Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Framework To Determine Domain Specific Knowledge from Text

Niharika Jakka*
M. Tech 2nd year, Dept of CSE
AITS, Tirupati, India.

B. Rupa Devi
Assistant Professor, Dept of IT
AITS, Tirupati, India.

Abstract— In many knowledge intensive applications, it is necessary to have extensive domain-specific knowledge in addition to general-purpose knowledge bases. This paper presents a methodology for discovering domain-specific concepts and relationships in an attempt to extend WordNet. The method was tested on five seed concepts selected from the financial domain: interest rate, stock market, inflation, economic growth, and employment..
Keywords—

I. INTRODUCTION

The knowledge is infinite and no matter how large a knowledge base is, it is not possible to store all the concepts and procedures for all domains. Even if that were possible, the knowledge is generative and there are no guarantees that a system will have the latest information all the time. And yet, if we are to build common-sense knowledge processing systems in the future, it is necessary to have general-purpose and domain-specific knowledge that is up to date. Our inability to build large knowledge bases without much effort has impeded many ANLP developments. The most successful current Information Extraction systems rely on hand coded linguistic rules representing lexico-syntactic patterns capable of matching natural language expressions of events. Since the rules are hand-coded it is difficult to port systems across domains. Question answering, inference, summarization, and other applications can benefit from large linguistic knowledge bases. The basic idea A possible solution to the problem of rapid development of flexible knowledge bases is to design an automatic knowledge acquisition system that extracts knowledge from texts for the purpose of merging itwith a core ontological knowledge base. The attempt to create a knowledge base manually is time consuming and error prone, even for small application domains, and we believe that automatic knowledge acquisition and classification is the only viable solution to large scale ,knowledge intensive applications.

This paper presents an interactive method that acquires new concepts and connections associated with user-selected *seed* concepts, and adds them to the Word Net linguistic knowledge structure (Fellbaum1998). The sources of the new knowledge are texts acquired from the Internet or other corpora. At the present time, our system works in a semi-automatic mode, in the sense that it acquires concepts and relations automatically, but their validation is done by the user.

.II. RELATED WORK

This work was inspired in part by Marti Hearst's paper (Hearst 1998) where she discovers manually lexico-syntactic patterns for the HYPERNYMY relation in WordNet. Much of the work in pattern extraction from texts was done for improving the performance of Information Extraction systems. Research in this area was done by (Kim and Moldovan 1995) (Riloff 1996),(Soderland 1997) and others. The MindNet (Richardson 1998) project at Microsoft is an attempt to transform the Longman Dictionary of Contemporary English (LDOCE) into a form of knowledge base for text processing.

Woods studied knowledge representation and classification for long time (Woods 1991), and more recently is trying to automate the construction of taxonomies by extracting concepts directly from texts(Woods 1997).The Knowledge Acquisition from Text (KAT) system is presented next. It consists of four parts: (1)discovery of new concepts, (2) discovery of new lexical patterns, (3) discovery of new relationships reflected by the lexical patterns, and (4) the classification and integration of the knowledge discovered with a WordNet - like knowledge base.

2. KAT SYSTEMS

2.1 DISCOVER NEW CONCEPTS

Select seed concepts. New domain knowledge can be acquired around some seed concepts that a user considers important. In this paper we focus on the financial domain, and use: *interest rate, stock market, inflation, economic growth,* and *employment as* seed concepts. The knowledge we seek to acquire relates to one or more of these concepts, and consists of new concepts not defined in WordNet and new relations that link these concepts with other concepts, some of which are in WordNet.For example, from the sentence: *When the US economy enters a boom, mortgage interest rates rise,* the system discovers: (1) the new concept *mortgage interest rate* not defined in WordNet but related to the seed concept *interest rate,* and (2) the state of the *US economy* and the value of *mortgage interest rate are* in a DIRECT

RELATIONSHIP. In WordNet, a concept is represented as a synset that contains words sharing the same meaning. In our experiments, we extend the seed words to their corresponding syn set. For example, *stock market* is synonym with *stock exchange and securities market,* and we aim to learn concepts related to all these terms, not only to *stock market* .Extract sentences. Queries are formed with each seed concept to extract documents from the Internet and other possible sources. The documents retrieve dare further processed such that only the sentences that contain the seed concepts are retained. This way, an arbitrarily large corpus .4 is formed of sentences containing the seed concepts. We limit the size of this corpus to 1000 sentences per seed concept. Parse sentences. Each sentence in this corpus is first part-of-speech (POS) tagged then parsed. We use Brill's POS tagger and our own parser.

2.2 **DISCOVER LEXLCO-SYNTACTIC PATTERNS**

Texts represent a rich source of information from which in addition to concepts we can also discover relations between concepts. We are interested in discovering semantic relationships that link the concepts extracted above with other concepts, some of which may be in WordNet. The approach is to search for lexico-syntactic patterns comprising the concepts of interest. The semantic relations from WordNet are the first we search for, as it is only natural to add more of these relations to enhance the WordNet knowledge base. However, since the focus is on the acquisition of domain-specific knowledge, there are semantic relations between concepts other than the WordNet relations that are important. These new relations can be discovered automatically from the clauses and sentences in which the seeds occur.

2.3 **KNOWLEDGE CLASSIFICATION AND INTEGRATION**

Next, a taxonomy needs to be created that is consistent with WordNet. In addition to creating a taxonomy, this step is also useful for validating the concepts acquired above. The classification is based on the *sub sumption* principle (Schmolze and Lipkis1983), (Woods 1991).This algorithm provides the overall steps for the classification of concepts within the context of Word-Net. Figure 1 shows the inputs of the Classification Algorithm and suggests that the classification is an iterative process. In addition to WordNet, the inputs consist of the corpus ,4, the sets of concepts Cs and Cn, and the relationships 7~. Let's denote with C = Cs U Cn the union of the seed related concepts with the new concepts. All these concepts need to be classified.

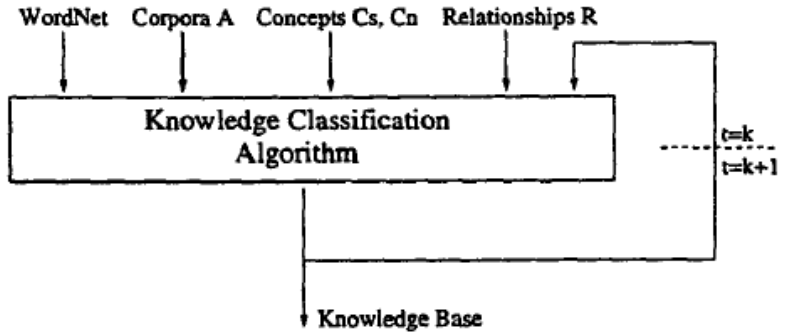


Figure 1: The knowledge classification diagram

Step 1. From the set of relationships 7"~ discovered in Part 3, pick all the HYPERNYMY relations. From the way these relations were developed, there are two possibilities:(1) A HYPERNYMY relation links a WordNet concept Cw with another concept from the set C denoted with *CAw* , or(2) A HYPERNYMY relation links a concept Cs with a concept Cn. Concepts C~w are immediately linked to Word-Net and added to the knowledge base. The concepts from case (2) are also added to the knowledge base but they form at this point only some isolated is lands since are not yet linked to the rest of the knowledgebase.

Step 2. Search corpus `4 for all the patterns associated with the HYPERNYMY relation that may link

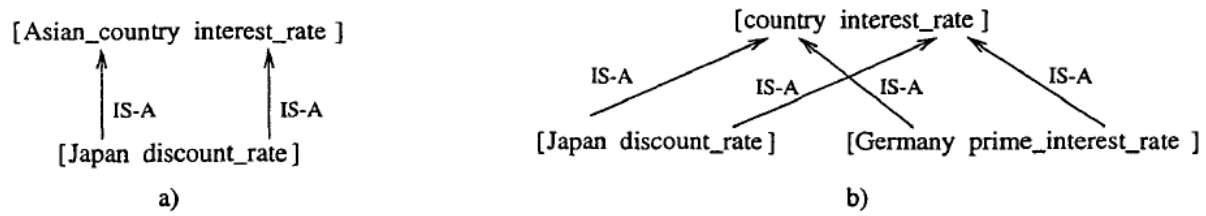


Figure 2: Relative classification of two concepts

concepts in the set Cn with any WordNet concepts. Altough concepts C, are not seed-based concepts ,they are related to at least one Cs concept via a relationship(as found in Task 3). Here we seek to find HYPERNYMY links between them and WordNet concepts. If such C,~ concepts exist, denote them with C~w. The union Chw = C~w LJ C2w represents all concepts from the set C that are linked to WordNet without any further effort. We focus now on the rest of concepts, Cc - - C N Chw, that are not yet linked to any WordNet concepts.

Step 3. :
Classify all concepts in set Ce using Procedures4.1 through 4.5 below.

Step 4.:
Repeat Step 3 for all the concepts in set Cc several times till no more changes occur. This reclassification is necessary since the insertion of a concept into the knowledge base may perturb the ordering of other surrounding concepts in the hierarchy.

Step 5.
Add the rest of relationships 7~ other
than the HYPERNYMY to the new knowledge base. The HYPERNYMY relations have already been used in the Classification Algorithm, but the other relations ,i.e. INFLUENCE, CAUSE and EQUIVALENT need to be added to the knowledge base.

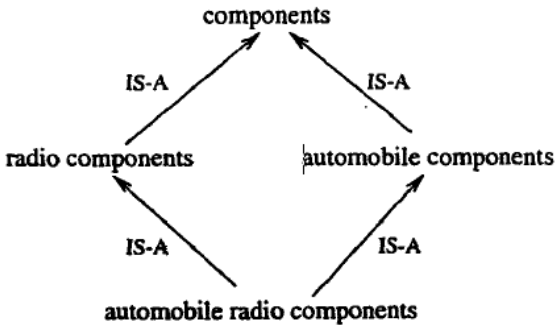


Figure 3: Classification of a compound concept with respect to its HYPERNYM concepts

Procedure : Merge a structure *of concepts with* the rest *of* the *knowledge base*
It is possible that structures consisting of several inter-connected concepts are formed in isolation of the main knowledge base as a result of some procedures. The task here is to merge such structures with the main knowledge base such that the new knowledgebase will be consistent with both the structure and the main knowledge base. This is done by

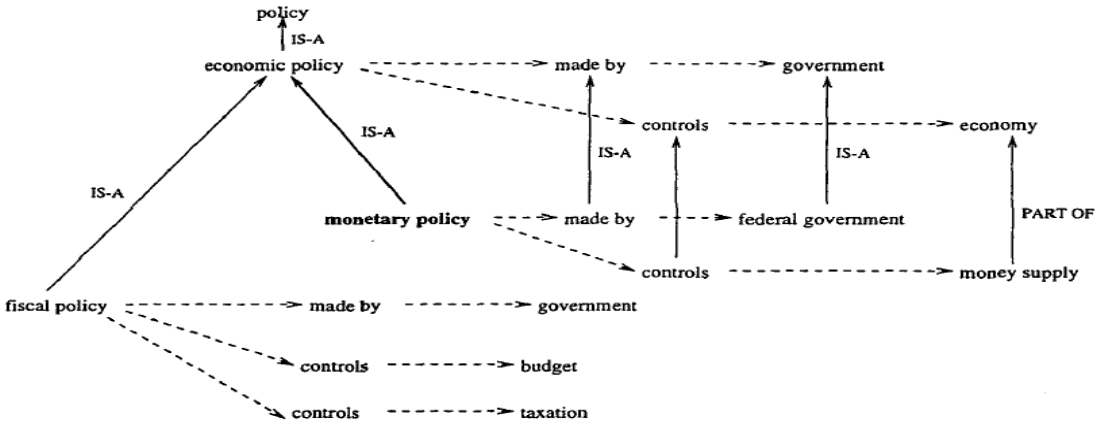
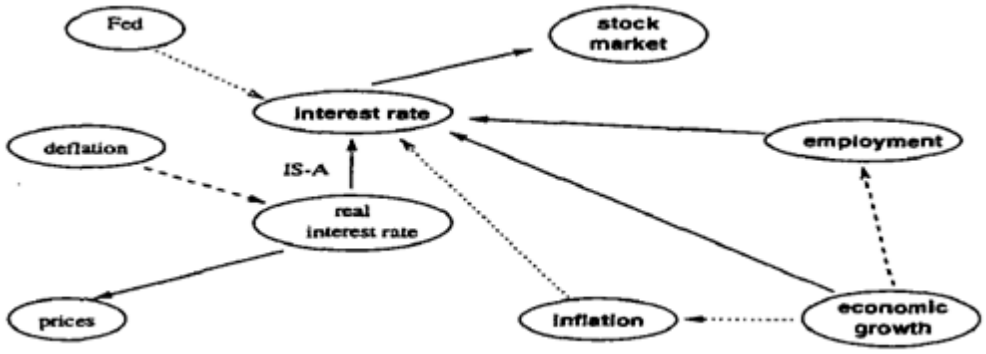


Figure 4: Classification of the new concept *monetary policy*

III. APPLICATIONS

An application in need of domain-specific knowledge is Question Answering. The concepts and the relationshipsacquired can be useful in answering difficult questions that normally cannot be easily answered just by using the information from WordNet. Consider the processing of the following questions after the new domain knowledge has been acquired:

QI: What factors have an impact on the *interstate?*
Q2: What happens with the *employment* when the *economic growth* rises?
Q3: How does *deflation* influence *prices?*



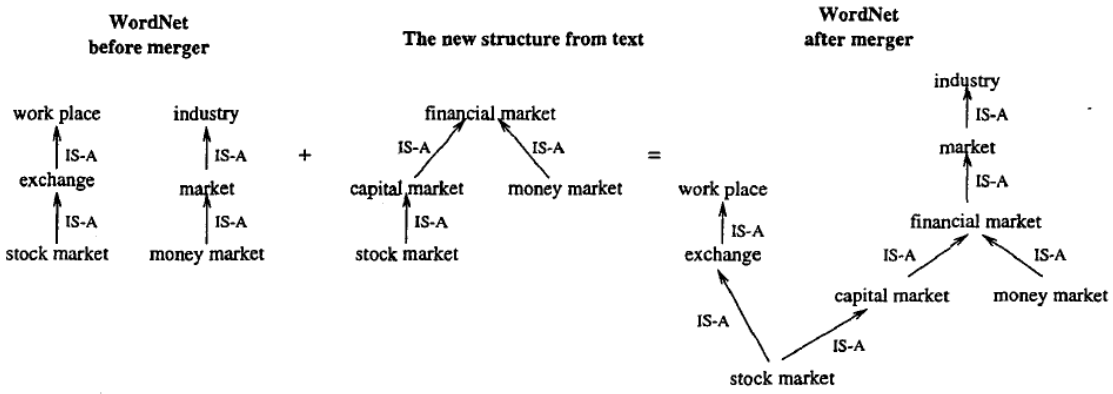


Figure 5: Merging a structure of concepts with WordNet

Figure 6: A **sample of concepts and relations acquired from the 5000 sentence corpus. Legend: continue lines represent *influence inverse proportionally,* dashed lines represent *influence direct** proportionally,* **and dotted lines represent *influence* (the direction** of the relationship was not specified in the text).

The last two questions ask for more detailed information about the complex relationship among these concepts. Following the path from the *deflation* concept up to *prices,* the system learns that *deflation* influences direct proportionally *real interest rate,* and *real interest rate* has an inverse proportional impacton *prices.* Both these relationships came from the sentence: *Thus, the deflation and the real interest rate are positively correlated, and so a higher real interest rate leads to falling prices.*

This method may be adapted to acquire information when the question concepts are not in the knowledge base. Procedures may be invoked to discover these concepts and the relations in which they may be used.

IV.CONCLUSIONS

The knowledge acquisition technology described above is applicable to any domain, by simply selecting appropriate seed concepts. We started with five concepts *interest rate, stock market, inflation, economic growth,* and *employment* and from a corpus of 5000 sentences we acquired a total of 362 concepts which 319 contain the seeds and 43 relate to these via selected relations. There were 22 distinctle:dco-syntactic patterns discovered used in 63instances. Most importantly, the new concepts can be integrated with an existing ontology. The method works in an interactive mode where the user accepts or declines concepts, patterns and relationships. The manual operation took on average40 minutes per seed for the 5000 sentence corpus. KAT is useful considering that most of the knowledgebase construction today is done manually. Complex linguistic phenomena such as co reference resolution, word sense disambiguation, and others have to be dealt with in order to increase the automation of the knowledge acquisition system. Without a good handling of these problems the results are not always accurate and human intervention is necessary..

REFERENCES

1. Christiane Fellbaum. WordNet - An Electronic LezicalDatabase, MIT Press, Cambridge, MA, 1998.
2. Marti Hearst. Automated Discovery of WordNet Relations.In WordNet: An Electronic Lezical Databaseand Some of its Applications, editor Fellbaum, C.,MIT Press, Cambridge, MA, 1998.

3. J. Kim and D. Moldovan. Acquisition of LinguisticPatterns for knowledge-based information extraction. IEEE Transactions on Knowledge and Data Engineering7(5): pages 713-724.
4. R. MacGregor. A Description Classifier for the PredicateCalculus. Proceedings of the 12th National Conference on Artificial Intelligence (AAAI94), pp. 213-220, 1994.
5. Stephen D. Richardson, William B. Dolan, Lucy Vanderwande.MindNet: acquiring and structuring semantic information from text. Proceedings of ACL-Coling1998, pages 1098-1102.
6. Ellen Riloff. Automatically'Generating Extraction Patternsfrom Untagged Text. In Proceedings of the Thirteenth National Conference on Artificial Intelligence,1044-1049. The AAAI Press/MIT Press.
7. J.G. Schmolze and T. Lipkis. Classification in the KLONEknowledge representation system. Proceedings of 8th Int'l Joint Conference on Artificial Intelligence(IJCAI83), 1983.
8. S. Soderland. Learning to extract text-based informationfrom the world wide web. In the Proceedings ofthe Third International Conference on Knowledge Discover#and Data Mining DD-97).Text REtrieval Conference. http://trec.nist.gov 1999
9. W.A. Woods. Understanding Subsumption and Taxonomy:A Framework for Progress. In the Principles of Semantic Networks: Explorations in the Representationof Knowledge, Morgan Kaufmann, San Mateo,Calif. 1991, pages 45-94.W.A. Woods.
10. A Better way to Organize Knowledge.Technical Report of Sun Microsystems Inc., 1997.
11. Yi-ming G., and Zhi-jun W.: A vertical format algorithm for mining frequent itemsets. In IEEE transactions, pp. 11-13 (2010)
12. Zaki M. J.: Parallel and distributed association mining: A survey. In IEEE concurrency, pp. 14-25 (1999)
13. Ruggieri S.: Frequent Regular Itemset Mining. In: ACM KDD (2010)
14. Zaki M.J., Gouda K.: Fast Vertival Mining using Diffsets. In. ACM SIGKDD'03 (2003)

