An Automated Learner for Extracting New Ontology Relations

Amaal Saleh Hassan Al Hashimy
Sultan Qaboos University, College of Science, Computer Science Department
Oman
amaalh@squ.edu.om

Narayanan Kulathuramaiyer
University of Malaya Sarawak UNIMAS, Sarawak, Malaysia
nara@fit.unimas.my

***Abstract*— Recently, the NLP community has shown a renewed interest in automatic recognition of semantic relations between pairs of words in text which called lexical semantics. This approach to semantics is concerned with psychological facts associated with the meaning of words. Lexical semantics is an important task with many potential applications including but not limited to, Information Retrieval, Information Extraction, Text Summarization, and Language Modeling. As this task "automatic recognition of semantic relations between pairs of words in text" can be used in many NLP applications, its implementation are demanding and may include many potential methodologies. And as it includes semantic processing, the results produced still need enhancements and the outcome was always limited in terms of domain or coverage.**

In this research we developed a buffered system that handle the whole process of extracting causation relations in general domain ontologies. The main achievement of this work is the heavy analysis of statistical and semantic information of causation relation context to generate the learner. The system also builds relation resources that made it possible to learn from itself, were each time it runs the resources incremented with new relations information recording all the statistics of such relation, making its performance enhanced each time it runs. Also we present a novel approach of learning based on the best lexical patterns extracted, besides two new algorithms the CIA and PS that provide the final set of rules for mining causation to enrich ontologies.

Keywords-lexical relations, mining causation relations, ontology learning.

I. INTRODUCTION

The notion of ontology now a day is a dominant research area in the field of computer science. While this gain its main features as a "world representation scheme" from the philosophy in past, it is now gaining specific role in AI, computational linguistics, and DB theory.

Ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. It is used to reason about the entities within that domain, and may be used to describe the domain. In theory, ontology is an "explicit specification of a shared conceptualization"[11].

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 (figure 1). 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 [13].

The next sections will explore related works and provide insights on approaches in handling lexical semantics for the purpose of ontology learning. Subsequently we propose an enhanced framework for ontology learning.

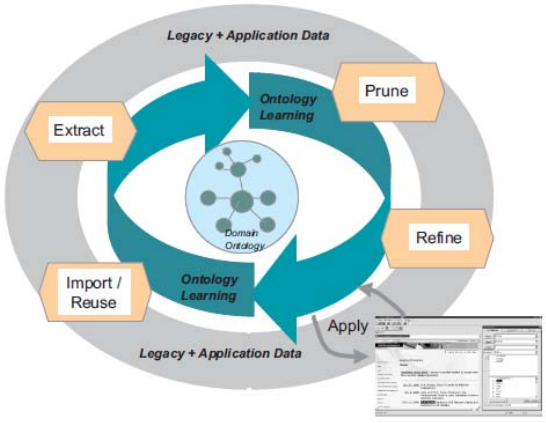

Figure1. Sample of ontology learning process
(Source: Maedche et.al. 2005)

II. RELATED WORK

Developing systems based on this kind of information is not new. Many systems have been developed Garcia in 1997 used verbs as causal indicators for causal knowledge acquisition in French. Khoo in [10] acquired causal knowledge with manually created syntactic patterns specifically for the MEDLINE text database. Works by Girju in [8] explored the acquisition of causal knowledge by using connective markers.

Ontologies wide-spread usage is still hindered by ontology acquisition being rather time consuming and, hence, expensive [3]. A number of proposals have been made to facilitate ontological acquisition through automatic discovery from

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.95

domain-specific natural language texts [2]. Nevertheless, most of these approaches have concentrated on 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, major efforts in ontology engineering are required to be dedicated to the definition of non-taxonomic conceptual relationships, such as Has_Part, Cause_Effect and Contain_Container relations. Despite this, the methods that address the non-taxonomic relations have also not come up with a state level in enhancing the process of classifying and extracting semantic lexical relations [3]. Most of the systems concentrate on just classifying relations without consideration for how it can be created. Works that provide methods for creating the relations have also not considered the context in which the relations can occur in [4].

Several approaches have then been proposed for covering the different phases involved in ontology mining, the phase of extraction of non-taxonomic relationships has been recognized as one of the most difficult and least explored problems [5]. Non-taxonomic discovery of 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 stages :

- 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 [13]. Moreover, even if the semantics is clear, it might still be difficult to guess which among several synonymous labels are preferred by a certain community for the task at hand [7].

III. OVERVIEW OF LEXICAL SEMANTICS

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.

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.
- Learning approaches using different learning algorithms like Generative models for semantic roles , Decision trees, Neural networks.
- Knowledge based methods that rely on the available many lexical resources like MRD, lexical ontology i.e. wordnet, framenet, and annotated corpuses.
- Hybrid approaches that make use a combination of the previously mentioned methods.

IV. DISCOVERING CAUSATION FROM TEXT

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

1. Using *causal links* to link (see classification below based on Altenburg [1]).
2. Using *causative verbs* .
3. Using *resultative* constructions like the pattern V-NP-ADJ).
4. Using *conditionals*, i.e. "if ... then ..." constructions.

We then adopt Altenberg [1] 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*.

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. As a fully automatic acquisition of knowledge by machines remains in the distant future, we consider the overall process of ontology learning as semi-automatic with human intervention in specific places e.g. to validate input samples.

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

As any NLP system, this project has made use of many of knowledge resources to support its flow of process and to improve the performance of its algorithms taking into consideration domain generality, relation relatedness, and confidence.

The framework will make use of variety knowledge sources which 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.

The system implemented into two main stages (with a number of models implemented in each stage) proposing new approaches in each one as follows

Stage 1, first, the system creates a certainty factor CF to evaluate the importance of the relation lexical patterns, so that the learner in the second stage will learn from the best patterns before the least. Second, the system creates relation DB that records the semantic information of the context where the relation found. This DB is a very useful tool in preserving the world around which the relation extracted. Also it records some statistical information that gives some validity for the learner later in the second stage.

Stage 2, the system use a novel approaches using the Conditional iterative abstraction algorithm (CIA) and the propagation schema (PS). The CIA provides the best semantically abstracted set of seen and un ambiguous examples

representative of causation relation in the possible shortest iteration. The PS uses the results of CIA and provides best set of rules to classify the unseen relation examples in any domain. The system incorporates the decision tree C5.0 to classify the correct rules for each relation pattern.

A. Stage one, causation patterns acquisition (figure 2)

The input knowledge sources are of different formats as mentioned before. So each one need a specific kind of preprocessing to derive 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), were she provided a general semantic cover set of features for cause and effect relations from SemEval2007. The approach is appropriate as the proposed set covers a good percentage of the SemEval2007 data set. The semantic cover set makes use of WN hierarchies that represent a class of word senses related in the hyponame chain. The cover set will include:

- 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 stage 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. Analyze the sentences guided by the causation general patterns to extract linguistics patterns for cause an effect from the resultant set of sentences.
6. Apply the certainty factor CF to the extracted patterns according to the level from which the pattern was extracted (were the highest CF is 4 which represent the best representative patterns of the relation, and the least CF is 0 assigned to the least representative patterns of the relation).
7. Create the relation DB which records the semantic information for each relation found in the different input resources. also record some statistics of the relation including its occurrences in the corpus.

The resultant patterns do not just specify causation relations in text; they also specify the direction of the relation indicating which the *cause* is and which the effect is. For example, the pattern

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

This DB builds a record for each relation detected in the input corpus and certified by our algorithm. The record will look like the following

DB_record(cause[lexical word, wn_sense #, hypernym #],effect[lexical word, wn_sense#, hypernym #], no_of _occurrence).

Relation DB

After identifying a sentence with causation relation and extracting the lexico-syntactic pattern, the cause and effect boundaries will be known. So a new relation extracted and a new record will be added to the DB. This DB can be used at different levels of learning as a guide for the learner in resolving ambiguous cases (explained in detail in chapter 5).

The most important feature in this DB is that it is not fixed, but dynamic (incremental). Each time the system works a new set of records will be added to increase its coverage of the relation.

B. Stage two, learning relation rules (figure 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 hypernym 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 conditional iterative abstraction algorithm (CIA) and the propagation schema (PS) 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 a number of categories according to the patterns components.

Each category of patterns will pass through the following steps:

- Select set of positive and negative examples for the sake of learning; this can be done by using the causation patterns mined in the first stage to extract sentences with and without causation relation from the corpus.
- Analyze the sentences to extract cause and effect.

- Abstract all the examples to the most general semantic features possible using hypernyme relations. The result will be two sets ambiguous and unambiguous
- Handle the ambiguous examples by using the CIA to get the best unambiguous set of semantic patterns from the seen examples.
- Use the resulting unambiguous examples as input to the PS to learn the new rules to identify the unseen relation rules.

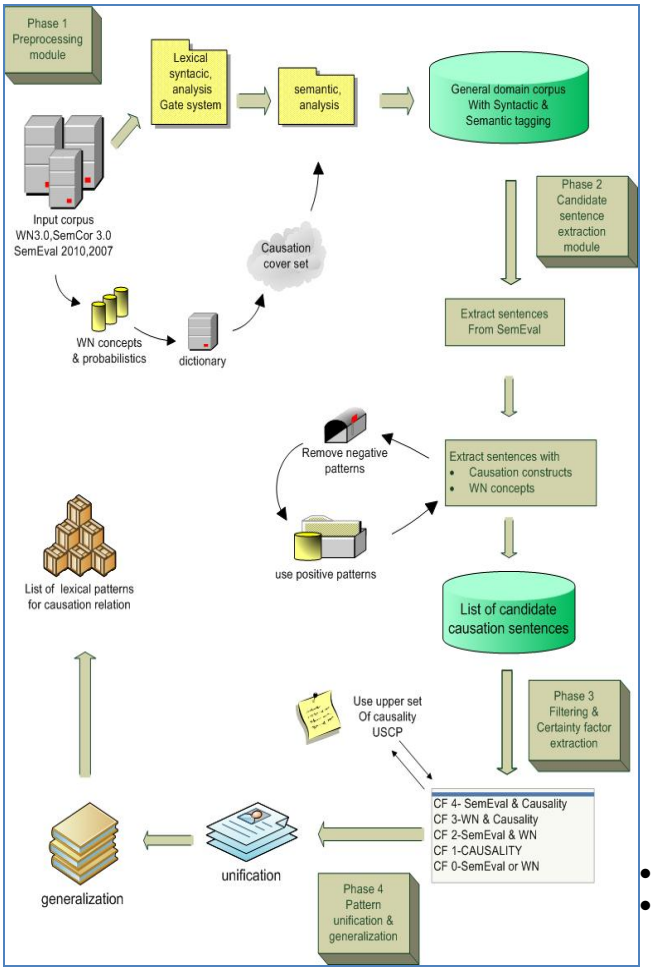

Figure. 2. Causation patterns acquisition

C. *The conditional iterative abstraction algorithm (CIA)*

For each category of patterns the system will start by abstracting each example to the highest class of semantic representation using the hypernym semantic relation in WN. The result will be set of ambiguous and un ambiguous examples.

To handle the ambiguous examples, the CIA algorithm will build a bottom-up tree from the leaves till the most abstract possible level of unambiguous semantic representation for the input ambiguous examples. The algorithm will work according to the certainty factor measure assigned to the patterns were each CF value require different processing. Also we will calculate the ambiguity degree GD in each level of abstraction while building the tree.

$$GD = \frac{\text{no of ambiguous examples in this level}}{\text{no of all examples in this level}}$$

Also the algorithm will calculate the distance D in case of low CF looking for more confident information before making a decision regarding the ambiguous example. As the shortest path indicates more valuable information, we will need it in some nodes in the bottom-up tree to resolve the ambiguity. It will go through from the leaf to the current node level trying to make a decision on the best match of the relation.

$$D = \frac{\text{shortest path in positive examples}}{\text{shortest path in negative examples}}$$

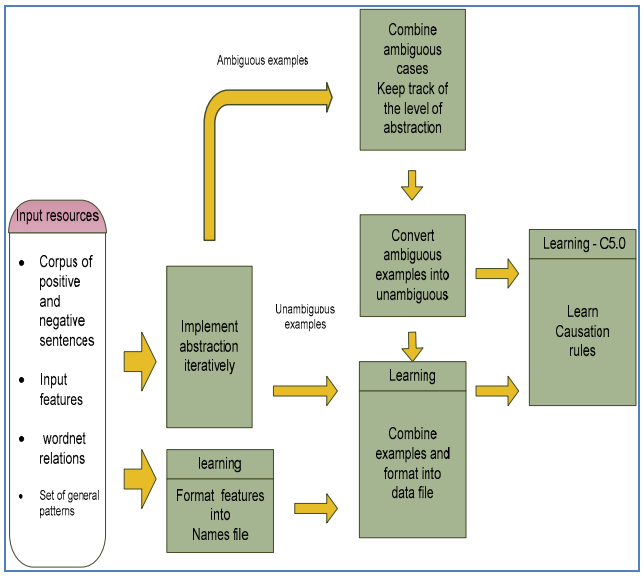

Figure. 3. Framework for learning rules

For each category of patterns perform the following, the result will be the list L of unambiguous examples:

- Abstract one level using hypernym relation.
- If ambiguous examples encountered then
  - begin
  - Calculate ambiguity degree GD of the level ($GD = \frac{\text{no of ambiguous examples in this level}}{\text{no of all examples in this level}}$)
  - If GD> T (T= threshold indicate unaccepted level of ambiguity) then
    1. ***Begin handling only ambiguous examples***
    2. If CF =3 then
       a. Begin
       b. If positive ex more than negative then assign relation to positive else

negative – ***add result examples to L***

c. end

3. If CF=2 then
   a. begin
   b. If negative ex more than positive then
      - If negative ex in DB then ***reject***

   c. If positive ex more than negative then
      - If positive ex synonyms in DB ***or*** If D<=1 ($D = \frac{\text{shortest path in positive examples}}{\text{shortest path in negative examples}}$) ***then*** assign relation to positive– ***add examples to L***

   d. If negative ex more than positive then relation negative – ***add result ex L***
   e. end
4. If CF= 1 then
   a. Begin
   b. If negative ex more than positive then
      - If negative ex in DB ***then reject***

   c. If positive ex more than negative then
      - If positive ex synonyms in DB then assign relation to positive – ***add result examples to L***

   d. If negative ex more than positive then
      - If D>1 ($D = \frac{\text{shortest path in positive examples}}{\text{shortest path in negative examples}}$) then assign relation to negative – ***add result ex L***

   e. End
5. If CF= 0 then
   a. Begin
   b. If negative ex more than positive then
      - If negative ex in DB ***then reject***.

   c. If positive ex more than negative then
      - If positive ex synonyms in DB ***and***
      - If D<=1($D = \frac{\text{shortest path in positive examples}}{\text{shortest path in negative examples}}$) then assign relation to positive– ***add ex to L***

   d. If negative ex more than positive then
      - If D>1 ($D = \frac{\text{shortest path in positive examples}}{\text{shortest path in negative examples}}$) then assign relation to negative – ***add ex to L***

   e. End
6. ***<u>end handling only ambiguous examples</u>***

- Add unambiguous examples to the list L.
- end

This algorithm will be implemented for each set of ambiguous examples results under the same top node. The result list L at each time will produce set of unambiguous examples that will be collected all together in a level table.

Then the propagation schema PS will implement the learning process using the C5.0 learning tool, and depending on the level table. 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.

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 a 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:

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

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

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

TABLE 1. 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 2. 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 [What is UI?], which was marginally better. 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. Discussion and Conclusions

In this research we focused on learning semantic relations patterns between word meanings by taking into consideration the surrounding context information in the general domain for discovering causation relations. We believe that extracting learning rules is much more effective than just discovering the relations, because the learned system can then be used to discover new relations not only the set dedicated to a specific corpus. The other advantage is that these patterns can further be used to generate new relations regardless of the domain, because mainly they characterize the syntactic and semantics of the context rather than on the specific meaning within the domain.

To validate our approach, we used as an input resources to learn causation patterns set Wordnet relations beside SemEval2010 training set. 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.

Our approach of evaluating the weight of different lexical syntactic patterns extracted for causation relations was very useful in the learning process later. As learning from the best lexical pattern first, will guide the learning process to more valuable results, and this will of course serve very well in provide a rich set of variety patterns for the causation relation.

Also our approach in building DB to preserve all the semantic and lexical information of the causation examples found in the corpus, was a good guide in judging the ambiguous cases and rejecting confusing patterns. And the semantic information recorded in it was useful tool in enforcing the same causation semantic constraints when the learner was unable to evaluate the patterns.

References

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

[2]. R. Byrd and Y. Ravin. "Identifying and extracting relations from text", NLDB'99 , 4th International Conference on Applications of NLto IS, 1999.

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

[4]. 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:, SemEval2010,2010.

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

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

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

[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]. 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.

[11]. Shamsfard, M., & Barforoush, A. A. The state of the art in ontology learning, A framework for comparison. Knowledge Engineering Review, 18(4), 293–316, 2003.

[12]. Shamsfard, M., & Barforoush, A. A. , Learning ontologies from natural language texts. International Journal of Human–Computer Studies, 60(1), 17–63, 2004.

[13]. Maedche, A. D. , Ontology learning for the semantic Web. Norwell, MA, USA:,Kluwer Academic Publishers, 2002.

[14]. George A. Miller, Communications of the ACM November 1995/Vol. 38, No. 11, 1995.