ABAKÓS
PUC Minas
Instituto de Ciências Exatas e Informática

Hybrid Semantic Annotation: Rule-based and Manual Annotation of the Open American National Corpus with Top-Level Ontology[*]

Anotação Semântica Híbrida: Anotação Baseada em Regras e Manual do Open American National Corpus com Ontologias de Nível topo

Guidson Coelho de Andrade[1]
Alcione de Paiva Oliveira[2]
Alexandra Moreira[3]

Resumo

O processamento de linguagem natural ainda enfrenta o desafio de fazer com que as máquinas compreendam o significado contido nas palavras que ocorrem em uma frase. A anotação semântica ajuda nesse processo adicionando metadados que atribuem significado aos lexemas. Existem diversos aspectos semânticos que podem ser anotados, tais como função, papel semântico e categorias ontológicas. As categorias ontológicas de nível superior adicionam informações sobre a natureza do conceito denotado pelo lexema e permitem eliminar ambiguidades. A proposta de trabalho é uma abordagem híbrida de anotação semântica baseada em ontologias de nível topo aplicadas a um *corpus* em inglês americano. A pesquisa é dividida em duas etapas de anotação, ambas usando as categorias de alto nível topo do Schema.org como rótulos de anotação. Na primeira etapa é criado um anotador baseado em regras, e na segunda etapa é feita uma anotação manual para correção e adição de rótulos no *corpus* anotado na etapa anterior. A contribuição deste trabalho é a geração de um *corpus* anotado que pode ser usado no treinamento de anotadores automáticos.

Palavras-chave: Anotação semântica Híbrida. Anotação Manual. Anotação baseada em regras.

[*] Submetido em 18/12/2018 - Aceito em 19/06/2019

[1] Mestre em Ciência da Computação. Universidade Federal de Viçosa, Brasil– guidson.c.andrade@gmail.com
[2] Doutor em Informática. Universidade Federal de Viçosa, Brasil– alcione@dpi.ufv.br
[3] Doutora em Linguística. Universidade Federal de Viçosa, Brasil– xandramoreira@gmail.com

Abstract

Natural language processing still faces the challenge of getting machines to understand the meaning expressed by words that occur in a sentence. Semantic annotation helps in this process by adding metadata that attaches meaning to lexemes. There are several semantic aspects that can be annotated, such as function, semantic role and ontological categories. Top-level ontological categories add information about the nature of the concept denoted by the lexeme and allow eliminating ambiguities. The proposed work is a hybrid semantic annotation approach based on top-level ontologies applied to an American English Corpus. The research is divided into two annotation steps, both using the top-level categories of Schema.org as annotation labels. In the first step a rules-based annotator is created, and in the second step a manual annotation is made for correction and addition of labels in the corpus annotated by rule annotator. The contribution of this work is the generation of an annotated corpus that can be used in the training of automatic annotators.

Keywords: Hybrid Semantic Annotation. Rule-based Annotation. Manual Annotation. Top-Level Ontology.

1 INTRODUCTION

Corpora are important resources for natural language processing systems. *Corpora* are composed of a set of corpus which, in turn, consist of a set of records of natural language in spoken or written form that can be interpreted by computer applications (PUSTEJOVSKY; STUBBS, 2012). A *corpus* is defined as the set of documents written or spoken in a natural language in order to represent a specific idiom or its linguistic diversity (LEECH, 1997). Typically, to be named as a *corpus*, the records must obey certain criteria and go through standardization and annotation processes to allow information patterns to be found and to facilitate the execution of inferences about textual information (PUSTEJOVSKY; STUBBS, 2012).

In computational linguistics, the term “annotation” refers to the process of adding metadata in a given *corpus* in order to promote the process of extracting information by the applications (PUSTEJOVSKY; STUBBS, 2012). The annotation process is responsible to add value to a raw *corpus*, so it is crucial because the contribution made to it allows any *corpus* to be a source of linguistic data for eventual researches and applications (LEECH, 1997). There are several types of linguistic characteristics that can be added to a *corpus* such as lexical, morphological, syntactic, semantic, among others features to increase its information value (LU, 2014). In this paper the focus is the semantic annotation that assigns some meaning to the content annotated.

Semantic annotation is an annotation task that attempts to explore the meaning of the elements being marked (WILSON; THOMAS, 1997). Semantic annotation searches for elements of the text and classifies them according to their meaning in the fragment in which it is inserted. It allows data to be interpreted by applications in such a way that machines can also capture part of the meaning inherent in the context. Information retrieval is also facilitated by the semantic annotation because it becomes easy to access and understand the structure of the document being analyzed (KIRYAKOV et al., 2004). There are several semantic facets that can be attached to textual elements, such as context, semantic roles, and ontological categories.

In Computer Science, GRUBER (1993) states that ontology is “a specification of a conceptualization”, in other words, it is a description of concepts and relationships that exist between these concepts. There are a few types of ontologies, but in this paper, we are interested in the type called top-level ontologies, also known as upper-level ontologies. According to GUARINO (1998) top-level ontologies describe very general concepts like space, time, matter, object, event, action etc., which are independent of a particular problem or domain. The top-level ontological categories add information about the nature of a term and allow us to distinguish that a term such as “bank” is related to a financial institution rather than a slope. Thus, besides assigning meaning, ontological annotation helps in the disambiguation of terms.

Annotating texts with the concepts originated from a top-level ontology can bring advantages for the semantic enrichment of texts and websites (KIRYAKOV et al., 2004). According to Erdmann et al (2000) ontologies can help guide the process of annotation and its categories can be used as tags for annotating the words. From the annotation and ontology junction, it emerges the semantic annotation based on ontologies that is the method of adding semantic meaning to terms using as guidance concepts and formal specifications of an ontology (ERDMANN et al., 2000). Ontology-based semantic annotating can also be very useful for future applications that needs a substantial amount of data to train algorithms.

However, the process of annotating large textual *corpus* is long and expensive, leading to a lack of semantically annotated datasets. This hinders the development of machine learning systems geared towards natural language. Contributions on the development of semantic annotated *corpora* would help to improve researches in the automatic annotation field. That said, the proposal of this research was to develop a rule-based semantic annotator adopting the concepts of a top-level ontology. It was used to add semantic tags to a pre-selected *corpus*. The rules were created taking into account linguistic features of American English to classify lexemes into the domain of the selected top-level ontology. A second objective of this work was to complement the annotation doing a manual annotation of the same *corpus* pre-annotated in the first place, so to improve the accuracy of the annotation. After those two processes the research yielded a semantically annotated *corpus* based on top-level ontology suitable for future applications and researches on machine learning and for training fully automatic semantic annotators.

This paper is organized as follows: the next section presents the works previously developed that are related to this research; Section 3 describes the materials and methods applied in the research; Section 4 presents the results obtained; and Section 5 presents our final remarks.

2 RELATED WORK

The area of ontological semantic annotation is a new field of research and it is constantly receiving contributions. This section addresses the current works that have some relationship with semantic annotation, rule-based annotation, and the use of ontologies to provide semantics to terms.

The work proposed by ANDRADE et al. (2017) is quite similar to the current research, but they used another approach. They created a rule-based semantic annotator to tag lexemes of a portion of the Open American National *Corpus* (OANC). The work was only applied to a small parcel of the OANC, and it was annotated only by the rule-based annotator. The semantic annotation, although guided by the use of an ontology, differs from this work because the researchers used a general domain ontology called SUMO (Suggested Upper Merged Ontology)[4]. The top-level concepts selected for annotation describe broad concepts applicable to any domain. Our proposal, on the other hand, although it uses a general domain ontology, it relies on ontology created from evidence of use by occurrence in websites and therefore has a more practical character.

ŞIMŞEK et al. (2017) proposed a rule-based tool to validate annotations made using the Schema.org concepts. They performed a task to evaluate annotation performance under a specific domain. Using a tourism domain, the research made a Schema.org annotation validation

[4] <http://www.adampease.org/OP/>

under two aspects: completeness of the annotations and semantic consistency of the annotated values. The focus of the research is to evaluate the rule-based annotator rather than creating an annotated *corpus*.

ALEC et al. (2016) presented a methodology of semantic annotation of documents guided by ontologies. The proposal used domain-specific ontologies to classify the entire document under this domain. The approach made use of rules to establish relations of ontological concepts present in the document and then to assign the corresponding label to the document. This work differs in some points to our research because the proposed approach is to annotate documents by rules rather than words that occur in the document. Another distinction is that domain ontologies are used to assign the tags instead of top-level ontologies.

The work developed by COHEN et al. (2017) focused in the biomedical domain, but has a similar approach to the current work. The authors created a semantically annotated *corpus* using full-text biomedical journal articles and, as a final product, it resulted in The Colorado Richly Annotated Full Text *Corpus* (CRAFT). The *corpus* identifies lexemes related to the ontological concepts of the biomedical area and multi-model annotation task, and the annotation was performed manually through guidelines based on these ontologies. The work resembles ours due to the proposal to produce a final product annotated semantically and ontologically, although the domain and the range of the annotation are different.

KLIEN (2007) reports the development of a semantic annotation strategy based on rules for the geospatial domain. To guide the geodata annotation process, she used specific domain ontologies from the geospatial context. A set of rules, using logical concepts, defines conditions for tagging terms of the geospatial domain. Her research presents a similar approach to our proposal because it uses a set of rules based on logical concepts to annotate the terms of the proposed ontology. The difference is that her rules are based on geospatial aspects, while our rules rely on the linguistic structures of words and sentences.

Similar to our research, KHALILI; AUER (2015) enriched semantically unstructured content using Schema.org. In particular, they embed metadata into unstructured documents on the Web. Its implementation uses RDFace and the MCE Tiny plugin WYSIWSM (What You See Is What You Mean) in text fragments. Excerpts found in web blogs created in WordPress are enriched by the properties of Schema.org. The last step of the model is to use NLP patterns and routines, that is, APIs for automatic annotation of named entities. They do not assess the benefits of their implementation by claiming that they were presenting the initial phase of the research.

3 MATERIALS AND METHODS

The research presented here, as previously mentioned, addresses the annotation of a *corpus* in a hybrid manner using a top-level ontology. Based on this initial requirement, the Schema.org ontology was selected to lend its ontological classes as tags for the semantic annotation, and the OANC project was selected as the *corpus* for the annotation process. Both projects will be described in the following paragraphs.

Schema.org is a collaborative project aiming to create structured data schemas for on-page markup in order to help search engines understand the information on web pages and provide richer search results[5]. A shared markup vocabulary makes easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts. Markups can also enable new tools and applications that make use of the structure. In this sense web pages may have machine-understandable information since their data uses the markup vocabulary proposed by Schema.org (PATEL-SCHNEIDER, 2014). It is formed by a set of categories, organized under a hierarchical structure and establishing nested relations between them. The categories are based on sets of vocabularies widely used by consumers and web page editors. In that regard, the hierarchy proposed by Schema.org has a close relationship with the use of the web and the need to have single integrated schema capable to cover the vocabulary of a wide range of topics contained in the internet (GUHA et al., 2016).

Schema.org, although it is not considered a formal global ontology that aims to classify all things in the world, as stated in its website[6], it presents itself in a hierarchy of categories, each one with its own properties and relations (PATEL-SCHNEIDER, 2014). According to RONALLO (2012), Schema.org is defined as a "middle ontology", in other words, it is an ontology that does not intend to cover everything that exists in the world, and neither to go deep into a specific area. Schema.org's main intention is to create a hierarchy capable of addressing content from the web common cases and applicable to the promotion of web information that can be understood by machines and search engines. From this perspective, Schema.org was selected as the ontology to be used in the proposed semantic annotation.

As already mentioned, Schema.org is structured as a hierarchy, being organized by upper and lower categories. To better explore and understand the Schema.org hierarchy, a solar burst chart was constructed based on the Full Hierarchy of Schema.org[7], as can be seen in Figure 1. The graph shows in gray the top class "thing" and the proportion of the division of the other sub-classes. Our interest is focused on the sub-classes immediately below the top class. In its first level, it has the top-level type called "thing", in gray, that includes all the other classes. The root node has 8 sub-categories (action, creative work, event, intangible, organization, person, place and product), as shown in the chart, so they are considered top-level categories of the ontology and have been selected as the tags for semantic annotation. After the ontological classes have been defined, the next step was to carry out the *corpus* selection.

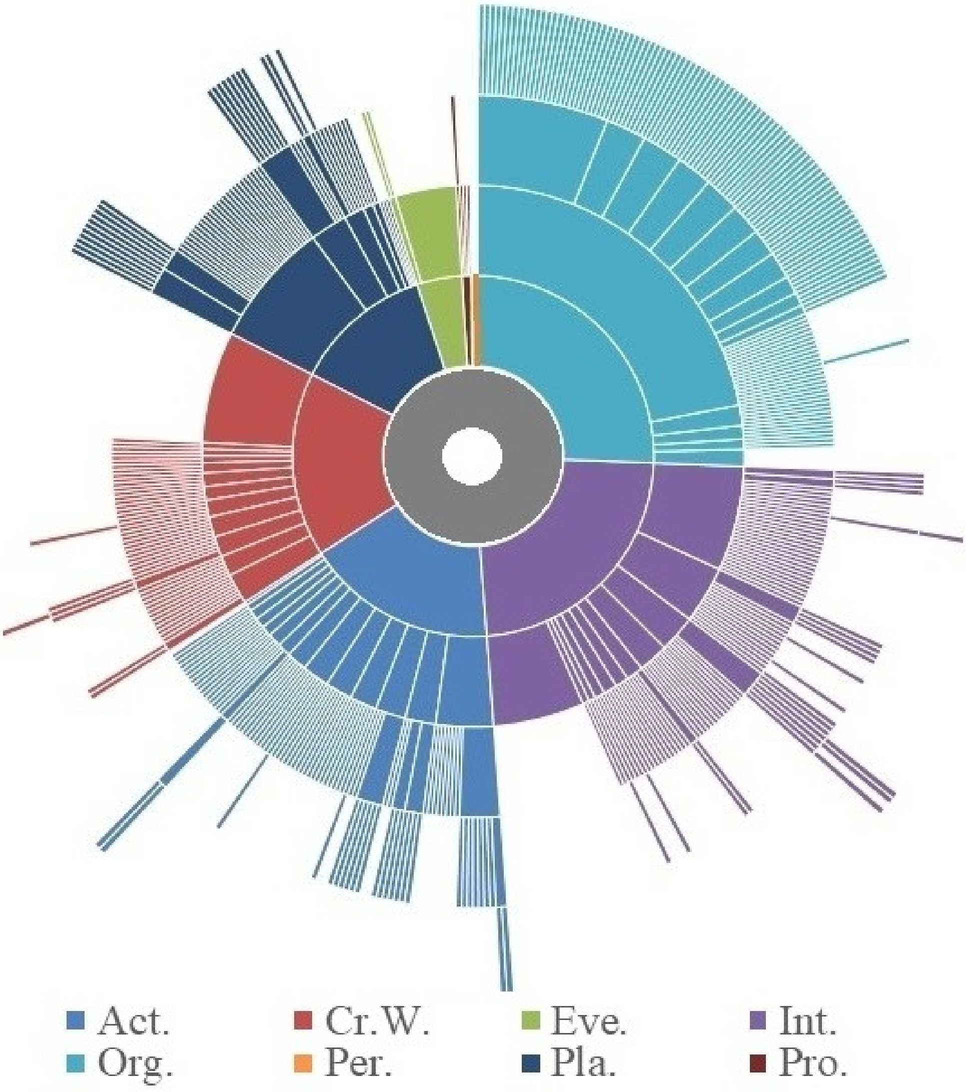
In order to express the linguistic variety of the American English we chose the OANC (IDE; SUDERMAN, 2004) as our standard *corpus*. The *corpus* is composed of a wide range of texts from technical articles, letters, governmental transcripts to oral speech data such as phone calls and face to face conversations (IDE; SUDERMAN, 2004). This diversity of genres allows to express a large number of words in American English and covers expressions of

[5] <http://schema.org/docs/faq.html>

[6] http://schema.org/docs/datamodel.html

[7] http://schema.org/docs/full.html

Figure 1 – Schema.org Sunburst Graphic. Act=action, cr. W.=creative work, Eve=event, Int=intangible, Org=organization, Per=person,Pla=place, and Pro=product



Source: <http://blog.schema.org/>.

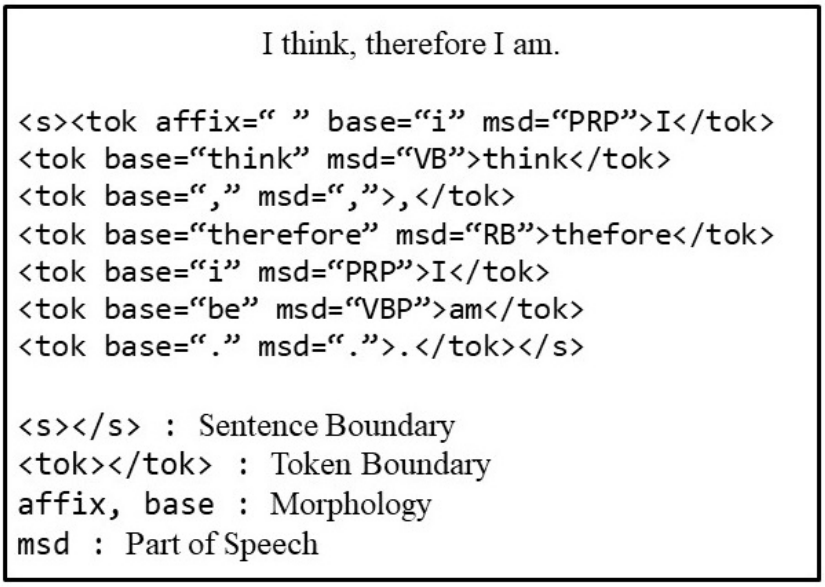
general and specific domains. The dimension of the OANC is approximately 15 million of words distributed among over eight thousand files. The *corpus* also provides some previous annotations such as structural markup, sentence boundaries, part of speech, noun chunks, and verb chunks, produced automatically using annotation systems (IDE; SUDERMAN, 2004). The entire *corpus* is available for download at the American National *Corpus* website free of any charge. All characteristics mentioned above justifies the choice of the *corpus* for this research.

The OANC is provided in .xml format according to the standard required by (ISO 24612), Linguistic Annotation Format (LAF). It was necessary to pre-process the entire *corpus* so that it could be used as input to the annotation system. First, it was necessary to remove all the markup tags related to the XML language and annotations related to the document structure, such as titles, paragraphs, sections, topics, identifications, and everything that was irrelevant. It was necessary to make some adjustments on the *corpus* as well. Due to the *corpus* annotations having been made automatically, some corrections were needed in order to repair some spurious annotations. The most frequent error encountered were in sentence boundaries, because in some cases, such as decimal numbers, acronyms, suspension points, colon and other cases that involve period signs, the annotator committed errors related to sentence ending. Another correction was the joining of numbers and words previously arranged as separated tokens and the other way around. In documents created from speech data it was necessary to change the marking labels and to identify the speaker of each sentence. Some acronyms and proper noun abbreviations were specially treated, considering the letters followed by a period as a single token. Some special symbols were replaced by their name to avoid tagging mistakes. To perform all corrections, it was created a python code that was applied to all documents in the *corpus*. After the adjustments all sentences were placed in a single line in each document.

During the adjustment phase it was observed that some files have errors that were beyond repair, such as, words broken by spaces and unreadable files. Therefore, some documents were discarded in order to not generate problems during the execution of the program. Afterwards, the *corpus* was analyzed to measure the number of types and tokens and whether, in its final state, it was suitable for the annotation process.

After the pre-processing phase was executed, all the documents were transformed into a plain UTF-8 file. Each cleaned file consisted of a set of sentences, each sentence occupying a single line. The sentences are formed by a set of tokens together with their respective characteristics, as can be seen in the Figure 2. A second type of file was also created, but only made up of plain unlabeled sentences. These files were designed for use in future annotations tasks.

Figure 2 – Sentence formatting. It shows the sentence “I think, therefore I am” with the annotations after the pre-processing. The captions are shown at the bottom of the figure



I think, therefore I am.

```
<s><tok affix=“ ” base=“i” msd=“PRP”>I</tok>
<tok base=“think” msd=“VB”>think</tok>
<tok base=“,” msd=“,”>,</tok>
<tok base=“therefore” msd=“RB”>thefore</tok>
<tok base=“i” msd=“PRP”>I</tok>
<tok base=“be” msd=“VBP”>am</tok>
<tok base=“.” msd=“.”>.</tok></s>
```

`<s></s>` : Sentence Boundary
`<tok></tok>` : Token Boundary
`affix, base` : Morphology
`msd` : Part of Speech

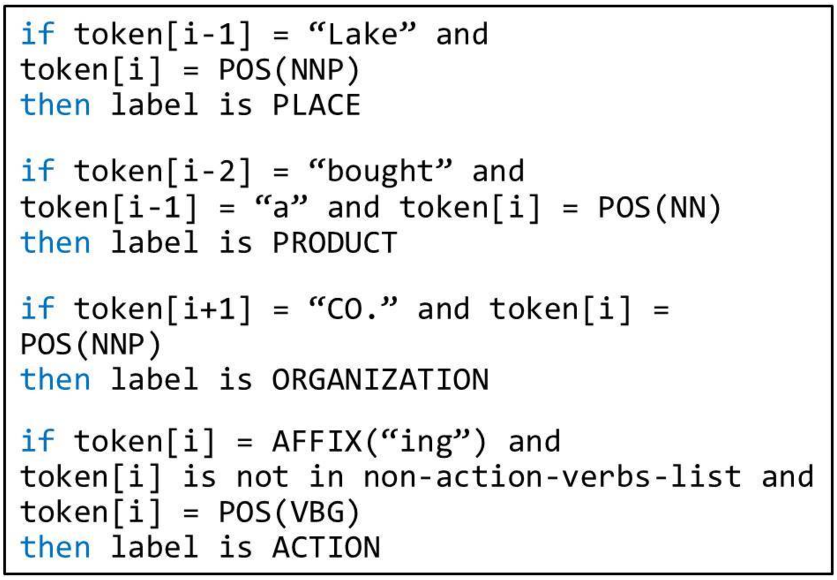
Source: The authors.

It was noticed the need to also annotate the named entities to assist in the process of identifying the associated ontological classes. To annotate named entities it was used the Stanford Named Entity Recognizer (SNER) (FINKEL et al., 2005). The SNER annotates named entities in a given text in one of three following categories: person, location, or organization. The tool uses an automatic annotation approach based on Conditional Random Field (CRF) (FINKEL et al., 2005). The annotations made by the SNER were used as features for the rules of our annotator.

With the aim of facilitating the processing of the *corpus* by the annotator the documents were transformed into a dictionary structure containing a list of sentences. Each sentence composing a set of tokens, and each token formed by the word and the set of linguistic characteristics of that word.

The annotator proposed by us uses a set of rules to produce the annotations. The rules rely upon a set of features related to each token, such as, syntactic classification, morphology, neighboring words, sentence positioning, named entity tags, lists of occurrences, words already tagged among others. The set of rules checks the token features related to the selected 8 top-level classes of the ontology and assigns a label to the token as shown in Figure 3. The figure shows some examples of rules that illustrate the process of labeling. The token and its characteristics are analyzed and if some rule of the class is satisfied the token receives the tag.

Figure 3 – Rules examples



```
if token[i-1] = “Lake” and
token[i] = POS(NNP)
then label is PLACE

if token[i-2] = “bought” and
token[i-1] = “a” and token[i] = POS(NN)
then label is PRODUCT

if token[i+1] = “CO.” and token[i] =
POS(NNP)
then label is ORGANIZATION

if token[i] = AFFIX(“ing”) and
token[i] is not in non-action-verbs-list and
token[i] = POS(VBG)
then label is ACTION
```

Source: The authors.

The first phase of the annotation process was to pass the *corpus* through annotation rules. The rule-based annotator scanned the entire *corpus*, analyzing words and expressions and assigning labels to them when the criteria expressed by some rule was met. The application was designed to assign only one label per word making switches of labels during the annotation process. Afterwards, statistical analyses were carried out to check the behavior and accuracy of the annotator. The results are shown in the next session.

The second phase of the annotation was done manually to fulfill two purposes: correcting possible spurious annotations and to annotate terms that were not tagged by the rule-based annotator. A portion of the *corpus* documents was selected to be read and analyzed manually. This sub-*corpus* was composed of the various literary genres that make up the entire *corpus*, totaling 425 documents, about 5% of each genre. Each sub-*corpus* document was read, annotated and corrected manually. All the adjustments made to a single document manually were applied to the whole *corpus*. An implementation analyzes each manually annotated document and promotes changes in all files. At the end of this process it was guaranteed that all documents were analyzed and verified.

At the end of this phase we got a verified tagged *corpus* according to the Schema.org top-level categories. We also added the changes to the original *corpus* in the standard format. The statistical data about the second phase and the final results from the annotation will be further detailed in the next session.

4 RESULTS AND DISCUSSION

After the entire annotation and verification process, some statistical analyzes were performed on the resulting *corpus*. These results can be seen in the Table 1. The number of documents, sentences, tokens and types was reduced in relation to the original *corpus* due to the corrections made. Lexical diversity is the measure of the lexical richness of the *corpus*, that is, the number of distinct words. The value is given by dividing the number of types by the number of tokens.

Table 1 – Tagged *corpus* in Numbers

Files	8,740
Sentences	732,816
Tokens	16,280,694
Types	195,050
Lexical Diversity	0.012

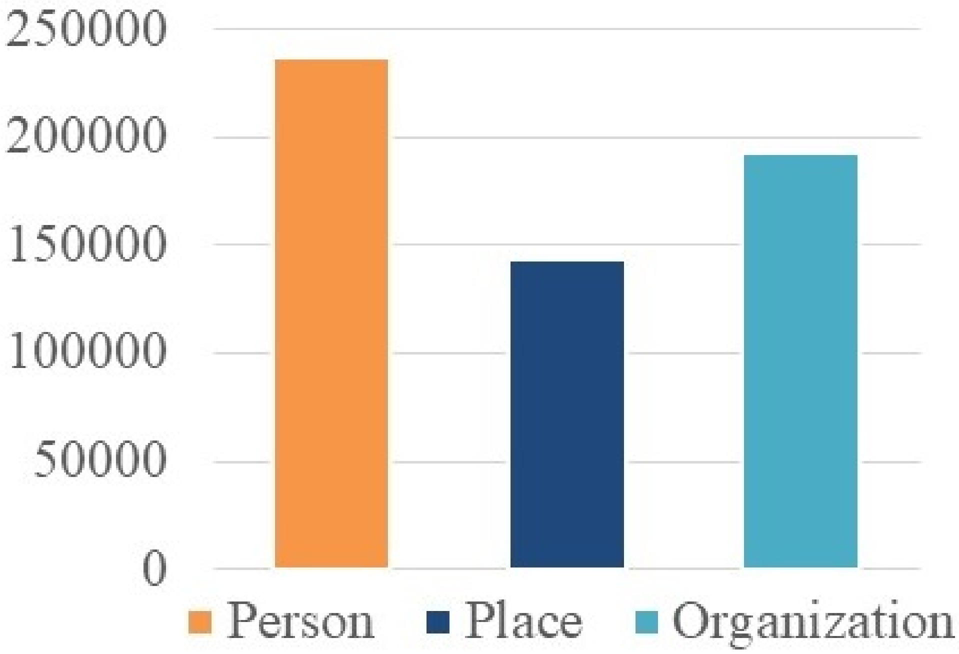
Source: The authors.

As mentioned in the previous section, the SNER tool was applied to the *corpus* documents in order to identify named entities. The number of tokens annotated by the tool is shown in the Figure 4, separated in three categories: Persons, places, and organizations. The pre-annotation was done to supply features for the rules of the later stage.

The rule-based annotator was responsible for assigning labels, according to the set of rules for each category. For each of the eight categories there were created a set of rules that tries to match with features of the word, of its neighborhood and grammar class. Each word is labeled as <*category*> if the annotator identifies the class, or <*o*> otherwise. At the end, the rule-based annotator tagged a total of 1,010,312 tokens distributed according to the first columns of the Figure 5.

At the end of the first annotation phase, using the rule-based annotator, the need to make some corrections and to re-annotate unmarked terms was noticed. The manual annotation of the entire *corpus* would be unfeasible due to the lack of time and human resources. Instead of re-annotating the entire *corpus*, a document category selection approach was used. The selected documents were annotated manually and the markings were applied to all the documents of the

Figure 4 – Number of items annotated by SNER

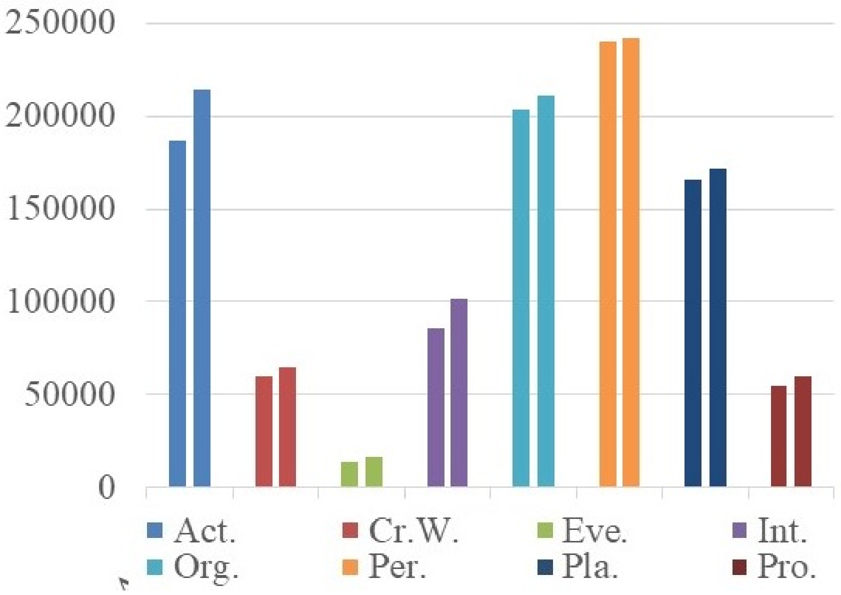


Source: The authors.

corpus through a tool created for this purpose. This approach ensured that all documents could have been indirectly revised. The results of the second phase demonstrated the need for the *corpus* annotation, and ensured that other tokens, previously not annotated, could be tagged. The improvements granted by this phase can be seen in the second columns of Figure 5. Its possible to see that occurred improvements in all categories.

It was noticed that the SNER tool was not sufficient to annotate all tokens belonging to the three categories. The number of Persons, places and organizations was superior to the results presented in the pre-annotation. The SNER also misclassified some terms that were corrected on the second phase. The combination of our proposal with the tool produced better results, justifying the necessity of not using SNER alone for the annotation.

Figure 5 – Rule-based and Manual Annotation. Act=action, cr. W.=creative work, Eve=event, Int=intangible, Org=organization, Per=person,Pla=place, and Pro=product



Source: The authors.

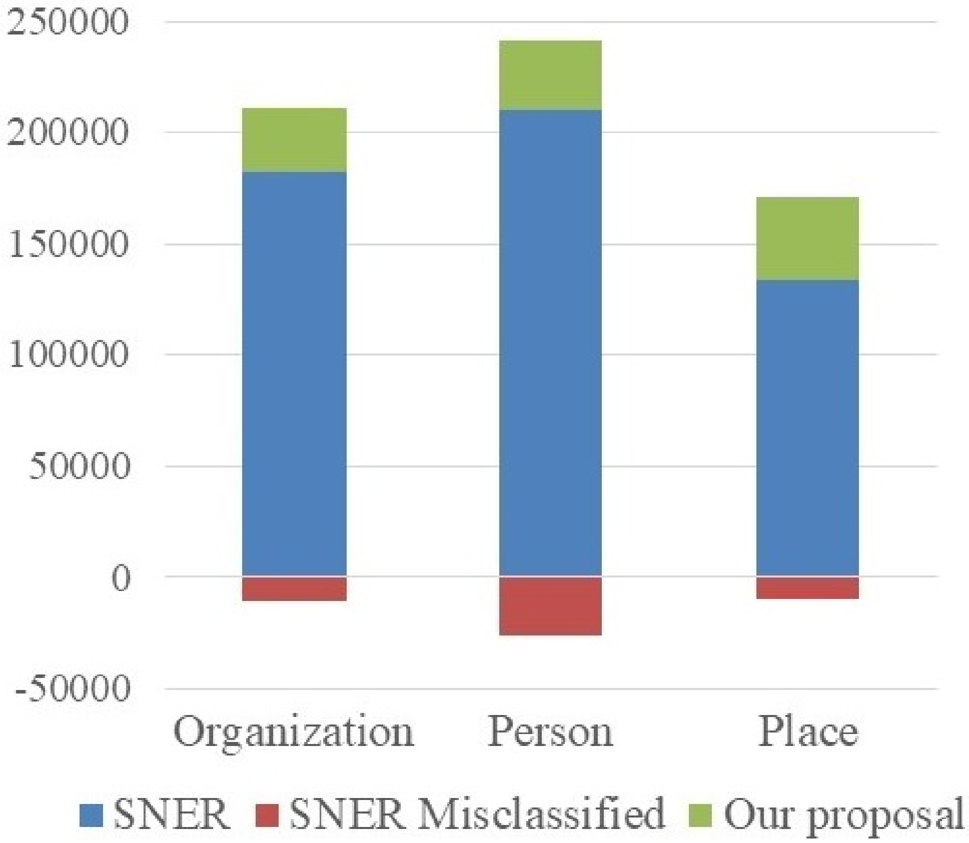
It is important to emphasize that the most important contribution of this research is the complete annotation of the OANC documents. The OANC documents received at the end of this research a semantic annotation referring to the eight top-level classes of the Schema.org ontology. At the end, 1,080,464 types were annotated, distributed according to the second columns of Figure 5. The Figure 6 shows the total number of organizations, Persons and places accomplished by SNER, in blue, and our annotation, in green. The real value of each class is represented by the sum of colors blue and green. The values below 0, in red, shows the number of tokens misclassified by the SNER. The *corpus* can be used for many activities related to natural language processing, attesting the importance of the contribution of this research.

5 CONCLUSIONS

Semantic annotation of lexical items is an important tool to help computational devices understand natural language texts. However, there is a shortage of annotated *corpus* both to function as gold *corpus* as well as to train annotators based on machine learning. This is especially true in the case of annotations based on ontological categories that play an important role in determining the context of a statement. Nonetheless, building a *corpus* manually, which is the traditional form of *corpus* construction, is very time consuming and expensive, and is often not feasible for the construction of large corpus aimed at machine learning.

In this paper, we have described the process of semantic annotation of an English *corpus* using as elements for annotation ontological categories. The semantic annotation of the *corpus* used a hybrid approach: automatic rule-based, and human inspection. The *corpus* produced,

Figure 6 – Organization, Person and Place annotation



Source: The authors.

the main contribution of the research, is available at <https://goo.gl/5AXD8n>, and can be used to train annotators based on machine learning techniques.

Being a rule-based, manually verified approach, it turns out to be a customized approach, making it difficult to compare its accuracy with other methods. Each corpus and domain may require their own custom rules and experts. Therefore, the main contribution of the proposal, in addition to the corpus itself, is in the methodological aspect, allowing for a faster generation of corpora. The next step is to use the *corpus* as a measurement parameter and to aid in the creation of other semantic annotators.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and also by the funding agencies FAPEMIG and CNPq.

REFERENCES

ALEC, Céline; REYNAUD-DELAÎTRE, Chantal; SAFAR, Brigitte. An ontology-driven approach for semantic annotation of documents with specific concepts. In: SPRINGER. **International Semantic Web Conference**. Heidelberg, 2016. p. 609–624.

ANDRADE, Guidson Coelho; OLIVEIRA, Alcione; MOREIRA, Alexandra. A rule-based semantic annotator: Adding top-level ontology tags. In: SBC. **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology**. Uberlândia, 2017. p. 53–62.

COHEN, K Bretonnel et al. The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. In: **Handbook of Linguistic Annotation**. Dordrecht: Springer, 2017. p. 1379–1394.

ERDMANN, Michael et al. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In: ACL. **Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content**: Association for computational linguistics. Luxembourg, 2000. p. 79–85.

FINKEL, Jenny Rose; GRENAGER, Trond; MANNING, Christopher. Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL. **Proceedings of the 43rd annual meeting on association for computational linguistics**. Ann Arbor, 2005. p. 363–370.

GRUBER, Tom. What is an ontology. **Knowledge Systems Laboratory**, 1993. Disponível em: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>. Acesso em: 25 oct. 2017.

GUARINO, Nicola. Formal ontology and information systems. In: **Proceedings of FOIS**. Trento: IOS Press, 1998. v. 98, p. 81–97.

GUHA, Ramanathan V; BRICKLEY, Dan; MACBETH, Steve. Schema. org: Evolution of structured data on the web. **Communications of the ACM**, ACM, v. 59, n. 2, p. 44–51, 2016.

IDE, Nancy; SUDERMAN, Keith. The american national corpus first release. In: CITESEER. **LREC**. Genoa, 2004.

KHALILI, Ali; AUER, Sören. Wysiwym–integrated visualization, exploration and authoring of semantically enriched un-structured content. **Semantic Web**, IOS Press, v. 6, n. 3, p. 259–275, 2015.

KIRYAKOV, Atanas et al. Semantic annotation, indexing, and retrieval. **Web Semantics: Science, Services and Agents on the World Wide Web**, Elsevier, v. 2, n. 1, p. 49–79, 2004.

KLIEN, Eva. A rule-based strategy for the semantic annotation of geodata. **Transactions in GIS**, Wiley Online Library, v. 11, n. 3, p. 437–452, 2007.

LEECH, Geoffrey N. Introducing corpus annotation. In: GARSIDE, Roger; LEECH, Geoffrey N; MCENERY, Tony (Ed.). **Corpus annotation: linguistic information from computer text corpora**. 1. ed. London, UK: Longman, 1997. cap. 1, p. 1–18.

LU, Xiaofei. **Computational methods for corpus annotation and analysis**. 1. ed. New York, NY: Springer, 2014.

PATEL-SCHNEIDER, Peter F. Analyzing schema. org. In: SPRINGER. **International Semantic Web Conference**. Cham, 2014. p. 261–276.

PUSTEJOVSKY, James; STUBBS, James Amber. **Natural Language Annotation for Machine Learning: A guide to corpus-building for applications**. 1. ed. Sebastopol, CA: O'Reilly Media, Inc., 2012.

RONALLO, Jason. Html5 microdata and schema. org. **Code4Lib Journal**, v. 16, 2012.

ŞIMŞEK, Umutcan et al. Domain specific semantic validation of schema. org annotations. In: SPRINGER. **International Andrei Ershov Memorial Conference on Perspectives of System Informatics**. Cham, 2017. p. 417–429.

WILSON, Andrew; THOMAS, Jenny. Semantic annotation. In: GARSIDE, Roger; LEECH, Geoffrey N; MCENERY, Tony (Ed.). **Corpus annotation: linguistic information from computer text corpora**. 1. ed. London, UK: Longman, 1997. cap. 1, p. 53–65.