Domain-Specific Knowledge Acquisition from Text

Dan Moldovan, Roxana Girju and Vasile Rus
Department of Computer Science and Engineering
University of Southern Methodist University
Dallas, Texas, 75275-0122
{*moldovan, roxana, rus*}*@seas.smu.edu*

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*.

1 Desiderata for Automated Knowledge Acquisition

The need for knowledge

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 it with 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 WordNet linguistic knowledge structure (Fellbaum 1998). 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.

We believe that domain knowledge should not be acquired in a vacuum; it should expand an existent ontology with a skeletal structure built on consistent and acceptable principles. The method presented in this paper is applicable to any Machine Readable Dictionary. However, we chose WordNet because it is freely available and widely used.

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 System

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 synset. 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 retrieved are further processed such that only the sentences that contain the seed concepts are retained. This way, an arbitrarily large corpus $\mathcal{A}$ 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. The output of the POS tagger for the example above is:
When/WRB *the*/DT *U.S.*/NNP *economy*/NN *enters*/VBZ *a*/DT *boom*/NN ,/, *mortgage*/NN *interest_rates*/NNS *rise*/VBP ./.

The syntactic parser output is:
TOP (S (SBAR (WHADVP (WRB *When*)) (S (NP (DT *the*) (NNP *U.S.*) (NN *economy*)) (VP (VBZ *enters*) (NP (DT *a*) (NN *boom*) (, ,))))) (NP (NN *mortgage*) (NNS *interest_rates*)) (VP (VBP *rise*)))

Extract new concepts. In this paper only noun concepts are considered. Since, most likely, one-word nouns are already defined in WordNet, the focus here is on *compound nouns* and *nouns with modifiers* that have meaning but are not in WordNet.

The new concepts directly related to the seeds are extracted from the noun phrases (NPs) that contain the seeds. In the example above, we see that the seed belongs to the NP: *mortgage interest rate.*

This way, a list of NPs containing the seeds is assembled automatically from the parsed texts. Every such NP is considered a potential new concept. This is only the "raw material" from which actual concepts are discovered.

In some noun phrases the seed is the head noun, i.e. [*word, word,..seed*], where *word* can be a noun or an adjective. For example, [*interest rate*] is in WordNet, but [*short term nominal interest rate*] is not in WordNet. Most of the new concepts related to a seed are generated this way. In other cases the seed is not the head noun i.e. [*word, word,..seed, word, word*]. For example [*interest rate peg*], or [*international interest rate differential*].

The following procedures are used to discover concepts, and are applicable in both cases:

Procedure 1.1. *WordNet reduction.* Search NP for words collocations that are defined in WordNet as concepts. Thus [*long term interest rate*] becomes [*long_term interest_rate*], [*prime interest rate*] becomes [*prime_interest_rate*], as all hyphenated concepts are in WordNet.

Procedure 1.2. *Dictionary reduction.* For each NP, search further in other on-line dictionaries for more compound concepts, and if found, hyphenate the words. Many domain-specific dictionaries are available on-line. For example, [*mortgage interest_rate*] becomes [*mortgage_interest_rate*], since it is defined in the on-line dictionary *OneLook Dictionaries* (http://www.onelook.com).

Procedure 1.3. *User validation.* Since currently we lack a formal definition of a concept, it is not possible to completely automate the discovery of concepts. The human inspects the list of noun phrases and decides whether to accept or decline each concept.

2.2 Discover lexico-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.

Pick a semantic relation R. These can be WordNet semantic relations or any other relations defined by the user. So far, we have experimented with the WordNet HYPERNYMY (or so-called IS-A) relation, and three other relations. By inspecting a few sentences containing *interest rate* one can notice that INFLUENCE is a frequently used relation. The two other relations are CAUSE and EQUIVALENT.

Pick a pair of concepts C_i, C_j among which R holds. These may be any noun concepts. In the context of finance domain, some examples of concepts linked by the INFLUENCE relation are:
interest rate INFLUENCES *earnings*, or
credit worthiness INFLUENCES *interest rate*.

Extract lexico-syntactic patterns C_i $\mathcal{P}$ C_j. Search any corpus $\mathcal{B}$, different from $\mathcal{A}$ for all instances where C_i and C_j occur in the same sentence. Extract the lexico-syntactic patterns that link the two concepts. For example, from the sentence : *The graph indicates the impact on earnings from several different interest rate scenarios*, the generally applicable pattern extracted is:
impact on NP2 from NP1
This pattern corresponds unambiguously to the relation R we started with, namely INFLUENCE. Thus we conclude: INFLUENCE(NP1, NP2).

Another example is: *As the credit worthiness decreases, the interest rate increases.* From this sentence we extract another lexical pattern that expresses the INFLUENCE relation:
[as NP1 vb1, NP2 vb2] & [vb1 and vb2 are antonyms]
This pattern is rather complex since it contains not only the lexical part but also the verb condition that needs to be satisfied.

This procedure repeats for all relations R.

2.3 Discover new relationships between concepts

Let us denote with C_s the seed-related concepts found with Procedures 1.1 through 1.3. We search now corpus $\mathcal{A}$ for the occurrence of patterns $\mathcal{P}$ discovered above such that one of their two concepts is a concept C_s.

Search corpus $\mathcal{A}$ for a pattern $\mathcal{P}$. Using a lexico-syntactic pattern $\mathcal{P}$, one at a time, search corpus $\mathcal{A}$ for its occurrence. If found, search further whether or not one of the NPs is a seed-related concept C_s.

Identify new concepts C_n. Part of the pattern $\mathcal{P}$ are two noun phrases, one of which is C_s. The head noun from the other noun phrase is a concept C_n we are looking for. This may be a WordNet concept, and if it is not it will be added to the list of concepts discovered.

Form relation $R(C_s, C_n)$. Since each pattern $\mathcal{P}$ is a linguistic expression of its corresponding semantic relation R, we conclude $R(C_s, C_n)$ (this is interpreted "C_s is relation R C_n)"). These steps are repeated for all patterns.

User intervention to accept or reject relationships is necessary mainly due to our system inability of handling coreference resolution and other complex linguistic phenomena.

2.4 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 *subsumption* principle (Schmolze and Lipkis 1983), (Woods 1991).

This algorithm provides the overall steps for the classification of concepts within the context of WordNet. 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 $\mathcal{A}$, the sets of concepts $\mathbf{C_s}$ and $\mathbf{C_n}$, and the relationships $\mathcal{R}$. Let's denote with $\mathbf{C} = \mathbf{C_s} \cup \mathbf{C_n}$ the union of the seed related concepts with the new concepts. All these concepts need to be classified.

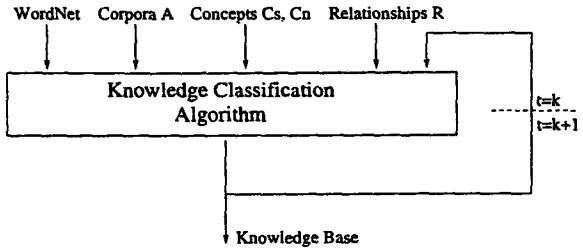

Figure 1: The knowledge classification diagram

Step 1. From the set of relationships $\mathcal{R}$ 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 C_w with another concept from the set $\mathbf{C}$ denoted with C^1_{hw}, or
(2) A HYPERNYMY relation links a concept $\mathbf{C_s}$ with a concept $\mathbf{C_n}$.

Concepts C^1_{hw} are immediately linked to WordNet 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 islands since are not yet linked to the rest of the knowledge base.

Step 2. Search corpus $\mathcal{A}$ for all the patterns associated with the HYPERNYMY relation that may link

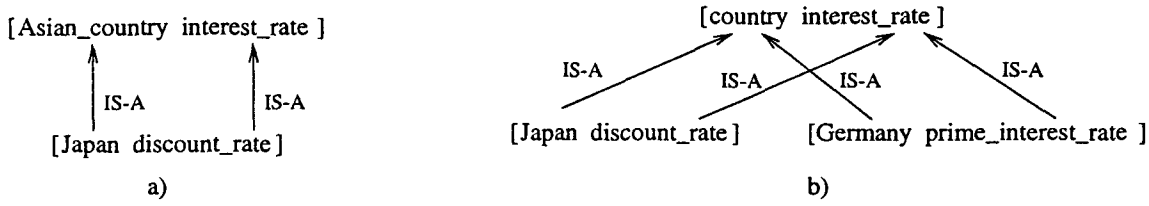

Figure 2: Relative classification of two concepts

concepts in the set $\mathbf{C}_n$ with any WordNet concepts. Altough concepts $\mathbf{C}_n$ are not seed-based concepts, they are related to at least one $\mathbf{C}_s$ concept via a relationship (as found in Task 3). Here we seek to find HYPERNYMY links between them and WordNet concepts. If such $\mathbf{C}_n$ concepts exist, denote them with $\mathbf{C}^2_{hw}$. The union $\mathbf{C}_{hw} = \mathbf{C}^1_{hw} \cup \mathbf{C}^2_{hw}$ represents all concepts from the set $\mathbf{C}$ that are linked to WordNet without any further effort. We focus now on the rest of concepts, $\mathbf{C}_c = \mathbf{C} \cap \overline{\mathbf{C}}_{hw}$, that are not yet linked to any WordNet concepts.

Step 3. Classify all concepts in set $\mathbf{C}_c$ using Procedures 4.1 through 4.5 below.

Step 4. Repeat Step 3 for all the concepts in set $\mathbf{C}_c$ 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 $\mathcal{R}$ 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.

Concept classification procedures

Procedure 4.1. *Classify a concept of the form [word, head] with respect to concept [head].*
It is assumed here that the [head] concept exists in WordNet simply because in many instances the "head" is the "seed" concept, and because frequently the head is a single word common noun usually defined in WordNet. In this procedure we consider only those head nouns that do not have any hyponyms since the other case when the head has other concepts under it is more complex and is treated by Procedure 4.4. Here "word" is a noun or an adjective.

The classification is based on the simple idea that a compound concept [word, head] is ontologically subsumed by concept [head]. For example, *mortgage_interest_rate* is a kind of *interest_rate*, thus linked by a relation HYPERNYMY(interest_rate, mortgage_interest_rate).

Procedure 4.2. *Classify a concept [$word_1$, $head_1$] with respect to another concept [$word_2$, $head_2$].*
For a relative classification of two such concepts, the ontological relations between $head_1$ and $head_2$ and between $word_1$ and $word_2$, if exist, are extended to the two concepts. We distinguish here three possibilities:

1. $head_1$ subsumes $head_2$ and $word_1$ subsumes $word_2$. In this case [$word_1$, $head_1$] subsumes [$word_2$, $head_2$]. The subsumption may not always be a direct connection; sometimes it may consist of a chain of subsumption relations since subsumption is (usually) a transitive relation (Woods 1991). An example is shown in Figure 2a; in WordNet, *Asian_country* subsumes *Japan* and *interest_rate* subsumes *discount_rate*. A particular case of this is when $head_1$ is identical with $head_2$.
2. Another case is when there is no direct subsumption relation in WordNet between $word_1$ and $word_2$, and/or $head_1$ and $head_2$, but there are a common subsuming concepts, for each pair. When such concepts are found, pick the most specific common subsumer (MSCS) concepts of $word_1$ and $word_2$, and of $head_1$ and $head_2$, respectively. Then form a concept [MSCS($word_1$, $word_2$), MSCS($head_1$, $head_2$)] and place [$word_1$ $head_1$] and [$word_2$ $head_2$] under it. This is exemplified in Figure 2b. In WordNet, *country* subsumes *Japan* and *Germany*, and *interest_rate* subsumes *discount_rate* and *prime_interest_rate*.
3. In all other cases, no subsumption relation is established between the two concepts. For example, we cannot say whether *Asian_country discount_rate* is more or less abstract then *Japan interest_rate*.

Procedure 4.3. *Classify concept [$word_1$ $word_2$ head].*
Several possibilities exist:

1. When there is already a concept [$word_2$ head] in the knowledge base under the [head], then place [$word_1$ $word_2$ head] under concept [$word_2$ head].
2. When there is already a concept [$word_1$ head] in the knowledge base under the [head], then place [$word_1$ $word_2$ head] under concept [$word_1$ head].
3. When both cases 1 and 2 are true then place [$word_1$ $word_2$ head] under both concepts.

4. When neither [word$_1$ head] nor [word$_2$ head] are in the knowledge base, then place [word$_1$ word$_2$ head] under the [head]. The example in Figure 3 corresponds to case 3.

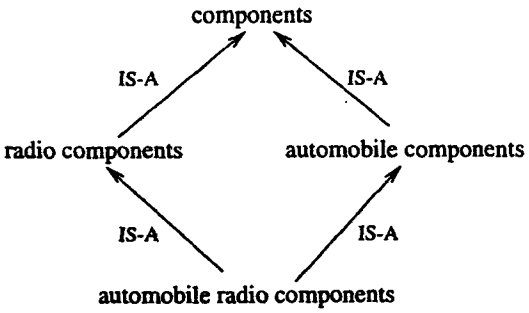

Figure 3: Classification of a compound concept with respect to its HYPERNYM concepts

Since we do not deal here with the sentence semantics, it is not possible to completely determine the meaning of [word$_1$ word$_2$ head], as it may be either [((word$_1$ word$_2$) head)] or [(word$_1$ (word$_2$ head))] often depending on the sentence context.

In the example of Figure 3 there is only one meaning, i.e. [*(automobile radio) components*]. However, in the case of [*performance skiing equipment*] there are two valid interpretations, namely [*(performance skiing) equipment*] and [*performance (skiing equipment)*].

Procedure 4.4 *Classify a concept [word$_1$, head] with respect to a concept hierarchy under the [head].*
The task here is to identify the most specific subsumer (MSS) from all the concepts under the head that subsumes [word$_1$, head]. By default, [word$_1$ head] is placed under [head], however, since it may be more specific than other hyponyms of [head], a more complex classification analysis needs to be implemented.

In the previous work on knowledge classification it was assumed that the concepts were accompanied by rolesets and values (Schmolze and Lipkis 1983), (Woods 1991), and others. Knowledge classifiers are part of almost any knowledge representation system.

However, the problem we face here is more difficult. While in build-by-hand knowledge representation systems, the relations and values defining concepts are readily available, here we have to extract them from text. Fortunately, one can take advantage of the glossary definitions that are associated with concepts in WordNet and other dictionaries. One approach is to identify a set of semantic relations into which the verbs used in the gloss definitions are mapped into for the purpose of working with a manageable set of relations that may describe the concepts restrictions. In WordNet these basic relations are already identified and it is easy to map every verb into such a semantic relation.

As far as the newly discovered concepts are concerned, their defining relations need to be retrieved from texts. Human assistance is required, at least for now, to pinpoint the most characteristic relations that define a concept.

Below is a two step algorithm that we envision for the relative classification of two concepts A and B. Let's us denote with AR_aC_a and BR_bC_b the relationships that define concepts A and B respectively. These are similar to rolesets and values.

1. Extract relations (denoted by verbs) between concept and other gloss concepts.
   $AR_{a1}C_{a1}$ $BR_{b1}C_{b1}$
   $AR_{a2}C_{a2}$ $BR_{b2}C_{b2}$
   ...
   $AR_{am}C_{am}$ $BR_{bn}C_{bn}$
2. A subsumes B if and only if
   (a) Relations R_{ai} subsume R_{bi}, for $1 \leq i \leq m$.
   (b) C_{ai} subsumes or is a meronym of C_{bi}.
   (c) Concept B has more relations than concept A, i.e. $m \leq n$.

Example: In Figure 4 it is shown the classification of concept *monetary policy* that has been discovered. By default this concept is placed under *policy*. However in WordNet there is a hierarchy *fiscal policy* - IS-A - *economic policy* - IS-A - *policy*. The question is where exactly to place *monetary policy* in this hierarchy.

The gloss of *economic policy* indicates that it is MADE BY *Government*, and that it CONTROLS *economic growth* - (here we simplified the explanation and used *economy* instead of *economic growth*). The gloss of *fiscal policy* leads to relations MADE BY *Government*, CONTROLS *budget*, and CONTROLS *taxation*. The concept *money supply* was found by Procedure 1.2 in several dictionaries, and its dictionary definition leads to relations MADE BY *Federal Government*, and CONTROLS *money supply*. In WordNet *Government* subsumes *Federal Government*, and *economy* HAS PART *money*. All necessary conditions are satisfied for *economic policy* to subsume *monetary policy*. However, *fiscal policy* does not subsume *monetary policy* since *monetary policy* does not control *budget* or *taxation*, or any of their hyponyms.

Procedure 4.5 *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 knowledge base will be consistent with both the structure and the main knowledge base. This is done by

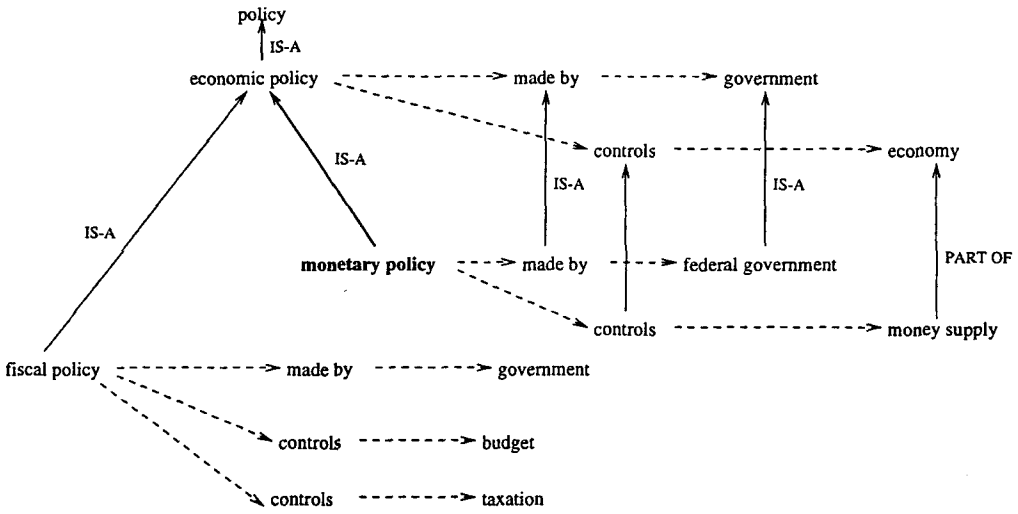

Figure 4: Classification of the new concept *monetary policy*

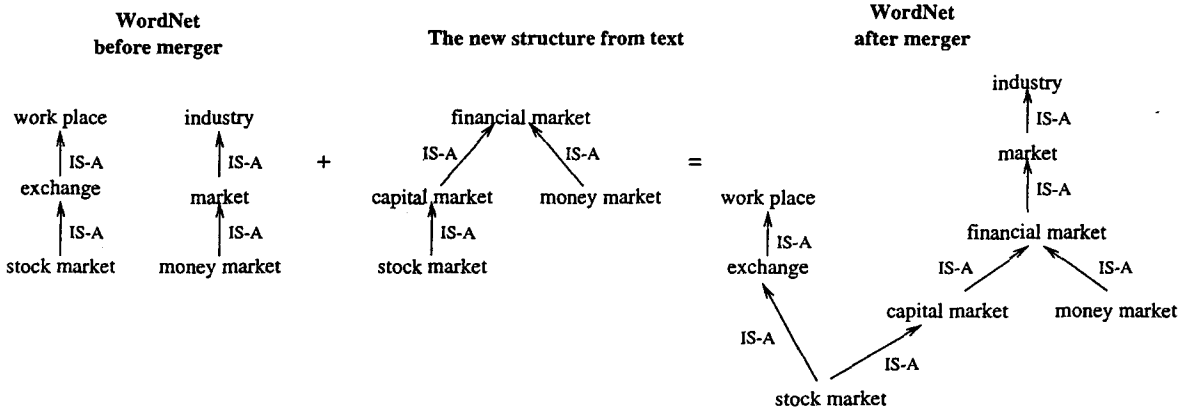

Figure 5: Merging a structure of concepts with WordNet

bridging whenever possible the structure concepts and the main knowledge base concepts. It is possible that as a result of this merging procedure, some HYPERNYMY relations either from the structure or the main knowledge base will be destroyed to keep the consistency. An example is shown in Figure 5.

Example : The following HYPERNYMY relationships were discovered in Part 3:
HYPERNYMY(financial market,capital market)
HYPERNYMY(financial market,money market)
HYPERNYMY(capital market,stock market)
The structure obtained from these relationships along with a part of WordNet hierarchy is shown in Figure 5. An attempt is made to merge the new structure with WordNet. To these relations it corresponds a structure as shown in Figure 5. An attempt is made to merge this structure with WordNet. Searching WordNet for all concepts in the structure we find *money market* and *stock market* in WordNet where as *capital market* and *financial market* are not. Figure 5 shows how the structure merges with WordNet and moreover how concepts that were unrelated in WordNet (i.e. *stock market* and *money market*) become connected through the new structure. It is also interesting to notice that the IS-A link in WordNet from *money market* to *market* is interrupted by the insertion of *financial market* in-between them.

3 Implementation and Results

The KAT Algorithm has been implemented, and when given some seed concepts, it produces new concepts, patterns and relationships between concepts in an interactive mode. Table 1 shows the number of concepts extracted from a 5000 sentence corpus, in which each sentence contains at least one of the five seed concepts.

The NPs were automatically searched in WordNet and other on-line dictionaries. There were 3745 distinct noun phrases of interest extracted; the rest contained only the seeds or repetitions. Most of the

Relations	Lexico-syntactic Patterns	Examples
	WordNet Relations	
HYPERNYMY	NP1 [<be>] a kind of NP2 ⇒ HYPERNYMY(NP1,NP2)	Thus, *LIBOR* is a kind of *interest rate*, as it is charged on deposits between banks in the Eurodolar market.
	New Relations	
CAUSE	NP1 [<be>] cause NP2 ⇒ CAUSE(NP1,NP2)	Phillips, a British economist, stated in 1958 that high *inflation* causes low *unemployment rates*.
INFLUENCE	NP1 impact on NP2 ⇒ INFLUENCE(NP1,NP2)	The Bank of Israel governor said that the *tight economic policy* would have an immediate impact on *inflation* this year.
	As NP1 vb, so <do> NP2 ⇒ INFLUENCE(NP1,NP2)	As the *economy* picks up steam, so does *inflation*.
	NP1 <be> associated with NP2 ⇒ INFLUENCE(NP1,NP2) INFLUENCE(NP2,NP1)	Higher *interest rates* are normally associated with weaker *bond markets*.
	As/if/when NP1 vb1, NP2 vb2. + vb1, vb2 = antonyms / go in opposite directions ⇒ INFLUENCE(NP1,NP2)	On the other hand, if *interest rates* go down, *bonds* go up, and your bond becomes more valuable.
	the effect(s) of NP1 on/upon NP2 ⇒ INFLUENCE(NP1,NP2)	The effects of *inflation* on *debtors* and *creditors* varies as the actual inflation is compared to the expected one.
	inverse relationship between NP1 and NP2 ⇒ INFLUENCE(NP1,NP2) ⇒ INFLUENCE(NP2,NP1)	There exists an inverse relationship between *unemployment rates* and *inflation*, best illustrated by the Phillips Curve.
	NP2 <be> function of NP1 ⇒ INFLUENCE(NP1,NP2)	*Irish employment* is also largely a function of the past high *birth rate*.
	NP1 (and thus NP2) ⇒ INFLUENCE(NP1,NP2)	We believe that the *Treasury bonds* (and thus *interest rates*) are in a downward cycle.

Table 2: Examples of lexico-syntactic patterns and semantic relations derived from the 5000 sentence corpus

		a	b	c	d	e
Total potential concepts (NPs)		773	382	833	921	836
Total concepts extracted with Procedure1						
Concepts found in WordNet		2	0	1	0	2
Concepts found in on-line dictionaries, but not in WordNet	Concepts with seed head	6	0	3	0	0
	Concepts with seed not head	7	0	1	1	1
Concepts accepted by human		78	62	58	60	37

Table 1: Results showing the number of new concepts learned from the corpus related to (a) *interest rate*, (b) *stock market*, (c) *inflation*, (d) *economic growth*, and (e) *employment*.

processing in Part 1 is taken by the parser. The human intervention to accept or decline concepts takes about 4 min./seed.

The next step was to search for lexico-syntactic patterns. We considered one WordNet semantic relation, HYPERNYMY and three other relations that we found relevant for the domain, namely INFLUENCE, CAUSE and EQUIVALENT. For each relation, a pair of related words was selected and searched for on the Internet. The first 500 sentences/relation were retained. A human selected and validated semi-automatically the patterns for each sentence. A sample of the results is shown in Table 2. A total of 22 patterns were obtained and their selection and validation took approximately 35 minutes/relation.

Next, the patterns are searched for on the 5000 sentence corpus (Part 3). The procedure provided a total of 43 new concepts and 166 relationships in which at least one of the seeds occurred. From these relationships, by inspection, we have accepted 63 and rejected 102, procedure which took about 7 minutes. Table 3 lists some of the 63 relationships discovered.

Relationships Examples
HYPERNYMY(interest rate, LIBOR) HYPERNYMY(leading stock market, New York Stock Exchange) HYPERNYMY(market risks, interest rate risk) HYPERNYMY(capital markets, stock markets)
CAUSE(inflation, unemployment) CAUSE(labour shortage, wage inflation) CAUSE(excessive demand, inflation)
INFLUENCE_DIRECT_PROPORTIONALY(economy, inflation) INFLUENCE_DIRECT_PROPORTIONALY(settlements, interest rate) INFLUENCE_DIRECT_PROPORTIONALY(U.S. interest rates, dollars) INFLUENCE_DIRECT_PROPORTIONALY(oil prices, inflation) INFLUENCE_DIRECT_PROPORTIONALY(inflation, nominal interest rates) INFLUENCE_DIRECT_PROPORTIONALY(deflation, real interest rates) INFLUENCE_DIRECT_PROPORTIONALY(currencies,inflation)
INFLUENCE_INVERSE_PROPORTIONALY(unemployment rates, inflation) INFLUENCE_INVERSE_PROPORTIONALY(monetary policies, inflation) INFLUENCE_INVERSE_PROPORTIONALY(economy, interest rates) INFLUENCE_INVERSE_PROPORTIONALY(inflation, unemployment rates) INFLUENCE_INVERSE_PROPORTIONALY(credit worthiness, interest rate) INFLUENCE_INVERSE_PROPORTIONALY(interest rates, bonds)
INFLUENCE(Internal Revenue Service, interest rates) INFLUENCE(economic growth, share prices)
EQUIVALENT(big mistakes, high inflation rates of 1970s) EQUIVALENT(fixed interest rate, coupon)

Table 3: A part of the relationships derived from the 5000 sentence corpus

4 Applications

An application in need of domain-specific knowledge is Question Answering. The concepts and the relationships acquired 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:
Q1: What factors have an impact on the *interest rate*?
Q2: What happens with the *employment* when the *economic growth* rises?
Q3: How does *deflation* influence *prices*?

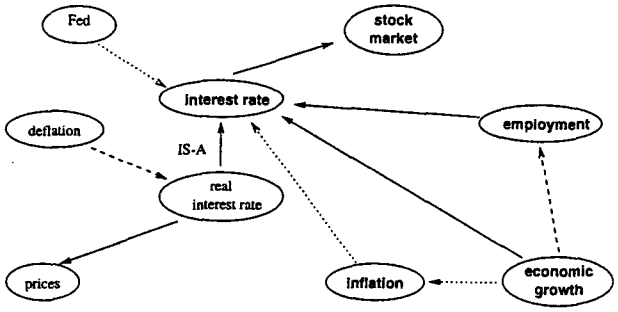

Figure 6: A sample of concepts and relations acquired from the 5000 sentence corpus. Legend: continue lines represent *influence inverse proportionally*, dashed lines represent *influence direct proportionally*, and dotted lines represent *influence* (the direction of the relationship was not specified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

Figure 6 shows a portion of the new domain knowledge that is relevant to these questions. The first question can be easily answered by extracting the relationships that point to the concept *interest rate*. The factors that influence the *interest rate* are *Fed, inflation, economic growth*, and *employment*.

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 impact on *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.

5 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 of which 319 contain the seeds and 43 relate to these via selected relations. There were 22 distinct lexico-syntactic patterns discovered used in 63 instances. 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 average 40 minutes per seed for the 5000 sentence corpus. KAT is useful considering that most of the knowledge base construction today is done manually.

Complex linguistic phenomena such as coreference 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

Christiane Fellbaum. WordNet - *An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998.

Marti Hearst. Automated Discovery of WordNet Relations. In *WordNet: An Electronic Lexical Database and Some of its Applications*, editor Fellbaum, C., MIT Press, Cambridge, MA, 1998.

J. Kim and D. Moldovan. Acquisition of Linguistic Patterns for knowledge-based information extraction. *IEEE Transactions on Knowledge and Data Engineering 7(5): pages 713-724.*

R. MacGregor. A Description Classifier for the Predicate Calculus. *Proceedings of the 12th National Conference on Artificial Intelligence* (AAAI94), pp. 213-220, 1994.

Stephen D. Richardson, William B. Dolan, Lucy Vanderwende. MindNet: acquiring and structuring semantic information from text. Proceedings of ACL-Coling 1998, pages 1098-1102.

Ellen Riloff. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1044-1049. The AAAI Press/MIT Press.

J.G. Schmolze and T. Lipkis. Classification in the KL-ONE knowledge representation system. *Proceedings of 8th Int'l Joint Conference on Artificial Intelligence* (IJCAI83), 1983.

S. Soderland. Learning to extract text-based information from the world wide web. In the *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97).*

Text REtrieval Conference. http://trec.nist.gov 1999

W.A. Woods. Understanding Subsumption and Taxonomy: A Framework for Progress. In the *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, Morgan Kaufmann, San Mateo, Calif. 1991, pages 45-94.

W.A. Woods. *A Better way to Organize Knowledge.* Technical Report of Sun Microsystems Inc., 1997.