Extracting Causation Knowledge from Natural Language Texts

陳祈
CHAN Ki, Cecia

A Thesis
Submitted in Partial Fulfilment of the Requirements for the Degree of
Master of Philosophy
in
Systems Engineering and Engineering Management

August 2002

Abstract

SEKE is a semantic expectation-based knowledge extraction system for extracting causation knowledge from natural language texts. It is inspired by human behavior on analyzing texts and capturing information with semantic expectations. The framework of SEKE consists of different kinds of generic templates organized in a hierarchical fashion. There are semantic templates, sentence templates, reason templates and consequence templates. The design of templates is based on the expected semantics of causation knowledge. They are robust and flexible. The semantic template represents the target relation. The sentence templates act as a middle layer to reconcile the semantic templates with natural language texts. With the designed templates, SEKE is able to extract causation knowledge from complex sentences. Another characteristic of SEKE is that it can discover unseen knowledge for reason and consequence by means of pattern discovery. Using simple linguistic information, SEKE can discover extraction pattern from previously extracted causation knowledge and apply the newly generated patterns for knowledge discovery. To demonstrate the adaptability of SEKE for different domains, we investigate the application of SEKE on two domain areas of news articles, namely the Hong Kong stock market movement domain and

the global warming domain. Although these two domain areas are completely different, in respect to their expected semantics in reason and consequence, SEKE can effectively handle the natural language texts in these two domains for causation knowledge extraction.

摘要

SEKE系統能從自然語言文句擷取與因果關相關的資訊，其靈感來自我們分析文章中選取資訊的行爲模式。此系統採用了以樣板組成的分層結構，當中包括語意樣板，語句樣板，成因樣板及後果樣板。SEKE透過這些設計了的樣板，擷取文字中的因果資訊。樣板是根據所預期的因果關係和語意而設計。語意樣板說明所期望的關係，語句樣板處於結構的中層，能將語意樣板與文句相配合。SEKE的另一特點是能從文句中抽取未曾見過的成因及後果。它能利用簡單的句法資料和已擷取的因果，發掘擷取資訊用的樣型。這些樣型會被應用在發掘新的資訊。我們將系統分別應用於香港股票市場的價格變動及溫室效應這兩種不同類別的新聞文章上，更藉此證明系統能適應不同類別的文章，並有效地取得其中的因果資訊。

Acknowledgments

I would like to express my thanks to my supervisors, Doctor Wai Lam and Doctor Kai Pui Lam, whose advice, helped me to get through all the challenges that I faced. I am deeply in debt to Dr. Wai Lam, as he has devoted so much time and effort in teaching me both in this research and in writing it. Without him, I would never have been able to accomplish this research and this thesis.

I would also like to thank Professor Boon Toh Low, who has given me the opportunity to work on this research and valunable help in starting the research. He has not only inpsired me and given me valuable ideas in this research, but also generously shared with me his view of life; that is, to enjoy your life and to enjoy your work.

I wish to express my warmest gratitude to all my collegues from the department of Systems Engineering And Engineering Management for their help and support. Moreover, I would like to thank my friends, Christine and Yanny, for encouraging me when I was frustrated, and cheering me up when I was under great pressure.

Above all, I would like to give my special thanks to my parents for their love and support. They are always there by my side, giving me the strength to come through the good times, bad times and uncertain times in my life. They have also given me the confidence and courage to make my own plans and pursue my own goals.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advance of information technology, and the rapid growth of the Internet, we are able to receive vast amounts of information via electronic means, in the form of texts, graphics, sound, etc. Among different forms, textual information constitutes a major part, as it conveys a large amount of context and preserves much of the human intelligence.

Texts in electronic format may come from different sources, such as newspapers, journal articles, manuals, e-mail messages, and so on. Extracting useful information is useful for users to digest the textual information. While humans can extract information rather easily, it is not such an easy task for computers. Moreover, natural language texts usually are expressed in different forms making the task more challenging. Natural language is a fundamental aspect of human behavior and is crucial to our lives. It represents the interface for humans to communicate with each other. For a long time, researchers, such as philosophers, linguists, and scientists, have believed that language has some influence on the way a person thinks. Therefore, by exploring the natural language, we can understand more about how human thinks.

Natural language text is the written form of natural language. It preserves human knowledge from generation to generation. The handling of natural language text becomes a challenging issue for many computational linguists, as text is intricate and complex, and is filled with ambiguity and variations. What makes it more challenging is that language is always changing.

There are different kinds of relations commonly found in textual documents, such as whole-part, conditional, causation, and so on. In particular, causation relation plays an important role in human cognition, as it greatly influences people's decision making. It represents the relation between cause and effect, which is the basis of our expectation. Causation knowledge, which is part of human knowledge, is mostly recorded and conveyed through texts.

The aim of our research is to develop an approach in capturing causation information automatically from natural language texts. It is observed that humans can read a lengthy text and obtain the information that he or she needs with little effort. Therefore, learning from how humans extract information can help develop an effective information extraction system. Humans make decisions relying on expectations. Based on some expected semantics, one can perform searching accordingly, and analyze the information. Inspired by this observation, we propose an expectation-based approach to capture causation information. Our framework is called SEKE (Semantic Expectation-based Knowledge Extraction), which is a semantic expectation-based knowledge extraction system [23, 24, 5].

The framework of SEKE consists of different kinds of generic templates organized in a hierarchical fashion. The top most level one is the semantic template of target relation and is domain independent. The second level template consists of sentence templates handling different sentence styles and complex sentences. They also act as a middle layer to reconcile the semantic template to the bottommost level templates associated with the expected semantics of the domain and the relation. As a result, SEKE can extract causation knowledge from complex sentences without full-fledged syntactic parsing, by the association of a causation semantic template with a set of sentence templates.

SEKE can extract expected semantics from seed knowledge with the predesigned templates. As new knowledge or unpredicted information appears from time to time, we cannot solely depend on the coverage of the initial lexicons. Therefore, another characteristic of SEKE is that it can discover unseen knowledge. This is achieved by incorporating two tasks. The first task is to make use of an electronic thesaurus to identify similar concepts. The second task is to make use of automatically generated patterns to discover unseen knowledge. By applying the discovered patterns, SEKE can extract new reasons or consequences from texts. As a result, the performance of the causation extraction is improved. Moreover, the generation of patterns does not require manual annotations, which means that no extra preparations or human efforts are needed. The newly discovered knowledge can also become part of the domain specific lexicons.

To demonstrate the adaptability of SEKE for different domains, we study the application of SEKE on two domain areas of news stories, namely the Hong Kong stock market and global warming.

1.1 Our Contributions

Semantic expectation-based extraction: We make use of different kinds of generic templates organized in a hierarchical fashion. The design of templates requires mainly semantic knowledge and simple linguistic clues related to causation. Changes can be made easily on the templates and the robustness as well as flexibility of the system can be achieved.

Unseen reason and consequence discovery: In practice, it is impossible to encode a complete lexicon manually in practice. To enhance the extraction performance, it is necessary to capture the unseen knowledge. SEKE can automatically discover extraction patterns with simple linguistic information and successfully extracted reasons and consequences. These discovered patterns are applied to extract new reasons and consequences from texts. No manual annotation is used in the discovery of patterns, and hence no extra human efforts are required.

Adaptability: We demonstrate the adaptability of SEKE by applying it to two domain areas, namely the Hong Kong stocks market and global warming. Although these two areas are completely different, in respect to their expected semantics in reasons and consequences, SEKE can effectively handle the natural language texts in these two domains for causation knowledge extraction.

1.2 Thesis Organization

The organization of this thesis will be as follows:

Chapter 1: Introduction. This chapter provides an introduction to the work including aims of the thesis.

Chapter 2: Related Work. This chapter describes some previous works in extracting causation knowledge from natural language texts. It also discusses the differences between SEKE and other existing works.

Chapter 3: Semantic Expectation-based Knowledge Extraction. In this chapter, we describe in detail the first part of SEKE system, including the basic ideas, the most important techniques, the framework in handling the semantic expectation-based extraction of causation knowledge.

Chapter 4: Using Thesaurus and Pattern Discovery for SEKE. This chapter presents the second part of the SEKE system, namely the technique of using theaurus and automatic pattern discovery.

Chapter 5: Applying SEKE on Hong Kong Stock Market Domain. This chapter presents the application of the system in the Hong Kong stock market movement domain. It also reports the results and evaluation for this domain.

Chapter 6: Applying SEKE on Global Warming Domain. This chapter presents the application of the system in the global warming domain. It also reports the results and evaluation of this domain.

Chapter 7: Conclusions and Future Directions. This chapter presents the conclusions and discusses some future directions.

Chapter 2

Related Work

In this chapter, we will describe some related works. First, we will provide some background information of each area and then look into some representative previous studies.

Full natural language understanding has long been recognized as an important topic. However, we are still far away from a truly versatile and general system. By focusing on a narrower domain, researchers work on different applications, such as text categorization, information extraction from texts, and text summarization. Information extraction has been drawing a lot of attention recently [10, 30, 6, 7]. Information extraction systems, such as the NYU proteus system [11] and the SRI FASTUS system [2], use syntactic parsing as the main technology, while some use syntactic parsing with the aid of semantic analysis to tackle the problem [14].

Causation knowledge plays an essential role in human decision making. It is concerned with what people's beliefs are. Many philosophers have given different definitions of causation. One of the most influential one is Hume's definition of causality [13]: "We may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or, in other words where, if the first object had not been, the second never had existed."

In modern discussion, causation, or alternatively, causality refers to "the relation between two items one of which is a cause of the other"[12, 36]. It represents the relations between causes and effects, which are the basis of our expectation. As causation knowledge, which is part of human knowledge, is mostly recorded and transmitted through texts, many researchers wish to extract the causation knowledge from texts. Previously, many researchers attempted to use knowledge-based inference techniques to build-up models for implicit causation knowledge in texts. Later, they focused on explicitly indicated causation knowledge. Recently, researcher tends to use linguistic techniques to achieve the task.

Explicitly indicated causation relations in texts can be expressed in the following ways:

- using causal links to link two phrases, clauses or sentences
- using causative verbs, verbs that denote causing something to happen
- using causative affixes

which are linguistic clues of the presence of causation relations.

2.1 Using Knowledge-based Inferences

Many studies in extracting causation knowledge from texts made use of knowledge-based inference technique to detect the causation knowledge in texts [32, 15, 21]. Kaplan and Berry-Bogghe [16], acquired causation knowledge from scientific texts. Although they used linguistic patterns to identify the causation relations, the grammar, the lexicon as well as the patterns for the system are all hand-crafted for a particular domain. The input for analysis of causation knowledge are manually designed with a set of propositions.

With a large amount of hand-coded domain knowledge, it is difficult to scale up for realistic applications. Moreover, the required rigid knowledge-base makes this approach suitable for only very limited domains and a very small amount of text.

2.2 Using Linguistic Techniques

Later, a lot of research focused on extracting explicitly indicated causation knowledge in texts using linguistic techniques.

2.2.1 Using Linguistic Clues

Garcia developed an automatic system, COATIS [8], to acquire causation knowledge from texts by using linguistic indicators of causality in sentences. It was designed to locate expressions that denote actions and are linked by causation relations in French texts. The system identifies the causation relations expressed by causative verbs of the French language. It manually classified the causative verbs, the indicator verbs of causality, into twenty-three kinds of causality, such as "to result", "to lead to", etc. The presence of an indicator invokes the system to detect the presence of the causation relations.

Khoo et al. [18, 19] developed an automatic extraction of cause-effect information from newspaper texts using linguistic clues and without any domain knowledge. It uses simple pattern matching without knowledge-based inferencing and without extensive parsing of sentences. A set of linguistic patterns that usually indicate the presence of a causation relation was constructed and used for pattern matching. The linguistic patterns were constructed based on manual analysis of the documents. The patterns were then refined by applying the patterns to sample sentences. The system is able to identify which part of the text is the cause and which part is the effect. It reported a recall of 68%.

2.2.2 Using Graphical Patterns

Khoo, Chan and Niu [17] developed a knowledge extraction system that extracts causation knowledge from texts using graphical patterns based on syntactic parsing. It uses a parser to construct syntactic parse trees for the sentences. Information is extracted from those parse trees by graphical patterns of causation knowledge. They focused on the extraction of causation knowledge from medical texts. Cause-effect templates are defined for different medical areas, by specifying different attributes involved in cause and effect. The attributes are also specified in the graphical patterns. The graphical patterns are constructed from manually analyzing training examples. A list of causal indicators is obtained from the examples of sentences with explicitly indicated causal relation. Based on the indicators, graphical patterns are obtained for each indicator by analyzing the ways a causal relation is explicitly expressed in a sentence. It reported an F-measure of 0.51 for extracting the cause and 0.58 for extracting the effect.

2.2.3 Using Lexicon-syntactic Patterns of Causative Verbs

Girju and Moldovan [9] developed an approach to automatically identifying lexicon-syntactic patterns which express the causal relation and semi-automatically validate the patterns. It focused on the syntactic patterns of a pair of noun phrases connected by causative verbs. The discovery of lexico-syntactic patterns of causation is done by first picking a pair of noun phrases with causation relations, then searching among texts for all the patterns where the pair of noun phrases are connected by a verb/verb expression. As not all of these patterns refer to causation, it imposes semantic constraints on both the verbs and noun phrases. The patterns are ranked by making use of WordNet's semantic information, analyzing the causation classes to which the cause and effect nouns belong, and analyzing the ambiguity of the verbs.

2.2.4 Comparisons with Our Approach

The above works mainly focus on identifying a causation relation in texts. Garcia as well as Girju and Moldocan attempted to identify the causation relation expressed by causative verbs. Khoo et al. focused more on the linguistic clues and syntactic pattern of causation knowledge. Their methods attempt to identify the presence of causation relation in simple sentences. Our research focuses on not only identifying the causation relation but also extracting the causes and effects in texts.

The approach of Khoo, Chan and Niu is able to extract more detailed information of cause and effect by incorporating the cause-effect template, which specifies different roles/concepts in cause and effect into the graphical pattern. However, as the cause-effect template is domain specific, the graphical patterns are also domain specific. To adapt the system to another domain, it requires the reconstruction of the graphical patterns. Moreover, expert knowledge of syntactic parse trees is required to construct the graphical patterns.

We attempted to design a more flexible approach for the extraction of cause and effect. Instead of matching with one set of domain specific patterns, we separated the extraction process into the identifying of causation relations, and the extraction of cause and effect information. The identification of causation relations is domain independent. To adapt the extraction process to another domain, we only have to update the template for the extraction of cause and effect information.

2.3 Discovery of Extraction Patterns for Extracting Relations

The causation relation can also be viewed as a tuple relation of reason and consequence. Even, reasons and consequences both have many different kinds of relations within. Due to the complexity of the relations, it is difficult to manually create all their patterns. Automatic extraction pattern discovery is useful for this problem. Some studies have attempted to develop systems for learning extraction patterns [33]. Many of the previous systems required the use of manually-tagged training data [28, 34]. Later, to reduce the manual effort, some researchers explored the area of automatic generation of extraction patterns [20, 29, 31, 26]. These systems is designed to discover extraction patterns for a certain relation. Two examples, namely Snowball and DIRT, aim to discover extraction patterns of relations for tuples in text.

2.3.1 Snowball system

Agichtein and Gravano [1] developed the Snowball system, which uses a handful of training examples to generate extraction patterns. The extraction patterns in turn extract new tuples from texts. During the extraction, Snowball evaluates the patterns and tuples to eliminate unreliable ones. It applies the system to the organization-location scenario, where a tuple represents the headquarters of some organizations. For the generation and matching of patterns, Snowball uses a named-entity tagger to identify phrases likely to be connected with organization and location. The Snowball patterns are a 5-tuple pattern indicating the tuple of named-entity and the texts on the left, middle and right of the tuple, and with weights assigned to the terms. It evaluates the patterns and tuples by updating the weight assigned to the patterns and calculating the confidence of the patterns and tuples.

2.3.2 DIRT system

Lin and Pantel [22] developed the system, DIRT, which aims to discover inference rules from text. The inference rules can be regarded as variants of patterns for a certain relation. It uses distributional hypothesis, which is the algorithm for finding similar words based on the idea that words tend to have similar meanings if they tend to occur in the same context. Instead of applying it to words, it applies it to paths in dependency trees. A path is a binary relation between two entities. It hypothesizes that the meanings of two paths are similar if they tend to link to the same sets of words. The similarity between two paths is computed from the frequency counts of all the slot fillers, the words filling the slots of the paths. Hence, two paths have a high similarity if there are a large number of common slot fillers.

2.3.3 Comparisons with Our Approach

Although Snowball is able to automatically generate extraction patterns and evaluate the tuples and patterns, it does not capture every instance of possible tuples, and only focuses on generating valid tuples. Therefore, it may consider some of the useful information as invalid. For example, if a tuple conflicts with user provided example tuples, it is considered invalid. This may not be true in practice, as in some domains, both of the tuples are possible.

Moreover, Snowball has limited applications as it only focuses on named-entity and its pattern representation can only handle two attributes. Similarly, DIRT also has the same limitations of only focusing on binary relations of nouns. In contrast, our approach not only extracts tuples with more than two attributes, but also identifies the causation relation between two kinds of tuple relations.

Chapter 3

Semantic Expectation-based Knowledge Extraction

SEKE (Semantic Expectation-based Knowledge Extraction) uses a semantic expectation-based knowledge extraction approach for extracting causation knowledge from texts [23, 24, 5]. In this approach, a set of generic templates organized in a hierarchical fashion is designed. In this chapter, we present in detail the characteristics and structures of each kind of template, and how the templates are organized to facilitate the causation knowledge extraction.

3.1 Semantic Expectations

Humans can extract precise information from texts readily and easily because of the following two characteristics:

1. There is always an expected semantic in mind, and

2. The expected semantic is used to guide the search and understanding.

For example, for someone who is interested in knowing the latest stock market movement, he or she may want to read about the analysis of cross-market influences, and what causes the recent market to move up or down. These are the semantic concepts expected in one's mind. While a person is reading newspaper articles, he or she will be particularly paying attention to the information related to their expectations. Inspired by this human behavior, we have developed an effective causation knowledge extraction system, called SEKE.

There are many relations expressed in natural language. The causation relation is an important one for human reasoning. Humans preserve their knowledge in texts and much of the knowledge is related to causation knowledge which helps us understand our world. Our goal is to find a way to extract the causation knowledge for effective understanding and reasoning.

Even though there are many different ways to present the information, a limited number of semantic structures are preserved for a particular type of relation. The encapsulated knowledge based on the expected semantics of the relation can be extracted by the use of semantic templates which specify the linking of actions. A causation semantic template states the linkage between reasons and consequences. It represents the highest level of templates. Furthermore, it is domain and language independent. In the next level of the hierarchy, it is associated with some sentence templates which act as a middle level to reconcile the semantic template, consequence and reason templates to a particular language. With some expected concepts of consequences and reasons, in the form of consequence and reason templates, we can model detailed content regarding the causation.

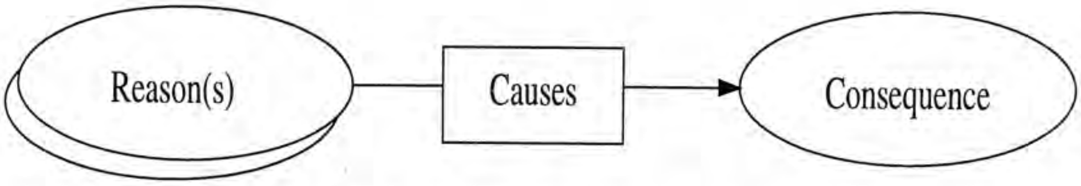

Figure 3.1: Causation Semantic Template

3.2 Semantic Template

Semantics in natural language processing refers to the meaning conveyed from texts. It is generally agreed that some of the basic semantic relations, such as causation, negation, and so on, are usually expressed in structured forms in different languages [27, 35]. We observe that the same kind of semantics are preserved for a particular type of relation. To process a certain type of relation in texts, we represent the expected semantics of those semantic relations by semantic templates. They capture the existence of different entities or actions and their linkage.

3.2.1 Causation Semantic Template

Causation relation is usually regarded as one of the fundamental semantic relations. The knowledge it captures is a kind of important logical concept. A causation relation typically has two kinds of entities, namely a reason and a consequence. These entities are linked by a directional causation indicator. For causation knowledge, the expected semantics are reason and consequence. A basic semantic template is shown in Fig. 3.1. This semantic template captures the fact that one or more reasons cause the occurrence of a consequence.

3.3 Sentence Templates

The semantic template is language independent. However, we need to deal with natural texts written in a particular language. In particular, a variety of sentence styles can express causality relationships. Sentence templates, associated with a semantic template, are introduced to handle different styles of sentences. They represent the characteristics of expressing a relation in texts.

The following two sentences are both obtained from the same piece of news article from Reuters on April 4, 2002, which illustrate different writing styles in expressing almost the same content. The semantics of the two sentences is concerned with the linking of two events. Specifically, the fall of Wall Street is followed by the fall of the Hong Kong stock market.

1. "Hong Kong stocks are set to open lower on Thursday following a dismal performance on Wall Street."

2. "HK stocks set for weak start after Wall Street slide."

We examined English training news articles in two domains, namely the Hong Kong stock market movement and global warming. Sample sentences conveying causation knowledge were further investigated. The causation knowledge, expressed in English sentences in texts, can be categorized into simple sentences and complex sentences according to the organization of the reasons and consequences.

1. Simple Sentence:

   A simple sentence consists of a single or multiple reasons, for example:

   (a) Single-reason Sentence:

> *"Hong Kong stocks closed higher on Friday helped by another overnight rise in the Dow Jones Industrial average."*

The above sentence consists of only one reason: *"rise in the Dow Jones Industrial average"*.

(b) Multiple-reasons Sentence:

> *"The benchmark Hang Seng Index extended losses on Monday morning, sinking 654.77 points or 4.05 percent to 15,531.17 on Wall Street weakness and interest rate jitters."*

This sentence consists of multiple reasons: *"Wall Street weakness"* and *"interest rate jitters"*.

2. Complex sentence:

A complex sentence has a more complicated structure. It consists of single or multiple reasons like simple sentences, but the consequence or reason itself is more complex, as it contains a causation relation within it.

Here is an example of a sentence with a complex reason.

- Complex-reason sentence:

> *"Hong Kong stocks fell on Monday for a third day, taking a cue from Friday's dive in U.S. stocks as investors heeded Federal Reserve Chairman Alan Greenspan's warning that interest rates will probably rise."*

Fig. 3.2 shows the sentence templates in the SEKE system. The first two sentence templates are used to model the sentence structure of simple sentences.

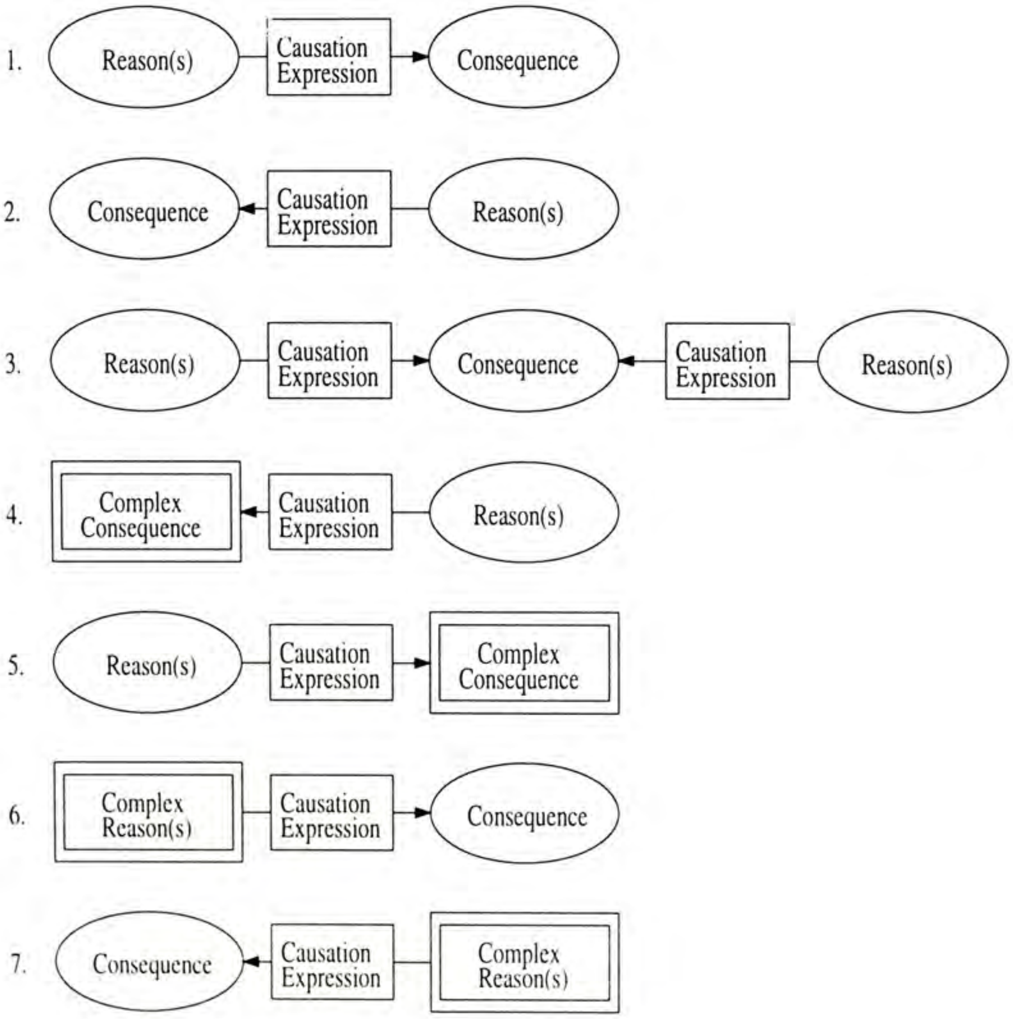

Figure 3.2: Sentence Templates

The sentence template 1 illustrates that some reasons cause a consequence, while the sentence template 2 illustrates that a consequence is caused by some reasons. One can see that the sentence template 1 is in the same order as the causation semantic template in Fig. 3.1, whereas the sentence template 2 is in the reverse order. The sentence template 3 states that some reasons cause a consequence, and those reasons can come before and after the consequence. It is observed that causation semantic templates have to be used also for extracting causation relation among reasons and consequences, and so have the sentence templates. Hence, one characteristic of the sentence templates is the recursive structure of the causation relation among reasons. Some sentence templates in Fig. 3.2 contain "Complex Consequence". It means that the consequence may consist of one of the first three sentence templates. In the last two sentence templates, "Complex Reason(s)" refers to the existence of one of the first three sentence templates in the reason(s). "Causation Expression" in the above templates refers to phrases for linking a consequence to a reason. Some examples are {*as, due to, because of, because, cause, caused by, helped by*}. Reasons are joined among themselves with the conjunction terms such as {*and*}.

Here are some examples of simple sentence templates in the Hong Kong stock movement domain:

- **Factor(s) with its movement causes the movement of stock**
  Example:

  "The increase of interest rate caused Hang Seng Index surged."

  where *"The increase of interest rate"* is the reason, *"Hang Seng Index* *surged"* is the consequence, and *"caused"* is the causation expression which links the reason to the consequences.

- **The movement of stock is caused by factor(s) with its movements**

  Example:

  > *"Hang Seng Index rose as Wall Street gained."*

  where *"Hang Seng Index rose"* is the consequence, *"Wall Street gained"* is the reason, and the *"as"* is the causation expression which links the reason to the consequence.

In the above two examples, the order of the reason and consequence is different. This shows that there are different sentence templates associated with the same causation semantic template. These two examples can be represented by the first two templates in Fig. 3.2. They each have only one reason, but usually more than one reason exists within a sentence.

Examples of sentence template for the complex sentence for the Hong Kong stock market movement domain and the global warming domain are shown as follows:

- **The movement of stock is caused by a factor with movement which is caused by another factor with movement.**

  Example:

  > *"Hong Kong stocks made into positive territory by midday on Thursday as investors picked up property plays, banking on a possible cut in US interest rate later this month."*

- **The movement of stock caused by a factor with movement which is caused by some factors with movements.**

  Example:

  > *"Hong Kong stocks took heart from strong gains on global telecom stocks and sprinted higher on Wednesday, driven by telecom heavy weight China Mobile which climbed over eight per cent."*

- **A factor with action caused by global warming causes a factor with movement.**

  Example:

  > *"An increase in temperatures as a result of global warming may lead to significantly higher."*

3.4 Consequence and Reason Templates

Similarly, reasons and consequences usually contain expected concepts or information. We design reason templates and consequence templates to represent their semantics. A variety of concepts can exist among consequences or reasons. For example, a reason or a consequence may have concepts including factors, movements, modifier of movements, time, duration, people, etc. The concepts for a consequence or a reason depend on what are the expected semantics, and also the focus of the causation relation.

If the causation relation focuses on finding the set of possible reasons, the consequence and reason templates have the following characteristics:

- The consequence template consists of one main concept and other additional concepts. This main concept of the domain is the same for every such causation relation.

- The reason template consists of at least one main concept and other additional concepts. The main concept can be one of the expected semantics, and need not be the same for all such causation relations.

Conversely, if the causation relation focuses on finding the set of consequences, the reason template will then have the same main concept of the domain for every such relation.

Using the Hong Kong stock market as an example, we focus on finding the set of possible reasons for the stock market movement. The consequence template includes "Hang Seng Index" or similar concept terms as the main concept, and other concepts such as what the market movement is, when the movement takes place, and how it moves. The reason template includes the factor(s) as the main concept and how it moves as the secondary concept. The two templates are shown in Fig. 3.3. Here, the factors are linked by a set of conjunction terms.

Here are some examples of reason templates and consequence templates for the Hong Kong stock market movement domain:

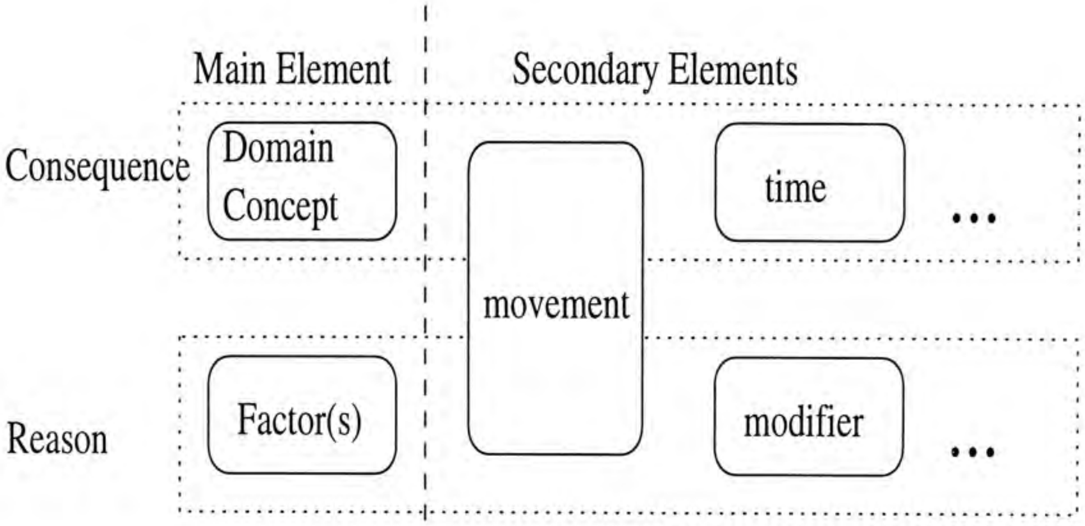

Figure 3.3: Consequence & Reason Template

Reason Template:	Factor	Movement	
	“Wall Street market”	*“gains”*	
	Factor	Movement	time
	“Wall Street market”	*“surged”*	*“yesterday”*
	Factor		Place
	“situation”	*in*	*“Argentina”*
Consequence Template:	Hang Seng Index	Movement	
	“Hang Seng Index”	*“rises”*	
	Hang Seng Index	Movement	Modifier
	“Hang Seng Index”	*“rises”*	*“sharply”*

where *“Wall Street market”* is the factor, *“Hang Seng Index”* is the consequence, *“gains”*, *“surged”* and *“rise”* are the movements. *“sharply”* is a modifier which describes the movement. *“yesterday”* is about the time of occurrence of the reason.

3.5 Causation Knowledge Extraction Framework

Based on the above observations, we have developed a basic framework of SEKE which can extract causation knowledge automatically from texts in a particular domain. The basic framework consists of three stages, namely template design, sentence screening, and semantic processing. Among these three stages, the first one is done manually by analyzing a training corpus containing relevant sentences of the domain. The remaining two stages are processed automatically. Fig. 3.4 shows the basic framework of SEKE.

3.5.1 Template Design

A training corpus is first prepared. It contains relevant sentences about a particular domain for causation knowledge extraction. Among those sentences, the ones expressing causation knowledge are picked out and analyzed for designing the templates. Moreover, initial lexicons for the expected semantics based on the causation semantic templates of that particular domain are prepared manually by examining those selected sentences. Optionally, the initial lexicons can be enhanced by items directly provided by users. They act as initial activations of SEKE extractions. The work involved in designing each kind of templates is explained as follows:

1. Sentence Template:

   Different sentence styles with different causation expression. It is used to identify the position of reason and consequence in a sentence, and the phrases for linking consequences to reasons.

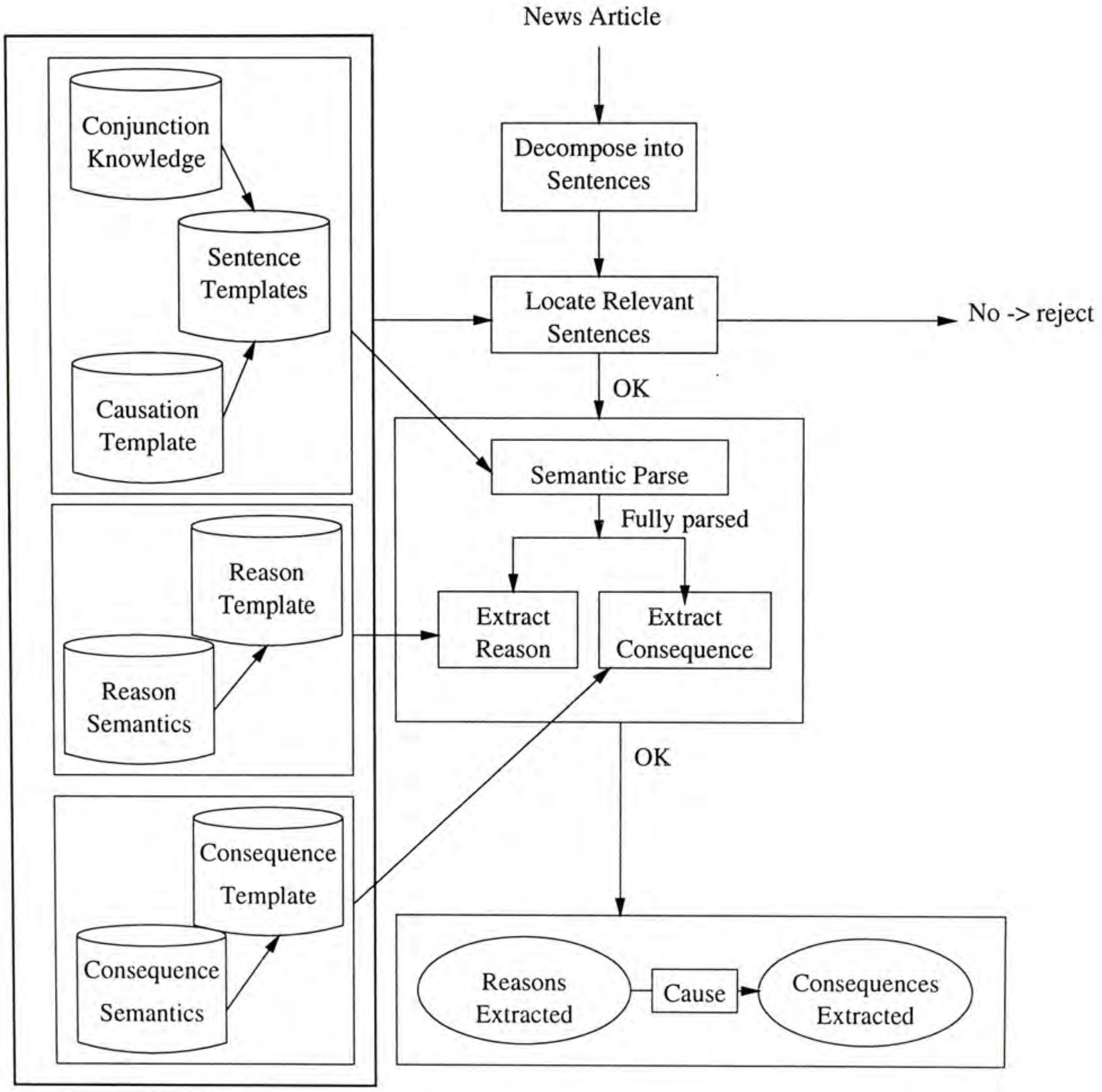

Figure 3.4: The basic framework of SEKE system

2. Reason Template:

   The reason concepts included in the reason template are first defined according to the expectation of the user. It refers to the kind of information the user wants. Then, initial reason lexicons including the seed concept terms for the expected concepts in the reason template are collected.

3. Consequence Template:

   Similarly, the consequence template is defined which specifies the concepts contained. Initial consequence lexicons with seed concept terms for the expected concepts in the consequence template are also collected from the training corpus.

3.5.2 Sentence Screening

Once the templates are designed, SEKE can process the documents automatically. The texts are first segmented into sentences. SEKE processes each sentence and attempts to screen out contexts that are irrelevant to causation knowledge. The steps involved are:

1. Using the templates, for each article being fed to the system, unrelated information is filtered out using the concept terms of the domain, and possibly relevant sentences are collected.

2. If the sentence can match with the expected semantic template and sentence templates, it is regarded as containing causation knowledge. It will be passed to the next stage.

3.5.3 Semantic Processing

After relevant sentences are filtered out, this stage is to conduct automatic semantic processing. The steps of semantic processing are as follows:

1. The collected sentences will be semantically parsed into reasons and consequences by the corresponding sentence template.

2. The reasons and consequences identified will be matched with the semantic templates again to see if they are complex reasons or consequences. If yes, it implies that the semantics of causation exist. Therefore, repeat procedure 1 to extract the causation knowledge in those reasons or consequences. Otherwise, it will move on to extract the information in reasons and consequences.

3. The reasons and consequences parsed will be matched with the reason template and consequence template respectively to extract the concepts of reasons and consequences. They are again parsed according to the reason and consequence templates, and are searched for the existence of the concepts in the reason and consequence templates.

4. The system will identify all the possible instances (terms) for each concept in the reason and consequence templates.

5. It can be observed from samples of causation sentences that:

    - if the number of expected concepts with a reason or a consequence template is more than one, and

    - if example a consequence consists of two concepts, A and B, and within a phrase, there are more than one possible candidate of A :

$\{a_1, a_2, ...\}$ and $B : \{b_1, b_2, ...\}$,

- then the pair of (a_x, b_y) with shortest distance between each other in the corresponding phrase has a higher possibility that b_y is the movement of a_x.

Therefore, among all the possible instances, the pair of terms with the closest distance between their positions in the phrase are regarded as the extracted reason or consequence.

6. Incomplete reason or consequence extracted will be passed to the next part of SEKE to discover unseen knowledge.

An Example

The following sentence is used to illustrates the steps of semantic processing:

> *"HK stocks ended practically unchanged on Monday, off earlier highs as the territory's top airline Cathay Pacific fell on 2002 earnings worries and property stocks slipped on concern that prices in the sector remain week."*

Step 1: The sentence is first matched with the sentence template:

[consequence] as [reason]

It is parsed into reason and consequence.

Consequence: *"HK stocks ended practically unchanged on Monday, off earlier highs"*

Reason: *"the territory's top airline Cathay Pacific fell on 2002 earnings worries and property stocks slipped on concern that prices in the sector remain week"*

For the consequence, the sentence template of multiple reason is matched, resulting in two reasons, Γ_1 and Γ_2.

Γ_1 *"the territory's top airline Cathay Pacific fell on 2002 earnings worries"*

Γ_2 *"property stocks slipped on concern that prices in the sector remain week"*

Step 2: Both the reasons consist of a causation relation, step 1 will be repeated, and the sentence template matched is:

[consequence] on [reason]

Γ_1 *"the territory's top airline Cathay Pacific fell on 2002 earnings worries"*

consequence: *"the territory's top airline Cathay Pacific fell"*

reason: *"2002 earnings worries"*

Γ_2 *"property stocks slipped on concern that prices in the sector remain week"*

reason: *"property stocks slipped"*

consequence: *"concern that prices in the sector remain weak"*

Step 3: The reason and consequence parsed will be matched with the reason template and consequences template respectively to extract the concepts of reason and consequence.

Consequence: *"HK stocks ended practically unchanged on Monday, off earlier highs"*

Consequence template: Hong Kong stock & movement

Consequence of Γ_1: *"the territory's top airline Cathay Pacific fell"*

Reason of Γ_1: *"2002 earnings worries"*

Consequence of Γ_2: *"property stocks slipped"*

Reason of Γ_2: *"concern that prices in the sector remain weak"*

Reason template: factor & movement

Step 4: The possible instances (terms) for each concept in the reason and consequence templates are identified. For example, "lower" is the instance of the movement concept.

Consequence:	Hong Kong stock	*"HK stocks"*
	movement:	"unchanged", *"highs"*
Consequence of Γ_1:	factor	*"Cathay Pacific"*
	movement:	"fell"
Reason of Γ_1:	factor	unidentified
	movement:	"worries"
Consequence of Γ_2:	factor	*"property stocks"*
	movement	"slipped"
Reason of Γ_2:	factor	*"prices"*
	movment	*"concern", "weak"*

Step 5 Multiple instances of the concept, movement, are identified for both the reason of Γ_2 and the consequence. Following the procedures in step 5, we compare the distances between the pair of conepts:

	factor	movement
$pair_1$	*"HK stocks"*	*"unchanged"*
$pair_2$	*"HK stocks"*	*"highs"*

$pair_2$ has a shorter distance than $pair_1$, hence $pair_2$ is regarded as the extracted consequence. Similarly, for the reason of Γ_2, the pair, "prices in the sector" and "concern", is extracted.

Step 6 As reason concepts for the reason of Γ_1 are not completely extracted, the next function of SEKE is to discover the unseen knowledge.

Representing the unseen knowledge as K, the following causation relations are extracted:

Consequence		Reason		Reason	
"HK stocks"	unchanged	*"Cathay Pacific"*	down	*"property stocks"*	down
"Cathay Pacific"	down	Knowledge, K	*"worries"*		
"property stocks"	down	*"prices"*	*"concern"*		

Chapter 4

Using Thesaurus and Pattern Discovery for SEKE

In the previous chapter, we illustrate how the basic framework of SEKE can extract expected semantics from seed knowledge. The basic framework of SEKE can extract causation knowledge buried in texts based on the pre-designed templates. The knowledge extracted depends solely on the coverage of initial lexicons. However, it is not possible to encode a complete lexicon manually in practice. New knowledge or unpredicted information appears from time to time.

We wish to enhance the extraction performance by incorporating two tasks into the basic framework of SEKE. The first task is to make use of a general-purpose knowledge base, such as an electronic thesaurus, to identify similar concepts, while the second task is to make use of automatically generated patterns to discover unseen knowledge. SEKE can extract new reasons or consequences from texts by applying the discovered patterns. Moreover, the generation of patterns does not require manual annotations. It means that no extra preparations or human efforts are needed. The basic framework of SEKE together with the use of a thesaurus and the discovery of patterns results in the complete framework as shown in Fig. 4.1. With this complete framework, we are able to improve the performance of the causation extraction and discover unseen causation knowledge. The newly extracted knowledge can also become part of the domain specific lexicons.

4.1 Using a Thesaurus

If the knowledge extracted from the phrase is incomplete or failed, it will be passed to this stage to search for similar concepts. In this stage, an electronic thesaurus, WordNet [25, 3], is used to identify similar concepts and conglomerate those terms by using the corresponding synonyms provided. Sentences or phrases are decomposed into words and phrases. WordNet is used for providing the synonyms for each of them according to their corresponding part of speech information provided by the tagger[1] . The part of speech information is used to reduce the ambiguities of words, restrict the synonyms obtained, and restrict the matching of synonyms with the existing concept terms. If its synonyms match with an existing concept term, that word or phrase is regarded as a similar concept and is accepted as part of the causation knowledge. It is also absorbed into the system and merged with the initial lexicons.

The steps of using the thesaurus are illustrated in Fig. 4.2. Optionally, human identification for similar concepts can be incorporated. After this stage, if the knowledge extracted is still incomplete or none of the words are identified as similar concepts, the phrases will be passed to the next stage for applying the patterns discovered.

[1] More descriptions about the part of speech tagger will be given in Chapter 4.2.2

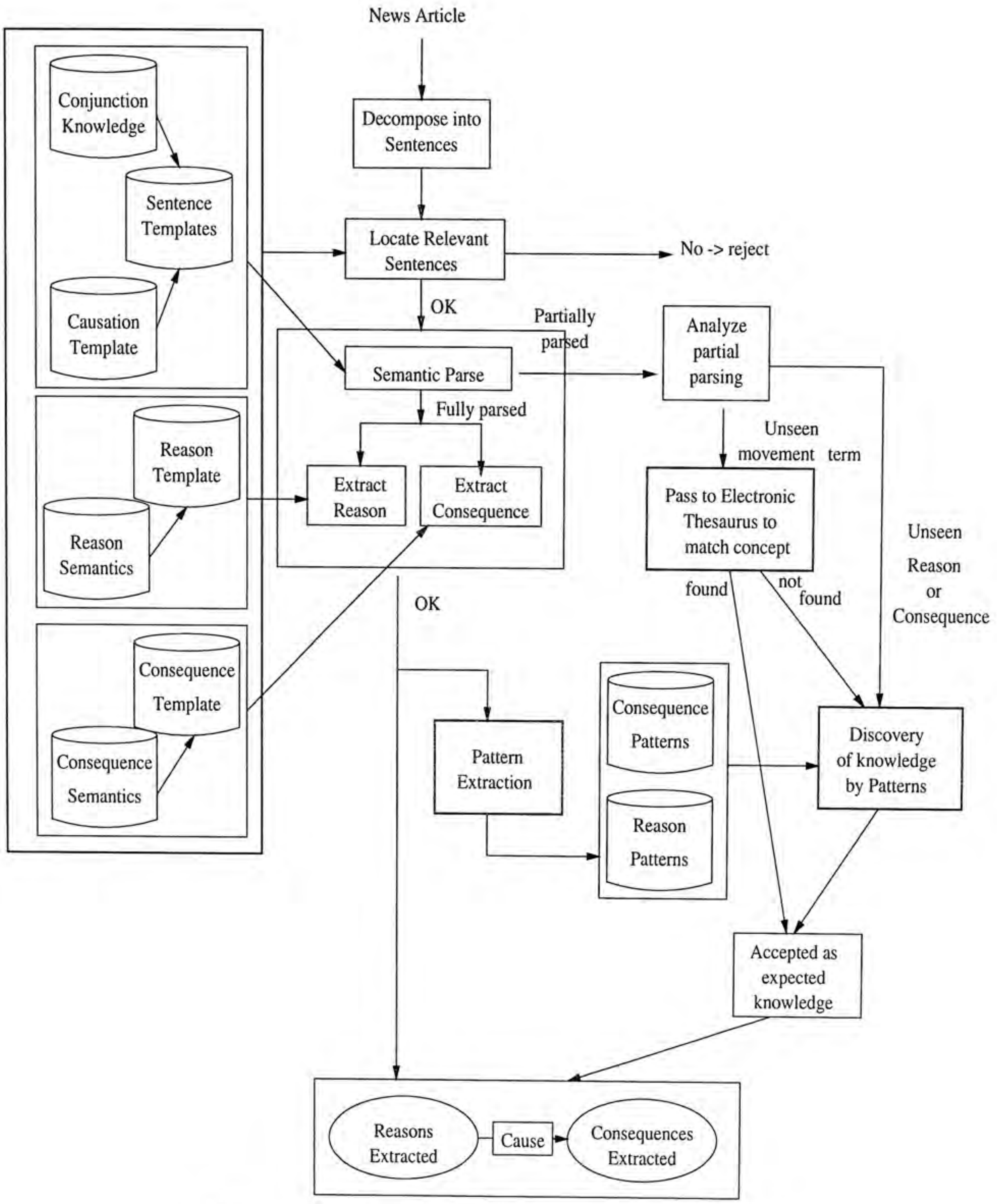

Figure 4.1: The complete framework of SEKE

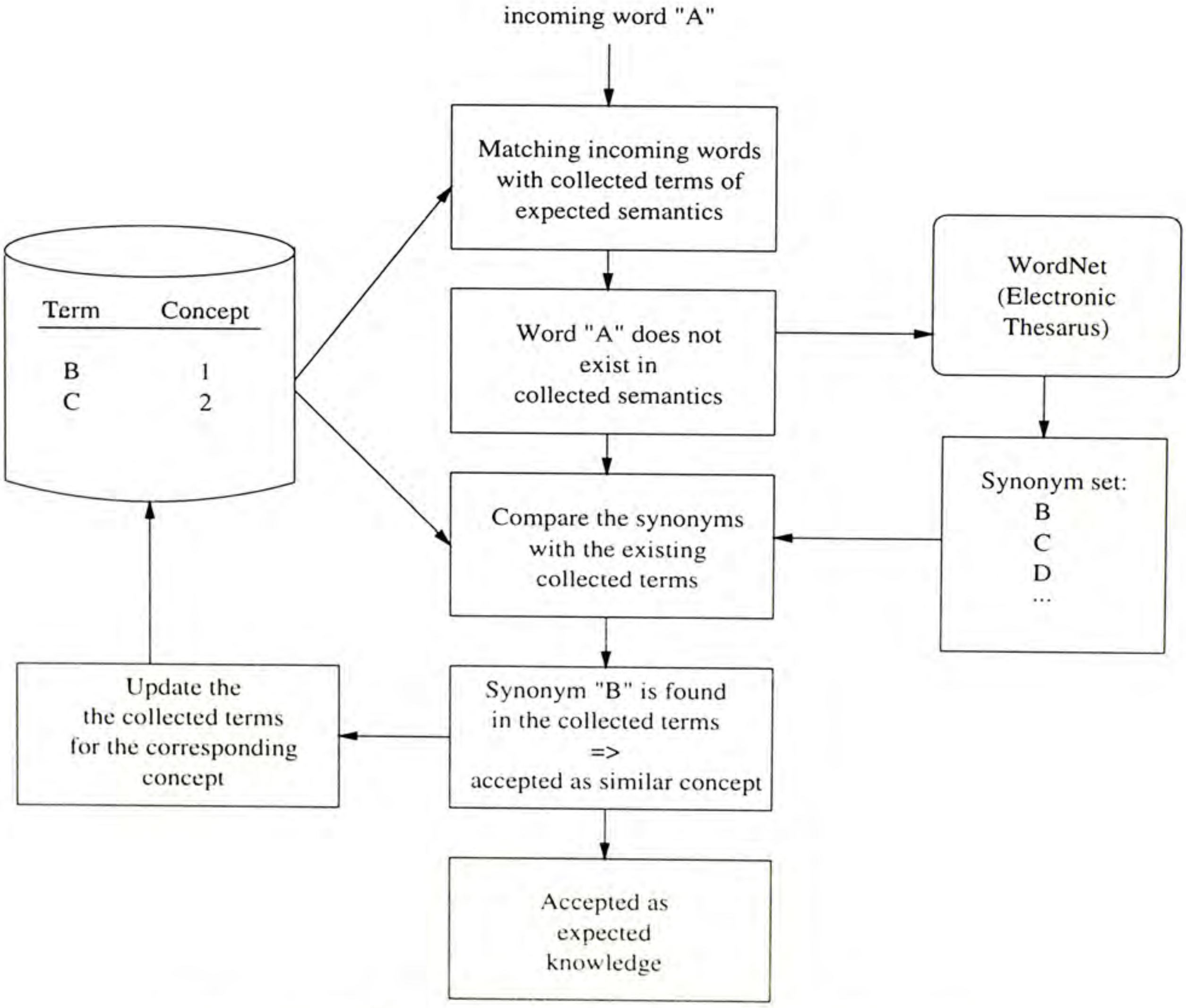
incoming word "A"
Matching incoming words with collected terms of expected semantics
Term
Concept
B
1
C
2
Word "A" does not exist in collected semantics
WordNet (Electronic Thesarus)
Synonym set:
B
C
D
...
Compare the synonyms with the existing collected terms
Update the the collected terms for the corresponding concept
Synonym "B" is found in the collected terms => accepted as similar concept
Accepted as expected knowledge

Figure 4.2: Processing with the thesaurus, WordNet

4.2 Pattern Discovery

For causation relation, a cause and an effect are usually not only simple noun phrases, they are more complex. The objective of pattern discovery in SEKE is to flexibly generate the patterns for the effects and causes automatically. The pattern discovery stage can automatically generate extraction patterns and update the support for each pattern. For each sentence where its causation knowledge is being extracted successfully, patterns for the reason and the consequence are generated. The newly generated patterns are then compared and combined with existing patterns. At the discovery step, a set of patterns with their corresponding support values are created.

Fig. 4.3 shows the process for pattern discovery. The detailed descriptions of the steps in the pattern discovery process are given in the subsequent sections.

4.2.1 Use of Semantic Expectation-based Knowledge Extraction

The pattern discovery process makes use of previously extracted causation knowledge. In SEKE, causation semantic template, sentence templates, consequence template and reason template are the expected semantics. With these expected semantics, causes and effects are extracted.

We use those previously extracted causes and effects for generating extraction patterns. For example, the reason template for the Hong Kong stock market movement is composed of a factor and a movement. For a successfully extracted reason, it will be a pair of terms referring to the factor and

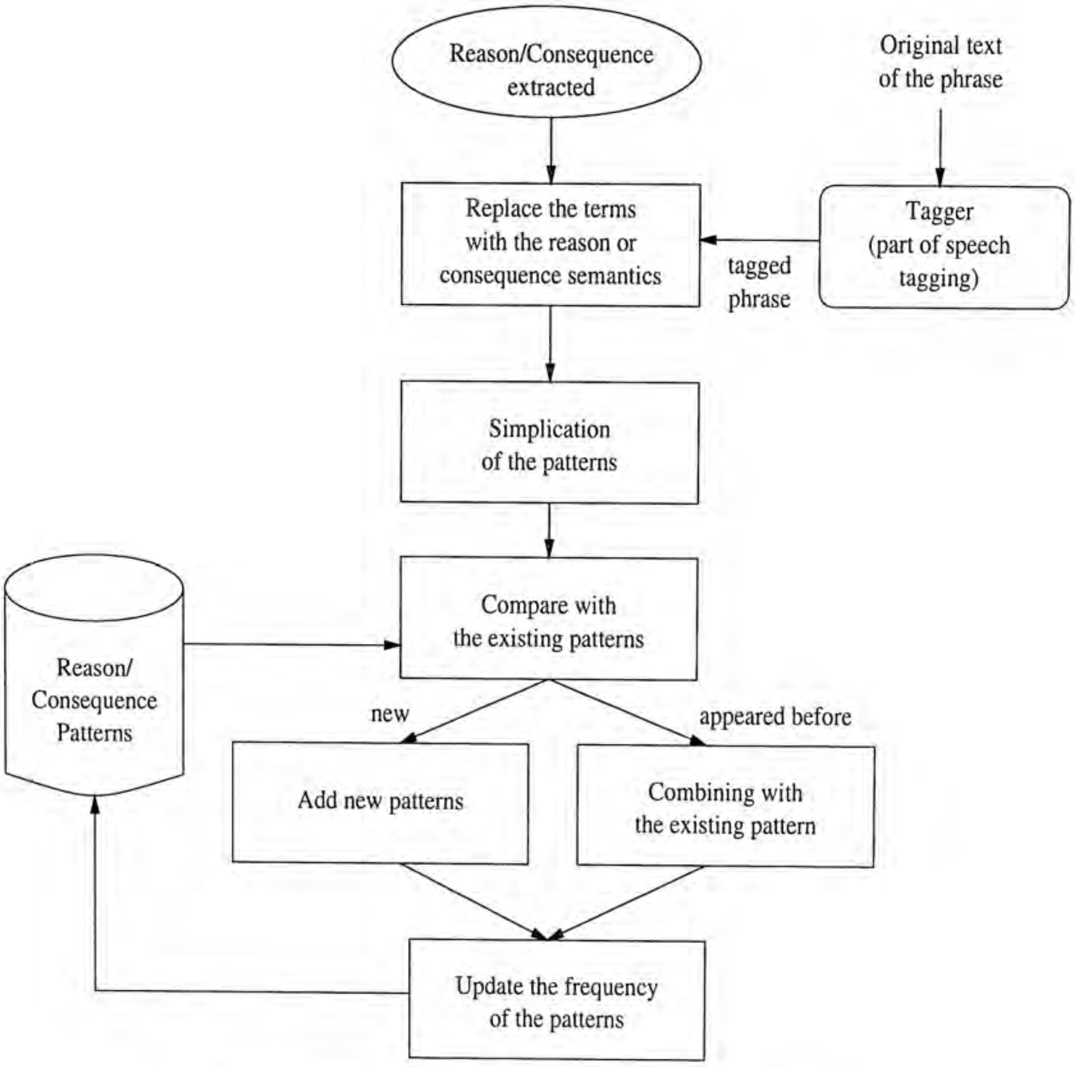
Reason/Consequence extracted
Original text of the phrase
Replace the terms with the reason or consequence semantics
tagged phrase
Tagger (part of speech tagging)
Simplication of the patterns
Compare with the existing patterns
Reason/ Consequence Patterns
new
appeared before
Add new patterns
Combining with the existing pattern
Update the frequency of the patterns

Figure 4.3: The pattern discovery process

the movement. This pair of terms can also be regarded as a factor-movement relation. We observed that there are some kinds of regularities in expressing that relation. Therefore, we have explored these regularities and developed an automated procedure to capture the extraction patterns.

4.2.2 Use of Part of Speech Information

There are many different styles in expressing a relation in text. To capture the regularities is to discover those different structures in expressing the relation. Therefore, linguistic information is useful in the construction of the patterns. In SEKE, we make use of a transformation-based part of speech tagger [4] to provide linguistic information. The part of speech tags adopted by the tagger are listed in the Appendix A.

4.2.3 Pattern Representation

The followings are symbols used in patterns. They are referred to as elements in a pattern.

1. Concept labels, in the following form: *[concept_label]/syntactic_tags*.
   Concept label represents the expected concept in the pattern.
   Example: *[factor]/NN*, which represent the expected concept is the "factor" with the syntactic tag of a noun.

2. Sample labels, in the following form: *(sample_terms)/syntactic_tags*.
   Sample terms in the sample label can be a list of words which appeared in different samples of the same pattern.
   Example: *(in,of)/IN*, which shows that the appeared terms are the perpositions,"in" and "of".

3. Syntactic tags are the expected syntactic information of the label. It can be a list of tags. For example, "/JJ/NN" means that the expected terms corresponding to the label should include an adjective followed by a noun.

   All the syntactic tags in the pattern will be only in their base forms, meaning that for a verb, even it is tagged as VBD(verb in past-tense), it will be expressed in the base form(VB) in the pattern. Moreover, for consecutive tags which are the same, they are grouped and were given a single syntactic tag.

4. "*", a wild card character which can match any of the terms with secondary speech tags. Secondary speech tags are those considered as modifiers in grammar such as determinants, adjectives, adverbs and so on.

4.2.4 Constructing the Patterns

For each sentence where its causation knowledge is extracted successfully, some patterns are constructed, one for the reason and one for the consequence. The extraction pattern is generated from:

- the original phrase,
- the extracted terms with their corresponding concepts,
- the tagged sentence or phrases.

Details of the pattern construction will be illustrated with the following sentence as an example:

> *Hong Kong stocks closed higher helped by an overnight rise in the Wall Street.*

Procedure 1: The extracted terms in the tagged phrases are generalized by **concept labels**, and their corresponding tags are transformed into **syntactic tags**. Using the above example for illustrations:

- Reason: *"an overnight rise in the Wall Street"*

  Extracted semantics: **Factor** *"Wall Street"*
  Movement *"rise"*

  Tagged phrase:

  an/DT overnight/JJ rise/NN in/IN the/DT Wall/NNP Street/NNP

  After procedure 1:

  an/DT overnight/JJ [movement]/NN in/IN the/DT [factor]/NN

- Consequence: *"Hong Kong stocks closed higher"*

  Extracted semantics: **Hong Kong stocks** *"Hong Kong stocks"*
  Movement *"higher"*

  Tagged phrase:

  Hong/NNP Kong/NNP stocks/NNS closed/VBD higher/JJR

  After procedure 1:

  [HongKongStocks]/NN closed/VBD [movement]/JJ

As both *"Wall/NNP Street/NNP"* and *"Hong/NNP Kong/NNP stocks/NNS"* are noun phrases. Therefore they are represented by *[factor]/NN* and *[HongKongStocks]/NN* respectively. After procedure 1, "Wall Street" and "Hong Kong stocks", the extracted terms of the concept "factor", are transformed into the concept labels, *[factor]/NN*. As "rise" is tagged as a noun(NN) and "higher" is tagged as an adjective(JJ), they are transformed in to the concept labels, *[movement]/NN* and *[movement]/JJ* respectively. Hence, both of the transformed phrases contain two concept labels.

Procedure 2: Replace the remaining terms having secondary speech tags with the wildcard "*". Again, using the above example, "overnight" and "the", are tagged as an adjective(JJ) and a determinant(DT) which are regarded as secondary speech tags. After procedure 2, they are replaced with the wild card, "*".

- Reason:
  Before procedure 2:
  an/DT overnight/JJ [movement]/NN in/IN the/DT [factor]/NN
  After procedure 2:
  * [movement]/NN in/IN * [factor]/NN
- Consequence:
  Before procedure 2:
  [HongKongStocks]/NN closed/VBD [movement]/JJ
  After procedure 2:
  [HongKongStocks]/NN closed/VBD [movement]/JJ

Procedure 3: For the remaining terms, they are considered as **sample labels**. For example, the preposition(IN), "in", is the remaining term, and it is transformed into the **sample label**, *(in)/IN*.

- Reason:
  Before procedure 3:
  * [movement]/NN in/IN * [factor]/NN
  After procedure 3:
  * [movement]/NN (in)/IN * [factor]/NN

- Consequence:

  Before procedure 3:

  [HongKongStocks]/NN closed/VBD [movement]/JJ

  After procedure 3:

  [HongKongStocks]/NN (closed)/VB [movement]/JJ

After the above procedures, two patterns, one for reason, and one for consequence, are constructed:

- Reason:

  * [movement]/NN (in)/IN * [factor]/NN

- Consequence:

  [HongKongStocks]/NN (closed)/VB [movement]/JJ

4.2.5 Merging the Patterns

Since many different patterns may be generated, similar patterns should be merged to reduce the number of patterns. The followings are rules for comparing the patterns:

- The wildcards(*) and terms in **sample labels** are ignored in the comparison.

- Only the **concept labels** and the syntactic tags of **sample labels** are compared.

For example, consider the following two patterns:

1. * [movement]/NN (in)/IN * [factor]/NN

2. * [movement]/NN (of)/IN * [factor]/NN *

They are both in the form of *"[movement]/NN ()/IN [factor]/NN"* by the above rules. Therefore, they are regarded as the same pattern. If two patterns are considered to be the same after the comparison, they are merged into one pattern. The combined pattern retains the characteristics of both patterns. This is done by retaining and combining the **sample labels** and *****(wild cards). The merged pattern for the above example is:

* [movement]/NN (in, of)/IN * [factor]/NN *

For each newly generated pattern, it is compared with existing patterns, and two sets of patterns, one for the reason and one for the consequence are generated automatically.

4.3 Pattern Matching

By pattern matching, unseen reasons and consequences could be discovered automatically. A phrase which is identified semantically as a reason or a consequence is matched with the corresponding patterns. For example, a reason phrase is matched against the set of generated reason patterns. A phrase may be matched with more than one pattern. To decide which pattern is used among the candidate patterns for extraction, two factors are considered:

1. The matching score which evaluates how well a phrase is matched with a particular pattern.

2. The support of the generated patterns.

An overall score for each pattern is then computed from the above factors using the following formula:

$$C(P_i) = wM_i + (1 - w)S_i \tag{4.1}$$

where $C(P_i)$ is the overall score for pattern P_i, M_i is the matching score for P_i, S_i is the support for P_i. w is a weight parameter controlling the relative importance of one factor to another.

The candidate pattern with the highest overall score is selected, and the reason or the consequence text fragment is extracted by the pattern. However, the confidence of the extracted knowledge is also affected by the relevancy of the sentence templates. A confidence value is calculated for the knowledge discovered by the patterns as follows:

$$F(K) = C(P_{\text{selected}})R(T_k) \tag{4.2}$$

where $F(K)$ is the confidence of the knowledge K discovered by pattern P_{selected} and processed by sentence template T_k. $R(T_k)$ is the relevancy of the sentence template T_k used in semantic parsing of the corresponding sentence.

If the knowledge discovered has a low confidence, it means that it is more likely to be irrelevant information. Therefore, with the computed confidence of the knowledge discovered by patterns, we can eliminate those output having a low confidence by setting a threshold, H. As a result, only those discovered knowledge with a confidence larger than a threshold, H, will be regarded as relevant causation knowledge.

In the following sections, detailed descriptions of the three factors, namely the matching score, the support of patterns, and the relevancy of sentence templates will be presented.

4.3.1 Matching Score

The matching score is used to measure how well the phrase is matched against a pattern by evaluating the similarity of elements in the phrase and the pattern.

Here are some considerations for computing the matching score:

1. An element in the phrase can either be a concept label or a word with its tags, in the form of (term)/tag, e.g. *(economy)/NN*.
2. For different elements in the phrase, different element weights, $s(j)$, are assigned in matching with the elements in the pattern.
3. The maximum score for each element in a phrase is 1.
4. If the same term is appeared in both the phrase and the pattern, the similarity is higher.

Firstly, for a reason or consequence phrase, if any concept of the reason or consequence is already extracted, it is processed in the same way as the procedure 1 in Chapter 4.2.4. The extracted terms in the tagged phrases are generalized by **concept labels**, and their corresponding tags are transformed into **syntactic tags**.

The element weight for each element j in the input phrase, $s(j)$, is assigned as follows:

- $s(j) = 1.0$,

  if element j is the same as the corresponding concept label in the pattern.

- $s(j) = \mu_1$,

  if the **tag** of element j is the same with the corresponding syntactic tag in the **sample labels** and the term of element j is within the list of sample terms in the **sample labels**.

- $s(j) = \mu_2$,

  if the **tag** of element j is the same with the corresponding syntactic tag in the **sample term**.

- $s(j) = \mu_2$,

  if the element j is a secondary speech tag and is matched with the corresponding "*" in the pattern.

- $s(j) = 0$, otherwise

The only constraint for μ_1 and μ_2 is: $0 < \mu_2 < \mu_1 < 1.0$. In our experiments, we set μ_1 to 0.8 and μ_2 to 0.5.

The matching score, M_i, of pattern, P_i, is defined as:

$$M_i = \frac{\sum_{j=1}^{j=n_i} s(j)}{n_i} \tag{4.3}$$

where $s(j)$ is the element weight for element j, and n_i is the number of elements in the phrase. The following is an example of computing the matching score.

Example: For the input reason phrase:

(a)/DT (weak)/JJ (performance)/NN (in)/IN (Dow)/NNP (Jones)/NNP (Industrial)/NNP (average)/NNP

Examples of pattern:

P_1 * [movement]/NN (in, of)/IN * [factor]/NN *

P_2 * [factor]/NN [movement]/VB *

The total number of elements in the input phrase is 8. Therefore,

$$M_1 = \frac{\sum_{j=1}^{j=8} s(j)}{8} = 0.54$$
$$M_2 = \frac{\sum_{j=1}^{j=8} s(j)}{8} = 0.125$$

4.3.2 Support of Patterns

Some patterns appear more often than the others indicating that the pattern has a higher confidence or support. The support, S_i of the pattern P_i is measured by the normalized frequency of the patterns. During the generation of patterns, new patterns are collected and the frequency of the occurrence for each pattern is recorded. It is defined as:

$$S_i = \frac{f_i}{\max(f)} \tag{4.4}$$

where f_i is the frequency of the pattern P_i and $\max(f)$ is the maximum frequency among the set of patterns.

The support of patterns is useful in evaluating the matched pattern. If the M_i of a phrase for matching two patterns is the same, the pattern having a higher S_i is a better choice because it is more likely to generate the correct information.

4.3.3 Relevancy of Sentence Templates

A sentence template is used for identifying the existence of a causation relation in a sentence. It is also used for parsing the sentence into a reason phrase and a consequence phrase. It consists of causation expression linking a consequence with a reason. It is possible that even though a sentence is matched with a certain sentence template, it does not contain a causation relation. Therefore, we estimate the relevancy of a sentence template by the probability of the sentence matching the sentence template to be a causation relation by computing the ratio:

$$R_k = \frac{f_{\text{relevant}}(k)}{f(k)} \tag{4.5}$$

where $f_{\text{relevant}}(k)$ refers to the frequency that a sentence is activated by the sentence template k to be relevant to causation knowledge. $f(k)$ refers to the frequency that the sentence template k appears.

4.4 Applying the Newly Discovered Patterns

This section describes how pattern discovery and pattern extraction procedures are incorporated into the basic framework of SEKE leading to a complete framework of SEKE. It also describes how new concept terms are generated automatically for a lexicon.

For every successfully extracted knowledge by SEKE, they will be passed to the pattern discovery stage, while those incomplete ones will be passed to the knowledge discovery stage. The knowledge extracted by this stage will also be accepted as part of the causation knowledge and as new concept terms for the expected semantic lexicons.

A phrase will be processed for the discovery of knowledge after semantic processing under the following conditions:

- It is identified as a reason or a consequence in the semantic processing stage of SEKE, but the knowledge extracted is incomplete, or,
- It is identified as a reason or a consequence in the semantic processing stage of SEKE, but the extraction of the reason or consequence has failed.

Generated patterns are used in the knowledge discovery stage. Fig. 4.4 shows the procedures for knowledge discovery from patterns.

1. The reason or consequence phrase will be processed according to the details in Chapter 4.3. For example: *"(a)/DT (weak)/JJ (performance)/NN (in)/IN (Dow Jones Industrial average)/NN"*,

2. Among all the candidate reason patterns or consequence patterns, select the one with the highest score.

3. With the selected pattern, reason or consequence concepts can be identified from the phrase. For example, the pattern *"* [movement]/NN (in, of)/IN * [factor]/NN *"* is selected for the phrase, *"(a)/DT (weak)/JJ (performance)/NN (in)/IN (Dow Jones Industrial average)/NN"*. The extracted reason is *"weak performance"*, and the extracted consequence is *"Dow Jones Industrial average"*.

4. A confidence value is computed for each of the identified concept. Only those with a confidence value higher than a pre-specified threshold, H, is regarded as part of the causation knowledge and is extracted.

5. The extracted knowledge is combined with those previously extracted by SEKE as a causation relation.

6. The discovered knowledge can be inserted into the expected semantic lexicons automatically as new concepts.

7. Optionally, human verification for the discovered knowledge can be incorporated.

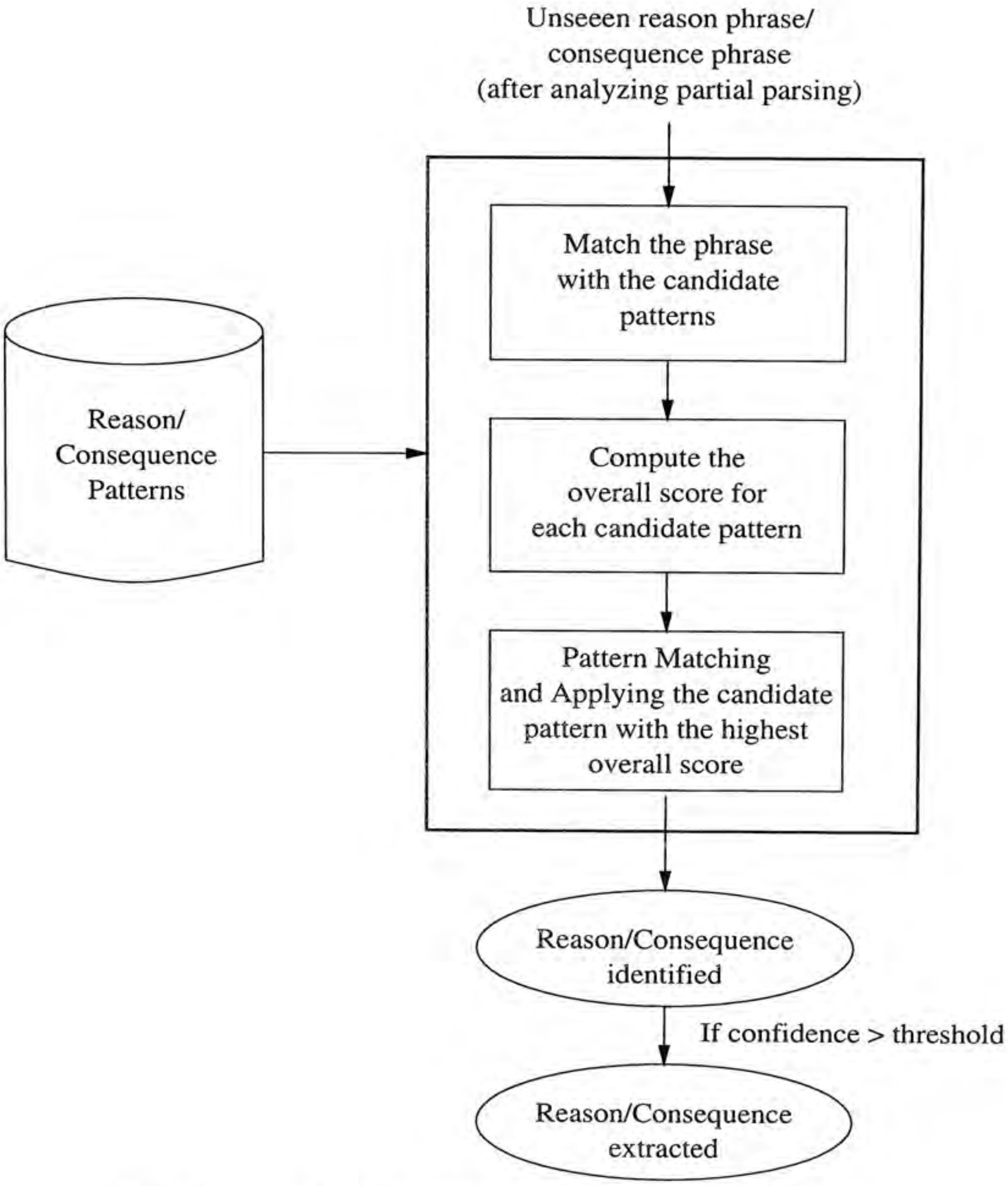

Figure 4.4: Applying the newly discovered patterns

Chapter 5

Applying SEKE on Hong Kong Stock Market Domain

As the framework of SEKE is domain independent, it can be applied to different domains. Two studies using the SEKE framework are carried out for two different domains. English news articles are collected for the studies. This chapter describes the investigation of SEKE in extracting causation knowledge for the Hong Kong stock market movement. The causation semantics in this domain are the reasons affecting the movement of Hang Seng Index (HSI) in the Hong Kong stock market. First, we will describe the tasks including template design, sentence screening, semantic processing, and pattern discovery. Then, we will present the experimental results.

News articles are provided by Reuters newsfeed. A total of 730 Hong Kong stocks related news articles between December 2001 to mid-April 2002 were collected as training data. News articles from mid-April to May 2002 were used as the testing data set. The testing set includes 365 pieces of Hong Kong stocks related news.

5.1 Template Design

From the training data, we analyze the sentences expressing Hong Kong stock market movements with their influencing reason, and design the templates.

5.1.1 Semantic Templates

The causality relation is about how the movements of some factors affect the Hong Kong stock market movement. The movement is mainly measured by Hang Seng Index(HSI). Thus, the causation semantic is composed of one or more factors with movements and the occurrence of the Hang Seng Index movement. Causation relation may also exist in the reasons and consequences themselves. Therefore, the semantic templates for the complex reason and consequence are also defined. The causation semantic templates for causation relation of a complex reason states that a reason which is a factor with movement causes the occurrence of another factor with movement. For the complex consequence, it can be either of the following two semantic templates. The first one is composed of one or more factors with movements and the occurrence of the Hang Seng Index movement. The other one states that a factor with movement causes the occurrence of another factor with movement.

5.1.2 Sentence Templates

Based on the observation of news articles in the training set, a set of sentence templates are designed. They are listed in Table 5.1.

Since causation sentences in the Hong Kong stock market movement domain do not only include simple structure but also complex structure, two

	Templates	
[consequence]	because	[reason]
[consequence]	after	[reason]
[consequence]	as	[reason]
[consequence]	ahead of	[reason]
[consequence]	due to	[reason]
[consequence]	thanks to	[reason]
[consequence]	with	[reason]
[consequence]	by	[reason]
[consequence]	following	[reason]
[consequence]	tracking	[reason]
[reason]	helping to	[consequence]
[reason]	help	[consequence]
[reason]	dragging	[consequence]
[reason]	push	[consequence]
[consequence]	on	[reason]
	Multiple Reason Template	
[reason]	and	[reason]
[reason]	,	[reason]

Table 5.1: Causation sentence templates for the Hong Kong stock market movement domain

features are associated with the sentence templates. Firstly, as there may exist more than one reasons in the causation relation, each sentence template is associated with multiple reason templates for handling of the multiple-reasons sentences. Secondly, reasons and consequences themselves can be a causation relation and they are referred to as complex reasons and consequences. Therefore, causation structure may occur recursively within the reasons or consequences. Causation sentence templates are matched recursively to the sentences for complex consequences or complex reasons. Examples of associating the sentence templates with multiple reason templates and recursively applying the templates are shown below:

Simple Sentence

Reason causes *Consequence.*

Consequence is caused by *Reasons.*

Multiple-reason Sentence

Consequence Multiple Reasons

Consequence is caused by $Reason_1$ and $Reason_2$.

Complex Sentence

Reason Complex Consequence

$Reason_1$ causes *Consequence* are caused by $Reason_2$.

Consequence Complex Reason

Consequence is caused by $Reason_1$ caused by $Reason_2$.

5.1.3 Consequence and Reason Templates:

In the causation relation, reasons affecting the performance of Hang Seng Index could be the performance of other stock markets, other stocks, other financial

Factor	**Concept Terms**
Overseas Market	Wall Street, U.S. , Asian markets, overseas markets, Nasdaq China issues
Interest rates	U.S. interest rates, interest rates
Individual stocks	HSBC , Cathay Pacific, China Mobile, Juniper Networks, Johnson, Motorola China's two cellular phone operators, Hutchison, China Shares, China Unicom, China telecoms Henderson, Legend Holdings, telcos
Financial sectors y	property, technology, moribund real estate market, Japanese yen, US sales data margins, window dressing, earnings, retail sales, banks, flat sales, profit-taking, oversupply
Economic	economy, economic downturn
Others	holidays, cyclicals, investors, pressure

Table 5.2: Initial lexicon for "factor" in the Hong Kong stock market movement domain

instruments, the actions of investors, the government, and so on. Therefore, the consequence template refers to Hang Seng Index with movement. The reason template refers to "factor" with "movement". The concept "movement" is common to both consequences and reasons and can be divided into four categories. The categories are upward movement, downward movement, no change and activity. SEKE requires the use of initial lexicons for capturing each concept in the reason and consequence templates. The initial lexicons for "factor" and "movement" are listed in Tables 5.2 and 5.3 respectively.

Movement	Concept Terms
upward	higher, gain ground, gain, rise, rose, up, boosted, gaining, rally, boost, go up, surge, recovery, reversing earlier losses, reversed early losses
downward	lower, down, cut, losses, loss, drop, falls, fall, slipping, slid, plunge, sink, cuts, lost ground, slip, fell, weak, slimmer, falling, slipped, slide, weaker, weakness, weakening, drop, decline, decrease
no change	steady, tight range, little changed, flat, range-bound, consolidating, mixed performance, mixed, consolidate, stabilize, unchanged
activity	fear, hope, concern, profit woe, bottoming out, lock in profits, cautious, bailed out, sell, emerged, weigh, concerned about, worries, sideline, worrying, worried about, suffer, lack of clear signs, further, concerned over, worried, revitalize, eye, lend support, lends support, shrugged off negative news, brisk, plague

Table 5.3: Initial lexicon for "movement" in Hong Kong stock market movement domain

5.2 Pattern Discovery

We aim to discover new factors in affecting the performance of Hang Seng Index. Hence, we focus on discovering patterns for extracting reasons only. With the designed templates, news articles are fed into SEKE for the extraction of causation knowledge, which is then used for discovering the patterns.

5.2.1 Support of Patterns

In the Hong Kong stock market movement domain, we observe that each reason consists of a factor and a movement. Therefore, a pattern is generated from a reason phrase if both factor and movement are identified. During the discovery of patterns, the number of occurrence of each pattern is recorded for computing the support of patterns. The 12 most frequent patterns discovered are shown in Table 5.4.

5.2.2 Relevancy of Sentence Templates

We estimate the relevancy of a sentence template by computing the ratio in Equation 4.5. The relevancy of the sentence templates is shown in Table 5.5.

5.3 Causation Knowledge Extraction Result

After the automatic sentence screening of testing data set, 365 relevant news articles were collected and 774 sentences were identified to be related to Hong Kong stock market.

Frequency	Patterns
43	* [factor]+[up/down/no_change/activity]/NN *
23	* [up/down/no_change/activity]/NN (about,from,on,in,of)/IN * [factor]+(stocks,tech+board,Holdings)/NN *
10	[factor]/NN [up/down/no_change/activity]/VB *
6	* [up/down/no_change/activity]/NN (in,on)/IN (China)+[factor]/NN * (evaporated,triggered,were,are)/VB
4	* [up/down/no_change/activity]/NN (from)/IN * [factor]/NN * (week,JNPR.O)/NN *
4	* (renewed,lingering,expected)/VB (quarter+percentage+point)+[up/down/no_change/activity]/NN (over,that,in)/IN * [factor]/NN *
4	(worries,concerns)/NN (over)/IN * [factor]/NN (erased,capping)/VB * [up/down/no_change/activity]/NN
4	[up/down/no_change/activity]/JJ [factor]+(consumer+confidence+data)/NN (and)/CC (remained+focused)/VB
3	* (record+drop)/NN (in)/IN * (sales)/NN (in)/IN * [factor]/NN (aggravated,exacerbated)/VB [up/down/no_change/activity]/NN (that)/IN * (territory)/NN ('s)/PO * (trading+partner)/NN (may)/MD (be+mired)/VB (in)/IN * (slump,slowdown)/NN *
3	[up/down/no_change/activity]/NN (over)/IN (its)/PR (exposure)/NN (to)/TO * [factor]/NN *
3	(investors)/NN (chose)/VB (to)/TO (lock)/VB (in)/IN [factor]/NN (from)/IN * [up/down/no_change/activity]/NN
3	(investors)/NN [up/down/no_change/activity]/VB [factor]/NN * (0941.HK)/CD *

Table 5.4: Examples of patterns discovered for Hong Kong stock market movement domain

Relevancy	Templates		
1.00	[consequence]	because	[reason]
0.79	[consequence]	after	[reason]
0.71	[consequence]	as	[reason]
0.73	[consequence]	ahead of	[reason]
1.00	[consequence]	due to	[reason]
1.00	[consequence]	thanks to	[reason]
0.68	[consequence]	with	[reason]
0.43	[consequence]	by	[reason]
1.00	[consequence]	following	[reason]
1.00	[consequence]	tracking	[reason]
1.00	[reason]	helping to	[consequence]
1.00	[reason]	help	[consequence]
1.00	[reason]	dragging	[consequence]
1.00	[reason]	push	[consequence]
0.34	[consequence]	on	[reason]
	Multiple Reason Template		
	[reason]	and	[reason]
	[reason]	,	[reason]

Table 5.5: Relevancy for the causation sentence templates in the Hong Kong stock market movement domain

5.3.1 Evaluation Approach

In order to evaluate the extraction performance, we manually examine the testing data and extract the causation knowledge. The causation knowledge discovered by SEKE will be compared with the items extracted manually.

The performance for the extraction of knowledge is evaluated using three performance metrics, namely the precision, recall and F-measure to measure the effectiveness of the system.

Recall, R = the number of correct slots filled by the system divided by the number of slots filled by human analyst

Precision, P = the number of correct slots filled by the system divided by the total number of slots filled by the system

F-measure, $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$

The β in F-measure is used for controlling the relative importance of recall and precision. In our experiments, β is set as 1 as we treat the recall and precision as in equal weight for combining the two metrics.

5.3.2 Parameter Investigations

Before carrying out the experiment on the testing data set, we investigated on finding the optimum value for the parameters. The optimum value for the parameters will be applied for the experiment of the testing data set.

The parameters used for the use of pattern discovery in SEKE are:

1. the weight parameter, w, in equation 4.1;

2. the threshold, H.

Weight parameter, w	**No. of patterns**
1.0	97
0.9	91
0.8	63
0.7	32
0.6	4
0.0 to 0.5	1

Table 5.6: Number of different patterns applied for different values of the weight parameter, w

The decision is made by observing the effect of the parameters. It is done by setting different values of parameters, and analyzing the corresponding performance of the training data set.

Fig. 5.1 shows the extraction performance of the training data set with different values of the weight parameter, w. As w increases, the F-measure value (with β=1) increases. It becomes stable when w reaches 0.8. Hence for the weight parameter, the value 0.8 is chosen.

The weight parameter, w, is to control the relative importance of the two factors in computing the overall score for pattern matching. The chosen value of 0.8 indicates that the matching score is a more important factor than the support of the patterns. This can be explained by a more detailed analysis on examining the patterns applied during pattern matching. Table 5.6 shows the number of different patterns applied for different values of the weight parameter, w, on the training data set. When w ranges from 0 to 0.6, the number of discovered patterns selected is very limited, only those with the highest support value are chosen. Usually, these patterns could not discover the correct knowledge. This is also the reason for the stable trend when w is less than 0.6.

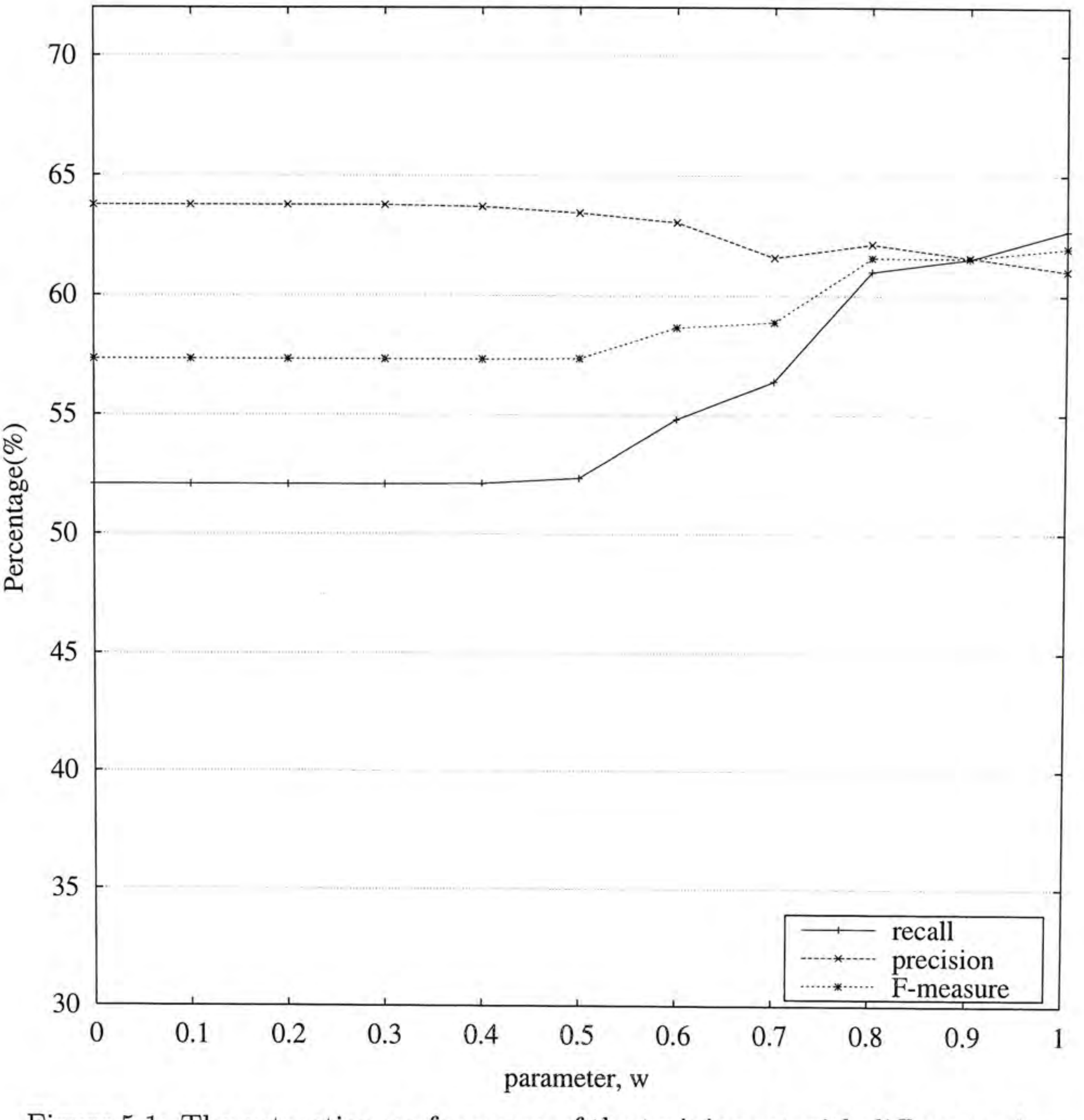

Figure 5.1: The extraction performance of the training set with different values of the weight parameter, w

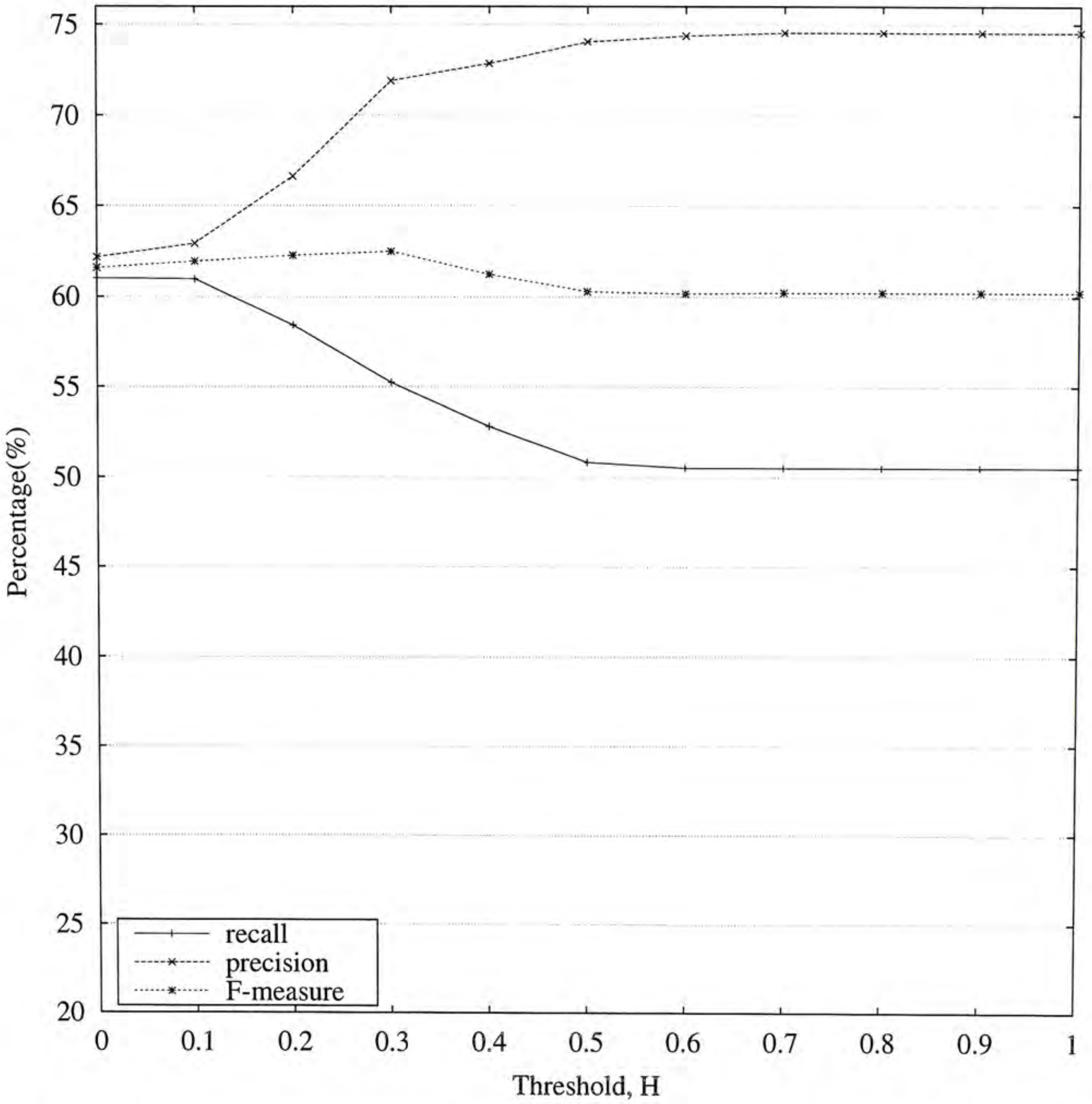

Figure 5.2: The extraction performance of the training set with different values of threshold, H.

The threshold, H, discussed in 4.3. is used to prune the information extracted by patterns. Fig. 5.2 shows the extraction performance of the training data set with different values of the threshold, H. From the figure, the value of F-measure (β=1) is maximum when H equals to 0.3, hence, the value 0.3 is chosen for the threshold, H.

With H increases, the value of F-measure (β=1) slightly increases. This is due to the filtering out of irrelevant knowledge. However, when H is larger than 0.3, the value of F-measure decreases, as much of the relevant information is filtered out. The trend finally stops at 0.6, as 99% of the discovered knowledge have confidence value smaller than 0.5. The distribution of the confidence value for the discovered unseen knowledge is shown in Fig. 5.3. Setting H at 0.6 means filtering out all the knowledge discovered by patterns.

5.3.3 Experimental Results

Recall that the basic framework of SEKE includes three stages, namely template design, sentence screening, and semantic processing. The complete framework of SEKE includes two additional stages, namely using a thesaurus and pattern discovery. In the stage of pattern discovery, 166 patterns are generated automatically from 299 reasons. For the experiment on the complete framework, the weight parameter, w, is set to 0.8 and the threshold, H is set to 0.3 according to the parameter investigation. The performance of SEKE is shown in Table 5.7.

With the use of pattern discovery and thesaurus, the complete framework of SEKE is able to extract 16% more causation knowledge than the basic framework. Despite of the drop of the precision by 10%, the F-measure is still

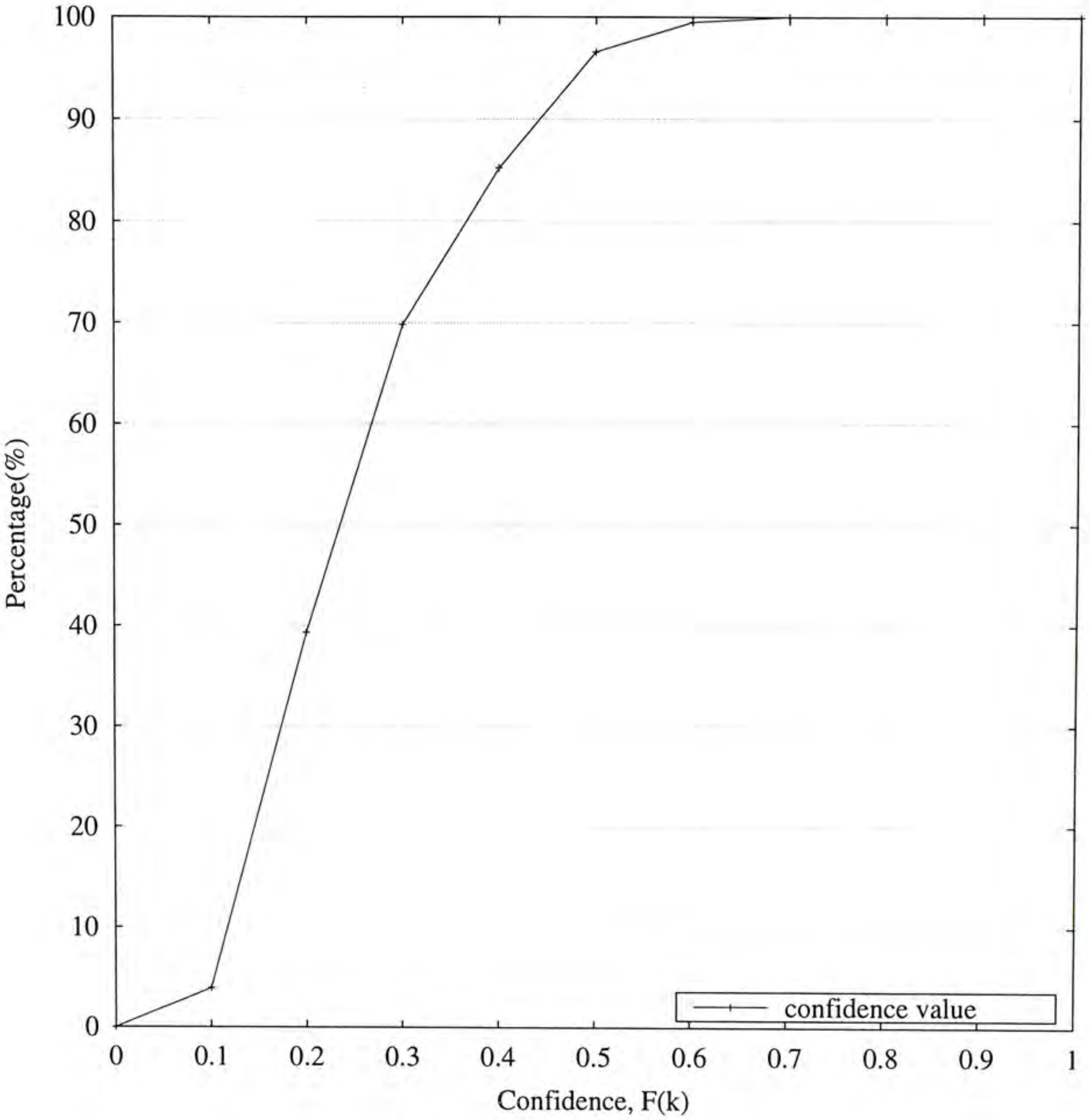

Figure 5.3: The accumulative distribution of confidence, $F(k)$, for the discovered unseen knowledge

	recall	precision	F-measure($\beta = 1$)
basic framework	30.1%	82.0%	43.9%
complete framework	45.9%	71.7%	56.0%

Table 5.7: Experimental results of SEKE in the Hong Kong stock market movement domain

Discovered Factors
equities/ equity markets
Argentina's economic problems
Industry/ industry competition
export picture/ exporters
sector outlook
debt loads
terrorism threats
Overseas market: London
China Resources

Table 5.8: Unseen reasons discovered of SEKE in the Hong Kong stock market movement domain

12% higher. This shows that SEKE can discover unseen reasons successfully. Some unseen reasons discovered are depicted in Table 5.8. The decrease in precision is due to the fact that some irrelevant information are extracted with the use of patterns for discovery of reasons.

However, both the recall and precision cannot reach 100%. From the following examples, we could observe some typical errors.

For the complete framework of SEKE, the errors in precision indicate that some information extracted are incorrect:

- Not able to be identified as similar concepts by the thesaurus.

  "HK stocks seen pulling back after Wednesday's rally." (18APR2002)

  In this example, the movement "pulling back" cannot be identified as similar concepts by the thesaurus.

- Extracting the wrong movement

  "HONG KONG, May 7 - Hong Kong stocks are expected to open lower on Tuesday after losses on Wall Street raised concern about the short-term direction of global equities." (07MAY2002)

The wrong movement “raised” is extracted for the factor “Wall Street”, which the correct one should be “losses”.

- Extracting a irrelevant reason due to the ambiguity of the sentence templates.

  “HONG KONG, May 22 - Hong Kong stocks were steady in early trade on Wednesday, with investors taking a breather after sharp losses in the previous trading session.” (22MAY2002)

  The extracted reason, “losses” of “trading session”, which is just a previous event to the Hong Kong stock and not a true reason.

For the complete framework of SEKE, the errors in recall indicate that some relevant information are not extracted, which is due to the problem that no generated pattern is matched.

- *“HONG KONG, April 17 - Hong Kong stocks opened higher on Wednesday after Wall Street logged its biggest gain in seven months on some upbeat corporate forecasts.” (17APR2002)*

  The reason “some upbeat corporate forecasts” cannot be identified as no pattern is matched for the phrase “some/DT upbeat/JJ corporate/JJ forecasts/NNS ./.”

5.3.4 Knowledge Discovered

SEKE is able to discover causation relations described explicityly by the authors of the news articles. We present some examples of discovered causation knowledge in the Hong Kong stock market domain. They are the information extracted from the text of the testing data set by SEKE.

Table 5.9 shows the reasons identified in the news articles in the testing data set for causing Hong Kong stock market to go upwards. The first row in the table shows that "Wall Street" is one of the factor for causing the Hong Kong stock market to go upwards. 4.5% of the extracted reasons for Hong Kong stock going upwards is caused by Wall Street with an upward movement and 4.5% is caused by the downward movement of Wall Street. 2.3% of the extracted reasons states "Wall Street" as the reason without the mentioning of movements. The percentage in the reasons of the rise(upward movement) and the fall(downward movmenet) of Wall Street are the same. This unclear situation is due to the errors of extracting the wrong movement occured in SEKE. However, we could still see from the total percentage that, the Wall Street factor accounts for 11.3% among all the reasons, and hence it is a very important reason in affecting the Hong Kong stock market.

The last several rows in the table show some multiple-reasons. For example, the factor, "Wall Street", together with the current economic situation causes the Hong Kong stock market to move upwards. Another multiple-reason is the combined effect of the activity of the property sector with the current economic situation. Each of the multiple-reasons account for 1.1% of all the reasons.

Next, we discuss the causes for the downward movement for the Hong Kong stock market in Table 5.10. 13.5% of those causes belongs to the fall(a downward movement) of Wall Street. In total, 15.7% of the reasons extracted for causing the Hong Kong stock market to go down consist of the factor of Wall Street. The downward movement of Wall Street, internal factors such as the concern(an example of activity) on some corporate earnings, cause the Hong Kong stock market to go downwards.

	Movement (%)					
Factors	**up**	**down**	**no change**	**activity**	**no movement**	**total**
Wall Street	4.5	4.5			2.3	11.3
property sector				1.1	8.0	9.1
economic situation	4.5			2.3	2.3	9.1
HSBC	2.3	1.1			3.4	6.8
US	3.4			1.1	1.1	5.6
interest rate			0.6	1.1		1.7
Cathay Pacific	0.6					0.6
corporate forecasts					1.1	1.1
telecom stocks					1.1	1.1
industry competition				1.1		1.1
futures	1.1			1.1	1.7	3.9
corporate earnings	1.1					1.1
Hutchison				1.1		1.1
HK Chinese Ltd.		1.1				1.1
China Resources		1.1				1.1
banks				1.1	1.1	1.1
Goldman Sachs					1.1	1.1
equity markets				1.1		1.1
Reason Multiple	US AND Dow Jones					1.1
	property sector + activity AND economic situation					1.1
	economic situation AND incoming fund + activity					1.1
	Wall Street And economic situation					1.1
	interest rate + no change And Hutchison					1.1
other						34.4

Table 5.9: Causation knowledge discovered for the Hong Kong stock market's upward movement

	Movement (%)					
Factors	**up**	**down**	**no change**	**activity**	**no movement**	**total**
Wall Street	1.1	13.5			1.1	15.7
property sector		3.2		4.3	1.1	8.6
corporate earning	1.6	0.5		3.2	3.2	8.5
HSBC	1.1	1.1		2.2	3.2	7.8
US	1.1	2.2			2.2	5.5
telecom sector				1.1	1.1	2.2
First Pac chungs					1.1	1.1
profit-taking	1.1		1.1		1.1	3.3
interest rate				1.1		1.1
US consumer confidence		1.1				1.1
pressure				1.1		1.1
futures				2.2		2.2
economic situation	2.2	1.1		1.1		4.4
bank					2.2	2.2
support barrier				2.2		2.2
oil giant CNOOC	1.1	1.1				2.2
garment trader/exporters					1.1	1.1
Multiple Reasons	properties+down AND HSBC					1.1
	US AND Dow Jones					1.1
	US AND bank					1.1
	profit-taking AND HSBC					1.1
	Cathay Pacific+down AND futures					1.1
	Indonesia AND CNOOC					0.5
	shares activity AND US down					1.1
other						23.6

Table 5.10: Causation knowledge discovered for the Hong Kong stock market's downward movement

	Movement (%)					
Factors	**up**	**down**	**no change**	**activity**	**no movement**	**total**
economic situations	10.3				3.4	13.7
Wall Street		3.4				3.4
US				3.4	3.4	6.8
bank	3.4	3.4				6.8
HSBC		3.4				3.4
corporate earnings	1.7					1.7
profit-taking				3.4		3.4
pay-TV				3.4		3.4
Multiple Reason	China Mobile AND HSBC+up					6.9
other						50.5

Table 5.11: Causation knowledge discovered for a stable Hong Kong stock market

The reasons for causing a stable Hong Kong stock market are shown in Table 5.11. The main reasons are the economice situations (13.7%) and the U.S. market (6.8%). Other factors include the Wall Street Market, the banks and so on.

For some of the cases, the effect in the causation relation in the Hong Kong stock market movement domain does not include the concept of "movement". For these cases, the reasons are depicted in Table 5.12. It shows that Wall Street, US and economic situation are the main factors in affecting the Hong Kong stocks market.

Moreover, SEKE also discovered some complex reasons for the Hong Kong stock market movement domain. Those causation knowledge are depicted in Table 5.13. For example, the activity of terrorism causes Wall Street fall and the activity of Argentina's economic situations affects the performance of HSBC.

<table>
<tr><th></th><th colspan="6">Movement (%)</th></tr>
<tr><th>Factors</th><th>up</th><th>down</th><th>no change</th><th>activity</th><th>no movement</th><th>total</th></tr>
<tr><td>Wall Street</td><td>1.8</td><td>1.8</td><td></td><td>1.8</td><td>1.8</td><td>7.2</td></tr>
<tr><td>US</td><td>2.7</td><td></td><td></td><td></td><td>4.9</td><td>7.6</td></tr>
<tr><td>economic situation</td><td>2.7</td><td>1.8</td><td></td><td>1.0</td><td>1.0</td><td>6.5</td></tr>
<tr><td>corporate earnings</td><td>4.0</td><td>0.4</td><td></td><td></td><td>0.4</td><td>4.8</td></tr>
<tr><td>property sector</td><td></td><td></td><td></td><td>1.8</td><td>1.0</td><td>2.8</td></tr>
<tr><td>HSBC</td><td></td><td></td><td></td><td></td><td>1.0</td><td>1.0</td></tr>
<tr><td>telecom stocks</td><td></td><td></td><td></td><td>1.0</td><td>1.0</td><td>2.0</td></tr>
<tr><td>Asian Mkt</td><td></td><td></td><td></td><td></td><td>1.0</td><td>1.0</td></tr>
<tr><td>corporate forecasts</td><td></td><td></td><td></td><td></td><td>1.0</td><td>1.0</td></tr>
<tr><td>investors</td><td></td><td></td><td></td><td></td><td>1.0</td><td>1.0</td></tr>
<tr><td>China Mobile</td><td></td><td></td><td></td><td></td><td>1.8</td><td>1.8</td></tr>
<tr><td>banks</td><td>1.0</td><td>1.0</td><td></td><td>1.8</td><td>0.4</td><td>3.2</td></tr>
<tr><td>industry competition</td><td></td><td></td><td></td><td>1.0</td><td></td><td>1.0</td></tr>
<tr><td>futures</td><td>1.8</td><td></td><td></td><td></td><td>1.0</td><td>2.8</td></tr>
<tr><td>equity markets</td><td></td><td></td><td></td><td>1.0</td><td></td><td>1.0</td></tr>
<tr><td>profit-taking</td><td>0.4</td><td>0.4</td><td></td><td></td><td>2.7</td><td>3.5</td></tr>
<tr><td>holiday</td><td></td><td></td><td></td><td></td><td>1.0</td><td>1.0</td></tr>
<tr><td>working purposes</td><td></td><td></td><td></td><td></td><td>1.0</td><td>1.0</td></tr>
<tr><td rowspan="9">Multiple Reasons</td><td colspan="5">US AND economic situation + activity</td><td>1.0</td></tr>
<tr><td colspan="5">futures AND earnings</td><td>1.0</td></tr>
<tr><td colspan="5">economic situation AND banking shares + activity</td><td>1.0</td></tr>
<tr><td colspan="5">Nasdaq + up AND U.S. + up</td><td>1.8</td></tr>
<tr><td colspan="5">Wall Street + activity AND bad debt</td><td>1.0</td></tr>
<tr><td colspan="5">Wall Street + down AND earnings</td><td>1.8</td></tr>
<tr><td colspan="5">Wall Street + down AND U.S.</td><td>0.4</td></tr>
<tr><td colspan="5">HSBC AND earnings</td><td>0.4</td></tr>
<tr><td colspan="5">HSBC AND profit-taking</td><td>0.4</td></tr>
<tr><td>other</td><td></td><td></td><td></td><td></td><td></td><td>40.8</td></tr>
</table>

Table 5.12: Causation knowledge discovered for the Hong Kong stock market

Consequence	movement	Reason	movement
US		profit-taking	activity
US		earnings	activity
Wall Street	up	properties	
Wall Street	down	US	
Wall Street	down	terrorism	activity
Wall Street	down	economy	up
Telecom stocks		banks	activity
Telecom stocks	activity	HSBC down AND oil giant CNOOC	
HSBC		UBS Warburg	
HSBC		economic problems	activity
HSBC	up	Indonesia	activity
HSBC	down	economic problems	activity
properties	activity	Legend Group	activity
properties		interest rate	activity
properties	down	Cathay Pacific	
properties	down	earnings	down
properties	down	profit-taking	down

Table 5.13: Complex reasons discovered for Hong Kong stock market movement domain

5.3.5 Parameter Effect

Besides carrying out the experiment using the optimal values of parameter choosen in Section 5.3.2, we investigated the effect of the other parameter values for the testing data set in the complete framework of SEKE.

In Fig. 5.4, we explore the tradeoff between recall, precision, and F-measure for different values of the weight parameter w, which is used for computing the overall score for patterns in Equation 4.1. The higher the value of w, the more the score depends on the matching score and less on the support of the pattern.

As the value of the weight parameter w decreases from 1.0, the recall decreases from 58% to 48%, and the precision increases from 64% to 70%. The trend stops when the value of w decreases to about 0.5. It is because when w is less than 0.5, the support of the patterns becomes the critical factor in the score of matching with patterns. Therefore, only the pattern with the highest support value is chosen which in most of the times could not discover the knowledge. A more detailed analysis is conducted by examining the patterns. Table 5.14 shows the number of different patterns applied for different values of the weight parameter, w. When w ranges from 0 to 0.5, the number of discovered patterns selected is 1. This results in the low recall and the small difference between the extraction results when the weight parameter ranges from 0 to 0.5. Conversely, when w is larger than 0.5, the percentage of applying the pattern with the highest support decreases and the number of patterns selected increase.

We explore the tradeoff between recall and precision for different values of this threshold in Fig. 5.5. The threshold, H, discussed in Chapter 4.3, is

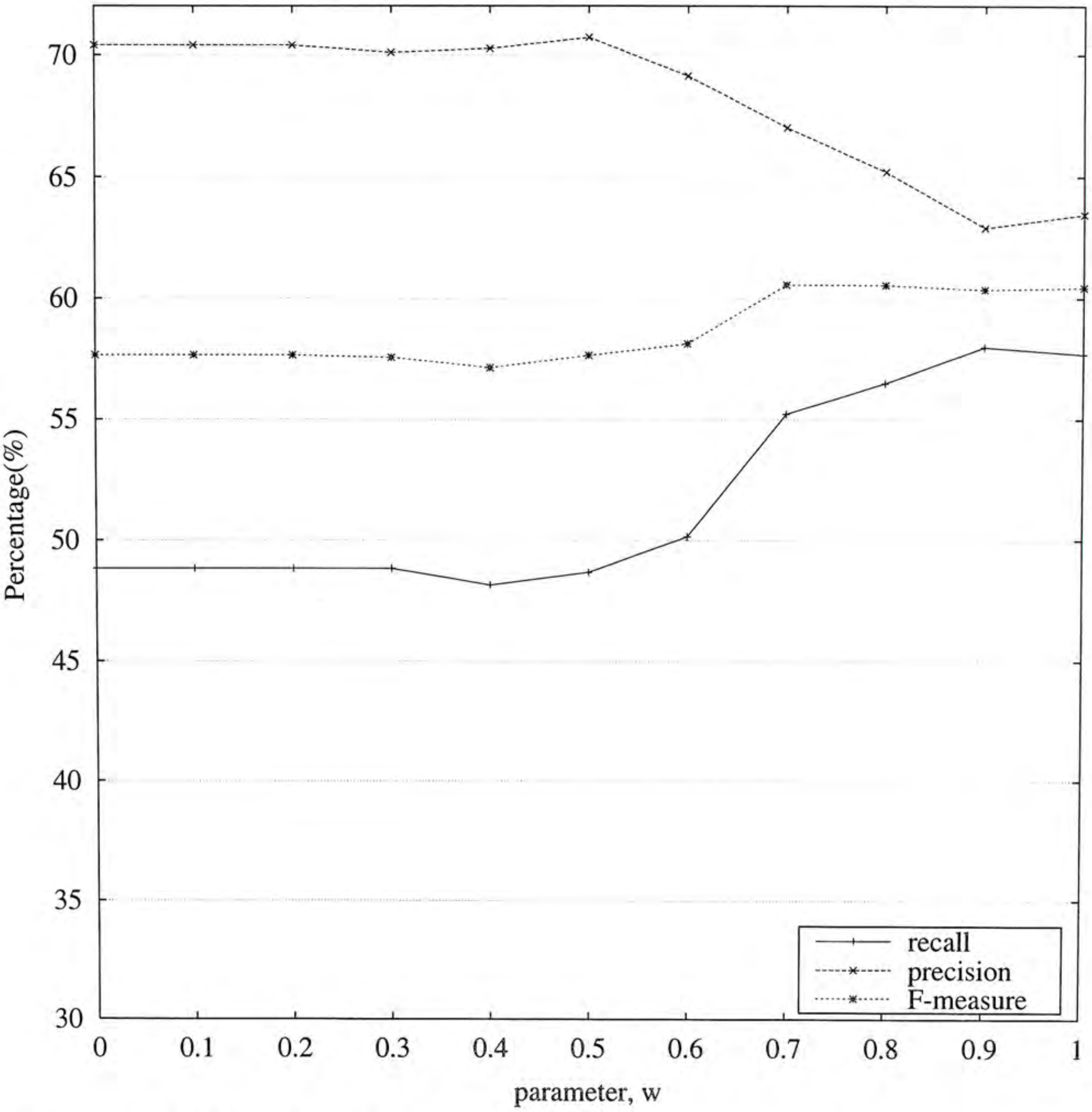

Figure 5.4: The effect of the weight parameter, w, on the extraction performance

Weight parameter, w	**No. of patterns**
1.0	83
0.9	75
0.8	49
0.7	15
0.6	2
0.0 to 0.5	1

Table 5.14: Number of different patterns applied for different values of the weight parameter, w

used to prune the information extracted by patterns. The results show that by setting the threshold higher, we can obtain a higher precision. However, the tradeoff is a lower recall, and vice versa. For example, by setting the threshold at 0.3, the recall will be 46.1% and precision will be 72.3%. The increase in precision is due to the filtering out of irrelevant knowledge. For example, by setting the threshold at 0.2, 46.7% of the irrlevant knowledge discovered by patterns is being filtered out. However, at the same time, 45.3% of the relevant knowledge discovered by patterns is being eliniminated, which causes the decrease in recall.

Fig. 5.6 shows the accumulative distribution of the confidence value for the discovered unseen knowledge, which can further explains the effects of threshold. When H is within the range of 0.1 to 0.2, there is a sharp increase of precision and decrease of recall. This is due to the fact that when H is set as 0.2, the number of discovered knowledge having confidence value below 0.2 accounts for 50%, while the percentage of those having confidence value below 0.1 is about 2%. 99% of the discovered knowledge have confidence value smaller than 0.5. This explains why the trend stops when H reaches 0.5.

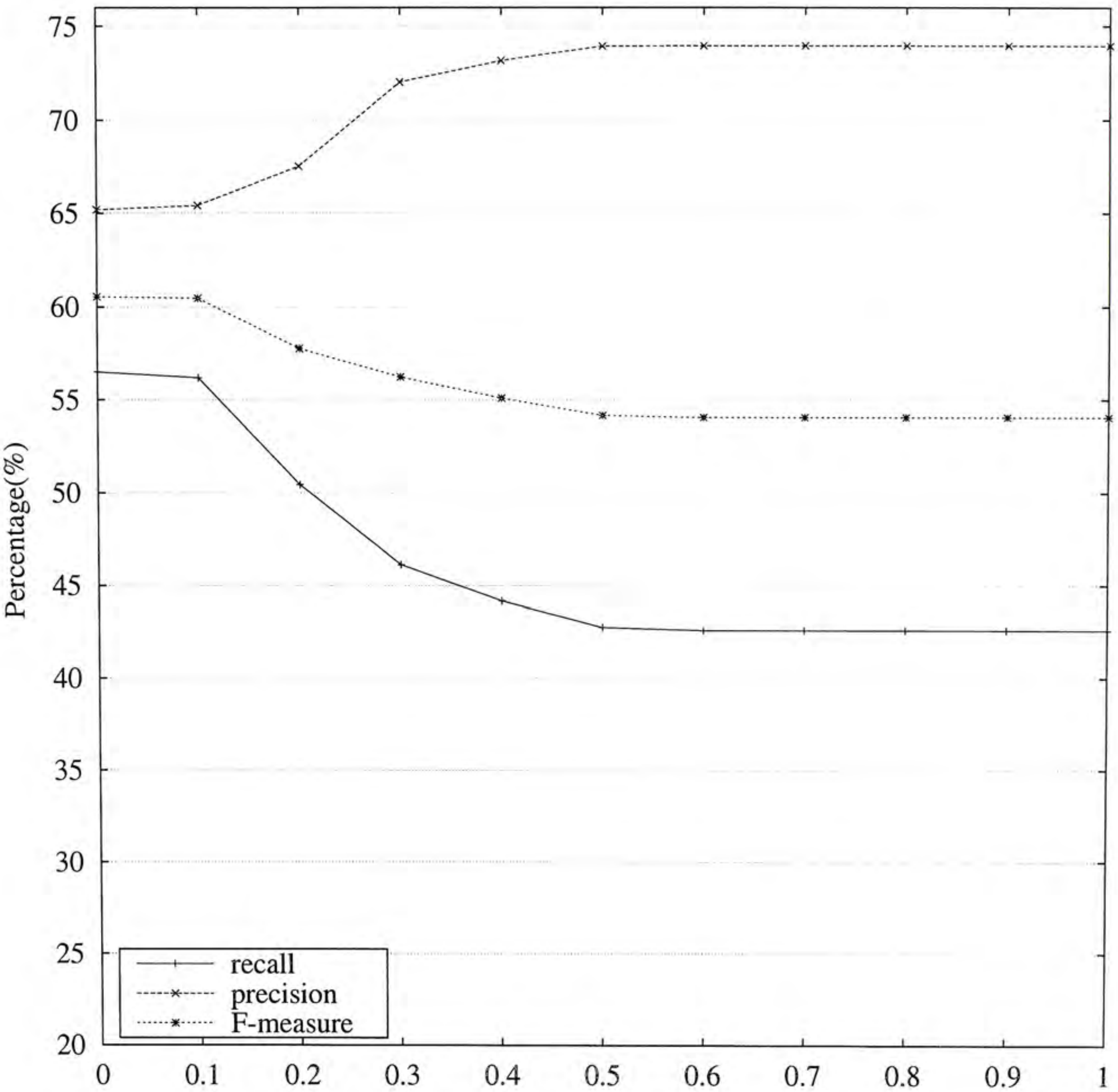

Figure 5.5: The effect of threshold H on the extraction performance

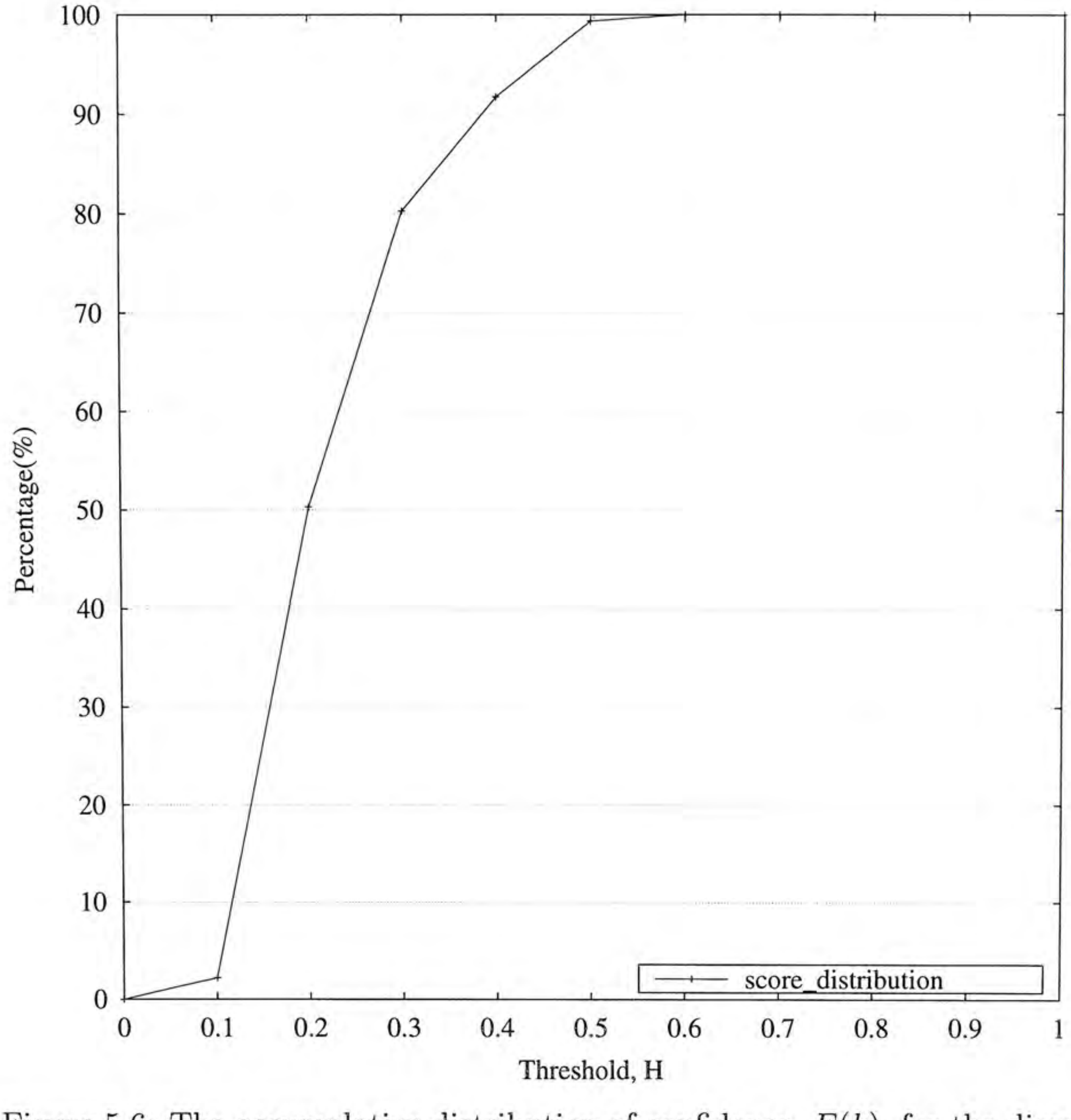

Figure 5.6: The accumulative distribution of confidence, $F(k)$, for the discovered unseen knowledge

Chapter 6

Applying SEKE on Global Warming Domain

In this chapter, we apply SEKE on another domain which is the global warming domain. We will present the experimental results of causation knowledge extraction and discovery. Again, news articles for this experiment were obtained from the Reuters newsfeed. The training data set include news articles collected from 2 September 2001 to 13 March 2002. It consists of 425 pieces of news related to global warming. The testing data set include news articles from 14 March 2002 to 31 May 2002, which includes 207 pieces of global warming related news.

6.1 Template Design

To design the templates, we analyze the sentences expressing the influencing reasons to global warming from the training data.

6.1.1 Semantic Templates

The causation relation is about what the reasons for affecting global warming are. In this causation relation, the consequence is global warming. Reasons are the factors which cause the changes in global warming such as worsening or reducing the problem of global warming. The semantic template states that some factors with actions cause the occurrence of global warming. Causation relation may also exist in the reasons or consequences themselves, therefore, causation semantic templates for complex consequences or reasons are present. They state that some factors with actions cause the occurrence of another factor with actions.

6.1.2 Sentence Templates

Based on the observation of causation sentences in the training set, a list of sentence templates are designed. They are listed in Table 6.1.

Since causation sentences in the global warming domain do not only include simple structure but also complex structure, two features are associated with the sentence templates. Firstly, as there may exist more than one reasons in the causation relation, each sentence template is associated with multiple reason templates for handling of the multiple-reasons sentences. Secondly, reasons and consequences themselves can be a causation relation and they are referred to as complex reasons and consequences. Therefore, causation structure may occur recursively within reasons or consequences. Causation sentence templates are matched recursively to the sentences for complex consequences or complex reasons. Examples of associating the sentence templates with multiple reason templates and recursively applying the templates are shown below:

	Templates	
[consequence]	caused by	[reason]
[reason]	cause of	[consequence]
[reason]	cause for	[consequence]
[reason]	cause	[consequence]
[consequence]	because	[reason]
[reason]	contribute to	[consequence]
[consequence]	blame on	[reason]
[reason]	blame for	[consequence]
[consequence]	result of	[reason]
[reason]	resulting in	[consequence]
[reason]	result in	[consequence]
[consequence]	resulting from	[reason]
If	[reason],	[consequence]
[reason]	lead to	[consequence]
[consequence]	associated with	[reason]
[reason]	attribute to	[consequence]
[reason]	affect	[consequence]
[consequence]	due to	[reason]
[consequence]	depend on	[reason]
[reason]	push	[consequence]
[consequence]	related to	[reason]
[reason]	produce	[consequence]
[reason]	is a major player in	[consequence]
[reason]	effect on	[consequence]
[reason]	is the main culprit behind	[consequence]
[reason]	the driving force behind	[consequence]
since	[reason],	[consequence]
[reason]	responsible for	[consequence]
[reason]	playing a role in	[consequence]
[reason]	impact on	[consequence]
[reason]	impact	[consequence]
[consequence]	by	[reason]
[consequence]	as	[reason]
[consequence]	with	[reason]
	Multiple Consequence/Reason Template	
[reason]	and	[reason]
[reason]	,	[reason]

Table 6.1: Causation sentence templates for the global warming domain

Simple Sentence

Reason	result in	*Consequence.*
Consequence	caused by	*Reason.*

Multiple-reason Sentence

Consequence		Multiple Reasons
$Reason_1$ and $Reason_2$	result in	*Consequence.*

Complex Sentence

Reason		Complex Consequence
$Reason_1$	result in	*Consequence* caused by $Reason_2$.
Consequence		Complex Reason
Consequence	is caused by	$Reason_1$ caused by $Reason_2$.

6.1.3 Consequence and Reason Templates

In the causation relation, reasons affecting global warming can be human activities, the amount of greenhouse gases, and so on. Therefore, the consequence template refers to "global warming" with "change". Reason template refers to "factor" with "change/action". The concept, "change", is common to both consequences and reasons and can be divided into "increase" or "decrease'. The concept, "action" is applicable with factors only. The initial lexicons for "factor" and "change/action" for the global warming domain are listed in Table 6.2 and 6.3 respectively.

Factor	**Concept Term**
Greenhouse gases	greenhouse gases, greenhouse gas, carbon dioxide levels, carbon dioxide, gases atmospheric methane
Human activity	human activity, human activities industrialized nations, factory, automobile
Fuels	fossil fuels,
Pollution	pollutants, pollution, air pollution
Ocean	iron-treated ocean, ocean, North Atlantic Oscillation
Temperatures	global temperatures
Others	mercury, phytoplankton, heat

Table 6.2: Initial lexicon for "factor" in the global warming domain

Change/Action	**Concept Term**
Increase	increase, rise, warmer, higher, warming, high, raising, increased, warm, rising, searing
Decrease	weakening
Action	burning, emission, record, change, severe, greening, fluctuations, melting, buildup, growing, uniform way, release, extinction, greener, development, sizzling, violent

Table 6.3: Initial lexicon for "change/action" in global warming domain

6.2 Pattern Discovery

We aim to discover new factors in affecting global warming. Hence, we focus on discovering patterns for extracting reasons. With the designed templates, a set of patterns is discovered for the reasons.

6.2.1 Support of Patterns

In the global warming domain, a pattern is generated if a factor is identified in the reason phrase. The factor is not necessarily accompanied by an identified "change/action" concept. During the discovery of patterns, the number of the occurrence of patterns are recorded for computing the support of patterns. The 12 most frequent patterns generated are shown in Table 6.4.

6.2.2 Relevancy of Sentence Templates

We estimate the relevancy of a sentence template by computing the ratio in Equation 4.5. The relevancy of the sentence templates is shown in Table 6.5.

6.3 Global Warming Domain Result

After the automatic sentence screening of the testing data set, 207 articles with 460 sentences were identified to be relevant to global warming.

6.3.1 Evaluation Approach

We manually examine the testing data and extract the causation knowledge. The causation knowledge extracted by SEKE will be compared with the items extracted manually.

Frequency	Patterns
13	* (marine)+[factor]+(effect)/NN *
11	[increase/decrease/action]/VB [factor]/JJ/NN
6	[factor]/JJ/NN *
6	* [factor]/NN (increase,are+blamed,are,moderate)/VB *
6	* (Kyoto+treaty)/NN (on)/IN (cutting)/VB * [factor]/NN
6	(Bush)/NN (presented)/VB * (plan)/NN (in,on)/IN (mid-February,Thursday)/NN (to)/TO (slow)/VB * [increase/decrease/action]/NN (of)/IN * [factor]/NN
4	* (Bush,power,gasoline,Kyoto+protocol)/NN (to)/TO (tackle,reduce,curb)/VB [factor]+[increase/decrease/action]/NN *
4	[increase/decrease/action]/VB [factor]/NN
4	(which)/WD (come)/VB * (from)/IN [increase/decrease/action]/VB [factor]/JJ/NN
3	(Environmentalists)/NN (say)/VB (kerosene)/NN ('s)/PO * (status)/NN (is)/VB * (subsidy)/NN (to)/TO (one)/CD (of)/IN * (growing)/VB (sources)/NN (of)/IN [factor]+[increase/decrease/action]/NN * (gas)/NN (blamed)/VB
3	(part)/NN (of)/IN (its)/PR (bid,plans,strategy)/NN (to)/TO (curb,reduce)/VB [factor]+[increase/decrease/action]/NN
3	* (scientists,Scientists)/NN (say)/VB [increase/decrease/action]/NN (of)/IN [factor]/NN
3	* [factor]+(levels)/NN (to)/TO [increase/decrease/action]/VB

Table 6.4: Examples of patterns discovered for the global warming domain

Relevancy	**Templates**		
1.00	[consequence]	caused by	[reason]
1.00	[reason]	cause of	[consequence]
1.00	[reason]	cause for	[consequence]
1.00	[reason]	cause	[consequence]
1.00	[consequence]	because	[reason]
1.00	[reason]	contributor to	[consequence]
1.00	[reason]	contribute to	[consequence]
1.00	[consequence]	blame on	[reason]
1.00	[reason]	blame for	[consequence]
1.00	[consequence]	result of	[reason]
1.00	[reason]	resulting in	[consequence]
1.00	[reason]	result in	[consequence]
1.00	[consequence]	resulting from	[reason]
1.00	If	[reason],	[consequence]
1.00	[reason]	lead to	[consequence]
1.00	[consequence]	associated with	[reason]
1.00	[reason]	attribute to	[consequence]
1.00	[reason]	affect	[consequence]
1.00	[consequence]	due to	[reason]
1.00	[consequence]	depend on	[reason]
0.67	[reason]	push	[consequence]
1.00	[consequence]	related to	[reason]
1.00	[reason]	produce	[consequence]
1.00	[reason]	is a major player in	[consequence]
1.00	[reason]	effect on	[consequence]
1.00	[reason]	is the main culprit behind	[consequence]
1.00	[reason]	the driving force behind	[consequence]
0.29	since	[reason],	[consequence]
0.67	[reason]	responsible for	[consequence]
0.33	[reason]	playing a role in	[consequence]
1.00	[reason]	impact on	[consequence]
0.36	[reason]	impact	[consequence]
0.26	[consequence]	by	[reason]
0.25	[consequence]	as	[reason]
0.13	[consequence]	with	[reason]
	Multiple Consequence/Reason Template		
	[reason]	and	[reason]
	[reason]	,	[reason]

Table 6.5: Relevancy for the causation sentence templates in the global warming domain

	recall	**precision**	**F-measure($\beta = 1$)**
basic framework	36.9%	76.6%	49.4%
complete framework	56.4%	63.5%	59.8%

Table 6.6: Experimental results of SEKE in the global warming domain

Discovered Factors
Industry/Aviation/circuit manufacturing
traffic growth
the burning of coal
people's shopping action
environment quality
edge technology (which enhance environment quality)

Table 6.7: Unseen reasons discovered of SEKE in the global warming domain

The three performance metrics, recall, precision and F-measure discussed in Chapter 5.3.1 are used for the evaluation. β is set to 1 in out experiments.

6.3.2 Experimental Results

We have evaluated the performance of both the basic and complete framework of SEKE. In the complete framework, a total of 199 patterns is discovered from 290 reasons in the training data during the pattern discovery stage. For the experiment of the complete framework, we use the same parameter settings as that of the Hong Kong stock market movement domain. The two parameters, w and H, are 0.8 and 0.3 respectively. The performance of SEKE is shown in Table 6.6.

The increase in the recall value shows that more causation relations are discovered by the complete framework of SEKE. Some unseen reasons discovered are depicted in Table 6.7.

The improved performance from the basic to the complete framework of SEKE in the global warming domain is not as obvious as that in the Hong Kong stock market domain. One possible reason is that the range of reasons for affecting global warming is smaller than that for the Hong Kong stock market movement. Therefore, the number of unseen factors is relatively small. Also, the variation of sentence structure is very large in the global warming domain. The number of pattern discovered is 20% more than that of the Hong Kong stock market movement domain. Moreover, as the patterns discovered for the global warming domain can contain only one concept, some patterns generated are too simple. These reasons attribute to the relatively low precision result of SEKE in the global warming domain.

6.3.3 Knowledge Discovered

SEKE extracted the causation knowledge explicitly stated by the authors from the news articles. We present some examples of those discovered causation knowledge in the global warming domain, which are identified by SEKE automatically from the text of testing data set. Table 6.8 shows the reasons for causing global warming to increase. From the table, the factor, "worst case scenario", accounts for 23.1% of all the reasons for causing an increase in global warming. Another 15.4% is due to the reason "greenhouse gases". They are the two main factors.

Usually, the effect in the causation relation in the global warming domain does not include the concept of "change/action". For these cases, the reasons are depicted in Table 6.9. The table shows that greenhouse gases accounts for 45% of such reasons. Within this 45%, 1.1% is related to the increase and 27.7%

	Movement %			
Factors	**increase**	**action**	**no change/action concept**	**Total**
worst case scenario	0	0	0	23.1
greenhouse gases	0	15.4	0	15.4
temperature	7.7	0	7.7	15.4
pollution	0	0	7.7	7.7
others	0	0	0	38.4

Table 6.8: Causation knowledge discovered for the increase of global warming

is related to the action concerned, such as emissions, with greenhouse gases. It also shows some multiple reasons extracted for global warming. One example consists of two factors, "greenhouse gases" and "circuit manufacturing". The two factors together affect global warming. The emission of greenhouse gases is again the main reason for causing global warming. Other factors include climate changes, human activities, such as the burning of fossil fuels releasing greenhouse gases and pollution.

Moreover, SEKE also discovered some complex reasons for the global warming domain. Those causation knowledge are depicted in Table 6.10. For example, the increase in greenhouse gases is caused by the industry or by pollution.

	Movement %			
Factors	**increase**	**action**	**no change/action concept**	**Total**
greenhouse gases	1.1	27.7	16.0	44.7
climate	1.1	14.9	5.3	21.3
temperatures	4.3	0.0	0.0	4.3
pollution	0.0	0.0	2.1	2.1
ice caps/ocean	1.1	0.0	1.1	2.2
Aviation	0.0	0.0	1.1	1.1
Automobile/traffic	0.0	0.0	1.1	1.1
Energy source	0.0	0.0	1.1	1.1
Scientists	0.0	0.0	1.1	1.1
Edge technology	1.1			
Multiple Reasons	government/society AND greenhouse gases			3.2
	government/society AND greenhouse gases+action			2.1
	government/society AND solutions			1.1
	government/society AND scientists			1.1
	government/society AND United States			2.1
	greenhouse gases AND circuit manufacturing			1.1
	greenhouse gases AND donation			2.1
	greenhouse gases AND scientists AND ice caps/ocean			1.1
	Others			7.1

Table 6.9: Causation knowledge discovered for affecting global warming without the concept of "change/action"

Consequence	**movement**	**Reason**	**movement**
greenhouse gases	increase	industry	
greenhouse gases	increase	pollution	
greenhouse gases	increase	coal	
Climate		greenhouse gases	action
Climate		pollution	
Energy		greenhouse gases	action
Pollution		mercury	
United States		climate	action
United States		greenhouse gases	action

Table 6.10: Complex reasons discovered for global warming domain

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

We have developed the framework of SEKE, a semantic expectation-based knowledge extraction system, for extracting causation knowledge from natural language texts. The basic framework of SEKE is composed of different kinds of generic templates organized in a hierarchical fashion. There are semantic templates, sentence templates, reason templates and consequence templates. The design of the templates is based on some expected semantics and simple linguistic clues related to causation. The semantic template represents the target relation. The sentence templates act as a middle layer to reconcile the semantic templates with natural language texts. With the designed templates and initial lexicons, the basic framework is able to extract causation knowledge buried in texts.

To enhance the extraction performance with limited size of initial lexicons, two techniques are used to extend the basic framework leading to the complete framework of SEKE. The first technique is to make use of a thesaurus, while the second technique makes use of automatically discovered patterns. The use of a thesaurus enables us to identify unseen concepts terms and causation knowledge from texts. The patterns are discovered from previously extracted cases and hence do not require the use of extra manual annotations. By applying the automatically discovered patterns, unseen reasons and consequences can be extracted.

We have applied both the basic framework and the complete framework of SEKE on two domain areas, namely the Hong Kong stock market movement domain and the global warming domain. The experimental results show that the recall of the complete framework is higher than that of the basic framework in both domain areas. It demonstrates that SEKE is able to discover unseen causation knowledge and also the adaptability of SEKE on different domain areas of texts for extracting causation knowledge.

7.2 Future Directions

The current approach of SEKE only uses causal links in extracting explicitly indicated causation relation in texts. To improve the coverage of SEKE, one future direction is to explore the use of other kinds of linguistic clues of causation, such as causal verbs and causative affixes in the templates. It involves issues such as how to solve the problem of capturing and resolving the ambiguities of the causal verbs.

Another direction is to explore a technique for validating the knowledge discovered and transforming the unseen knowledge into a reliable domain specific lexicon.

The complete framework of SEKE generates patterns to discover unseen reasons and consequences. We can explore the possibility of automatically discovering sentence templates. One possible way is to make use of previously discovered reason patterns and consequence patterns. The regularities in the structure between a reason pattern and a consequence pattern within a sentence may provide some hints for discovering a sentence template.

Causation relation is only one of many semantic relations. Since the framework of SEKE is based on expected semantics, by modifying the templates design, we can capture other semantic relations. Therefore, we can explore the possibility for adapting the framework of SEKE to other semantic relations.

Bibliography

[1] E. Agichtein and L. Gravano. "Snowball: Extracting Relations from Large Plain-Text Collections". In *Proceedings of Digital Libraries*, pages 85–94, San Antonio, 2000.

[2] D. Appelt, J. Bear J. Hobbs, D. Israel, M. Kameyama, and M. Tyson. "SRI: Description of the JV-Fastus System Used for MUC-5". In *Proceedings of the Fifth Message Understanding Conference*, pages 221–236, 1993.

[3] A. Bagga, J. Y. Chai, and A. W. Biermann. "The Role of WordNet in the Creation of a Trainable Message Understanding System". In H. Shrobe and T. Senator, editors, *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pages 941–948, Menlo Park, California, 1996. AAAI Press.

[4] Eric Bill. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics*, 21(4), 1995.

[5] K. Chan, B.T. Low, W. Lam, and K.P. Lam. "Extracting Causation

Knowledge from Natural Language Texts". In *Proceedings of 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2002*, pages 555–561, 2002.

[6] J. Cowie and Y. Wilks. "Information Extraction". In R. Dale, H. Moisl, and H. Soners, editors, *Handbook of Natural Language Processing*, 2000.

[7] R. Gaizauskas and Y. Wilks. "Information Extraction: Beyond Document Retrieval". *Journal of Documentation*, 54(1):70–105, 1998.

[8] D. Garcia. "COATIS, an NLP System to Locate Expressions of Actions Connected by Causality Links". In *Proceedings of the 10th European Workshop in Knowledge Acquisition, Modeling and Management, EKAW '97*, pages 347–352, 1997.

[9] R. Girju and D. Moldovan. "Text Mining for Causal Relations". In *Proceedings of Florida Artificail Intelligence Research Society, FLAIRS 2002*, Pensacola, Florida, May 2002.

[10] R. Grishman. "Information Extraction: Techniquies & Challenges". International Summer School on Information Extraction, SCIE-97. Springer, 1997.

[11] R. Grishman and J. Sterling. "New York University: Description of the Proteus System as used for MUC-5". In *Proceedings of the Fifth Message Understanding Conference*, pages 181–194, 1993.

[12] T. Honderich, editor. *The Oxford Companion to Philosophy*. Oxford University Press, 1995.

[13] D. Hume. *An Enquiry Concerning Human Understanding*. Oxford, 1975.

[14] P. S. Jacobs and L. F. Rau. "A Friendly Merger of Conceptual Expectations and Linguistic Analysis in a Text Processing System". In *Proceedings of IEEE 88*, pages 351–356, 1988.

[15] L. Joskowiscz, T. Ksiezyk, and R. Grishman. "Deep domain models for discourse analysis". In *The Annual AI Systems in Government Conference*. Silver Sprint, MD: IEEE Computer Society, 1989.

[16] R.M. Kaplan and G. Berry-Bogghe. "Knowledge-based Acquisition of Causal Relationships in Text". *Knowledge Acquisition*, 3(3):317–337, 1991.

[17] C.S.G. Khoo, S. Chan, and Y. Niu. "Extracting Causal Knowledge from a Medical Database Using Graphical Pattern". In *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

[18] C.S.G. Khoo, S. Chan, Y. Niu, and A. Ang. "A Method for Extracting Causal Knowledge from Textual Database". *Singapore Journal of Library & Information Management*, 29:48–63, 1999.

[19] C.S.G. Khoo, J. Kornfit, R.N. Oddy, and S.H. Myaeng. "Automatic Extraction of Cause-effect Information from Newspaper Text Without Knowledge-based Inferencing". *Literary and Linguistic Computing*, 13(4):177–186, 1998.

[20] J. Kim and D. Moldovan. "Acquisition of Semantic Patterns for Information Extraction from Corpora". In A. Ram and K. Moorman, editors, *Proceedings of the Nineth IEEE Conference on Artificial Intelligence for Applications*, Los Alamitos, CA, 1993. IEEE Computer Society Press.

[21] J. Kontos and M. Sidiropoulou. "On the Acquisition of Causal Knowledge from Scientific Texts with Attribute Grammars". *Expert Systems for Information Management*, 4:31–48, 1991.

[22] D. Lin and P. Pantel. "DIRT - Discovery of Inference Rules from Text". In *Proceedings of Knowledge Discovery and Data Mining*, pages 323–328, San Francisisco, 2001.

[23] B.T. Low, K. Chan, M.Y. Choi, L.L. Choi, and S.L. Lay. "A Semantic-based Acquisition Study on Hong Kong Stock Market Movement". In *Proceedings of 5th World Multiconference on Systemics, Cybernetics and Informatics, SCI 2001*, 2001.

[24] B.T. Low, K. Chan, M.Y. Choi, L.L. Choi, and S.L. Lay. "Semantic Expectation-based Causation Knowledge Extraction: A Study on Hong Kong Stock Market Movement Analysis". In *Proceedings of 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2001*, pages 114–123, 2001.

[25] G.A. Miller. "WordNet: A Lexical Database". *Communications of the ACM*, 38(11):39–41, 1995.

[26] C. Nobata and S. Sekine. "Towards Automatic Acquisition of Patterns for Information Extraction". In *International Conference of Computer Processing of Oriental Languages*, 1999.

[27] V. I. Podlesskaya. "Causatives and causality: towards a semantic typology of causal relations". In B. Comrie and M. Polinsky, editors, *Causatives and transitivity*, pages 165 - 175. J. Benjamins Pub. Co., 1993.

[28] E. Riloff. "Automatically Constructing a Dictionary for Information Extraction Tasks". In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 811–816. AAAI Press/The MIT Press, 1993.

[29] E. Riloff. "Automatically Generating Extraction Patterns from Untagged Text". In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1044–1049. AAAI Press/The MIT Press, 1996.

[30] E. Riloff. "Information Extraction as a Stepping Stone toward Story Understanding". In A. Ram and K. Moorman, editors, *Computational Models of Reading and Understanding.* The MIT Press, 1999.

[31] E. Riloff and R. Jones. "Learning Dictionaries for Information Extraction by Multi-level Boostrapping". In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 474–479, 1999.

[32] M. Selfridge. "Toward a Natural Language-based Causal Model Acquisition System". *Applied Artificial Intelligence*, 3:107–128, 1989.

[33] S. Soderland. "Learning Information Extraction Rules for Semi-Structured and Free Text". *Machine Learning*, 34:233–272, 1999.

[34] S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert. "CRYSTAL: Inducing a conceptual dictionary". In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1314–1319. AAAI Press/The MIT Press, 1995.

[35] J. J. Song. *Causatives and Causation: A Universal-typological perspective.* Longman, 1996.

[36] E. Sosa and M. Tooley, editors. *Causation.* Oxford University Press, 1993.

Appendix A

Penn Treebank Part of Speech Tags

Listed below are the standard tags used in the Penn Treebank:

No	Tag	**Description**
1	CC	conjunction, coordinating
2	CD	numeral, cardinal
3	DT	determiner
4	EX	existential there
5	FW	foreign word
6	IN	preposition or conjunction, subordinating
7	JJ	adjective or numeral, ordinal
8	JJR	adjective, comparative
9	JJS	adjective, superlative
10	LS	list item marker
11	MD	modal auxiliary
12	NN	noun, common, singular or mass
13	NNP	noun, proper, singular
14	NNPS	noun, proper, plural
15	NNS	noun, common, plural

No	Tag	Description (continued)
16	PDT	pre-determiner
17	POS	genitive marker
18	PRP	pronoun, personal
19	PRP$	pronoun, possessive
20	RB	adverb
21	RBR	adverb, comparative
22	RBS	adverb, superlative
23	RP	particle
24	SYM	symbol
25	TO	“to” as preposition or infinitive marker
26	UH	interjection
27	VB	verb, base form
28	VBD	verb, past tense
29	VBG	verb, present participle or gerund
30	VBN	verb, past participle
31	VBP	verb, present tense, not 3rd person singular
32	VBZ	verb, present tense, 3rd person singular
33	WDT	WH-determiner
34	WP	WH-pronoun
35	WP$	WHpronoun, possessive
36	WRB	Wh-adverb
37	$	dollar
38	“	opening quotation mark
39	”	closing quotation mark
40	(	opening parenthesis
41	)	closing parenthesis
42	,	comma
43	-	dash
44	.	sentence terminator
45	:	colon or ellipsis