electronics

MDPI

Article

Complement Recognition-Based Formal Concept Analysis for Automatic Extraction of Interpretable Concept Taxonomies from Text

Stefano Ferilli

Department of Computer Science, University of Bari, 70125 Bari, Italy; stefano.ferilli@uniba.it; Tel.: +39-080-5442293

Abstract: The increasing scale and pace of the production of digital documents have generated a need for automatic tools to analyze documents and extract underlying concepts and knowledge in order to help humans manage information overload. Specifically, since most information comes in the form of text, natural language processing tools are needed that are able to analyze the sentences and transform them into an internal representation that can be handled by computers to perform inferences and reasoning. In turn, these tools often work based on linguistic resources for the various levels of analysis (morphological, lexical, syntactic and semantic). The resources are language (and sometimes even domain) specific and typically must be manually produced by human experts, increasing their cost and limiting their availability. Especially relevant are concept taxonomies, which allow us to properly interpret the textual content of documents. This paper presents an intelligent module to extract relevant domain knowledge from free text by means of Concept Hierarchy Extraction techniques. In particular, the underlying model is provided using Formal Concept Analysis, while a crucial role is played by an expert system for language analysis that can recognize different types of indirect objects (a component very rich in information) in English.

Keywords: text mining; conceptual taxonomies; formal concept analysis

Citation: Ferilli, S. Complement Recognition-Based Formal Concept Analysis for Automatic Extraction of Interpretable Concept Taxonomies from Text. *Electronics* **2023**, *12*, 2137. https://doi.org/10.3390/electronics12092137

Academic Editors: Juan M. Corchado, Carlos A. Iglesias, Byung-Gyu Kim, Rashid Mehmood, Fuji Ren and In Lee

Received: 24 March 2023
Revised: 28 April 2023
Accepted: 4 May 2023
Published: 7 May 2023

1. Introduction

The proliferation of computers and the widespread success of the Internet have led to increased exploitation of digital documents in place of legacy paper ones, which represent a huge source of information for our everyday life. Most of this information is included in the form of textual components, and is often hidden by the complexity of natural language. Hence, there is an increasing need for automatic tools to analyze documents and extract the underlying concepts and knowledge, and a related need for natural language processing (NLP) tools that are able to analyze the sentences and transform them into an internal representation that can be handled by computers to perform inferences and reasoning. This is a hard task due to the intrinsic ambiguity and variety of natural languages, and to the large amount of implicit knowledge underlying natural language utterances. Several sub-tasks can be identified, often requiring the availability of corresponding linguistic resources, from morphological and lexical levels to more syntactic and semantic aspects.

Text mining, or knowledge discovery from text [1], aims at discovering new and hidden information in large amounts of text in natural language. The availability of domain knowledge plays a fundamental role in this effort. Text mining can be seen from many different perspectives. As an Information Extraction (IE) process, it aims at extracting useful information from single words, sentences and passages, where the kind of semantic information that is being looked for is not known in advance. As a data mining process, it applies algorithms and methods of statistical machine learning to text in order to identify useful patterns (sequences of words) [2,3]. As a knowledge discovery in databases process,

it often does not actually discover information, but just extracts it [4]. The Web is a very interesting source of non-structured textual data for IE [5].

Text mining is carried out in two steps. The former exploits NLP and IE techniques to transform non-structured text documents into an intermediate semi-structured (e.g., conceptual graphs) or structured (i.e., relational) representation; the latter deduces from such a representation, patterns of knowledge [6].

An important task in the field of IE consists of the retrieval of relevant 'concepts' in a set of documents concerning a specific domain. However, extracting concepts from pure text is very difficult; additionally, there is no agreement on how such an activity is to be carried out. Possible perspectives include:

- Extensional, where the meaning of a concept or term is intended as its 'extension', i.e., the set of objects that fall in its definition;
- Intensional, where the meaning of a concept or term is described according to a set of attributes or properties that characterize it, and that are often expressed in their textual definition (e.g., the OntoLearn system [7]).

After extracting concepts from text, their retrieval, reuse and understanding must be supported by structuring the information into categories and by identifying and expressing relationships among them in an abstract and compact way that allows reasoning. Being the core of ontological resources, this is a fundamental issue for any knowledge base, in order to specify axioms, rules and implications among facts. However, the 'knowledge acquisition bottleneck' (gaining domain knowledge from experts and formally representing it is a very hard task) is well-known in the literature, and several techniques have been proposed to tackle such a problem. General-purpose lexical resources such as WordNet [8] are very common, but are often insufficient to support specific knowledge domains. Hence, the motivation for automatically building such resources. Methods for the (semi-)automatic creation of concept hierarchies from text can be grouped into two broad categories:

- Similarity-based approaches, which exploit a similarity (or distance) measure to assess the similarity between pairs of terms and decide whether they can be grouped together; this class includes hierarchical clustering algorithms based on Harris' distributional hypothesis [9] or on the analysis of term co-occurrence.
- Set-theoretical approaches, which exploit a partial order on objects based on their set of attributes; this class includes Formal Concept Analysis (FCA) to be exploited in this work.

Document processing aimed at extracting and organizing the concepts expressed in natural language text is an important challenge for research in text mining. The need for such abilities has grown as the number of available documents has become so huge that they cannot be handled manually. The application of Formal Concept Analysis to obtain such a result might provide interesting results towards the understanding of the text semantics and the creation of a domain ontology or conceptual taxonomy or conceptual network to be shared among different users. However, the application of the FCA is based on the availability of a formal context representing meaningful relations between relevant objects and attributes. This work proposes to leverage the syntactic and logical structure of sentences to provide such a context. In particular, the focus is on grammatical complements as a very expressive component of the sentence structure. For this reason, an existing proprietary expert system for grammatical parsing of English sentences was extended to recognize particular kinds of complements and provide the data for building a formal context on which to run FCA algorithms.

Summing up, the objectives of this paper are:

1. Proposing a syntax-based approach to automatically build interpretable concept taxonomies starting from pure text;
2. Using the learned taxonomies to understand the content of the document from which it was extracted.

Objective 2 will indirectly contribute to assessing the effectiveness of objective 1 as well.

The rest of this paper is organized as follows. After recalling relevant background information and related work in the next section, we describe our approach in Section 3, and then provide a qualitative and quantitative evaluation in Section 4. Finally, Section 5 concludes the work and provides an overall discussion.

2. Background and Related Work

While there is no universal agreement about fully automatic ontology acquisition from text being a realistic goal (some only consider systems that support knowledge engineers [10,11]), many attempts have been made in the literature to build concept taxonomies by mining (possibly large amounts of) text.

2.1. Clustering Approaches Based on the Vector Space Model

A widespread technique for representing a collection of text documents is the Vector Space Model (VSM) [12], which uses the terms occurring in the corpus as dimensions of a space in which documents are represented as points (i.e., vectors) whose coordinates are determined by the occurrences of terms therein (another interpretation might be that terms correspond to points in a space whose dimensions are the documents in the collection). Thus, the similarity (or distance) between terms (or documents) can be computed according to classical techniques borrowed from geometry, such as Euclidean distance or cosine similarity.

Therefore, given a geometric space and associated distance or similarity measures, clustering algorithms can be exploited to automatically infer semantic classes. Clusters are often represented by their centroids (the average of all their elements) or medoids (the actual element therein that is most representative of the whole cluster). However, the computation of such representatives might be biased by spurious elements or outliers. For this reason, the contexts of terms might be exploited to better define them and group those that usually occur in the same context. This corresponds to reducing the number of dimensions to the most representative terms only, which can be considered as keywords, which also improves efficiency.

Dimensionality reduction techniques try to capture the concepts underlying a collection of documents and to express the documents according to such concepts, instead of the terms appearing therein. Latent Semantic Analysis [13] is a statistical technique based on the principle that the set of contexts in which a word appears, or does not appear, represents a set of constraints according to which it can determine the similarity among groups of words. The classical term-document matrix provided by the Vector Space Model undergoes a Singular Value Decomposition process [14], which identifies a set of 'latent' concepts underlying the collection that become the dimensions of a new space (much less than the number of terms), in which both documents and terms can be placed. In this semantic space, similar documents and terms are placed near each other.

Similarly, Concept Indexing (CI) aims at representing each document as a function of the concepts appearing in the document collection [15] by first discovering groups of similar documents, each of which is intended to be a different concept in the collection, and then using these groups as dimensions for the reduced space. In supervised mode, CI identifies these groups based on previously available classes of documents, while in the unsupervised case, they are obtained using clustering techniques.

A shortcoming of these techniques is their inability to provide deep insight into the extracted concept because only isolated terms are considered.

2.2. Specific Approaches Aimed at Building Concept Taxonomies

Some approaches exploit external resources to bridge the gap between the syntactic level of text and the semantics behind it. For example, TEXT-TO-ONTO [10,11] semi-automatically discovers concepts and taxonomic relationships (only) between them. This is

achieved by leveraging user contributions and external resources in the form of constraints and background knowledge at various language levels (from morphology up to pragmatics). The process of knowledge acquisition is performed semi-automatically by means of external resources and user contributions. OntoLearn [7] works in four steps: terminology extraction, derivation of term sub-trees by means of string inclusion, Word Sense Disambiguation and combination of the sub-trees into a taxonomy. It uses WordNet [16], a static and general-purpose resource, which prevents the learning of domain-specific taxonomies. The approach proposed in [17] uses a combination of VerbNet [18] and WordNet to shift the representation to the semantic level. It works in two steps: first, the semantic roles in a sentence are identified, and then these roles are used, together with semi-automatically compiled domain-specific knowledge, to build the concept graph.

More recently, the Automatic Taxonomy Construction from Text (ATCT) framework [19] uses a filtering approach to select the most relevant terms for a specific domain among those extracted from a corpus of documents and disambiguates their sense using a technique based on WordNet; then, it generates the concepts and organizes them in a generalization taxonomy determined using a subsumption technique based on concept co-occurrences in a text. Ref. [20] focuses on concept formation and hierarchical relation learning, obtained by partitioning and grouping the extracted concepts through Hierarchical Agglomerative Clustering, informed by syntactic matching and semantic relatedness functions and based on a novel automated dendrogram pruning technique, which is dynamic to each partition. The explicitly claimed novelty of this work lies in its exploitation of external knowledge bases. As witnessed by the references reported in these papers, the problem is still a relevant one, but no recent research has been carried out to tackle it.

When no structured and machine-readable external knowledge is available, approaches based only on what is expressed in the text are needed. Among these, ref. [21] defines a language to build formal ontologies by deductive discovery such as in logic programming. In particular, the author defines both a specific language for manipulating web pages and a logic program to discover the concept lattice. Ref. [22] uses text understanding to automatically expand a small, manually built ontology kernel that defines the primitive concepts, relations and operators. The features are domain/application-independent. The ontology (including words, concepts, taxonomic and non-taxonomic relations and axioms) is learned by applying a hybrid approach consisting of logic, linguistic and semantic analysis methods. Finally, most relevant to the work proposed in this paper, ref. [23] automatically acquires taxonomies or concept hierarchies from a text corpus using Formal Concept Analysis (FCA) to group nouns/objects based on verbs/attributes. It models the context of a term as a vector representing syntactic dependencies that are automatically extracted from the text corpus using a linguistic parser. Based on this context information, FCA produces a lattice that is converted into a concept hierarchy by defining a partial order. Due to the lack of standard datasets and golden standards, the learned taxonomies have been compared to hand-crafted ones.

Regardless of the approaches, both with and without the use of external resources, the main issue is the validation of the conceptual graph obtained. Since there are no standard formal techniques [24] capable of validating the quality of a conceptual graph learned from text, it can be assessed by the social consensus of domain experts or by the usability of the conceptual graph with respect to business objective.

2.3. BLA-BLA and ConNeKTion

The work presented in this paper continues a flow of research aimed at building a general system for the automatic generation of linguistic resources from pure text, called BLA-BLA (an acronym for 'Broad-spectrum Language Analysis-Based Learning Application') [25]. The resources learned by BLA-BLA may be used as returned by the system, and/or be taken as a basis for further manual refinements. It currently includes several techniques that allow us to learn in a fully automatic way linguistic resources for language identification [26], stopword removal [27], term normalization [28], syntax checking [29]

and concept taxonomies [30]. Whenever more texts become available for the language, it is easy to run BLA-BLA again and obtain updated resources.

In particular, the component of BLA-BLA aimed at extracting concepts and organizing them into taxonomies is ConNeKTion (an acronym for 'CONcept NEtwork for Knowledge representaTION') [30]. It also extracts several kinds of non-taxonomic relationships. ConNeKTion integrates a mix of existing and novel tools and techniques that are brought to cooperation in order to reach its objectives. In addition to learning the concept network, it embeds tools inspired by human abilities in text understanding and concept formation to support several exploitation tasks, among which are assessing the relevance of the concepts expressed in a given text; obtaining formal and human-readable descriptions of the concepts underlying the terms; generalizing concepts; applying different kinds of reasoning on the learned network (e.g., associative reasoning based on indirect relationships between concepts, deductive reasoning based on formal logics, etc.); identifying relevant keywords that are present in a text; helping the user to retrieve useful information from a document corpus; and recognizing the author of a document based on the writing style and structure of the sentences. It comes with a graphical control panel that allows its users to apply its concept network learning, consultation and exploitation tools and to comfortably visualize, filter according to various criteria and use the learned graph in order to obtain information on the document collection and its content. While the implemented prototype works in English, the methodologies embedded in ConNeKTion are completely general and applicable to any language.

ConNeKTion was developed in opposition to approaches that need external resources to learn the concept network. Differently from TEXT-TO-ONTO and [17], it completely avoids human intervention. With respect to TEXT-TO-ONTO, it also extracts non-taxonomic relationships associated with verbs, while differently from OntoLearn, it can exploit all recognized concepts, both generic and domain-specific. However, it was also developed to overcome some limitations of the existing approaches that do not use external resources. Differently from [21], it does not learn a pre-defined set of relationships, but any new relationship is added to the taxonomy as soon as it is found. Compared to [23], it builds the concept graph relying on the whole set of concepts and relationships, rather than only on shared attributes as in their taxonomic representation of concepts.

Like [23], it associates concepts with terms and relationships with verbs. Moreover, it also associates attributes with adjectives. It can learn propositional (expressed as feature vectors or attribute-value pairs) or relational (expressed using first-order logic formulæ) concept definitions, so as to be able to handle different levels of complexity and expressiveness in concept representation. It can further structure and enrich the network of explicit (taxonomic or non-taxonomic) relationships among concepts extracted from the text, as well as in the presence of missing or partial knowledge (a typical problem in small collections), with additional taxonomic relationships obtained by clustering and/or logic generalization of concepts.

3. Concept Extraction Based on Complement Recognition

Inspired by [23], in this work, we expanded the taxonomy restructuring and refinement abilities of ConNeKTion by adding a new approach based on FCA to build and organize a concept taxonomy associated with a single document or to a set of documents belonging to one or more domains. Differently from [23], we more thoroughly exploit the syntactic information that is present in the text, and more specifically focus on the added value that an analysis of complements can provide. According to the Oxford Dictionary, a complement is "any word or phrase that is governed by a verb and usually comes after the verb in a sentence".

3.1. Formal Concept Analysis Basics

Formal Concept Analysis (FCA) [31] is a data analysis method to identify concept taxonomies. It groups objects having similar attributes, whose intensional descriptions

are human-readable [32]. Although based on a very simple knowledge representation, its outcome is a very complex and powerful lattice expressing concepts and their generalization/specialization relationships [33]. Compared to other kinds of structures, such as hierarchies, lattices allow multiple inheritances, can be translated into different representations that are more suitable to some kinds of tasks and are endowed with a fully-fledged algebraic structure.

Objects and associated attributes are grouped, based on Galois connections, to determine which attributes and objects make up consistent entities called concepts, each endowed with an extension (the set of objects that belong to the concept) and an intension (the set of attributes owned by all objects that belong to the concept). The starting point is the set of all pairs (object, attribute) such that the object owns an attribute, which can be compactly represented as Boolean matrix objects $\times$ attributes where a cell (i, j) is marked whenever object i owns attribute j. More formally:

- A *formal context* it is a triple $K = (G, M, I)$ where
  - G is a set of *objects*,
  - M is a set of *attributes*, and
  - $I \subseteq G \times M$ is a relationship such that $(g, m) \in I$ if "object g has attribute m";
- Given K, a *formal concept* is a pair (X, Y), where
  - $X = \{x \in G | \forall y \in Y : (x, y) \in I\} \subseteq G$ is its *intension* and
  - $Y = \{y \in M | \forall x \in X : (x, y) \in I\} \subseteq M$ is its *extension*.

The set of formal objects associated with a formal concept identifies its extension, while the set of formal attributes represents its intension. The adjective 'formal' suggests that it is just a theoretical distinction between two roles (i.e., in some domains, what would be intuitively considered an object might play the role of a formal attribute, and what would be usually considered an attribute might play the role of a formal object).

For a given formal context, extensions and intensions of the corresponding formal concepts are univocally determined. A partial ordering relationship can be defined on the set of formal concepts according to set theory: two formal concepts are related if the intension of the former is a subset of the intension of the latter (and thus the extension of the former is a superset of the intension of the latter). The former is said to be a super-concept of the latter, and the latter a sub-concept of the former. This induces a *concept lattice*, in which the union and intersection of concepts can be computed, always yielding another super- or sub-concept, respectively. The top element in the lattice contains all formal objects (a kind of 'universal' concept), while the bottom element owns all formal attributes (often resulting in an empty extension).

Several algorithms were developed to extract all concepts from a given formal context, and to generate the corresponding conceptual lattice. Some of them work in batch mode, producing the final result from scratch, proceeding either top-down (concepts having larger extensions are generated first (e.g., [34,35]) or bottom-up (e.g., [36]); others work incrementally, generating at the nth step the conceptual lattice for the first n objects in the context (e.g., [37–39]).

Figure 1 graphically shows the procedure for identifying the formal concepts. On the top-left, the formal context is shown, in which the rows are the objects and the columns are the features. On the right-hand side, the sequence of concepts extracted is shown, by progressive levels of aggregation (i.e., of abstraction). Each concept is described as a pair (e, i), where e is its extension (the objects belonging to it) and i is its intension (the features owned by the concept). In the first level, most specific concepts are extracted, associated with single objects. For example, the first row/object produced the first concept in the list, having as an extension only that object ($\{1\}$) and as the intension all of its features ($\{a, c, f, h\}$). In the second level, concepts obtained from pairs of concepts in the first level are created. For example, concepts $\{2\}$ and $\{5\}$ are merged into a new concept with extension $\{2, 5\}$ and intension the intersection of their features ($\{g\}$). Similarly, concepts $\{3\}$ and $\{5\}$ are merged into a new concept with extension $\{3, 5\}$ and intension

the intersection of their features, which is again $\{g\}$. With the intension of this new concept being the same as for concept $\{2,5\}$, the two concepts are actually the same, and their extensions are merged taking their union ($\{2,3,5\}$). Then, the procedure continues until the top concept, with extension of the whole set of objects E and intension of the empty set (i.e., no specific features) is obtained. Opposite to this, the bottom concept $(\emptyset, F)$, involving all features F and with extension of the empty set (since in this case no object in the context owns all features), is generated.

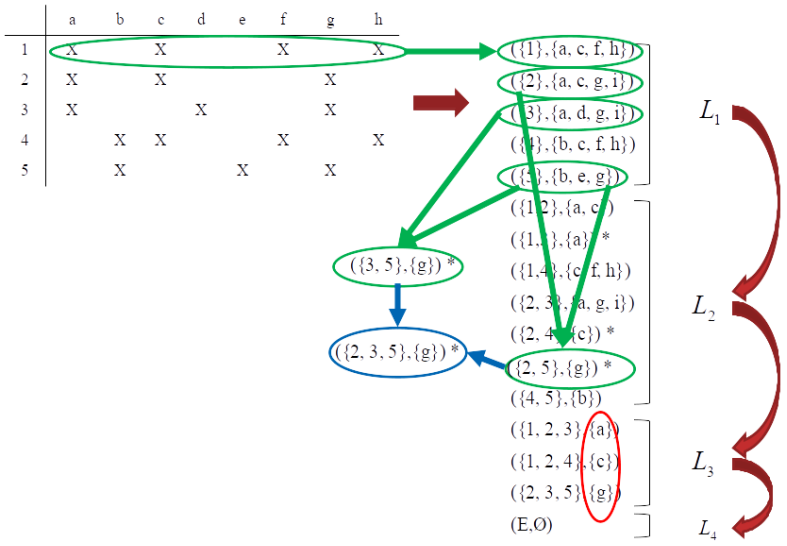

Figure 1. Formal Concept Analysis procedure.

3.2. Complement-Based Concept Taxonomy Extraction

Complement recognition can be carried out as a further processing step, after a preliminary standard parsing of the sentences that make up the text. In fact, parsers usually identify noun groups, verb groups and prepositional phrases, without further specifying the type of the latter. We believe that complements may provide extremely useful information that contributes to the definition of concepts because they express different ways in which concepts relate to each other and to their attributes.

Among the many kinds of complements available in natural language grammar, in this work, we decided to handle the following ones because they are considered the most widespread and relevant:

- Subject Complement (SC);
- Direct Object (DO);
- Indirect Object (IO);
- Object Complement (OC);
- Object of Preposition (OoP), in turn further specialized as follows:
  - OoP Time (a noun group expressing time);
  - OoP Place (a noun group concerning locations);
  - OoP Cause-Purpose (a noun group or verb group expressing a cause or a purpose);
  - OoP Possessive Phrase (a verb group preceded by preposition 'of').

To obtain such a detailed output, which is not provided by parsers available in the literature, we purposely developed a rule-based system, implemented in CLIPS language. While a detailed description of the parser goes beyond the scope of this paper, Figure 2 shows the overall pipeline underlying its execution, applied to the sample sentence "Mary photographed the woman with the children". The original text annotated with Part-of-Speech tags is translated into a logic format that is input into the complement recognition rule-based system, which enriches the file by adding information on larger and larger grammatical aggregates that are progressively recognized based on rules such as R5_DO

in the figure, stating that "in a sequence 'verb–noun phrase–prepositional phrase' the noun phrase is the direct object". The resulting file is input into another rule-based CLIPS module that extracts the features in the format required by the subsequent processing steps of our approach.

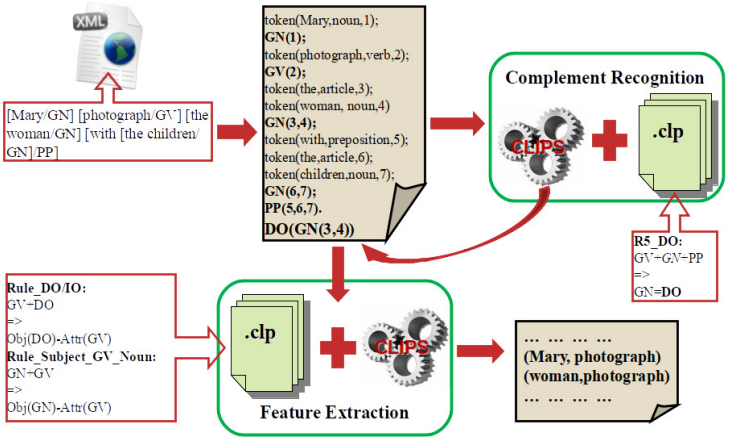

Figure 2. Complement recognition and feature extraction pipeline.

Then, the identified complements are processed to obtain a formal context by applying different patterns. Some are modeled after syntactic-functional dependencies derived by the grammatical roles played by each syntagm in the sentence:

- (Subject, verb);
- (Object, verb);
- (Indirect_Object, verb);
- (Object_of_Preposition, verb).

Others are based on semantic connections (Location, Temporal, Reason_Cause, Is_a, Part_Of), and allow us to associate the different kinds of complements with further object–attribute pairs. Specifically, the following strategy was used:

- Subject Complement: represents additional information about the subject; hence, the subject plays the role of a formal object while the attribute is the subject complement.
- Object Complement: comes along with a DO and further specifies it; hence, the DO represents the object while the attribute consists of the object complement.
- Object of Preposition: introduced by a preposition, represents an object, whose associated attribute is made up by the verb and preposition that introduces the OoP, in particular:
  - Prep of Time/Time Complement: the noun represents a temporal expression; hence, the associated noun becomes a formal object associated to attribute **Temporal_Term**,

    (noun, **Temporal_Term**)
  - Prep of Place: the noun plays the role of the object associated with the attribute **Location_Term**,

    (noun, **Location_Term**).
  - Prep of Cause/Purpose: the noun is the object associated with the attribute **Causal_Term**,

    (noun, **Causal_Term**).
  - Prep of Possessive Phrase: the associated noun adds details about the noun preceding it; hence, the former is associated with the latter.
  - Undefined Complement: objects of these complements are associated with an attribute made up by the 'verb + preposition' introducing them, except when introduced by the preposition 'with', which might denote a relation of inclusion

(meronymy) between two nouns where the concept underlying the OoP noun might be included in the concept associated to the noun preceding it.

Operationally, the elements that make up the formal context are extracted from the text starting from all possible attributes. Then, the formal objects associated with the identified attributes are collected, along with additional pairs determined by the complement semantics. By analyzing the neighborhood of all recognized complements, the 'verb+preposition' preceding a complement is taken as a formal attribute, and associated with the formal object obtained as the noun that is the object of the complement; conversely, in the case of the complement immediately following the verb, the object is associated with the attribute made up by the verb alone. Complement analysis might generate further pairs by associating the noun that is the object of the complement, as a formal object, to the formal attribute identified by the meaning of the complement (e.g., in the case of an Object of Preposition of Time the system will take the noun identifying such a complement as a formal object, and will associate it to a formal attribute **Temporal_term**). Finally, the object–attribute pairs associated with the subject–verb dependency are obtained by analyzing the neighborhood of all extracted attributes to identify possible nouns that precede the verbs classified as attributes (that represent the formal objects associated with those attributes).

In determining which pairs satisfy the chosen syntactic-functional and semantic dependencies, the verbs 'to be' and 'to have' are assumed to play the role of auxiliaries whenever associated with other verbs in a verb group, and in such a case, are not exploited to introduce a separate formal attribute.

The final concept taxonomy consists of a hierarchy in which a 'universal' root concept (called the *Domain*) can be identified, while the various leaf nodes represent the extracted concepts, and intermediate nodes along the paths from the Domain to specific concepts correspond to attributes that identify that concept. Obviously, one document might yield several unrelated portions of this hierarchy, but the presence of the same concept in different parts of the document will allow us to merge some of these portions together, thus obtaining a more connected and structured taxonomy. In the same way, the partial taxonomies coming from different documents from the same domain can be merged together in a larger taxonomy by leveraging shared concepts. This will result in a stratified organization, in which several levels of conceptualization can be identified.

We defined a convenient pattern to represent such a structure. We formally express the pattern using XML language. Compared to other options (e.g., JSON), we considered XML more suitable for various reasons, among which: (i) it is more expressive, allowing us to specify more information as attributes for the tags; (ii) it is not strictly connected to any specific language (while JSON, albeit processable in all languages, has a direct connection to JavaScript); (iii) it can be easily integrated with other representations used by our overall system BLA-BLA. It is structured as follows:

- Concept x
  {**extension of** x}
  - Concept y
    {**extension of** y}
    * Concept z
      extension of z}
      {intension of z};
    * ...

    {intension of y};
  - ...

  {intension of x};
- ...

where formal concepts are identified by integers (in the following we will denote concept n as C-n, for short) and described by two lists of terms: the former representing their extension, and the latter their intension. Sub-concepts are represented as a nested list of

concepts (e.g., in the above pattern, concept C-z is a sub-concept of concept C-y, which in turn is a sub-concept of concept C-x). Note that the list of terms that express the extension of a concept might even be missing, in case the concept is not univocally determined by any term, but just by the extension of its sub-concepts.

4. Evaluation

The correctness of the proposed technique, and its practical viability in real contexts, has been evaluated by building a prototype of the system and running it on selected documents to obtain both quantitative and qualitative indications of performance. This also allowed us to study the system behavior and compare its results on different kinds of input.

Evaluation was carried out running the system on a PC endowed with an Intel Core 2 Duo T7250 processor at 2.0 Ghz, 2 GB RAM and the Microsoft Windows Vista operating system. This obsolete platform was used to show that the approach can satisfactorily run even on low-performance computers. Reported figures for time are actually an approximation because the time of access to the disk for DBMS operations, which depends on the DBMS load, was not specifically estimated. Considering the characteristics of the computer used for the evaluation, the execution time is not relevant as a whole, but it is reported to highlight the difference between short and long documents.

Since the quality of the system outcome strongly depends on the quality of the parsing/complement recognition phase, the expert system parser outcome was manually fixed by human experts before providing it to the FCA sub-module in order to avoid errors in the automatic parsing affecting the final output of the proposed procedure. Indeed, the discussion of the final results reported below will confirm that the correctness of the parsing step is a crucial requirement for the success of the whole procedure.

The evaluation took into account six documents that were different in terms of subject (although for sport and politics two documents were included to check whether the information extracted from them is related and consistent) and structure:

1. American Football, with references to rules and origins of such a discipline;
2. The phenomenon of globalization, including quotations of definition from scholars and interesting aspects of the phenomenon;
3. Rugby Union, with references to rules and history of such a discipline;
4. Aspects of Artificial Intelligence, such as origins, history and cinematographic references;
5. Main aspects of governmental politics in Italy, with references to the legislative, executive and judicial branches in the Italian system;
6. Main aspects of governmental politics in the U.S., with references to the legislative, executive and judicial branches in the U.S. system.

Both short (up to 20 sentences) and long (more than 20 sentences) documents were included to check the impact of text length on performance. The complete texts are reported in Appendix A.

4.1. Quantitative Evaluation

We would like to emphasize that the aim of this paper is to propose a new approach to extract concepts and their taxonomic organization from single documents in pure text, and showing that it can extract relevant knowledge. The aim is not extracting all the possible information available in the text. We are more interested in checking what kind of concepts can be found using this approach than in having a pre-defined list of concepts to look for. Indeed, to the best of our knowledge, there is no other system that uses the complements in the same way as we do and for the same purposes. Therefore, there is no ground truth available against which to measure the quality (e.g., completeness or correctness) of the extracted information. Even should this be available, the problem might be in the parser, not in the concept extraction approach. For these reasons, our quantitative evaluation can report only general statistics on what was extracted to check if it is sensible and substantial. Preparing a large dataset and annotating it with ground truth would

require a specific research effort, which is out of the scope of the aims of this paper. Instead, we selected a limited number of documents with different topics, styles and lengths.

Our quantitative analysis of the experimental results is reported in Table 1, concerning document size (number of sentences and terms), number of formal objects and attributes extracted, number of formal concepts generated and number of relationships identified among them.

Table 1. Statistics about documents used for evaluation.

Doc	#Sent	Length	#Term	#Obj	#Attr	#Conc	#Rel	Runtime (s)
D-1	14		295	54	50	82	77	31.2
D-2	17	short	419	72	62	94	77	37
D-3	16		383	53	62	91	91	54.4
D-4	21		601	121	106	151	99	98.4
D-5	26	long	524	67	75	110	103	84.6
D-6	23		494	83	85	121	105	113.5

Figure 3 shows the impact of each processing phase on the overall runtime. In some documents, the portion of runtime needed for feature extraction is larger, due to a more complex syntactic structure of the sentences to be parsed by the expert system in order to recognize complements. In all cases, the prevailing activity is FCA, while the effort for generation of the conceptual taxonomy marginally affects runtimes and is proportional to the number of concepts and relations included in the taxonomy. Overall, the performance in efficiency can be considered as satisfactory; short documents are processed, on average, in about 40 s, while long ones take 1 min and 40 s on average. The runtime for D-3 and D-6 is higher than the average due to the large number of concepts and relations identified therein. The long runtime required for processing D-4 can be explained by the large amount of information born by that document, because it extensively deals with the subject of Artificial Intelligence, ranging from considerations about ethics to the associated cinematography. This is confirmed if the number of relationships identified among the extracted concepts (91) is considered, which is a small value compared to the number of extracted concepts (137), indicating that in the considered document there is no strong consistency among concepts, differently from other cases (e.g., D-6) where a very specific subject defined by strongly correlated concepts is dealt with.

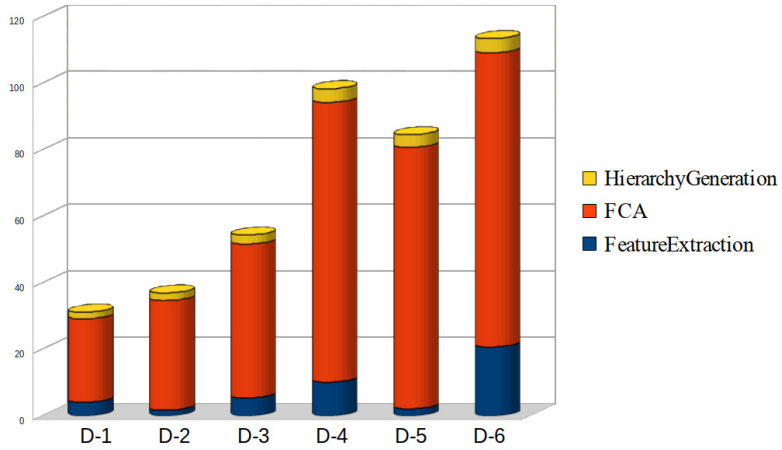

Figure 3. Proportion of runtimes for different processing phases.

Table 2 reveals that the average frequency of formal objects and attributes extracted from each document is acceptable, because each sentence may contain several nouns that concur in describing one or more concepts. In particular, the average values of objects and attributes extracted from D-5 and D-6 are not very different from each other because the

two documents concern very similar subjects (politics and government in Italy and USA, respectively). Moreover, the average number of concepts extracted from D-4 is large, confirming what was pointed out above, while the larger average number of relationships among concepts is obtained for D-3 because the concepts are highly inter-related due to the very specialized subject it discusses. Overall, we can claim that the Concept Hierarchy Extraction task is carried out effectively by the prototype.

Table 2. Averages of extracted items from sample documents.

Doc	#Obj/#Sent	#Attr/#Sent	#Conc/#Sent	#Rel/#Sent
D-1	3.86	3.57	5.86	5.5
D-2	4.24	3.65	5.53	4.53
D-3	3.31	3.88	5.69	5.69
D-4	5.76	5.05	7.19	4.71
D-5	2.58	2.88	4.23	3.96
D-6	3.61	3.7	5.26	4.57

4.2. Qualitative Evaluation

Now, let us turn to an evaluation and discussion of the correctness, quality and reliability of the conceptual taxonomies actually extracted from those documents. Due to space limitations, a detailed analysis of all 649 concepts and 552 relations cannot be reported in the paper, and even providing an overview of the taxonomies would be uncomfortable to read. Instead, in the following, we will discuss some elements extracted that we consider more interesting, also providing a comparison between items extracted from different documents when relevant. The complete information can be obtained by sending a request to the contact author.

4.2.1. Concept Extraction from Single Documents

We start with an analysis of the elements extracted from single documents. Specifically, we focus on D-2 among short documents, and on D-4 among long ones, since both provided the largest number of concepts and allow us to check for possible correspondences.

Among the several concepts extracted, let us consider C-207 "internet access, e-commerce and coop bring in Third World nation" (see Table 3). It comes from the sentence "[...] Internet access and e-commerce have brought small-scale coops in Third World nations [...]" in D-2, and confirms the importance of correctly identifying the concept 'Third World nation' as a location_term based on the semantics of complements. Concept C-142 'world' adds further information; indeed, it is specified by the attribute 'about business', which establishes an association identified by concept C-196 'business world' in the taxonomy, which is again based on the same C-142.

In a different fragment of the taxonomy (reported in Table 4), C-189 identifies the concept 'Tom Palmer of Cato Institute'. This is interesting because it shows that the system can identify entities as concepts (in this case 'Tom Palmer' thanks to attributes 'of Cato' and 'of Institute'). Moreover, exactly those attributes allow us to find a possible relationship to C-209 'Cato Institute define globalization on elimination or diminution about restriction across border and system'. C-188 adds further information by specifying that the concept 'border and system' refers to market ('about exchange') and to production ('about production'). The same holds for the concept C-96 represented by the term 'globalization', because some of the attributes describing it add further information, as in the case of attribute 'is_a process', where the concept identified by the term 'process' is in turn described in the taxonomy by C-104 as a combination of transformations ('is_a combination', 'about transformation). Here, we notice, again, that information associated with a concept can, in turn, be described by other formal concepts in the taxonomy, and can contribute to extending the acquired knowledge. The above concept is indeed correct, as it is expressed in the document text: "[...] Tom Palmer of the Cato Institute defines globalization as the

diminution or elimination of state-enforced restrictions on exchanges across borders and the increasingly integrated and complex global system of production and exchange [...]".

Table 3. Fragment of the conceptual network extracted from document D-2 (I).

- C-207
  - C-168
    {**third**};
    - C-95
      {**nation**}
      {mean_among};
    - C-142
      {**world**}
      {has_on,about_business,shrink_around,exert_on,mark_by};

    {location_term};
  - C-205
    {**internet,access,e_commerce,coop**}
    {bring_into};

  {bring_in}.
- C-196
  - C-124
    {**business**};
    {bring_into};
  - C-142
    {**world**}
    {has_on,about_business,shrink_around,exert_on,mark_by};

  {location_term, mark_by}

Table 4. Fragment of the conceptual network extracted from document D-2 (II).

- C-189
  {**palmer,tom**}
  {of_cato,of_institute};
- C-209
  - C-173
    {**cato,institute**};
    - C-174
      {**elimination,diminution**}
      - C-96
        {**globalization**}
        {has_on,is_into,is_a_process,refer_to, use,has_impact,use_in, shrink_around, mark_by,affect};

      {about_restriction};

    {define_on};
  - C-188
    {**system,border**}
    {location_term,about_production,about_exchange};

  {define_across}.
- C-104
  {**process**}
  {is_a_combination,about_transformation};

Moreover, in the case of C-219 ("Internationalization use the term globalization and refer to the importance", see Table 5) the corresponding fragment of conceptual taxonomy closely recalls what the text aims at expressing through the sentence "[...] The term 'internationalization' refers to the importance of international trade, relations, treaties, etc. [...]". The analyzed concept is defined by more specialized concepts, such as that

defined by the term 'importance' and described by attributes 'about trade', 'about relation' and 'about treaty', in this case contributing to add more information to the main concept by specifying what the importance refers to. Among the attributes that describe the concept 'globalization', 'has impact' specifies that globalization determines some kind of impact, where 'impact' is in turn defined by C-123, specifying that such an impact concerns a flattening ('about flattening'). In this case it was associated with the concept/entity 'Thomas Friedman' through the attribute 'examine', describing the concept expressed in the sentence "[...] Thomas Friedman examines the impact of the flattening' [...]" that is present in the document text. Again, the system identified the person entity Thomas Friedman as a concept, proving its ability to carry out some kind of entity recognition task (albeit in a very approximate way, but sufficient to provide a meaningful result).

Table 5. Fragment of the conceptual network extracted from document D-2 (III).

- C-219
  - C-146
    {**importance**};
    {about_trade, about_relation, about_treaty};
  - C-186
    {**internationalization**}
    - C-101
      {**globalization**}
      {has_on,is_into, is_a_process, define_across, define_on, about_restriction, has_impact, use_in, shrink_around, mark_by, affect};
    - C-144
      {**term**}
      {location_term, characterize_in, about_age, about_gender, about_background};

    {use};

  {refer_to}.
- C-214
  - C-123
    {**impact**};
    {has_on, about_ flattening, quicken_on, continue_on, exert_on};
  - C-190
    {**thomas,friedman**}
    {argue};

  {examine}

Finally, let us note that in D-2, the concept 'globalization' is one of those with the most detailed description in terms of the number of attributes that make up its intension. Therefore, it can be considered the main subject of the document, which is in fact true.

Based on the results obtained in D-2, we can conclude that processing outcomes on short documents are satisfactory.

As to the long document D-4, only a few of all concepts that are present in the taxonomy will be considered in order to focus the evaluation on particularly interesting aspects. The selected concepts are more complex compared to those extracted from the short document, which is an indication that the length and complexity of a document actually affect the definition of the final conceptual taxonomy.

One can easily guess that C-477 'Artificial Intelligence' represents the main concept in D-4 since it is described by a large number of attributes compared to the other concepts expressed in the taxonomy. This confirms that an analysis of the concepts represented in the taxonomy allows us to identify those that can be reasonably assumed to be the main concepts in the document. Since C-477 is related to many other concepts extracted from this document, we report it only once in Table 6, and will just refer to it in the other fragments. By observing the attributes that define the concept 'Artificial Intelligence', we note that some contribute to better specifying it. In particular, it is defined as a 'target' (through

attribute 'is_a goal'), and in turn the latter concept is described in the taxonomy as a 'goal about research'. This implies that 'Artificial Intelligence' is conceptually related to the concept 'research', and this relationship is confirmed by the portion of taxonomy reported in Table 6, which defines "the research of Artificial Intelligence is technical and specialized".

Table 6. Fragment of the conceptual network defining 'Artificial Intelligence', extracted from document D-4 (I).

- Concept 518
  {**research**}
  - Concept 477
    {**artificial**}
    - Concept 344
      {**intelligence**}
      {about_machine, about_branch, believe, build_in};

      {argue, is_a_goal, is_a_intelligence, define, is_a_subject, organize_around, include, appear_in, present_by, consider_in, transcend, reach, has_by};

    {is_technical, is_specialized}.

The portion of taxonomy defined by C-455 (Table 7) describes that "Artificial Intelligence is organized around difference about opinions, and around applications or problems about tools". It is worth noting that more detailed information that contributes to further specifying the concept being described can be deduced from the attributes of each sub-concept considered. In particular, concept C-329 'difference' is specified by attribute 'about_opinion', while the reference of C-496 'application' and C-370 'problem' to concept 'tool' is expressed by attribute 'about_tool'. By observing concept C-370 'problem', one can note that its attributes include 'of_Artificial' and 'of_Intelligence', which together are further confirmation of the previously noted relationship between concepts 'problem' and 'Artificial Intelligence'.

Table 7. Fragment of the conceptual network extracted from document D-4 (I).

- C-455
  - C-456
    - C-329
      {**difference**}
      {about_opinion};
    - C-496
      {**application**}
      - C-370
        {**problem**}
        {provide_for, of_artificial, of_intelligence};

      {about_tool};

    {location_term};
  - C-477 ['Artificial Intelligence'...]

  {organize_around}.

The taxonomy resulting from processing D-2 is not based only on the content of the text. It is also enriched by information deduced by the semantics associated with the kind of complements that were identified. The complexity of such a task is strictly related to the intrinsic complexity of natural language, which might mislead the system and cause it to return to incorrect outcomes. For instance, in this case, by observing the taxonomy, one can note that the system has wrongly identified concepts C-329 'difference' and C-496

'application, problem' as a 'location_term'; the blame for such kinds of errors can be put on the complexity of the English language, and not on processing errors of the system.

Another fragment of conceptual taxonomy (shown in Table 8) extracted from D-2 describes the concept expressed in the passage of the document "[...] the film of Artificial Intelligence considers a machine in the form of a small boy [...]" and additionally specifies some involved concepts. For instance, C-474 associates the concepts 'machine' and 'Artificial Intelligence', by specifying that in this context concept C-343 'machine' refers to machines appearing in the Artificial Intelligence field. By observing the attributes that describe this concept, we note that the system associated 'machine' with the concept 'computer science' and with the concepts/entities 'Yan Shi', 'Hero Alexandria', 'Al-Jazari' and 'Wolfang vonKempelen'. This reflects the following sentence that appears in the text: "[...] including the machines of Yan Shi, Hero of Alexandria and Al-Jazari or Wolfgang von Kempelen [...]".

Table 8. Fragment of the conceptual network extracted from document D-4 (II).

- C-462
  - C-334
    {**form**}
    {location_term, about_boy};
  - C-474
    - C-343
      {**machine**}
      {about_computer, about_science, simulate_by, of_yan, of_shi, of_hero, of_alexandria, of_al-jazari, of_wolfgang, of_vonkempelen};
    - C-477 ['Artificial Intelligence'...]
      {include, appear_in};

  {consider_in}.

Another example of how the taxonomy enriches the information expressed is represented by concept C-334 'form', described through attribute 'about boy' specifying that, in this particular context, the appearence of these machines is that of a boy, while the attribute 'location_term' is an error, since it would identify the concept as a place.

The last portion of taxonomy we will analyze (shown in Table 9) represents quite a complex concept C-457, textually described as "Texbooks define Artificial Intelligence as a field about the study and design of agent or as engineering and science on the intelligence about machine". In this case, the attributes concur in defining such a concept in more detail. Concept C-458, represented by the terms 'field, design, study', is described by 'about agent', which allows us to define more accurately the concept itself and also establishes a conceptual association between 'field, design, study' and the concept 'agent', which in this context is defined as a system that perceives the external environment. Such a definition is induced by analyzing and integrating the description associated with the concept 'agent' with that associated with the concept 'system'.

Summing up, we may say that in both short and long documents the information expressed by the extracted taxonomy satisfactorily reflects the contents of the document from which they were extracted. Nevertheless, a few errors are still present due to the complexity of natural language, concerning the description of some concepts, which might mislead the reader in interpreting the associated information.

4.2.2. Concept Extraction from Groups of Documents

We now consider the combination of several documents from the same domain, both in the short category and in the long one. In this case, we will not show the fragments of taxonomy, but will just provide their description and interpretation.

Table 9. Fragment III of the conceptual network extracted from document D-4.

- C-457
  {**textbook**}
  - C-458
    {**design**,**study**}
    - C-330
      {**field**}
      {find_on};

    {about_agent};
  - C-472
    {**engineering**}
    - C-41
      {**science**}
      {location_term, create, aim, provide_in};
    - C-344
      {**intelligence**}
      {argue, is_a_goal, about_branch, is_a_intelligence, is_a_subject, is_technical, is_specialized, organize_around, include, appear_in, believe, build_in, present_by, consider_in, transcend, reach, has_by};

    {about_machine};
  - C-477 ['Artificial Intelligence'...]

  {define}.

As to the analysis of short documents, let us consider the taxonomies extracted from D-1 and D-3, both dealing with sports subjects that are very close to each other. The former deals with aspects of American Football concerning the rules of the game, while the latter discusses Rugby Union with references to the regulations of the specific sport discipline. Some 'larger' concepts, as to the number of attributes used to describe them, can be identified. They represent the main concepts in those documents and allow us to frame the subject these documents deal with. Indeed, by analyzing the taxonomy associated with D-1, the relevant concept is represented by the term 'American football', and in the taxonomy extracted from D-3 the relevant concept turns out to be 'rugby union', which perfectly catch the subjects of the two documents. The probability of identifying a list of relevant concepts in a taxonomy is strictly related to the accuracy and precision by which the subject is discussed in the document itself. If the subject is generic, outstanding concepts will be hardly identified in the conceptual taxonomy.

The concept identified by the term 'rugby football' turns out to be shared by both documents; while in the former, it is associated with the concept 'sport', identifying a significantly wider concept that defines American Football as "a sport born from Rugby Football". In the latter, it is associated with the concepts 'rugby union' and 'region' (where the latter is correctly identified by the system as a place), to describe the concept that "in some regions by Rugby Union people refers to Rugby Football". As a consequence, 'rugby football' does not represent the only concept shared by both documents. Indeed, among attributes describing the concept 'rugby football' there is 'is_a league', which defines a relationship between 'rugby football' and 'league', and a relationship between the concepts 'rugby football' and 'sport' that, in addition to being present in D-1 as well, is also identified in D-3, in that 'is_a sport' is one of the many attributes describing concept 'football rugby'. Moreover, the concepts 'sport' and 'league' describe the domain of both documents and indeed represent frequent concepts found in both taxonomies.

In both documents, the system identified an association between 'sport' and 'team', which, in D-3, defines rugby as "a sport played by teams', and in D-1, additionally associates it with the concepts 'team play' and 'strategy', and that define American Football as a "team sport famous for team play with strategies". By observing the attributes describing

the concept 'team', it is particularly interesting to note that in both taxonomies there is a relationship to the concept 'player' induced by the presence of attributes 'about player' and 'has player' that describe the concept 'team'. Thus, not only can one conclude that both 'team' and 'player' allow us to specify the domain of both documents but also that the relationships identified in both documents between the concepts 'team' and 'sport' and that between 'team' and 'player' are specific to the domain to which the two documents belong.

The concept 'ball' is associated with the concept 'zone' in one document, and to the concept 'area' in the other, both defined by attributes 'about opponent' and 'location_term'; this means that the two concepts are equivalent, and related to the concept 'ball'. Both express the same idea that "the ball must be kicked or grounded in the opponent's area (or zone)". A similar aspect can be highlighted from the results concerning the concept 'goal'. Indeed, it is associated in both documents with concepts described by the attributes 'location_term' and 'about_opponent'. While in D-1 it turns out that the concept 'goal' is associated both with the concept 'line' to define the concept 'goal line', and with the concept 'post', which describes the concept 'goal posts'; conversely, in D-3 it turns out that the concept 'goal' is associated with 'crossbar' and defines the concept 'goal crossbar'. This combination of results allows us to determine a similarity between the concepts 'goal line', 'goal post' and 'goal crossbar'.

As a final evaluation of the results obtained on the selected short documents, it turns out that the shared concepts identified by the system in both documents are able to generally define the portion of knowledge shared between the two documents, proving that the subjects they described were actually closely related to each other.

As to long documents, we compared the taxonomies extracted from D-5 and D-6, both concerning the general subject of a national politics system (the Italian one in the former case and that of the USA in the latter). Differently from the previous documents, here, in both cases, it is difficult to identify outstanding concepts in the taxonomies that allow us to properly describe the subject of the two documents. This is further confirmation that the subjects described in documents are identified by the concepts that are relevant in the respective taxonomies.

Some concepts can be found in both documents. The concept 'president' in D-5 is one of the most relevant/frequent, and one of the most detailed in terms of describing attributes as well, and is also present in the taxonomy associated with D-6, although with a lesser prominence in terms of relationships in the taxonomy. This allows us to identify 'president' as a domain concept. In both results, the concept 'president' is described, among others, by the attribute 'is_a commander-in-chief', which establishes a relationship between the concepts 'president' and 'commander-in-chief'. This latter concept is similarly defined in the two taxonomies: in D-5 as a 'commander in chief of armed forces', and in D-6 as a 'commander in chief of the army'. In both taxonomies, three inter-related concepts are defined. In D-5, the concept 'judicial, power' is described by the attribute 'is independent' and establishes a relationship to the concept 'independent', which in turn is specified by the concepts 'executive' and 'legislative' to define that "the judicial power is a power independent from the legislative and executive ones"; conversely, in D-6, the concepts 'executive', 'legislative' and 'judicial' are associated with the concepts 'branch' and 'government' to express the fact that "government is made up by the legislative branch, the executive branch and the judicial one".

Another very interesting aspect concerns the concept 'republic', exploited to define the political organization both in Italy and in the U.S. While in one document/taxonomy it is associated with the concept 'Italy' and 'framework' to say that "Italy is structured as a democratic republic of parliamentary kind with a many-partitic system", in the other, it is associated with the concepts "representative, democracy' and 'united, states' to specify that "the United States are a federal republic structured as a representational democracy". In both cases, there is a reference to the concept 'democracy', but on one side specifies the kind of republic that is present in Italy, and the other defines how the U.S. federal republic is structured.

Another concept that is shared by both documents, and is somehow also related to the reference context, is that defined by the terms 'electoral system'. This is associated in D-5 to the concepts 'representation' and 'majority prize' to specify that in Italy "the election system combines a proportional representation with majority shares", and in D-6 to concepts 'college', 'vote' and 'state' to express the fact that in the USA "the election system is of a collegial kind, where votes are distributed among states". In both documents, the system identified the concept 'supreme, court', which hence should be considered as a domain concept.

Summing up, we can claim that the concepts extracted by the system in both taxonomies are able to describe a shared portion of knowledge between the two considered documents, which confirms what we observed during the evaluation carried out on short documents.

5. Conclusions

This work contributes to the investigation of the possibility of extracting conceptual taxonomies from text documents in the English language. Such an extraction was carried out through cooperation between two inter-related steps: a sentence parsing that is able to distinguish different kinds of linguistic complements and their semantics, whose outcome is exploited by Formal Concept Analysis techniques to extract the actual concepts and arrange them in a taxonomy. Specifically, here, the way these two steps are put together is of interest. FCA techniques are well-known and widely exploited in the literature, while the architecture of the parsing system is outside the scope of this paper.

The final results of the Concept Hierarchy Extraction task on sample documents having different properties are satisfactory and encouraging from both a quantitative and a qualitative performance perspective. Overall, the results are somehow affected both by the document size and by the kind of subject it deals with. When processing different documents dealing with similar subjects, the Concept Hierarchy Extraction task is able to identify some common concepts in both taxonomies, which allows us to define portions of shared knowledge between the processed documents.

These outcomes allow us to identify further work directions aimed at improving and optimizing the proposed technique and its implementation. Among such future work directions: supporting the Concept Hierarchy Extraction step by further techniques (e.g., Keyword Extraction) that can point out prominent or important concepts to be included in the taxonomy in order to reduce the amount of data to be considered by selecting peculiar formal concepts and attributes; investigating the possibility of expanding the generated taxonomy through the recognition of instances of the taxonomy concepts in the documents; integrating the Concept Hierarchy Extraction with Name Entity Recognition capabilities that allow us to identify the concepts expressed in the text as entities of the document domain; and introducing algorithms to compress the generated conceptual taxonomy, so that the most relevant concepts in the document can be considered. We are also aware that the representation of concepts is not very intuitive, and that an overall view of the taxonomy is complex to grasp. For this reason, additional future work will also concern developing techniques that will allow the interpretation and browsing of the extracted elements.

Funding: This research received no external funding.

Data Availability Statement: Data used in this work are reported in the Appendix A.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CI	Concept Indexing
FCA	Formal Concept Analysis
IE	Information Extraction
NLP	Natural Language Processing

Appendix A. Full Text of Documents Used for Evaluation

D-1

American football is known in the United States and Canada simply as football and it is a competitive team sport known for mixing strategy with physical play. The objective of the game is to score points by advancing the ball into end zone of the opposing team. The ball can be advanced by carrying it, a running play, or by throwing it to a teammate, a passing play. Points can be scored in a variety of ways, including carrying the ball over goal line of the opponent, catching a pass from beyond that goal line, kicking the ball through the goal posts at end zone of the opponent and tackling an opposing ballcarrier within his end zone. The winner is the team with the most points when the time expires. The sport is also played outside the United States. National leagues exist in Germany, Italy, Switzerland, Finland, Sweden, Japan, Mexico, Israel, Spain, Austria and several Pacific Island nations. The National Football League is the largest professional American football league in the world. American football is closely related to Canadian football, but with significant differences. Both sports originated from rugby football. Game play in American football consists of a series of downs, individual plays of short duration, outside of which the ball is dead or not in play. These can be plays from scrimmage—passes, runs, punts, or field goal attempts—or free kicks such as kickoffs. Substitutions can be made between downs, which allows for a great deal of specialization as coaches choose the players best suited for each particular situation. Each team should have 11 players on the field during a play, and each player has specific tasks assigned for that specific play.

D-2

Globalization (or globalisation) is the process of transformation of local or regional phenomena into global ones. This process is a combination of economic, technological, sociocultural and political forces. Globalization is often used to refer to economic globalization, that is, integration of national economies into the international economy through trade, foreign direct investment, capital flows, migration, and the spread of technology. Tom Palmer of the Cato Institute defines "globalization" as "the diminution or elimination of state-enforced restrictions on exchanges across borders and the increasingly integrated and complex global system of production and exchange that has emerged as a result". Thomas Friedman "examines the impact of the 'flattening' of the globe", and argues that globalized trade, outsourcing, supply-chaining, and political forces have changed the world permanently, for both better and worse. Thomas Friedman argues also that the pace of globalization is quickening and will continue to have a growing impact on business organization and practice. Noam Chomsky argues that the word globalization is also used, in a doctrinal sense, to describe the neoliberal form of economic globalization. Herman Daly argues that sometimes the terms internationalization and globalization are used interchangeably but there is a slight formal difference. The term "internationalization" refers to the importance of international trade, relations, treaties etc. International means between or among nations. Globalization has had extensive impact on the world of business. In a business environment marked by globalization, the world seems to shrink, and other businesses halfway around the world can exert as a great impact on a business as one right down the street. Internet access and e-commerce have brought small-scale coops in Third World nations into the same arena as thriving businesses in the industrialized world, and visions of low-income workers handweaving rugs on primitive looms that compete with rug dealers in major cities are not totally far-fetched. Globalization has affected workforce demographics, as well. Today's workforces are characterized by greater diversity in terms of age, gender, ethnic and racial background, and a variety of other demographic factors. In fact, management of diversity has become one of the primary issues of 21st-century business. Trends such as outsourcing and offshoring are a direct offshoot of globalization and have created a work environment in which cultural diversity can be problematic.

D-3

Rugby union is an outdoor contact sport played with a prolate spheroid-shaped ball by two teams of 15 players. It is one of the two main codes of rugby football, the other being rugby league. There is also a seven-a-side variant named rugby sevens, which is played under modified rules and with only seven players per team. Rugby union is often referred to as simply rugby or football, and in regions where rugby league is played, as union. A rugby union match lasts for 80 min, (plus stoppage time), with a short interval (not more than 10 min) after the first 40 min. At under-19 level and below, games are limited to a maximum of 70 min, with an interval after 35 min. A match is controlled by a referee, who is assisted by two touch judges or assistant referees. For professional matches, a Television Match Official commonly called the video referee is often employed, usually to advise the referee on matters pertaining to the scoring of tries and dropped goals. The players form then a ruck in order to win the ball back. On some (usually rare) occasions, a team may be awarded a penalty try, if their opponents commit a foul which is deemed by the referee to have prevented a probable try, for example collapsing a scrum or maul close to the try line. The object of the game is to score as many points as possible. The winner is the team that scores the greater number of points. Points are awarded for scoring a try or kicking a goal. A try, which is worth 5 points, is scored when the ball is grounded by a player on the attacking team within the in-goal area of opponent. A goal is scored by kicking the ball over the crossbar of the goal of opponent while remaining between the posts. There are three ways to score a goal: a dropped goal (scored in open play where the ball must hit the ground immediately before it is kicked), a penalty goal (awarded after the opposing side infringes against the laws of rugby and may be kicked from a stationary ground position or by drop kick), and a conversion (awarded after a try is scored) by either a drop kick or a place kick.

D-4

Artificial intelligence is the intelligence of machines and the branch of computer science which aims to create it. Major Artificial Intelligence textbooks define the field as "the study and design of intelligent agents", where an intelligent agent is a system that perceives its environment and takes actions which maximize its chances of success. John McCarthy, who coined the term in 1956, defines it as "the science and engineering of making intelligent machines". The field was founded on the claim that a central property of human beings, intelligence—the sapience of Homo sapiens—can be so precisely described that it can be simulated by a machine. This raises philosophical issues about the nature of the mind and limits of scientific hubris, issues which have been addressed by myth, fiction and philosophy since antiquity. Artificial intelligence has been the subject of breathtaking optimism, has suffered stunning setbacks and, today, has become an essential part of the technology industry, providing the heavy lifting for the most difficult problems in computer science. Artificial Intelligence research is highly technical and specialized, so much so that some critics decry the "fragmentation" of the field. Subfields of Artificial Intelligence are organized around particular problems, the application of particular tools and around long standing theoretical differences of opinion. The central problems of Artificial Intelligence include such traits as reasoning, knowledge, planning, learning, communication, perception and the ability to move and manipulate objects. General intelligence (or "strong Artificial Intelligence") is still a long term goal of (some) research. Thinking machines and artificial beings appear in Greek myths, such as Talos of Crete, the golden robots of Hephaestus and Pygmalion's Galatea. Human likenesses believed to have intelligence were built in every civilization, beginning with the sacred statues worshipped in Egypt and Greece, and including the machines of Yan Shi, Hero of Alexandria and Al-Jazari or Wolfgang von Kempelen. It was widely believed that artificial beings had been created by Geber, Judah Loew and Paracelsus. Stories of these creatures and their fates discuss many of the same hopes, fears and ethical concerns that are presented by Artificial Intelligence. The idea also appears in modern science fiction: the film of Artificial Intelligence considers a machine in the form of a small boy which has been given the ability to feel human emotions

including the capacity to suffer. This issue, now known as "robot rights", is currently being considered by California's Institute for the Future, although many critics believe that the discussion is premature. Another issue explored by both science fiction writers and futurists is the impact of Artificial Intelligence on society. In fiction, Artificial Intelligence has appeared as a servant (R2D2 in Star Wars), a comrade (Commander Data in Star Trek), a conqueror (in The Matrix), an exterminator (in Terminator, Battlestar Galactica), a race (Asurans in Stargate Atlantis). Academic sources have considered such consequences as: a decreased demand for human labor, the enhancement of human ability or experience and a need for redefinition of human identity and basic values. Several futurists argue that Artificial Intelligence will transcend the limits of progress and fundamentally transform humanity. Ray Kurzweil has used law of Moore (which describes the relentless exponential improvement in digital technology with uncanny accuracy) to calculate that desktop computers will have the same processing power as human brains by the year 2029, and that by 2045 Artificial Intelligence will reach a point where it is able to improve itself at a rate that far exceeds anything conceivable in the past, a scenario that science fiction writer Vernor Vinge named the "technological singularity".

D-5

The politics of Italy take place in a framework of a parliamentary, democratic republic and a multi-party system. Executive power is exercised collectively by the Council of Ministers, which is led by a president, informally referred to Premier or Prime Minister. Legislative power is vested in the two houses of parliament primarily, and secondarily on the Council of Ministers. The judicial power is independent from the executive and the legislative. Italy has been a democratic republic since 2 June 1946, when the monarchy was abolished by popular referendum. The constitution was promulgated on 1 January 1948. The president of the Italian Republic is elected for seven years by the parliament sitting, jointly with a small number of regional delegates. The president represents the unity of the nation and has many of the duties previously given to the King of Italy. The president serves as a point of connection between the three branches of power. The president is elected by the lawmakers, appoints the executive, is the president of the judiciary, and the president is also the commander-in-chief of armed forces. The president nominates the Prime Minister, who proposes the other ministers and formally named by the president. The Council of Ministers must obtain a confidence vote from both houses of parliament. Legislative bills may originate in either house and must be passed by a majority in both. Italy elects a parliament consisting of two houses, namely the Chamber of Deputies, which has 630 members, and the Senate of the Republic, that comprise 315 elected members and a small number of senators for life. As of 15 May 2006, there are seven life senators of whom three are former presidents. Both houses are elected for a maximum of five years, but both may be dissolved by the president before the expiration of their normal term, if the parliament cannot elect a stable government. Legislation may originate in either house and must be passed in identical form by a majority in each. The houses of parliament are popularly and directly elected through a complex electoral system which combines proportional representation with a majority prize for the largest coalition (Chamber). All Italian citizens older than 18 can vote. However, to vote for the senate, the voter must be at least 25 or older. The electoral system of the Senate is based upon regional representation. The Italian Parliament has a peculiarity, that is the representation given to Italian citizens permanently living abroad, about 2.7 million people. Among the 630 Deputies and the 315 Senators there are, respectively, 12 and 6 elected in four distinct overseas constituencies. Those members of parliament were elected for the first time in April 2006 and they have the same rights as members elected in Italy. The Italian judicial system is based on Roman law that it is modified by the Napoleonic code and later statutes. The Supreme Court of Cassation is the court of last resort for most disputes. The Constitutional Court rules on the conformity of laws with the Constitution and is a post-Second World War innovation.

D-6

The United States is oldest surviving federation of the world. The United States is a constitutional republic, "where majority rule is tempered by minority rights protected by law". The United States is fundamentally structured as a representative democracy, though American citizens are excluded from voting for federal officials. The government is regulated by a system of checks and balances defined by the American Constitution, which serves as the country's supreme legal document and as a social contract for the American people. In the American federalist system, citizens are usually subject to three levels of government, federal, state, and local. The federal government is composed of three branches: Legislative, Executive and Judicial. The bicameral Congress is made up of the Senate and the House of Representatives, and makes federal law, declares war, approves treaties, has the power of the purse and impeachment, by which the Congress can remove sitting members of the government. The president is the commander-in-chief of the military, can veto legislative bills before they become law and appoints the Cabinet and other officers, who administer and enforce federal laws and policies. The Supreme Court and lower federal courts, whose judges are appointed by the president and with Senate approval, appoints and interpret laws, and can overturn laws they deem unconstitutional. The House of Representatives has 435 members, each representing a congressional district for a two-year term. House seats are apportioned among the states by population after tenth year. As of the 2000 census, seven states have the minimum of one representative, while California, the most populous state, has fifty-three representatives. The Senate has 100 members with each state having two senators, elected at-large to six-year terms, one third of Senate seats are up for election every other year. The president serves a four-year term and may be elected no more than twice. The president is not elected by direct vote, but by an indirect electoral college system where the determining votes are apportioned by state. The Supreme Court, led by the Chief Justice of the United States, has nine members, who serve for life. The state governments are structured in roughly similar mode, Nebraska uniquely has a unicameral legislature. The governor (chief executive) of each state is directly elected. All laws and procedures of both state and federal governments are subject to review, and any law ruled in violation of the Constitution by the judiciary is voided. The original text of the Constitution establishes the structure and responsibilities of the federal government and its relationship with the individual states. Article One protects the right to the "great writ" of habeas corpus, and Article Three guarantees the right to a jury trial in all criminal cases. Amendments to the Constitution require the approval of three-fourths of the states. The Constitution has been amended twenty-seven times, the first ten amendments, which make up the Bill of Rights, and the Fourteenth Amendment form the central basis of Americans' individual rights.

References

1. Feldman, R.; Dagan, I. Kdt—knowledge discovery in texts. In Proceedings of the First International Conference on Knowledge Discovery (KDD), Montreal, QC, Canada, 20–21 August 1995; pp. 112–117.
2. Gaizauskas, R. *A Research Perspective on Text Mining: Tasks, Technologies and Prototype Applications*; Technical Report; University of Sheffield: Sheffield, UK, 2003.
3. Nahm, U.; Mooney, R. Text mining with information extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, Palo Alto, CA, USA, 25–27 March 2002.
4. Hearst, M. Untangling text data mining. In Proceedings of the ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, USA, 20–26 June 1999.
5. Etzioni, O. The Word Wide Web: Quagmire or gold mine? *Commun. ACM* **1996**, *39*, 65–68. [CrossRef]
6. Hotho, A.; Nurnberger, A.; Paaß, G. A Brief Survey of Text Mining. *J. Lang. Technol. Comput. Linguist.* **2005**, *20*, 19–62. [CrossRef]
7. Velardi, P.; Navigli, R.; Cucchiarelli, A.; Neri, F. Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In *Ontology Learning from Text: Methods, Applications and Evaluation*; IOS Press: Amsterdam, The Netherlands, 2005.
8. Miller, G.A. WORDNET: A Lexical Database for English. In Proceedings of the Workshop on Speech and Natural Language, Harriman, NY, USA, 23–26 February 1992; Morgan Kaufmann: Burlington, MA, USA, 1992.
9. Harris, Z. *Mathematical Structures of Language*; Wiley: Hoboken, NJ, USA, 1968.
10. Maedche, A.; Staab, S. The TEXT-TO-ONTO Ontology Learning Environment. In Proceedings of the ICCS-2000—Eight International Conference on Conceptual Structures, Software Demonstration, Darmstadt, Germany, 14–18 August 2000.

11. Maedche, A.; Staab, S. Mining Ontologies from Text. In Proceedings of the EKAW, Juan-les-Pins, France, 2–6 October 2000; pp. 189–202.
12. Salton, G.; Wong, A.; Yang, C. A Vector Space Model for Automatic Indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
13. Landauer, T.; Foltz, P.; Laham, D. An Introduction to Latent Semantic Analysis. *Discourse Process.* **1998**, *25*, 259–284. [CrossRef]
14. Wall, M.; Rechtsteiner, A.; Rocha, L. Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*; Berrar, D., Dubitzky, W., Granzow, M., Eds.; Kluwer Academic Publishers: New York, NY, USA, 2003; pp. 91–109.
15. Karypis, G.; Han, E.H.S. Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization & Retrieval. In Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM), McLean, VA, USA, 6–11 November 2000; pp. 12–19.
16. Fellbaum, C., Ed. *WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998.
17. Hensman, S. Construction of conceptual graph representation of texts. In Proceedings of the Student Research Workshop at HLT-NAACL, Boston, MA, USA, 2–7 May 2004; HLT-SRWS '04; Association for Computational Linguistics: Cedarville, OH, USA, 2004; pp. 49–54.
18. Kipper, K.; Dang, H.T.; Palmer, M. Class-Based Construction of a Verb Lexicon. In Proceedings of the 17th NCAI and 12th IAAI Conference, Austin, TX, USA, 30 July–3 August 2000; AAAI Press: Washington, DC, USA, 2000; pp. 691–696.
19. Meijer, K.; Frasincar, F.; Hogenboom, F. A semantic approach for extracting domain taxonomies from text. *Decis. Support Syst.* **2014**, *62*, 78–93. [CrossRef]
20. Hoxha, J.; Jiang, G.; Weng, C. Automated learning of domain taxonomies from text using background knowledge. *J. Biomed. Inform.* **2016**, *63*, 295–306. [CrossRef] [PubMed]
21. Ogata, N. A Formal Ontology Discovery from Web Documents. In *Web Intelligence: Research and Development, First Asia-Pacific Conference (WI 2001)*; Number 2198 in Lecture Notes on Artificial Intelligence; Springer: Berlin/Heidelberg, Germany, 2001; pp. 514–519.
22. Shamsfard, M.; Barforoush, A. Learning Ontologies from Natural Language Texts. *Int. J. Hum.-Comput. Stud.* **2004**, *60*, 17–63. [CrossRef]
23. Cimiano, P.; Hotho, A.; Staab, S. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Int. Res.* **2005**, *24*, 305–339. [CrossRef]
24. Hasegawa, R.; Kitamura, M.; Kaiya, H.; Saeki, M. Extracting conceptual graphs from Japanese documents for software requirements modeling. In Proceedings of the APCCM '09, Sixth APCCM, Wellington, New Zealand, 20–23 January 2009; Volume 96, pp. 87–96.
25. Ferilli, S.; Izzi, G.L.; Franza, T. Automatic Stopwords Identification from Very Small Corpora. In *Intelligent Systems in Industrial Applications*; Stettinger, M., Leitner, G., Felfernig, A., Ras, Z.W., Eds.; Springer: Cham, Switzerland, 2021; pp. 31–46.
26. Ferilli, S.; Esposito, F.; Redavid, D.; Angelastro, S. Language Identification as Process Prediction Using WoMan. In *Digital Libraries and Archives*; Grana, C., Baraldi, L., Eds.; Springer: Cham, Switzerland, 2017; pp. 159–172.
27. Ferilli, S. Automatic Multilingual Stopwords Identification from Very Small Corpora. *Electronics* **2021**, *10*, 2169. [CrossRef]
28. Ferilli, S.; Esposito, F.; Grieco, D. Automatic Learning of Linguistic Resources for Stopword Removal and Stemming from Text. *Procedia Comput. Sci.* **2014**, *38*, 116–123. [CrossRef]
29. Ferilli, S.; Angelastro, S. Towards a Process Mining Approach to Grammar Induction for Digital Libraries. In *Digital Libraries: Supporting Open Science*; Manghi, P., Candela, L., Silvello, G., Eds.; Springer: Cham, Switzerland, 2019; pp. 291–303.
30. Rotella, F.; Leuzzi, F.; Ferilli, S. Learning and Exploiting Concept Networks with ConNeKTion. *Appl. Intell.* **2015**, *42*, 87–111. [CrossRef]
31. Wille, R. Restructuring lattice theory: An approach based on hierarchies of concepts. In *Ordered Sets*; Rival, I., Ed.; Reidel Dordrecht: Boston, MA, USA, 1982; pp. 445–470.
32. Davis, R.; Shrobe, H.; Szolovits, P. What is a knowledge representation? *AI Mag.* **1993**, *14*, 17.
33. Ganter, B.; Wille, R. *Formal Concept Analysis: Mathematical Foundations*; Translation of Formale Begriffsanalyse Mathematische Grundlagen; Springer: Berlin/Heidelberg, Germany, 1999.
34. Bordat, J. Calcul pratique du treillis de Galois d'une correspondance. *Math. Sci. Hum.* **1986**, *96*, 31–47.
35. Ganter, B. *Two Basic Algorithms in Concept Analysis*; Technical Report FB4-Preprint No.831; TH Darmstadt: Darmstadt, Germany, 1984.
36. Chein, M. Algorithme de recherche des sous-matrices premières d'une matrice. *Bull. Math. Soc. Sci. Math. Répub. Soc. Roum.* **1969**, *13*, 21–25.
37. Godin, R.; Missaoui, R.; Alaoui, H. Incremental Concept Formation Algorithms Based on Galois (concept) Lattices. *Comput. Intell.* **1995**, *11*, 246–267. [CrossRef]
38. Norris, E. An Algorithm for Computing the Maximal Rectangles in a Binary Relation. *Rev. Roum. Math. Pures Appl.* **1978**, *23*, 243–250.
39. Nourine, L.; Raynaud, O. A Fast Algorithm for Building Lattices. *Inf. Process. Lett.* **1999**, *71*, 199–204. [CrossRef]