information

MDPI

Article

Ontological Semantic Annotation of an English Corpus Through Condition Random Fields

Guidson Coelho de Andrade †, Alcione de Paiva Oliveira *,† and Alexandra Moreira †

Departamento de Informática—Centro de Ciências Exatas e Tecnológicas, Universidade Federal de Vicosa, Viçosa MG 36570-900, Brazil; guidson.c.andrade@gmail.com (G.C.d.A.); xandramoreira@gmail.com (A.M.)

* Correspondence: alcione@dpi.ufv.br; Tel.: +55-31-3899-2396

† These authors contributed equally to this work.

Received: 27 February 2019; Accepted: 30 April 2019; Published: 9 May 2019

Abstract: One way to increase the understanding of texts by machines is through adding semantic information to lexical items by including metadata tags, a process also called semantic annotation. There are several semantic aspects that can be added to the words, among them the information about the nature of the concept denoted through the association with a category of an ontology. 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. Considering that annotation is an arduous task that demands time and specialized personnel to perform it, much is done on ways to implement the semantic annotation automatically. The use of machine learning techniques are the most effective approaches 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 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.

Keywords: information extraction; semantic annotation; ontology; condition random fields

1. Introduction

In order to develop natural language processing systems, it is necessary to construct language resources that capture as much as possible the linguistic diversity present in a natural language. These natural language resources, called *corpus*, are usually provided with meta-data containing information about the tokens and the documents the forms up the *corpus* [1]. The addition of meta-data to a *corpus* is called annotation or labelling. It is the process of adding information to plain textual data. Annotations that aggregate information to a *corpus* can be applied to a document as a whole, to its sentences, its terms and words and can be performed manually or automatically [2]. Annotations can facilitate the development of various types of applications related to the understanding of natural language, ranging from information extractors to automatic language translators. The reason for this is that annotations (syntactic and/or semantic) provide additional information that help establish the context of the statement where the lexical item is inserted and helps eliminate ambiguities.

Corpus annotation may be applied to the various levels of the linguistic structures. So annotations can express the grammar class of the annotated elements (Part-of-Speech), their morphology, correlation phenomena, the aspects of phonetics and so on [2]. The annotation may also cover

other aspects related to the structure of the annotated text and its content as a whole. There are several aspects of a lexical item that can be annotated, but we can basically divide it into syntactic and semantic aspects. Syntactic annotation aims to add information related to the form of the lexical item, such as its part-of-speech tagging or its dictionary form (lemma form). On the other hand, semantic annotation is the process in which are added to the terms significant references to express their meaning. The annotation task attempts to capture the essence of the meaning of the tagged object [3]. In general, the main goal of semantic annotation is to make texts capable of being understood by machines. In the case of semantic annotation, a set of labels whose meaning has already been formally defined or is well-known 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 [4]. 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 a set of words [5]. 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. The main contribution of semantic annotation is to eliminate ambiguities regarding the meaning of words by computer devices. Establishing the meaning of a lexical item is still a challenging task due to its polysemic character [6]. For example, the word " bank ", according to Wordnet [7], has ten meanings as a noun and eight meanings as a verb. Some meanings are listed below:

- Noun

    1. S: (n) bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank";
    2. S: (n) depository financial institution, bank, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) "he cashed a check at the bank";
    3. S: (n) bank (a supply or stock held in reserve for future use (especially in emergencies)).

- Verb

    1. S: (v) bank (do business with a bank or keep an account at a bank) "Where do you bank in this town?"
    2. S: (v) bank (cover with ashes so to control the rate of burning) "bank a fire"
    3. S: (v) count, bet, depend, swear, rely, bank, look, calculate, reckon (have faith or confidence in) "you can count on me to help you any time".

Polysemy is a complex area of study within linguistics and has divergent theories formulated by linguistics classical and cognitive linguistics [8,9]. Regardless of the underlying theory, in order to establish the correct meaning of a lexical item it is necessary to know the context where it occurs, that is, the words that occur in its neighborhood. According to [10] *apud* [11], " You shall know a word by the company it keeps. " In the sentence "he cashed a check at the bank" it is possible to infer that the word "bank" refers to a banking institution because of the co-occurrence of the words "sit" and "check".

The semantic annotation, which is the focus of this article, can be further divided. Pustejovsky and Stubbs [12] divide the semantic annotation into annotation of semantic roles and annotation of semantic types. They say "we can distinguish two kinds of annotation for semantic content within a sentence: what something is, and what role something plays." In Semantic Typing annotation a language structure is labeled with a type identifier, from a reserved vocabulary or ontology, indicating what it denotes, whereas in Semantic Role Labeling a language structure is identified as playing a specific semantic role relative to a role assigner, such as a verb [12].

As one of the branches of philosophy, the term ontology is related to the study of many things that exist in the world, and its main function is the organization of what exists in a set of categories.

In the words of [13], an ontology, from a Computer Science perspective, is a specification of a conceptualization. In this sense, the ontology tries to describe the concepts existing in a domain and to relate them according to their characteristics [14]. As a Computer Science object, 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 to which they propose to specify, they may adequately be used to define part of the meaning of a lexical item.

When using ontologies in the annotation process, the ontological classes become the labels and the contents specify which objects should be annotated [15]. The task of annotation is closely related to the ontology domain. There are several types of ontology, which are classified according to their function and their scope. Generally, in the semantic annotation task, when aimed at annotating texts in a specific domain, one should use classes from an ontology corresponding to that domain. On the other hand, if the annotation comprehends broader concepts, a more generic ontology that can incorporate a vast number of domains should be used. General domain ontologies, also called top-level ontologies, are extremely extensive and can define concepts applicable to any domain. Top level ontologies describe broader and more abstract concepts regardless of a particular problem or particular domain [16]. Due to its expressiveness and to the large number of classes that compose such ontologies, commonly only fragments are selected from them, usually the top-level, to form the label set and then to perform the annotation task. Annotating text with the concepts of a top-level ontology can be the starting point for deepening semantic annotation at more specific levels of a general ontology or a domain ontology.

In addition to producing semantically annotated *corpus*, semantic annotation based on top-level ontologies can also be useful for the enrichment of Web content. One of the main applications of semantic annotation on the Internet sphere is the contribution that it can offer to the Semantic Web [5]. Semantic Web 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 [17]. 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 [18]. Attaining this goal would facilitate, for instance, search engines in providing a response to users of what is most relevant to their needs. Ontology vocabulary in an important element and valuable tool for organizing the data of a domain and enriching it by adding meaning. In this sense, semantic annotation based on ontologies plays a fundamental role in the process of semantic enrichment of web content to support the Semantic Web [19].

Moreover, semantic annotation, particularly annotation based on ontologies, can help improve results of other applications that currently rely more on syntactic information and word relationships. This is the case of the applications addressed by [20,21]. Both papers deal with the classification of documents based on information extracted from the style of writing. In the first work the author tries to find out whether a scientific article was written by an automatic text generator or not, while the second work seeks to identify the authorship of a text. Both researches produced expressive results in their tasks, with accuracy greater than 88%, but it would be interesting to see if semantically annotated texts can improve these numbers. Looking beyond the analysis of feelings, Preoţiuc-Pietro et al. [22] presented the results of their research that aimed to predict the political ideology of tweeters from the analysis of their posts. To carry out the predictions they used as language features, Unigrams, Linguistic Inquiry and Word Count (LIWC), Word2Vec clustering, Political Terms, and words associated with six emotions. The results showed that Word2Vec clusters obtain the highest predictive accuracy for political leaning, and for political engagement, political Terms and Word2Vec clusters obtain similar predictive accuracy. In works such as this, the joining of words with their ontological types before the construction of the model has the potential to generate results with greater accuracy. Liu et al. [23] investigated the feasibility of career path prediction from social network data. The approach they proposed was a multi-source learning framework with a fused lasso penalty (MSLFL), thus the predicted results from individual sources should be the same or similar, otherwise a penalty should

take place. As the model fuses information distributed over multiple social networks to characterize users from multiple views, it could benefit from semantically annotated information to make a more appropriate merging. Estival et al. [24] developed a project in which ontologies are part of the reasoning process used for information management and for the presentation of information. According to the authors, "*users access to information and the presentation of information to users are both mediated via natural language, and the ontologies used in the reasoning component are coupled with the lexicon used in the natural language component*". We believe that the system can be even more efficient if the information base had previously been annotated with ontological tags.

Semantic annotations are valuable and help many types of NLP applications, however, according to [25], semantic annotation is an extremely time and resource consuming task. 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 could provide a solution [12]. So, in order to optimize the task and decrease such complexity, researchers from the NLP area uses methods that learn from previous annotated *corpora* using machine learning techniques. The learning algorithms have the ability to, after training under the use of previously annotated *corpus*, perform the annotation of new text documents. Nonetheless, to make use of automatic annotation techniques via learning algorithms training material is needed and there is a shortage of annotated *corpora* for this task [15]. 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 semantic annotation serve as motivation for doing research in the area. The main objective of this work is to make use of a machine learning method to perform the semantic annotation based on top-level ontologies of an American English *corpus*. Specifically, we have constructed a model capable of classifying the selected top-level ontology types with a satisfactory prediction rate to apply it in the semantic annotation task. This paper is organized as follows. The next section gives an overview of the works that have a relation with this research, presenting the advances and highlighting the points that can be improved. Section 3 are divided into three subsections. Schema.org ontology subsection presents the process of ontology selection as well as its characterization and definition as the top-level ontology responsible for generating the classification labels. The subsection Corpus, introduces more details of the *corpus* adopted. The CRF approach subsection describes the chosen classification model, and the preprocessing stages, and the classification process. The results achieved in the classification stage, and a discussion about them are presented on the Section 4, and finally the conclusions are presented at Section 5.

2. Related Works

Automatic semantic annotation based on top-level ontologies is a recent research area, becoming feasible by novel advances in hardware architectures and, therefore, there are few papers available in the literature for comparison. In this section, we present an overview of related work concerning the semantic annotation of texts, even though some do not specifically address ontological knowledge. The following papers can be divided according to the type of annotation in three different groups: semantic role annotation, named entity recognition and ontology-based annotation. Although the first two groups have different definition and applicability, they were considered because they use similar techniques and share similar challenges. The third group has the same goals of the work presented in this paper.

Semantic role labeling (SRL) is the task of identifying the semantic arguments of a predicate and labeling them with their semantic roles [26]. You can distinguish annotation based on macro-roles such as *agent* and *patient* or micro-roles can be adopted such as those defined by the frame semantic theory [27]. FitzGerald et al. [26] proposed a method for semantic role annotation in which arguments and semantic roles were jointly embedded in a shared vector space for a given predicate.

The embedding was produced by a neural network model. Training their model jointly on both FrameNet and PropBank data, they achieved the best result to date on the FrameNet test set.

Named Entity Recognition (NER) holds a certain relation to ontological annotation since the tags used in NER such as PEOPLE, ORGANIZATION and REGION are a subset of the ontological categories of some top-level ontologies. Hence, NER can be considered a particular case of ontological annotation. The work presented by [28] describes the use of the Conditional Random Field algorithm for named entity recognition. Their work comes close to ours since, in addition to recognize entities, they add semantic features before performing the annotation. This approach differs from the usual one to NER and brings better results, augmenting the semantic information supplied to the model. The recognition method was performed combining standard training features and semantic information gathered from the Cogito linguistic analysis engine (http://www.expertsystem.com/). 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 NER corpus (https://www.clips.uantwerpen.be/conll2003/ner/) that was manually annotated using five categories, PER, LOC, ORG, MISC and O. Throughout the experiment the authors compared results obtained with and without the use of the semantics. They also employed *corpus* of different sizes to analyze the performance. According to the authors, the results were considerably better when compared to the usual approach. Without semantics, they obtained an average of 0.8507 for precision, 0.8188 for recall and 0.8336 for F1-measure. Adding the semantic information they obtained an average of 0.8629 for precision, 0.8392 for recall and 0.8505 for F1-measure. Hence, the research showed that by combining semantic information with training features resulted in a positive effect on the outcome of the NER task.

Skeppstedt et al. [29] proposed the 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 quite a similar work considering the annotation process. The procedure aimed to recognize clinical entities from medical texts written in Swedish. It is common to annotate Health records with those classes in order to assist the patient's analysis and medical hypothesis construction. The contribution of the proposal is to aid in medical knowledge extraction in a language distinct from English. Because of that, it was done a comparative study to figure out how well clinical entities previously annotated in English are recognized in the Swedish clinical *corpus*. The main reason why this research was selected as a related work is because of its automatic annotation approach performed by the same machine learning algorithm. After the *corpus* selection, training, and test set distribution, the CRF algorithm was applied to annotate using the four selected categories. The results produced by the algorithm using the best features, settings and its ability to generalize to held-out data was an F-score of 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.

The work proposed by [30] focused on ontological annotation. The study came up with a self-adaptive system for automatic ontology-based annotation of unstructured documents in the context of digital libraries. Different from our proposal, this work has an approach of annotating an entire document over an ontological perspective and not the terms, being the use of ontologies what correlate both works. The authors aim to create a system capable to automate the ontological-based annotation process of texts from digital libraries. The work is based on the STOLE [31], an ontology-based digital library created from documents about the history of public administration in Italy in 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 preprocessing phase was applied to the *corpus* providing necessary information to build the features, such as sentence boundary, part-of-speech, named entity recognition. The system used 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 had precision of 0.80, recall of 0.53, F-measure of 0.63. Although the results are considerably 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 used an ontology to annotate terms related to a domain is described in the article of [32]. 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 had as its focus the extraction of information about bridges maintenance and deficiencies, naming entities related to the theme. As an object of analysis, eleven bridge inspection reports were used which had 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 age, rendering a *corpus* appropriate for the proposed technique. The authors carried out a preprocessing phase, making the documents readable by the application. Subsequently, the feature extraction phase of was accomplished, taking into account aspects related to the part of speech, stem, and semantic characteristics. Finally, came the last step corresponding to the extraction of information through the semi-supervised conditional random fields. In the evaluation phase, they took into account 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 tags in the extraction process, so the ontology assisted 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 in being based on the use of ontologies to annotate text and for using CRF as a learning model.

3. Materials and Methods

This section describes the techniques used in the research. To contextualize the techniques used, some concepts will be described in the following subsections as well. The first subsection describes the ontology used as well as the selection of its top-level concepts for the constitution of the eight labels used in the annotation process. Afterward, the *corpus* used in the experiments is presented and the grounds for its choice are also made clear. Finally, the last subsection presents the training algorithm and the reason for its choice.

3.1. Schema.org Ontology

Before describing the use of Schema.org ontology, it is necessary to introduce the concept of Linked Data. According to [33] 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 cover a multitude of domains, allowing the development of a variety new applications. The connection established between the data allows navigation along the links, optimizing tasks such as search, and more appropriate query results according to content sought [34]. In this sense, all data published on the web from different sources, connected somehow with internal and external dataset, with meaning explicitly defined and machine-readable is characterized as Linked Data. Liked Data relies on files containing information expressed in RDF (Resource Description Framework) format to record typed statements that link distinctive data sources [33]. The main goal of feeding the Web with this huge amount of data is to aid the understanding of Web content by machines, giving rise to the so called Semantic Web. The Semantic Web is defined as a web environment where machines directly or indirectly process the data and understand it [18].

A practical example of the utilization of Liked Data principles to construct a set of connected data is the Linking Open Data (https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData) project. The project's main purpose is to aid applications and researches devised for the Semantic Web by collecting open web data, converting to RDF format and finally publishing it

on the Web [33]. 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 charge. When a 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 [34]. 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 the 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 among others. Schema.org is an initiative to create a collaborative project that supports Linked Open Data by creating structured data schemas [35]. The goal behind the project was to offer a single and integrated approach capable of covering a wide range of frequent web topics and to propose a structured vocabulary for web page marking. The vocabulary is organized in 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 [36]. The version used by this is the core Vocabulary (http://schema.org/docs/full.html) composed of 606 types so far, according to the full hierarchy shown on its website.

Based on the use of Web content, the proposed types were organized in a 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 [35]. However, in order to make the supertype/type relation more natural, a given type may have more than one parent type, although this is not common. According to [36] the relations are polymorphic in the sense they have one or more domains and one or more ranges. Schema.org can be seen as an ontology where the terms are based on evidence taken from the Web. As reported by [37], Schema.org is defined as a "middle-level ontology", it means that its purpose is not to cover all the existing things in the world nor to describe to a specific domain. He further adds that the main intention of middle-level ontologies, such as Schema.org, is to produce such a broad scope capable of covering the most frequent cases of a general domain. Based on this point of view the ontological classes presented by Schema.org are used to markup web pages, and can also be easily applied text documents adding semantic content to them. Because of that the Schema.org was chosen to be the ontology that provided the classes for the automatic annotation phase of this research. However, to enable the use of a subset of the ontology classes in the annotation process, the first step is to establish the level of the hierarchy of concepts that will be used.

In our research, the categories used for the annotation were those located at the top level of the Schema.org top-level. The ontology most generic concept is denoted by the term "Thing" (http://schema.org/Thing). This category is divided into eight major categories that have been defined from textual evidence. These eight categories are those that have been adopted for annotation. The definitions of each category are explained below.

- Action: An action executed by a direct or indirect agent(s) upon a direct object producing a result (http://schema.org/Action);
- Creative Work: The most generic kind of creative work, including books, movies, photographs, software programs, etc. (http://schema.org/CreativeWork);
- Event: An event happening at a certain time and location (http://schema.org/Event);
- Intangible: Object That can not be tanger (http://schema.org/Intangible);
- Organization: Institution that is intended to carry out acts in the various spheres of society (http://schema.org/Organization);
- Person: A person (alive, dead, undead, or fictional) (http://schema.org/Person);
- Place: Entities that have a somewhat fixed, physical extension (http://schema.org/Place);
- Product: Any offered product or service (http://schema.org/Product).

3.2. The Corpus

In order to be able to perform the automatic annotation proposed by this research it was necessary to select a *corpus* with some specific requirements such as lexical coverage and dimension. The *corpus* chosen was the one proposed by [38], though not the original one, but the one that has been altered and standardized by [39]. The Open American National Corpus (OANC) is an American English *corpus* available free, and that comprises texts and oral conversations of several categories such as fiction, documents, scientific articles among others [38]. The main goal of the *corpus* is to gather different text sources to achieve the most comprehensive linguistic diversity as possible and offer a valuable resource for researchers in the Natural Language Processing area. The *corpus* is distributed through 8293 files from different sources all under the same XML format standard. Each document is accompanied by annotations of 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* were performed automatically through annotation tools, which generated errors of different types. The errors do not compromise the development of applications. However, they negatively influenced the accuracy of the results obtained by these applications.

The contribution of the work done by [39] was the production of a new version of OANC corpus. First, the authors tried to eliminate the annotation errors of the *corpus*. After the corrections and some adjustments, the *corpus* exhibited a total of 16,280,694 tokens and 214,827 types. Subsequently, the documents were submitted to a standardization process in order to be suitable as input to the system. Finally, they carried out a hybrid approach annotation process in the *corpus*, aiming to label the lexemes with the Schema.org top-level categories. The first phase of this last step relies on a rule-based annotator that to labels terms of the *corpus* according to the selected classes. For each class was constructed a set of rules which analyzes the necessary condition to assign the Schema.org tag to a candidate term. Once this phase was completed, the annotation performed by the rule-based annotator was straightened out. Hence, the second phase consisted in inspecting the documents and correcting them manually. At the end of the task, Andrade [39] accomplished a total of 1,080,464 terms annotated under the eight Schema.org top-level ontology classes.

In the standardized *corpus* each word of a sentence occupied a single line, followed by its tags, and each sentence of a document was separated by a blank line. The *corpus* was divided in ten sets of data consisting of 1%, 10%, 20%, 30% and son on up to 100% of the total size of the original *corpus*. To ensure the lexical diversity of each set of data, the documents were carefully separated by categories. In this way, it was possible to ensure that each dataset was composed of a portion of texts from each category. Subsequently, each dataset was further divided into two subsets, the training set and the test set. The training set and test set were randomly organized containing 80% and 20% of the documents respectively.

3.3. The CRF Approach

The model used to train the datasets was the Conditional Random Fields [40], so before reporting the process it is important explain a bit about its operation. The model proposed by [40] is a probabilistic method based on conditional approach for segmenting and labeling sequence data. Although the CRF model is highly computationally complex at the training stage of the algorithm it has the advantage of combining the ability to compactly model multivariate data with the ability to leverage a large number of input features for prediction [41]. The model is trained to maximize the conditional probability of the outputs given the inputs. The method has the form of undirected graph model which, given a data sequence for training, it defines a single log-linear distribution over label sequences [40]. It has a high flexibility to integrate a large number of arbitrary and non-independent input resources, and this is one of its main advantages [42]. In relation to generative models, such as Hidden Markov Models, discriminative models like CRF do not attempt to model the joint probability distribution $p(y|x)$. In fact, they try to directly model $p(y|x)$ without wasting efforts on modeling the observations, which does not contribute for the goal $\hat{y} = argmax_y p(y|x)$. Also, it does not assume that the features

are independent of each other in order to ensure a treatable inference [40]. In relation to Maximum Entropy Markov Models, which are also discriminative, CRF avoids the so called Label Bias Problem, which generates a bias towards states with few successor states. Despite the advantages of the CRF approach when it comes to predicting sequences, it has the disadvantage of being computationally expensive in the training step. We chose the CRF approach since it allows us to consider contextual information through features involving distant observations, which is fundamental for annotations in natural language texts. Equation (1) shows the general CRF formula for sequence.

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{z(\mathbf{x})} \prod_{t=1}^{T} exp \left\{ \sum_{k=1}^{k} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\} \quad (1)$$

where $Z(\mathbf{x})$ is a sum of all possible state sequences such that the total becomes 1, in order to generate a probability score. This is the hard part to calculate. θ_k are the weights, and f_k are the features.

The Figure 1 graphically shows the main differences between the most common approaches to sequence tagging. The figure shows an attempt to establish the sequence of ontological classes for a sequence of words (observations), in this case the sentence "Olusegun Obasanjo emerged years military dictatorship". Figure 1a shows the Hidden Markov Model approach (HMM) where, as shown by the direction of the arrow, it tries to establish the probability of $p(y|x)$ from $p(x|y)$. In addition it operates locally, taking into account the current observation and the previous annotation. Such models try to model calculate the joint probability $p(x, y)$ which in turn might impose highly dependent features which are very often intractable to model and compute. In the case of Figure 1b it shows that in the Maximum Entropy Markov Model approach the probability of $p(y|x)$ is calculated directly from the observations, as shown by the direction of the arrow. In addition, it shows that the calculation of probability takes into account features over distant observations. Finally, Figure 1c shows the CRF approach where, as shown undirected graph, establishes a conditional random field where the random variables Y, conditioned on X obey the Markov property: the probability of y depends only on neighboring nodes. Also, in the same way as the MEMM approach, the calculation of probability may takes into account features over distant observations.

As previously mentioned, CRF has a high computational complexity in training time and, therefore, scale poorly for tasks with large numbers of states. According to [43] this is a consequence of inference having a time complexity which is at best quadratic in the number of states. However, maintaining the small number of labels controls, in part, this complexity. On the other hand, once trained, the model is relatively efficient and scalable at annotation time, having time complexity of $O(T|S|^2)$ (T is the size of the observations, whereas S is the number of labels. in case of using Viterbi algorithm for inference).

For the execution of the CRF model, it was used the sklearn-crfsuite (https://sklearn-crfsuite.readthedocs.io/en/latest/index.html) package, which is a wrapper over CRFsuite. The tool provides an interface similar to the one furnished by the scikit-learn, which allows to save/load CRF models using joblib or to use of scikit-learn model selection utilities ((cross-validation, hyper-parameter optimization). The sklearn-crfsuite has available five implementations of the CRF algorithm: *lbfgs* (Gradient descent using the L-BFGS method), *l2sgd* (Stochastic Gradient Descent with L2 regularization term), *ap* (Averaged Perceptron), *pa* (Passive Aggressive), and *arow* (Adaptive Regularization Of Weight Vector). All the implementations were tested, and the *lbfgs*, which is a gradient descent using the L-BFGS method, presented the best performance. Hence, it was chosen to run the experiments. L-BFGS stands for Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm [44], and is an optimization algorithm in the family of quasi-Newton methods designed to perform using a limited amount of computer memory. It also has the advantage of being fast when compared to other training algorithms. Sklearn-crfsuite also provides a set of typical machine learning metrics (accuracy, precision, recall, F1-measure) to validate the model and analyze its predictive performance. The package is written in Python and requires version 2.7 or above, 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, hyper-parameter optimization, retraining using best parameters, and finally learning analysis.

To define the features to be used, we analyzed the most relevant rules conceived in the work of [39]. The features were defined taking into account the sentence structure, surrounding words, syntax and word morphology, word shape, among others textual elements. After selecting the features, they were extracted from the training and test files to create the respective datasets. Once the extraction of the features was finished, the training phase was started. 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 [45]. After a certain training time, it is possible to carry out a pre-evaluation of the model through the metrics of the initial results. The results of the class "O" were removed so that they would not influence the results of the other classes that are of greater interest. Afterwards, aiming to obtain better results and to improve the quality of the model, it was performed a hyper-parameter optimization through randomized search over the parameters. During this process a 3-fold cross-validation was also applied. The value related to the cross-validation technique was defined empirically and considering the limitation of computational resources. Then, the results achieved by the model were again verified to check the improvements.

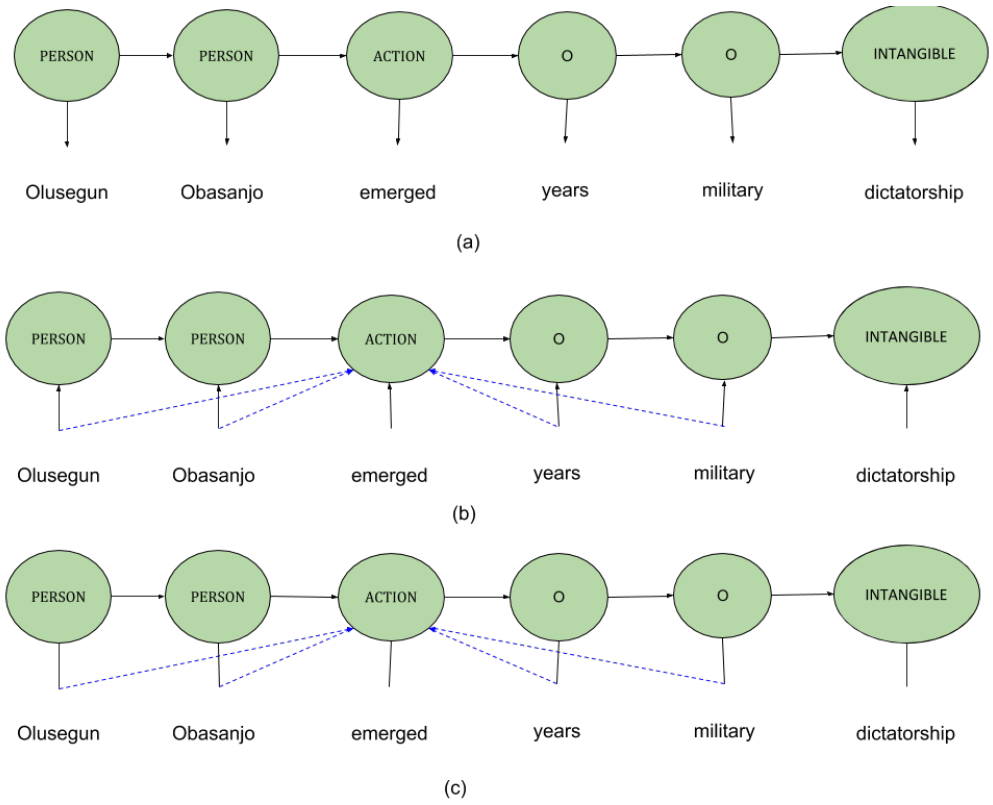

Figure 1. Differences between the main approaches for sequence tagging. (**a**) HMM: it tries to establish the probability of $p(y|x)$ from $p(x|y)$; (**b**) MEMM: the probability of $p(y|x)$ is calculated directly from the observations, as shown by the direction of the arrow. The dotted arrows implies that the calculation of probability takes into account features over distant observations; (**c**) CRF: represented by an undirected graph. It also takes into account features over distant observations.

4. Results and Discussion

This section presents the results obtained in the annotation phase using the CRF classifier. As mentioned earlier, the focus of the annotation phase was to tag lexemes using features over the nine tag classes. However, the results described here take into account only the eight classes that really

assign semantic value to the annotated words. The class O, standing for lexemes classified as OTHER, does not add semantic value to the words, so it is not relevant to the purpose of this research. Another important point to consider is that the *corpus* used as a training and test set does not have a balanced distribution between classes, which can lead to a biased weight distribution. Finally, we carried out two types of test. The first used a simple CRF configuration and the second used hyper-parameter optimization and cross-validation, both approaches applied to all datasets. Recalling that the *corpus* was divided into subsets that gradually increased in size to analyze the behavior of the model during the experiments.

The Table 1 presents the precision, recall, and F1-measure values for the eight annotated classes. The results relate to the execution performed with 100% of the *corpus* using the standard CRF model. The training phase comprised the total of 6997 documents, encompassing the various literary models stored in the *corpus*. After the training stage, the test set was submitted for evaluation, producing the described results. All classes showed satisfactory results in the annotation process obtaining a score higher than 85% in the F1-measure. The class that presented the best results was ACTION, probably due to the fact that this class comprises nouns that have some distinguished features, such as having the suffix "ing" or being preceded by the verb "to be." Although the EVENT class presented the lowest number of words for evaluation, it obtained results similar to those of the other classes. The class that presented the lowest scores was ORGANIZATION, with an F1-measure of 0.875, even though it has a relatively high number of occurrences in the *corpus*. At the end of the tests, the model reached a general average of 0.940 for precision and 0.929 for recall. These values yielded to an F1-measure of 0.935, an impressive result for the task of automatic semantic annotation.

Table 1. The results obtained from the execution of CRF model using the standard configuration.

	Precision	**Recall**	**F1-Measure**	**Occurrences**
ACTION	0.996	0.997	0.996	48,058
PERSON	0.924	0.913	0.918	41,346
PLACE	0.914	0.898	0.906	33,202
INTANGIBLE	0.981	0.967	0.974	26,543
CREATIVE_WORK	0.912	0.879	0.895	11,059
ORGANIZATION	0.883	0.867	0.875	40,572
PRODUCT	0.965	0.954	0.959	15,675
EVENT	0.945	0.958	0.951	4785
AVERAGE	0.940	0.929	0.935	

After the execution, the tool used to run the CRF Model outputs a list containing the most prominent features detected by the learning process. The features were ranked according to their weight relative to the classification of each class. The Table 2 presents the features that are most relevant to the prediction for each of the top-level classes of Schema.org. The features listed are related to the processing of the whole dataset. The elements showed in the Table 2 can be interpreted as follows: the "affix" refers to a token suffix or prefix; "postag" refers to the part-of-speech of the token, and the value received matches the nomenclature used in the Penn Treebank Project (https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html); the "−" and "+" signs denote the position related to the current token, with "−" indicating a previous position in the sequence, and the "+" indicating a posterior position in the sequence; the number refers to the number of positions from the current token. There are many other features, but these were selected to depict the behavior of the model during the prediction phase.

Thus, in the sentence "top things to do in Kansas City", the word "Kansas" is preceded by the preposition "in", which matches the "−1:word.lower():in" feature, as is followed by the noun "city" with the "+1:word.lower():city" feature. Those two features reinforce the likelihood of the word "Kansas" falling into the category of PLACE. Table 3 summarizes the hyper-parameters used in the training, with and without hyper parameter optimizer.

Table 2. Most prominent features of each class detected during the training phase.

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

Table 3. Hyper-parameters used in the experiment.

Hyper-Parameter Optimization	Training Algorithm	Regularization	Coefficient for L1 and L2	Number of Iterations	Number of Features
no	GD with L-BFGS	elastic net (L1 + L2)	0.1 and 0.1	100	43,987
yes	GD with L-BFGS	elastic net (L1 + L2)	0.319 and 0.005	100	43,987

The Figure 2 shows the F1-measure scores for each class and for each dataset size. From the Figure 2 it is possible to note that the score increases as the size of the dataset gets bigger. All classes have a relatively low F1-measure for datasets with sizes of 1% and 10% of the whole *corpus*. However, as the size of the dataset increases, the value of F1-measure increases significantly up to 50% of the whole *corpus*. From 50% the F1-measure continues to grow but more slowly. This demonstrates that although the bigger the *corpus* the better the results, a data set with 50% of the total *corpus* size is enough to evaluate the model.

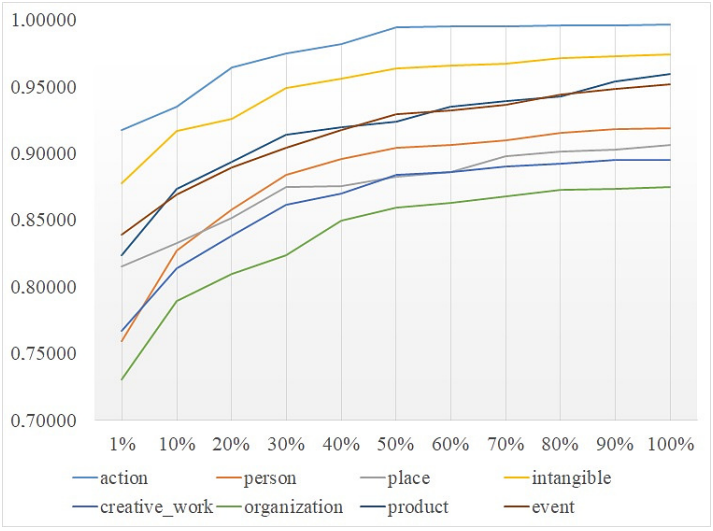

Figure 2. F1-measure increase along with the *corpus* size growth.

Due to the restrictions of computational resources to execute the second approach of the CRF model on the whole *corpus*, the results presented in Table 4 are partial. The tests performed in the second approach comprised 50% of the *corpus*. As mentioned earlier, the second phase of the model corresponds to the use of hyper-parameter optimization and cross-validation in order to improve the performance of the annotation process. These techniques require additional computational power to perform the training of the model. Therefore, since the *corpus* used is large, the available resources

were only able to handle half of the set. Regarding the results obtained, it is important to note that the performance of the model using this approach is similar to the results obtained using the whole *corpus*, although slightly better. The tests obtained a value of 0.927 for precision, 0.918 for recall and 0.923 for F1-measure using 116,530 tokens tagged over the eight classes. On the average, the results achieved using hyper-parameter optimization and cross-validation were higher, suggesting that it is beneficial the use optimization module of the tool.

Table 4. Results with use of hyper-parameter optimization and cross validation.

	Precision	**Recall**	**F1-Measure**	**Support**
ACTION	0.994	0.995	0.994	26,866
PERSON	0.913	0.895	0.904	21,625
PLACE	0.887	0.878	0.882	17,748
INTANGIBLE	0.974	0.953	0.963	14,830
CREATIVE_WORK	0.896	0.872	0.884	5099
ORGANIZATION	0.867	0.852	0.859	19,026
PRODUCT	0.919	0.928	0.923	8537
EVENT	0.914	0.945	0.929	2799
AVERAGE	0.927	0.918	0.923	

5. Conclusions

The meaning of words has many facets and levels of details. One of these facets is the information captured by ontologies which is capable of providing some insight about the nature of the concept denoted by the lexeme. Nonetheless, there are several levels of details, and ontological aspects that can be explored. The top-level ontology categories capture general concepts that attempt to present what exists in the world. In this sense, the use of these categories for annotating texts helps to categorize lexemes broadly. The semantic annotation can be performed either manually or automatically, but doing it manually is expensive as it demands skilled labor and time to perform the annotation. Aiming to mitigate this problem, this research proposed an automatic approach of semantic annotation based on top-level through the use of a supervised machine learning model.

Semantic annotation based on top-level ontologies using the supervised machine learning approach may contribute considerably to the successful execution of the task. For the research execution, it was necessary to define a *corpus*, being that in this case was selected the OANC *corpus*. Also, it was necessary to clear some errors in the annotated words, and to format the *corpus* according to an appropriate pattern. To supply the categories for the annotation process it was selected Schema.org, an ontology directed to organize the most common types of the Web. From Schema.org was chosen its top-level categories, composed of 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 relate features that occur far-off in a sequence, which is the case in natural language statements. It is a system that performs well, provided proper feature engineering is carried out. After having been trained the system can perform text annotation in real-time. Finally, the classification was performed, and the results for the different versions of the training and test sets were analyzed.

The results obtained were encouraging, although they are difficult to compare with other studies, since there is a lack of related works in the area of semantic annotation based on ontology. In general, the CRF model presented excellent results when annotating the *corpus* using the eight selected classes, achieving an F1-measure above 85% in each class and an average of 93.5% considering all classes. Comparing these results with the state of the art, which is reported in the section of related works, we can see that the proposal produces equivalent or superior results. That said, the use of CRF has some advantages over other techniques that are currently suggested. Regarding the maximum-entropy Markov model (MEMM), CRF does not suffer from the "label bias problem", where states with low-entropy transition distributions ignore their observations. However, CRF takes considerably longer to train. Regarding the deep learning technique, CRF has the advantage of not behaving like a black box and still presenting competitive results. Nonetheless, in fact, it presents the drawback

of needing a feature engineering phase. Another significant outcome of the research are the results obtained for the classes, PERSON, PLACE and ORGANIZATION, commonly used in classifications of named entities. In these classes we achieved results comparable to the state of the art. Although computational resources prevented the use of the entire database using hyper-parameter optimization and cross-validation, the results were positive enough to justify the approach. To conclude, after analyzing all the results obtained it is possible to conclude that, although it is still a new approach for automatic text annotation text based on top-level ontologies, the results of this research were quite promising suggesting the continuity of the research in this direction.

For future work, we plan to use more powerful computer resources that are able to deal with the whole *corpus*. Also, we plan to try hyper-parameter optimization techniques and cross-validation in order to improve the results and the processing power of the model. Another approach that can be explored is the use of other machine learning techniques to compare the results obtained. One of these techniques which has become popular is the deep learning model approach. Intending to deepen the process of ontology-based semantic annotation, one proposal to be analyzed is to use lower level categories of the Schema.org in order to assign meaning to the lexemes in a greater level of detail.

Author Contributions: Conceptualization, G.C.d.A. and A.d.P.O.; Funding acquisition, A.d.P.O.; Investigation, G.C.d.A.; Methodology, G.C.d.A., A.d.P.O. and A.M.; Resources, G.C.d.A.; Supervision, A.d.P.O. and A.M.; Validation, G.C.d.A.

Funding: This research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001, and also by the funding agencies FAPEMIG and CNPq.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sardinha, T.B. Lingüística de corpus: histórico e problemática. *Delta* **2000**, *16*, 323–367. [CrossRef]
2. Leech, G. Corpus annotation schemes. *Lit. Linguist. Comput.* **1993**, *8*, 275–281. [CrossRef]
3. Kiryakov, A.; Popov, B.; Terziev, I.; Manov, D.; Ognyanoff, D. Semantic annotation, indexing, and retrieval. *Web Semant. Sci. Serv. Agents World Wide Web* **2004**, *2*, 49–79. [CrossRef]
4. Reeve, L.; Han, H. Survey of semantic annotation platforms. In Proceedings of the 2005 ACM Symposium on Applied Computing, Santa Fe, NM, USA, 13–17 March 2005; pp. 1634–1638.
5. Handschuh, S.; Staab, S. *Annotation for the Semantic Web*; IOS Press: Amsterdam, The Netherlands, 2003; Volume 96.
6. Norvig, P.; Lakoff, G. Taking: A study in lexical network theory. In *Annual Meeting of the Berkeley Linguistics Society*; Berkeley Linguistics Society: Berkeley, CA, USA, 1987; Volume 13, pp. 195–206.
7. Miller, G.A. WordNet: a lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
8. Gries, S.T. Polysemy. In *Handbook of Cognitive Linguistics*; Dabrowska, E., Divjak, D., Eds.; Walter de Gruyter GmbH & Co KG: Berlin, Germany, 2015; Volume 39.
9. Ravin, Y.; Leacock, C. *Polysemy: Theoretical and Computational Approaches*; OUP Oxford: Oxford, UK, 2000.
10. Firth, J.R. A synopsis of linguistic theory, 1930–1955. In *Special Volume, Philological Society*; Oxford University Press: Oxford, UK, 1957; pp. 1–32.
11. Monaghan, F. Judging a word by the company it keeps: The use of concordancing software to explore aspects of the mathematics register. *Lang. Educ.* **1999**, *13*, 59–70. [CrossRef]
12. Pustejovsky, J.; Stubbs, A. *Natural Language Annotation for Machine Learning*; O'Reilly Media, Inc.: Newton, MA, USA, 2012.
13. Gruber, T.R. A translation approach to portable ontology specifications. *Knowl. Acquis.* **1993**, *5*, 199–220. [CrossRef]
14. Guarino, N. Formal ontology and information systems. In *Proceedings of FOIS*; IOS Press: Trento, Italy, 1998; Volume 98, pp. 81–97.
15. Uren, V.; Cimiano, P.; Iria, J.; Handschuh, S.; Vargas-Vera, M.; Motta, E.; Ciravegna, F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semant. Sci. Serv. Agents World Wide Web* **2006**, *4*, 14–28. [CrossRef]

16. Guarino, N. Some organizing principles for a unified top-level ontology. In *AAAI Spring Symposium on Ontological Engineering*; AAAI Press: Menlo Park, CA, USA, 1997; pp. 57–63.
17. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
18. Berners-Lee. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*; HarperBusiness: New York, NY, USA, 2001.
19. Maedche, A.; Staab, S. Ontology learning for the semantic web. *IEEE Intell. Syst.* **2001**, *16*, 72–79. [CrossRef]
20. Amancio, D.R. Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics* **2015**, *105*, 1763–1779. [CrossRef]
21. Akimushkin, C.; Amancio, D.R.; Oliveira, O.N., Jr. Text authorship identified using the dynamics of word co-occurrence networks. *PLoS ONE* **2017**, *12*, e0170527. [CrossRef] [PubMed]
22. Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; Ungar, L. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 729–740. [CrossRef]
23. Liu, Y.; Zhang, L.; Nie, L.; Yan, Y.; Rosenblum, D.S. Fortune teller: Predicting your career path. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
24. Estival, D.; Nowak, C.; Zschorn, A. Towards Ontology-based Natural Language Processing. In Proceeedings of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology, Barcelona, Spain, 25 July 2004; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 59–66.
25. Gries, S.T.; Berez, A.L. Linguistic annotation in/for corpus linguistics. In *Handbook of Linguistic Annotation*; Springer: Berlin, Germany, 2017; pp. 379–409.
26. FitzGerald, N.; Täckström, O.; Ganchev, K.; Das, D. Semantic role labeling with neural network factors. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; ACL: Lisbon, Portugal, 2015; pp. 960–970.
27. Fillmore, C.J. Frame semantics. *Cogn. Linguist. Basic Read.* **2006**, *34*, 373–400.
28. Bergamaschi, S.; Cappelli, A.; Circiello, A.; Varone, M. Conditional random fields with semantic enhancement for named-entity recognition. In Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, Amantea, Italy, 17–22 June 2017; ACM: New York, NY, USA, 2017; p. 28.
29. Skeppstedt, M.; Kvist, M.; Nilsson, G.H.; Dalianis, H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *J. Biomed. Inform.* **2014**, *49*, 148–158. [CrossRef] [PubMed]
30. Pandolfo, L.; Pulina, L. ADnOTO: 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*; Springer: Arras, France, 2017; pp. 495–501.
31. Adorni, G.; Maratea, M.; Pandolfo, L.; Pulina, L. An Ontology-Based Archive for Historical Research. In *Description Logics*; EUR-WS: Athens, Greece, 2015.
32. Liu, K.; El-Gohary, N. Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Autom. Constr.* **2017**, *81*, 313–327. [CrossRef]
33. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data-the story so far. *Int. J. Semant. Web Inf. Syst.* **2009**, *5*, 1–22. [CrossRef]
34. Heath, T.; Bizer, C. Linked data: Evolving the web into a global data space. *Synth. Lect. Semant. Web Theory Technol.* **2011**, *1*, 1–136. [CrossRef]
35. Patel-Schneider, P.F. Analyzing schema.org. In *International Semantic Web Conference*; Springer: Riva del Garda, Italy, 2014; pp. 261–276.
36. Guha, R.V.; Brickley, D.; Macbeth, S. Schema.org: Evolution of structured data on the web. *Commun. ACM* **2016**, *59*, 44–51. [CrossRef]
37. Ronallo, J. HTML5 Microdata and Schema. org. *Code4Lib J.* **2012**, *16*, 1.
38. Ide, N.; Suderman, K. The American National Corpus First Release. In *LREC*; ELRA: Paris, France, 2004.
39. Andrade, G.C. Hybrid Semantic Annotation: Rule-Based and Manual Annotation of the Open American National Corpus with a Top-Level Ontology. Ph.D. Thesis, Universidade Federal de Viçosa, Viçosa, Brazil, 2017.

40. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 282–289.
41. Sutton, C.; McCallum, A. An introduction to conditional random fields. *Found. Trends Mach. Learn.* **2012**, *4*, 267–373. [CrossRef]
42. Dietterich, T.G. 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)*; Springer: Cham, Switzerland, 2002; pp. 15–30.
43. Cohn, T. Efficient Inference in Large Conditional Random Fields. In *Machine Learning: ECML 2006*; Fürnkranz, J., Scheffer, T., Spiliopoulou, M., Eds.; Springer: Berlin/Heidelberg, Germnay, 2006; pp. 606–613.
44. Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **1980**, *35*, 773–782. [CrossRef]
45. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]