Review of: "RelTopic: A Graph-Based Semantic Relatedness Measure in Topic Ontologies and Its Applicability for Topic Labeling of Old Press Articles"

Silvio Peroni[1]

1 University of Bologna

Potential competing interests: The author(s) declared that no potential competing interests exist.

Metadata

Title: RelTopic: A Graph-Based Semantic Relatedness Measure in Topic Ontologies and Its Applicability for Topic Labeling of Old Press Articles

Author: Mirna El Ghosh, Nicolas Delestre, Jean-Philippe Kotowicz, Cecilia Zanni-Merk, Habib Abdulrab

Submitted to: Semantic Web

Preprint: http://semantic-web-journal.net/content/reltopic-graph-based-semantic-relatedness-measure-topic-ontologies-and-its-applicability

Review

The authors describe a relatedness measure (a.k.a. RelTopic) based on graphs (which represent an ontology) that can be used to label texts with concepts defining their topics. They accompany the presentation of RelTopic with a section about its evaluation, highlighting the pros and cons of the implementation and tests they provided.

It is undoubtedly an interesting article that addresses a particular problem, i.e. classifying texts through an ontology. The whole process proposed starts from a requirement (and hypothesis) which is to have a good representation of a text via a set of entities available in a knowledge graph (in this context, Wikidata). Thus, supposing that this set is provided for each text to label, the approach proposed is fully automatic.

However, I have some issues with two aspects of the article that must be addressed appropriately. The first one is about the narrative, and the second one is about the evaluation. Both of them are introduced as follows, followed by a final remark.
Thus, while I value the authors' article, I believe additional work is needed to be appropriate for publication in Semantic Web.

Narrative

Section 2 of the article ("Problem Definition") highlights the particular research problem the authors are trying to address. The whole work is done in the context of a larger project (ASTURIAS) in which the approach proposed by the authors is just one part of the story (i.e. that of work package 3). It would be better to start from the beginning of the introduction on presenting the ASTURIAS, which research problems it poses, and how the authors want to address them (or part of them), thus focussing on the RelTopic algorithm they propose for labelling texts. In this way, the authors can provide a more concrete and pragmatic setting of the general framework they propose. Then, section 2 should be dedicated only to introducing the research problem into consideration and its initial hypothesis.

Also, since the hypothesis that "disambiguated" entities can be a good proxy for representing a text is a crucial aspect of RelTopic proposal, it is essential to convince the reader that this hypothesis is valid, in particular in the context of the project. I know that this information should come from the people working in WP2 (and they are not necessarily the same authors of this article), but having a clear view about the validity of that hypothesis is crucial to claim the robustness of the approach proposed in this article.

Finally, I think it necessary to specify why RelTopic and the entire framework for labelling text have been proposed in the project context. I believe – but this is not explicitly stated in the article – that the framework's goal is to replace humans with software for labelling, in principle, a vast number of texts, something that would require too much human effort to do it manually. If this is the real goal to achieve, it should be explicitly stated - and, honestly, it does have consequences on the evaluation (see below).

Evaluation

The quality of Topic-OPA and the uses of the scores obtained by using RelTopic indeed is difficult to assess. The best approach to adopt in these cases to check the quality of a whole set of technologies is to see if (a) Topic-OPA + (b) RelTopic + (c) the automatic labelling process enable one to obtain results comparable to what humans do - in particular, if it is true that the goal is to substitute humans with software in the labelling of texts. Of course, this will not provide a specific evaluation for each framework component (i.e. a-c). Still, it would be appropriate to assess the framework overall, when all its components are used in combination, which is adequate for the project from which this work is part of.

However, to do that, one should organise an appropriate testing session with several humans. Something is described in section 8.2.1, but it is very general, and it seems it is not enough to convince a reader that the approach works fine - i.e. in a way which is similar to what humans do. Since even humans can be in contradiction for specific texts, it would be essential to involve, a least, three distinct annotators for each text used in the evaluation, to measure also if there are possible agreement/disagreement between them. My perception is that the authors' framework should show an agreement similar to that shown by humans when they are asked to do the same job. In particular, I think it is not enough to measure how much of the topics identified by the framework matches those returned by (one?) human.

It is also crucial to measure how much the initial hypothesis (i.e. the fact that disambiguated entities may be a proxy of an

article) has consequences on the results obtained. In particular, have the entities used as proxies for the texts in the evaluation been obtained using the software developed in WP2? Or, have they been selected by the authors of this article by hand? In both cases, how are the authors sure that such entities represent good proxies for the texts used in the evaluation?

Thus, a clear explanation of how the testing session has been organised, which information has been provided to humans, which newspapers have been considered (better if more than one, unless the project only focuses on Le Matin), which data have been collected, which results have been obtained, must be specified. Reading the text seems that such a testing session was done naively without considering all these variables. If that is the case, the authors should run a new testing session to gather meaningful data. Otherwise, it is necessary to clarify all the passages in the paper if they already did it.

The proposition of such an evaluation should be: to convince a reader that the authors' framework behaves in a way which is comparable with humans.

Finally, it would be good if all the data gathered in the evaluation, including the software implementing the framework and the documents describing the protocol used for the testing session, could be available online for replicability purposes (e.g. in Zenodo, Figshare, Protocols, GitHub).

A final remark

The approach adopted by the authors seems to be appropriate and generalisable for implementing labelling activities for any text, not only newspaper articles. Is that the case? To what extent? May the authors elaborate more on this aspect?