Extracting Causal Knowledge from a Medical Database Using Graphical Patterns

Christopher S.G. Khoo, Syin Chan and Yun Niu
Centre for Advanced Information Systems, School of Computer Engineering
Blk N4, Rm2A-32, Nanyang Avenue
Nanyang Technological University
Singapore 639798
assgkhoo@ntu.edu.sg; asschan@ntu.edu.sg; niuyun@hotmail.com

Abstract

This paper reports the first part of a project that aims to develop a knowledge extraction and knowledge discovery system that extracts causal knowledge from textual databases. In this initial study, we develop a method to identify and extract cause-effect information that is explicitly expressed in medical abstracts in the Medline database. A set of graphical patterns were constructed that indicate the presence of a causal relation in sentences, and which part of the sentence represents the cause and which part represents the effect. The patterns are matched with the syntactic parse trees of sentences, and the parts of the parse tree that match with the slots in the patterns are extracted as the cause or the effect.

1 Introduction

Vast amounts of textual documents and databases are now accessible on the Internet and the World Wide Web. However, it is very difficult to retrieve useful information from this huge disorganized storehouse. Programs that can identify and extract useful information, and relate and integrate information from multiple sources are increasingly needed. The World Wide Web presents tremendous opportunities for developing knowledge extraction and knowledge discovery programs that automatically extract and acquire knowledge about a domain by integrating information from multiple sources. New knowledge can be discovered by relating disparate pieces of information and by inferencing from the extracted knowledge.

This paper reports the first phase of a project to develop a knowledge extraction and knowledge discovery system that focuses on causal knowledge. A system is being developed to identify and extract cause-effect information from the Medline database – a database of abstracts of medical journal articles and conference papers. In this initial study, we focus on cause-effect information that is explicitly expressed (i.e. indicated using some linguistic marker) in sentences. We have selected four medical areas for this study – heart disease, AIDS, depression and schizophrenia.

The medical domain was selected for two reasons:

1. The causal relation is particular important in medicine, which is concerned with developing treatments and drugs that can effect a cure for some disease
2. Because of the importance of the causal relation in medicine, the relation is more likely to be explicitly indicated using linguistic means (i.e. using words such as *result*, *effect*, *cause*, etc.).

2 Previous Studies

The goal of information extraction research is to develop systems that can identify the passage(s) in a document that contains information that is relevant to a prescribed task, extract the information and relate the pieces of information by filling a structured template or a database record (Cardie, 1997; Cowie & Lehnert, 1996; Gaizauskas & Wilks, 1998).

Information extraction research has been influenced tremendously by the series of Message Understanding Conferences (MUC-5, MUC-6, MUC-7), organized by the U.S. Advanced Research Projects Agency (ARPA) (http://www.muc.saic.com/proceedings/proceedings_index.html). Participants of the conferences

develop systems to perform common information extraction tasks, defined by the conference organizers.

For each task, a template is specified that indicates the slots to be filled in and the type of information to be extracted to fill each slot. The set of slots defines the various entities, aspects and roles relevant to a prescribed task or topic of interest. Information that has been extracted can be used for populating a database of facts about entities or events, for automatic summarization, for information mining, and for acquiring knowledge to use in a knowledge-based system. Information extraction systems have been developed for a wide range of tasks. However, few of them have focused on extracting cause-effect information from texts.

Previous studies that have attempted to extract cause-effect information from text have mostly used knowledge-based inferences to infer the causal relations. Selfridge, Daniell & Simmons (1985) and Joskowsicz, Ksiezyk & Grishman (1989) developed prototype computer programs that extracted causal knowledge from short explanatory messages entered into the knowledge acquisition component of an expert system. When there was an ambiguity whether a causal relation was expressed in the text, the systems used a domain model to check whether such a causal relation between the events was possible.

Kontos & Sidiropoulou (1991) and Kaplan & Berry-Rogghe (1991) used linguistic patterns to identify causal relations in scientific texts, but the grammar, lexicon, and patterns for identifying causal relations were hand-coded and developed just to handle the sample texts used in the studies. Knowledge-based inferences were also used. The authors pointed out that substantial domain knowledge was needed for the system to identify causal relations in the sample texts accurately.

More recently, Garcia (1997) developed a computer program to extract cause-effect information from French technical texts without using domain knowledge. He focused on causative verbs and reported a precision rate of 85%. Khoo, Kornfilt, Oddy & Myaeng (1998) developed an automatic method for extracting cause-effect information from Wall Street Journal texts using linguistic clues and pattern matching. Their system was able to extract about 68% of the causal relations with an error rate of about 36%.

The emphasis of the current study is on extracting cause-effect information that is explicitly expressed in the text without knowledge-based inferencing. It is hoped that this will result in a method that is more easily portable to other subject areas and document collections. We also make use of a parser (Conexor's FDG parser) to construct syntactic parse trees for the sentences. Graphical extraction patterns are constructed to extract information from the parse trees. As a result, a much smaller number of patterns need be constructed. Khoo et al. (1998) who used only part-of-speech tagging and phrase bracketing, but not full parsing, had to construct a large number of extraction patterns.

3 Initial Analysis of the Medical Texts

200 abstracts were downloaded from the Medline database for use as our training sample of texts. They are from four medical areas: depression, schizophrenia, heart disease and AIDs (fifty abstracts from each area). The texts were analysed to identify:

1. the different roles and attributes that are involved in a causal situation. *Cause* and *effect* are, of course, the main roles, but other roles also exist including *enabling conditions*, *size of the effect*, and *size of the cause* (e.g. dosage).
2. the various linguistic markers used by the writers to explicitly signal the presence of a causal relation, e.g. *as a result, affect, reduce*, etc.

3.1 Cause-effect template

The various roles and attributes of causal situations identified in the medical abstracts are structured in the form of a template. There are three levels in our cause-effect template, Level 1 giving the high-level roles and Level 3 giving the most specific sub-roles. The first two levels are given in Table 1. A more detailed description is provided in Khoo, Chan & Niu (1999).

The information extraction system developed in this initial study attempts to fill only the main slots of *cause*, *effect* and *modality*, without attempting to divide the main slots into subslots.

Table 1. The cause-effect template

Level 1	Level 2
Cause	Object State/Event Size
Effect	Object
	State/Event
	Size
Polarity *(e.g. "Increase", "Decrease", etc.)*	
Condition	Object
	State/Event
	Size
	Duration
	Degree of necessity
Modality *(e.g. "True", "False", "Probable", "Possible", etc.)*	
Evidence	Research method
	Sample size
	Significance level
	Information source
	Location
Type of causal relation	

Table 2. Common causal expressions for *depression* & *schizophrenia*

Expression	No. of Occurrences
causative verb	69
effect (of) …(on)	51
associate with	35
treatment of	31
have effect on	28
treat with	26
treatment with	22
effective (for)	14
related to	10

Table 3. Common causal expressions for *AIDs & heart disease*

Expression	No. of Occurrences
causative verb	119
have effect on	30
effect (of)…(on)	25
due to	20
associate with	19
treat with	15
causative noun (including nominalized verbs)	12
effective for	10

3.2 Causal expressions in medical texts

Causal relations are expressed in text in various ways. Two common ways are by using causal links and causative verbs. Causal links are words used to link clauses or phrases, indicating a causal relation between them. Altenburg (1984) provided a comprehensive typology of causal links. He classified them into four main types: the adverbial link (e.g. *hence, therefore*), the prepositional link (e.g. *because of, on account of*), subordination (e.g. *because, as, since, for, so*) and the clause-integrated line (e.g. *that's why, the result was*). *Causative verbs* are transitive action verbs that express a causal relation between the subject and object or prepositional phrase of the verb. For example, the transitive verb *break* can be paraphrased as *to cause to break*, and the transitive verb *kill* can be paraphrased as *to cause to die*.

We analyzed the 200 training abstracts to identify the linguistic markers (such as causal links and causative verbs) used to indicate causal relations explicitly. The most common linguistic expressions of cause-effect found in the *Depression* and *Schizophrenia* abstracts (occurring at least 10 times in 100 abstracts) are listed in Table 2. The common expressions found in the AIDs and Heart Disease abstracts (with at least 10 occurrences) are listed in Table 3. The expressions listed in the two tables cover about 70% of the explicit causal expressions found in the sample abstracts. Six expressions appear in both tables, indicating a substantial overlap in the two groups of medical areas. The most frequent way of expressing cause and effect is by using causative verbs.

4 Automatic Extraction of Cause-Effect Information

The information extraction process used in this study makes use of pattern matching. This is similar to methods employed by other researchers for information extraction. Whereas most studies focus on particular types of events or topics, we are focusing on a particular type of relation. Furthermore, the patterns used in this study are graphical patterns that are matched with syntactic parse trees of sentences. The patterns represent different words and sentence structures that indicate the presence of a causal relation and which parts of the sentence repre-

sent which roles in the causal situation. Any part of the sentence that matches a particular pattern is considered to describe a causal situation, and the words in the sentence that match slots in the pattern are extracted and used to fill the appropriate slots in the cause-effect template.

4.1 Parser

The sentences are parsed using Conexor's Functional Dependency Grammar of English (FDG) parser (http://www.conexor.fi), which generates a representation of the syntactic structure of the sentence (i.e. the parse tree). For the example sentence

Paclitaxel was well tolerated and resulted in a significant clinical response in this patient.

a graphical representation of the parser output is given in Fig. 1. For easier processing, the syntactic structure is converted to the linear conceptual graph formalism (Sowa, 1984) given in Fig. 2.

A conceptual graph is a graph with the nodes representing concepts and the directed arcs representing relations between concepts. Although the conceptual graph formalism was developed primarily for semantic representation, we use it to represent the syntactic structure of sentences. In the linear conceptual graph notation, concept labels are given within square brackets and relations between concepts are given within parentheses. Arrows indicate the direction of the relations.

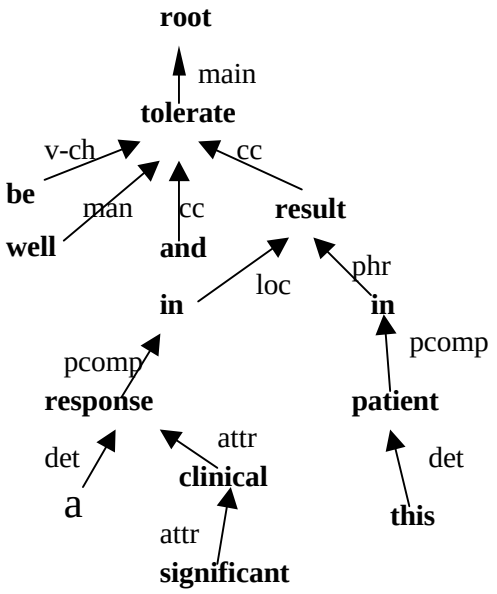

Fig. 1. Syntactic structure of a sentence

4.2 Construction of causality patterns

We developed a set of graphical patterns that specifies the various ways a causal relation can be explicitly expressed in a sentence. We call them causality patterns. The initial set of patterns was constructed based on the training set of 200 abstracts mentioned earlier. Each abstract was analysed by two of the authors to identify the sentences containing causal relations, and the parts of the sentences representing the cause and the effect. For each sentence containing a causal relation, the words (causality identifiers) that were used to signal the causal relation were also identified. These are mostly causal links and causative verbs described earlier.

Example sentence

Paclitaxel was well tolerated and resulted in a significant clinical response in this patient.

Syntactic structure in linear conceptual graph format

```
[tolerate]-
  (vch)->[be]->(subj)->[paclitaxel]
  (man)->[well]
  (cc)->[and]
  (cc)->[result]-
      (loc)->[in]->(pcomp)->[response]-
          (det)->[a]
          (attr)->[clinical]->(attr)
                  ->[significant],
      (phr)->[in]->(pcomp)->[patient]
                  ->(det)->[this],,.
```

Example causality pattern

```
[*]-
  &(v-ch)->(subj)->[T:cause.object]
  (cc|cnd)->[result]+-
      (loc)+->[in]+->(pcomp)
              ->[T:effect.event]
      (phr)->[in]->(pcomp)
              ->[T:effect.object],,.
```

Cause-effect template

Cause: paclitaxel
Effect: a significant clinical response in this patient

Fig. 2. Sentence structure and causality pattern in conceptual graph format

We constructed the causality patterns for each causality identifier, to express the different sentence constructions that the causality identifier can be involved in, and to indicate which parts of the sentence represent the cause and the effect. For each causality identifier, at least 20 sentences containing the identifier were analysed. If the training sample abstracts did not have 20 sentences containing the identifier, additional sentences were downloaded from the Medline database. After the patterns were constructed, they were applied to a new set of 20 sentences from Medline containing the identifier. Measures of precision and recall were calculated. Each set of patterns are thus associated with a precision and a recall figure as a rough indication of how good the set of patterns is.

The causality patterns are represented in linear conceptual graph format with some extensions. The symbols used in the patterns are as follows:

1. Concept nodes take the following form: *[concept_label]* or *[concept_label: role_indicator]*. *Concept_label* can be:
   - a character string in lower case, representing a stemmed word
   - a character string in uppercase, refering to a class of synonymous words that can occupy that place in a sentence
   - "*", a wildcard character that can match any word
   - "T", a wildcard character that can match with any sub-tree.

   Role_indicator refers to a slot in the cause-effect template, and can take the form:
   - *role_label* which is the name of a slot in the cause-effect template
   - *role_label = "value"*, where value is a character string that should be entered in the slot in the cause-effect template (if "value" is not specified, the part of the sentence that matches the *concept_label* is entered in the slot).
2. Relation nodes take the following form: *(set_of_relations)*. *Set_of_relations* can be:
   - a *relation_label*, which is a character string representing a syntactic relation (these are the relation tags used by Conexor's FDG parser)
   - *relation_label | set of relations* ("|" indicates a logical "or")
3. *&subpattern_label* refers to a set of subgraphs.

Each node can also be followed by a "+" indicating that the node is mandatory. If the mandatory nodes are not found in the sentence, then the pattern is rejected and no information is extracted from the sentence. All other nodes are optional. An example of a causality pattern is given in Fig. 2.

4.3 Pattern matching

The information extraction process involves matching the causality patterns with the parse trees of the sentences. The parse trees and the causality patterns are both represented in the linear conceptual graph notation. The pattern matching for each sentence follows the following procedure:

1. the causality identifiers that match with keywords in the sentence are identified,
2. the causality patterns associated with each matching causality identifier are shortlisted,
3. for each shortlisted pattern, a matching process is carried out on the sentence.

The matching process involves a kind of spreading activation in both the causality pattern graph and the sentence graph, starting from the node representing the causality identifier. If a pattern node matches a sentence node, the matching node in the pattern and the sentence are activated. This activation spreads outwards, with the causality identifier node as the center. When a pattern node does not match a sentence node, then the spreading activation stops for that branch of the pattern graph. Procedures are attached to the nodes to check whether there is a match and to extract words to fill in the slots in the cause-effect template. The pattern matching program has been implemented in Java (JDK 1.2.1). An example of a sentence, matching pattern and filled template is given in Fig. 2.

5 Evaluation

A total of 68 patterns were constructed for the 35 causality identifiers that occurred at least twice in the training abstracts. The patterns were applied to two sets of new abstracts downloaded from Medline: 100 new abstracts from the original four medical areas (25 abstracts from each area), and 30 abstracts from two new domains (15 each) – digestive system diseases and respi-

ratory tract diseases. Each test abstract was analyzed by at least 2 of the authors to identify "medically relevant" cause and effect. A fair number of causal relations in the abstracts are trivial and not medically relevant, and it was felt that it would not be useful for the information extraction system to extract these trivial causal relations.

Of the causal relations manually identified in the abstracts, about 7% are implicit (i.e. have to be inferred using knowledge-based inferencing) or occur across sentences. Since the focus of the study is on explicitly expressed cause and effect within a sentence, only these are included in the evaluation. The evaluation results are presented in Table 4. Recall is the percentage of the slots filled by the human analysts that are correctly filled by the computer program. Precision is the percentage of slots filled by the computer program that are correct (i.e. the text entered in the slot is the same as that entered by the human analysts). If the text entered by the computer program is partially correct, it is scored as 0.5 (i.e. half correct). The F-measure given in Table 4 is a combination of recall and precision equally weighted, and is calculated using the formula (MUC-7):

*2*precision*recall / (precision + recall)*

Table 4. Extraction results

Slot	Recall	Precision	F-Measure
Results for 100 abstracts from the original 4 medical areas			
Causality Identifier	.759	.768	.763
Cause	.462	.565	.508
Effect	.549	.611	.578
Modality	.410	.811	.545
Results for 30 abstracts from 2 new medical areas			
Causality Identifier	.618	.759	.681
Cause	.415	.619	.497
Effect	.441	.610	.512
Modality	.542	.765	.634

For the 4 medical areas used for building the extraction patterns, the F-measure for the cause and effect slots are 0.508 and 0.578 respectively. If implicit causal relations are included in the evaluation, the recall measures for cause and effect are 0.405 and 0.481 respectively, yielding an F-measure of 0.47 for cause and 0.54 for effect. The results are not very good, but not very bad either for an information extraction task.

For the 2 new medical areas, we can see in Table 4 that the precision is about the same as for the original 4 medical areas, indicating that the current extraction patterns work equally well in the new areas. The lower recall indicates that new causality identifiers and extraction patterns need to be constructed.

The sources of errors were analyzed for the set of 100 test abstracts and are summarized in Table 5. Most of the spurious extractions (information extracted by the program as cause or effect but not identified by human analysts) were actually causal relations that were not medically relevant. As mentioned earlier, the manual identification of causal relations focused on medically relevant causal relations. In the cases where the program did not correctly extract cause and effect information identified by the analysts, half were due to incorrect parser output, and in 20% of the cases, causality patterns have not been constructed for the causality identifier found in the sentence.

We also analyzed the instances of implicit causal relations in sentences, and found that many of them can be identified using some amount of semantic analysis. Some of them involve words like *when*, *after* and *with* that indicate a time sequence, for example:

- The results indicate that changes to 8-OH-DPAT and clonidine-induced responses occur quicker *with* the combination treatment than *with* either reboxetine or sertraline treatments alone.
- There are also no reports of serious adverse events *when* lithium is added to a monoamine oxidase inhibitor.
- Four days *after* flupenthixol administration, the patient developed orolingual dyskinetic movements involving mainly tongue biting and protrusion.

Table 5. Sources of Extraction Errors

A. Spurious errors (the program identified cause or effect not identified by the human judges)

A1. The relations extracted are not relevant to medicine or disease. (84.1%)

A2. Nominalized or adjectivized verbs are identified as causative verbs by the program because of parser error. (2.9%)

A3. Some words and sentence constructions that are used to indicate cause-effect can be used to indicate other kinds of relations as well. (13.0%)

B. Missing slots (cause or effect not extracted by program), incorrect text extracted, and partially correct extraction

B1. Complex sentence structures that are not included in the pattern. (18.8%)

B2. The parser gave the wrong syntactic structure of a sentence. (49.2%)

B3. Unexpected sentence structure resulting in the program extracting information that is actually not a cause or effect. (1.5%)

B4. Patterns for the causality identifier have not been constructed. (19.6%)

B5. Sub-tree error. The program extracts the relevant sub-tree (of the parse tree) to fill in the cause or effect slot. However, because of the sentence construction, the sub-tree includes both the cause and effect resulting in too much text being extracted. (9.5%)

B6. Errors caused by pronouns that refer to a phrase or clause within the same sentence. (1.3%)

In these cases, a treatment or drug is associated with a treatment response or physiological event. If noun phrases and clauses in sentences can be classified accurately into treatments and treatment responses (perhaps by using Medline's Medical Subject Headings), then such implicit causal relations can be identified automatically.

Another group of words involved in implicit causal relations are words like *receive*, *get* and *take*, that indicate that the patient received a drug or treatment, for example:

- The nine subjects who *received* p24-VLP and zidovudine had an augmentation and/or broadening of their CTL response compared with baseline ($p = 0.004$).

Such causal relations can also be identified by semantic analysis and classifying noun phrases and clauses into treatments and treatment responses.

6. Conclusion

We have described a method for performing automatic extraction of cause-effect information from textual documents. We use Conexor's FDG parser to construct a syntactic parse tree for each target sentence. The parse tree is matched with a set of graphical causality patterns that indicate the presence of a causal relation. When a match is found, various attributes of the causal relation (e.g. the cause, the effect, and the modality) can then be extracted and entered in a cause-effect template.

The accuracy of our extraction system is not yet satisfactory, with an accuracy of about 0.51 (F-measure) for extracting the *cause* and 0.58 for extracting the *effect* that are explicitly expressed. If both implicit and explicit causal relations are included, the accuracy is 0.41 for cause and 0.48 for effect. We were heartened to find that when the extraction patterns were applied to 2 new medical areas, the extraction precision was the same as for the original 4 medical areas.

Future work includes:

1. Constructing patterns to identify causal relations across sentences
2. Expanding the study to more medical areas
3. Incorporating semantic analysis to extract implicit cause-effect information
4. Incorporating discourse processing, including anaphor and co-reference resolution
5. Developing a method for constructing extraction patterns automatically
6. Investigating whether the cause-effect information extracted can be chained together to synthesize new knowledge.

Two aspects of discourse processing is being studied: co-reference resolution and hypothesis confirmation. Co-reference resolution is important for two reasons. The first is the obvious reason that to extract complete cause-effect information, pronouns and references have to be resolved and replaced with the information that they refer to. The second reason is that quite often a causal relation between two events is expressed more than once in a medical abstract,

each time providing new information about the causal situation. The extraction system thus needs to be able to recognize that the different causal expressions refer to the same causal situation, and merge the information extracted from the different sentences.

The second aspect of discourse processing being investigated is what we refer to as hypothesis confirmation. Sometimes, a causal relation is hypothesized by the author at the beginning of the abstract. This hypothesis may be confirmed or disconfirmed by another sentence later in the abstract. The information extraction system thus has to be able to link the initial hypothetical cause-effect expression with the confirmation or disconfirmation expression later in the abstract.

Finally, we hope eventually to develop a system that not only extracts cause-effect information from medical abstracts accurately, but also synthesizes new knowledge by chaining the extracted causal relations. In a series of studies, Swanson (1986) has demonstrated that logical connections between the published literature of two medical research areas can provide new and useful hypotheses. Suppose an article reports that A causes B, and another article reports that B causes C, then there is an implicit logical link between A and C (i.e. A causes C). This relation would not become explicit unless work is done to extract it. Thus, new discoveries can be made by analysing published literature automatically (Finn, 1998; Swanson & Smalheiser, 1997).

References

Altenberg, B. (1984). Causal linking in spoken and written English. *Studia Linguistica, 38*(1), 20-69.

Cardie, C. (1997). Empirical methods in information extraction. *AI Magazine, 18*(4), 65-79.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80-91.

Finn, R. (1998). Program Uncovers Hidden Connections in the Literature. *The Scientist, 12*(10), 12-13.

Gaizauskas, R., & Wilks, Y. (1998). Information extraction beyond document retrieval. *Journal of Documentation, 54*(1), 70-105.

Garcia, D. (1997). COATIS, an NLP system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management, 10th European Workshop, EKAW '97 Proceedings* (pp. 347-352). Berlin: Springer-Verlag.

Joskowsicz, L., Ksiezyk, T., & Grishman, R. (1989). Deep domain models for discourse analysis. In *The Annual AI Systems in Government Conference* (pp. 195-200). Silver Spring, MD: IEEE Computer Society.

Kaplan, R. M., & Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition, 3*(3), 317-337.

Khoo, C., Chan, S., Niu, Y., & Ang, A. (1999). A method for extracting causal knowledge from textual databases. *Singapore Journal of Library & Information Management, 28*, 48-63.

Khoo, C.S.G., Kornfilt, J., Oddy, R.N., & Myaeng, S.H. (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing, 13*(4), 177-186.

Kontos, J., & Sidiropoulou, M. (1991). On the acquisition of causal knowledge from scientific texts with attribute grammars. *Expert Systems for Information Management, 4*(1), 31-48.

MUC-5. (1993*). Fifth Message Understanding Conference (MUC-5).* San Francisco: Morgan Kaufmann.

MUC-6. (1995*). Sixth Message Understanding Conference (MUC-6).* San Francisco: Morgan Kaufmann.

MUC-7. (2000). *Message Understanding Conference proceedings (MUC-7)* [Online]. Available: http://www.muc.saic.com/proceedings/muc_7_toc.html.

Selfridge, M., Daniell, J., & Simmons, D. (1985). Learning causal models by understanding real-world natural language explanations. In *The Second Conference on Artificial Intelligence Applications: The Engineering of Knowledge-Based Systems* (pp. 378-383). Silver Spring, MD: IEEE Computer Society.

Sowa, J.F. (1984). *Conceptual structures: Information processing in man and machine*. Reading, MA: Addison-Wesley,.

Swanson, D.R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine, 30*(1), 7-18.

Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence, 91*, 183-203.