GUIDSON COELHO DE ANDRADE

SEMANTIC ENRICHMENT OF AMERICAN ENGLISH *CORPORA* THROUGH AUTOMATIC SEMANTIC ANNOTATION BASED ON TOP-LEVEL ONTOLOGIES USING THE CRF CLASSIFICATION MODEL

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
A553s
2018

Andrade, Guidson Coelho de, 1992-
Semantic enrichment of American English c*orpora* through automatic semantic annotation based on top-level ontologies using the CRF classification model / Guidson Coelho de Andrade. – Viçosa, MG, 2018.
xiii, 74 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Inclui apêndices.
Orientador: Alcione de Paiva Oliveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 58-62.

1. Banco de dados. 2. Linguagem de programação (Computadores) - Sintaxe. 3. Semântica. 4. Ontologia. 5. Computação semântica. 6. Processamento de linguagem natural (Computação) . I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. II. Título.

CDD 22.ed. 005.75

GUIDSON COELHO DE ANDRADE

SEMANTIC ENRICHMENT OF AMERICAN ENGLISH CORPORA THROUGH AUTOMATIC SEMANTIC ANNOTATION BASED ON TOP-LEVEL ONTOLOGIES USING THE CRF CLASSIFICATION MODEL

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae.*

APROVADA: 26 de abril de 2018.

Jugurta Lisboa Filho

Ely Edilson da Silva Matos

Alcione de Paiva Oliveira
(Orientador)

I dedicate this work to my beloved family, Célio, Odete e Stéfane and my sweetheart Igor.

Acknowledgments

First of all, I want to thank my advisor, Professor Alcione de Paiva Oliveira, for his availability and for the accompaniment he carried out during this work. He was of great importance during the execution of this project and my training in the master's degree program being an example to be followed.

To my co-advisor, Alexandra Moreira for her receptivity, for the patient work reviewing the writing of articles and the dissertation and her infinite solicitude.

To Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the essential financial support, allowing me to devote myself fully to this work.

To Universidade Federal de Viçosa for the welcoming and beautiful environment, for providing such great intellectual knowledge and for providing me with unforgettable moments. Studying here was a dream come true.

To all the professors of the UFV's Department of Informatics that somehow contributed to my intellectual and professional growth. I could not have done it is wasn't for their help and teaching.

To my parents Célio and Odete for all financial support and for being my greatest example, giving up pleasures and comfort for better education for their children. These people are the foundation of my life and their advice, caring and example are the fuels for my dreams.

To my sister, Stéfane, for always being ready to help me and for following me since I was little. She, with her sensitivity, realized my needs and, unlike many, instead of giving me advice to do what they think is right, she supported me to do what I want and believe.

To my boyfriend, Igor, for the immense love, patience, and encouragement shown during the master's degree. Even in the most difficult times when the time to repay his love was scarce, he was by my side. Without him, I would not have made it this far. Every day, helping me to be a better person, and when I am not, he is ready to be my safe haven.

To my friends from UFV, Aly, Bárbara, Carol, César, Dâmaris, Daniel, Fábio,

Joana, Liliane, Paulo, Rubens and Vinício, companions of work and brothers in friendship that were part of my formation and will continue to be present in my life. These friends provided me with memorable moments, in parties, the peteca group and in our “rolês” of fifth grade.

To my life friends. I am not going to name them, so I do not commit injustices, but true ones will know who they are. You are the best and at all times have been by my side, supporting me, making me laugh, and getting up when I fell.

To my former undergraduate instructors, who have always encouraged me and believed in my potential.

To all of you and also to those who I may have forgotten to mention the name, because they directly or indirectly made me grow up and get here.

Finally, I direct my thanks to God, who empowers and sustains me with each new day. His providence gives me inner strength to overcome difficulties. To Him all honor, glory and praise.

Contents

List of Figures vii

List of Tables viii

List of Abbreviations and Acronyms ix

Abstract x

Resumo xii

1 Introduction 1

1.1 The problem and its importance . . . 3
1.2 Hypothesis . . . 5
1.3 Objective . . . 5
1.4 Dissertation Structure . . . 5

2 A Rule-based Semantic Annotator: Adding top-level ontology Tags 9

2.1 Introduction . . . 9
2.2 Related work . . . 11
2.3 Materials and Methods . . . 12
2.4 Results . . . 16
2.5 Conclusions . . . 19

3 Hybrid Semantic Annotation: Rule-based and Manual Annotation of the Open American National *Corpus* with a Top-Level Ontology 20

3.1 Introduction . . . 21
3.2 Related work . . . 23
3.3 Materials and Methods . . . 24
3.4 Results and Discussion . . . 29

3.5 Conclusions 32

4 CRF model applied to semantic annotation of top-level ontologies of Schema.org using an American English *corpus* 33

4.1 Introduction 34
4.2 Related Works 37
4.3 Materials and Methods 41
4.3.1 Schema.org Ontology 41
4.3.2 *Corpus* 44
4.3.3 CRF Approach 45
4.4 Results and Discussion 48
4.5 Conclusions 51

5 General Conclusions and Future Works 54

Bibliography 58

Appendix A Hybrid annotation proposal 63

Appendix B Rules 65

B.1 Action 67
B.2 Creative Work 68
B.3 Event 69
B.4 Intangible 70
B.5 Organization 71
B.6 Person 72
B.7 Place 73
B.8 Product 74

List of Figures

2.1 Top layers of the SUMO Ontology 12
2.2 Rules examples 13
2.3 Sentence previous annotation sample. 14
2.4 Dictionary structure. 15
2.5 Sentence annotated with the ontological categories. 16

3.1 Schema.org Sunburst Graphic 26
3.2 Sentence formatting 28
3.3 Rules examples 29
3.4 Number of items annotated by SNER. 30
3.5 Rule-based and Manual Annotation 31
3.6 Organization, People and Place annotation 31

4.1 F1-score progression from the Schema.org top-level classes 50

A.1 Development of a rule-based semantic annotator using the SUMO ontology. 63
A.2 Annotation of the Open American National Corpus using the hybrid proposal. 64

B.1 Initial code 66
B.2 Action Rules 67
B.3 Creative Work Rules 68
B.4 Event Rules 69
B.5 Intangible Rules 70
B.6 Organization Rules 71
B.7 Person Rules 72
B.8 Place Rules 73
B.9 Product Rules 74

List of Tables

2.1 Confusion matrix 17

3.1 Tagged *corpus* in Numbers. 29

4.1 Results from the pattern execution of CRF 49
4.2 Most prominent features of every Schema.org top-level class. 49
4.3 Results with use of hyperparamter optimization and cross validation on the CRF 51

List of Abbreviations and Acronyms

CAPES:	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq:	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CRF:	Conditional Random Fields
FAPEMIG:	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
ISO:	International Organization for Standardization
L-BFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
LAF:	Linguistic Annotation Format
LOC:	Location
MISC:	Miscellaneous
NER:	Named entity recognition
NLP:	Natural Language Processing
O:	Other
OANC:	Open American National *Corpus*
ORG:	Organization
PER	Person
POS:	Part of speech
RDF:	Resource Description Framework
SNER	Stanford named entity recognition
SUMO:	Suggested Upper Merged Ontology
UFV:	Universidade Federal de Viçosa
WYSIWYW:	What You See Is What You Mean
XML:	Extensible Markup Language

Abstract

Andrade, Guidson Coelho de, M.Sc., Universidade Federal de Viçosa, April, 2018. **Semantic enrichment of American English *corpora* through automatic semantic annotation based on top-level ontologies using the CRF classification model**. Advisor: Alcione de Paiva Oliveira. Co-advisor: Alexandra Moreira.

Textual databases carry with them human-perceived meanings, but those meanings are difficult to be interpreted by computers. In order for the machines to understand the semantics attached to texts, and not only their syntax, it is necessary to add extra information to these *corpora*. Semantic annotation is the task of incorporating this information by adding metadata to lexical items. This information can be ontological concepts that help define the nature of the word in order to give it some meaning. However, annotating texts according to an ontology is still a task that requires time and effort from annotators trained for this purpose. Another approach to be considered is the use of automatic semantic annotation tools that use machine learning techniques to classify annotated terms. This approach demands a database for training the algorithms that in this case are *corpora* pre-annotated according to the semantic dimension to be explored. However, this methodological lineage has limited resources to meet the needs of learning methods. There is a large lack of semantically annotated *corpora* and an even larger absence of ontologically annotated *corpora*, hindering the advance of the area of automatic semantic annotation. The purpose of the present work is to assist in the semantic enrichment of American English texts by automatically annotating them based on top-level ontology through the Conditional Random Fields (CRF) supervised learning model. After the selection of the Open American National *Corpus* as a linguistic database and Schema.org as an ontology, the work had its structure divided into two stages. First, the pre-processed and corrected *corpus* was submitted to a hybrid annotation, with

a rule-based annotator, and later manually. Both annotation tasks were driven by the concepts and definitions of the eight classes from the top-level of the selected ontology. Once the *corpus* was written ontologically, the automatic annotation process was started using the CRF learning method. The prediction model took into account the linguistic and structural features of the terms to classify them under the eight ontological types. The results obtained during the evaluation of the model were very satisfactory and reached the objective of the research. The work, although it is a new approach of semantic annotation and with little margin of comparison, presented promising results for the advance of the research in the area of automatic semantic enrichment based on top-level ontologies.

Resumo

Andrade, Guidson Coelho de, M.Sc., Universidade Federal de Viçosa, abril de 2018. **Enriquecimento semântico de *corpora* do Inglês americano através de anotação semântica automática baseada em ontologias de nível topo utilizando o modelo de classificação CRF**. Orientador: Alcione de Paiva Oliveira. Coorientadora: Alexandra Moreira.

O significado de bases de dados textuais é de fácil percepção para as pessoas, mas de difícil interpretação por parte dos computadores. Para que as máquinas possam compreender a semântica associada aos textos e não somente a sintaxe, é necessário a adição de informações extras a esses *corpora*. A anotação semântica é a tarefa que incorpora essas informações por meio da adição de metadados aos itens lexicais. Essas informações podem ser conceitos ontológicos que ajudam a definir a natureza da palavra a fim de atribuir-lhe algum significado. No entanto, anotar textos segundo uma determinada ontologia ainda é uma tarefa que demanda tempo e esforço de anotadores treinados para esse fim. Outra abordagem a ser considerada é o desenvolvimento de ferramentas de anotação semântica automática que utilizem técnicas de aprendizado de máquina para classificar os termos anotados. Essa abordagem demanda uma base de dados para treinamento dos algoritmos que nesse caso são *corpora* pré-anotados segundo a dimensão semântica a ser explorada. Entretanto, essa linhagem metodológica dispõe de recursos limitados para suprir as necessidades dos métodos de aprendizado. Existe uma grande carência de *corpora* anotados semanticamente e, particularmente, uma ausência ainda maior de *corpora* ontologicamente anotados, dificultando o avanço da área de anotação semântica automática. O objetivo do presente trabalho é auxiliar no enriquecimento semântico de textos do Inglês americano, anotando-os de forma automática baseando-se em ontologia de nível topo através do modelo de aprendizagem supervisionada Conditional Random Fields (CRF). Após a seleção do Open American National *Corpus*

como base de dados linguística e da Schema.org como ontologia, o trabalho teve sua estrutura dividida em duas etapas. Primeiramente, o *corpus* pré-processado e corrigido foi submetido a uma anotação híbrida, com um anotador baseado em regras e, posteriormente, uma anotação complementar manual. Ambas as tarefas de anotação foram dirigidas pelos conceitos e definições das oito classes provenientes do nível topo da ontologia selecionada. De posse do *corpus* anotado ontologicamente, iniciou-se o processo de anotação automática via uso do método de aprendizagem CRF. O modelo de predição levou em consideração as características linguísticas e estruturais dos termos para classificá-los sob os oito tipos ontológicos. Os resultados obtidos durante a avaliação do modelo foram muito satisfatórios e atingiram o objetivo da pesquisa. O trabalho, embora seja uma nova abordagem de anotação semântica e com pouca margem de comparação, apresentou resultados promissores para o avanço da pesquisa na área de enriquecimento semântico automático baseado em ontologias de nível topo.

Chapter 1

Introduction

The advent of the Internet, added to the possibility of any connected person providing information voluntarily and / or involuntarily, led to the generation of an immense amount of information in natural language stored on digital format. However, this information is of little use if it can not be interpreted by computational devices. For this, it is necessary to add, in some way, semantic information to the textual elements. One of the undertakings in this sense was called the Semantic Web. Formally, the Semantic Web is a web environment where the information has been organized in a way that not only human beings can understand it, but machines as well (Berners-Lee et al., 2001). To achieve this level of performance it is necessary that the web content be semantically enriched so that the machines are able to manipulate the data in such a way to produce inferences. The semantic enrichment of content presented in the web environment is the main subsidy for the development of computer understanding.

The composition of texts and documents made available digitally are, most of the times, arranged for only the human interpretation. The natural language present in these texts, regardless of idiom, does not consist only of a set of structured words organized syntactically. Humans are apt to understand such texts because they know the intrinsic meaning in that sequence of characters, in other words, they are able to abstract the semantics contained therein. On the other hand, computers and applications do not have such capability and could hardly interpret documents consisting of simple sets of sentences. To assist in this task, computational techniques use metadata capable of assigning additional information to the terms. Although it seems a simple task from the human point of view, producing texts propitious to be fully understood by computers is still a difficult challenge.

Any document originated from a natural language and prepared to be manip-

ulated by computational techniques is called *corpus*. A *corpus* can be defined as a set that the data is based on oral content or natural language texts, which can be interpreted by the computer and able to provide a sample of the linguistic variation present in a domain (Leech, 1997). The content of a *corpus* should represent the linguistic aspect of the studied phenomenon to obtain better performance from the use of computational techniques (Sardinha, 2000). To improve this performance, as previously mentioned, it is used the addition of metadata to the content of the *corpus* in order to add more information to it.

In the field of Natural Language Processing (PLN), science which studies linguistic phenomena linked to computation, the practice of incorporating extra information into texts is called *corpus* annotation. According to Pustejovsky and Stubbs (2012) annotation refers to the process of adding metadata in a given text in order to give a computer the ability to manipulate the text using techniques and algorithms of the area. *Corpus* annotation has the main purpose of offering computers extra information so that it is possible to increase the performance of data processing. The annotation can cover several spheres of linguistics, being able to incorporate syntactic, morphological, semantic, structural value to the annotated object. To contribute in the aggregation of meaning to the annotated text, the dimension to be explored by the annotation should focus mainly on semantic aspects.

Semantics is the field of linguistics that studies the meaning of words, as well as the interpretation of this meaning, distinguishing also the difference of meaning that words present in different contexts and over time. In this sense, the type of annotation that evidences the meaning included in the term is the semantic annotation. The semantic annotation process is responsible for assigning meaning to the annotated content, through labels that express concepts and definitions inherent to the term (Kiryakov et al., 2004). Many elements can be used as marking labels during the process of annotation, among them stands out the use of ontologies.

Ontologies are, in the words of Gruber (1993), a "specification of a conceptualization". More clearly and paraphrasing, ontologies is a description of concepts of all things existing within a domain and the relationship between them (Guarino, 1998). In this sense, entering deeper into this concept, ontologies are responsible for classifying things present in some context, establishing types, definitions and relations between these types. Thus, from the perspective of semantic annotation, ontologies play a primordial role in the process of defining tags for marking, because they already express a conceptualization required by the annotation task.

The ontologies expand by several domains and their classification is given by the scope related to this domain. Ontological types can describe very specific do-

mains of an area as well as general domains that attempt to encompass all things existing in the world (Guarino, 1998). Ontologies of general domain, usually define types that are primarily more generic and go deeper into more specific concepts. The more generic types of this ontological genre guide the structure of the ontology and are defined as the top-level (Guarino, 1998). Ontologies of the top-level express what exists of more generic in the general domain and provides classes able to integrate a larger number of elements.

Taking into account the factors related to top-level ontologies, they can guide the annotation process by transforming into the marking labels of terms. In this sense, the types that came from a top-level ontology lead to annotation according to their definitions and concepts. Once using top-level ontologies, the annotated terms could express more than just a set of letters, they would convey the meaning inherent to it by the class to which it was tagged. Although it seems to be a simple task to execute, semantic annotation based on top-level ontologies faces difficult challenges.

1.1 The problem and its importance

As previously highlighted, computers have limited resources in the area of interpretation of natural languange and there are still many problems that surround the theme and make it difficult for semantic annotation to aid in this process. First of all, semantic annotation is a task that can be performed in two ways, manually and automatically by the use of computational techniques. Both methods face difficulties in the implementation process and limit the development of the area. In this subsection it will describe the problems faced by the area as well as an explanation of why the use of top-level ontologies is not yet exploited for automatic semantic enrichment of texts.

The semantic annotation process, which is responsible for assigning some meaning to the content, is still commonly performed manually. Annotators, who must have some experience in the field, are responsible for breaking down a textual base into smaller segments and identifying specific entities, events, attributes, or facts mentioned in the documents, and assigning to them a semantic label. However, because this activity is closely related to the human factor, it is subject to errors and failures. Different annotators may have different interpretations of a content, generating inconsistency in the annotation, or even omissions. Another problem confronted by the annotators is time, since such a task requires substantial

effort to be fully completed.

Computational tools helps to save time and people to perform the semantic annotation and execute it more accurately, so those tools could contribute substantially to increase the volume of digital documents annotated semantically. Usually computational resources used to annotate documents adopt machine learning methodologies. This approach, although it seems feasible to be implemented, addresses the problem of lack of subsidy for its practice. As it is well known to all, machine learning methods rely on already classified database to train their algorithms and subsequently label an unclassified data set. The adversity that inhibits the progress of automatic methods based on machine learning is exactly the lack of resources to train classifier algorithms.

To train learning algorithms, previously annotated *corpus* must be used, containing information necessary for the intended application. For machine learning methods to perform efficiently it is necessary to meet two requirements related to the *corpus* used as training database. First, the *corpus* must be accurately annotated to ensure a database free of errors or with as few errors as possible. Secondly, the size of the *corpus* should be representative enough and capable of expressing the linguistic variance of the domain to be classified. Both problems also characterize automatic semantic annotation through learning techniques and they are responsible for making it difficult to advance.

Considering the automatic semantic annotation, the *corpus* used to train learning algorithms should be annotated taking into account some semantic aspects. However, there is a huge shortage of resources to support learning techniques for semantic annotation. Another factor of greater magnitude that further aggravates the problem is the lack of a *corpus* annotated according to a specific ontology. After a search for *corpus* annotated according to a top-level ontology, none were found that presented sufficient characteristics to apply it in automatic learning methods. The annotation based on top-level ontologies, although not much explored to create *corpora*, could provide, under some level of abstraction, meaningful content to be used as an algorithm training database. In this sense, all the problematic that involves the automatic semantic annotation turns mainly around the lack of resources to supply the need of the methods of machine learning. In addition, the non-exploitation of benefits derived from ontologies in the semantic annotation process leads to the lack of *corpora* enriched ontologically for the development of future applications and researches.

Both the problem related to human resources and the lack of *corpus* to train learning algorithms make it difficult to develop semantically enriched content to

furnish the need of the Web. The Web to reach the level proposed by the Semantic Web still needs to notably enrich its content. The internet will be only prescribed as Semantic Web from the moment a computer can read a block of semantically enriched digital information and from inferences with other blocks generate a greater knowledge. It is a fact that this reality is still far from being fully achieved, but research and work related to the subject can contribute strongly to its development. Semantically enriched content and tools capable of performing the semantic annotation process are important initiatives to contribute to the progress of the area. The use of top-level ontologies, although it is a small step towards semantic enrichment, may be the bottom line contribution to the expansion of the search field.

1.2 Hypothesis

The hypothesis that was tested in this research was the possibility of performing automatic semantic annotation based on concepts of top-level ontologies on American English documents, using as the training object of the supervised learning model a previously annotated *corpus* by the use of hybrid approaches of semantic annotation.

1.3 Objective

The general objective of this research is to use a classification model built from a supervised machine learning technique in order to perform automatic semantic annotator based on top-level ontologies to assist semantic enrichment of American English documents. Specifically, the objectives of this research are:

- Develop a rule-based semantic annotator using the chosen top-level ontology definitions to specify the assignment rules.
- Annotate the chosen *corpus* in a hybrid way (rule-based and manually) using the annotator created in the previous objective.
- Use the CRF supervised learning method to construct an automatic semantic annotation model based on top-level ontologies.

1.4 Dissertation Structure

This dissertation was elaborated and written according to one of the formats recommended by the Commission of the Graduate Program in Computer Science of

the Federal University Viçosa. The body of the text is organized as a collection of articles resulting from the development of the research. This first chapter presented a general introduction of the problem to be treated in this dissertation, as well as its hypothesis and general and specific objectives. It is important to mention that the introduction presents only some concepts that will be better explained during the reading of the articles. The following chapters, composed of three articles, published, submitted and about to be submitted, respectively, fully illustrate all the work developed in the dissertation.

The dissertation is structured as follows:

Chapter 2 entitled "A Rule-based Semantic Annotator: Adding top-level ontology Tags" presents the article published on the proceedings of the XI Brazilian Symposium in Information and Human Language Technology (STIL-2017). The article proposes a method of semantic annotation based on the use of rules to capture elements that correspond to the selected top-level ontology. Because it is an annotation proposal, the tests performed are significantly simple in order to validate only the rule-based annotator. The results obtained and presented in the article describe the execution of the rules-based annotator under a sub-*corpus* originating from the main *corpus*.

Because the article has page limitation, the model of the proposal expressed in the form of development flow has been removed from the article. The figures that illustrate this flow can be seen in the Appendix A, as well as a brief explanation about them. It is also important to point out that the ontology selected in this phase of the research does not correspond to the one used in the next stages of the study. Technical issues explored in the next paragraph, and further detailed in the next chapter, justify the circumstances responsible for the change. In this way the main intention of the article is to legitimize and prove the importance of semantic annotation based on top-level ontologies, besides introducing this new topic to the scientific community. The full reference to the article can be visualized and quoted as follows:

ANDRADE, Guidson Coelho; OLIVEIRA, Alcione de Paiva; MOREIRA, Alexandra. A Rule-based Semantic Annotator: Adding top-level ontology Tags. In: Brazilian Symposium in Information and Human Language Technology, 11, Uberlândia. **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology and Colocated Events.**. Uberlândia: [s.n.], 2017. p. 53-62.

The second article already submitted to Language Resources and Evaluation Journal corresponds to the content of Chapter 3, named as "Hybrid Semantic Annotation: Rule-based and Manual Annotation of the Open American National *Corpus* with a Top-Level Ontology". The article describes the entire process proposed in the previous article applied effectively to the main *corpus*, but with a different ontology perspective. The ontological bias sought by the research should aim at sufficiently generic concepts capable of being later used by computational applications in understanding the semantically annotated documents. The ontology previously used in the first article presents very abstract concepts which represented some difficulty in creating semantic elements for the composition of an effective enrichment of the documents. Another factor of great relevance for selection of the ontology is its popularity, applicability and its genuine relation with the content commonly used in the Web. In this sense, from the second article, the ontology used as basis of the work was the Schema.org. In general, the second article describes the effective execution of the proposal reported in the first article, except for some changes during the course of the research.

Still summarizing Chapter 3, its content brings a hybrid semantic annotation based on top-level ontologies of an entire American English *corpus*. Initially in the work, a pre-processing of the *corpus* was executed in order to format it to the assigned standard for the development of future rules. Corrections and formatting imposed on the *corpus* resulted in a standardized database for the application of the rules that were elaborated according to the definitions provided by the top-level of the selected ontology. In the process of annotating the *corpus* before the rule-based annotator is executed, a third-party annotation tool was used with the purpose of assisting in the process and as a confronting aspect between the two approaches. Finally, after applying the rules and with the purpose of correcting and adding new tags to terms not captured by the rules-based annotator, a manual annotation of the *corpus* was manipulated. At the end of the article, as a result of the hybrid annotation approach, the *corpus* was entirely annotated according to the criteria defined for the execution of the next phase of the research.

The set of rules elaborated for the annotation of the terms extracted from the corpus were only exemplified during session of methodology of the article. As a way to further enrich the dissertation, Appendix B presents in more detail a larger set of rules that exemplifies the method used. Although not all the rules used, the appendix helps to illustrate the methodology applied for setting the rule-based annotator.

Chapter 4, which presents the most important element of the research, de-

scribes the article "CRF model applied to semantic annotation of top-level ontologies of Schema.org using an American English *corpus*". The reported work refers to the use of the *corpus* produced in the previous stage as a database to be manipulated by machine learning techniques. As a classification tool, after a state-of-the-art bibliographic survey, the Conditional Random Fields learning model is considered to be the most appropriate one to be used in the research. The method is then applied to different portions of the *corpus*, properly formatted to the recommended standard, in order to obtain the best set of features capable of matching the expected results. The ontological types used to annotate the *corpus*, become the classification labels of the model. The article also describes the execution of two approaches using the classifier model and compares their respective results. Finally, it is described the results obtained from the automatic semantic annotation based on top-level ontologies performed by the model.

Finally, chapter 5 discusses the results obtained in the research in general, narrating some conclusions about the new line of research suggested and introduced by this work. The conclusion still reflects on how the use of top-level ontologies for automatic text enrichment can impact the semantic information attributed to these texts. In this chapter are left some open research opportunities that may provide future works on the field.

For organizational purposes, the references of the articles and other parts of the dissertation were assembled in a single Bibliography section.

Chapter 2

A Rule-based Semantic Annotator: Adding top-level ontology Tags

Abstract

Understanding natural language texts is a simple task for human beings, but despite recent advances, it is still a challenge for computational devices. An important step in allowing machines to understand texts in natural language is to annotate lexemes with semantic information. Semantic information has several levels and aspects, but a type of semantic annotation that has the ability to help determine the context of the statement is the ontological information. However, annotating texts according to an ontology is still a task that requires time and effort from annotators trained for this purpose. The goal of the project is to assist in the semantic enrichment of texts, through a rule-based annotator. Given an entry in the format required by the annotator, the tool returns a document annotated according to the concepts proposed by the SUMO ontology. The project consists in elaborating a semantic annotator based on rules that is able to annotate a *corpus* using the selected top-level ontologies.

2.1 Introduction

Assigning semantic information to lexemes is a task that has made significant progress recently, mainly due to the increase in computational power and to the availability of large linguistic *corpora* to train automatic learning tools. Nowadays, there are commercial devices such as smartphones or Amazon® Echo that are capable of answering questions made in natural language. In order for these devices

to function properly, some semantic information must be attributed to the utterances, even if implicitly through statistical analysis. Another way to aid in the understanding of texts by computer devices is to explicitly add semantics to textual information by annotating lexemes with semantic information.

There are different annotation granularities that range from associating a label to a full text to associating a label with each phrase or even word Leech (1997). Semantic annotation is an annotation that attempts to unveil the meaning of the things being marked (Reeve and Han, 2005). Semantic annotation searches for text elements and classifies them according to their meaning in the fragment in which they are inserted (Mitkov, 2005). Semantics has several levels and aspects, but a type of semantic information that has the ability to help determine the context of a statement is the ontological information. The term ontology can be defined as a specification of a conceptualization (Gruber, 1993). In other words, it is a description of concepts of existing entities in the world and relationships that exist between these entities (Uschold and Gruninger, 1996). Ontology studies the various entities that exist, a description of types and structures of things, their properties, events, relationships and processes throughout the real world (Guarino, 1998). The ontologies are used to classify the objects of some domain, according to some pre-established criteria (Maedche, 2012). The annotation based on ontologies, although not much explored, could provide, under a certain level of abstraction, contextual information to the annotated textual object (Handschuh and Staab, 2003). However, annotating texts according to an ontology is still a task that requires time and effort from annotators trained for this purpose, and it is still commonly elaborated through manual work (Pustejovsky and Stubbs, 2012). The exhaustive work caused by this task is responsible for the lack of ontology annotated *corpora*. A tool capable of semantically annotating texts based on ontology would be very useful and would help increase the availability of *corpora* annotated with this type of information.

The objective of the research presented in this article was to construct a semantic annotator based on rules capable of annotating the terms of a given *corpus* under the concepts of a top-level ontology. It uses the concepts of the chosen ontology level and classifies the terms of the *corpus* to annotate them according to the ontology. The tool developed is domain independent but has been implemented with focus on the English language.

This paper is organized as follows: the next section presents the work previously developed that are related to this research; Section 2.3 describes the materials and methods applied in the research; Section 2.4 presents the results obtained; and

Section 2.5 presents the final remarks.

2.2 Related work

The semantic annotation field is quite active, but most of the work deals with annotation of semantic roles. Here we will discuss some recent work dealing with ontological annotation.

Asooja et al. (2016) developed a system to automatically annotate texts of the regulatory sector for different industries using the semantic frames via FrameNet, which is, in a certain sense, a lexical ontology. The application of the FrameNet lexical base contributed to the increased performance of its results. The system also made use of POS annotation and n-grams. The difference in relation to our work is that we choose to use a formal ontology rather than a lexical ontology (the distinction is that lexical ontology has inspiration in what is enunciated rather than in what exists). Another distinct point of the work of Asooja et al. (2016) is that they used classification by means of statistical techniques while we used rules.

Alec et al. (2016) proposed an ontology-driven approach for semantic annotation of documents from a *corpus* where each document describes an entity of a same domain. The focus of the work was more to annotate documents rather than the words of the documents. In addition, the researches used domain ontologies instead of a top-level ontology.

Moreira et al. (2016) proposed a system that extracts the terms of a text and links them to an ontology (SUMO ontology in that case). The system could be used to annotate the text but was not used for this purpose. In addition, the system analyzes only terms that originate from noun phrases, which is a more limited scope than the current research.

Pham et al. (2016) presented a domain-independent approach to automatic semantic labeling that uses machine learning techniques. Similarly to our proposal the domain-independent feature was the novelty of their approach. Unlike our approach which is a rule-based method, the authors used similarity metrics as features to compare against labeled domain data and learns a matching function to infer the correct semantic labels for data. They also focused on domain ontology rather than on top-level ontology.

2.3 Materials and Methods

The Suggested Upper Merged Ontology (SUMO) Pease et al. (2002) was the top-level ontology chosen for this project. Its choice was based on being an ontology with a certain degree of maturity, with a broad scope and for being well formalized. SUMO was first released in December 2000 and defines a hierarchy of classes, rules, and relationships (Niles and Pease, 2001). It is intended to be an ontology that underpins a variety of computer information processing systems (Pease et al., 2002). Although it is an ontology that addresses some domains, in our work we focus only on top-level concepts, because we believe that this first step is essential for a later annotation focused on a specific domain.

This work was developed using the concepts of the first three levels of SUMO ontology. The top-level of the SUMO ontology contains 12 classes distributed in the three levels, as shown in Figure 2.1. The first level displays the root class named *Entity*. The *Physical* and *Abstract* classes compose the second level. And finally, on the third level of the ontology there are the classes, *Object*, *Process*, *Quantity*, *Attribute*, *Set Or Class*, *Relation*, *Proposition*, *Graph* and *Graph Element*. Each class has a formal definition, allowing to distinguish which entity can belong to the class. The semantic annotator was constructed by creating rules to assign the lexemes of a text to their respective ontological class in the SUMO ontology, some examples of rules are provided in Figure 2.2.

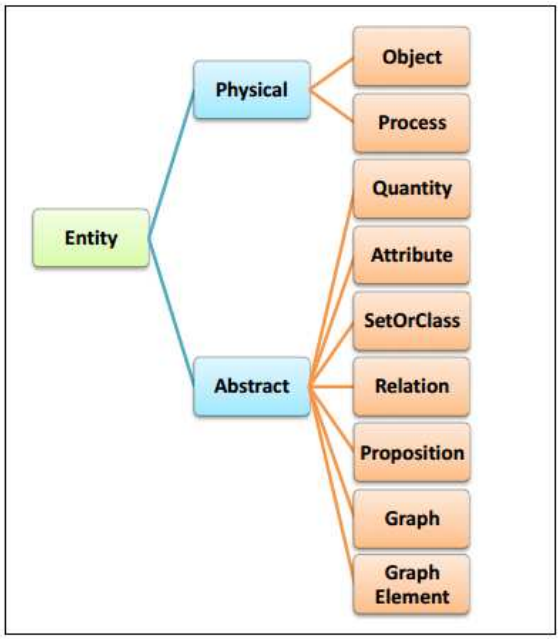

Figure 2.1: Top layers of the SUMO Ontology

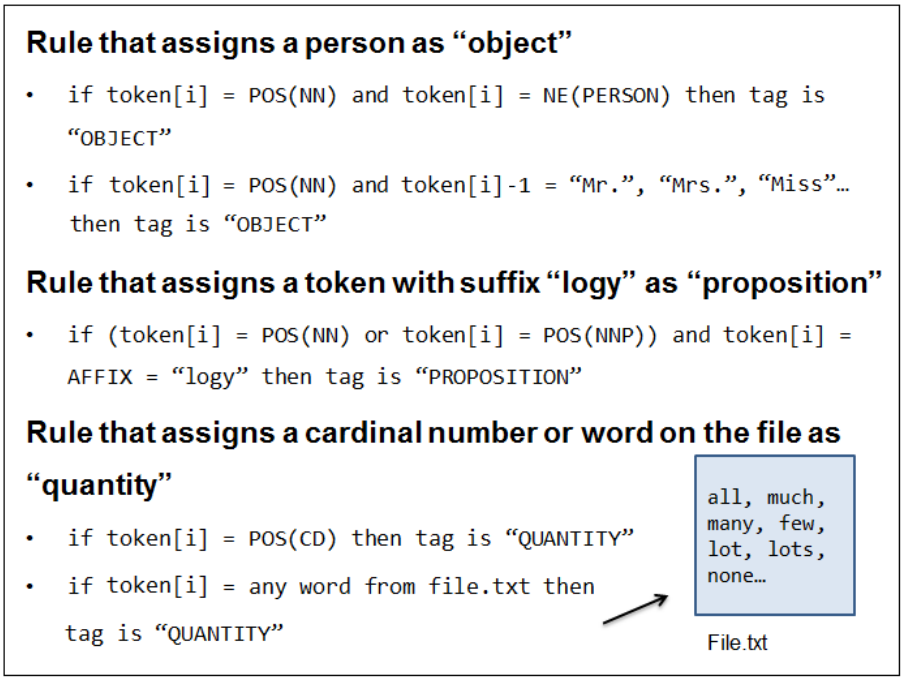

Figure 2.2: Rules examples

The Open American National *Corpus* (OANC) Ide and Suderman (2004) was selected to carry out training and tests of the annotator. The linguistic diversity of the Open American National *Corpus* allows expressing a wide range of language expressions and covering the largest number of words in American English. OANC is a *corpus* composed of 5 million words derived from various textual and oral genres of American English (Ide and Suderman, 2004). It is free of charge and available for download. It is annotated according to structural markup, sentence boundaries, part of speech, noun chunks and verb chunks, which justifies the choice of the *corpus* for the application (Ide and Suderman, 2004). The annotation provided by the *corpus* served as the basis for the construction of the rules of this work.

Due to the massive amount of documents, it was necessary to make a snippet of the *corpus* to turn the application development more manageable. The sub-*corpus* chosen was initially the text entitled "Who Killed Martin Lutter King?". The sample was used to illustrate the procedure adopted by the application to perform a properly annotation. The document in .xml format was extracted containing the annotations provided by the *corpus*. The .xml document, as well as all the *corpus* files, are in the Linguistic Annotation Format (LAF) (ISO 24612) standard for creating annotated *corpus*. In order to process the document it was necessary to normalize it, excluding paragraph markings, white space, headers and structural tags. The output of the normalization was a .txt document containing only annotated sentences.

The file generated in the previous phase went through further transformations. An important information for the application being developed is the named entity annotation, however the OANC does not provide this type of annotation. In order to add this layer of annotation to the *corpus* it was used the Stanford NER, a named entities annotator. Stanford NER is a annotator created by the Stanford Natural Language Processing Group, and it annotates entries under the categories “PERSON", “ORGANIZATION" and “LOCATION", using the Conditional Randon Field (CRF) approach (Finkel et al., 2005). The outcome of this phase was a .txt file having the annotations and the format required for the development of the semantic annotator.

The semantic annotator proposed has three phases, formatting for annotation, annotation and post annotation. The formatting phase formats the input document into a structure capable of being interpreted by the annotator. The annotation is the step that marks the elements present in the text according to a SUMO ontology category. Finally, post annotation uses the already annotated structure to create the annotated .txt document. The details of each phase will be described in the following paragraphs.

A document consists of a series of sentences, which in turn is composed of a series of tokens. Each sentence token has become a dictionary entry where the key is the number of the sentence and the value of the entry is a list of pair $< token, attributes >$. The *attributes* is a set of syntactic and semantic information about the token. Figure 2.3 shows a sentence with the annotations and Figure 2.4 shows the dictionary structure.

```
<s><tok base="last" msd="JJ" ne="O">Last</tok> <tok base="week" msd="NN" ne="O">week</tok><tok base="," msd="," ne="O">,</tok> <tok base="a" msd="DT" ne="O">a</tok> <tok affix="s" base="memphi" msd="NNP" ne="LOCATION">Memphis</tok> <tok base="jury" msd="NN" ne="O">jury</tok> <tok affix="ed" base="find" msd="VBD" ne="O">found</tok> <tok base="that" msd="DT" ne="O">that</tok> <tok base="restaurant" msd="NN" ne="O">restaurant</tok> <tok base="owner" msd="NN" ne="O">owner</tok> <tok base="lovd" msd="NNP" ne="PERSON">Lovd</tok> <tok affix="s" base="jower" msd="NNP" ne="PERSON">Jowers</tok> <tok affix="ed" base="be" msd="VBD" ne="O">was</tok> <tok affix="ed" base="involve" msd="VBN" ne="O">involved</tok> <tok base="in" msd="IN" ne="O">in</tok> <tok base="a" msd="DT" ne="O">a</tok> <tok base="conspiracy" msd="NN" ne="O">conspiracy</tok> <tok base="to" msd="TO" ne="O">to</tok> <tok base="kill" msd="VB" ne="O">kill</tok> <tok base="martin" msd="NNP" ne="PERSON">Martin</tok> <tok base="luther" msd="NNP" ne="PERSON">Luther</tok> <tok base="king" msd="NNP" ne="PERSON">King</tok> <tok base="jr" msd="NNP" ne="PERSON">Jr</tok><tok base="." msd="." ne="O">.</tok></s>
```

Figure 2.3: Sentence previous annotation sample.

After the formatting for annotation phase, the actual annotation phase began. The annotation phase consists of applying rules that evaluate whether a token belongs to an ontological category. The rules evaluate several aspects of the token, such as its affixes, POS, named entity annotation, neighborhood tokens and occur-

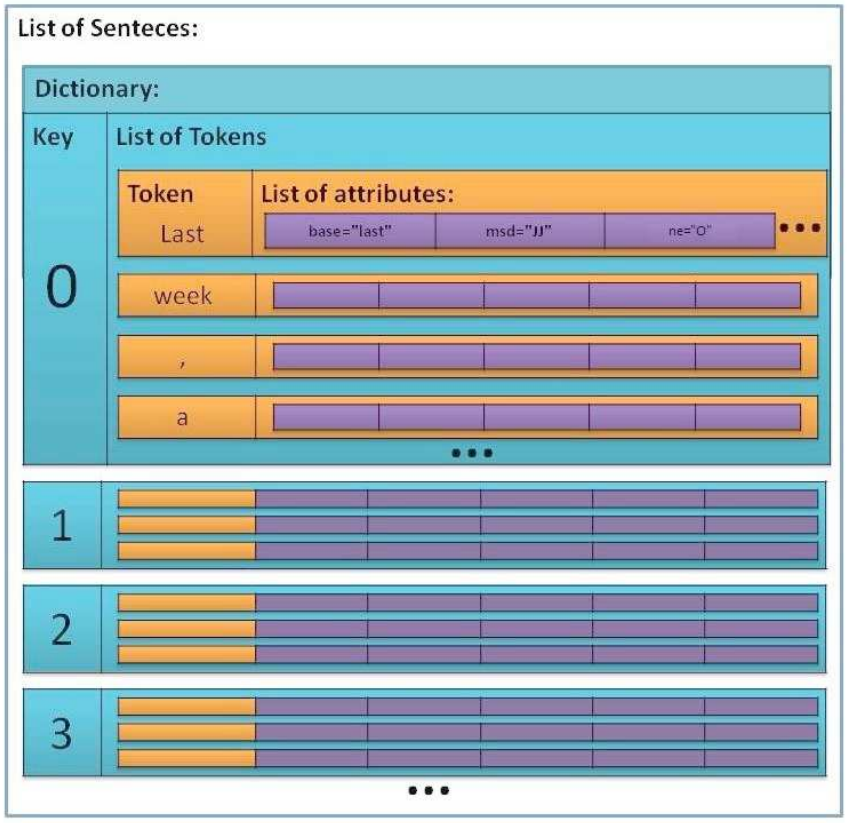

Figure 2.4: Dictionary structure.

rence in gazetteers lists. When the token is classified under a given rule, the class tag is added to the attribute list of the token. The classification occurs inversely, from the third to the first level, because if the token is classified under a class of the third level, it is already possible to say its second and first level.

At the end of the classification the token receives three labels (so1stl, for the first level of the SUMO ontology, so2ndl for the second level of the SUMO ontology, and so3rdl for the third level of the SUMO ontology). The classification adds the tags to the attributes list of the token, receiving the “CLASS NAME" if the rule applies to the token or “O" if the rules does not apply to it. If applicable to the first level of the SUMO ontology the annotator adds the “ENTITY" tag. At the second level the annotator may mark the token as “PHYSICAL" or “ABSTRACT". At the third-level annotator tags are “OBJECT", “PROCESS", “QUANTITY", “ATTRIBUTE", “SET OR CLASS", “RELATION", “PROPOSITION", “GRAPH" and “GRAPH ELEMENT".

The tokens of the document receive annotation related to the ontology, changing in the form exemplified by Figure 2.5. The third step restructures the entire

list of sentences in a document annotated according to the ISO format Linguistic Annotation Format (LAF) (ISO 24612) model. This phase is necessary because it is important that the document outputted by the annotator be in a standard format that can be used by other applications.

<s><tok base="last" msd="JJ" ne="O" so1stl="ENTITY" so2ndl="ABSTRACT" so3rdl="ATRIBUTE">Last</tok> <tok base="week" msd="NN" ne="O" so1stl="ENTITY" so2ndl="O" so3rdl="O">week</tok> <tok base="," msd="," ne="O" so1stl="O" so2ndl="O" so3rdl="O">,</tok> <tok base="a" msd="DT" ne="O" so1stl="O" so2ndl="O" so3rdl="O">a</tok> <tok affix="s" base="memphi" msd="NNP" ne="LOCATION" so1stl="ENTITY" so2ndl="PHYSICAL" so3rdl="OBJECT">Memphis</tok> <tok base="jury" msd="NN" ne="O" so1stl="ENTITY" so2ndl="O" so3rdl="O">jury</tok> <tok affix="ed" base="find" msd="VBD" ne="O" so1stl="O" so2ndl="O" so3rdl="O">found</tok> <tok base="that" msd="DT" ne="O" so1stl="O" so2ndl="O" so3rdl="O">that</tok> <tok base="restaurant" msd="NN" ne="O" so1stl="ENTITY" so2ndl="O" so3rdl="O">restaurant</tok> <tok base="owner" msd="NN" ne="O" so1stl="ENTITY" so2ndl="O" so3rdl="O">owner</tok> <tok base="loyd" msd="NNP" ne="PERSON" so1stl="ENTITY" so2ndl="PHYSICAL" so3rdl="OBJECT">Loyd</tok> <tok affix="s" base="jower" msd="NNP" ne="PERSON" so1stl="ENTITY" so2ndl="PHYSICAL" so3rdl="OBJECT">Jowers</tok> <tok affix="ed" base="be" msd="VBD" ne="O" so1stl="O" so2ndl="O" so3rdl="O">was</tok> <tok affix="ed" base="involve" msd="VBN" ne="O" so1stl="O" so2ndl="O" so3rdl="O">involved</tok> <tok base="in" msd="IN" ne="O" so1stl="O" so2ndl="O" so3rdl="O">in</tok> <tok base="a" msd="DT" ne="O" so1stl="O" so2ndl="O" so3rdl="O">a</tok> <tok base="conspiracy" msd="NN" ne="O" so1stl="ENTITY" so2ndl="O" so3rdl="O">conspiracy</tok> <tok base="to" msd="TO" ne="O" so1stl="O" so2ndl="O" so3rdl="O">to</tok> <tok base="kill" msd="VB" ne="O" so1stl="O" so2ndl="O" so3rdl="O">kill</tok> <tok base="martin" msd="NNP" ne="PERSON" so1stl="ENTITY" so2ndl="PHYSICAL" so3rdl="OBJECT">Martin</tok> <tok base="luther" msd="NNP" ne="PERSON" so1stl="ENTITY" so2ndl="PHYSICAL" so3rdl="OBJECT">Luther</tok> <tok base="king" msd="NNP" ne="PERSON" so1stl="ENTITY" so2ndl="PHYSICAL" so3rdl="OBJECT">King</tok> <tok base="jr" msd="NNP" ne="PERSON" so1stl="ENTITY" so2ndl="PHYSICAL" so3rdl="OBJECT">Jr</tok> <tok base="." msd="." ne="O" so1stl="O" so2ndl="O" so3rdl="O">.</tok> </s>

Figure 2.5: Sentence annotated with the ontological categories.

2.4 Results

In this section we present the results of the test conducted with the annotator on a text sample to illustrate its performance. The system has been tested with 39 sentences and 908 tokens from the text sample mentioned in the previous chapter after it being manually annotated. In addition is shown the confusion matrix, precision and recall measurements performed for each ontological class, assuming the annotation results from the chosen sub-*corpus*.

Because of the huge size of the entire chosen *corpus* used to build the rules, it was necessary to select a sample text to perform the test. The test was conducted by manually adding tags to the documents according to the top-level ontology and then comparing it with the same sample annotated by the application. Although it is a limited fragment of the *corpus*, the text exemplifies how the annotator would perform if the *corpus* was already manually annotated.

Table 2.1 shows the confusion matrix for the third-level classes. A special class, called “NO ONTOLOGY" was added to handle tokens that did not fit into any class. The vast majority of tokens were annotated correctly. But the result was not good with the terms related to the class “SET OR CLASS". A minority of tokens related with the special class “NO ONTOLOGY" were misclassified. Note that all

Table 2.1: Confusion matrix

	OBJEC	PROCE	QUANT	ATTRI	SETCL	RELAT	PROPO	GRAPH	GRAEL	NONON
OBJEC	97	0	0	0	0	0	0	0	0	0
PROCE	0	14	0	0	0	0	0	0	0	0
QUANT	0	0	15	0	0	0	0	0	0	0
ATTRI	0	0	0	46	0	0	0	0	0	0
SETCL	4	0	0	0	1	0	0	0	0	0
RELAT	0	0	0	0	0	0	0	0	0	0
PROPO	0	0	0	0	0	0	0	0	0	0
GRAPH	0	0	0	0	0	0	0	0	0	0
GRAEL	0	0	0	0	0	0	0	0	0	0
NONON	69	14	7	0	0	0	0	0	0	641

On the vertical are the classes that should be assigned to the tokens and horizontally those that were assigned by the annotator.

elements of this class that were erroneously classified ended up being classified into class "OBJECT". This shows that the rule is not distinguishing properly when the concept denoted by a token focuses more on the aspect of the elements than on the parts.

The annotation accuracy relative to the third level of the SUMO ontology was 89.65%. As one can infer from the confusion matrix the precision, recall and F1 measures had good results except for the class “SET OR CLASS".

```
-----PRECISION-----

OBJECT PRECISION: 0.5705882352941176
PROCESS PRECISION: 0.5
QUANTITY PRECISION: 0.6818181818181818
ATTRIBUTE PRECISION: 1.0
SET_OR_CLASS PRECISION: 1.0
RELATION PRECISION: 0.0
PROPOSITION PRECISION: 0.0
GRAPH PRECISION: 0.0
GRAPH_ELEMENT PRECISION: 0.0
NO ONTOLOGY PRECISION: 1.0


-----RECALL-----
```

```
OBJECT RECALL: 1.0
PROCESS RECALL: 1.0
QUANTITY RECALL: 1.0
ATTRIBUTE RECALL: 1.0
SET_OR_CLASS RECALL: 0.2
RELATION RECALL: 0.0
PROPOSITION RECALL: 0.0
GRAPH RECALL: 0.0
GRAPH_ELEMENT RECALL: 0.0
NO ONTOLOGY RECALL: 0.8768809849521204


-----F1 MEASURE-----

OBJECT MEASURE: 0.7265917602996255
PROCESS MEASURE: 0.6666666666666666
QUANTITY MEASURE: 0.8108108108108109
ATTRIBUTE MEASURE: 1.0
SET_OR_CLASS MEASURE: 0.33333333333333337
RELATION MEASURE: 1.0
PROPOSITION MEASURE: 1.0
GRAPH RECALL: 1.0
GRAPH_ELEMENT MEASURE: 1.0
NO ONTOLOGY MEASURE: 0.934402332361516
```

The statistical results provided in this section refers only to the text sample and it does not apply to the *corpus*. The sub-*corpus* was used only to exemplify the behavior of the annotator comparing to a manually annotated text. To verify the overall metrics of the *corpus* it would be necessary to hand-annotate all files and afterwards compare them with the rule-annotated documents generated by the application.

2.5 Conclusions

Semantic annotation allows data to be interpreted by applications in such way that machines can capture the underlying meaning of an utterance. However, annotating documents to help express aspects of their semantic meaning is still challenging, due to the lack of applications that assist the task. Notably, there is some difficulty of finding tools capable of executing semantic annotation in text documents using ontological concepts, this was the main reason for the development of this research. Manual annotation is a task that takes time and knowledgeable staff to carry it out, and the proposal of a rules-based annotator can be of great help.

Therefore, the proposal of this research was the creation of a tool that would aid in the process of semantic annotation based on top-level ontological classes. The tool makes use of a set of rules elaborated according to the concepts described by the ontology and by making use of previous annotations layers provided by the *corpus*.

Although the experiment described in this paper only used a single document to demonstrate viability of the proposal, it is possible to apply the same technique to the whole OANC. The annotation of the whole *corpus* helps to enrich it in an ontological dimension, so the text files can be used in futures researches on the semantic annotation field.

The importance of this work is the possibility of increasing the number of annotated *corpus* with ontological information, which may facilitate the training annotators based on supervised machine learning techniques, enabling a new generation of semantic annotators with higher performance and accuracy.

Chapter 3

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

Abstract

Natural language processing still faces the challenge of making machines understand the meaning expressed by a set of words occurring in a sentence. Semantic annotation aids in this process by adding metadata that assign meaning to terms. There are several semantic facets that can be annotated to terms, such as context, semantic roles, and ontological categories. Top-level ontological categories add information about the nature of a term and allow us to distinguish, for example, the different meanings of the same word. The work proposal is a hybrid approach to semantic annotation based on top-level ontologies applied to an American English *Corpus*. The research is divided into two annotation phases, both directed by the top-level use of Schema.org, an ontology that underlies the Linked Data approach. The first method is the development of a rule-based annotator and the second is a manual annotator for correction and addition of labels to terms not previously captured. The results obtained are used to the Open American National *Corpus* and can be later adopted for new applications in the context of semantic annotation.

3.1 Introduction

Corpora are indispensable resources when they come to researching and developing products related to human natural language. *Corpora* consist of a set of records of natural language in spoken or written form that can be interpreted by computers applications (Pustejovsky and 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 and Stubbs, 2012).

In computational linguistics, the term annotation refers to the process of adding meta-data in a given *corpus* in order to promote the process of extracting information by the applications (Pustejovsky and 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 that attempts to explore the meaning of the things being marked (Wilson and Thomas, 1997). The semantic annotation searches for elements of the text and classifies them according to their meaning in the fragment in which it is inserted. The semantic annotation allows data to be interpreted by applications in such a way that machines can also interpret 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 annotated to terms, 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 or 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 "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). Erdmann et al. (2000) explains that ontologies are responsible for guiding the process of annotation becoming the classes which the words will be tagged. 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). The task of ontology-based semantic annotating can also be very useful for future applications that needs a substantial amount of data to train learn algorithms.

However, the process of annotating large textual *corpus* is a long and expensive process, 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 the work was to complement the annotation doing a manual annotation of the same *corpus* pre-annotated by the first phase in order to improve the accuracy of the annotation. After those two process 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 work previously developed that are related to this research; Section 3.3 describes the materials and methods applied in the research; Section 3.4 presents the results obtained; and Section 3.5 presents the final remarks.

3.2 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 annotated 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)[1]. 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 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

[1] http://www.adampease.org/OP/

and multi-model annotation task 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 and Auer (2015) enriched semantically unstructured content using Schema.org. In particular, they embed metadata into unstructured documents in 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 justifying that this would only be a phase of the research.

3.3 Materials and Methods

The research presented here, as previously mentioned, address 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[1]. 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

[1]http://schema.org/docs/faq.html

(Patel-Schneider, 2014). It is formed by a set of types, organized under a hierarchical structure and establishing nested relations between them. The types 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 it is stated in its website[2], it presents itself in a hierarchy of types, each 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 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 types. To better explore and understand the Schema.org hierarchy, a solar burst chart was constructed based on the Full Hierarchy of Schema.org[3], as can be seen in Figure 3.1. 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-types (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 has been defined, the next step was to carry out the *corpus* selection.

In order to express the linguistic variety of the American English it was selected the OANC (Ide and Suderman, 2004). 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 and Suderman, 2004). This diversity of genres allows to express a large number of words in American English and covers expressions of 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 as structural markup, sentence boundaries, part of speech, noun chunks, and verb chunks, produced au-

[2]http://schema.org/docs/datamodel.html
[3]http://schema.org/docs/full.html

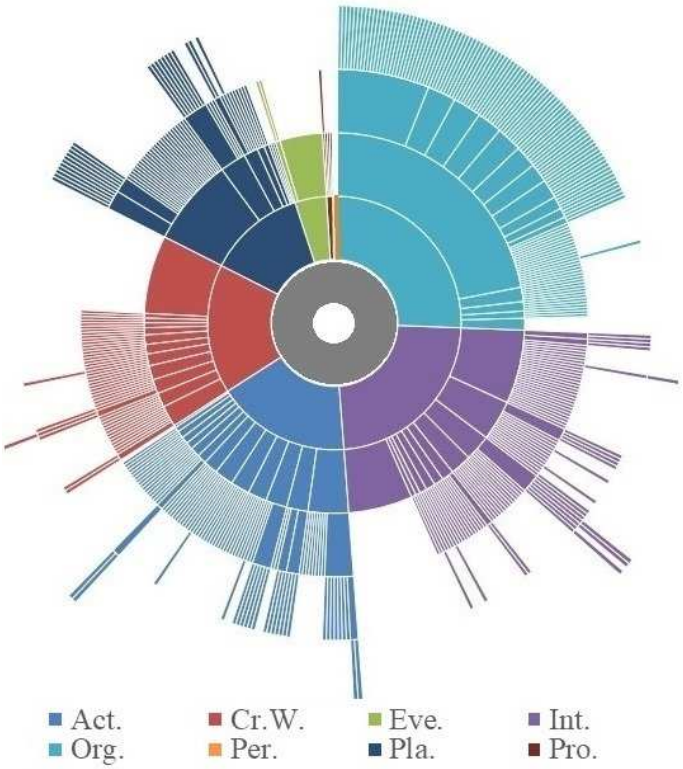

Figure 3.1: Schema.org Sunburst Graphic

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.

tomatically using annotation systems (Ide and Suderman, 2004). The entire *corpus* is available for download at the American National *Corpus* website free of any charge. All characteristics above mentioned 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). For the development of the application it was necessary a pre-processing of the whole *corpus* to shape it in a standard that could be handled by the task. First, it was necessary to remove all the markings 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 been made automatically, some corrections were needed in order to repair some spurious annotations. The most frequently error encountered were sentence boundary, 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 the 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 nouns 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 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 application execution. 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 being 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 3.2. A second type of file was also created, but only made up of plain unlabeled sentences. These files was designed for use in future annotations tasks.

It was also noticed the need to annotate named entities to assist in the process of identifying the associated ontological classes. For the accomplishment of this task, 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 annotation 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

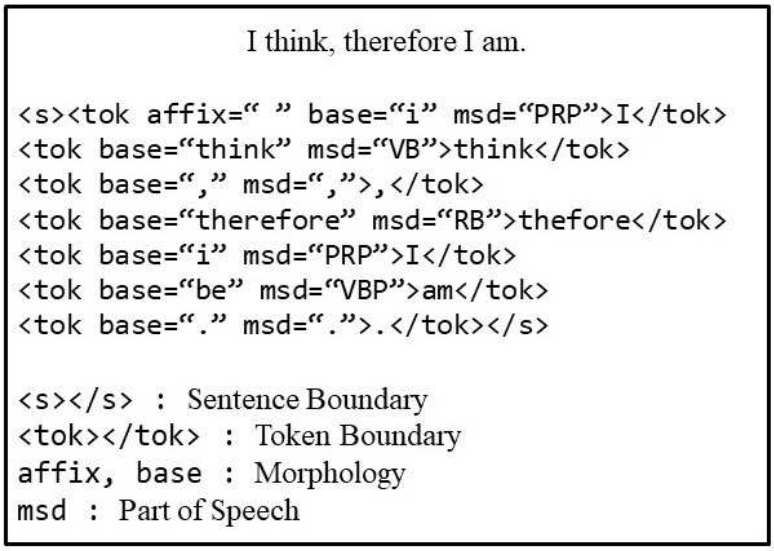

Figure 3.2: Sentence formatting

The figure 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.

token features related to the selected 8 top-level classes of the ontology and assign a label to the token as shown in Figure 3.3.

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 is met. The application was designed to assigned 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 annotation second phase 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 promote 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

```
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
```

Figure 3.3: Rules examples

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.

Schema.org top-level types. 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.

3.4 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 3.1. 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 3.1: Tagged *corpus* in Numbers.

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

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 3.4, separated in three categories: people, places, and organizations. The pre-annotation was done to supply features for the rules of the later stage.

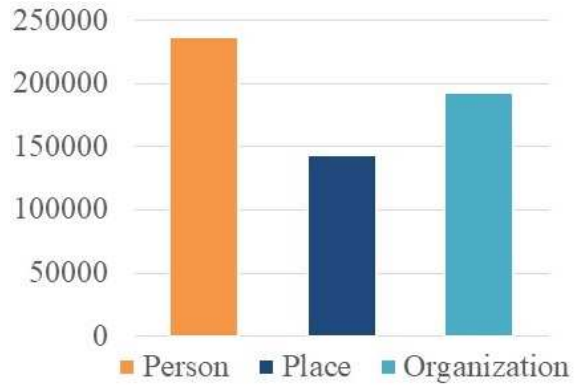

Figure 3.4: Number of items annotated by SNER.

The rule-based annotator was responsible for assigning labels, according to the set of rules for each category. For each of the eight categories 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 rule-based annotator tagged a total of 1,010,312 tokens distributed according the first columns of the Figure 3.5.

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 noted. The manual annotation of the entire *corpus* would be unfeasible due to the lack of time and human resources. Instead of re-annotate 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 *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 3.5.

It was noticed that the SNER tool was not enough to annotate all tokens belonging to the three types. The number of people, place and organization was superior to the results presented by the pre-annotation. The SNER also misclassified some terms that were corrected on the second phase. The combination of our

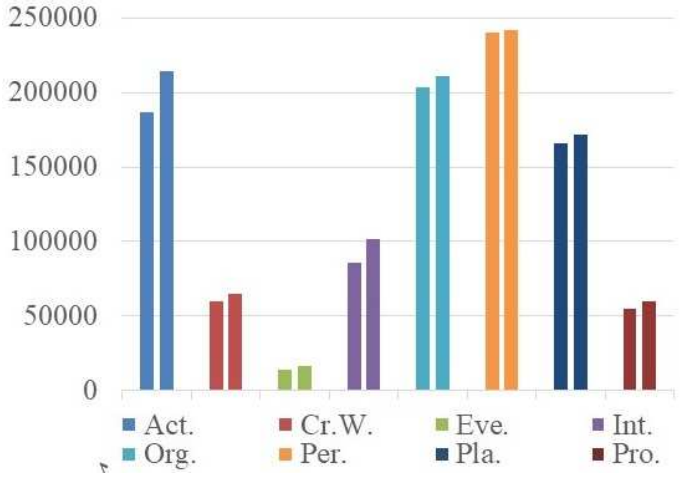

Figure 3.5: Rule-based and Manual Annotation

Its possible to see that occurred improvements in all categories.

proposal with the tool produced better results, justifying the necessity of not using SNER alone for the annotation.

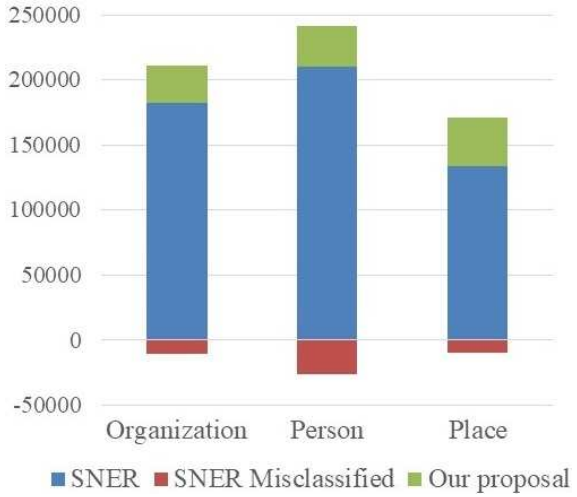

Figure 3.6: Organization, People and Place annotation

The figure shows the total number of organization, people and place 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.

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, it was annotated 1,080,464 types distributed according to the second columns of Figure 3.5. The *corpus* can be used for many activities related to natural language processing, attesting the importance of the contribution of this research.

3.5 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.

In this paper, we have described the process of semantic annotation of a *corpus* in the English language, using as elements for annotation ontological categories. The semantic annotation of the *corpus* used a hybrid approach: automatic rule-based and manually checked. The *corpus* produced, 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. The next step of the research is to use the *corpus* as a measurement parameter and to aid in the creation of other semantic annotators.

Chapter 4

CRF model applied to semantic annotation of top-level ontologies of Schema.org using an American English *corpus*

Abstract

On the information age, various documents and texts, especially those available on the internet, are still appropriately designed for human interpretation only. Understanding natural language texts by computers with the same human performance is still a challenging task. The ability to make text documents interpretable by machines can be facilitated through the technique of semantic annotation, which aggregates metadata to the content of the text in order to express the meaning contained therein. The semantic annotation must be led by the use of labels that can capture the essence present in the annotated element, such as the use of types from ontologies. The application of ontologies in the annotation task can span multiple domains, however this particular research focused its approach on top-level ontologies due to its generalizing characteristic. Because it is an arduous task that demands time and specialized manpower to perform it, much is done on ways to implement the semantic annotation automatically. The use of machine learning techniques are an effective approach in the annotation process. Another factor of great importance for the success of the training process of the supervised learning algorithms is the use of a sufficiently large *corpus* and able to condense the linguistic variance of the natural language. In this sense, this article aims to

present an automatic approach to enrich documents from the American English *corpus* through a CRF model for semantic annotation of ontologies from Schema.org top-level. The research uses two approaches of the CRF model obtaining promising results for the development of semantic annotation based on top-level ontologies. Although it is a new line of research, the use of top-level ontologies for automatic semantic enrichment of texts can contribute significantly to the improvement of text interpretation by machines.

4.1 Introduction

A key resource for the area of natural language processing is the linguistic *corpora*. These natural language samples, also called *corpus* (in singular), are usually accompanied by meta-data containing information about the tokens and their sets (Sardinha, 2000). The addition of meta-data to the *corpus* is called annotation, in other words, it is the process of adding new information in source texts. The annotations that aggregates information to the *corpus* can be applied to the document with a whole, its sentences, terms and words and can be performed manually by humans or automatically by machine techniques (Leech, 1997). The annotation allows the search for linguistic phenomena and from these phenomena generate statistics and correlations between them. All information extracted from a *corpus* coming from the annotation process, serves as the basis for searching for new inferences and studies in the area of Natural Language Processing (NLP). In general, the main purpose of annotation in the NLP field is to serve as input for methods of machine learning.

The *corpus* annotation embraces several spheres of linguistics, so the annotations can represent grammar class of the annotated elements, their morphology, correlation phenomena, aspects of phonetics and so on (Leech, 1997). The annotation may also cover other aspects related to the structure of the annotated text and its content itself. Under these circumstances, over the years much has developed in the area of NLP thanks to the benefits provided by adding information regarding grammar aspects of the annotated elements. However, a set of annotated words may provide much more information regarding its interpretation and significance that can be used by computational techniques (Pustejovsky and Stubbs, 2012). In this sense, it is common knowledge that every set of words organized in a structured way carries with it an inherent meaning, and in the field of linguistics the study of this meaning is the responsibility of semantics. Since semantics studies the meaning and interpretation of natural language terms, this property was also

used to contribute to the annotation process. In this way, a new annotation layer can be added to a previously annotated *corpus* introducing semantic perspectives to the document.

Semantic annotation is the process in which are added to the terms significant references to express their meaning. Not unlike other forms of annotation, semantic annotation is performed by labeling elements, but the annotation task attempts to capture the essence of the meaning of the tagged object (Kiryakov et al., 2004). In general, the main goal of semantic annotation is to make texts interpreted by humans also containing content understandable to computers. For semantic annotation execution, a set of labels pertaining to a domain is selected and such labels are assigned to the terms annotated by their meaning. This annotation category can instantiate words, sentences, paragraphs, or full text and can incorporate one or more domains (Reeve and Han, 2005). There are many advantages of semantic annotation, such as allowing context comprehension, assisting search tools, establishing correlations, and the main one which is to offer meaning to the set of words Handschuh and Staab (2003). The semantic annotation establishes a network of concepts allowing to infer the context referring to the annotated context. Since the semantic annotation is performed on a certain document, it can be easily interpreted by machines allowing a multitude of applications.

To perform the annotation task, it must first specify the domain to be annotated and then select the annotation labels. The selected tags will guide the annotation process by labeling the candidates terms. There are several alternatives to guide the process of semantic annotation, but the most common of them is the use of ontologies. Philosophically speaking, the term ontology is related to the study of all things in the world, and its main function is the organization of these things under different types. In the words of Gruber (1993), an ontology in the field of Computer Science is a "specification of a conceptualization". In this sense the ontology tries to describe the concepts of the things existing in a domain and to relate them according to their characteristics (Guarino, 1998). To become an object of computing field, the categories, concept and definitions of an ontology need to be constructed under a formal specification, representing an abstract real-world model capable of being machine readable. Since the ontologies are formed by concepts, properties and relations of the domain which they propose to specify, they are strictly adequate to guide the process of semantic annotation.

When using ontologies in the annotation process, the ontological classes become the labels and the specify which objects should be annotated (Uren et al., 2006). The task of annotation is closely related to the domain to which the ontol-

ogy proposes to encompass. There are several types of ontology, which are classified according to their function and their scope. Generally in the semantic annotation when motivated to tag a specific domain uses classes of an ontology corresponding to that domain. On the other hand if the annotation comprehends more broad concepts are used more generic ontologies that can incorporate a vast number of domains. General domain ontologies are extremely extensive and can define concepts applicable to any domain. Due to its expressiveness and to the large number of classes that compose such ontologies, it is selected ontology fragments from them, usually the top-level, to form the label set and then to perform the annotation task. Top-level ontologies describe broader and more abstract concepts regardless of a particular problem or particular domain (Guarino, 1997). Because top-level ontologies have fewer classes and generic concepts, they may play an important role in semantic annotation. Annotating top-level ontologies can be the starting point for deepening semantic annotation at more specific levels of a general ontology.

More than producing semantically annotated *corpus*, semantic annotation based on top-level ontologies can also be conducted primarily for the enrichment of Web content. One of the main applications of semantic annotation in the internet sphere is the contribution that it can offer to the Semantic Web (Handschuh and Staab, 2003). The Semantic Web is a web perspective and follows the principle that all information made available on the Internet should be labeled in such a way that computers are able to understand the content (Berners-Lee et al., 2001). The ultimate goal of the Semantic Web is to enable machines to perform more useful tasks by developing a network of connected data through standard vocabularies and definitions that have semantic meaning with them (Berners-Lee and Fischetti, 2001). All this action would facilitate for example search engines in providing a response to users that is most relevant to their desire. Vocabularies from an ontology are important elements and valuable tools for organizing the data of a domain and enriching it by adding meanings. In this sense the semantic annotation based on ontologies plays a fundamental role in the process of semantic enrichment of web content to support the Semantic Web (Maedche, 2012).

Despite all the advantages described regarding semantic annotation, it still faces a difficult challenge. In the process of annotation performed through human work, factors related to the time, cost or heterogeneity of linguistics itself still prevent the task from being performed optimally. Automation of annotation routine using computational tools comes to be a solution (Pustejovsky and Stubbs, 2012). In order to optimize the task and decrease such complexity, the PLN area applies methods that learn from annotated *corpora* using machine learning tech-

niques. The learning algorithms have the ability to after training under the use of previously annotated *corpus* perform the annotation of new documents through the generated models. However, to make use of automatic annotation techniques via learning algorithms training material is needed and there is large lack of annotated *corpora* in this sense (Uren et al., 2006). Another difficulty is to find top-level ontologies capable of specifying appropriate domains to guide the semantic annotation process. These factors that hinder the development of the task serve as motivation for the development of research in the area.

Faced all the problems reported in the previous paragraphs, the objective behind this research emerges in order to initiate research aimed at the area. The main purpose of this work is to make use of a machine learning classificatory methodology to perform the semantic annotation based on top-level ontologies of an American English *corpus*. The main intention is to construct a model capable of classifying the selected top-level ontology types with a satisfactory prediction rate to apply it in the semantic annotation task. The present work that reports the research is organized as follows. The section 4.2 gives an overview of the works that have a relation with this research, presenting the advances and the points of improvement of the area. The section 4.3 are divided into three subsections. Schema.org ontology subsection presents the process of ontology selection as well as its characterization and definition of the top-level responsible for generating the classification labels. The next subsection, called *Corpus*, introduces more details about obtaining the *corpus* and information about it. The CRF approach subsection presents the definition of the chosen classification model and the pre-processing stages and the classification process. The results achieved in the classification stage and a discussion about them are discussed on the next session and finally the conclusions on the last section.

4.2 Related Works

The semantic annotation based on top-level ontologies is a new field of study with limited resources on the literature to compare the approach of this work. Because of that, in this section, we present an overview of the recent and existing studies concerning the extraction and annotation of semantic knowledge from texts considering or not ontological aspects. The following paragraphs address about three different methodologies, semantic annotation, named entity recognition and ontology-based annotation. Although the first two techniques have different definition and applicability, they were considered because of sharing the same processes on automatic

annotation using classifier algorithms. The third methodology reported on this section has a direct connection with our work presenting annotation approaches that consider the use of ontological classes based on their definitions.

To present a overview about the potentiality of the inherent characteristics of the text to assist in the task of semantic annotation stands out the work of (Boella et al., 2014). The study intents to demonstrate how syntactic and lexical components from texts can contribute to their semantic enrichment. Their research derived from the difficulty in automatically obtaining a structured information for ontological construction and semantic search. They initially bring together common and well-resolved techniques from Natural Language Processing to extract syntactic information about the text such as Part-Of-Speech tagging and Syntactic Parsing. This extracted information served as a basis for the construction of features for application in a machine learning classifier in order to capture and label semantic units. For automate the task of annotation, it was used the Support Vector Machine algorithm that automatically selects features from the training set. For case of study they narrow the research to a specific domain recruiting texts from legal sphere. The classifier seek for three different labels all related with the legal texts. The proposed model presents as average result from the three models the following values, precision of 85.3, recall of 74.1, F-score of 77,3. Those results is a general average, however the model fits better to some labels than others accomplishing for example 99.2 as F-score for one of the labels. After that the same approach was used also to identify hypernyms and hyponyms relations in Wikipedia entries. The second phase achieved F-score of 85.81 for hypernyms and 79.42 for hypernyms. They concluded their work discussing about the importance of consider the lexical and syntactic aspects of sentences for the semantic extraction process and for the ontology retrieval. Our approach is a practical example of potentiality introduced by this study. Although the study has a different domain from ours, it helps to justify that the use of grammatical characteristics of the sentence can bring valuable features to support the automatic semantic annotations and also on the ontological dimension.

As mentioned above, even though named entity recognition has no ontological baseline to relate to our work, it is relevant to compare because both studies make use of the same machine learning algorithm. The work presented by Bergamaschi et al. (2017) reports the semantic enhancement for named entity recognition applying the Conditional Random Fields algorithm. This work especially comes close to ours because in addition to recognize entities, they add semantic features before performing the annotation. These technique differs from the usual NER and

brings better results furthermore adding a deeper semantic aspect to this task. The recognition method was performed from the combination of standard training features and semantic information gathered from the Cogito[1] linguistic analysis engine. Cogito semantic analysis creates a network associating words which are related to each other via semantic links. The experiments were applied to the CoNLL 2003 *corpus*[2]. The *corpus* is a manually annotated over five categories, PER, LOC, ORG, MISC and O, none of them having any ontological background. Throughout the experiment the authors compared models obtained from the use and absence of semantics and they also consider different sizes of the *corpus* to analyze the approach performance. The achieved results were considerably better comparing to the usual methodology. Without semantics they obtained a average of 0.8507 for precision, 0.8188 for recall and 0,8336 for F1-score and through the use of semantics, 0.8629 for precision, 0.8392 for recall and 0.8505 for F1-score. As conclusion, the research shows that by combining semantic information with training features resulted in a positive effect on the NER task. This work helps to demonstrate how the use of semantic networks can help in the task of recognition, so the same concept could also be applied to the use of ontologies.

Skeppstedt et al. (2014) proposes a use of machine learning techniques to recognize and annotate disorders, findings, pharmaceuticals and body structures from clinical text. Although the research has a different aspect because it has no ontology background, the authors performed a quite similar work considering the annotation process. The procedure aims to recognize clinical entities from medical texts written in Swedish. Health records have usually been annotated by these perspective in order to support the patient overview generation and medical hypothesis construction. The scientific contribution of the proposal is to support medical knowledge extraction in a language that is not English. Because of that is explored a comparative of how well clinical entities previously annotated in English are recognized in Swedish clinical *corpus*. The main reason why this research was selected as a related work is its automatic annotation approach performed by the same machine learning algorithm. After the *corpus* selection and the training and test set distribution, the CRF algorithm was applied to annotate the four selected entities. The results of the F-score from the algorithm performance by using best features, settings and its ability to generalize to held-out data was 0,81 for Disorder, 0.69 for Finding, 0.88 for Pharmaceutical Drug, 0.85 for Body Structure and 0.78 for the combined category Disorder + Finding. They did not achieve better results than clinical annotation of

[1]http://www.expertsystem.com/

[2]https://www.clips.uantwerpen.be/conll2003/ner/

entities applied to English medical records, but the obtained results are considerable whereas is a new study on the Swedish field.

Considering now ontological aspects for annotation, it can highlighted the work proposed by (Pandolfo and Pulina, 2017). The study come up with a self-adaptive system for automatic ontology-based annotation of unstructured documents on the context of digital libraries. Different from our proposal, this work has an approach of annotating an entire document over an ontological sphere and not only terms, like ours, but the use of ontologies is what correlate both works. The authors aims to create a system capable to automate the ontological-based annotation process of texts from digital libraries. The work is based on the STOLE [3], an ontology-based digital library created from documents about the history public administration in Italy on the 19th and 20th centuries. For annotation purposes they considered classes from STOLE to perform the experiment, Article, Event, Institution, Legal System, and Person. In order to execute the task, 20 documents manually annotated were selected. A pre-processing phase is applied to the *corpus* to provide necessary information to build the features, such as sentence boundary, part-of-speech, named entity recognition. The system uses its own algorithm capable of annotating automatically from features extracted from the document. The algorithm also has a self-adaptive approach. After all tests, it was noticed that the application is sensitive to the entry order of the documents producing different results for each entry. The best results achieved by the tests was precision of 0.80, recall of 0.53, F-measure of 0.63. Although the results are considerable low and the system does not use any machine learning approach, the study is important to introduce the ontological-based annotation as a new field of study for both specific domain and general domain.

Another work that also considers the use of ontologies to annotate terms related to a domain is described in the article of (Liu and El-Gohary, 2017). In their work, the authors used a semi-supervised Conditional Random Fields based on ontology for automated information extraction from bridge inspection reports. The research has as main focus the extraction of information about maintenance and deficiencies of bridges, naming entities related to the theme. As an object of analysis, eleven bridge inspection reports are used which has sufficient amount of content, considered by the authors to carry out the research. These reports generated a total of 1866 sentences on different aspects of bridge maintenance, complexity, conditions and years, making the *corpus* appropriate for applying the proposed methodology.

[3] ADORNI, Giovanni et al. An ontology-based archive for historical research. In: 28th International Workshop on Description Logics. 2015. p. 322.

The work has its beginning in the pre-processing of the text, making it readable by the application. Subsequently, the extraction phase of features was introduced, taking into account aspects related to the part of speech, stem, and semantic characteristics. Finally, the last step corresponds to the extraction of information through the semi-supervised CRF modeling that derives the sentences in a set of training and tests containing the classes to be annotated. In the evaluation phase, they take into account the constituted model and the classification of eleven classes in the set of tests that reached an average precision, recall and, F-1 measure of 94.1%, 87.7%, and 90.7%, respectively. The ontological aspect of the research is related to the definition of the classes tagged in the extraction process, so the ontologies assist in analyzing the context based on a specific domain. Again, the research differs from our work because it is a specific domain of annotation, but it is similar to being based on the use of ontologies to annotate texts using CRF as a learning model.

4.3 Materials and Methods

This section details the techniques used to develop the research. To contextualize the components used, some concepts will also be reported in the following subsections. The first subsection describes the ontology of the research as well as the selection of its top-level for the constitution of the eight labels. Next, the *corpus* used in the experiments is presented and the justification for its choice is also reported. Finally, the last subsection exposes the training algorithm and why it was used. The methodology in which the experiments were performed are reported over the course of the following three sessions.

4.3.1 Schema.org Ontology

Before report the use of Schema.org ontology, it is necessary to introduce the concept of Linked Data. According to Bizer et al. (2009) the term refers to all structured data on the Web connected and published through of the use of good and standardized practices. A global data space containing billions of structured data has been expanded over the years using those practices. The web connected data covers a multitude of domains allowing the development of several and new applications. The connection established between the data allows a navigation along links optimizing tasks such as search, and more appropriate responses according to content sought (Heath and Bizer, 2011). In this sense, all data published on the web from different sources, connected some how with internal and external data set, with meaning

explicitly defined and machine-readable is characterized a Linked Data. To make typed statements that link distinctive data sources the Linked Data relies on files containing data in RDF (Resource Description Framework) format (Bizer et al., 2009). The main goal of feeding the Web with this huge amount of data is support the understanding of contents by machines, the called Semantic Web. The Semantic Web is defined as an web environment where machines directly or indirectly process the data and understand it (Berners-Lee and Fischetti, 2001).

A practical example of the utilization of Liked Data principles to construct a set of connected data is the Linking Open Data[4] project. The project's main intention is to support applications and researches for the Semantic Web by collecting open web data, converting to RDF format and finally publishing it on the Web (Bizer et al., 2009). The content gathered by the project relies on different fields of study, but they are released under an open license making possible reuse free of charges. The project is disposed under a cloud that shows its content and all the network that links them. When data is published on the web by the Linked Open principles, the content is allocated on global data space, which allows the data to be identified by semantic queries used by various applications (Heath and Bizer, 2011). One example of structured data supported by Linked Open Data is the Schema.org project.

To encourage the further development of Linked Data to support Semantic Web, there was a need for webmasters to add semantic information about the content published on the web. However, there was no standardized semantic vocabulary to accomplish the task. In this sense and in order to offer a solution to this problem, the Schema.org was launched in 2011 by the most prestigious search engines corporations such as Bing, Google, Yahoo and others. Schema.org is an initiative to create a collaborative project that supports Linked Open Data by creating structured data schemas (Patel-Schneider, 2014). The goal behind the project was to offer a single and integrated approach capable to cover the wide range of frequent web topics and propose a structured vocabulary for web page marking. The vocabulary is organized by a hierarchy of types that comes from the content commonly consumed by users of web pages. Because of that the Schema.org is not a static project, so the number of classes and relations are growing over the years (Guha et al., 2016). The version used by this work is the Core Vocabulary[5] composed by 606 types so far according to the full hierarchy shown on its web site.

From the use of web content, the proposed types were being organized through

[4]https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[5]http://schema.org/docs/full.html

their concepts in a nesting structure forming the Schema.org hierarchy. The organization by a hierarchy or tree format guarantee that each type may have at least one father type (Patel-Schneider, 2014). The types of this hierarchy have conceptual relation between them and their definitions is what characterizes an object to be layered under this or that type. Types derived from the same supertype can invoke definitions having the same concept, but different applications, thus producing polymorphism relations. All of these characteristics earlier reported make Schema.org also defined an ontology (Guha et al., 2016). As reported by Ronallo (2012) Schema.org is defined as a "middle ontology", it means that its intention is not to cover all the existing things in the world nor to turn to a specific domain. He further adds that the main intention of middle ontologies such as Schema.org is to produce such a broad scope capable of covering the most frequent cases of a general domain. From this point the Schema.org will be considered as an ontology which has a bias towards search engine and commercial use cases commonly used by internet users. Following this perspective the ontological classes presented by Schema.org is used to markup web pages, so it can also be easily applied text documents adding semantic context to them. Because of that the Schema.org was chosen to be the ontology that provides the classes of automatic annotation phase from this research, but before that it was necessary select the ontology top-level.

In order to narrow the amplitude of the research and follow its the main objective, it was considered only the Schema.org top-level. At the first level of ontology is the Thing[6] type which encompasses the most generic type of the ontology. The first class spread into eight general domain ontology classes which have their own definitions and properties. For the proposed clipping was considered the second level of ontology addressing those eight types. Those eight types became the classes of the automatic annotation process and also they were selected because they were the annotation labels found on the *corpus* used in the experiment. According to Schema.org the definitions of each type is explained bellow with the intention to make understandable the object of annotation.

- Action: An action executed by a direct or indirect agent(s) upon a direct object producing a result;[7]
- Creative Work: The most generic kind of creative work, including books, movies, photographs, software programs, etc;[8]

[6]http://schema.org/Thing
[7]http://schema.org/Action
[8]http://schema.org/CreativeWork

- Event: An event happening at a certain time and location;[9]
- Intangible: Object that can not be tanger;[10]
- Organization: Institution that is intended to carry out acts in the various spheres of society;[11]
- Person: A person (alive, dead, undead, or fictional);[12]
- Place: Entities that have a somewhat fixed, physical extension;[13]
- Product: Any offered product or service;[14]

4.3.2 *Corpus*

To perform the automatic annotation phase proposed by this research it was necessary to select a source with specific characteristics for the case of study. The *corpus* with those properties can be found on the work proposed by Ide and Suderman (2004) and after some changes by (Andrade et al., 2018). The Open American National *Corpus* (OANC) is a American English *corpus* available free of charges that brings together texts and oral conversations of diverse categories like fiction, documents, scientific articles among others (Ide and Suderman, 2004). The main goal of the *corpus* is to gather different text sources to achieve the greatest linguistic diversity possible and offer an ideal product for researches in the Natural Language Processing area. The *corpus* is distributed through 8293 files from different sources all under the same XML standard. Each document is accompanied by annotations of a lexical, morphological and syntactic nature, such as sentence limit, part-of-speech, affixes, bases and suffixes and other annotations. The annotations provided by the *corpus* was performed automatically through annotation tools which results errors of different aspects. The errors do not compromise the development of applications, however they influence negatively the accuracy of the results obtained by these applications.

The research proposed by Andrade et al. (2018) presents a new version of the OANC plus annotations based on the top-level of the Schema.org ontology. Before adopt any annotation method the authors performed a correction of errors resulting

[9] http://schema.org/Event
[10] http://schema.org/Intangible
[11] http://schema.org/Organization
[12] http://schema.org/Person
[13] http://schema.org/Place
[14] http://schema.org/Product

from the annotation already provided by the *corpus*. After the corrections and some adjustments made in the *corpus*, it resulted in a linguistic plurality composed of 16,280,694 tokens and 214,827 types. Subsequently the documents were submitted to a standardization to finally develop the proposal of the work. Their came up with a hybrid approach to annotate the *corpus* under a top-level ontology using two complementary techniques. The top-level ontology chosen to execute the experiment was the second level of Schema.org containing the eight same classes, object of this research. The first phase of their project relies on rule-based annotation to label terms from the *corpus* according the selected classes. For each class was constructed a set of rules which analyzes the necessary parameters to assign the Schema.org tag to the candidates terms. Once the first phase was completed, there was a need to revise the annotation performed by the rule-based annotator. In this way, the second phase consisted in inspecting a portion of the documents and annotate them manually. So that the all documents go through an indirect review, the changes made in the fraction of the manually annotated *corpus* were later overlapped over the integral *corpus*. At the end of the paper, Andrade et al. (2018) accomplished a total of 1,080,464 terms annotated under the eight Schema.org top-level ontology classes distributed non-homogeneously.

In possession of the *corpus* annotated to start the experiments related to the machine learning phase, some pre-processing tasks were required. First, each document of the *corpus* was conditioned to an input standardization of the algorithm. In this standardization, all the words of a sentence occupied a single line, followed by their linguistic attributes and to separate the sentences was typed a blank line. Thereafter, were created eleven sets of data consisting of 1%, 10%, 20%, 30% respectively up to 100% of the documents. To maintain the textual variety of each set of data, the documents were carefully separated by categories. In this way, it was possible to ensure that each data set was composed of a portion of texts from each category. Subsequently, each data set was divided into two sub-sets, the training set and the test set. Obeying the rule of containing all categories, the training set and test set were randomly organized containing 80% and 20% of the documents in that order. After the standardization of the documents in the subsets, the training stage of the algorithm was started.

4.3.3 CRF Approach

The model used to train the data sets was the Conditional Random Fields Lafferty et al. (2001), so before reporting the process it is crucial explain a bit about its

operation. The framework proposed by Lafferty et al. (2001) is a probabilistic method based on conditional approach for segmenting and labeling sequence data. Although the model has an exponential configuration equivalent to the maximum entropy models, it presents effective methods for complete inference and training (Dietterich, 2002). The algorithms based on the model are trained to maximize the conditional probability of the outputs given the inputs. The method has the form of non-directed graph model which given a data sequence for training defines a single logarithmically linear distribution on label sequences (Lafferty et al., 2001). The method has a great flexibility to integrate a large number of arbitrary and non-independent input resources, and this is one of its main advantages (Dietterich, 2002). Different and more advantageous than other methods such as the Hidden Markov Model (HMM), the CRF admits a conditional perspective because it results in the softening of assumptions about the independence of states in order to ensure a treatable inference (Lafferty et al., 2001). Algorithms based on CRF has frequently been used on challenges in the area of natural language processing, and have been achieving success in most cases because of their ability to deal with data disposed under a dependent sequence.

The CRF model includes some characteristics that justify its choice for the research. Tran et al. (2017) state that among the class of probabilistic structured output methods, the CRF is popular until then and Krishnan and Manning (2006) claim that it is still a current method regarding the use of sequence models for semantic classification producing effective results. The CRF was also chosen because its probabilistic method analyses the sequence of tokens as a single input, considering each token as part of the whole, instead of assuming the token as a single element. To perform the prediction function the CRF takes into account the dependence among the input entries like a sentence, for example. This characteristic makes the algorithm suitable for labeling the ontological classes on a sentence. For accomplishment of the task the method takes into account the information contained in the sentence as well as the attributes of each isolated word. Due to the size and organization of the acquired *corpus* and the characteristics presented above, the CRF is the indicated model to perform the automatic annotation of top-level ontologies.

For the execution of the CRF model, it was used the sklearn-crfsuite[15] package, which is a wrapper over CRFsuite. The tool provides an interface similar to scikit-learn which allows to save/load CRF models using joblib or to use of scikit-

[15]https://sklearn-crfsuite.readthedocs.io/en/latest/index.html

learn model selection utilities (cross-validation, hyperparameter optimization). The sklearn-crfsuite has five available implementations of CRF algorithms, *lbfgs*, *l2sgd*, *ap*, *pa*, and *arow*. All the implementations were tested, and the *lbfgs* which is a gradient descent using the L-BFGS method presented best performance, so in order to build a better model, it was chosen to run the experiments. Sklearn-crfsuite also provides a set of typical machine learning metrics (accuracy, precision, recall, F1-score) to validate the model and analyze its predictive performance. The package runs on Phython and requires a version over 2.7, in this specific case it was used the Python 3.5.4 version. The experiments using the sklearn-crfsuite tool were organized into the following phases, features selection, training, evaluation, hyperparameter optimization, retraining using best parameters and finally learning analysis.

To define the features to be used, we analyzed the most relevant rules produced in the work of (Andrade et al., 2018). The features were extracted empirically taking into account the sentence structure, neighborhood words, word syntax and morphology, word shape, among others. After selecting the features, they were extracted from the training and test files to create the respective data sets. Once the extraction of the features was finished, the training phase was started. Firstly, the input parameters were configured to execute the *lbfgs* algorithm with elastic net (L1 + L2) regularization. The elastic net is a regularized regression approach from the statistics that linearly combines the L1 and L2 penalties, so it solves the limitations of lasso and ridge methods (Zou and Hastie, 2005). After a certain training time it is possible to do a pre-evaluation through the metrics from the results. In order to remove the weight that the class O would influence in the results, it was removed to direct focus only on the interest classes of the annotation. Subsequently the pre-evaluation, it was performed hyperparamter optimization by using randomized search over the parameters. The method was executed in the interest to obtain better results and to improve quality of the model. During this process a 3-fold cross-validation was also applied. The value related to the cross validation technique was defined empirically and taking into account the limitation of computational resources for task processing. Then, the results achieved by the model were again verified to check the improvements. After all, it was checked and the assumptions about the automatic annotation process and what learning the method was able to get.

4.4 Results and Discussion

Based on the experiments performed with the CRF classifier this section presents the results obtained from its performance in the data set. As mentioned earlier, the focus of the annotation task was to tag text terms, using features capable of capturing the nine tag classes. However, the results described here take into account only the eight classes that really assign semantic value to the annotated terms. Class O, for terms classified as OTHER, does not add value to the result, so it is not relevant to the purpose of the research. Another important point to consider is that the *corpus* used as a training and test set does not have all the balanced classes that can cause in classes of greater weights than others. Finally, the results are supported by the two classification approaches, the first using a simple CRF configuration and the second using hyperparamter optimization and cross validation, both approaches applied to all data sets. Recalling that the *corpus* was divided into sub sets that gradually increased in size to analyze the behavior of the model during the experiments.

The Table 4.1 presents the precision, recall, and f1-score values for the 8 annotated classes. The described results refer to the execution of the standard CRF obtained from the test performed with 100% of the *corpus* used in both the test set and the training set. The support for the training phase comprises the total of 6997 documents, of the various literary models offered by the *corpus*. This total of documents corresponds to the classes distributed heterogeneously in the eight classes of Schema.org. After the training stage, the test set is submitted to the evaluation of the model producing the described results. As support for the tests, a number of 859,224 tokens are counted referring to the total of each class. All classes obtained satisfactory results in the annotation process reaching a margin higher than 85% F1-score. The class that presented the best results was ACTION due to its characterization of describing nominal elements that express action, target of easy annotation. Although the EVENT class presents the lowest support for evaluation, the obtained results were able to follow the performance of the other classes. The class that presented the lowest result was ORGANIZATION, with F1- score of 0.875 although its support is still relatively high. At the end of the tests, the model reached a general average of 0.940 for precision, 0.929 for recall. Both values revert to a F1-score of 0.935, a result of great impact to the automatic annotation process due to the ability to correct labeling of the model.

After the construction of the model, the tool used to execute the CRF presents a relation with the features of greater impact for the classification. The features are ranked according to their weight relative to the classification of each class. The table

Table 4.1: Results from the pattern execution of CRF

	precision	recall	f1-score	support
ACTION	0.996	0.997	0.996	48058
PERSON	0.924	0.913	0.918	41346
PLACE	0.914	0.898	0.906	33202
INTANGIBLE	0.981	0.967	0.974	26543
CREATIVE_WORK	0.912	0.879	0.895	11059
ORGANIZATION	0.883	0.867	0.875	40572
PRODUCT	0.965	0.954	0.959	15675
EVENT	0.945	0.958	0.951	4785
AVERAGE	0.940	0.929	0.935	

4.2 presents the features that are most relevant to the prediction model for each of the top-level classes of Schema.org. The features refer to execution generated from 100% of the data set. As an explanation of the table, it is sub-divided into each class and for each one the five most notable features chosen by the model are presented. The terms presented by the features are translated as follows: affix matches the token suffix and base with its stem; postag refers to part of speech of the token and the value received matches the nomenclature found in the Alphabetical list of part-of-speech tags used in the Penn Treebank Project[16]; the - and + sign symbolize the forward and back tokens and the number refers to the degree of distance from the token. There are many other features, but these were selected to exemplify the behavior of the model during the prediction phase.

Table 4.2: Most prominent features of every Schema.org top-level class.

ACTION	**CREATIVE_WORK**	**EVENT**	**INTANGIBLE**
affix:ing	postag:NN	postag:NNP	postag:NN
postag:VBG	postag:NNP	-1:postag:JJ	postag:NNS
postag:NN	postag:NNS	-1:word.lower():book	-1:postag:JJ
-1:word.lower():is	-1:word.lower():book	base:conference	word.lower():hate
+1:word.lower():is	-1:word.lower():movie	base:party	base:law
ORGANIZATION	**PERSON**	**PLACE**	**PRODUCT**
postag:NNP	postag:NNP	postag:NNP	postag:NN
postag:NNPS	-1:word.lower():ms.	-1:word.lower():in	postag:NNS
-1:word.lower():at	-1:word.lower():mr.	+1:word.lower():city	-2:word.lower():bought
+1:word.lower():organization	+1:word.lower():smith	word.lower():country	-2:word.lower():sold
base:institution	+1:word.lower():.	word.lower():apartment	-1:word.lower():a

The Figure 4.1 shows the F1-score progression extracted from the execution made from the generated data sets. From the Figure 4.1 it is possible to note that the evolution of all classes follow the same pattern of growth as the size of the data

[16]https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

set increases. All classes have a relatively low F1-score in the first executions with 1% and 10%, however with each increment made to the data set, it results in a significant increase until the 40% and 50%. Subsequently the set with 50% of the data, the F1-score growth remains little unstable, with few peaks of elevation. This fact leads to the conclusion that although the execution of the entire data set is necessary to improve the construction of the model, the amount of data existing in 50% of the *corpus* was enough to result in a satisfactory model. This does not induce that half of the data set is disposable to construct a model with good prediction, but it justifies that the second approach carried out in the research has a sufficiently creditworthy result even though it uses only 50% of the *corpus*.

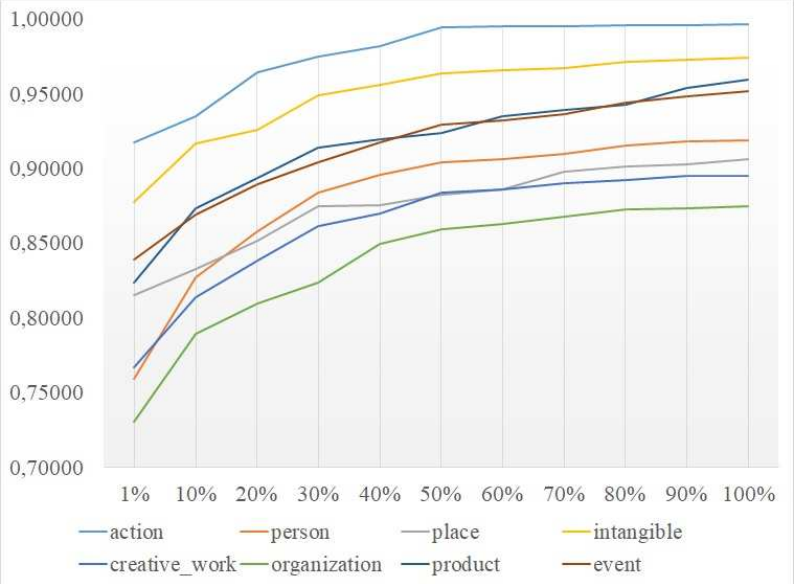

Figure 4.1: F1-score progression from the Schema.org top-level classes

Due to the limitations of computational resources to execute the second approach of the CRF model, the results presented in Table 4.3 are partial. Although the results presented are limited to only 50% of the *corpus*, it is necessary to analyze the behavior of the model from the use of the second approach. As mentioned earlier, the second phase of the model corresponds to the use of hyperparamter optimization and cross validation in order to improve the performance of the annotation process. These techniques require enough computational resources to perform the training of the model, so as the *corpus* is relatively large, the available resources were only able to handle half the set. Regarding the results obtained, it is important to observe that the performance of the model resembled the results referring to the use of all *corpus*, although with slightly lower levels of F1-score. In order to compare the results, in general the results corresponding to the experiment using the standard

model in only 50% of the *corpus* obtained an average of 0.919, 0.911, 0.914 for precision, recall and F1-score respectively. In the evaluation of the second approach, it was reached 0.927 for precision, 0.918 for recall and 0.923 for f1-score under a total support of 116530 tokens tagged by the eight classes. The average achieved results using hyperparamter optimization and cross validation were considerably higher confirming the importance of its use for optimization of the model.

Table 4.3: Results with use of hyperparamter optimization and cross validation on the CRF

	precision	recall	f1-score	support
ACTION	0.994	0.995	0.994	26866
PERSON	0.913	0.895	0.904	21625
PLACE	0.887	0.878	0.882	17748
INTANGIBLE	0.974	0.953	0.963	14830
CREATIVE_WORK	0.896	0.872	0.884	5099
ORGANIZATION	0.867	0.852	0.859	19026
PRODUCT	0.919	0.928	0.923	8537
EVENT	0.914	0.945	0.929	2799
AVERAGE	0.927	0.918	0.923	

4.5 Conclusions

The task of adding metadata to texts in the field of natural language processing is also known as semantic annotation, and can be driven by several objects of semantic content. Generally, ontologies are the main structures used for the attribution of labels to texts, however there are several layers and ontological spheres that can be explored. Top-level ontologies are responsible for abstracting and relating more generic definitions of the types of things in the world. In this sense, in the task of semantic annotation such levels of ontologies fit perfectly due to their ability to organize and conceptualize things under types that relate to each other. The use of top-level ontologies attached as semantic metadata of texts and web pages can contribute to the enrichment of content, however this is not an easy task to implement. The semantic annotation can be performed either manually or automatically, but manually it is expensive to demand skilled labor and time to perform the annotation. In summary, this research proposed an automatic approach of semantic annotation based on top-level ontologies in order to optimize the annotation practice through the use of a supervised machine learning model.

Semantic annotation based on top-level ontologies led by the use of supervised machine learning techniques can contribute considerably to task optimization. For the development of the supervised classification process, it was previously necessary to define the *corpus*, in which the OANC was selected, an English language *corpus* preliminarily annotated and formatted according to a appropriate standard. As annotation object was selected Schema.org, an ontology responsible for organizing most common types of web environment. From Schema.org was chosen its top-level which appears the eight types that have become the classification tags used by the machine learning model. This work focused on the use of the CRF model, due to its ability to consider the sequential context to which the word to be annotated is inserted. The methodology proposed by this probabilistic model is appropriate because it takes into account a set of features derived from the tokens, and still allows to take advantage of the fact that the classified elements are arranged in a sequential way in a sentence. Finally, the classification was performed and the results for the different versions of the training and test sets were analyzed.

The results obtained from this research, although difficult to compare with other studies, because it was a new annotation approach, were very favorable. In general, the CRF model presented excellent prediction results when classifying the eight selected classes, achieving results above 85% of f1-score for all classes and a general average of 93.5%. Another relevant point of the research is the results obtained for the classes, PERSON, PLACE and ORGANIZATION, commonly used in classifications of named entities. These classes achieved results that equate to the state of the art in what concerns automatic classification using learning algorithms, but now with an ontological perspective of the types classified under these three classes. Although computational resources limited the execution of the entire database using hyperparamter optimization and cross validation, the results were satisfying enough to prove that such techniques can add better results to the standard model. After all the development of the research and analysis of all the results obtained it is possible to conclude that although it is still a new approach to annotating top-level ontologies using automatic methods, the results of this research were quite promising for the development of new studies in the area.

For future work, the use of more powerful machinery for the execution of the whole *corpus* can be initially proposed using the techniques of hyperparamter optimization and cross validation. Another perspective that can be explored is the use of other machine learning techniques to compare the results obtained. One of these other techniques that is in vogue in the state of the art is the use of the deep learning model, however it is questionable that the size of the *corpus* used in this

research is sufficiently large for such technique. In order to broaden the studies in the field of automatic annotation of top-level onotologies, one proposal to be analyzed is to involve lower levels of Schema.org to discriminate less general types of the ontology. The range of new studies that can emerge from this new approach are innumerable, but what really should be focused is the contribution that semantic annotation can add to annotated contents. What should always be kept in mind is that the main purpose of using semantic annotation based on top-level ontologies is to favor the enrichment of texts or web pages making them capable of being read by machines.

Chapter 5

General Conclusions and Future Works

The work presented in this dissertation, discussed through three articles, contributes significantly to the area of semantic enrichment of texts. The work presents itself as an initiative to develop studies on the use of top-level ontologies to semantically annotate American English documents. The general objective of the proposal was reached as well as those specific that subsidized the phases for the conclusion of all work. The results obtained were very satisfactory and validated the hypothesis raised at the beginning of the project. Finally, the execution of the work resulted in some contributions that will be discussed in this session and opens the way for new research opportunities in the area.

Initially the main purpose of the research was to promote and contribute to the automatic semantic enrichment of American English texts, using top-level ontologies to guide the annotation process. To perform automatic annotation, the work also proposed the use of a learning model from the use of the Conditional Random Fields method. Although it appeared to be a simple task to complete, the work faced some difficulties in its course which resulted in the definition of new specific objectives to be achieved before the overall goal. The lack of resources to train the learning algorithms has changed the course of the research, forcing it to make a curve before executing the main objective.

After an arduous research for semantically annotated *corpus* under some ontological conceptualization, no database was found that met the criteria of the project. Due to this lack of resources, it was necessary to change the course of the research to the elaboration of a linguistic database based on the selected top-level ontology. In this sense, a *corpus* of American English, the Open American National *Corpus*, was

selected, which has sufficient size and characteristics to proceed with the next steps. The *corpus* after a pre-processing phase, corrections and standardization was submitted to a hybrid semantic annotation to assign labels corresponding to top-level ontology classes. This hybrid annotation consisted of two phases, the first using a rule-based annotator, developed during the research, and the second a manual annotation for correction and addition of new labels to unmarked terms. This step resulted in the first effective contribution of the research.

The semantic annotation of the Open American National *Corpus* according to the top-level of the Schema.org ontology is one of the concrete contributions of this work. After the annotation phase mentioned in the previous paragraph, the *corpus* became the first general domain *corpus* to receive semantic labels from a top-level ontology. The *corpus* was further formatted and corrected and later made freely available on the internet at https://goo.gl/5AXD8n. The database is structured in such a way that it allows the development of new research in the area, contributing to the advancement of semantic annotation based on top-level ontologies.

Once the *corpus* was annotated and sufficient resources were left to continue the research, the second phase began. The second phase corresponded to the implementation of the actual objective of the project. The *corpus* was divided into portions of databases and each one was subdivided into two sets, training and testing. Later the CRF learning method was executed under the data sets in order to produce a sufficiently precise model to meet the expectations of the proposal. Numerous feature sets were used until the best model was obtained. From the best set, and using 100% of the database, an overall average of 93.5% F1-score was obtained, a relatively high value and subject to credibility. The results obtained from the first tests were enough to test the hypothesis and reach the general objective.

In addition, new trainings were performed this time using hyperparamter optimization and cross validation on the CRF. Such techniques are used in order to optimize the model and produce better results and with a higher reliability rate. However, to execute the method under these conditions requires high computational power and time. These features were limited, and unfortunately it was impracticable to perform the experiments with 100% of the database. Nevertheless, the analyzes were possible to be executed with 50% of the *corpus* and produced very promising results. The obtained model reached a general rate of 92.3% of f1-score. This result proves that the use of such techniques can add even more value to the model contributing to better results. Although the resources had been limited, this factor was not a great obstacle to prove the effectiveness of the technique.

The main contribution of this research was the implementation of a model ca-

pable of semantic annotating American English texts under the guidance of top-level ontologies. Although it is a new semantic annotation approach, its application can contribute substantially to the semantic enrichment of texts. The use of ontologies in the annotation process adds semantic value to the terms, since the ontological types already carry with them a whole conceptualization, a unique feature of the ontologies. Moreover, the use of top-level ontologies has a more generic level of abstraction, which captures a larger number of terms and classifies them. The approach through the use of top-level ontologies can support the initiation of semantic enrichment research based on general domain ontologies and be the starting point for further advancement in the field.

The semantic enrichment of texts brings countless benefits to computer interpretation. Content written in a natural language and enriched semantically via metadata are the fuel for the machine understanding. Semantic Web can take advantage of this type of content, for example, read it, understand it and establish new inferences more productive and precise. Another useful application is the use of enriched documents to generate new contents from the use of machine learning techniques. New annotated *corpus* could be of great value for the repercussion of new resources to subsidize the understanding by the machines. From this perspective, the work described in this dissertation, although it is a new and recent approach, presents promising results and opens a range of opportunities for new experiments.

Many different theories, proposals, and experiments have been left for the future due to limitations of resources and time and the scope of the research was sufficient to test the proposed hypothesis. However, it is important to highlight the ways to direct new research possibilities. The first important aspect to be analyzed is to use more computational resources to extend what could not be done in this work. New tests using higher computational power could bring even better results.

The annotated *corpus* constituted in this work was only a sample for conducting the research. The idea and approach used can help generate many other datasets in order to increase the resources available for use in learning methods. A larger set of data could provide the use of more robust machine learning techniques such as deep learning methods. These techniques require more data for training, that is, an even larger annotated *corpus*, and higher computational resources, but it is a possibility to be tested.

Finally, one last approach to be explored is the use of ontologies. New paths are to use the top-level of other general domain ontologies that do not specify only the most accessed web content. Other ontologies have different perspectives, and can contribute significantly to the semantic enrichment of texts. Or, following the same

line of this research, one possibility would be to deepen into more particular levels of Schema.org in order to increase the level of specificity of the annotated content. This approach could be applied either to the prediction model for automatic annotation, or to the manually or rules-based *corpora* annotation.

Bibliography

Alec, C., Reynaud-Delaître, C., and Safar, B. (2016). An ontology-driven approach for semantic annotation of documents with specific concepts. In *International Semantic Web Conference*, pages 609--624. Springer.

Andrade, G. C., Oliveira, A. d. P., and Moreira, A. (2017). A rule-based semantic annotator: Adding top-level ontology tags. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 53--62.

Andrade, G. C., Oliveira, A. d. P., and Moreira, A. (2018). Hybrid semantic annotation: Rule-based and manual annotation of the open american national corpus with a top-level ontology. *Manuscript submitted for publication at Language Resources and Evaluation*.

Asooja, K., Bordea, G., and Buitelaar, P. (2016). Using semantic frames for automatic annotation of regulatory texts. In *International Conference on Applications of Natural Language to Information Systems*, pages 384--391. Springer.

Bergamaschi, S., Cappelli, A., Circiello, A., and Varone, M. (2017). Conditional random fields with semantic enhancement for named-entity recognition. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, page 28. ACM.

Berners-Lee, T. and Fischetti, M. (2001). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. DIANE Publishing Company.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific american*, 284(5):34--43.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1--22.

Boella, G., Di Caro, L., Ruggeri, A., and Robaldo, L. (2014). Learning from syntax generalizations for automatic semantic annotation. *Journal of Intelligent Information Systems*, 43(2):231--246.

Cohen, K. B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. E. (2017). The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. In *Handbook of Linguistic Annotation*, pages 1379--1394. Springer.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 15--30. Springer.

Erdmann, M., Maedche, A., Schnurr, H.-P., and Staab, S. (2000). From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, pages 79--85. ACL.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363--370. Association for Computational Linguistics.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199--220.

Guarino, N. (1997). Some organizing principles for a unified top-level ontology. In *AAAI Spring Symposium on Ontological Engineering*, pages 57--63. AAAI Press Menlo Park.

Guarino, N. (1998). Formal ontology and information systems. In *Proceedings of FOIS*, volume 98, pages 81--97.

Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema. org: Evolution of structured data on the web. *Communications of the ACM*, 59(2):44--51.

Handschuh, S. and Staab, S. (2003). *Annotation for the semantic web*, volume 96. IOS Press.

Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1--136.

Ide, N. and Suderman, K. (2004). The american national corpus first release. In *LREC*. Citeseer.

Khalili, A. and Auer, S. (2015). Wysiwym–integrated visualization, exploration and authoring of semantically enriched un-structured content. *Semantic Web*, 6(3):259--275.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49--79.

Klien, E. (2007). A rule-based strategy for the semantic annotation of geodata. *Transactions in GIS*, 11(3):437--452.

Krishnan, V. and Manning, C. D. (2006). An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1121--1128. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282--289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Leech, G. (1997). *Introducing corpus annotation*. Addison Wesley Longman.

Liu, K. and El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Automation in Construction*, 81:313--327.

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Springer, New York, NY, 1 edition.

Maedche, A. (2012). *Ontology learning for the semantic web*, volume 665. Springer Science & Business Media.

Mitkov, R. (2005). *The Oxford handbook of computational linguistics*. Oxford University Press.

Moreira, A., Lisboa-Filho, J., and Oliveira, A. P. (2016). Automatic ontology generation for the power industry the term extraction step. In *Proceedings of the*

21 International Conference on Applications of Natural Language to Information Systems, pages 415--420. Springer.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2--9. Association for Computing Machinery.

Pandolfo, L. and Pulina, L. (2017). Ad n oto: A self-adaptive system for automatic ontology-based annotation of unstructured documents. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 495--501. Springer.

Patel-Schneider, P. F. (2014). Analyzing schema. org. In *International Semantic Web Conference*, pages 261--276. Springer.

Pease, A., Niles, I., and Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28.

Pham, M., Alse, S., Knoblock, C. A., and Szekely, P. (2016). Semantic labeling: a domain-independent approach. In *International Semantic Web Conference*, pages 446--462. Springer.

Pustejovsky, J. and Stubbs, A. (2012). *Natural language annotation for machine learning*. O'Reilly Media, Inc.

Reeve, L. and Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1634--1638. ACM.

Ronallo, J. (2012). Html5 microdata and schema. org. *Code4Lib Journal*, 16.

Sardinha, T. B. (2000). Lingüística de corpus: histórico e problemática. *Delta*, 16(2):323--367.

Şimşek, U., Kärle, E., Holzknecht, O., and Fensel, D. (2017). Domain specific semantic validation of schema. org annotations. In *International Andrei Ershov Memorial Conference on Perspectives of System Informatics*, pages 417--429. Springer.

Skeppstedt, M., Kvist, M., Nilsson, G. H., and Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *Journal of biomedical informatics*, 49:148--158.

Tran, T., Phung, D., Bui, H., and Venkatesh, S. (2017). Hierarchical semi-markov conditional random fields for deep recursive sequential data. *Artificial Intelligence*, 246:53--85.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: science, services and agents on the World Wide Web*, 4(1):14--28.

Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(02):93--136.

Wilson, A. and Thomas, J. (1997). Semantic annotation. In Garside, R., Leech, G. N., and McEnery, T., editors, *Corpus annotation: linguistic information from computer text corpora*, chapter 1, pages 53–65. Longman, London, UK, 1 edition.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301--320.

Appendix A

Hybrid annotation proposal

The Figure A.1 shows the flow for the development of rules-based annotator. The left side of the flow display the required inputs, the ontology, the *corpus*, and the NLP techniques. As can be seen in the figure, and because it is still the first approach, the ontology assigned to the extraction of its top-level is SUMO. The *corpus* accompanied by the annotations referring to the syntax, morphology and structure of sentences and words is the Open American National *Corpus*. Natural language processing techniques correspond to the use of the structure of sentences and words, as well as their linguistic attributes for the construction of rules. The joining of the three input elements constitute the set of rules established for annotation of the *corpus*. Finally the set of rules created in the previous phase results in the rules-based annotator.

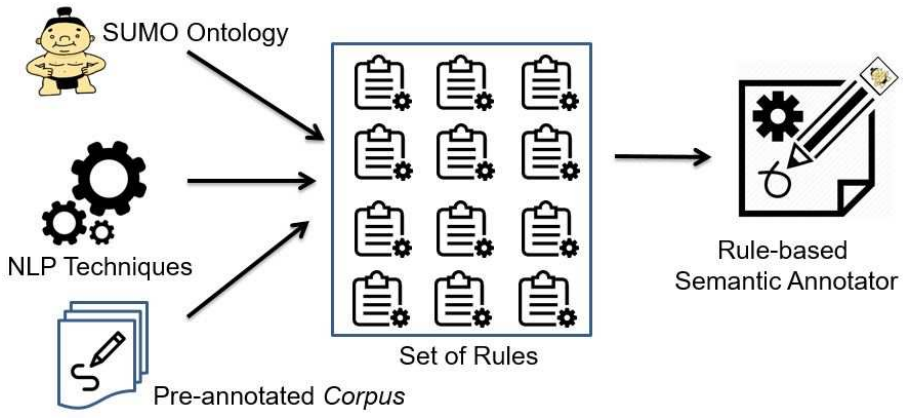

Figure A.1: Development of a rule-based semantic annotator using the SUMO ontology.

The Figure A.2 presents the flow for annotation do *corpus* using the hybrid

approach. The Open American National *Corpus*, pre-formatted and corrected is submitted to an adding of labels from the top-level of the SUMO ontology. The annotation corresponds to the application of two approaches to the *corpus*. First, the rules-based annotator adds the labels to the terms whose characteristics match with their ontological type. Subsequently, it is performed the manual annotation in order to further enrich the *corpus* and correct possible errors of the previous phase. At the end of the two phases of annotation hybrid the result generated is the OANC plus a semantic annotation according to the top-level of the SUMO ontology.

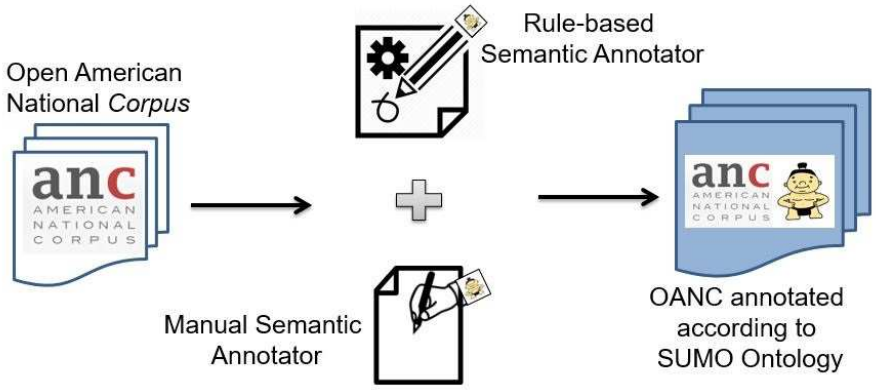

Figure A.2: Annotation of the Open American National Corpus using the hybrid proposal.

Appendix B

Rules

The rules-based annotator developed in the research consists of a set of rules that assign labels referring to the Schema.org ontology to the selected terms. The rules use the characteristics derived from the pre-annotated corpus to label the terms. Each rule describes a single ontological type, and each term only receives a single label. The appendix provides examples of rules used to annotate the corpus and each session describes examples used for each ontological type. The rules are encoded in Python language, but simplified for better understanding. All of the example codes present some errors related to logic or even syntax, but do not stick to that detail. The figure B.1 shows the structure of the corpus to be manipulated by the rules. The following subsections present the continuation of the code, but with the rules of each type from Schema.org top-level.

```
# The part of the code responsible for assignment of the rules and reading
# the corpus was removed.

sentence = {} #Each sentence of the corpus
sentence = {0: ['word1',['attr1','attr2', 'attr3']],
            1: ['word2',['attr1','attr2', 'attr3']],
            2: ['word2',['attr1','attr2', 'attr3']],
            3: ['word2',['attr1','attr2', 'attr3']]}

# Assingment funtion
# x is the current sentence
# y is the token that will receive the label
assingment(x,y)

for i in range (0,Len(sentence)):

    token = sentence[i] # Each token that will be analized plus its attributes
    word = token[0] # The current word
    attr = token[1] # The attributes from the current token

    tokenPos = sentence[i+1] # Posterior token
    wordPos = token[0] # Posterior word
    attrPos = token[1] # Posterior attributes

    tokenPre = sentence[i+1] # Previous token
    wordPre = token[0] # Previous word
    attrPre = token[1] # Previous attributes

    #Rules....

```

Figure B.1: Initial code

B.1 Action

```
29    # Rules....
30
31    # File containdg non action verbs.
32    file = open ("ClassFiles\\NonAction.txt", 'r', encoding="utf-8")
33
34    if attr[0] == 'affix="ing"' and attr[2] == 'mds="NN"':
35        for nonaction in file:
36            if word.upper() == nonaction[:len(word)-1]:
37                assingment(sentence, token)
38
39    if attr[0] == 'affix="ing"' and
40    (attr[2] == 'mds="VBG"' or attr[2] == 'mds="VBN"'):
41        for nonaction in file:
42            if word.upper() == nonaction[:len(word)-1]:
43                assingment(sentence, token)
44
45    if wordPre[0].upper() == "AM" and attr[0] == 'affix="ing"':
46        for nonaction in file:
47            if word.upper() == nonaction[:len(word)-1]:
48                assingment(sentence, token)
49
50    if wordPre[0].upper() == "IS" and attr[0] == 'affix="ing"':
51        for nonaction in file:
52            if word.upper() == nonaction[:len(word)-1]:
53                assingment(sentence, token)
54
55    if wordPre[0].upper() == "BE" and attr[0] == 'affix="ing"':
56        for nonaction in file:
57            if word.upper() == nonaction[:len(word)-1]:
58                assingment(sentence, token)
59
60    if wordPre[0].upper() == "WERE" and attr[0] == 'affix="ing"':
61        for nonaction in file:
62            if word.upper() == nonaction[:len(word)-1]:
63                assingment(sentence, token)
64
65    # Continue...
```

Figure B.2: Action Rules

B.2 Creative Work

```
29    # Rules....
30
31    if wordPre == 'BOOK' and word.isupper() and
32    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
33        assingment(sentence, token)
34
35    if wordPre == 'PAINTING' and word.isupper() and
36    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
37        assingment(sentence, token)
38
39    if wordPre == 'ARTICLE' and word.isupper() and
40    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
41        assingment(sentence, token)
42
43    if wordPre == 'MOVIE' and word.isupper() and
44    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
45        assingment(sentence, token)
46
47    if word == 'PHOTOGRAPH':
48        assingment(sentence, token)
49
50    if word == 'REPORT' and attr[2] == 'mds="NN"':
51        assingment(sentence, token)
52
53    if wordPos == 'MAP' and word.isupper() and
54    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
55        assingment(sentence, token)
56
57    # Continue...
```

Figure B.3: Creative Work Rules

B.3 Event

```
    # Rules....

    if wordPos == 'CONFERENCE' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'CONGRESS' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'SYMPOSIUM' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPre == 'CEREMONY' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if word == 'HOLIDAY':
        assingment(sentence, token)

    if word == 'SPETACLE':
        assingment(sentence, token)

    if word == 'BIRTHDAY':
        assingment(sentence, token)

    if wordPre == 'PUBLIC' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[3] == 'schema="EVENT"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPos[3] == 'schema="EVENT"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    # Continue...
```

Figure B.4: Event Rules

B.4 Intangible

```
    # Rules....

    if attr[0] == 'base="life"' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if attr[0] == 'base="love"' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if attr[0] == 'base="peace"' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if attr[0] == 'base="fear"' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if attr[0] == 'base="hate"' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if attr[0] == 'base="idea"' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    # Continue....
```

Figure B.5: Intangible Rules

B.5 Organization

```
    # Rules....

    if attr[2] == 'ne="ORGANIZATION"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'CORPORATION' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'ORGANIZATION' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPre == 'PUBLIC' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'ASSOCIATION' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'COMPANY' and word.isupper():
        assingment(sentence, token)

    if attrPre[3] == 'schema="ORGANIZATION"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPos[3] == 'schema="ORGANIZATION"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    # Continue...
```

Figure B.6: Organization Rules

B.6 Person

```
    # Rules....

    if attr[2] == 'ne="PERSON"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPre == 'Mr.' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPre == 'Mrs.' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[2] == 'ne="PERSON"' and word == "A." and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[2] == 'ne="PERSON"' and word == "S." and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'ORGANIZATION' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[2] == 'ne="PERSON"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPos[2] == 'ne="PERSON"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[3] == 'schema="PERSON"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPos[3] == 'schema="PERSON"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'SMITH' and word.isupper() and attr[2] == 'mds="NNP"':
        assingment(sentence, token)

    # Continue...
```

Figure B.7: Person Rules

B.7 Place

```
    # Rules....

    if attr[2] == 'ne="LOCATION"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPre == 'at' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[2] == 'ne="LOCATION"' and wordPre == 'in' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'LAKE' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'PARK' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[2] == 'ne="LOCATION"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPos[2] == 'ne="LOCATION"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPre[3] == 'schema="PLACE"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if attrPos[3] == 'schema="PLACE"' and word.isupper() and
    (attr[2] == 'mds="NNP"' or attr[2] == 'mds="NNPS"'):
        assingment(sentence, token)

    if wordPos == 'USA' and word.isupper():
        assingment(sentence, token)

    if wordPos == 'ILLINOIS' and word.isupper():
        assingment(sentence, token)

    # Continue...
```

Figure B.8: Place Rules

B.8 Product

```
    # Rules....

    tokenPre2 = sentence[i+1] # 2 Previous token
    wordPre2 = token[0] # 2 Previous word

    if wordPre2 == 'bought' and wordPre2 == 'a' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if wordPre2 == 'sell' and wordPre2 == 'the' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if wordPre2 == 'buy' and wordPre2 == 'an' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if wordPre2 == 'sold' and wordPre2 == 'many' and
    (attr[2] == 'mds="NN"' or attr[2] == 'mds="NNS"'):
        assingment(sentence, token)

    if wordPos == 'CAR' and word.isupper():
        assingment(sentence, token)

    if wordPos == 'COMPUTER' and word.isupper():
        assingment(sentence, token)

    # Continue...
```

Figure B.9: Product Rules