Summarizing information by means of causal sentences through causal questions

C. Puente[1], A. Sobrino[2], E. Garrido[1], J. A. Olivas[3]
[1] Advanced Technical Faculty of Engineering ICAI
Pontifical Comillas University, Madrid, Spain
{Cristina.Puente, Eduardo.Garrido}@upcomillas.es
[2] Faculty of Philosophy. University of Santiago de Compostela
Santiago de Compostela, Spain
alejandro.sobrino@usc.es
[3] Information Technologies and Systems Dept
University of Castilla-La Mancha, Ciudad Real, Spain
Joseangel.olivas@uclm.es

Abstract. The aim of this paper is to introduce a set of algorithms able to configure an automatic answer from a proposed question and summarize information from a causal graph. This procedure has three main steps. The first one is focused in the extraction, filtering and selection of those causal sentences that could have relevant information for the process. The second one is focused in the composition of a suitable causal graph, removing redundant information and solving ambiguity problems. The third step is a procedure able to read the causal graph to compose a suitable answer to a proposed causal question by summarizing the information contained in it.

Keywords: Causal questions, Causality, Causal Sentences, Causal Representation, Causal summarization.

1 Introduction to Causality

Causality is a key notion in science and philosophy. Physics laws are often expressed in terms of a causal relation, helping in the relevant job of explanation and prediction. For example, Newton's second law predicts the force necessary (cause) to perform a desired acceleration (effect). In Philosophy, the relevance of causality was highlighted by Aristotle. In Posterior Analytics, he asserted that: we think we have knowledge of a thing only when we have grasped its cause (APost. 71 b 9-11. Cf. APost. 94 a 20). Aristotle also advanced the distinction of causality in four mayor types or classes: material cause, formal cause, efficient cause and final cause.

In the history of thought, it is possible to trace two main conceptions of causality: (i) causality as an illusion, argued by Hume and (ii) causality as a basis for the epistemological foundations of science, argued by Kant or Mill. In Treatise [1], Hume argues that the existence of the cause can be inferred from the existence of the effect thanks to the persistence in time of the association between both phenomena. Although persistence presupposes regularity in nature, this fact is not an empirical truth, but something grounded in the psychological constitution of human nature.

Kant [2] agrees with Hume in that causal connections cannot necessarily be found empirically. For him, the causal principle is housed on the plane of understanding, not at the level of sensibility, and it comes from the transcendental deduction of categories. The cause-effect relationship is the condition of the objective validity of our empirical judgments. The principle of causality is an a priori synthetic judgment that enables us to distinguish mere subjective successions from objective associations. Mill [3] raises the importance of causality in experimental sciences. Inductive inferences from a limited number of observed cases to general instances are feasible because nature is governed by causal laws.

Usually, causality is characterized as a relationship following the schema: 'A causes B', where A is the cause, B the effect and 'cause', the causal particle. Traditionally, any causal relationship follows these guidelines [4]:

- Temporality: causes generally precede their effects.
- Contiguity: causes are contiguous to the immediate effects.
- Evidential: causes and effects provide evidence of each other.

To these traditional ideas, we would like to add another:

- Imperfection: causes, effects and the cause-effect link are usually qualified by different degrees of strength.

This last property is endorsed by the presence of vague words in the three aforementioned undisputed properties of causation, as 'generally precede', 'immediate effects' or 'is evidence of'. It is a fact that, in many cases, causality is imperfect in nature and causal relations are a matter of degree.

Causal statements in texts frequently denote relevant content. Usually, causality is linked to laws or general statements, describing outstanding knowledge about an addressed topic. Thus, to isolate and to identify causal sentences in texts seems to be a profitable goal.

2 Causal Questions

Causal explanation is commonly seen as a static process: if we do not understand an event or fact, an explanation of it should be furnished. But causal explanation is alternatively seen as a dynamic act: something provides an explanation to someone about something. In this view, causal explanation adopts the form of some kind of specialized conversation. The conversational model of causal explanation radically diverges from the more usual one called 'attributional model', which describes causal processes as out of context explanations.

Despite current searchers do not include true interrogative facilities in natural language, performing web queries a rudimentary dialogue is established between the computer and the human being. Having a search engine able to discriminate interrogative particles and able to direct the answer to what is pointed at by the particle seems to be a primary goal. In this paper we focus on the relevance of interrogative lexicons in order to provide appropriate causal descriptions.

The way we ask a question is relevant to widen or narrow the range of potential answers. Comparing a yes/no question with a when or a how question; the required answer to the first seems to be less complex than the response to the second ones.

Interrogative particles involved in interrogative sentences are, among others, which, who, when, where, what, how and why. The following pyramid arranges those particles depending on the potential complexity of their answers:

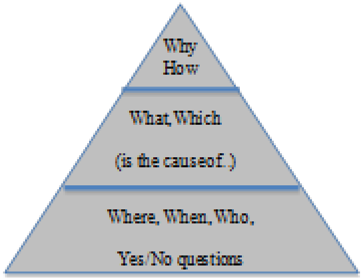


Fig. 1. Pyramid of questions' complexity.

Ascending in the pyramid using interrogative particles there is more and more demand for complex answers to questions, stimulating reflective thinking and a deeper level of conversation.

Those types of questions direct to concrete motives, causes or reasons about what is asked for. The identification of causal patterns -that we will address in a subsequent point of this paper- will be a useful tool to isolate and match causal slots containing causal content. Finally, questions including interrogative particles above the line are related to deepest answers; evoking rather than an answer the justification of the response. How questions frequently refer to a process or mechanism that show the way the answer is reached. In turn, What...is the cause, refers to the cause or causes that are asked for. Last, why questions usually presuppose some external knowledge about the query in order to answer it and are related to the prior cause or to the minimum path in the mechanism that must be followed to get the answer.

Taking these premises into account, in this paper we present three algorithms to extract causal knowledge from texts, create a causal graph and compose a summary of information using that graph as an answer to a causal question. The first algorithm is focused on the extraction of causal sentences from texts belonging to different genres or disciplines, using them as a database of knowledge about a given topic. Once the information has been selected, a question is proposed to choose those sentences where this concept is included. These statements are treated automatically in order to achieve a graphical representation in form of causal graph. The second algorithm is in charge of the generation of an answer by reading the information represented by the causal graph obtained in the previous step. Redundant information is removed, and the most relevant information is classified using several algorithms such as collocation algorithms like SALSA or classical approaches like keywords depending on the context, TF-IDF algorithm. The third algorithm generates an answer in natural language thanks to another procedure able to build phrases using a generative grammar.

3 Algorithm to extract and represent causal sentences

In [5], Puente, Sobrino, Olivas & Merlo described a procedure to automatically display a causal graph from medical knowledge included in several medical texts.

A Flex and C program was designed to analyze causal phrases denoted by words like ‘cause’, ‘effect’ or their synonyms, highlighting vague words that qualify the causal nodes or the links between them. Another C program received as input a set of tags from the previous parser and generated a template with a starting node (cause), a causal relation (denoted by lexical words), possibly qualified by fuzzy quantification, and a final node (effect), possibly modified by a linguistic hedge showing its intensity. Finally, a Java program automated this task. A general overview of the extraction of causal sentences procedure is the following:

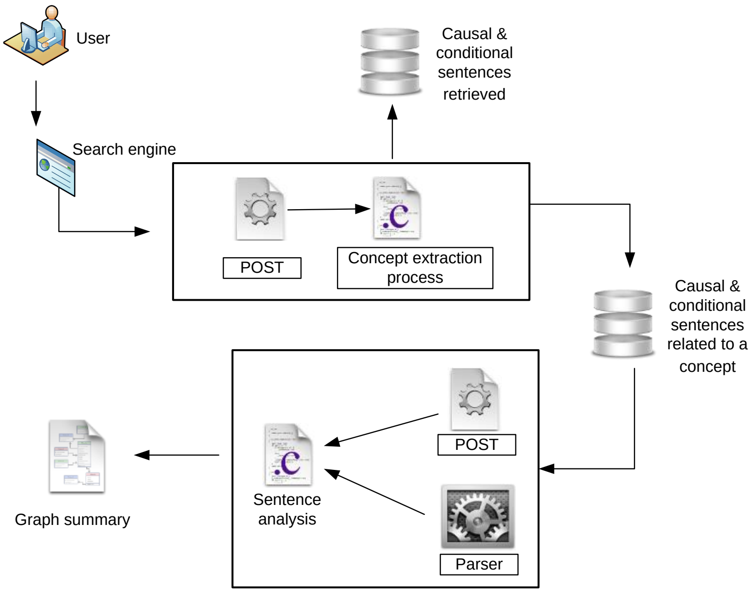

User
Causal & conditional sentences retrieved
Search engine
POST
Concept extraction process
Causal & conditional sentences related to a concept
POST
Sentence analysis
Parser
Graph summary

Fig. 2. Extraction and representation of causal sentences.

Once the system was developed, an experiment was performed to answer the question *What provokes lung cancer?*, obtaining a set of 15 causal sentences related to this topic which served as input for a causal graph representation. The whole process was unable to answer the question directly, but was capable of generating a causal graph with the topics involved in the proposed question as shown in figure 3.

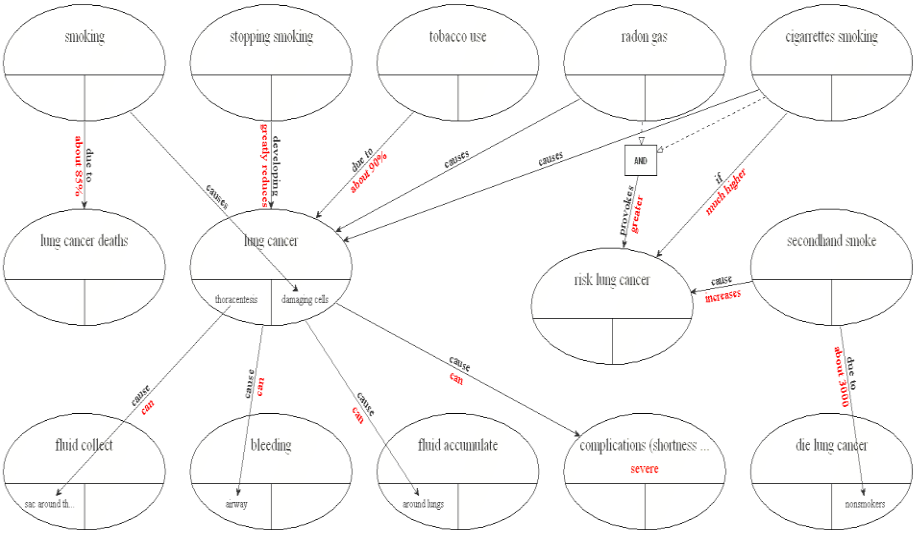


Fig. 3. Causal representation related to the question *What provokes lung cancer?*

With this causal graph, and the analysis of causal questions we want to go a step further in this paper to generate the answer to the proposed question by means of a summary, processing the information contained in the causal nodes and the relationships among them.

4 Summarizing the content of the causal graph

The ideal representation of the concepts presented in Fig. 3 would be a natural language text. This part of the article presents the design of a possible approach to do so. The size of the graph could be bigger than the presented one as not all the causal sentences are critical to appear in the final summary. It is necessary to create a summary of the information of the graph in order to be readable by a human as if it was a text created by other human.

The causal graph presented in Fig. 3 has the problem that the concepts represented could have a similar meaning in comparison to other concepts. For example, “smoking” and “tobacco use” have the similar meaning in the graph so one of these concepts could be redundant. Not only is this relationship of synonymy but other semantic relations such as hyperonymy or meronymy are important as well.

To solve this problem, we created a process to read the concepts of the graph sending them to an ontology like Wordnet or UMLS and retrieving different similarity degrees according to each relation [6].

To produce a summary, we need several processes to read the graph, reduce the redundancy, and produce the summary. The following diagram shows the design of the summary system that is created to solve this issue, including the main processes and the main tools needed.

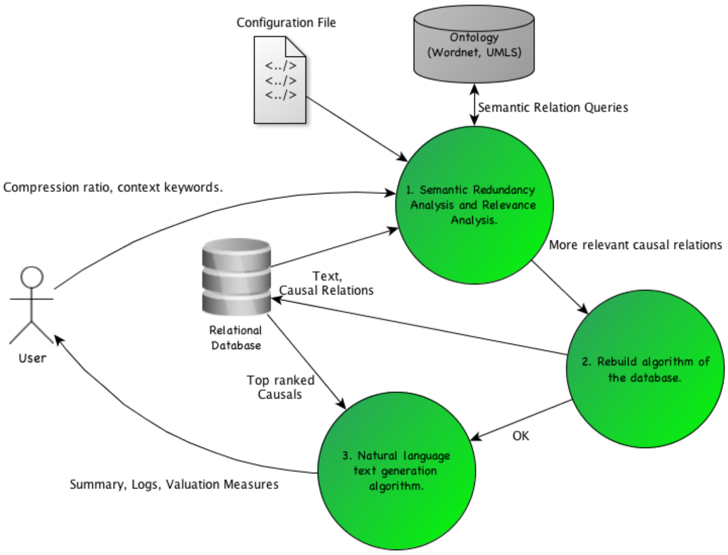

Configuration File
<../>
<../>
<../>
Ontology
(Wordnet, UMLS)
Semantic Relation Queries
Compression ratio, context keywords.
1. Semantic Redundancy Analysis and Relevance Analysis.
More relevant causal relations
Text,
Causal Relations
Relational Database
User
2. Rebuild algorithm of the database.
Top ranked Causals
OK
3. Natural language text generation algorithm.
Summary, Logs, Valuation Measures

Fig. 4 Basic design of the summarization process.

The first problem to solve is the redundancy among nodes. A redundancy analysis process is created to solve this problem taking into account the multiple synsets of every word of the concepts that is been analyzed. It is also taken into account the context of the text having keywords of every context and other measures.

To do so, Wordnet synsets are queried from Java thanks to Jwnl and RiWordnet tools to find out the meaning of these terms. The output of the process consists of possible relations between all pairs of entities compared, declaring the type and intensity of this relationship.

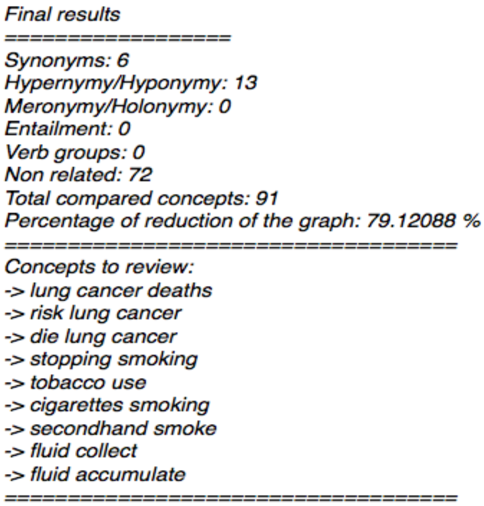
The degree of their similarity with other concepts is computed as well, being a measure to take into account in the relevance analysis. Different algorithms of similarity between concepts such as Path Length, Leacock & Chodorow [7] or Wu Palmer [8], are executed through platforms like Wordnet::Similarities.

A comparison matrix is then created with all this information, showing the similarity between terms according to different semantic relations. Those concepts with higher similarity degrees with others are the redundant ones.

$$M = \begin{pmatrix} 0 & m_{12} & m_{13} & . & . & . & m_{1n} \\ 0 & 0 & m_{23} & . & . & . & m_{2n} \\ . & . & . & . & . & . & m_{3n} \\ . & . & . & . & m_{ij} & . & m_{in} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & 0 & m_{n-1n} \\ 0 & 0 & 0 & . & . & 0 & 0 \end{pmatrix}$$

Fig. 5. Comparison matrix built by the semantic redundancy algorithm.

After running the whole process, a list of semantic relations between entities is obtained. This list contains all the information of the relations and a list of semantic entities containing the entities which are going to be deleted. This is the entry for the graph reconstruction algorithm. Additionally, a report is obtained on this first version with the final results:



```
Final results
==================
Synonyms: 6
Hypernymy/Hyponymy: 13
Meronymy/Holonymy: 0
Entailment: 0
Verb groups: 0
Non related: 72
Total compared concepts: 91
Percentage of reduction of the graph: 79.12088 %
===================================
Concepts to review:
-> lung cancer deaths
-> risk lung cancer
-> die lung cancer
-> stopping smoking
-> tobacco use
-> cigarettes smoking
-> secondhand smoke
-> fluid collect
-> fluid accumulate
===================================
```

Fig. 6. Final results.

Once a relation has been found, the problem is choosing which term is the most relevant. In the example mentioned above, the question would be, what is the most important concept, “smoking” or “tobacco use”. In [9] we proposed a mechanism that gives the answer to that question performing an analysis of the relevance of each concept. To do so, classical measures that analyses the appearance of the concepts in the text like TF-Algorithm are used. Connective algorithms that analyses the graph as SALSA or HITS [10] are also used.

When a causal relation is going to be moved to other concept due to the fact that this concept is going to be erased according to the semantic redundancy or relevance ranking algorithm, if the causal relation also exists in the concept which is not going to be erased then two different grades exists. In order to see which implication degree is the resultant one an expression is proposed:

NGa = (1-s)*Ga + s*(relA/(relA+relB)*Ga + relB/(relA+relB)*Gb);
/ NGa [0,1] ∀ {s,relA,relB,Ga,Gb} [0,1]

Being NGa the new degree of the concept A and being the concept B the one which is going to be erased, s [0,1] the semantic similarity between the two concepts, Ga and Gb [0,1] the implication degree of the concepts, relA and relB [0,1] the relevance of both terms according to the relevance ranking algorithm. Using this expression the new implication degree is calculated in function of all of the parameters of both nodes.

If the implication does not exist in the node which is not going to be erased then the expression of the new degree is the following one:

NGa = s*(relB/(relA+relB)*Gb);
/ NGa [0,1] ∀{s,relA,relB,Gb} [0,1]

After these analyses, the information of the graph has been summarized obtaining the following graph:

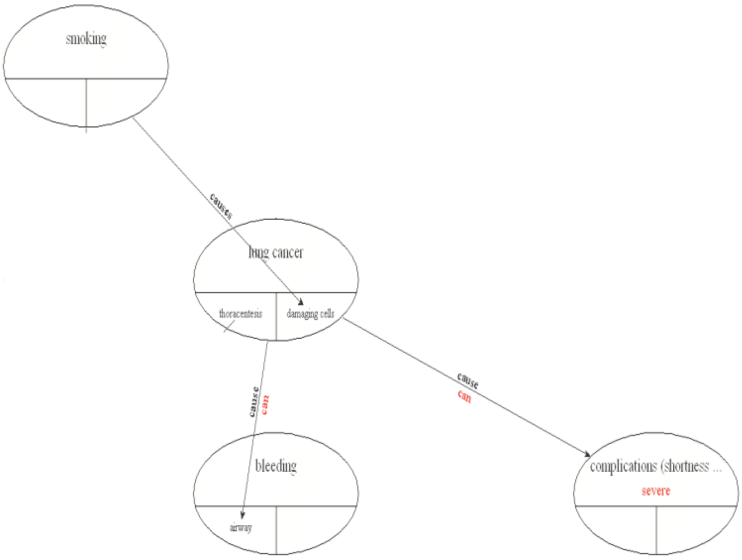


Fig. 7. Causal graph summarized.

The summary process has a configuration module depending on the user's preferences and the nature and context of the text to be analyzed. All the modules and measures can be parameterized by means of a weight-value algorithm. In order to have a better reading of the graph, the information needs to be expressed in natural language. The last process consists of an algorithm that generates natural language given the top ranked causals by the semantic redundancy analysis and relevance analysis. We have performed two experiments varying the compression rate to evaluate the obtained results and check the configuration of the algorithm. In the first experiment, we used a compression rate of 0.3, obtaining as a result the following summary:

"*Cigarettes smoking causes die lung cancer occasionally and lung cancer normally. Tobacco use causes lung cancer constantly and die lung cancer infrequently. Lung cancer causes die lung cancer seldom and fluid collect sometimes. It is important to end knowing that lung cancer sometimes causes severe complication.*"

The original text length is 1497 characters and the summary length is 311 so the system has been able to achieve the compression rate, being the summary less than the 30% of the original text. In this case, the main information has been included, removing redundant information. The system has chained sentences with the same causes to compose coordinate sentences and reduce the length of the final summary. As seen, the grammatical and semantic meaning is quite precise and accurate, without losing relevant information. In the second experiment, the compression rate was the lowest, to remove all the redundant and irrelevant information, it was set so the summary represents a 10% of the original text, so the result was the following:

"*Lung cancer is frequently caused by tobacco use. In conclusion severe complication is sometimes caused by lung cancer.*"

In this case, the system just takes the information of the three most relevant nodes, one cause, one intermediate node, and an effect node, creating a summary with the most relevant nodes included in the graph. As it can be seen, the length of the summary is of 118 characters, what represents less than a 10% of the length of the original text, the system is able to modify its behavior according to different configurations of the weights of the redundance and relevance algorithms and the compression rate.

5 Conclusions and Future Works

The massive amount of information, growing constantly, is a problem that should be treated using systems like the one proposed in this paper. This system is able to extract the most relevant knowledge contained in texts to create a causal database related to a given topic.

Using this database, the representation algorithm is able to create a causal graph containing the main concepts of a proposed question. With this graph, we have developed a procedure to remove the irrelevant information for an automatic answer. By removing the redundancy, we are able to compose an answer suitable to the proposed question, with different levels of compression.

As future works, we would like answer more complex questions, like how questions, which require for a more complex mechanism to be answered.

References

1. Hume D., A Treatise of Human Nature. D. F. Norton & U. J. Norton, (eds.), Oxford University Press, 2000.
2. Kant I., Critique of Pure Reason. 2nd ed., 1787, trans. by N. K. Smith, London, McMillan, 1929.
3. Mill J. S., A system of logic, ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation. N. Y. Harper & Brothers, 8 ed.
4. Bunge M., Causality and modern science. Dover, 1979
5. Puente, C., Sobrino A., Olivas J. A., Merlo R., Extraction, Analysis and Representation of Imperfect Conditional and Causal sentences by means of a Semi-Automatic Process.

Proceedings IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010). Barcelona, Spain, pp. 1423-1430, 2010

6. Varelas G., Voutsakis E., Raftopoulou P., Petrakis G. M. E., Milios E.. Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. WIDM'05, Bremen, Germany. November 5, (2005).
7. Leacock C., Chodorow M., Combininglocal context and WordNet similarity for word sense identification. In Fellbaum 1998, pp. 265–283.
8. Wu, Z., Palmer, M. Verb semantics and lexical selection. In 32nd Annual Meeting of the Association for Computational Linguistics, pp.133-138, Resnik, 1994.
9. Puente C., Garrido E., Olivas J. A., Seisdedos R., Creating a natural language summary from a compressed causal graph.Proc. of the Ifsa-Nafips'2013. Edmonton, Canada, 2013.
10. Najork M., Comparing The Effectiveness of HITS and SALSA. CIKM'07 Proc. of the sixteenth ACM conference on Conference on information and knowledge management. November 6–8, 2007, Lisboa, Portugal. (2007).