CEUR-WS.org/Vol-3882/yoda-3.pdf

Refining predicates for relation extraction through thesaurus integration (abstract)

Ewan Hannaford[1], Youcef Benkhedda[2], Marc Alexander[1], Goran Nenadic[2] and Riza Batista-Navarro[2,*]

[1]School of Critical Studies, University of Glasgow, Glasgow, G11 6EW, UK
[2]Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Abstract

Using natural language processing (NLP) approaches combined with humanities expertise, the *Our Heritage, Our Stories* project seeks to connect community archives from across the UK — archives run for and by local communities that contain unique datasets of community-generated digital content (CGDC) — with materials in The National Archives (TNA) of the UK. Foundational to this research is the refinement of NLP methods for their application to CGDC, in order to extract entities appearing in materials (people, places, dates, organisations, etc) and relations between them, and link these across collections.

A key step in this process is the definition of predicates (i.e., relation types) that form the basis of the automated extraction of relationships between named entities. As there may exist many different linguistic formulations of the same relationship, normalisation of these heterogeneous, yet synonymous, forms to a prototypical relation is often necessary to adequately capture similar relationships appearing in diverse materials. This ensures that materials expressing the same meaning in different ways are recognised as doing so, and that connections can be drawn across materials as a result. Wikidata, a canonical knowledge base for linked data approaches, contains an extensive list of relation types, which researchers can map relations appearing in their materials to as a means of relation normalisation. In the *Our Heritage, Our Stories* project, a set of ~30 key relations relevant to CGDC were identified, which covered the core relationships between entities typically presented in community archive materials.

However, these relations, as prototypical expressions, do not, and were not intended to, comprehensively cover the diverse ways in which such relations can be expressed in CGDC. As a result, in order to prevent relations that appeared in CGDC but not in lists of canonical relations going unrecognised and unrepresented, a broader range of expressions was required to capture the full range of linguistic manifestations of relationships between entities in materials. Using the *Historical Thesaurus of English* — the world's largest thesaurus of English — the project team identified and enriched the set of predicates selected from Wikidata with synonymous terms from similar semantic fields, subsequently integrating these expanded relation types into annotation and processing.

This talk discusses this work, explaining how key relations were selected, how synonymous terms for these relations were identified from the *Historical Thesaurus*, and how these were integrated into the project's NLP approaches to improve the automated interpretation of relationships appearing in CGDC. In doing so, it delineates how linguistic thesauri may be used as a means of refining relation extraction and linking, enabling more comprehensive capture of relations whilst maintaining normalisation necessary for linked data approaches. Consequently, it proposes a hybrid approach to relation extraction, demonstrating how NLP methods can be supported by manual integration of linguistic expertise and resources.

Proceedings of the Joint Ontology Workshops (JOWO) - Episode X: The Tukker Zomer of Ontology, and satellite events co-located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), July 15-19, 2024, Enschede, The Netherlands.

[*]Corresponding author.

ewan.hannaford@glasgow.ac.uk (E. Hannaford); youcef.benkhedda@manchester.ac.uk (Y. Benkhedda); marc.alexander@glasgow.ac.uk (M. Alexander); gnenadic@manchester.ac.uk (G. Nenadic); riza.batista@manchester.ac.uk (R. Batista-Navarro)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073