Learning Non-taxonomic Relationships of Financial Ontology

Omar El Idrissi Esserhrouchni[1], Bouchra Frikh[2] and Brahim Ouhbi[1]
[1]*Lab LM2I, ENSAM, Moulay Ismaïl University, Marjane II, B.P. 4024, Meknès, Morocco*
[2]*Lab LTTI, ESTF, Sidi Mohamed Ben Abdellah University, B.P. 2427, Route d'imouzer, Fès, Morocco*

Keywords: Ontology Learning, Financial Ontology, Non-taxonomic Relationships Extraction, Knowledge Acquisition, Open Information Extraction.

Abstract: Finance ontology is, in most cases, manually addressed. This results in a tedious development process and error prone that delay their applicability. This is why there is a need of domain ontology learning methods that built the ontology automatically and without human intervention. However, in this learning process, the discovery of non-taxonomic relationships has been recognized as one of the most difficult problems. In this paper, we propose a new methodology for learning non-taxonomic relationships and building financial ontology from scratch. Our new technique is based on using and adapting Open Information Extraction algorithms to extract and label domain relations between concepts. To evaluate our new method effectiveness, we compare the extracted non-taxonomic relations of our algorithm with related works in the same finance corpus. The results showed that our system is more accurate and more effective.

1 INTRODUCTION

Financial ontologies are created to solve widespread financial problems such as support of financial decision making, market research and analysis, investment recommendations or assessing the financial health of a company.

Financial ontology engineering is typically performed manually, requiring the intervention of (1) financial domain expert who provide the financial knowledge, (2) and knowledge engineers who are able to formalize that knowledge. However, the field of finance is a conceptually rich domain where information is diverse, huge in volume and obtained from a variety of heterogeneous sources. A massive amount of valuable information is produced worldwide every day in the web. In fact, the manual development of these ontologies is costly, time consuming, tedious and error prone task, which delay the applicability of the resulting ontologies (Shamsfard and Barforoush, 2003; Sanchez and Moreno, 2008).

Due to these reasons, nowadays, there is a need for methods and processes that can build finance ontology automatically and without human intervention. In this sense, domain ontology learning was identified as the process of building domain ontology from scratch, enriching, or adapting an existing ontology in an unsupervised way (Maedche and Staab, 2001). This process reduces the time and effort needed in the ontology development process. Domain ontology learning is still an emerging field, which aims at assisting knowledge engineers in ontology construction.

In the literature several approaches have been proposed for learning domain ontology. Nevertheless, most of these approaches address only the way to learn the taxonomic part of domain ontologies. The phase of extraction of non-taxonomic relationships has been recognized as one of the most difficult and least tackled problems (Villaverde et al., 2009). This phase includes two different problems: (1) discovering the existence of relevant relationships between concepts and (2) labeling these relationships according to their semantic meaning. Moreover, in the ontology learning process, non-taxonomic relationships identification layer uses background knowledge from domain taxonomy in order to extract non-taxonomic relations from domain corpus (Buitelaar et al. 2005). However, the precision and the quality of the extracted ontology relations will highly depend on the accuracy of its domain taxonomy given as input.

In previous works (El Idrissi Esserhrouchni et al., 2014; Frikh et al., 2011), we covered the

Esserhrouchni, O., Frikh, B. and Ouhbi, B..
Learning Non-taxonomic Relationships of Financial Ontology.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 479-489
ISBN: 978-989-758-158-8
Copyright © 2015 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

learning of taxonomic relationships for domain ontology and introduced an efficient system that builds financial taxonomy more accurately than other benchmark algorithms. In this paper, we present a new methodology for learning non-taxonomic relationships and building financial ontology from scratch. Our new technique is based on integrating and adapting Open Information Extraction (Open IE) algorithms to extract and label domain relations between concepts. Indeed, Open IE algorithms, extract automatically and without human intervention all existing relations from a large text corpus. Due to its open-domain and open-relation nature, gross use of Open IE algorithms is unable to relate the extracted relations to domain ontology. However, an adaptation of Open IE algorithms is necessary to overcome this limitation.

Open IE approaches were introduced recently by Banko et al., (2007). So there is only a small amount of projects using it. To our knowledge, they were never used as a part of the process of learning domain ontology from scratch. Our work will be the first one that adapts these types of algorithms to learn domain ontology, especially for finance domain.

The rest of the paper is organized as follows. Section 2 presents related works. Section 3 describes the general steps for learning domain ontology. Then in Section 4, our proposed algorithm is described. In Section 5, the architecture of our system is presented. In Section 6, an evaluation of our proposed algorithm is carried out in term of precision. We integrate our system with three well-known Open IE tools: Reverb (Fader et al., 2011), Ollie (Schmitz et al., 2012) and ClausIE (Del Corro and Gemulla, 2013). Then we compare it with the algorithm of Sanchez and Moreno (2008), in finance corpus, to evaluate the relevance of the resulting non-taxonomic relationships of each algorithm.

2 RELATED WORKS

Some techniques have already been proposed for learning non-taxonomic relationships of domain ontology. Most of them combine different levels of machine learning and linguistic analysis.

(Maedche and Staab, 2000) propose a semi-automatic method for learning domain ontology. It uses generalized association rules to identify relationships between pairs of words and propose the best possible level in the hierarchy where to add the relationships. Nevertheless, this technique does not address the problem of labeling relationships. As consequence, users have to complete this task manually and without any assistance. In the same context, (Villaverde et al., 2009) use the strength of the association between concepts and verb given by POS-tagging rules to suggest multiple labels relationship between concepts. They identify non-taxonomic relationships by the presence of two concepts of the taxonomy in the same sentence with a verb between them at maximum of *N* terms. One limitation of this method is that the authors refer to the verbs and the concepts as single words when in fact, in most cases, they appear in the form of verb phrases. However, the accuracy will be reduced and the recall will be low.

Alternatively, Jiang and Tan (2005) attempt to acquire non-taxonomic relationships between concepts using regular expressions and natural language processing. Their algorithm performs a full parsing of the entire corpus using the Berkeley Parser (Petrov et al., 2006) and inspects the entire corpus to identify instances of patterns that indicate non-taxonomic relationships. However, the use of regular expressions limits the identification of the relations in the corpus. It may lack relevant relationships that are not recognized by the used patterns. An other important element to note is that their method involves a full parsing from text, which allows better extraction of concepts, but it is very expensive in time.

Other works combine natural language processing and statistical measure to extract and select relevant non-taxonomic relationships for domain ontology. OntoGain (Drymonas et al., 2010) selects verbs frequently occurring in the context of pairs of concepts for labeling semantic relationships. In that work, the association between concepts is assessed using a measure of the conditional frequency of a pair of concepts given a verb. In Sanchez and Moreno (2008) paper, non-taxonomic relationships extraction is based on (1) morphological and syntactic analysis to extract verbs and concepts that have a relationship with the domain keyword, (2) web scale statistics to refine the extracted relations. To avoid the natural language complexity, they apply some restriction, for example: only sentences with present tenses are used, verb phrases containing modifiers in the form of adverbs are rejected. Thus, to learn non-taxonomic relationships, for each sentence that contain the domain keyword, relationships represented by tuples in the form *(np1, vp, np2)* are extracted and evaluated for every verb phrase *(vp)* with the first noun phrase to its left *(np1)* and the second noun phrase to its right *(np2)*.

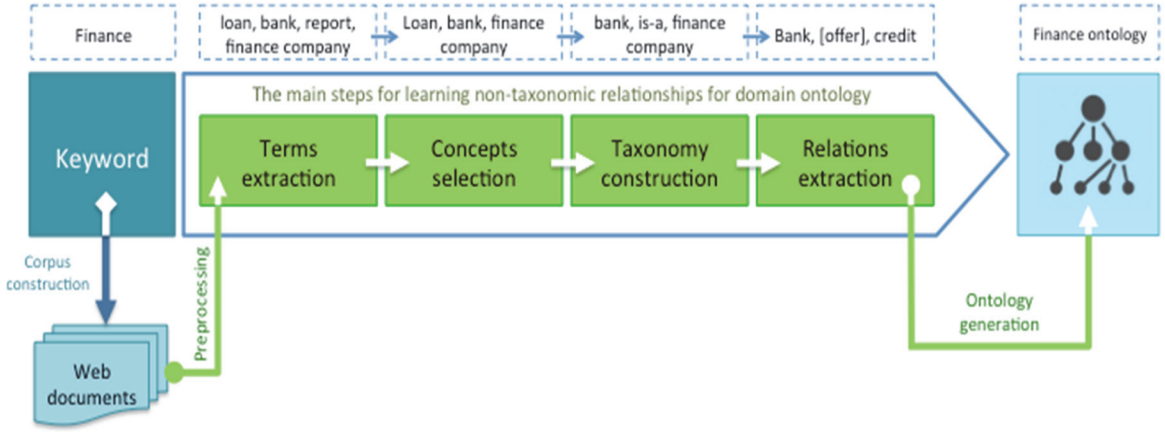

Figure 1: The main steps for learning non-taxonomic relationships for domain ontology.

In a more recent work, (Serra et al., 2013) use on their work Natural Language Processing (NLP) and statistics to extract non-taxonomic relationships of domain ontology. The system provides three types of extraction rules: the Sentence Rule (SR), the Sentence Rule with Verb Phrase (SRVP) and the Apostrophe Rule (AR). The importance of the extracted relations is measured using co-occurrence frequency.

In term of financial ontologies, following the research that we have conducted, we found that most financial ontologies were built manually (Mellouli et al., 2010; Wang et al., 2011). Rarely are the researches topics that address the automatic construction of this domain ontology. The OntoPlus methodology introduced by Novalija et al., (2011) is one of the few works that deals with automating the extraction of finance ontology. However, OntoPlus relates to the automatic extension of an existing ontology and not to learn it from scratch. It introduces a new methodology for semi-automatic ontology extension for analyzing business and financial news. OntoPlus used mining techniques to automatically identify candidate concepts in the ontology to relate to the new knowledge from the domain. However, general non-taxonomic relationships were not included in the process of ontology extension.

In the present work, we propose a new method to overcome all the limitations listed below and to learn financial ontology from scratch. We incorporate, for the first time in the literature, Open IE algorithms in the domain ontology extraction process. Thus, we benefit from the performance and the experience of these algorithms in the area of non-taxonomic relationships extraction.

3 OVERVIEW OF THE STEPS OF LEARNING NON-TAXONOMIC RELATIONSHIPS FOR DOMAIN ONTOLOGY

From the ontology-learning point of view (Buitelaar et al., 2005), the main steps of learning non-taxonomic relationships for domain ontology are the following (see Figure.1):

Step 1 - Extraction of Terms that Represent the Domain: It's needed to identify terms that are likely related to the studied domain. This step constitutes a principal prerequisite for unsupervised concept acquisition. In our work, the term extraction is based on the neighborhood of an initial keyword that characterizes the input corpus. In our finance domain example (Figure 1), the initial keyword given as input for building the ontology is *"Finance"* and the terms *"bank, loan, report, finance company"* are examples of the extracted terms in this phase of the ontology learning process.

Step 2 - Selection of Domain Concepts: This step aims to select the most relevant terms from those previously extracted. The filtered terms constitute the concepts of the studied domain. Concepts selection can be performed using statistic-based techniques (Maedche and Volz, 2001; Makrehchi and Kamel, 2007), linguistic-based techniques or a hybrid one (Meijer et al., 2014).

In our work, we use a statistical measure that combine Chir statistic and Conditional Mutual Information measure to select the most relevant domain concepts (see section 4). In our example of the finance ontology, the word *"report"*, extracted in the first step, does not belong to the finance

domain. Using our combined method, previously cited, that noisy word will be excluded in this second stage of the learning process of ontology. Therefore, only the terms *"bank, loan, finance company"* will be retained as relevant concepts of the field of finance.

Step 3 - Construction of the Taxonomy: Learning or extracting taxonomic relations means finding hyponyms between the selected concepts with the goal of constructing a concept hierarchy. It can be performed in various ways such as using predefined relations from existing background knowledge (Lee et al., 2008), using hierarchical clustering (Maedche and Volz, 2001), relying on semantic relatedness between concepts (Pekar and Staab, 2002), or using linguistic and logical rules or patterns (Jiang and Tan, 2005).

Our proposed algorithm uses the term structure through string matching of the immediate posterior and anterior words of a keyword concept to extract their hierarchies. In our example, we extract the hyponym relation between the two concepts *"bank"* and *"finance company"*.

Step 4 - Extraction of Non-taxonomic Relationships: Discovering and labeling non-taxonomic relations are mainly reliant on the analysis of the structure and dependencies of candidate sentences. In this phase, verbs are good indicators for non-taxonomic relations and are used to label such relations.

Our main contribution in this work is related to this step of the ontology construction process by introducing a novel technique based on Open IE algorithms to extract and label non-taxonomic relations of domain ontology. In our finance domain example, the relation between *"bank"* and *"credit"* was identified and labeled with the verb *"offer"*.

4 OUR METHODOLOGY

4.1 Terms Extraction

Our algorithm is based on the analysis of a large number of web corpus files in order to find relevant terms of a domain. In the English language, the immediate posterior and anterior words of a keyword express a semantic specialization between them (Grefenstette, 1997). From this point of view, the method used to extract relevant terms (candidates concepts) is based on the neighborhood of an introduced keyword (initial keyword) that is enough representative for the studied domain. The algorithm selects it anterior and posterior words as relevant terms for the next step of concepts selection.

4.2 Concepts Selection

To identify domain concepts from the extracted terms, we used the hybrid measurement (Conditional S-Measure) based on the Chir-statistic and the conditional information similarity already defined in our previous work (El idrissi esserhrouchni et al., 2014). As we have shown in that paper, Conditional S-Measure statistics is more efficient than other benchmark algorithms for selecting domain concepts.

However in the following sections we remind the various components of the Conditional S-Measure.

4.2.1 Chir Statistic

The Chir statistic proposed by Li et al., (2008) is an extended variant of the $\chi 2$ statistic to measure the degree of dependency between a term w and a category c of documents. They showed that their method could improve the performance of text clustering by selecting the words that help to distinguish documents in different groups. Indeed, when the $\chi 2$ statistic measure the lack of independence between the terms in the category (Saengsiri et al., 2010), the Chir statistic selects only relevant terms that have strong positive dependency on certain categories in the corpus and remove the irrelevant and redundant terms. To define the term goodness of a term w in a corpus with m classes Li et al., (2008) use a combining formula of $\chi 2$ and a category dependency measure $R(w,c)$ defined by:

$$R_{w,c} = \frac{O(w,c)}{E(w,c)} \quad (1)$$

Where $O(w,c)$ is the number of documents that are in the category c and contain the term w, and $E(w,c)$ is the expected frequency of the category c to contain the term w. If there is a positive dependency, then $R(w,c)$ should be larger than 1. If there is negative dependency, $R(w,c)$ should be smaller than 1. In the case of the no-dependency between the term w and the category c, the term category dependency measure $R(w,c)$ should be close to 1. In summary, When $R(w,c)$ is larger than 1, the dependency between w and c is positive, otherwise, the dependency is negative.

The final formula of the Chir statistic was defined by:

$$r_{\chi^2}(w) = \sum_{j=1}^{m} p\left(R_{w,c_j}\right) \chi^2_{w,c_j} \quad \text{With } R_{w,c_j} > 1 \quad (2)$$

Where

$$p\left(R_{w,c_j}\right) = \frac{R_{w,c_j}}{\sum_{i=1}^{m} R_{w,c_i}} \quad \text{With } R_{w,c_i}>1 \quad (3)$$

is the weight of $\chi^2_{w,c}$ in the corpus in terms of Rw,c_j.

A bigger $r_{\chi^2}(w)$ value indicates that the term is more relevant. When there is a positive dependency between the term w and the category c_j.

4.2.2 Conditional Information Measure

In our ontology learning process, conditional mutual information is used to measure dependency between two terms w and w' conditioned by the occurrence of a parent term w_p. On the basis of Brun et al., (2002) work, two terms are considered similar if their mutual information with all terms in the vocabulary is nearly the same. Thus, we defined the conditional similarity by (El idrissi esserhrouchni et al., 2014):

$$Sim(w,w'/w_P) = \frac{1}{2|V|}\sum_{i=1}^{|V|}\left(\frac{\min(I(zi,w/w_P),I(zi,w'/w_P))}{\max(I(zi,w/w_P),I(zi,w'/w_P))}\right.$$

$$\left. + \frac{\min(I(w,zi/w_P),I(w',zi/w_P))}{\max(I(w,zi/w_P),I(w',zi/w_P))}\right) \quad (4)$$

where V is the vocabulary, and $I(z_i,w/w_p)$ is the Conditional Mutual Information of terms z_i and w conditioned by the presence of the term w_p. The Conditional Mutual Information formula is given by (Zhang et al., 2012):

$$I(z_i,w/w_P) = P_d(z_i,w,w_P)log\frac{P(w_P)P_d(z_i,w,w_P)}{P_d(z_i,w_P)P_d(w,w_P)} \quad (5)$$

Where d is the withdrawal, $P(w_p)$ is the probability of the term w_p. $P_d(z_i,w_p)$ and $P_d(w,w_p)$ are the probability of succession of the terms z_i and w_p, w and wp respectively in the window observation. $P_d(z_i,w,w_p)$ is the probability of succession of the terms z_i, w and w_p in the window observation. This probability can be estimated by ratio of the number of times that the term z_i is followed by the terms w and w_p within the window, with the cardinal of the vocabulary:

$$P(z_i,w,w_P) = \frac{f_d(z_i,w,w_P)}{|V|} \quad (6)$$

Where $f_d(z_i,w,w_p)$ is the number of times that the term z_i is followed by the terms w and w_p.

4.2.3 Conditional S-Measure

To identify the relevant concepts from those extracted, we have defined a hybrid measure based on the weighting model combining a component estimated from the Chir-statistic and another one from the similarity measure using conditional information. This new scoring measure was defined as:

$$S(w_c/w_p) = \lambda * r_{\chi^2}(w_c) + (1-\lambda) * sim(w_c,w_k/w_p) \quad (7)$$

Where λ is a weighting parameter between 0 and 1.

Since relevant concepts have strong dependency with the studied domain and convey semantically similar information with respect to their parent concept and to the initial keyword, the candidate concepts having strong score are likely to be relevant and should be integrated into the extracted taxonomy.

4.3 Taxonomy Construction

The taxonomy extraction process is summarized in the following steps:

a) Perform a k-means clustering algorithm on the set of all documents and get initial clusters.
b) Start with a keyword that has to be representative enough for the domain and a set of parameters that constrain the search and the concepts selection.
c) Extract all the candidate concepts by analyzing the neighborhood of the initial keyword; select the anterior words and posterior words as candidate concepts.
d) For each candidate concept, calculate its score $S(w/w_p)$ measure by using (7).
e) Sort the terms in descending order of their $S(w/w_p)$ measure.
f) Select the top l terms from the list.
g) The l extracted concepts are incorporated as classes or instances in the taxonomy.
h) For each concept incorporated in the taxonomy, a new keyword is constructed joining the new concept and the initial one. This process is repeated recursively until a selected depth level is achieved or no more results are found.
i) Finally, a refinement process is performed in order to obtain a more compact taxonomy and to avoid redundancy.

4.4 Non Taxonomic Relations Extraction

The main contribution of our methodology in this area is the integration of Open IE in the process of learning non-taxonomic relationships for domain ontology. Thus, we benefit from the performance

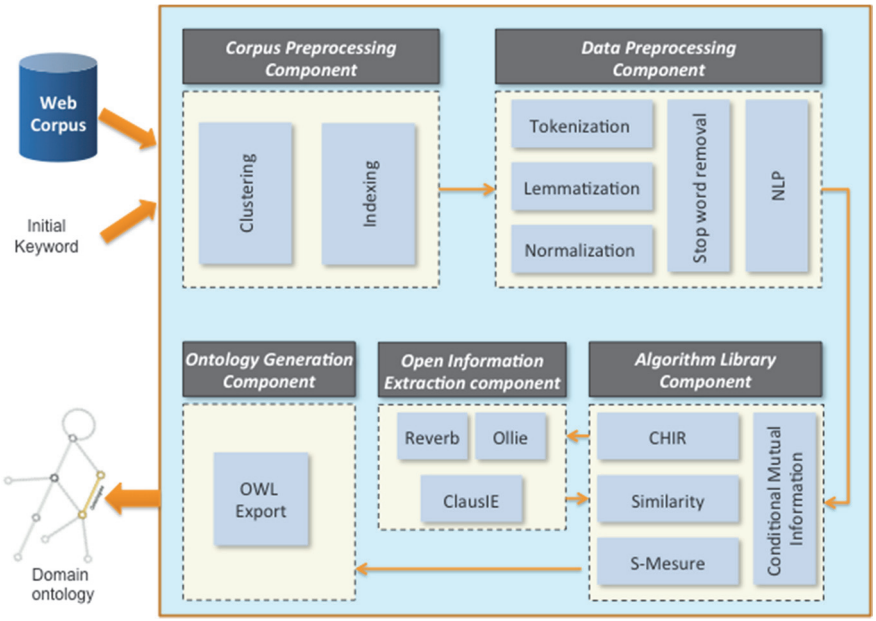

Figure 2: System architecture.

and the experience of these algorithms in the area of relations extraction.

Open IE approaches are relatively a recent paradigm for extracting relations form unstructured documents. They were introduced by Banko et al., (2007). Open IE systems facilitates domain independent discovery of relations. They extract all possible relations without any prerequisite or restriction. In the recent years, many systems for Open IE have been developed. For our non-taxonomic learning process, we integrate the most recent ones, mainly: Reverb (Fader et al., 2011), Ollie (Schmitz et al., 2012) and ClausIE (Del Corro and Gemulla, 2013).

Reverb uses shallow syntactic parsing to identify relations expressed by verbs. The system takes a sentence as input, identifies a candidate pair of noun phrase arguments *(arg1, arg2)* from the sentence, and then uses the learned extractor to label each word between the two arguments as part of the relation phrase or not.

In Ollie system, the authors use context analysis to extract relations in a given sentence. It extracts not only relations expressed via verb phrases, but also relations mediated by adjectives and nouns.

ClausIE is the most recent Open IE system. It differs from Reverb and Ollie approaches in that it separates the detection of useful information expressed in a sentence from their representation in terms of extractions. ClausIE exploits linguistic knowledge to first detect clauses in an input sentence and to subsequently identify the type of each clause according to the grammatical function of its constituents.

However, in our domain extraction context, due to its open-domain and open-relation, gross use of Open IE algorithms is unable to relate the extracted relations to domain ontology. Subsequently, an adaptation of the used Open IE algorithms is necessary to overcome this limitation.

To address this limitation, we have implemented a solution based on the concepts of the taxonomy already extracted in the previous stage. The proposed process for learning non-taxonomic domain relationships with Open IE tools is performed in three steps:

Step1: For each concept of the taxonomy, we extract from the corpus all the sentences where the concept *c* is found.

Step2: For each extracted sentence, we discover all possible relations using one of the proposed Open IE algorithms. As an output, we obtain a set of relational tuples <*Arg1, Rel, Arg2*> that describe the sentence verb relation *(Rel)* and its arguments *(Arg1 and Arg2)*.

Step3: Finally, we judged each tuple as related to the studied domain or not, based on whether it contains the concept *c* in one of the extracted arguments. The selected relations are incorporated into the result ontology.

This process is repeated until all concepts of the taxonomy are processed.

5 SYSTEM ARCHTECTURE

The proposed ontology learning system consists of five basic components, namely: Corpus Pre-processing Component (CPC), Data Pre-processing Component (DPC), Algorithm Library Component (ALC), Open IE Component (OIEC) and Ontology Generation Component (OGC) (see Figure 2).

5.1 Corpus Pre-processing Component

This component aims to import the corpus into the system and prepare it for processing. In a first stage, k-means clustering algorithm (Jain et al., 1999) is applied. It partitions the corpus into k clusters so that two documents within the same cluster are more closely related than two documents from two different clusters. In a second step, an indexing process is executed to index the full contents of the clustered documents. Based on the neighborhood of an initial keyword to select domain concepts and to extract their taxonomic and non-taxonomic relationships. The indexation allows an efficient and fast retrieval of this information.

5.2 Data Pre-processing Component

In the Data Preprocessing Component, text information is filtered and cleaned. The preprocessing includes the following elements:

- Tokenization: Splitting strings into their component words based on delimiters.
- Splitting compound words.
- Normalization: Elimination of stylistic differences due to capitalization, punctuation, word order, and characters not in the Latin alphabet.
- Lemmatization: Elimination of grammatical differences due to verb tense, plurals, etc. It's used to improve the recall of the domain ontology concepts in the corpus.
- Stop word removal: Remove of very common words. The Glasgow stop word list is used in this work.

5.3 Algorithm Library Component

The Algorithm Library consists of a statistical algorithm that (1) extracts candidate concepts from a collection of documents and stores the sentences where they are located; (2) evaluates the taxonomic dependency between key concepts; (3) selects the most relevant ones, (4) discovers non-taxonomic relationships of the selected concepts (5) and performs an iterative mining algorithm that constructs the ontology.

5.4 Open IE Component

The Open IE component proposes three algorithms to extract non-taxonomic relations: Reverb, Ollie and ClausIE. It communicates with the previous component "Algorithm Library" in input/output mode. According to the Open IE algorithm chosen at the beginning of the process, and from the sentences provided by the "Algorithm Library", the Open IE component identifies non-taxonomic relationships of each concept of the taxonomy and returns verbs and concepts that describe the identified relationships.

5.5 Ontology Generation Component

In this final phase, the resulted domain ontology is generated in an OWL file format by using the Jena toolkit. The resulting file can be visualized using an ontology editor such as Protegé.

6 EVALUATION

The evaluation process is an essential step that should be performed in any ontology learning approach. It's particularly important in automatic approaches as the present work. In this section, we evaluate the performance of our methodology for extracting non-taxonomic relationships for finance domain using a gold standard ontology. We implement our new system using three well-known Open IE tools in the literature: Reverb, Ollie and ClausIE. Then we compare it with Sanchez and Moreno (2008) algorithm, in a finance corpus, to evaluate the relevance of the resulting non-taxonomic relationships of each algorithm.

Due to the lack of an evaluation corpora and an ontology officially released and fully covering this area, we decided to build our own corpus and gold standard ontology. The corpus was automatically constructed from the financial news Website "Yahoo! Finance" and the gold standard ontology was developed in a collaboration with a domain specialist.

It should be noted that the comparison of the finance result taxonomy of our algorithm with that of Sanchez and Moreno and others related works was already presented in a previous work (El idrissi esserhrouchni et al., 2014). This comparison showed that our algorithm was more efficient in building

finance concept hierarchies.

6.1 Gold Standard Ontology

The gold standard ontology was developed in two steps. It was designed to give equal chances to all tested algorithms. In the first step, the four candidate algorithms (our three algorithms based on Open IE tools and the Sanchez and Moreno algorithm) were launched on the same constructed finance corpus. Then, the resulting ontologies were automatically merged into a single one. Redundant concepts and relations were exported only once. Our aim is to build a domain ontology based on the previous four ontologies. In the second step, the constructed ontology was presented to a domain specialist for validation. He performs a cleanup of the constructed ontology and removes erroneous concepts and invalid taxonomic and non-taxonomic relationships. The final ontology was used as a reference ontology for our evaluation process.

The number of the extracted non-taxonomic relationships are shown in Table 1. Table 2 shows the number of non-taxonomic relationships in the Gold standard ontology after the specialist validation and the refinement of the merged ontology.

6.2 Finance Domain Corpus

We built the test corpus from the financial news Website "Yahoo! Finance". A Java program was developed to retrieve financial articles from it. We based on a period of 100 days since January 1, 2015 to retrieve randomly new articles. 7213 documents were retrieved for that period, which corresponds to over 11 million words. One of the extracted document is shown in Figure 3.

6.3 Algorithms Implementation

We programmed our new algorithm using Java 8. Regarding the integration of Open IE algorithms in the domain ontology learning process, we used the original Java source code published by their owners and we adapted them to extract domain relations.

In the absence of the source code of the algorithm of Sanchez and Moreno, and to perform the comparison with our algorithm, we developed their algorithm from scratch in Java 8 by referring to its description in their paper.

6.4 Evaluation

To evaluate the relevance of the extracted non-taxonomic relationships by the studied algorithms we have used the Lexical Precision measure (LP). It measures the number of relevant relations extracted $e_{relevant}$ divided by the total number of relations extracted e_{all}. LP is defined as (Sabou et al., 2005):

$$LP = \frac{e_{relevant}}{e_{all}} \quad (8)$$

In the interest of a fair comparison, all algorithms were executed under the same conditions. The test was performed on a 64-bit windows machine, with 8 GB of RAM.

The results given by the different algorithms are summarized in Table 3.

Table 1: Number of generated non-taxonomic relationships.

Source	Our method using Reverb	Our method using Ollie	Our method using ClausIE	Sanchez & Moreno	Merged ontology
Number of extracted non-taxonomic relations	17	20	17	13	42

Table 2: Number of non-taxonomic relationships in the gold standard ontology.

Source	Gold standard ontology
Number of non-taxonomic relations	29

According to those results, we observe that our method, used with the three Open IE algorithms, is more precise than that proposed by Sanchez and Moreno with over 30% of difference.

However, using the open IE algorithm reverb, our method reaches its best performance and achieved an accuracy of 82.35%. It outperforms Ollie and ClausIE in term of precision of the extracted non-taxonomic relationships. Table 4 shows an extract of the learned non-taxonomic relationships using reverb with our proposed algorithm.

Table 3: Precision evaluation of non-taxonomic relationships performed on a financial corpus.

Algorithm	Number of rejected non-taxonomic relations	Number of accepted non-taxonomic relations	Precision
Our method using Reverb Open IE	3	14	82,35%
Our method using Ollie Open IE	4	16	80,00%
Our method using ClausIE Open IE	4	13	76,47%
Sanchez & Moreno	5	8	61,54%

Car loan: green light
Even if you can afford to pay for a car with cash, you may want to consider paying it out over time. If your credit is good, you may qualify for a very low-interest loan that could help your credit mix and eventually boost your credit score even higher.
You can get car loans from several different sources. Most dealers are affiliated with finance companies that offer loans. You can also get a loan through your bank or credit union. The better your credit score, the lower the interest rate you'll be offered.
If your score is less than stellar, consider using a credit union. Interest rates at federal credit unions are limited by law to 18 percent, and may be more reasonable than rates at buy-and-drive car lots or finance companies, says Rex Johnson, owner and founder of Lending Solutions Consulting Inc., a credit-union consulting firm in Elgin, Ill.

Figure 3: A partial example of an extracted document from Yahoo! Finance web site.

Table 4: An extract of the learned non-taxonomic relationships using reverb with our proposed algorithm.

The extracted relations		
Subject (NP)	Verb (VP)	Object (NP)
finance company	provide	loan
credit	to be get from	finance company
mortgage finance solution	include	loan
company	will pay	finance debt
corporate bond	help	company financing
dealer	Be affiliate with	Finance institution
broker	Be affiliate with	Finance institution

7 CONCLUSION

This paper presented a new methodology for learning non-taxonomic relationships and building financial ontology from scratch. One of the main difficulties of domain ontology learning is the identification of relevant non-taxonomic relations. The novelty of our approach in this field is that it is based on adjusting Open IE algorithms to extract and label domain relations between concepts. The obtained results show that using Open IE tools is an interesting way to extract ontological relationships with a higher degree of precision.

For the implementation of our methodology, we integrated three well-known Open IE algorithms in our process of extracting non-taxonomic domain relationships, namely: Reverb, Ollie and ClausIE. The approach has been evaluated in two steps: (i) using a gold standard ontology. In this case, the obtained results show that Reverb enables to achieve best performance and outstrips Ollie and ClausIE in terms of accuracy of the extracted non-taxonomic relationships. (ii) By comparing the discovered non-taxonomic relationships by our approach to those extracted by Sanchez and Moreno algorithm on the same finance corpus. In this latter test, our method was more precise and obtained the best results. However, the number of the extracted relationships by the different methods remains limited. This is due to the complexity of the financial domain and the number of documents in the corpus. Indeed, the size and the quality of the corpus are important criteria for the success of an ontology learning tool. Also, changing the corpus might be interesting to test if the performance of our approach remains the best from one domain to another.

In terms of perspective, our future work will consist of increasing the size of the finance corpus in order to build a richer ontology for finance domain. We plan its validation by using it in a decision making application in the financial sector. Also, our current work consist in the online publication of our algorithm as a web application on the address *www.ontologyline.com*. This will allow the ontological research community to test the algorithm and evaluate it in various domains.

REFERENCES

Banko, M., Cafarella, M. J., Soderland, S., Broadhead M., Etzioni O., 2007. Open information extraction from the Web. In *Proceedings of the 20th international joint conference on Artifical intelligence* (IJCAI), 2670-2676.

Brun, A., Smaïli, K., Haton, J. P., 2002. WSIM: une méthode de détection de thème fondée sur la similarité entre mots. *Actes de TALN*, 145-154.

Buitelaar, P., Cimiano, P., Magnini, B., 2005. Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*, IOS Press, Amsterdam, 3-12.

Del Corro, L., Gemulla, R., 2013. ClausIE: clause-based

open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, 355-366. International World Wide Web Conferences Steering Committee.

Drymonas, E., Zervanou, K., Petrakis, E. G. M., 2010. Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System. In *Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems*, 277-287. Springer.

El idrissi esserhrouchni, O., Frikh, B., Ouhbi, B., 2014. HCHIRSIMEX: An extended method for domain ontology learning based on conditional mutual information. In *Third IEEE International Information Science and Technology (CIST),* 91-95. IEEE.

Fader, A., Soderland, S., Etzioni, O., 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 1535-1545. Association for Computational Linguistics.

Frikh, B., Djaanfar, A. S., Ouhbi, B., 2011. A Hybrid Method for Domain Ontology Construction from the Web. *In KEOD*, 285-292.

Grefenstette, G., 1997. Short query linguistic expansion techniques: Palliating one-word queries by providing intermediate structure to text. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, 97-114. Springer.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264-323.

Jiang, X., Tan, A. H., 2005. Mining ontological knowledge from domain-specific text documents. In *Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE.

Lee, S., Huh, S. Y., McNiel, R. D., 2008. Automatic generation of concept hierarchies using WordNet. *Expert Systems with Applications*, 35(3), 1132-1144.

Li, Y., Luo, C., Chung, S.M., 2008. Text Clustering with Feature Selection by Using Statistical Data. *IEEE Transactions on Knowledge and Data Engineering*, 20(5), 641-652.

Maedche, A., Staab, S., 2000. Semi-automatic engineering of ontologies from text. In *Proceedings of the 12th international conference on software engineering and knowledge engineering,* 231-239.

Maedche, A., Staab, S., 2001. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*. 16 (2), 72-79.

Maedche, A., Volz, R., 2001. The ontology extraction and maintenance framework text-to-onto. In *Proceedings of the ICDM'01 Workshop on Integrating Data Mining and Knowledge Management*.

Makrehchi, M., Kamel, M. S., 2007. Automatic taxonomy extraction using google and term dependency. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 321-325. IEEE Computer Society.

Meijer, K., Frasincar, F., Hogenboom, F., 2014. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems,* 62, 78-93.

Mellouli, S., Bouslama, F., Akande, A., 2010. An ontology for representing financial headline news. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2), 203-208.

Novalija, I., Mladenie, D., Bradesko, L., 2011. OntoPlus: Text-driven ontology extension using ontology content, structure and co-occurrence information. *Knowledge-Based Systems*, 24(8), 1261-1276.

Pekar, V., Staab, S., 2002. Taxonomy learning: factoring the structure of a taxonomy in to a semantic classification decision. *19th International Conference on Computational Linguistics*, Vol. 1, 1-7. Association for Computational Linguistics.

Petrov, S., Barrett, L., Thibaux, R., Klein, D., 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics,* 433-440. Association for Computational Linguistics.

Sabou, M., Wroe, C., Goble, C., Mishne, G., 2005. Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In *Proceedings of the 14th international conference on World Wide Web*, 190-198. ACM.

Saengsiri, P., Meesad, P., Na Wichian, S., Herwig, U., 2010. *Comparison of Hybrid Feature Selection Models on Gene Expression Data*. *IEEE International Conference on ICT and Knowledge Engineering*, 13-18. IEEE.

Sanchez, D., Moreno, A., 2008. Learning non-taxonomic relationships from web documents for domain ontology construction. *Data and Knowledge Engineering*, 64(3), 600-623.

Schmitz, M., Bart, R., Soderland, S., Etzioni, O., 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 523-534. Association for Computational Linguistics.

Serra, I., Girardi, R., Novais, P., 2013. PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text. In *Proceedings of the 10th International Conference on Information Technology*. IEEE.

Shamsfard, M., Barforoush, A. A., 2003. The state of the art in ontology learning: A framework for comparison. *The Knowledge Engineering Review*, 18(4), 293-316.

Villaverde, J., Persson, A., Godoy, D., Amandi, A., 2009. Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. *Expert System with Applications*, 36(7), 10288-10294.

Wang, S., Xu, K., Liu, L., Fang, B., Liao, S., Wang, H., 2011. An ontology based framework for mining dependence relationships between news and financial instruments. *Expert Systems with Applications*, 38(10), 12044-12050.

Zhang Y., Zhang Z., 2012. Feature subset selection with cumulate conditional mutual information minimization. *Expert Systems with Applications*, 39 (5): 6078-6088. Elsevier.