Empirical Analysis of Semantic Metadata Extraction from Video Lecture Subtitles

Marcos Vinícius M. Borges[1], **Julio Cesar dos Reis**[1], **Guilherme P. Gribeler**[1]

[1]Institute of Computing – University of Campinas (UNICAMP) – Campinas – SP – Brazil

m211847@dac.unicamp.br, jreis@ic.unicamp.br, guigribeler@gmail.com

Abstract. *Video lectures improve the learning experiences considering individual's needs and learning styles. However, the large amount of educational content and their availability turns difficult the tasks of accessing these resources. The extraction of semantic metadata from a video subtitle involves challenges in dealing with informal aspects of language and the detection of semantic classes from the text. In this paper, we conduct an empirical analysis of semantic annotation approaches supported by ontologies in the extraction of relevant metadata from textual transcriptions of video lectures in Computer Science. The obtained results indicate that existing tools can be useful and the semantic metadata extraction process is highly influenced by the underlying ontologies.*

1. Introduction

The growth of information dissemination enabled the easy access for multimedia content that helps in the learning process, resulting in a significant increase in the amount of educational resources available to students. In this context, efforts are required by students to select the appropriate resources in the learning process. Potentially, video lectures from other courses or teachers may be interesting to replace or complement the concepts of a lesson. The filtering and searching of education contents could benefit from techniques exploring the meaning of concepts appearing in the video lectures.

The key challenge in this research is to investigate techniques that allow semi-automatically annotation of text transcriptions from video lectures based on Semantic Web standards and knowledge bases. The required techniques are complex and can be influenced by factors such as the quality of video transcriptions, language, ambiguity and context. This investigation addresses the challenge in creating annotations in text as metadata that associate concepts represented in an ontology with a particular piece of text or multimedia resource.

In this paper, we analyze existing semantic annotation tools to enable extraction of relevant semantic metadata from video lectures. These metadata must be able to describe the video well so it could be used as input to automatic semantic-enhanced recommendation methods. Our contribution enables further analysis of semantically annotated videos using existent annotation tools associated with general-domain or specific ontologies.

2. Study Design

We conducted experiments to assess the quality of semantic annotations obtained from a set of real-world video lectures in Computer Science area available on *Youtube*. The semantic metadata extraction process automatically retrieved the subtitles from a video

lecture in a textual format. The procedure used these textual subtitles as input for existing semantic annotation tools.

Our investigation considered software tools for semantically annotating texts such as *AutoMeta*[1], *CSO-Classifier*[Salatino et al. 2018], *NCBO Annotator* and OntoText [Kiryakov et al. 2004]. As support for the semantic annotation task we used the following ontologies: *DBpedia* ontology, Computer Network Ontology[2], and two releases of Computer Science Ontology (CSO)[Salatino et al. 2018].

3. Results

Table 1 presents the obtained results for a total of seven setups considered to conduct the evaluation, with average results and confidence interval for the mean of 95%, representing an interval of plausible values for population mean to analyze the overall effectiveness for each setup.

Table 1. Overall results. Table presents the tool's name, ontology used, distinct relevant terms (DRV), distinct terms annotated (DTA), distinct relevant terms correctly annotated (DRA), precision (Pr), recall (Re) and f-score

Tool	Ontology	DRV	DTA	DRA	Pr	Re	F-Score
AutoMeta	DBPedia	49 [36; 61]	62 [44; 81]	20 [18; 30]	0.318	0.416	0.360
AutoMeta	Computer Science V_1	49 [36; 61]	50 [26; 64]	13 [7; 20]	0.278	0.290	0.283
AutoMeta	Computer Science V_2	49 [36; 61]	21 [13; 28]	11 [5; 17]	0.553	0.243	0.337
CSO-Classifier	Computer Science V_1	49 [36; 61]	36 [22; 49]	13 [8; 18]	0.383	0.282	0.324
CSO-Classifier	Computer Science V_2	49 [36; 61]	25 [14; 36]	15 [7; 23]	0.633	0.324	0.428
Ontotext	DBPedia	49 [36; 61]	32 [20; 44]	7 [4; 10]	0.193	0.169	0.180
NCBO	Computer Network	49 [36; 61]	10 [5; 15]	6 [2; 10]	0.838	0.324	0.467

We found that the ontology used by the annotation tools plays an important role in the task of annotating the terms. A higher coverage of concepts matching with relevant terms in the video leads to better results. The results obtained with domain-specific ontologies (CSO V_1, CSO V_2 and Computer Network) showed that the number of distinct terms annotated (DTA) and the number of distinct relevant terms (DRA) decreased in general. However, the overall results for precision and recall were higher using these ontologies

4. Conclusion

Our findings point out that obtained annotations considering an ontology related to the specific domain can achieve more precise results, even though less domain-specific ontologies like *DBpedia* can help in the process. Our experimental results were relevant to understand which parts of the whole metadata extraction process can influence the most on the quality of the extracted metadata. Future work involves the development of further techniques to enrich a computer science ontology from book resources.

References

Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics*, 2(1):49–79.

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., and Motta, E. (2018). The computer science ontology: a large-scale taxonomy of research areas. In *International Semantic Web Conference*, pages 187–205. Springer.

[1]https://github.com/celsowm/AutoMeta

[2]https://bioportal.bioontology.org/ontologies/CN