A Prototypical Semantic Annotator for A Tribuna Newspaper

Elias de Oliveira[1], Wesley Silva[1] Juliana P. C. Pirovani[2], Jean-Rémi Bourguet[3]

[1]Programa de Pós-Graduação em Informática
Universidade Federal do Espírito Santo
Av Fernando Ferrari, 514, Goiabeiras – Vitória, ES 29075-910

[2]Departamento de Computação
Universidade Federal do Espírito Santo
Alto Universitário, s/n – Guararema – Alegre, ES 29500-000

[3]Departmento de Ciência da Compuação,
Universidade Vila Velha, Vila Velha, Brasil

jean-remi.bourguet@uvv.br, {elias,juliana}@lcad.inf.ufes.br

Abstract. *The issue of recommending an appropriate piece of information has become essential for the news portals. In this context, a well founded ontological layer represents actually an indispensable artifact to suggest relevant news for the readers. However, news agencies still need to mine their data in order to discover valuable knowledge. In this paper, we present a prototypal automatic semantic annotator for the regional Brazilian newspaper called A Tribuna. Founded on a set of inductive algorithms allowing to classify newspapers in Portuguese and extract named entities from them, our approach describes the standardized categorization and the semantic matching with DBpedia, the so-called nucleus of the Linked Open Data Cloud. We discuss the limitation of our prototype and draw some challenging perspectives to face them. Finally, our proposal paves the way to a new kind of recommendation-based systems.*

1. Introduction

Nowadays, we surely can encounter a large amount of information from various news portals around the world. News agencies have become modern mining platforms by continuously gathering new facts and producing new narratives into their printed or digital materials. One can use this font to look into what is happening in a city, state, or country. The traditional way of reading newspapers is by browsing their pages in order to discover interesting items. Nevertheless, with the recent development of advanced methodological and computational artifacts, the newspaper portals grow into producers of suggestions for their readers. As mentioned, this huge bunch of data has to be well managed to perform accurate recommendations by selecting particular items and discarding others. For example, if a given user explicitly indicates a preference for a kind of news concerning innovations in *Information Technology*, the recommendation system may recommend some new *Artificial Intelligence*-based tools. Such an approach is founded on content-based filtering [Pazzani and Billsus 2007]. On the other hand, it exists a collaborative filtering approach [Herlocker et al. 2004] in which the system generates a group of similar users in terms of interests producing a recommendation based on the analysis

of their characteristics. In order to be able to release a competitive recommendation-based system, a well founded ontological layer represents an indispensable contemporary artifact [Cantador et al. 2008]. Indeed, by taking advantage of accurate semantic annotations [Wetzker et al. 2009] and domain-specific semantic networks [Nguyen et al. 2014], such systems are able to recommend items which a priori would not be revealed through classical Vector Space Model techniques [Baeza-Yates and Ribeiro-Neto 2011].

In [Branquinho-Filho and Oliveira 2017], an experimental approach is released to classify journalistic documents published on a newspaper *A Tribuna*. Founded in 1938, *A Tribuna* is currently the main regional newspaper of Espírito Santo - Brazil. Later, another work extended this approach by providing a framework extracting named entities [Pirovani and Oliveira 2017]. In order to enrich semantically the news metadata, a Named Entity Recognizers (NER) is classically used to map elements of a document with some well known instances on the Semantic Web (see [Troncy 2008] for example). Although complex, NER is an important task to the *Natural Language Processing* (NLP). It can be understood as the task of identification and annotation of the *Named Entities* (NEs) in free-written text corpora. In turn, a NE can be defined as a sequence of words that is capable of representing a real-world entity [Zhang et al. 2019]. In Portuguese, few works have tackled NER-based tasks [Pirovani et al. 2019]. Among these proposals, the hybrid approach CRF+LG proposed in [Pirovani and Oliveira 2017] outclassed the results obtained by other systems performing under equivalent conditions. *Conditional Random Fields* (CRF) is a supervised statistical algorithm to predict an output vector $\mathbf{y} = \{y_0, y_1, \cdots, y_T\}$ based on the random variables given an observed features' vector $\mathbf{x}$, an input vector of features $\{x_0, x_1, x_2, \cdots, x_T\}$ [Sutton and McCallum 2011]. One of the features, y_s, can conceptualize the NE class. Then, the goal is to maximize the number of labels $y_s \in \mathbf{y}$ that are correctly classified, mapping $\mathbf{x} \mapsto y_s$, for each s. The CRF+LG algorithm works first classifying the entities using *Local Grammars* (LG), using the Unitex tool[1]. In addition to post-tagging structures, the LG result is used as an additional feature for training the CRF model. The key strength of this approach is the combination of the probabilistic and linguistic models. Even if the newspaper texts have a generic structure, they gather heterogeneous data with content related to economics, politics, sports, entertainment, among others. The literature already pointed out that the NEs present in these texts can represent efficient supports for information retrieval, whether used as indexing items [Pirovani et al. 2018], for clustering tasks [Spalenza et al. 2019], for classification tasks [Nadeau and Sekine 2007], or for automatic question generations [Pirovani et al. 2017].

In this work, we present a prototypal automatic semantic annotator for *A Tribuna*. Supported by the aforementioned seminal proposals, our approach describes a standardized categorization of the news articles and a semantic matching of their contents with DBpedia.

The rest of this paper is structured as follows: Section 2 presents some related works. Section 3 describes the upstream inductive approach of our automatic annotator while Section 4 presents the downstream semantic categorization and matching. Finally, Section 5 presents the main conclusions so far achieved at this phase of our prototype release.

[1]https://unitexgramlab.org/

2. Related Works

Semantic annotation is not a novelty for the journalists. For a long a time, they have been using their proprietary semantic tools to manually annotate contents of news items by filling in some forms. For example, BBC created its own ontology[2] for modeling its own news articles.

Early in 2000, the project *PlanetOnto* [Domingue and Motta 2000] extended a news server by providing support for ontology-driven document formalization integrating browsing and deductive knowledge retrieval, personalized news feeds and alerts, and proactive identification of potentially interesting news items. Soon after, *NAMIC (News Agencies Multilingual Information Categorisation)* [Basili et al. 2001] was released as an architecture to extract relevant facts from the news streams of large European news agencies and to support semantic inferences by aligning the extracted concepts with EuroWordNet objects [Vossen 1998]. Neptuno [Castells et al. 2004] presents an emergent semantic-based technologies to improve the processes of creation, maintenance, and exploitation of the digital archive of a newspapers based on a knowledge base supported by an ontology for the description of journalistic information, a semantic search module and a module for content browsing and visualization. *PENG (Personalised nEws coNtent programminG)* [Pasi et al. 2006] provides a set of functionalities for gathering, classifying and filtering heterogeneous news materials (television, radio, magazine) considering a number of individual interests.

Indeed, news items can be represented in a lot of formats like XML-based ones for instance. Nevertheless, News Industry Text Format (NITF) or NewsML remains two of the most widespread formats standardized by the International Press Telecommunications Council (IPTC). Thus, a relevant initiative related in [Troncy 2008] tries to bring the IPTC news architecture into the Semantic Web by designing a workflow to populate computational ontologies and enriching semantically the news metadata by processing text and performing visual analysis of photo and video of news items. Named Entity Recognizers such as GATE[3], SPROUT[4] or OpenCalais[5] have been used. Once the named entities are extracted, they are mapped to well known instances on the web (using Geonames for the locations and DBPedia for the persons and organizations).

As it is mentionned, the field of newspapers has already been tackled by approaches dealing with artificial intelligence and semantic web techniques in reason of potential social fallouts. For example, some works relate how the journalists can use the Semantic Web standards to support the news angles creation and news production [Heravi et al. 2012, Opdahl and Tessem 2020, Panagiotidis and Veglis 2020]. But, in [Moreno et al. 2015], the authors point out a certain distance between metadata standards identified in the literature review and those in the HTML tags of the newspaper industry emphasizing the importance and needs of semantic alignments. Different initiatives have proposed ontological-based infrastructures in the journalism domain focused on linked open data-based strategies.

[2]http://www.bbc.co.uk/ontologies/storyline
[3]http://gate.ac.uk/
[4]http://sprout.dfki.de/
[5]http://www.opencalais.com/

In [Hopfgartner and Jose 2010], the authors extract named entities from news videos teletext using OpenCalais. Moreover, OpenCalais WebService is used to compare the actual entity string with an up-to-date database of entities and their spelling variations. This disambiguation maps these entities with a Uniform Resource Identifier (URI) and their representation in DBpedia. Then, the authors exploit the Linked Open Data Cloud (i.e. by using the SKOS vocabulary in DBpedia) to identify similar news stories that match the users interest. In [Aksaç et al. 2012], the authors release a semantic web browser that allows users to browse news web pages and also access related data resources via annotation and a side-bar listing all found linked data resources. In [Papadokostaki et al. 2017], the authors present an integrated platform dedicated to news articles, providing storage, indexing and searching functionalities by using semantic web technologies and services.

Finally, one of the last and most complete proposal was realized through the European project NEWS (News Engine Web Services) [García et al. 2006] consisting of a set of facilities like automatic extraction of metadata from news items' contents, named entity disambiguation [García et al. 2012], storage, retrieval of news items. Their NEWS ontology [García et al. 2010] covers the different types of metadata that can be attached to a news item: management, categorization and content metadata. The project also produced components and algorithms [García et al. 2007] that automatically detect entities and events mentioned in a newspaper text and link them to instances in their NEWS ontology. Notice that the content annotation module of the NEWS ontology is partially inspired by SUMO [Niles and Pease 2001] and MILO [Niles and Terry 2004].

3. Upstream Inductive Approach

Our goal has two folds. Firstly, we want to assign one of the twenty-one possible subject topics to the news article: 1) *Atualidades (Current Affairs)*, 2) *Qual a Bronca? (What's up?)*, 3) *Cidades (Cities)*, 4) *Ciência e Tecnologia (Science and Technology)*, 5) *Concursos (Public-Exam Competitions)*, 6) *Economia (Economy)*, 7) *Esporte (Sports)*, 8) *Especial (Special)*, 9) *Família (Family)*, 10) *Imóveis (Real State)*, 11) *Informática (Computers & Electronics)*, 12) *Internacional (International)*, 13) *Minha Casa (My Home)*, 14) *Mulher (Woman)* 15) *Opinião (Opinion)*, 16) *Polícia (Police)*, 17) *Política (Politics)*, 18) *Regional (Regional County)*, 19) *Sobre Rodas (On Wheels)*, 20) *Tudo a Ver (Everything to do with)*, 21) *TV Tudo (All TV)*. To accomplish this goal, we will apply a similar but improved approach used in [Branquinho-Filho and Oliveira 2017], where each news article document was turned into a vector of weighted word-frequency, known as the bag-of-words approach.

Secondly, we use some NLP tools and methodologies to perform the meta-information extraction from the news free-text format. The meta-information we are interested in are any of the five possible named entities: 1) *Organization* (ORG), 2) *Person* (PER), 3) *Local* (PLC), 4) *Time* (TME), and 5) *Value* (VAL), mentioned in the news-article texts.

The news-article objects are all in PDF format at the site https://tribunaonline.com.br/. So, we downloaded them and extracted $45{,}908$ articles to perform the undermentioned algorithms.

3.1. The Classification of Topics Problem

The classification of documents is a hard task these days of information overload [Bawden and Robinson 2009]. The problem we have at hand is to deal with $45{,}908$ news articles, a tiny portion of the total newspaper archive of only one information source. Hence, arguments in favor of automation of this activity are unnecessary. Table 1 shows the number of documents in each class of this data set used for our experiments.

Class	#Docs	Class	#Docs	Class	#Docs
Atualidades	5617	Especial	1470	Opinião	1634
Qual a Bronca?	346	Família	442	Polícia	4671
Cidades	5234	Imóveis	124	Política	5918
Ciência e Tecnologia	470	Informática	1506	Regional	1802
Concursos	309	Internacional	2187	Sobre Rodas	352
Economia	6558	Minha Casa	37	Tudo a Ver	30
Esporte	6657	Mulher	103	TV Tudo	440

Table 1. The number of documents within each class.

While in the first, third, and fifth columns, we show the name of those classes also presented in the introductory part of Section 3, in the second, fourth, and sixth columns the quantities of documents for their respective class. Note that the majority of news article documents are about *Esporte* (Sports) class, with $6{,}657$ files – shown in the last line of the first column. Whereas, the least populated is *Tudo a ver*, with only thirty document files, in the line before the last, in the fifth column.

After transforming each news article in a weighted word-frequency vector, we selected a sample of these documents to serve as training, and the remaining used to test the classification algorithm. Within the training sample, we also subdivide this sample into two sets: one for the actual training and another for validation, plying as testing to maximize the classifier performance. We used the *Gradient Boosting* algorithm implemented into the *scikit-learn*[6] for the Python programming language.

The goal is thus to mimic humans assigning each news articles to the already known subject class to which they belong. The process took all together less than fifty minutes, and the quality measured – more than 98% of accuracy – is much encouraging, as to the best of our knowledge we do not know a similar performance figure to do a pair comparison when a similar task is carried out totally by humans.

3.2. The Named Entity Recognition Problem

Mining pieces of meta-information from free-texts is still a challenging task in the academia and industry [Augenstein et al. 2017, Nadeau and Sekine 2007]. Researchers are trying to catch up with the performance level reached by NERs systems in English also for the Portugues language [Collovini et al. 2019].

A hybrid approach combining CRF+LG proposed in [Pirovani et al. 2019] is the same strategy adopted in our experiments. This approach, besides the advantage already

[6]https://scikit-learn.org/stable/

mentioned in Section 1, requires lesser training data to achieve competitive results. For the sake of illustration, in our experiments, we manually annotated 100 newspaper documents of the *A Tribuna*[7] dataset. In total, we found $1{,}354$ NEs of the class Person, for instance. In Table 2, we depicted the results when using 80% of documents for training and the remaining for test.

Method		Metrics		
		P	R	F1
LG		60.72	46.41	52.61
CRF+LG		55.00	58.97	56.92
	ORG	18.17	47.35	26.26
	PER	58.64	70.45	64.01
	PLC	43.24	41.27	42.23
	TME	78.76	79.92	79.34
	VAL	53.80	33.21	41.07

Table 2. Results of automatic extracting NEs

The first column, in Table 2, shows the strategy used to yield the results in the following columns. The LG and the CRF+LG approaches are respectively in the third and fourth lines. From the fifth line onward, are the best results by the CRF+LG approach.

The Precision, Recall, and F1 metrics are respectively in the third, fourth, and fifth columns. The worst results for F1, an average of 26.26%, is that carried out by the CR+LG for the ORG entity class, in the fifth line. The best figure for recall, 79.34%, in the fifth column, eighth line, is obtained for the TME class. The LG was superior to the CRF+ÇG in precision, but the later was superior in all the remaining metrics.

Once the newspaper item output the upstream inductive approach, our prototype will perform standardized categorizations and semantic matchings with DBpedia.

4. Downstream Semantic Alignments

The initial approach presented in the Section 3 was thought to support an internal categorization of the news article for archival data storage. As presented in the subsection 3.1, there were 21 selected categories. Our approach plugs a semantic annotator downstream of the inductive algorithms. The IPTC NewsCodes defining 36 thesauri, we will focus on items subjects consisting of about 1400 terms organized into a taxonomy of three levels. Each Subject Reference is identified by an eight-digit decimal string. The terms are organized in a taxonomy as described in the Figure 1. Originally, the subsumption relationship is not explicit but instead encoded into the coding scheme identifying the terms. For example, "survey" (`subj:13006001`) is narrower than "research" (`subj:13006000`) which is narrower than "science and technology" (`subj:13000000`) because they share the two and the four first digits.

[7] http://www.inf.ufes.br/∼elias/dataSets/aTribuna-21dir.tar.gz

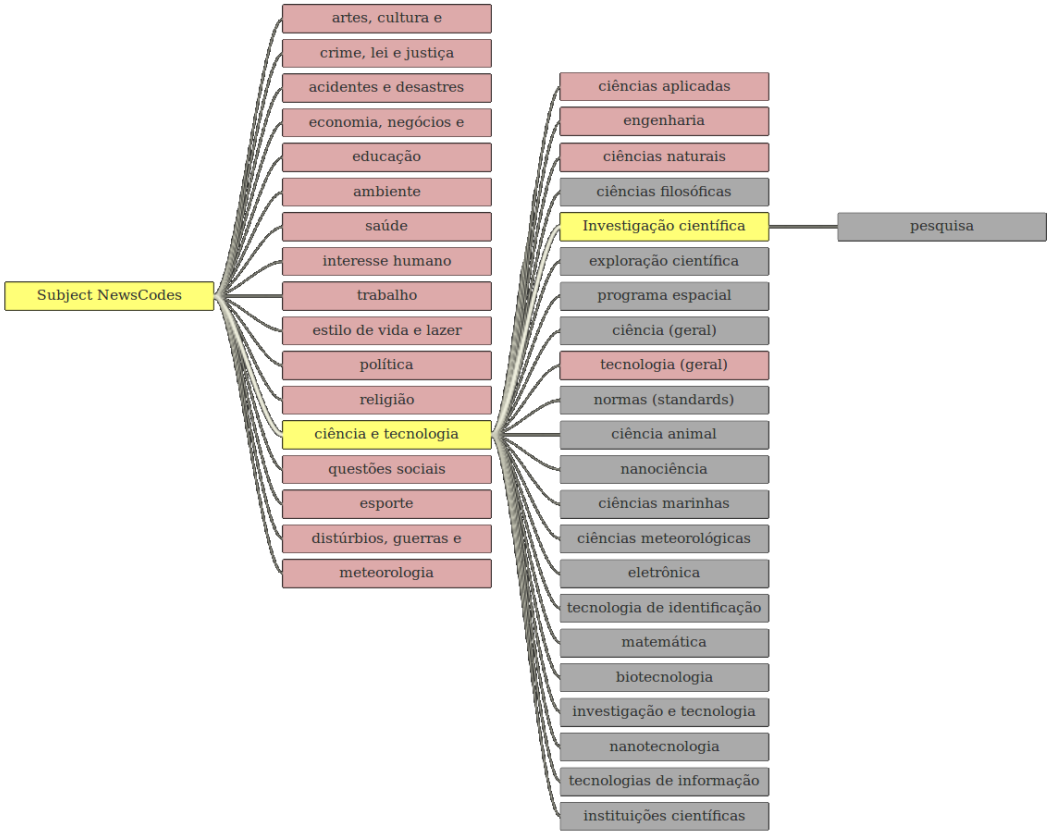

Figure 1. IPTC NewsCodes Subjects Code Taxonomy

IPTC shares its Controlled Vocabularies (CV) by a server at `http://dev.iptc.org/NewsCodes-CV-Server`. The users can use this server for the retrieval of full CVs or only single concepts. The datasets of the CVs and concepts are delivered in five different formats: HTML as human readable variant, and NewsML-G2 Knowledge Items (XML), RDF/XML or RDF/Turtle and JSON/JSON-LD as primarily machine readable variants. Thus, the thesauri are extractable into SKOS, an application of RDF, making the subsumption relationships explicit (*i.e.* `skos:narrower`, `skos:broader`). Each term is thus identified by a dereferencable URI. Moreover, SKOS allows to encode other semantic relations with other controlled vocabularies (from IPTC or from other vocabularies in the LOD).

For example, the two triples presented below and extracted from the IPTC NewsCodes Subjects vocabulary called `cptall-en-GB.rdf` described the fact that the topic *scientific research* (scientific and methodical investigation of events, procedures and interactions to explain why they occur, or to find solutions for problems) with the URI `medtop:20000735` is exactly similar to the subject *research* (a methodical investigation of events or procedures to explain why they occur, or to find solutions for problems) with the URI `subj:13006000`; and presents a semantic similarity with the subject *survey* (examination of public attitudes on various subjects or issues, such as the quality of goods, the value of services) with the URI `subj:13006001`.

```
subj:13006000 skos:exactMatch medtop:20000735
subj:13006001 skos:closeMatch medtop:20000735
```

One of the objectives of our framework is to provide an automatic semantic news categorization. Our semantic categorization module is producing assertions by using a own made controlled vocabulary called `cjat.rdf`. In this vocabulary, we encoded some alignments between the subjects from *a Tribuna* (those presented in Table 1) with the IPTC NewsCodes Subjects. In Figure 2, the vocabularies `cjat.rdf`, `cptall-en-GB.rdf` and their alignments are visualized by the *sparna-labs skos-play* [8].

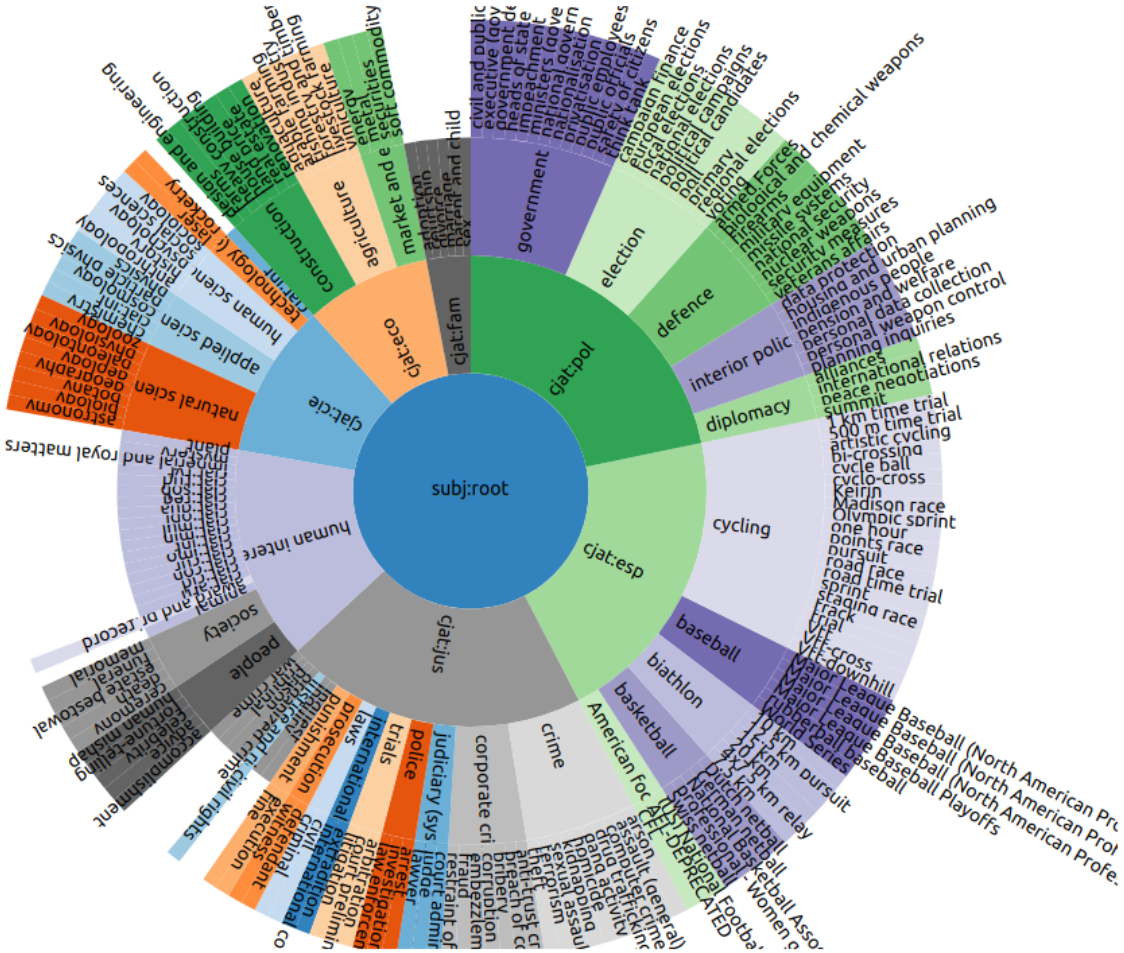

Figure 2. Alignments in `cjat.rdf`

All the outputs obtained by the NER algorithm presented in previous section will be used as potential matchers with resources on DBpedia. We use the DBpedia Lookup Service[9] to look up DBpedia URIs by related keywords (label or anchor text). Note that, in this prototype version, we retain the first ranked resource as matcher. Once the type of the article is encoded and all the matchers are gathered, our writer module will build the RDF representation of the news article. The URI of the article is formed by the concatenation of the *namespace da Tribuna* with the date of the edition, the beginning

[8] http://labs.sparna.fr/skos-play/
[9] https://wiki.dbpedia.org/lookup/

page and an integer representing the id of the article in this page. In the Figure 3, we present an article published on the 1st of March 2019 in the fourth page and identified by the URI `https://tribunaonline.com.br/010319p04a2`. After inputting the automatic semantic annotator and then the writer module, the RDF excerpt is produced. We chose to use the BBC Creative Work Ontology [10] to support this phase. The reasons are multiple: this is a core ontology well established and recognized, the object property `category` (resp. `tag`) allows to encode properly the categorization of the article (resp. the relation with the extracted NEs) and alignments are already encoded towards other vocabularies.

At the end of our current workflow, the RDF Graph of the article is obtained by using the `ontology-visualization` API[11].

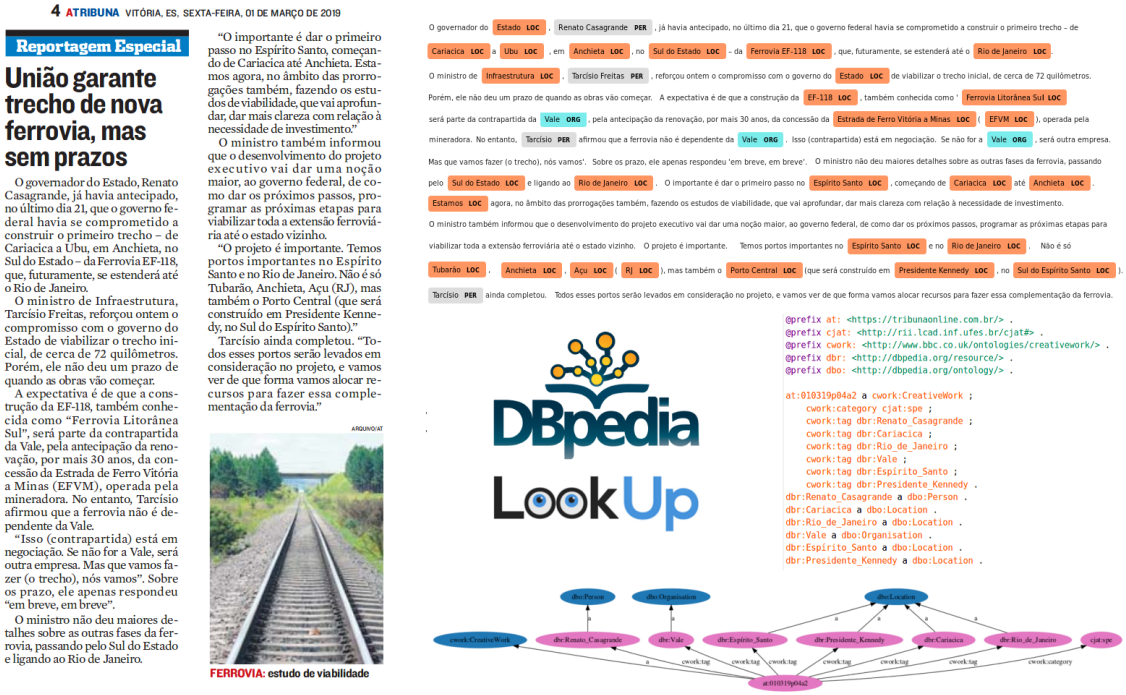

Figure 3. Prototype

Even if our approach is promising, some important issues have to be solved. A disambiguation can be required to select the right matcher. For example, `http://dbpedia.org/page/Vale` maps actually to more than 30 possible resources. Note that, such a disambiguation process could also be performed among the ranked output resources of the DBpedia Lookup output. We intend to browse a small part of the DBPedia semantic network in order to find some evidence to select (or not) the best matcher. Another recurrent issue is the absence of the proper resources in DBpedia but not in Wikipedia. In the news presented above, it was the case for `https://pt.wikipedia.org/wiki/EF-118`, `https://pt.wikipedia.org/wiki/Estrada_de_Ferro_Vit%C3%B3ria_a_Minas` and

[10] https://www.bbc.co.uk/ontologies/creativework
[11] https://github.com/fatestigma/ontology-visualization

https://pt.wikipedia.org/wiki/Ferrovia_Litor%C3%A2nea_Sul. As wikipedia webpages represent semi-structured semantic datasets, it could be interesting to process them also.

Finally, we plan to perform matchings with other dataset among the LOD Cloud. An intuitive perspective would be to perform the disambiguations of the locations by consulting Geonames or Geoplanet in addition. Implementing a system of vote among the matchers could support such a task.

5. Conclusions

In this paper, we presented a prototypal semantic annotator from a free-text archive of the regional Brazilian newspaper called *A Tribuna*. Founded on a set of inductive algorithms allowing to classify newspapers in Portuguese and extract NEs from them, our approach describes both an upstream inductive approach and a downstream semantic categorization alignments with IPTC NewsCodes Subjects taxonomy and NEs matching with DBpedia. We also discussed the current limitation of our prototype and draw some challenging perspectives to face them. Nevertheless, up to this stage, we are ready to automatically produce recommendations supported by an ontological layer browsing. In our example, we could easily imagine to recommend news concerning past governors by using the richness of the semantic network in DBpedia.

For future work, we intend to improve the non-symbolic part of our approach. This is a promising line of research especially with respect to the minimization of the manpower efforts [Oliveira et al. 2014, Spalenza et al. 2019]. We also intend to deal with a larger and more precise NER to improve the quality of the performance of our algorithms [Collovini et al. 2019, Pirovani et al. 2019]. We also project to deal with other domains like institutional repositories, educational materials and jobs offerings. Mentioning this, one of the possible horizons in terms of potentially suitable products would be to release a Graphical User Interface supporting the Semantic Web ideals [Bourguet 2017].

References

Aksaç, A., Ozturk, O., and Dogdu, E. (2012). A Novel Semantic Web Browser for User Centric Information Retrieval: PERSON. *Expert Syst. Appl.*, 39(15):12001–12013.

Augenstein, I., Derczynski, L., and Bontcheva, K. (2017). Generalisation in Named Entity Recognition: A Quantitative Analysis. *Computer Speech & Language*, 44:61 – 83.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. Addison-Wesley, New York, 2 edition.

Basili, R., Catizone, R., Padró, L., Pazienza, M. T., Rigau, G., Setzer, A., Webb, N., Wilks, Y., and Zanzotto, F. M. (2001). Multilingual Authoring: the NAMIC Approach. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management@ACL 2001, Toulouse, France, July 9-11, 2001*.

Bawden, D. and Robinson, L. (2009). The Dark Side of Information: Overload, Anxiety and other Paradoxes and Pathologies. *J. Inf. Sci.*, 35(2):180–191.

Bourguet, J. (2017). Worldwide scholarships spreading. In Rus, V. and Markov, Z., editors, *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, pages 670–675. AAAI Press.

Branquinho-Filho, D. and Oliveira, E. (2017). Automatic Classification of Journalistic Documents on the Internet. *TransInformação*, 29(3):245–255.

Cantador, I., Bellogín, A., and Castells, P. (2008). A Multilayer Ontology-Based Hybrid Recommendation Model. *AI Commun.*, 21(2-3):203–210.

Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, V. R., Contreras, J., and Lorés, J. (2004). Neptuno: Semantic Web Technologies for a Digital Newspaper Archive. In Bussler, C., Davies, J., Fensel, D., and Studer, R., editors, *The Semantic Web: Research and Applications, First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, May 10-12, 2004, Proceedings*, volume 3053 of *Lecture Notes in Computer Science*, pages 445–458. Springer.

Collovini, S., Santos, J., Consoli, B., Terra, J., Vieira, R., Quaresma, P., Souza, M., Claro, D. B., Glauber, R., and a Xavier, C. C., editors (2019). *Portuguese Named Entity Recognition and Relation Extraction Tasks at IberLEF 2019*, CEUR Workshop Proceedings. CEUR-WS.org.

Domingue, J. and Motta, E. (2000). PlanetOnto: from News Publishing to Integrated Knowledge Management Support. *IEEE Intelligent Systems and their Applications*, 15(3):26–32.

García, N. F., Arias-Fisteus, J., Sánchez, L., and López, G. (2012). IdentityRank: Named Entity Disambiguation in the News Domain. *Expert Syst. Appl.*, 39(10):9207–9221.

García, N. F., del Toro, J. M. B., Arias-Fisteus, J., Sánchez, L., Sintek, M., Bernardi, A., Fuentes, M., Marrara, A., and Ben-Asher, Z. (2006). NEWS: Bringing Semantic Web Technologies into News Agencies. In Cruz, I. F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *The Semantic Web - ISWC 2006, 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006, Proceedings*, volume 4273 of *Lecture Notes in Computer Science*, pages 778–791. Springer.

García, N. F., del Toro, J. M. B., Sánchez, L., and Bernardi, A. (2007). IdentityRank: Named Entity Disambiguation in the Context of the NEWS Project. In Franconi, E., Kifer, M., and May, W., editors, *The Semantic Web: Research and Applications, 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, 2007, Proceedings*, volume 4519 of *Lecture Notes in Computer Science*, pages 640–654. Springer.

García, N. F., Fuentes, D., Sánchez, L., and Arias-Fisteus, J. (2010). The NEWS Ontology: Design and Applications. *Expert Syst. Appl.*, 37(12):8694–8704.

Heravi, B. R., Boran, M., and Breslin, J. (2012). Towards Social Semantic Journalism. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

Hopfgartner, F. and Jose, J. M. (2010). Semantic User Profiling Techniques for Personalised Multimedia Recommendation. *Multimedia Syst.*, 16(4-5):255–274.

Moreno, M. J. B., Felipe, E. R., Sánchez, J. A. P., Béjar, R. M., and Lima, G. (2015). Metadatos en noticias: un análisis internacional para la representación de contenidos en periódicos. In *II Congreso ISKO España-Portugal. Organización del conocimiento: sistemas de información abiertos*, pages 290–303. Universidad de Murcia.

Nadeau, D. and Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Lingvisticæ Investigationes*, 30(1):3–26.

Nguyen, T. T. S., Lu, H., and Lu, J. (2014). Web-Page Recommendation Based on Web Usage and Domain Knowledge. *IEEE Trans. Knowl. Data Eng.*, 26(10):2574–2587.

Niles, I. and Pease, A. (2001). Towards a Standard Upper Ontology. In *2nd International Conference on Formal Ontology in Information Systems, FOIS 2001, Ogunquit, Maine, USA, October 17-19, 2001, Proceedings*, pages 2–9. ACM.

Niles, I. and Terry, A. (2004). The MILO: A General-purpose, Mid-level Ontology. In Arabnia, H. R., editor, *Proceedings of the International Conference on Information and Knowledge Engineering. IKE'04, June 21-24, 2004, Las Vegas, Nevada, USA*, pages 15–19. CSREA Press.

Oliveira, E., Basoni, H. G., Saúde, M. R., and Ciarelli, P. M. (2014). Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification. In *6th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Rome, Italy. IC3K.

Opdahl, A. L. and Tessem, B. (2020). Ontologies for Finding Journalistic Angles. *Software and Systems Modeling*, pages 1–17.

Panagiotidis, K. and Veglis, A. (2020). Transitions in Journalism–Toward a Semantic-Oriented Technological Framework. *Journal. Media*, 1:1.

Papadokostaki, K., Charitakis, S., Vavoulas, G., Panou, S., Piperaki, P., Papakonstantinou, A., Lemonakis, S., Maridaki, A., Iatrou, K., Arent, P., et al. (2017). News Articles Platform: Semantic Tools and Services for Aggregating and Exploring News Articles. In *Strategic Innovative Marketing*, pages 511–519. Springer.

Pasi, G., Bordogna, G., and Villa, R. (2006). The PENG System: Practice and Experience. In *17th International Workshop on Database and Expert Systems Applications (DEXA 2006), 4-8 September 2006, Krakow, Poland*, pages 445–449. IEEE Computer Society.

Pazzani, M. J. and Billsus, D. (2007). Content-Based Recommendation Systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 325–341. Springer.

Pirovani, J., Alves, J., Spalenza, M., Silva, W., Silveira Colombo, C., and Oliveira, E. (2019). Adapting NER (CRF+LG) for Many Textual Genres. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, volume 2421 of *CEUR Workshop Proceedings*, pages 421–433, Bilbao, Spain. CEUR-WS.org.

Pirovani, J., Nogueira, M., and Oliveira, E. (2018). Indexing Names of Persons in a Newspaper Large Dataset. In *13$^{\underline{th}}$ International Conference on the Computational Processing of Portuguese (PROPOR)*, volume 11122, Canela, RS. Springer.

Pirovani, J. and Oliveira, E. (2017). CRF+LG: A Hybrid Approach for the Portuguese Named Entity Recognition. In *17$^{\underline{th}}$ International Conference on Intelligent Systems Design and Applications: Intelligent Systems Design and Applications*, pages 102–113, Delhi, India. Springer, Springer International Publishing.

Pirovani, J., Spalenza, M., and Oliveira, E. (2017). Geração Automática de Questões a Partir do Reconhecimento de Entidades Nomeadas em Textos Didáticos. In *XXVIII Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1147–1156, Ceará, CE. SBC.

Spalenza, M., Pirovani, J., and Oliveira, E. (2019). Structures Discovering for Optimizing External Clustering Validation Metrics. In *19$^{\underline{th}}$ International Conference on Intelligent Systems Design and Applications: Intelligent Systems Design and Applications*, pages 102–113, Delhi, India. Springer, Springer International Publishing.

Sutton, C. and McCallum, A. (2011). Conditional Random Fields: An Introduction. *Foundations and Trends® in Machine Learning*, 4:267–373.

Troncy, R. (2008). Bringing the IPTC News Architecture into the Semantic Web. In *International Semantic Web Conference*, pages 483–498. Springer.

Vossen, P. (1998). A Multilingual Database with Lexical Semantic Networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Wetzker, R., Umbrath, W., and Said, A. (2009). A Hybrid Approach to Item Recommendation in Folksonomies. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 25–29.

Zhang, H., Boons, F., and Batista-Navarro, R. (2019). Whose Story is It Anyway? Automatic Extraction of Accounts from News Articles. *Information Processing & Management*, 56(5):1837 – 1848.