Towards an Enhanced Framework for Learning Semantic Relation Classification

Amaal Saleh Hassan
Sultan Qaboos University
Oman
amaalh@squ.edu.om

Narayanan Kulathuramaiyer
Universiti Malaysia Sarawak
94300 Kota Samarahan
Sarawak, Malaysia
nara@fit.unimas.my

Abstract—**Most of the research in this area depends on NLP techniques, machine learning, and statistical approaches, but the challenging issue here is to provide a general learning framework in an automated way that make use of different kinds of contextual knowledge, and to make use of that framework to enrich the ontology with new relations, hence facilitate ontology acquisition. We have developed a unique framework for acquiring ontology relations from a large collection of domain independent texts. Our experiments on the proposed approach have shown promising results, even at an early state. The main contribution of this work lies in the semantic validation of causation relationships based on thorough corpus linguistics.**

Keywords-ontology acquisition; ontology learning; lexical relations.

I. INTRODUCTION

Ontologies represent "an explicit specification of a conceptualization" where a "conceptualization" is the "set of objects, concepts, and other entities that are assumed to exist" in some domain (Gruber, 1995). They have shown their usefulness in a wide range of application such as intelligent information integration, question answering, NLP and many other applications. However, their wide-spread usage is still hindered by ontology acquisition being rather time consuming and, hence, expensive [1].

A number of proposals have been made to facilitate ontological acquisition through automatic discovery from domain-specific natural language texts [2]. Nevertheless, most of these approaches have only looked at how to learn the taxonomic part of ontologies. A typical approach collects relevant domain concepts and clusters them into a hierarchy using combinations of statistic and linguistic data. Though this in itself is helpful, but major efforts in ontology engineering are required to be dedicated to the definition of non-taxonomic conceptual relationships, e.g. Has_Part, Cause_Effect,Contain_Container, etc, relations between concepts. Even the methods that address the non-taxonomic relations did not come up with a state level in enhancing the process of classifying and extracting semantic lexical relations [1]. Most of the systems concentrate on just classifying the relations without giving a solution for how it can be created. And those that provide methods for creating the relations did not consider the context in which the relations might occur in [3].

Ontology learning from texts constitutes a promising means for ontology acquisition to significantly speed up the ontology building process. In this process, the phase of extraction of non-taxonomic relationships has been recognized as one of the most difficult (102) and least tackled problems [4]. Non-taxonomic relations between concepts "appear as a major building block" in common ontology definitions. In fact, their definition consumes much of the time needed for engineering ontology.

This phase can be divided into two different problems :

- Discovering the existence of a relationship between a pair of concepts .
- Labeling this relationship according to its semantic meaning .

The assignment of labels to relationships is also difficult since various relationships among instances of the same general concepts are possible [5]. Moreover, even if the semantic is clear, it might still be hard to guess which synonymous labels are preferred by a certain community [6].

According to these facts, this field would be greatly getting advantage if the learning algorithms used to classify the lexical semantic relations, are able to generate new relations automatically. There is also a requirement for the learning process to fulfill suitable conditions, i.e. no ambiguity, good learning examples, learning algorithm with all possibilities, contextual properties, and concerned features to generate a comprehensive learning model for that lexical relation. Also most of the projects have been focused on the construction of the lexical semantic relations within a specific domain, rather than generating the patterns that can work as a discoverer and generator of such relations.

In this research we focus on the learning of semantic relations patterns between word meanings by taking into consideration the surrounding context in the general domain for discovering causation relations. We believe that extracting learning criteria is much more effective than just discovering the relations, because the learned system can be used to

978-1-61284-130-4/11/$26.00 ©2011 IEEE

discover new relations not only the set dedicated to a specific corpus. The other advantage is that these patterns can be used to generate new relations regardless of the domain, because mainly they depend on the syntactic and semantics of the context rather than on the specific meaning of the domain.

And why causation relations, specifically these relations (we say relations because there is a range of relations classified under this kind different according to many researchers e.g. cause, effect, purpose, etc.) its importance came not just as being semantic relations that represent the meaning involved in the text, but also as a tool for extraction information in question answering systems and information retrieval

The remaining parts of the paper are organized as follows; second section provides literature review, third section provides current approaches in handling lexical semantics for the sake of ontology learning. Then section four provides the enhanced framework, with the constituent algorithms. Section five discusses the implementation issues and expected results.

II. Related Works

In recent years, the acquisition of ontologies from domain texts using machine learning and text mining methods has been proposed as a means of facilitating the ontology engineering process. In this context, ontology learning has been identified as an emerging field which aims at assisting knowledge engineers as well as end-users in ontology construction. It can be seen as a multi-disciplinary field, which integrates disciplines such as ontology engineering, machine learning, and natural language processing, among others. The use of these technologies is distributed in three main phases, lexical entry extraction, taxonomy extraction, and non-taxonomic relation extraction [6].

Developing systems based on this kind of information is not new. Many systems have been developed like (Garcia 1997) used verbs as causal indicators for causal knowledge acquisition in French. Reference [7] and (Khoo et.al.2000) acquired causal knowledge with manually created syntactic patterns specifically for the MEDLINE text database. Works by [8] and (Inui Takashi 2003) then explored the acquisition of causal knowledge by using connective markers.

It has to be noted that none of these systems actually made use of the semantic information that can be constructed from the sentence utterances in a causation scope. Most of the rules derived are hand crafted and these systems handle simple part of the expressions.The work of Khoo is an exception in that it had produced a comprehensive treatment of causation expressions. This work however did not subsequently incorporate any semantic information in their analysis.

In our system we will take both kinds of information into consideration, we will first extract the causation patterns automatically and then perform semantic analysis of its component to learn new rules that are supported by both syntactic and semantic information.

III. An overview of Lexical semantics

Most NLP systems based on the semantic representation of the texts. In order to represent the semantics of the texts we need to extract the semantic relations that connect different constituents of it. This is called lexical semantics. This semantic representation can be abstracted clearly through ontologies. But this process is not clearly identified and consumes time especially in such NLP kind of ontologies where so many factors and variables need to be included.

Lexical semantic representation of text meaning, facilitates inferences, reasoning, and greatly improves the performance of Question Answering, Information Extraction, Machine Translation and other NLP applications. Broadly speaking, semantic relations are unidirectional underlying connections between concepts and hence the words that representing that concept. For example, the noun phrase "car's engine" encodes a part-whole relation: the engine is a part of the car, and "Mary's brother" indicate kinship relation, while viral infection represent cause-effect relation.

There is a growing interest in text semantics field by the new wave of semantic technologies and ontology that aim at transforming unstructured text into structured knowledge. Many studies inducted for studying lexical semantics through different approaches, which include:

- Statistical approaches like Blaheta and Charniak , and Gildea.
- Learning approaches using different learning algorithms like Generative models for semantic roles (Thompson), Decision trees (Girju, Badulescu and Moldovan), (Rosario and Hearst, Neural networks (Rosario, Hearst and Fillmore).
- Knowledge based methods depending on the available many lexical resources like MRD,lexical ontology like wordnet, framenet, and annotated corpuses.
- Hybrid approaches that make use a combination of the previously mentioned methods.

But not all the lexical syntactic patterns clear in representing one syntactic relation, some of these patterns may represent more than one relation, this adds more labor in finding the suitable patterns that represent the relation, besides some semantic constrains to enforce such distinctions in discovering, for example:property (“Hussain's cleverness”), part-whole (“Hussain's hand”), depiction/depicted(“Hussain's photo”), source/from (“Hussain's birthplace”), location (“Hussain's city”), kinship ("Hussain's brother").

IV. Discovering Causation from text

The main representation scheme which expresses causation patterns can be represented as following (detailed description can be found in [9])

1. using *causal links* to link like (see classification below based on [10]).
2. using *causative verbs (complete list in appendix A).*
3. using *resultative* constructions like (like the pattern V-NP-ADJ).
4. using *conditionals*, i.e. "if ... then ..." constructions.
5. using *causative adverbs and adjectives(not used).*

All these categories are handled in this research except for the fifth category "causative adverbs and adjectives". We then [10] classification of causal links into four main types:

a. The adverbial link, e.g. *so, hence, therefore.*
b. The prepositional link, e.g. *because of, on account of.*
c. Subordination, e.g. *because, as, since.*
d. The clause-integrated link, e.g. *that's why, the result was*.

Mining causality knowledge induces knowledge of reasoning that is beneficial for our daily use in a variety of diagnosis problems. There are three main problems in the causality or cause-effect extraction; cause-effect identification, causality ordering and cause-effect boundary determination. The cause-effect identification and the causality ordering problems can be solved by learning different elementary discourse units and learning lexico syntactic pattern (i.e., if NP1 then NP2). The cause effect boundary can be solved by learning semantic constrains of the cause and effect part, though this problem needs more investigation.

V. PROPOSED FRAMEWORK

Our notion of Ontology Learning aims at the integration of a multitude of disciplines in order to facilitate the machine learning process. Because the fully automatic acquisition of knowledge by machines remains in the distant future, we will consider the overall process of ontology learning as semi-automatic with human intervention in specific places such as to validate input samples.

Our work focuses on the semantic relations within specific text constituents such as nominal phrases, and verbal phrases, where the causation relation may be expressed in various formats, among different combination of these phrases.

As the ontology is a wide system of a number of subsystems, we will focus on the classification of semantic relations between concepts and in finding an enhanced framework for learning more classification rules and the generalization of these rules with the aim of enriching the ontology with new relations in an automated way. Our learning framework is outline in Figure 1 below.

As any NLP system, our project made use of many of knowledge resources to support its flow of process and to improve the performance of its algorithms. The criteria in choosing these resources can be abstracted in domain generality, relation relatedness, and Confidence.

The framework makes use of variety knowledge sources. This includes NOMLEX (dictionary of nominalizations, Proteus Project, New York University), SemCore3.0, WordNet3.0, extended WordNet3.0 (WordNetGloss3.0), SemEval2007 data, and SemEval2010 data as a training corpus. Besides that we propose new learning approaches using the iterative semantic abstraction algorithm (ISA) and the pruning schema algorithm (PSA). Including the decision tree C5.0 to classify the correct rules for each relation pattern.

The framework will go in two main phases:

Phase one, causation patterns acquisition (fig. 2).

The input knowledge sources are of different format as we mention before. So each one need specific kind of preprocessing so that they all contain clearly annotated knowledge in terms of POS, WSD, WN senses and syntactic parsing.

For resources without WN sense annotations, we adopt the result of previous studies of causations provided by (Cristina Butnariu.et.al.2008). She provided some general semantic cover set of features for cause and effect relations after SemEval2007. The approach is appropriate as the proposed set covers a good percentage of the SemEval2007 data set.

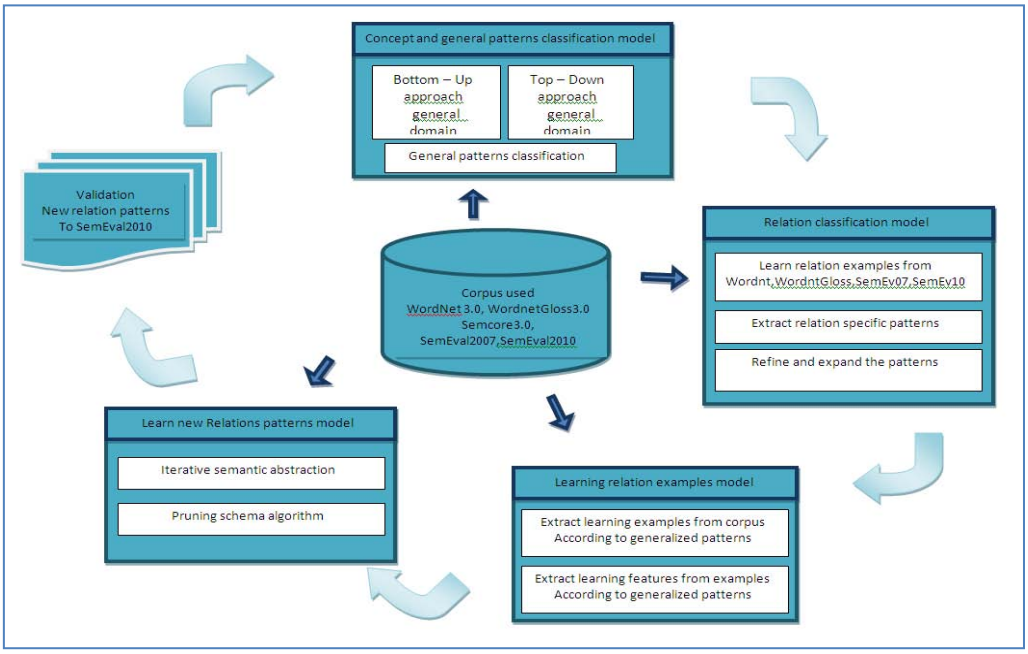

Figure 1. An enhanced framework for learning ontology relations

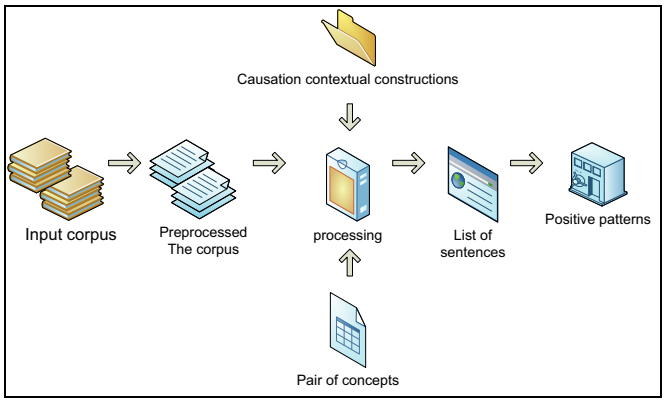

Figure 2. Causation patterns acquisition

The semantic cover set makes use of WN hierarchies that represents a class of word senses related in the hyponame chain. The cover set includes:

- Causes to include the following WordNet categories and their descendants: [{causal_agent},{psychological_feature},{attribute},{substance},{phenomenon},{communication}, {natural_action}, {organic_process}].
- Effects to include: [{psychological_feature},{attribute},{physical_process}, {phenomenon},{natural_action},{possession} , {organic_process}].

In implementing the cover set we have developed the following heuristics for assigning WN senses. For each term:

- Identify the sense that can be derived from one of its hypernyms leading to a member in the cover set.
- If more than one identified, choose the sense with the highest frequency of usage according to WN factor.

After preprocessing the resources phase one will go through the following steps:

1. Specify what causation contextual information to handle.
2. From corpus extract sentences that hold such information. Pass this set to step 4.
3. Specify set of concepts pairs of causation relation extracted from WordNet, SemEval2007, and SemEval2010.
4. From the annotated corpus extract the sentences that hold the pairs.
5. Apply set of negative patterns from [9] to help the system avoid the wrong identification.
6. Analyze the sentences guided by the causation general patterns to extract linguistics patterns for cause an effect from the resultant set of sentences.

These patterns not just specifying causation relations in text but also they specify the direction of the relation, which is cause and which is effect. For example, the pattern

[effect] *is the result of* [cause]
or [cause] *result in* [effect]

Several sets of patterns in this research have been developed trying to make a balance between the coverage of patterns for causation and the requirement of the learning tool C5.0. There will three main divisions of patterns in which each one will include specific set of causation general patterns.

1. Patterns involving a causal link and *if-then* conditional.
2. Patterns involving causative verbs.
3. Patterns involving resultative constructions.

Each pattern consists of a sequence of slots separated by a space. Each slot indicates a component like (particular word, POS, phrase, and a set of sub patterns).

For example the adverbial link (e.g. so, hence, therefore) patterns looks like

- $+ * C1 therefore ** C2 $-
- $+ * C1 ** therefore **C2*$-

Phase two, learning relation rules (fig. 3)

The learning process depends on a set of lexical, syntactic, and semantic features. These features control the classifier to generate certain rules. These features are

1. Lexical and contextual features
   a. Order of cause and effect.
   b. Causative constructions type.
2. Semantic features
   a. Wordnet hypername category.
   b. Cover set category.
   c. Verb ambiguity factor.

All these features are clear as its name explain it. Verb ambiguity factor is calculated depending on the number of senses of the verb and its frequency of usage provided by WordNet.

The iterative Semantic Abstraction (ISA) algorithm will classify the semantic relations and learn how to combine the input features in an automated unambiguous fashion. We will be dealing with patterns in three main categories and more subcategories, were each one have specific pattern.

For example the first main category of patterns is "Patterns involving a causal link and if-then conditional". This category sub divided into more five subcategories.

⇒ the adverbial link, e.g. (np|cause {so/hence/therefore/…} np|effect)
⇒ the prepositional link e.g. (np|effect { because of/ on account of/..} np|cause)
⇒ subordination, e.g. (because, as, since
⇒ the clause-integrated link, e.g. that's why, the result was.

These patterns will pass through the following steps:

- Select set of positive and negative examples for the sake of learning; this can be done by extracting sentences with and without causation relation from

the corpus. Using the causation patterns extracted in the previous phase.

- Analyze the sentences to extract cause and effect parts.
- Use wordnet semantic hypernym relations to create first abstract representation of the cause and effect parts.

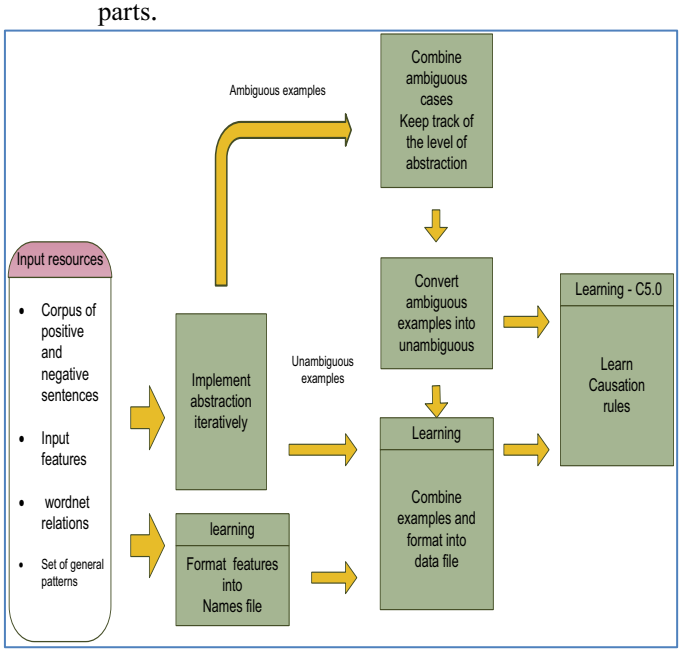

Figure 3. Learning rules framework

The Iterative Semantic Abstraction algorithm

To avoid ambiguity, the ISA algorithms will build the tree from the leaves till the root node. The algorithm will begin with abstracted input examples and using WordNet hypernym relations, continuously, will abstract the branch and after each level of abstraction the system examine the state to see if there is an ambiguous examples encountered. If no ambiguous examples encountered then the process continue until reaching the top nod of that hierarchy branch. Then using C5.0 learning tool, with the unambiguous hierarchies only the system can learn the rules from these examples. The record of data input to C5.0 will hold many slots as follows

[cause_abstract#,effect_abstract#,Order of cause and effect,Causative constructions type, Coverset category, Verb ambiguity factor,target_relation]

The output of this algorithm will be a set of rules in each nodes of the tree. To enhance the output of the learning process and to come up with more generalized classification rules, we created another algorithm for this purpose called the pruning schema algorithm (GSA). In this schema algorithm the system reformats the classification process it learned before into more reliable fashion with more generalization way without losing the main constrains of always producing the best quality features in the classification method it learned.

VI. Preliminary Results

Preliminary experiments have been carried out in exploring the proposed discovery of causation relationships. We have implemented each model separately. Table1 below shows the ability of the system in classifying causation relations from subset of the benchmark of data set provided by SemEval2010. We employ the SemEval2010 tagged dataset as a means to compare the performance of this phase with the systems participated in the competition. The subset was taken from the training corpus of 8000 sentences; our subset consists of 1000 sentence. We implement the *f.measure* as follows:

$$f.measure = \frac{2.(precision.recall)}{(precision + recall)}$$

$$precision = \frac{\text{no of relations correctly retreived}}{\text{no. of CE relations}}$$

$$recall = \frac{(no.of\ retreived\ relations)}{(no.of\ relations)}$$

TABLE I. Corpus Statistics

Corpus	No of sentences in corpus	No of sentences extracted by our system	No of sentences that hold the relation
SemEval2010	1000	130	90

TABLE II. Precision, Recall and F Measure

Corpus	precision	recall	F-measure
SemEval2010	0.69	0.76	0.72

Our results has been promising in that it has achieved a higher F measure that all the participants for the SEMEVAL 2010 competition except for UI…, which was marginally better. We are exploring enhancements which we will be able to report during the conference in July.

As we noticed and analyzed the patterns within the sample extracted we observe that in about 80% of the sentences, causation is expressed through causative constructs other than causative verbs. And about 20% of the sentences with causation relation make use of causative verbs.

We noticed also that most of the sentences produced by the system, are of malty components, several noun phrases and verb phrases before after and in between the terms, which make us adjust some of the general patterns set as we, did not use the expressions specified by SemEval for the relation only but also the surrounding information.

VII. 5. Discussion and Conclusions

One of the most expensive and demanding features of any NLP system is it requirement of a vast amount of knowledge sources for the sake of processing. We believe that a more useful long-term approach to the problem of knowledge acquisition for NLP ontologies would be to acquire what is needed from the outset; this would be represented by the

learned classified semantic relations that represent the core of the ontology.

To validate our approach, we used as an input resources to learn causation patterns set wordnet relations beside SemEval2010 training set. These will of course serve very well in provide a rich set of variety patterns for the causation relation.

Then usage of causation contextual information (e.g. causal links, causative verbs, etc.) will put more confidence in the proposed procedure, as these information comprehensively represent causation.

References

[1]. Ricardo Gacitua, Pete Sawyer, Scott Piao, Paul Rayson, "Ontology Acquisition Process: A Framework for Experimenting with different NLP Techniques ", 2007.

[2]. R. Byrd and Y. Ravin. "Identifying and extracting relations from text", NLDB'99 , 4th International Conference on Applications of Natural Language to Information Systems, 1999.

[3]. Iris Hendrickx , Su Nam Kimy , Zornitsa Kozarevaz , Preslav Nakovx , Diarmuid ´O S´eaghdha, Sebastian Pad´ok , Marco Pennacchiotti__, LorenzaRomanoyy, Stan Szpakowiczzz, SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominal, SemEval2010,2010.

[4]. D. Sanchez, A. Moreno, "Learning non-taxonomic relationships from web documents for domain ontology construction", Data & Knowledge Engineering 64 (3) (2008).

[5]. Maedche, Alexander, "Ontology Learning for the Semantic WebSeries" The Springer International Series in Engineering and Computer Science, Vol. 665, 2002.

[6]. Martin Kavalec,VojtěchSvatek, "A Study on Automated Relation Labeling in Ontology Learning", Ontology Learning from Text: Methods, Evaluation and Applications. IOS, (2005).

[7]. Khoo, C., Kornfilt, J., Oddy, R., Myaeng, S.H.: "Automatic extraction of cause-effect information from newspaper text without knowledge-based inference". Literary & Linguistic Computing, vol. 13(4), pp. 177—186, 1998.

[8]. Girju, R., Moldovan, M.: Text mining for causal relations. In: Proceedings of the FLAIRS Conference, pp. 360--364 (2002)

[9]. Khoo, C, PhD thesis, "automatic identification of causal relations in text and their use for improving precision in information retrieval", 1995.

[10]. Altenberg, B., "Causal Linking in Spoken and Written English", Studia Linguistica,38(1): 20-69, (1984).