Semantische Extraktion auf antiken Schriften am Beispiel von Keilschriftsprachen mithilfe semantischer Wörterbücher

Homburg, Timo
timo.homburg@gmx.de
Hochschule Mainz, Deutschland

Einleitung und Motivation

Semantische Extraktionsmechanismen (z.B. Topic Modelling) werden seit vielen Jahren im Bereich des Semantic Web und Natural Language Processings sowie in den Digital Humanities als Verfahren zur Visualisierung und automatischen Kategorisierung von Dokumenten eingesetzt. Oft ergeben sich durch den Einsatz neue Aspekte der Interpretation von Dokumentensammlungen die vorher noch nicht ersichtlich waren. Als Beispiele solcher Verfahren kommen häufig Machine Learning Algorithmen zum Einsatz, welche eine Grobeinordnung von Texten vornehmen können. Gepaart mit Metadaten von Texten können anschließend beispielsweise thematische Übersichten von Dokumenten mit geographischem Bezug auf Kartenmaterialien in GIS Systemen oder mittels historischer Gazetteers zeitliche Zusammenhänge automatisiert dargestellt werden. In dieser Publikation möchten wir die Möglichkeiten der semantischen Extraktion nutzen und diese auf einer Sammlung von Texten in Keilschriftsprachen anwenden.

Keilschriftsprachen

Keilschriftsprachen haben in den letzten Jahren ein größeres Interesse in der Digital Humanities und Linguistik Community erfahren. (Inglese 2015, Homburg et. al. 2016, Homburg 2017, Sukhareva et. al. 2017). Neben der andauernden Standardisierung in Unicode werden unter anderem Part Of Speech Tagger und Mechanismen der automatisierten Übersetzung erprobt um Keilschrifttexte besser mit dem Computer zu erfassen und zu interpretieren. Desweiteren wurde die Erlernbarkeit der Keilschriftsprachen durch digitale Tools wie Eingabemethoden oder Karteikartenlernprogramme verbessert. (Homburg 2015) Trotz all der erreichten Fortschritte verbleiben jedoch zahlreiche Probleme bei der maschinellen Verarbeitung von Keilschriftsprachen, die unter anderem mit der geringen Verfügbarkeit annotierter Ressourcen und der fehlenden Verfügbarkeit maschinenlesbarer und semantisch sowie linguistisch annotierter Wörterbücher zusammenhängt. Diese Limitierungen hindern viele Natural Language Processing und semantische Extraktionsalgorithmen daran ein besseres Ergebnis zu erzielen. Wir möchten mit dieser Publikation einen Beitrag leisten diese Situation zu verbessern und stellen das "Semantic Dictionary for Ancient Languages" vor, welches ein Versuch ist durch Annotierung vorhandener in der Forschungscommunity anerkannter Wörterbuchressourcen mit Unicode Characters, Semantic Web Konzepten, etymologischen Daten, gemeinsamen Vokabularen und POSTags eine semantische Ressource in RDF für die Optimierung solcher Algorithmen auf Basis der Sprachen Hethitisch, Sumerisch und Akkadisch zu schaffen.Das Wörterbuch basiert auf dem Lemon-Standard, ein W3C Standard der es erlaubt ebenfalls multilinguale Resourcen abzubilden. So können Entwicklungen der Sprache und gemeinsame Vokabulare wie zum Beispiel Akkadogramme und Sumerogramme in Hethitisch mit erfasst werden.

Semantisches Wörterbuch und Semantische Extraktion

Wir testen die Performance des Wörterbuchs auf einer der größten Sammlungen von digitalen Keilschrifttexten, der CDLI, aus der wir repräsentative Texte in hethitischer, sumerischer und akkadischer Keilschrift aus verschiedenen Epochen extrahieren und mittels Machine Learning klassifizieren, sowie verschlagworten. Das Ergebnis der semantischen Extraktion ist eine Sammlung von Themen pro Keilschrifttafel, die sich wiederum in Überkategorien gruppieren lassen und in einen zeitlichen, sprachlichen, dialektischen, sowie örtlichen Kontext gestellt werden können. Anhand der verschiedenen Metadaten der CDLI war es uns möglich eine thematische Karte der Fundorte der Keilschrifttafeln sowie deren Inhalt pro Epoche darzustellen aus der das relevante Fachpublikum schließen kann welche Themen zu welcher Zeit an welchem Fundort relevant für die Schreiber der jeweiligen Epoche waren. Im Zuge einer Weiterentwicklung möchten wir diese Informationen mit weiteren Metadaten wie beispielsweise der Jurisdiktion, den Daten der jeweiligen Herrscher sowie rekonstruierten Orten aus der antiken Zeit vervollständigen um Rückschlüsse auf interessante historische Ereignisse zu ziehen.

Aufbau des Posters

Auf unserem Poster möchten wir gerne den Prozess des Aufbaus, sowie die Struktur des semantischen Wörterbuchs sowie die Karte die durch unsere semantische Extraktion entstanden ist präsentieren um die jeweiligen Fachwissenschaftler zur Diskussion über die Entwicklung eines Semantic Web von Keilschriftsprachen und

Keilschriftartefakten einzuladen. Desweiteren soll unser Poster eine Reihe von Anwendungen demonstrieren die sich in Zukunft mit unserer semantischen Ressource entwickeln lassen können um einen Beitrag zu einem hoffentlich zukünftig existierenden LinkedData Datensatz für Keilschriftartefakte zur Dokumentation von Keilschrift zu leisten.

Bibliographie

Inglese, G. (2015): “Towards a hittite treebank. basic challenges and methodological remarks.” In: Corpus-Based Research in the Humanities (CRH) p. 59 1.1

Homburg, T. (2017): “Postagging and semantic dictionary creation for Hittite cuneiform.” In:DH2017 (2017)

Homburg, T., Chiarcos, C. (2016): “Word segmentation for Akkadian cuneiform.” In: LREC2016 1.1

Homburg, T., Chiarcos, C., Richter T., Wicke, D. (2015): “Learning Cuneiform the Modern Way.” In: DhD2015

Sukhareva, M., Fuscagni, F., Daxenberger, J., Görke, S., Prechel, D., Gurevych,I. (Aug 2017): “Distantly supervised pos tagging of low-resource languages under extreme data sparsity: The case of Hittite.” In: LaTeCH-CLfL ’17 Proceedings of the 11th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp.95–104 1.1