J Intell Inf Syst
DOI 10.1007/s10844-011-0149-4

Learning non-taxonomical semantic relations from domain texts

Janardhana Punuru · Jianhua Chen

Received: 2 November 2010 / Revised: 6 January 2011 / Accepted: 10 January 2011

Abstract Ontology of a domain mainly consists of concepts, taxonomical (hierarchical) relations and non-taxonomical relations. Automatic ontology construction requires methods for extracting both taxonomical and non-taxonomical relations. Compared to extensive works on concept extraction and taxonomical relation learning, little attention has been given on identification and labeling of non-taxonomical relations in text mining. In this paper, we propose an unsupervised technique for extracting non-taxonomical relations from domain texts. We propose the VF*ICF metric for measuring the importance of a verb as a representative relation label, in much the same spirit as the TF*IDF measure in information retrieval. Domain-relevant concepts (nouns) are extracted using techniques developed earlier. Candidate non-taxonomical relations are generated as (SVO) triples of the form (subject, verb, object) from domain texts. A statistical method with log-likelihood ratios is used to estimate the significance of relationships between concepts and to select suitable relation labels. Texts from two domains, the Electronic Voting (EV) domain texts and the Tenders and Mergers (TNM) domain texts are used to compare our method with one of the existing approaches. Experiments showed that our method achieved better performance in both domains.

Keywords Ontology learning · Semantical relation · Text mining

J. Punuru
Computer Science Department, Louisiana State University,
Baton Rouge, LA 70803-4020, USA
e-mail: punuru@csc.lsu.edu

J. Chen (✉)
Computer Science Department and Center for Computation and Technology,
Louisiana State University, Baton Rouge, LA 70803-4020, USA
e-mail: jianhua@csc.lsu.edu

Published online: 21 January 2011

1 Introduction

Ontology of a domain is a formal specification about the conceptualization of the domain. An ontology typically consists of several major components: (1) Concepts, which represent classes of objects that are deemed important for the domain; (2) Taxonomical relations that describe the hierarchical "IS-A" relationship between the concepts; (3) Other semantical relationships between the concepts. The following diagram (Fig. 1) illustrates an ontology for the "Electronic Voting" domain. In the diagram, the ellipses represent concepts and the arrows between the ellipses denote taxonomical or other semantic relations between the corresponding concepts. The concepts for the EV domain include "voter", "person", "ballot", "official", "voting machine", "county", etc. The taxonomical relations for the EV domain include pairs such as (voter, person) (meaning that a "voter" is a "person"), (voting machine, machine), etc. Other non-taxonomical semantic relations for the domain include triples such as (voter, cast, ballot) (meaning that "voter" is related to "ballot" via the relation "cast"). The Electronic Voting ontology was obtained semi-automatically using the techniques that we developed for ontology learning from domain texts. This paper reports our work on automatically learning non-taxonomical relations from domain texts.

Ontologies are widely used in various research and application fields like information retrieval, information extraction, conceptual mapping, and knowledge management. Recently, Semantic Web (Berners-Lee et al. 2001), proposed by Tim Berners-Lee inventor of the WWW, has attracted a great attention toward a new kind of representation for the Web using XML, RDF, and OWL. Semantic Web aims at representing the textual content of Web pages by ontologies, which allow machines

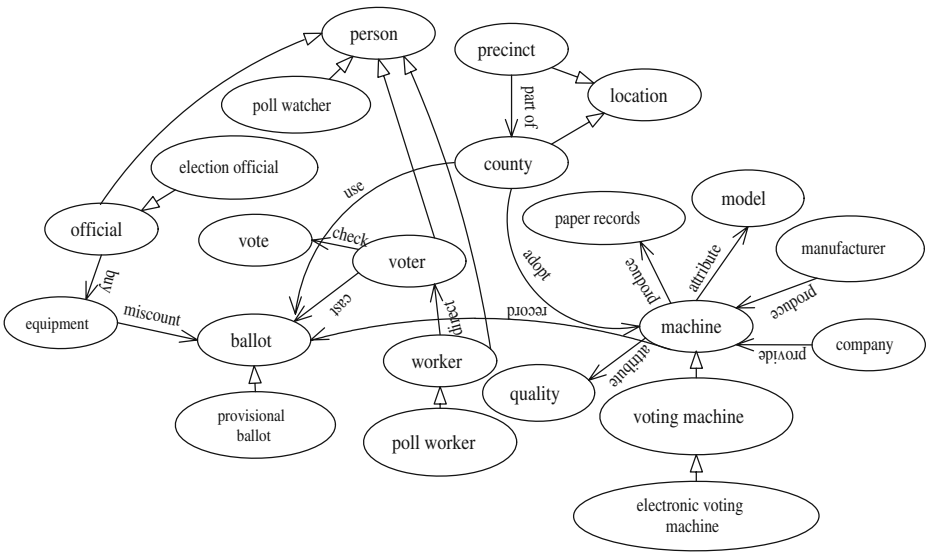

Fig. 1 Ontology for the electronic voting domain

to comprehend the semantics of documents and return more accurate answers to user queries. As of now, ontologies for domains of interests are developed manually in spite of the wide spread use of ontologies. This situation clearly indicates the urgent need for techniques that can automatically learn (construct) ontology from domain texts.

Various recent research projects have focused on automatically learning ontology from texts (Maedche and Volz 2001) to eliminate the high time and labor costs of the manual ontology construction. Concepts of a domain are often identified by domain relevant terms occurring in texts of the domain. Various techniques are presented in the literature (Tomokiyo and Hurst 2003; Pantel and Lin 2001; Punuru and Chen 2006) for concept extraction. Considerable attention has also been given to the task of taxonomy (hyponym/hypernym) relation extraction (Hearst 1992; Caraballo 1999; Cederberg and Widdows 2003). Compared to the abundance of works on concept extract and taxonomic relation learning, very little attention has been given to extracting non-taxonomic relations. Most existing techniques for the task make some simplifying assumptions, for example, focusing on some fixed relation types such as part-whole (Girju et al. 2003; Berland and Charniak 1996) or cause-effect (Girju and Moldovan 2002), or focusing on relations among some fixed set of concept types such as persons, locations, and organizations. Very few techniques are presented for extracting general non-taxonomical relations. Moreover, most of the existing techniques are *supervised* methods which require the availability of labeled training data. Details of some existing methods are discussed in the related works section.

To address the limitations of existing approaches, we develop an *unsupervised* method for extracting *general*, non-taxonomical relations from domain texts. We consider relations of the form $C_i \rightarrow Rl \rightarrow C_j$ as instances of non-taxonomic relations where Rl is a relation name different from "IS-A". If concepts C_i and C_j are indeed related with the relationship indicated by Rl, then the ordered triple (C_i, Rl, C_j) is considered a valid non-taxonomic relation in the domain. For example, the triple (voter, cast, ballot) indicates a valid non-taxonomical relation. For brevity we often use, from now on, the word "relations" to refer to non-taxonomical relations when no confusion is likely. The problem of relation extraction can be tackled as two sub problems. The first problem is identification of concept pairs(C_i, C_j) such that some semantic relationship holds from C_i to C_j. The other problem is determination of the label for the semantic relationship from C_i to C_j. To identify candidate concept pairs, the log-likelihood ratio measure (described in detail in Section 6) is applied to filter an initial set of concept pairs, which are constructed from domain specific concepts based on co-occurrence in sentences of domain texts. Candidate relationship labels are identified using the VF*ICF metrics, which is explained in more detail in Section 5. And again the log-likelihood ratio method is used to determine the relation label for each candidate concept pair.

Our approach has two main advantages. First, it is a completely *unsupervised* technique. That is, it does not require any pre-labeled training data. Second, our method is *general* in the sense that it is domain-independent - it does not assume any pre-defined list of relation types (such as part-whole, causal, etc.) or concept types. and does not use any external general purpose knowledge bases such as WordNet (Miller et al. 1990). These strengths make our approach applicable in many diverse domains. In addition, our method is computationally quite efficient.

Before moving on to the subsequent sections, we briefly describe the two domain text data sets which have been used in the experiments in this research. The Electronic Voting (EV) data set consists of 15 textual documents on electronic voting from New York Times website www.nytimes.com. This is a small text set, yet it contains reasonably rich information for demonstrating and verifying the ideas proposed in this study. The Tender offers, Mergers, and Acquisition (TNM) data set is a larger collection. The TNM Corpus is collected from TIPSTER Volume 1 corpus distributed by NIST. TIPSTER Volume 1 corpus consists of three year (1987, 1988, and 1989) news articles from Wall Street Journal. In TIPSTER corpus data each news article is labeled with the topic it describes. The TNM Corpus is obtained by collecting news articles with the topic label Tender offers, mergers, and acquisitions (TNM). It consists of 272 news articles which amounts to a total of 30 MB texts.

This paper is an extension of our early work (Punuru and Chen 2007). The rest of the paper is organized as follows. The next section gives a brief review on the existing methods for non-taxonomical relation extraction. In Section 3, we present an overview of the proposed method. Section 4 presents the VF*ICF measure for extracting candidate relation labels. The log-likelihood ratio method for relationship label assignment is presented in Section 5, along with a brief description of two variations of our method. Section 6 provides the experimental results of the proposed method. Sections 7 and 8 gives future directions and conclusions of the paper respectively.

2 Related works

One of the least tackled problems in ontology learning is extraction of non-taxonomical relations from domain texts. Relatively fewer works in the literature focus on this problem in comparison to the large amount of works on concept (term) extraction and taxonomical relation learning. Most existing works for non-taxonomical relation learning focused on finding relations between named entities (Hasegawa et al. 2004; Stevenson 2004; Yangarber et al. 2000; Riloff 1996). Thus such works concentrate on relations over a set of predefined entities such as *person*, *location*, *organization* and etc. These entities are fixed irrespective of the domain. This is rather limiting because actual concepts vary considerably across different domains. Techniques for ontological relation extraction should not rely on any predefined entities or concepts.

2.1 Works for learning semantic relations among non-predefined concepts

Very few methods are available for identification of relations among non-predefined concepts (Faure and Nedellec 1998; Ciaramita et al. 2005; Kavalec et al. 2004; Schutz and Buitelaar 2005). In 1998, Faure et al. considers the relation extraction problem as learning selection restrictions for verbs. In this method all terms occurring along with a single verb are clustered and each of the clusters are manually labeled. The methods in Ciaramita et al. (2005), Kavalec et al. (2004) and Schutz and Buitelaar (2005) exploit the syntactic structure and dependencies between words for relation extraction. Both Ciaramita et al. (2005) and Schutz and Buitelaar (2005) extract concept pairs occurring in pre-specified dependency relations. They use chi-square test

to verify the statistical significance of the co-occurrence of concepts in a concept pair. Ciaramita et al.'s (2005) work is experimented with the Molecular Biology domain texts. In Ciaramita et al. (2005), chi-square test is employed to learn the patterns for relations such as $SUBJ \rightarrow \text{bind} \rightarrow DIR\_OBJ$. Schutz and Buitelaar (2005) used the football domain texts for their experimentation. Their technique builds relation triples by extracting predicates along with related concept pairs. In 2009, Fu et al. used a two-stage approach combining the use of lexico-syntactic patterns and K-means clustering for semantic relation extraction. They focus on learning pre-defined semantic relations in the computer components trouble-shooting domain.

2.2 Statistical measure on co-occurrence of concept pairs and verbs

Kavalec et al.'s (2004) work extracts relation a relation label V for the relationship between concepts C_1 and C_2 based on the Above Expectation (AE) measure defined in (1). Intuitively, using the AE measure could be explained this way: If the co-occurrence of a concept pair (C_1, C_2) with a given verb V is more frequent than the individual concept's co-occurrence with V, then the verb V is probably semantically related to the concept pair, and thus should be a good candidate label for concept pair. The authors of Kavalec et al. (2004) used the tourism domain texts for their experiments.

$$AE\left(\left(C_1 \bigwedge C_2\right)\middle|V\right) = \frac{P((C_1 \bigwedge C_2)|V)}{P(C_1|V).P(C_2|V)} \tag{1}$$

We reimplemented and verified the above AE metric (1) with our domain texts. The results of this measure are compared with that of our approach. In Kavalec et al. (2004), another AE measure in (2) was also defined. This measure also intends to capture the connection strength of the verb V with a given pair of concepts C_1 and C_2. This AE measure should be quite useful in selecting candidate labels for relations, but the examples in Kavalec et al. (2004) appear to emphasize more on the use of the first AE measure (1).

$$AE\left(\left(V\middle|C_1 \bigwedge C_2\right)\right) = \frac{P((V\,|C_1 \bigwedge C_2)}{P(V\,|C_1).P(V\,|C_2)} \tag{2}$$

2.3 Supervised learning versus unsupervised learning of semantic relations

In recent years, more and more attention has been given to the task of learning non-taxonomical semantic relations. Various supervised and semi-supervised methods have been developed by researchers. For example, Zhou et al. (2008) proposed a semi-supervised method that uses both labeled and unlabeled relation instances for learning semantic relations between named entities. The works in Qian et al. (2009) improved the performance of such semi-supervised semantic relation learning via stratified sampling strategy. Our work is quite different from the above-mentioned works because we use an un-supervised approach. However it is quite possible to combine our approach with other supervised and un-supervised methods in producing a more powerful semantic relation extraction system. This can be addressed in the future work.

3 Overview of our approach

Our approach to the relation extraction task involves the following steps.

1. Text pre-processing (parts of speech tagging, noun phrase identification, morphological processing, etc.)
2. Extraction of domain specific concepts,
3. Identification concept pairs(C_i, C_j) such that C_i and C_j are related.
4. Extraction of the candidate labels, *Rl*, for the relations.
5. Assignment of labels, *Rl*, for the relations between the concepts.

The first step of our approach is, naturally, pre-processing of domain texts. We first perform Parts of Speech (PoS) tagging on the raw texts to label each word by its category (noun, adjective, verb, adverb, etc.) This is done using Brill's rule-based PoS tagger (Brill 1992). To identify noun phrases, BaseNP Chunker (Ramshaw and Marcus 1995) is used to mark all the noun phrases. Standard text processing operations such as morphological processing (word stemming) and stop-words removal, are also applied. This will produce a candidate list of noun phrases and verbs. Domain-specific concepts (noun phrases) are extracted using a technique described in Punuru and Chen (2006). Using the extracted concepts, concept pairs are identified based on the position of their occurrence in texts. Candidate relation labels (domain-specific verbs) are identified from texts. Finally, relation labels are assigned to concept pairs to yield non-taxonomical relations. The overall system architecture is shown in Fig. 2.

As mentioned earlier, various statistical (Pantel and Lin 2001; Tomokiyo and Hurst 2003) and syntactic based techniques (Jacquemin 1995) are in the literature for extracting domain specific concepts. For this part of the work, we use a combination of TF*IDF and lexical knowledge based technique presented in Punuru and Chen (2006). The concepts obtained may include compound terms such as "voting machine", where the head word for this compound term is "machine". The technique in Punuru and Chen (2006) has quite high precision and high recall according to the experiments in the Electronic Voting domain. To avoid error carry over, however, irrelevant terms (or concepts) retrieved by the concept extraction method are manually eliminated for our relation extraction study in the Electronic Voting domain. We did NOT eliminate irrelevant concepts in our larger scale study

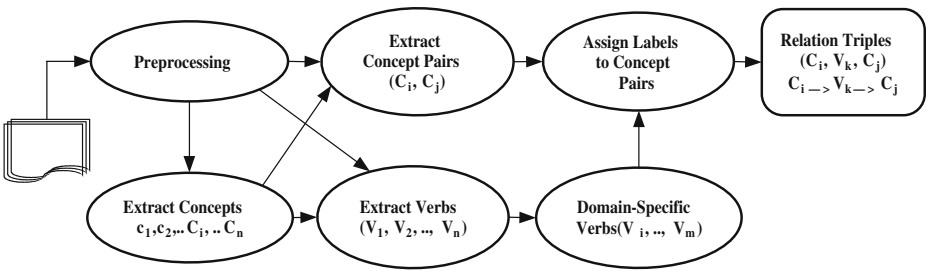

Fig. 2 Framework for relation extraction

over the Tenders and Mergers domain texts, which possibly explains the relative poor performance of our approach in that domain.

Candidate concept pairs are constructed by applying the log-likelihood ratio measure to filter an initial collection of concept pairs constructed from the set C of domain specific concepts. How do we get an initial set of concept pairs from C? Note that when C contains n concepts, there can be as many as $n * (n-1)$ pairs, which could be a big number if n is quite large. Instead of using such an exhaustive approach, we start with an initial set of concept pairs (C_1, C_2), each pair appearing at least once in a sentence, with both C_1, C_2 from C. Among such pairs, those with high log-likelihood ratio scores are considered as the candidate pairs. A variant called SVO triples method (described in more detail in Section 5.2) further requires that C_1 and C_2 in a concept pair must be the subject and an object in a sentence respectively. This, of course requires syntactical parsing of sentences. The MINIPAR (Lin 1999) shallow parser is used for this purpose.

The log likelihood ratio measure is used in two places: To filter the initial concept pairs, and to determine the relation label for a candidate concept pair. To avoid repetition, formulation for computing log-likelihood ratios is explained in Section 5. The next section describes the VF*ICF measure for finding domain specific verbs as candidate relation labels.

4 Finding candidate relation labels

To find the relations between the concepts in concept pairs, we first identify candidate relation labels and then map the labels to concept pairs. This section describes the method employed to identify candidate relation labels. Intuitively, verbs occurring between the concepts in a sentence could be useful for labeling the relationships. Thus, it is reasonable to consider frequent verbs as candidates relation labels. However, many high frequency verbs are of the form *do*, *is*, *have*,..etc; which do not signify any semantic relation of the domain. To find domain-specific verbs, we define the VF*ICF metric, shown in (3), which is in the spirit of the TF*IDF metric used information retrieval. Informally, VF*ICF metric can be explained as follows. Verbs occurring with only a few set of concepts are more significant because they tend to be domain specific; in contrast, verbs occurring along with too many concepts tend to be overly general and do not denote important domain specific semantic relations.

$$VF * ICF(V) = (1 + logVF(V)) * log\left(\frac{|C|}{CF(V)}\right) \tag{3}$$

In (3), $|C|$ is the total number of concepts, $VF(V)$ is the count of the occurrence of verb V in domain texts and $CF(V)$ count of the concepts with which the verb V is associated. A verb V is considered to be associated with a concept c, if both of them occur in a sentence. Table 1 shows the top 10 verbs with corresponding VF*ICF values in the Electronic Voting domain. Evaluation of the VF*ICF metric for discovering domain specific verbs is presented in the experiments section.

Table 1 Top 10 verbs with high VF*ICF value in the EV domain

Verb(V)	VF*ICF(V)
Produce	25.010
Check	24.674
Ensure	23.971
Purge	23.863
Create	23.160
Include	23.160
Say	23.151
Restore	23.088
Certify	23.047
Pass	23.047

5 Assigning relation labels

5.1 The log likelihood ratio measure

Another component in relation extraction from domain texts is the assignment of labels for relations between the concepts. Here we use the domain specific verbs, extracted with the VF*ICF metric, as candidate relation labels. Relation label assignment is performed using the log-likelihood ratio measure as criterion function. The log-likelihood ratios are computed with the assumption of hypotheses H_1 and H_2 separately. Here hypothesis H_1 formalizes that the occurrence of a verb V is independent of the occurrence of a concept pair(C_1, C_2). Whereas H_2 formalizes that the occurrence of V is dependent on the occurrence of (C_1, C_2).

- **Hypothesis1 (H_1).** $P(V|(C_1, C_2)) = P(V|\neg(C_1, C_2))$
- **Hypothesis2 (H_2).** $P(V|(C_1, C_2)) \neq P(V|\neg(C_1, C_2))$

Now the log-likelihood ratio is computed using the (4).

$$log\lambda = log\frac{L(H_1)}{L(H_2)} \tag{4}$$

Here $L(H_1)$ and $L(H_2)$ represent the likelihoods of H_1 and H_2 respectively. Below are the details on computing $L(H_1)$ and $L(H_2)$. Let $S(C_1, C_2)$ be the set of sentences in which both the concepts C_1 and C_2 have occurred. Similarly, let $S(V)$ be the set of sentences in which the verb V has occurred. As usual, we use $|S|$ to denote the cardinality (number of elements) in a set S. Let

$$n_C = |S(C_1, C_2)|, \quad n_V = |S(V)|,$$

$$n_{CV} = |S(V) \cap S(C_1, C_2)|, \quad \text{and}$$

$$N = \sum_{i=1}^{n} \sum_{j,k=1}^{|C|} |S(V_i) \cap S(C_j, C_k)|.$$

Where n is count of domain-specific verbs and $|C|$ is the count of concepts in the set of candidate concept pairs.

Assuming H_1 is true, then

$$P(V|(C_1, C_2)) \;=\; P(V|\neg(C_1, C_2)) \;=\; p \;=\; \frac{n_V}{N}$$

$L(H_1)$, the likelihood of H_1 is computed by

$$L(H_1) = \mathbf{B}(n_{CV}; n_C, p)\mathbf{B}(n_V - n_{CV}; N - n_C, p). \tag{5}$$

In the same way, assuming H_2 is true, then

$$P(V|(C_1, C_2)) \;=\; p_1 \;=\; \frac{n_{CV}}{n_C}$$

and

$$P(V|\neg(C_1, C_2)) \;=\; p_2 \;=\; \frac{n_V - n_{CV}}{N - n_C}.$$

$L(H_2)$, the likelihood of H_2 is given by

$$L(H_2) = \mathbf{B}(n_{CV}; n_C, p_1)\mathbf{B}(n_V - n_{CV}; N - n_C, p_2). \tag{6}$$

Here $\mathbf{B}(k; n, x) = \binom{n}{k}x^k(1-x)^{n-k}$. $L(H_1)$ and $L(H_2)$ are computed assuming a binomial distribution of the observed frequencies.

Similar formulation for collocation discovery using log-likelihood ratios is mentioned in Manning and Schütze (1999)(Section 5.3.4) and Dunning (1993). Since we want the triples with high $L(H_2)$ and low $L(H_1)$ scores, we multiply $\log\lambda$ with -2. It is also mentioned in Manning and Schütze (1999) that if λ is the likelihood ratio then the quantity $-2log\lambda$ is asymptotically χ^2 distributed. For our purposes, we consider the triples$(C_1,\ V,\ C_2)$ with high $-2log\lambda$ score as valid non-taxonomic relations of the domain. From now on, we call $-2Log\lambda$ the *log likelihood score*. The above formulation, likelihood estimate, is also used to determine whether concepts(C_1, C_2) are related, as mentioned in Section 3.

5.2 RCL and SVO methods

Using the above mentioned components, two variants of our method are developed for the relation extraction task. One is called Related Concepts Labeling(RCL) and the other is Subject-Verb-Object(SVO) Triples method. These two methods use a slightly different criterion in selecting initial concept pairs, while they use the same metric for assigning relation labels. In the RCL method, a concept pair (C_1, C_2) is considered an initial concept pair if both C_1 and C_2 are domain specific concepts, and they occur together in at least one sentence. The SVO Triples approach imposes a stronger requirement for an initial concept pair (C_1, C_2): on top of the conditions of the RCL method, the pair must also satisfy that C_1 occurs as the subject and C_2 as an object in a sentence. To determine the subject and object(s) of a sentence, MINIPAR (Lin 1999) shallow parser is used. Dependency relations produced by MINIPAR are analyzed (automatically) to identify the subject and object(s) of a sentence. Initial concept pairs produced by the RCL method are further filtered by the parsing results to generate the set of initial concept pairs by teh SVO method. Clearly the SVO method will produce fewer number of initial concept pairs.

Once initial concept pairs are obtained (either by RCL or SVO method), the following steps are performed, for both methods. First, from initial concept pairs, those with high log likelihood ratio scores are chosen as candidate concept pairs. Next, for each such concept pair, candidate triples are formed by the concept pair plus verbs with high VF*ICF scores, occurring together with the concept pair in a sentence. Again, log likelihood ratio score is used to identify valid triples. Evaluation of both RCL and SVO Triples methods along with the AE measure(described in related works section) is presented in the following section.

6 Experiments

The presented approaches for the extraction of related concept and the identification of relation labels are experimented with the *Electronic Voting* (EV) domain collected from New York Times. Later on, experiments are conducted with the Tenders and Mergers (TNM) domain texts, which is a much larger collection.

From the EV domain texts, a total of 164 concepts are obtained after automatic concept extraction and manual filtering. Experimental results of the VF*ICF metric for domain specific verbs and relations extraction methods(RCL and SVO) are shown as follows.

6.1 Evaluation of VF*ICF metric on the EV domain

Using extracted concepts/verbs, and texts from the EV domain, the VF*ICF score for each verb in the text is computed. We initially removed the stop words from the extracted verbs. The top 20% (by VF*ICF score) of the remaining verbs are considered domain specific, thus they are candidate relation labels. To assess the performance of the VF*ICF metric, each of the verbs is manually classified as either relevant or not. Whether a verb is relevant or not is determined based on the authors knowledge about the domain. After the manual classification, the precision score for top 20% of verbs is 57%. To give an intuition on what kind of verbs we considered as relevant to the domain, Some of the relevant and irrelevant verbs for relation labeling in the EV domain are shown in Table 2.

From the performance of the VF*ICF metric, we believe that further research is required in finding candidate relation labels. We also think that using only verbs for relation labeling is not sufficient. The section on discussions will touch more on this topic.

Table 2 Some relevant and irrelevant labels on the EV domain

Relevant	Irrelevant
Make	Say
Vote	Try
Produce	Ensure
Cast	Know
Certify	Tell
Install	Help
Count	Believe
Elect	Want

6.2 Evaluation of RCL and SVO methods on the EV domain

Because the lack of gold standard for domain relation extraction, it is difficult to verify the performance of methods for the task. For example, to compute the recall of a relation extraction method on a collection of domain texts, one needs to know all valid relations of the domain represented in the text collection. In our experiments, the criteria for evaluate a method's performance are the accuracies of the results produced. Here accuracy is defined as the percentage of correct relation labeling. Further more, the accuracy evaluation of the methods is performed based on the following three constraints.

- **Constraint 1**. In a concept pair(C_1, C_2), C_1 and C_2 are related by a non-taxonomical relation.
- **Constraint 2**. In a triple(C_1, V, C_2), V is the label for relation either $C_1 \rightarrow C_2$ or $C_1 \leftarrow C_2$.
- **Constraint 3**. In a triple($C_1 \rightarrow V \rightarrow C_2$), V is a label for the relation from $C_1 \rightarrow C_2$ only.

Constraint **1** verifies whether the concepts in the concept pair are related via a non-taxonomical semantic relation. Since both RCL and SVO Triples methods extract candidate concept pairs and then assign relation labels to candidate pairs, this evaluation is useful to verify the accuracy in extracting candidate concept pairs. Constraint **2** is useful for assessing whether the assigned label is valid for the relation on a concept pair without considering relation directions. Similarly constraint **3** tests whether the direction of the relation is maintained. Verification with respect to constraint **3** is required because the direction of a relation is important for ontological relations. For example in the triple `(voter, cast, ballot)`, the label `cast` indicates the relation from `voter` to `ballot` but not in the reverse direction.

As mentioned in Section 3, two concepts appearing together in at least one sentence are considered an initial concept pair by the RCL method. With the above notion, a total of 184 concept pairs are obtained from the EV domain texts. Of these pairs, the top 20% pairs (37 total) with high log-likelihood scores are considered as candidate pairs. For illustration, some candidate concept pairs, each verifying a non-taxonomic relation, are shown in Table 3.

Now the verbs produced via the VF*ICF metric are used to determine the relation labels for the candidate pairs. For each candidate concept pair, verbs with high VF*ICF values, occurring at least once together with the concept pair in a sentence, are considered as the candidate labels. Among all the candidate labels for a candidate concept pair, the verb with the highest log likelihood score is considered as the label

Table 3 Example concept pairs by RCL method on the EV domain

Concept pairs(C_1, C_2)
(Election, official)
(Company, voting machine)
(Ballot, voter)
(Manufacturer, voting machine)
(Polling place, worker)
(Polling place, precinct)
(Poll, security)

Table 4 RCL method example results on the EV domain

Concept(C_1)	Label(V)	Concept(C_2)
Machine	Produce	Paper
Ballot	Cast	Voter
Paper	Produce	Voting
Polling place	Show up	Voter
Polling place	Turn	Voter
Election	Insist	Official
Ballot	Include	Paper
Manufacturer	Install	Voting machine

for the relation between the concepts in the pair. Some of the concept pairs along with their semantic relation labels obtained by the RCL method are shown in Table 4.

Even though the RCL technique is able to extract suitable relation labels for candidate concept pairs, some of the labels are in the wrong direction. Namely, they indicate the relationship $C_2 \rightarrow C_1$ rather than $C_1 \rightarrow C_2$. For instance, rows 2, 3, 4, 5, and 6 in Table 4 show the relationship $C_2 \rightarrow C_1$. The main disadvantage of the RCL method is that it is unable to find the correct relationship direction assigned to concept pairs. From these observations, it is clear that RCL method performs poorly on satisfying the constraint **3**. Accuracy scores based on each of the constraints for RCL are shown in Table 6.

Another approach we employed to extract the relations is SVO Triples method. In SVO Triples method, in each of the candidate triples(C_1, V, C_2), C_1 has to be the subject and C_2 has to be an object in a sentence containing the verb V. Moreover, V must have a high VF*ICF score, and the pair (C_1, C_2) must have a high log likelihood score. Because of the above restrictions, very few(only 19) triples are obtained for the EV domain texts. But among the triples obtained, most of them are valid semantic relations satisfying constraint **3**, i.e. direction of the relationship maintained. For illustration, some triples obtained with SVO Triples approach are shown in Table 5. Accuracy results for SVO method over the EV domain texts are shown in Table 6.

Table 6 shows the performance of RCL, SVO and the AE measure (1) with respect to the constraints **1**–**3**. The first column shows the method applied. The second column shows accuracy of the methods according to the constraint **1**. Similarly, columns 3 and 4 indicate accuracies of the corresponding methods with respect to

Table 5 SVO method sample results on the EV domain

Concept(C_1)	Label(V)	Concept(C_2)
Machine	Produce	Paper
Voter	Cast	Ballot
Voter	Record	Vote
Official	Tell	Voter
Voter	Trust	Machine
Worker	Direct	Voter
County	Adopt	Machine
Company	Provide	Machine
Machine	Record	Ballot

Table 6 Evaluation of RCL and SVO methods on the EV domain

Method	(C_1,C_2)	(C_1, V, C_2)	$(C_1 \rightarrow V \rightarrow C_2)$
AE measure	89.00	6.00	4.00
RCL	81.58	30.36	9.82
SVO triples	89.47	68.42	68.42

constraints **2** and **3** respectively. The first row of Table 6 shows the results of the AE measure (1) presented in Kavalec et al. (2004). The AE measure is used in Kavalec et al. (2004) to select from an initial set of triples (C_1, V, C_2) such that C_1 and C_2 (both domain specific) appear within a pre-defined distance (eight words in Kavalec et al. (2004)) from V in texts. We also implemented this AE measure and applied it to the EV domain texts. One can see clearly that the AE measure performed rather poorly in assigning relation labels and in maintaining relationship directions as well, although it is able to extract related concepts with high accuracy. Also, the significant increase of accuracy by RCL and SVO method over the AE method with respect to constraint **2** (column 3) indicates that the VF*ICF measure useful for filtering some of the irrelevant relation labels.

The Second row in Table 6 shows the results of the RCL method described. The RCL method is able to extract concept pairs such that the individual components are semantically related. But in finding the relationship labels for such pairs, RCL is not quite successful. We believe the main reasons for such a low accuracy on finding the labels are as follows. Concepts in some of the concept pairs occur more as a compound term in texts rather than connected by some verb. For example, the compound term `voting machine` has occurred more often on its own than the concepts `voting` and `machine` are connected by some verb. Another reason is some of the concepts, occurring together more often, are connected by a preposition or a conjunction rather than a verb showing the relationship between them. For example, in the sentence `there were constant problems with the hardware and software`, the occurrence of concepts, `hardware` and `software`, does not signify a semantic relation labeled by a verb. Further more, using the verbs occurring along with the concept pair in a sentence may indicate a relationship between some other concepts in the sentence, rather than between the concepts in the pair. These reasons could be also used to explain why the AE measure performed poorly with respect to the constraints **2** and **3**, while it fares very well with respect to constraint **1**.

Because of the poor performance in satisfying the constraint **3** for RCL, restrictions on selecting candidate triples, mentioned before, are applied in SVO Triples method. With those restrictions, most of the concept pairs obtained are indeed related. Among the obtained concept pairs, very few of them got invalid labels. The few invalid labels might have been obtained due to parse errors. Further more, all of the valid relations obtained using SVO Triples method maintained the direction of the relationship $(C_1 \rightarrow C_2)$. Even though most of the relations obtained by SVO method are valid, SVO Triples method extracts only a small fraction of the total relations from domain texts. Hence SVO Triples method gives poor coverage (recall). From the experiments, we believe that the SVO method alone is not sufficient to find all non-taxonomic relations of the domain, even though it is quite useful for extracting some of them. Further research is needed to find semantic relations which do not occur as subject and object(s) in the texts.

6.3 Experiments with the tender offers and mergers domain texts

To further test the SVO Triples method, experimental evaluation is performed on Tender Offers, Mergers, and Acquisitions(TNM) domain texts. The TNM data is extracted from 1987, 1988, and 1989 years news articles of the Wall Street Journal corpus obtained from TREC. It consists of 272 news articles which amounts to a total of 30MB texts. For the TNM data, TOP 1% of the nouns with high Tf*Idf values are considered domain specific concepts (with NO manual removal of the irrelevant ones). A total of 1833 concepts are obtained. These concepts and domain texts are inputs to the SVO Triple method and to the AE measure for non-taxonomical relation extraction. Using the SVO triple approach, a total of 24,329 triples are formed. The top 200 triples with high log-likelihood ratio scores are selected and manually evaluated against the three constraints mentioned in the Section 6.2.

Similarly, to evaluate the AE measure (1) on the TNM data, AE triples(C_1, V, C_2) are formed such that concepts C_1 and C_2 within eight words apart from a verb V. With this approach a total of 3,629,981 triples formed. From this large number of triples, 30,000 triples with high frequencies are considered for computing the AE measure. Among these, 200 triples with high AE measure are evaluated against the three constraints mentioned in Section 6.2. The results of AE measure and SVO Triples method are shown in the Table 7.

From these results, it is clear that the AE measure outperformed our method in finding the concept pairs such that their components are semantically related. This is probably because focusing on the most frequent triples allows the AE measure to better weed out invalid concept pairs. On the other hand, with respect to constraints **2** and **3**, the SVO Triples method outperformed the AE measure. This is because, in SVO triples method, (C_1, V, C_2) triples are formed such that C_1 is the subject and C_2 is the object of V in a sentence, whereas in AE measure method, (C_1, V, C_2) triples are formed such that C_1 and C_2 are within eight words apart from V. Hence, even though C_1, C_2 are in the same sentence, they many not be related by the semantic relation described by V. This is the reason for such a low accuracy in relation labeling with the AE measure. Even though SVO Triples method has better performance compared to the AE measure in labeling the relations, its accuracy is not very high. The SVO triples method depends heavily upon the quality of the concepts extracted in concept extraction step. If the concepts are not domain specific, concepts pairs formed by the SVO method may not be semantically related. The conceptual terms extracted from the TNM data contain many general terms such as “short”, “class”, “delay”, and pronouns such as , “they”, “him”, “who”, etc. The concept pairs containing this kind of terms are not useful for labeling the relations. These pairs penalized the accuracy in finding non-taxonomically related concept pairs. This is an error carried over from concept extraction to relation extraction. Thus, an important lesson is that getting high quality concepts is crucial for obtaining good performance for the semantic relation extraction task.

Table 7 Evaluation of SVO and AE methods on the TNM domain

Method	(C_1,C_2)	(C_1, V, C_2)	($C_1 \rightarrow V \rightarrow C_2$)
AE measure	66.00	12.50	9.50
SVO triples	42.00	38.00	38.00

Upon removal of triples containing pronouns such as "they", "who", "him" from those top 200 triples produced by the SVO method, the accuracies with respect to constraints **1, 2, 3** by the SVO method become 50 percent, 45 percent and 45 percent respectively. This suggests that some simple post-processing could be done to further improve the accuracies of the SVO approach.

7 Discussion and future work

From the results of RCL and SVO Triples methods over the EV domain, it is clear that RCL method performs poorly for the domain specific relation extraction task. Even though SVO Triples method is able to identify non-taxonomical relations with reasonable accuracy, the relations extracted are limited to those represented by a SVO triple. To improve the coverage of of non-taxonomic relations, even better methods are required. Since relation labels obtained using VF*ICF measure has only 57% accuracy, we believe considering only the verbs as candidates for relational labels is not sufficient. As part of the future work, we are interested in using words with other parts of speech tags also as candidates for relation labeling. For example, prepositional phrases could be explored as candidates for finding semantic relations. The main idea of extracting relations from prepositional phrases is the following. In a given prepositional phrase, two concepts could be linked by a preposition, the relationship between the concepts is labeled based on the semantic classes of the concepts and the associated preposition. For example, in the phrase "`hand recount of paper ballots`", the semantic class of `hand recount` is "Action" and that of `paper ballot` is "Object". Using pre-learned patterns based on the semantic classes and the prepositions involved, relationships between the concepts could be automatically labeled as one of the predefined relation labels.

The performance of the SVO method on the TNM domain texts is not satisfactory yet. The poor performance may be attributed, for a significant part, to errors carried over from domain concept identification. This suggests the need for further works to improve the quality of domain concepts extracted. Moreover, one could possibly consider combining the AE measure with our SVO method for the relation extraction task, given that the AE measure performs quite well with respect to constraint **1**, and SVO method generally does better with respect to constraints **2** and **3**.

8 Conclusions

Extracting non-taxonomic semantic relations from domain texts is an important component for automated ontology learning. In spite of its importance, non-taxonomic relation extraction is relatively less tackled. In this paper, we present an unsupervised, general technique for this task. We consider the domain-specific verbs occurring along with domain specific concept pairs as the candidate relation labels. The VF*ICF metric, in the same spirit as the TF*IDF metric in information retrieval, is defined and applied to find domain-specific verbs. Candidate triples of the form (C_1, V, C_2) are evaluated by the log likelihood ratio score, a statistical technique for estimating the connection strength between the verb V and the concept pair (C_1, C_2).

Two variants (RCL and SVO) of our method are described. Empirical evaluations of each of the methods with respect to three different constraints are performed on the Electronic Voting (EV) domain. The SVO method is also compared with one of the existing methods using the AE measure, on both the EV domain texts and the TNM domain texts. The SVO method achieves better labeling accuracy in both domains, although its accuracy over the TNM domain is still not satisfactory. The experimental results indicate that the SVO Triples method is useful for finding non-taxonomic relations from domain texts. However, the SVO method alone is not sufficient to extract all non-taxonomical relations for a domain. Methods utilizing other syntactic/semantic information from texts are needed to improve the recall and precision of relation extraction. Better concept extraction methods could also significantly improve the quality of non-taxonomical relations extracted.

Acknowledgements This work was partially supported by the NSF grant ITR-0326387 and AFOSR grants FA9550-05-1-0454, F49620-03-1-0238, F49620-03-1-0239, and F49620-03-1-0241. The authors thank the reviewers for comments and suggestions that help to improve the presentation of the paper.

References

Berland, M., & Charniak, E. (1996). Finding parts in very large corpora. In *Proc. 37th annual meeting of association for computational linguistics* (pp. 57–64).

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). *The semantic web* (pp. 30–37). Scientific American.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Third conference on applied natural language processing*.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proc. 37th annual meeting of association for computational linguistics*.

Cederberg, S., & Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proc. of conference on natural language learning*. Edmonton, Canada.

Ciaramita, M., Gangemi, A., Ratsch, E., Jasmin, S., & Isabel, R. (2005). Unsupervised learning of semantic relations between concepts of molecular biology ontology. In *Proc. of 19th international joint conference on artificial intelligence*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*, 61–74.

Faure, D., & Nedellec, C. (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *Proceedings of LREC workshop on adapting lexical and corpus resources to sublangauges and applications* (pp. 5–12).

Fu, J., Fan, X., Mao, J., & Liu, X. (2009). Two stage semantic relation extraction. In *Proceedings of international conference on hybrid intelligent systems*. Shen Yang, China.

Girju, R., Badulescu, A., & Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proc. HLT/NAACL-03* (pp. 80–87). Edmonton, Canada.

Girju, R., & Moldovan, D. (2002). Text mining for causal relations. In *Proc. of FLAIRS conference* (pp. 360–364).

Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proc. of association of computational linguistics*.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. 14th international conference computational linguistics*. Nantes, France.

Jacquemin, C. (1995). A symbolic and surgical acquisition of terms through variation. In *Symbolic approaches to learning for natural language processing* (pp. 425–438).

Kavalec, M., Maedche, A., & Svatek, V. (2004). Discovery of lexical entries for non-taxonomic relations in ontology learning. In *SOFSEM—theory and practice of computer science. LNCS* (Vol. 2932). Springer

Lin, D. (1999). MINIPAR: A minimalist parser. In *Maryland linguistics colloquium*. University of Maryland, College Park.

Maedche, A., & Volz, R. (2001). The text-to-ontology extraction and maintenance system. In *ICDM-workshop on integrating data mining and knowledge management*. San Jose, California.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). An introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235–244.

Pantel, P., & Lin, D. (2001). A statistical corpus based term extractor. In E. Stroulia, & S. Martwin, (Eds.), *AI lecture notes in artificial intelligence* (pp. 35–46).

Punuru, J., & Chen, J. (2006). Automatic acquisition of concepts from domain texts. In *Proc. of IEEE granular computing Atlanta*.

Punuru, J., & Chen, J. (2007). Extraction of non-hierarchical relations from domain texts. In *Proc. of IEEE symposium on computational intelligence and data mining*. Honolulu.

Qian, L., Zhou, G., Kong, F., & Zhu, Q. (2009). Semi-supervised learning for semantic relation classification using stratified sampling strategy. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 1437–1445). Singapore.

Ramshaw, L., & Marcus, M. (1995). Text chunking using transformation-based learning. In *Third association for computational linguistics workshop on very large corpora*.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In *Proc. of 13th national conference on artificial intelligence* (pp. 1044–1049).

Schutz, A., & Buitelaar, P. (2005). RelExt: A tool for relation extraction from text in ontology extension. In *Proc. of 4h international semantic web conference (ISWC-2005)*. Galway, Ireland.

Stevenson, M. (2004). An unsupervised wordnet-based algorithm for relation extraction. In *Fourth international conference on language resources(LREC-04)*. Lisbon, Portugal.

Tomokiyo, T., & Hurst, M. (2003). A language model approach for keyphrase extraction. In *Proc. of ACL 2003 workshop on multiword expressions: Analysis, acquisition, and treatment* (pp. 33–40).

Yangarber, R., Grishman, R., Tapanainen P., & Huttunen, S. (2000). Unsupervised discovery of scenario-level patterns for information extraction. In *Proc. of applied natural language processing conference*. Seattle, WA.

Zhou, G., Li, J., Qian, L., & Zhu, J. (2008). Semi-supervised learning for relation extraction. In *Proceedings of international joint conference on natural language processing* (IJCNLP08) (pp. 32–38). Hyderabad, India.