nicht-kreuzende Alignierungen erzeugen kann. *Graph-based predicate alignment* (GPA; Roth, 2014, Roth & Frank, 2012) kann paarweise und kreuzende Alignierungen erzeugen. *Bayesian model merging* (BMM; Stolcke & Omohundro, 1993) ist der mächtigste Algorithmus, der Alignierungen beliebiger Länge mit Überkreuzungen erzeugen kann. Diese drei Algorithmen wurden in zwei Experimenten evaluiert: In einer intrinsischen Evaluation wurden die Ergebnisse mit einem von zwei Ritualwissenschaftlern parallel erzeugten Goldstandard verglichen (κ=0.61). Dabei erzielte BMM die besten Ergebnisse insgesamt und GPA die besten Ergebnisse auf einem Einzeldokumentpaar.
Im zweiten Experiment wurde aus den automatisch erzeugten Alignierungen ein Maß für Dokumentenähnlichkeit berechnet und in einem Clustering-Verfahren eingesetzt. Das Ergebnis des Clusterings – eine Einteilung der Dokumente auf Basis der errechneten strukturellen Ähnlichkeit – konnte dann mit der Gruppierung verglichen werden, die „natürlicherweise“ in den Korpora vorkommt (Ritualtypen bzw. ATU-Klassen). Dabei zeigten sich wieder GPA und BMM als die leistungsstärksten Algorithmen.

Visualisierung und Nutzung

Um es Wissenschaftlerinnen und Wissenschaftlern aus der Ritual- bzw. Märchenforschung zu ermöglichen die Analysen zu nutzen, haben wir Visualisierungen entwickelt, die eine systematische Untersuchung der gefundenen Ähnlichkeiten ermöglichen. Auf einer Vogelperspektive stellen wir die Dokumentenähnlichkeit in einer Heatmap dar. Auf interessante, dicht verknüpfte Stellen können wir hinweisen, indem für jedes Ereignis ein *connectivity score* in einem Diagramm anzeigt wird. Eine detaillierte Darstellung der Einzelereignisse (mit Teilnehmern und Kontext-Ereignissen) ist ebenfalls möglich. Direkt aus der Diskursrepräsentation können wir außerdem eine Visualisierung des sozialen Netzwerks erzeugen, in der wichtige Entitäten (Charaktere, Gegenstände und Materialien) in einem Netzwerk angezeigt und gemeinsam auftretende Figuren verknüpft werden.

Konklusion

Der Posterbeitrag präsentiert eine Methode zur Erkennung struktureller Ähnlichkeiten zwischen narrativen Texten. Die Ähnlichkeiten werden basierend auf computerlinguistischen Analysen vollautomatisch identifiziert und können zielgerichtet auf unterschiedlichen Granularitätsebenen dargestellt und manuell inspiziert werden. Damit eignet sich die Methode auch zur Analyse von größeren Datenmengen, ohne bestimmte Interpretationen vorwegzunehmen. Eine ausführliche Darstellung des Verfahrens sowie des geisteswissenschaftlichen Anwendungskontexts findet sich in Reiter (2014) und Reiter et al. (2014).
Auf einer methodischen Ebene zeigt sich in diesem Projekt, dass komplexe linguistische Analysen auch für nicht-kanonische Textsorten erstellt werden können und eine vielversprechende Ausgangsbasis für Analysen darstellen. Die Besonderheiten natürlicher Sprache (z.B. Ambiguität, Vielseitigkeit) stellen für automatische Verarbeitung eine große Herausforderung dar, werden aber in der Computerlinguistik bereits untersucht. Auf (computer-)linguistische Analysen aufzubauen erlaubt die Untersuchung komplexer semantischer Phänomene, die vergleichsweise eng mit den Zielkategorien vieler Geisteswissenschaften verwandt sind.

Bibliographie

Oliver Hellwig and Axel Michaels. Ritualgrammatik. In Christiane Brosius, Axel Michaels, and Paula Schrode, Hrsg., *Ritual und Ritualdynamik*, S. 144–150. Vandenhoeck & Ruprecht, Göttingen, Germany, 2013.

Saul B. Needleman and Christian D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3):443–453, March 1970.

Vladimir Yakovlevich Propp. *Morphology of the Folktale*. University of Texas Press, Austin, TX, 2nd edition, 1958. Translated by Laurence Scott (Original work published 1928).

Nils Reiter. *Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms*. PhD thesis, Heidelberg University, June 2014.

Nils Reiter, Anette Frank, and Oliver Hellwig. An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605, 2014.

Michael Roth. *Inducing Implicit Arguments via Cross-document Alignment – A Framework and its Applications*. PhD thesis, Heidelberg University, 2014.

Michael Roth and Anette Frank. Aligning predicates across monolingual comparable texts using graph-based clustering. In Jun'ichi Tsujii, James Henderson, and Marius

Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 171–182, Jeju Island, Korea, July 2012.

Andreas Stolcke and Stephen Omohundro. Hidden markov model induction by bayesian model merging. In Steve J. Hanson, J. D. Jack D. Cowan, and C. Lee Giles, Hrsg., *Advances in Neural Information Processing Systems*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, California, 1993.

Hans-Jörg Uther. *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. Number 284–286 in FF Communications. Suomalainen Tiedeakatemia, Helsinki, 2004.