AI-based Companion Services for Humanities [⋆]

Simon Schiff[1 [0000−0002−1986−3119]], Felix Kuhr[1 [0000−0001−5964−3738]], Sylvia Melzer[2 [0000−0002−0144−5429]], and Ralf Möller[1 [0000−0002−1174−3323]]

[1] Universität zu Lübeck, Institute of Information Systems, Ratzeburger Allee 160,
23562 Lübeck, Germany
{schiff,kuhr,moeller}@ifis.uni-luebeck.de
http://www.ifis.uni-luebeck.de

[2] Universität Hamburg, Centre for the Study of Manuscript Cultures,
Warburgstraße 26, 20354 Hamburg, Germany
{sylvia.melzer}@uni-hamburg.de
https://www.written-artefacts.uni-hamburg.de

Extended Abstract

Different social and cultural aspects have to be considered in order to evaluate manuscripts. It is well-known that even texts cannot always be translated *correctly* without context information. For instance existing Tamil dictionaries such as Google Translate [3] can be used for translating manuscripts from this century. However, using Google Translate for manuscripts from the 12th century results in a poor translation performance since the meaning of some words has changed over time, and certain words have a certain meaning only in a specific context. Generally, the reason for different translations of the same word is given by language change. Language change means that words of languages and their meanings change over time. In addition, the translated words, or more generally, research data arise from a specific context and often make sense only in a specific context. For example the English translation of the Tamil word “acai” could be “to move”, “to be weary”, or “to tie” [1]. We argue that services for humanities are required which consider context-specific information s.t. services can use the most probable translation obtaining context-specific *correct* Tamil translations.

Context-specific services could provide estimating the most probable dictionary for translating manuscripts from various centuries by considering different contexts. If there exist more than one dictionary for the same context a service is required for merging them. Manually processes such as merging dictionaries is a very laborious and time-consuming task. For this purpose we implemented AI-based companion services for automating different tasks, i.e. a service for merging dictionaries from the same context and proposing context-specific translations. Our services presuppose that data from different repositories are in a uniform and machine processable format. In practice, generally, data is formatted to be

[⋆] The research for the extended abstract was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2176 ‘Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures’, project no. 390893796.

readable by humans so that it is difficult to parse these documents by machines automatically. We solved this by implementing a parser for different kind of documents as a preprocessing step to load content from these documents into a project-specific database. For the implementation we use Word documents containing editions, translations, and vocabulary entries from NETamil repository [2]. The vocabulary entries derived from different editions and translations, and therefore these documents are written in the same context.

We assume that translations in the repository associated with additional data making the content more explicit by providing descriptions, references, or explanations about the translations. We denote these entries as location-specific subjective content descriptions (SCDs) which are annotation-specific descriptions in different representations. Location-specific means that each vocabulary entry is associated (i.e. located) with one or more words in the translations.

Kuhr et al. have shown that (location-specific) SCDs provide a value for gathering related documents in the context of a given repository, e.g. classifying new documents [4], or enriching documents with SCDs from other documents from the same repository [5]. These approaches provide retrieval of similar documents, however, retrieval of complementary ones could also be relevant. For instance, the SCDs for dictionaries and materials are differently described and could be similar in a specific context (cf. [6]).

For more effective performance, documents may be annotated with SCDs, then contexts are identified by the SCDs, the similarity of the documents are computed, and thus, context-specific information will be connected to the specific words in the document automatically. Given a new document, the question is: Does that document have anything of value to add in the given context? We are interested in providing services s.t. a person can upload new documents and the system supports the person e.g. using the best translation for words, represent the most similar documents in a given context, etc. Summarized our contributions are: (i) preprocessing of documents, (ii) suggesting context-specific translations for humanities, and (iii) extending a repository with new documents.

References

1. A Dravidian etymological dictionary. https://dsalsrv04.uchicago.edu/dictionaries/burrow/, Accessed: 2020-06-28
2. Going From Hand to Hand: Networks of Intellectual Exchange in the Tamil Learned Traditions. https://www.manuscript-cultures.uni-hamburg.de/netamil/, Accessed: 2020-06-28
3. Google Translate. https://translate.google.de/, Accessed: 2020-06-25
4. Kuhr, F., Braun, T., Bender, M., Möller, R.: To Extend or not to Extend? Context-specific Corpus Enrichment. In: Proceedings of AI 2019: Advances in Artificial Intelligence. pp. 357–368. Springer (2019)
5. Kuhr, F., Witten, B., Möller, R.: Corpus-Driven Annotation Enrichment. In: Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC-19). pp. 138–141 (Jan 2019)
6. Melzer, S.: Semantic Assets: Latent Structures for Knowledge Management. Ph.D. thesis, University of Lübeck, Department of Computer Sciences (Jun 2018)