A Lightweight Tool for Automatically Extracting Causal Relationships from Text

Stephen V. Cole
Benedictine College
Atchison, KS 66002
cole.steve@gmail.com

Matthew D. Royal
Pensacola Christian College
Pensacola, FL 32503
mroyal4904@students.pcci.edu

Marco G. Valtorta
University of South Carolina
Columbia, SC 29208
mgv@cse.sc.edu

Michael N. Huhns
University of South Carolina
Columbia, SC 29208
huhns@engr.sc.edu

John B. Bowles
University of South Carolina
Columbia, SC 29208
bowles@engr.sc.edu

Abstract

A tool that uses natural language processing techniques to extract causal relations from text and output useful Bayesian network fragments is described. Previous research indicates that a primarily syntactic approach to causal relation detection can yield good results. We used such an approach to identify subject-verb-object triples and then applied various rules to determine which of the triples were causal relations. Overall, precision and recall were low; however, causal relations with a subject-verb-object structure accounted for a low percentage of the total causal relations in the texts we analyzed. Our research shows that additional methods are needed in order to reliably detect explicit causal relations in text.

1. Introduction

Imagine that you are an intelligence analyst when the following scenario begins to unfold. First you learn that a routine analysis of data collected at a small border crossing indicated that a suspected terrorist might have been able to enter the country at that location. Then, twelve days later you see a report that a large quantity of dynamite was stolen from a construction site in a mid-western city and a large truck was rented in the same city. A colleague tells you that monitoring of messages in an on-line chat room sometimes used by members of a terrorist cell to coordinate their activities revealed that they had all made hotel reservations in a certain city, which you know is located on a river with a large hydroelectric power dam. From these facts you might infer that the dam is likely to be a terrorist target.

Intelligence analysts must sift through volumes of data like that in the preceding narrative every day and they must be able to identify the relevant facts and link them into useful information. Often, as in the above example, the data is from disparate sources and the relevant facts are buried in masses of extraneous data. Although we are presently far from being able to automatically search files of data, extract the relevant information, analyze the data, and causally link such data in a meaningful way, an automated system to assist analysts in this regard would be very beneficial.

Other projects have investigated using Bayesian networks to link apparently unrelated events [1]; our work is designed to create Bayesian networks automatically from causal relations extracted from text. Previously, the relationships necessary for building Bayesian networks had to be extracted by human analysts. We propose that natural language processing, specifically identifying causal relations in text, can be used to build useful Bayesian network fragments. If the process of building Bayesian networks could become mostly automated, the amount of time required to build such networks would be reduced, thereby increasing information throughput for intelligence analysts.

2. Previous Work

Much research is currently being conducted with the goal of developing tools and methods that will automatically extract the underlying semantic meaning of text. One such tool is Polaris, created by the Language Computer Corporation [2]. Polaris is able to detect several different semantic relations in text; however, the cause-effect relationship is not one of them. Research has also been conducted with the goal of identifying the cause-effect semantic relationship specifically, without regard for other semantic relationships. Much of this research has produced methods that require domain-specific knowledge or machine learning techniques. Some of it is directed towards developing methods that identify cause-effect relationships in any text without any background knowledge [3, 4]; this is the research that is most relevant to our project.

In [4] Khoo, et al identified five linguistic patterns which signify explicit cause-effect relationships. The first

0-4244-0169-0/06/$20.00 © 2006 IEEE.

pattern, the causal link pattern, consists of two distinct elements (phrases, clauses, etc.) with some causal-link joining them. For example, in the sentence "The daisies stayed bright because of frequent watering," "the daisies stayed bright" and "frequent watering" are the distinct elements joined by the causal link "because of." The second pattern consists of a subject-verb-object triple in which the verb is a synonym for "cause" or reflects a resulting effect in the object. The third pattern, referred to as a resultative construction, consists of the syntactic structure <Verb Phrase><Noun Phrase><Adjective>, where the adjective describes the effect of the verb. The fourth pattern consists of the conditional structure "if X, then Y", where X causes Y. The fifth pattern consists of causal relations within adjectives and adverbs, in which the adjective or adverb conveys an effect of a cause expressed in the element it modifies. All of these patterns are strictly syntactic and do not rely on machine-learned or domain-specific knowledge. Using these methods, Khoo achieved approximately 68% recall on a sample of Wall Street Journal texts [4]. Recall is the percentage of relations correctly found out of the total number of relations found by human analysts.

Girju [3] has further investigated Khoo's second causal pattern—i.e., the subject-verb-object triple. Using the general semantic WordNet categories of the subjects and objects found in the sentence [5], Girju formulated eight specific semantic patterns which can signify the presence or absence of additional causal relations in the syntactic structure <NP><VP><NP>. Using this method, Girju achieved approximately 66% recall on a test corpus generated from an archive of Los Angeles Times articles [3].

3. Methods

Rather than trying to identify causal relations based on semantic information already identified by existing semantic systems, we chose to construct a syntactic framework from which to extract causal relations directly.

We used the Apple Pie syntactic parser [6], which statistically analyzes grammar rules from the Penn Treebank [7] to determine the parts of speech in a sentence and the sentence's entire grammatical structure. Apple Pie parses one complete sentence at a time, and may make minor grammatical or punctuation corrections. In order to handle syntactically confusing elements such as verb phrases containing words normally functioning as prepositions, we manually replaced these elements with syntactically equivalent but simpler ones before sending them to the Apple Pie parser. After a sentence was parsed, the original elements were restored. From Apple Pie's parsed output, a tree representing the syntactic structure of the sentence was then constructed in memory. This tree served as the basis for our causal analysis, allowing syntactic cause-effect patterns to be matched against it. Figure 1 shows an example of the tree representation of a simple causal sentence. We restricted our efforts to identify only explicit causal relations, since inferred causal relations require background knowledge.

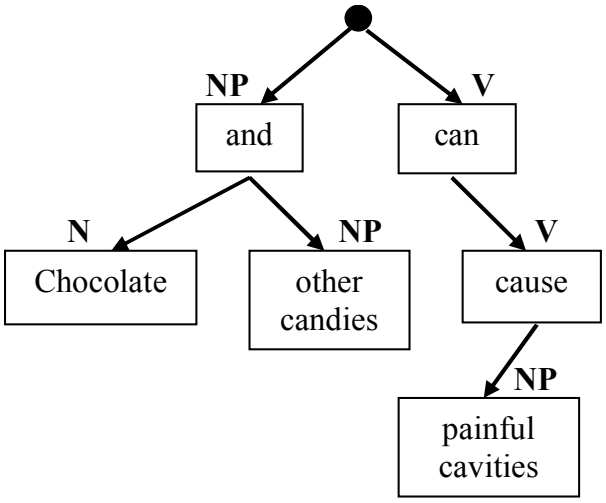

Figure 1. Syntactic tree representation of the sentence "Chocolate and other candies can cause painful cavities.

The primary causal relations we searched for were those involving subject-verb-object triples. These triples consist of the syntactic pattern <NP><VP><NP>. In finding all fragments of this structure, the tool finds all the subject-verb-object triples in all of a sentence's clauses. We then verified this method's accuracy by testing the machine-extracted results against human analysis of a test document [8].

Our primary criterion for determining whether a subject-verb-object triple is causal was that the verb phrase be synonymous with either "cause" or another verb indicating a change in the object, similar to Khoo's second pattern [4]. These verbs are maintained in an extensible list gathered primarily from an online thesaurus [9].

Our secondary criterion for determining whether an SVO triple is causal was that the subject and object fit certain patterns identified by Girju [3]. Of the nine patterns of semantic categories Girju identified as signifying causal relations, we implemented the three that she found to have the highest reliability. Like Girju, we attempted no word-sense disambiguation when referencing the WordNet semantic categories of the words.

We also transformed some passive constructions into active voice subject-verb-objects so that our causal criteria could also be applied to sentences using passive voice. For example, the sentence "The flooding of the Mississippi was caused by heavy rain" generates the subject-verb-object triple <heavy rain><caused><the flooding>.

We tested the precision and recall of these criteria in the same way we tested our subject-verb-object extraction, namely by comparing the machine's output with human analysis of the causal relations extracted from two test documents. One document was a hypothetical intelligence

report which represented the type of document this tool may eventually be used to analyze; the other was extracted from an article in the *Wall Street Journal* [10]. The *Wall Street Journal* article was selected mainly for its high number of explicit causal relations. The resulting machine-extracted Bayesian network fragments are represented as nodes in XML-BIF format. These fragments could then be integrated into a full Bayesian network. An overview of the tool is shown in Figure 2.

4. Results

The metrics we used to analyze our test results for subject-verb-object extraction and causal relation extraction were precision and recall. Precision is defined as the ratio of relations correctly extracted by the machine to the total number of relations it extracted, and recall is defined as the ratio of relations correctly found by the machine out the total number of relations extracted by human analysis. These measures produced a high precision and recall for the subject-verb-object testing, but a low precision and recall for the causal relation testing, as seen in Figures 3 and 4.

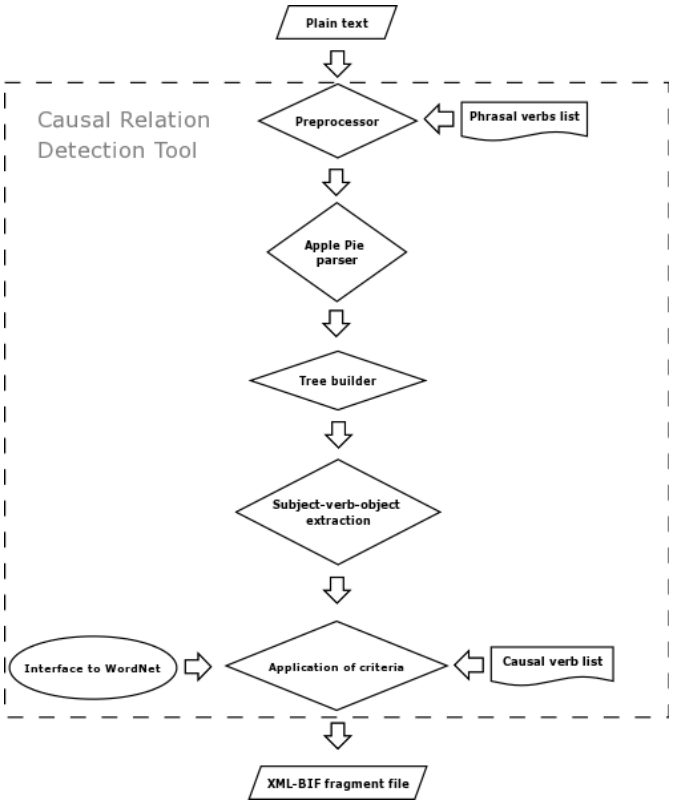

Figure 2. Causal relation detection process. All aspects of the process are automated, so that the tool takes a text file as input and outputs the final results in an XML-BIF file.

Actual number	55
Number found	36
Number correct	34
Precision	94.44%
Recall	61.82%

Figure 3. Automatic subject-verb-object detection results. Actual number is the number of SVO triples identified by human analysis, number found is the total number of SVO triples extracted by the tool, and number correct is the number of SVO triples identified by both the tool and human analysis.

5. Discussion

We chose to base our causal relation tool on a lightweight, syntactic textual analysis rather than a full semantic analysis. The existing semantic analysis tools we investigated were insufficient for identifying cause/effect relationships. While they identified many semantic relationships, they did not specifically address the cause/effect semantic relationship. Therefore, by building our own syntactic framework as a base, we were able to focus exclusively on the cause/effect relationship and customize our criteria for finding this relationship based on the syntactic structure itself or on additional semantic patterns.

We found that introducing the three WordNet criteria from Girju's research did not affect the recall percentage for the intelligence report, while it significantly increased the precision percentage. The only causal relation found in the Wall Street Journal article was found using the WordNet criteria, so removing these criteria reduced the precision and recall to 0%. Because of the nature of the future application of this tool, a higher precision seems more desirable than a higher recall—it is preferable for an analyst to have to spend time finding relations missed by the tool rather than spend additional time removing relations incorrectly found by the tool. Therefore, since removing the WordNet criteria reduced the number of incorrectly obtained relations by 54 for the intelligence report and 9 for the Wall Street Journal article, it seems that our tool was more useful without the WordNet criteria.

The recall percentage we obtained, 8.8% and 5.0% respectively, is low primarily because our tool was designed to handle causality within subject-verb-object triples, which accounted for a low percentage of the total causal relations in the documents we analyzed (see Figure 3). In these documents, subject-verb-object triples accounted for approximately 10% of total causal relations in the Wall Street Journal article and 18% in the intelligence report.

We suspect that the precision percentage we obtained is low primarily because many of the synonyms for “cause”

that our system is designed to detect do not always mean "cause" in context, and our lightweight syntactic tool did not attempt to disambiguate these contexts. Girju's WordNet criteria are designed to partially solve this problem for a wider range of verbs by identifying general contexts in which an ambiguous verb is causal; however, because of their generality, all verbs and contexts cannot be disambiguated by them, and therefore many non-causal relations were still identified as causal. Another possible source of low precision was parsing errors from the Apple Pie syntactic parser. According to Sekine [6], Apple Pie operates with a precision of 71.04% and a recall of 70.33%. Overall, our precision for finding causative verb relations was consistent with Khoo's 19% precision for these types of relations, indicating that to reliably identify causal relations, additional methods besides identifying causal verbs must be used.

While we did not specifically determine a margin of error for our measured precision and recall, we suspect the margin of error to be significant because of the small number of texts analyzed. However, it seems that the trends of low precision and recall for causal analysis and high precision and recall for subject-verb-object extraction were strong enough that they would remain consistent over a large body of analyzed text as well.

Our tool does generate some useful Bayesian network fragments, though until precision can be improved, it seems only marginally useful as a practical, time-saving application. However, our tool does provide a foundation for building a more complete one by effectively building and storing the syntactic structure of a sentence. The ability to add new patterns for extracting causal relations makes the tool useful for further development.

6. Future Work

The work of this project in identifying causal relations can be expanded in two ways: improving the existing subject-verb-object causal identification pattern and expanding the causal identification criteria to include other patterns. As it exists now, our tool identifies subjects and objects that are noun phrases; handling of subjects and objects that are clauses and other types of phrases would improve its performance. Expansion of our tool beyond the subject-verb-object pattern could be accomplished by adding the other patterns described by Khoo, specifically the causal-link, resultative construction, and conditional patterns. These patterns, which also identify inter-sentence relationships in which a cause and effect are in adjacent sentences, would be especially helpful in analyzing a text like our *Wall Street Journal* article, which contained predominantly causal-link and conditional causal relations.

The work of this project could also be expanded in ways that would assist its specific application to Bayesian networks. In addition to simply detecting the presence of causal relations, our work could be further extended to support identification of different degrees of causality based on textual clues such as the modifiers "always", "often", "sometimes", "occasionally", and "never". This would generate more descriptive Bayesian network fragments, thus further reducing the workload of intelligence analysts. To produce less ambiguous Bayesian network fragments, our work could also be extended to include antecedent support, matching pronouns with the words they refer to.

7. Acknowledgement

This work was done as part of the Research Experience for Undergraduates in Multidisciplinary Computing project at the University of South Carolina and supported in part by the National Science Foundation (award #0353637). We gratefully acknowledge discussions with Dr. Hua Li of Sarnoff Labs on the conceptual picture for the tool

	Hypothetical Intelligence Report		**Wall Street Journal Article**	
	With WordNet	*Without WordNet*	*With WordNet*	*Without WordNet*
Total relations	34		20	
SVO relations	6		2	
Number found	72	18	10	0
Number correct	3	3	1	0
Precision	4.2%	17%	10%	0%
Recall	8.8%	8.8%	5%	0%
SVO Recall	50%	50%	50%	0%

Figure 4. Automatic causal relation detection. Each document was tested both with and without our version of Girju's WordNet criteria. Note the number of causal relations that had a subject-verb-object construction and the adjusted recall when only considering these relations.

presented in this paper.

8. References

[1] K. B. Laskeya and T. S. Levitt, "Multisource fusion for opportunistic detection and probabilistic assessment of homeland terrorist threats," in *Proc. SPIE Vol.4708,* pp.80-89.

[2] D. Moldovan and P. Parker, "Towards Automatic Discovery of Semantic Relations," forthcoming.

[3] R. Girju, "Automatic Detection of Causal Relations for Question Answering," in *Proceedings of Association for Computational Linguistics Workshop on Multilingual Summarization and Question Answering,* 2003, pp. 76-83.

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

[5] C. Fellbaum, ed., *Word Net: An Electronic Lexical Database.* Cambridge, MA: MIT Press, 1998.

[6] S. Sekine, and R. Grishman, "A Corpus-based Probabilistic Grammar with Only Two Non-terminals," *Fourth International Workshop on Parsing Technology,* pp. 216-223*,* 1995.

[7] The Penn Treebank project, Available from http://www.cis.upenn.edu/~treebank/, accessed 2005-12-13.

[8] "Germany," *Wikipedia.* [Online]. Available from http://en.wikipedia.org/wiki/Germany/, [Last accessed: July 20, 2005].

[9] Thesaurus.com, [Online]. Available from http://thesaurus.reference.com/, [Last accessed: July 20, 2005].

[10] J. D. Opdyke, "The Dollar Strikes Back—Some Investors Start to Rethink Their Exposure To Foreign Funds as U.S. Currency Rises," *Wall Street Journal,* July 12, 2005, p D1.