Domain-Specific Knowledge Acquisition and Classification using WordNet

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

Abstract

For many knowledge intensive applications, it is necessary to have extensive domain-specific knowledge in addition to general-purpose knowledge bases usually built around Machine Readable Dictionaries. This paper presents a methodology for acquiring domain specific knowledge from text and classifying the concepts learned into an ontology that extends WordNet. The method was tested for three seed concepts selected from the financial domain: *interest rate*, *stock market*, and *inflation*. Queries were formed with each of these concepts and a small corpus of 500 sentences was extracted automatically from the Internet for each concept. The system learned a total of 151 new concepts and 69 new relations.

Domain Knowledge

The knowledge is infinite and no matter how large a knowledge base becomes, it is not possible to store all the concepts and procedures for all domains. Even if that was 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.

A possible solution to this problem is to build 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 knowledge intensive applications.

This paper presents a method that acquires new concepts and relations associated with some *seed* concepts, and classifies them using the WordNet linguistic knowledge structure. The sources of the new knowledge are texts acquired from the Internet. 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.

WordNet 1.6 is one of the largest machine readable dictionaries containing over 160,000 words grouped in 91,000 synsets and linked by 350,000 relations (Fellbaum 1998). WordNet is a general-purpose lexicon, meaning that it contains only the most commonly used words, and relations that always hold. Practically in any domain, being finance, scientific domains, or sports there are specific terms, acronyms, proper names and procedures commonly used by that domain experts that cannot be found in ordinary dictionaries. In many knowledge intensive applications, such as High Performance Knowledge Bases (Cohen et al.1998), Question Answering (TREC-8 1999), Information Extraction (Muslea 1999) and others, it is necessary to have extensive domain knowledge that is beyond what is currently available in WordNet. To facilitate reasoning, domain knowledge should not be acquired in vacuum, it should expand an existent ontology and WordNet provides a very good skeletal structure for this.

Learn new concepts

Select seed concepts

New domain knowledge can be acquired around some seed concepts that a user considers important. In this paper we focused on the financial domain, and used three seed concepts: *interest rate*, *stock market*, and *inflation*. The knowledge we seek to acquire relates to one or more of these concepts, and consists of new concepts that are not defined in WordNet and relations that link these concepts with other WordNet concepts.

For example, from the sentence: *When the US economy enters a boom, mortgage interest rates rise,* the system learns: (1) the new concept *mortgage interest rate* not defined in WordNet but related to the seed concept *interest rate*, and (2) that between the *state of the US economy* and the value of *mortgage interest rate* there is a DIRECT RELATIONSHIP.

Extract sentences containing the seed concepts
Queries are formed with each seed concept to extract documents from the Internet. The documents retrieved are further processed such that we retain only the sentences that contain the seed concepts. In our experiments, the number of sentences was arbitrarily limited to 500 for each seed.

Extract new concepts
In this paper only noun concepts are considered. Since most likely, one word nouns are defined in WordNet, the focus here is on compound nouns, or 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) in which the seeds reside. In our example, we see that the seed belongs to the *mortgage interest_rate* noun phrase. This way, from all the parsed sentences a list of NPs that contain the seeds is assembled. Every such NP is considered a potential new concept. This is only the raw material from which actual concepts are to be discovered.

Two distinct cases are considered.

Case 1. NP in which the seed is the head noun, [*word, word,..seed*], where *word* can be a noun or an adjective. For example, [*interest_ rate*] is in WordNet, but [*short term real interest_rate*] is not in WordNet. Most of the new concepts related to a seed are generated this way.

Case 2. NP in which the seed is not the head noun [*word, word,..seed, word, word*]; for example [*inflation risk*], or [*international interest_rate differential*].

The following procedure is used to extract concepts, and is applicable in both cases:

`Procedure 1:` Concept extraction.
`Input:` Noun phrases that contain the seed concept.
`Output:` New concepts constructed around the seed.

Step 1. *WordNet reduction.* Search NP for two or more consecutive words 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.

Step 2. *Dictionary reduction.* For each NP, further search in other on-line dictionaries for more compound concepts. 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).

The following example shows a partial list of concepts from the NPs containing *interest rate*:

[*long_term interest_rate*] [*prime_interest_rate*] [*mortgage_interest_rate*] [*US short_term interest_rate*] [*interest_rate_risk*] [*interest_rate analysis*] [*Federal_Reserve interest_rate hike*] [*interest_rate derivative_securities*]

Since we lack a formal definition of a concept, it is not feasible to completely automate the process of concept learning. The human may inspect the list with the remaining noun phrases and decide whether to accept or decline each concept.

Concept classification

The next step is to create a taxonomy for the newly acquired concepts that is consistent with WordNet. In addition to creating an ontology, this step is also useful to validate the concepts we acquired above. The classification is based on the subsumption relation.

`Procedure 2:` Classification by subsumption
`Input:` A list of NPs containing the seed as head noun
`Output:` An ontology of concepts under the seed

Step 1. Take out all adjectives from the NPs, with the following two exceptions:

1. if the adjective is part of a concept determined with Procedure 1 (e.g.: *nominal interest rate*), and

2. if the adjective comes from a proper noun (PN) and the relationship holds: POSSESSOR(PN, N) (e.g.: *European interest rate*).

Step 2. Classify concepts of the form [word, seed], where word is a noun or an adjective which belongs to one of the exceptions from Step 1. The classification is based on the simple idea that a compound concept [word, seed] is ontologically subsumed by concept [seed]. For example, *mortgage_interest_rate* is a kind of *interest_rate*, thus linked by a relation HYPERNYM(interest_rate, mortgage_interest_rate).

Similarly, for a relative classification of any two concepts [word1, seed] and [word2, seed], the ontological relation between word1 and word2 if it exists is extended to the two concepts. In the case that word1 subsumes word2, a relation is formed between the two concepts. An example is HYPERNYM(European interest_rate, German_interest_rate), since in WordNet there is the relation HYPERNYM(Europe, Germany). This applies also in the cases when word1 and word2 are linked via a chain of subsumption relations, since subsumption is usually a transitive relation.

In the case when between word1 and word2 there is no subsumtion relation, we seek to establish whether word1 and word2 have a common subsuming concept. If found, pick the most specific common subsumer (MSCS) concept of word1 and word2. Then form a concept [MSCS(word1, word2), seed] and place it under [seed], and also place [word1, seed] and [word2, seed] under [MSCS(word1,word2), seed]. For example, to classify [*German interest_rate*] and [*Japanese interest_rate*], we find that [*country interest_rate*] subsumes them both, as in WordNet [*country*] is the MSCS of [*Germany*] and [*Japan*].

Step 3. The final step here is to group the adjectives striped from the NPs in Step 1, based on the synonymy or antonymy relations defined in WordNet. The adjectives provide a set of attributes or features for the noun concepts. We observed that the adjectives modifying a concept in the ontology could also modify other concepts in the same ontology under the seed.

The following figure shows a part of the noun ontology learned for the *interest rate*, and the adjectives used to modify these concepts (the emphasized concepts were already in WordNet, while the concepts marked with a star (*) were discovered with Procedure 2, Step 2).

- **interest rate**
    - *prime interest rate*
    - Federal Reserve interest rate
    - market interest rate
    - nominal interest rate
    - variable interest rate
    - mortgage interest rate
    - real interest rate
    - *country interest rate
        - *Asian country interest rate
            - Japan interest rate
        - *European country interest rate
            - German interest rate
        - *North American country interest rate
            - Canadian interest rate
        - *South American country interest rate
            - Brazilian interest rate

Time
-short-term
-long-term
-annual
-overnight
-future
-current

Value
-high
-low
-maxim

Direction
-falling
-rising
-stable

In Table 1 we summarize the number of concepts extracted from the 1500 sentence corpus *A*. First, it is shown the number of unique NPs containing each seed, then the concepts acquired with Procedure 1, followed by the concepts accepted by human.

		a	b	c
Total potential concepts (NPs)		431	237	322
Total concepts extracted with Procedure 1				
Concepts found in WordNet		2	0	1
Concepts found in on-line dictionaries, but not in WordNet	Concepts with seed head	3	0	3
	Concepts with seed but not head	2	0	1
Concepts accepted by human		53	48	38

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

Learn Relations

Texts are a rich source of information from which in addition to concepts we can also learn relations between concepts. We are interested here on finding out the relations that link the concepts extracted above with other concepts from 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 an enhanced WordNet knowledge base. However, there are other semantic relations between concepts that do not have a correspondent in WordNet relations, thus through an iterative process we enlarge the set with new basic semantic relations that occur frequently in connection with the seed concepts.

In this section we present briefly the main steps of two related procedures that detect the relationships between the newly acquired concepts and other WordNet concepts. The procedure is broken down into two parts. Part (a) learns lexical patterns in which semantic relations can be expressed, and part (b) learns new connections between concepts. A new training corpus of 500 sentences, *B*, containing the seed concepts was extracted from the Internet (mostly from CNNfn).

Procedure 3a: Learn lexico-syntactic patterns
Input: Training corpus *B* and semantic relation types
Output: Basic lexico-syntactic patterns in which semantic relations can be expressed

Step 1. Pick a semantic relation from WordNet, R_w, (such as HYPERNYMY), or other relation that is known to be important for a seed, denoted as R_s. For example, by inspecting a few sentences containing *interest rate* one can notice that INFLUENCE is a frequently used relation. Start with one relation at a time.

Step 2. Pick a pair of words, not necessarily seeds, among which there is the relation *R* selected in Step 1. Let's consider the INFLUENCE relation:

interest rate INFLUENCES *earnings*, or
credit worthiness INFLUENCES *interest rate*.

Step 3. Search on the corpus for all instances when the pairs of two concepts selected above occur in the same sentence. Extract the lexico-syntactic patterns that link the two concepts in a pair. Examples are:

1. "The graph indicates the impact on earnings from several different interest rates". From this sentence extract the pattern that is generally applicable: [*impact on NP2 from NP1*] ⟹ INFLUENCE(NP1, NP2)

2. "As the credit worthiness decreases, the interest rate increases". From this sentence extract another pattern that is generally applicable: [*as NP1 vb1, NP2 vb2*] & [*vb1 and vb2 are antonymes*] ⟹INFLUENCE(NP1, NP2).

Step 4. Repeat from Step 1 for all the semantic relations in which one is interested.

Now that the patterns were extracted, they can be applied on a larger corpus in order to find new links between concepts.

`Procedure 3b`: Learn new relationships between concepts
`Input`: (1) The seed concepts list, (2) concepts learned with `Procedure 1` and classified with `Procedure 2`, (3) the corpus *A* and (4) the patterns acquired with `Procedure 3a`
`Output`: New relationships between concepts

Step 1. Using the new lexico-syntactic patterns found with `Procedure 3a`, search the corpus *A* for all concepts connected with a seed or a concept containing a seed via the same patterns. Repeat Step 1 for all seed concepts and for all concepts classified with `Procedure 2`.

Step 2. Repeat from Step 1 for all relations.

These Procedures were implemented and the results are shown in Table 2. In addition to the WordNet relation HYPERNYMY, two other semantic relations occurred frequently in our corpus: INFLUENCE, and CAUSATION. From the training corpus *B*, `Procedure 3a` discovered lexico-syntactic patterns for each of these semantic relations (Table 2 shows the most general ones). Then, according with `Procedure3b`, a search was implemented to automatically find these patterns on corpus *A*. The procedure provided a total of 147 patterns in which at least one of the seeds occurred. From these, by inspection we have accepted 69 and rejected 78. The columns **a**, **b**, and **c** in Table 2 indicate the exact number of occurrences for each pattern, along with an example. These 69 occurrences provide relationships between the seed concepts or concepts determined with `Procedure 1` and other WordNet concepts, and are added to WordNet.

Again, the intervention to accept or reject relationships is necessary mainly due to our system inability of handling coreference resolution and other complex linguistic phenomena. For example, the sentence ***The UK Budget has useful measures designed to help small businesses but its impact on interest rates is now seen as dependent on the General Election*** is rejected because we cannot determine automatically the referent.

Applications

An application in need of domain-specific knowledge is Question Answering. The concepts and the relations acquired can be useful in answering difficult questions that normally cannot be easily answered just by using the information from WordNet.

Consider the following questions:

1. What factors have an impact on the ***interest rate***?
2. What happens with the ***unemployment rate*** when ***inflation*** rises?
3. How does ***deflation*** influence ***prices***?

Figure 1 shows a small portion of the knowledge acquired that helps to answer these questions. `Procedure 1` learned the concepts ***real interest rate*** and ***nominal interest rate*** placed by `Procedure 2` under their hypernym ***interest rate***. The other unbolded concepts were already in WordNet. The relationships were learned with `Procedure 3`.

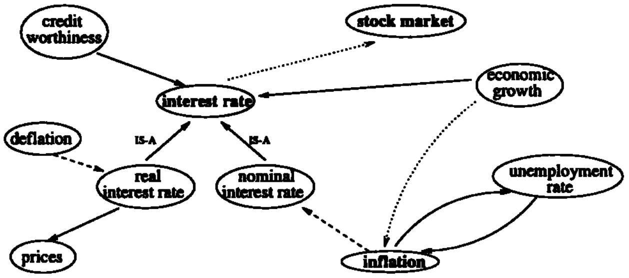

Figure 1: **A sample of concepts and relationships acquired from the 1500 sentence corpus. Legend: continue lines represent *influence inverse proportionally*, dashed lines represent *influence direct proportionally*, and dotted lines represent *influence* (the direction of the relation was not specified in the text).**

The first question can be easily answered by extracting the relationships that point to the ***interest rate*** concept. The factors that influence the ***interest rate*** are ***credit worthiness*** and ***economic growth***.

The last two questions ask for more detailed information about the complex relationships 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.***

Conclusions

The knowledge acquisition technology described above is applicable to any domain, by simply selecting appropriate seed concepts. We started with three concepts

Relations	Rules	a	b	c	Examples
WordNet Relations					
HYPERNYMY	NP1, including NP2,..NPn ⇒ HYPERNYMY(NPi,NP1)	3	0	0	There are a variety of *swaps products*, including *interest rate*, *currency*, and *basis swaps*.
	NP1 [<be>] a sort of NP2 ⇒ HYPERNYMY(NP1,NP2)	1	0	0	Thus, LIBOR is a kind of interest rate, as it is charged on deposits between banks in the Eurodolar market.
	NP1, such as NP2 ⇒ HYPERNYMY(NP1,NP2)	3	4	3	.. *regional stock markets* such as *Cincinnati Stock Exchange* and *Philadelphia Stock Exchange*.
New Relations					
CAUSE	NP1 [<be>] cause NP2 ⇒ CAUSE(NP1,NP2)	1	0	1	But if we don't consider external factors, we can assume that high *interest rates* are causing unemployment.
INFLUENCE	impact on NP2 from NP1 ⇒ INFLUENCE(NP1,NP2)	2	3	3	The graph indicates the impact on *earnings* from several different *interest rates*.
	As NP1 vb, so do/does NP2 ⇒ INFLUENCE(NP1,NP2)	0	0	1	As the *economy* picks up stream, so does *inflation*.
	NP1 <be> associated with NP2 ⇒ INFLUENCE(NP1,NP2), INFLUENCE(NP2,NP1)	2	2	6	Higher *interest rates* are normally associated with weaker *stock markets*.
	As NP1 vb1, NP2, vb2. + vb1, vb2 = antonymes ⇒ INFLUENCE(NP1,NP2)	1	1	0	As *the credit worthiness* decreases, *the interest rate* increases.
	NP1 (and thus NP2) ⇒ INFLUENCE(NP1,NP2)	1	0	0	We believe that the 30-year *Treasury bond* (and thus) *interest rates* is in a downward cycle.
	if NP1 vb1, NP2 vb2. + vb1, vb2 = go in the same direction ⇒ INFLUENCE(NP1,NP2)	3	0	1	Traders said that if *U.S. interest rates* rise, *dollars* would head north, erasing interest in the local market.
	if NP1 vb1, NP2 vb2. + vb1, vb2 = antonymes or go in opposite direction ⇒ INFLUENCE(NP1,NP2)	2	0	1	On the other hand, if *interest rates* go down, *bonds* go up, and your bond becomes more valuable.
	the effect of NP1 on/upon NP2 ⇒ INFLUENCE(NP1,NP2)	0	2	5	The effects of *inflation* upon *debtors* and *creditors* varies as the actual inflation is compared to the expected one.
	inverse relation between NP1 and NP2 ⇒ INFLUENCE(NP1,NP2), INFLUENCE(NP2,NP1)	0	0	1	There exists an inverse relationship between *unemployment rates* and *inflation*, best illustrated by the Phillips Curve.
	NP1 reflect NP2 ⇒ INFLUENCE(NP1,NP2)	2	1	3	As a rough approximation, at least, *nominal interest rates* reflect *rates of inflation*.
	NP1 lead to NP2 ⇒ INFLUENCE(NP1,NP2)	2	0	1	Thus a higher *inflation* (really, expected inflation since it's the future that matters) leads to higher *nominal interest rates*.
	NP1 and NP2 <be> negatively correlated ⇒ INFLUENCE(NP1,NP2)	1	0	0	The chart shows that *interest rate* and *equities* are negatively correlated: when *interest rates* climb, *equities* do poorly.
	NP2 <be> determined by NP1 ⇒ INFLUENCE(NP1,NP2)	0	0	1	Under the new system, the higher *taxation* will automatically be determined by *inflation* - as new car prices rise, the assessed tax benefit will rise in proportion.
	When NP1 vb1, NP2 vb2 + vb1, vb2 = antonymes / go in opposite directions ⇒ INFLUENCE(NP1,NP2)	2	0	3	Phillips, a British economist, stated in 1958 that when *inflation* rises the *unemployment rate* decreases and when *inflation* decreases the *unemployment rate* increases.

Table 2: Relations derived from the 1500 sentence corpus

interest rate, *stock market*, and *inflation* and acquired a total of 151 new concepts and 18 distinct lexico-syntactic relations used in 69 instances to link the seed ontologies with other WordNet concepts. Most importantly, the new concepts can be integrated with an existing ontology, and the type of the new relations is small which is helpful for reasoning activities.

At this time the method is not yet fully automated, but it gives the user choices and the user only has to accept or decline the concepts and relationships learned. The manual selection and validation of concepts and relationships took on average 25 minutes per seed. This is a big gain of time, considering that currently the acquisition of new knowledge is done manually.

References

Paul Cohen, Robert Schrag, Eric Jones, Adam Pease, Albert Lin, Barbara Starr, David Gunning and Murray Burke. The DARPA High Performance Knowledge Bases Project. In *AI Magazine*, Vol 18, No 4, pag. 25-49, 1998.

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, MIT Press, Cambridge, MA, 1998.

Lynette Hirschman, Marc Light, Eric Breck and John D. Burger. Deep Read: A Reading Comprehension System. In the *Proceedings of the 37th Meeting of the Association for Computational Linguistics (ACL-99)*, pag. 325–332, University of Maryland, 1999.

J. Kim and D. Moldovan. Acquisition of Linguistic Patterns for knowledge-based information extraction. *IEEE Transactions on Knowledge and Data Engineering*, 1995, 7(5): pag. 713-724.

G.A. Miller. WordNet: A Lexical Database. *Communication of the ACM*, vol 38: No11, pag. 39–41, Nov. 1995.

Ion Muslea. Extraction Patterns for Information Extraction Tasks: A Survey. In the *AAAI Workshop*, pag. 1-6, Orlando, Florida, 1999.

Ellen Riloff and Rosie Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In the *Proceedings of the 16th National Conference on Artificial Intelligence*, pag. 474-497, Orlando, Florida, 1999.

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), 1997.*

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

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