Journal of Applied Logic 24 (2017) 3–14

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

Summarizing information by means of causal sentences through causal graphs

C. Puente [a,*], A. Sobrino [b], J.A. Olivas [c], E. Garrido [d]

[a] *Advanced Technical Faculty of Engineering ICAI, Pontifical Comillas University, Madrid, Spain*
[b] *Faculty of Philosophy, University of Santiago de Compostela, Spain*
[c] *Information Technologies Systems Dept., University of Castilla-La Mancha, Ciudad Real, Spain*
[d] *University Autónoma of Madrid, Madrid, Spain*

ARTICLE INFO

Article history:
Available online 15 November 2016

Keywords:
Causal questions
Causality
Causal sentences
Causal representation
Causal summarization

ABSTRACT

The objective of this work is to propose a complete system able to extract causal sentences from a set of text documents, select the causal sentences contained, create a causal graph in base to a given concept using as source these causal sentences, and finally produce a text summary gathering all the information connected by means of this 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 system. The second one is focused on 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.

1. Introduction and justification

Providing causal contents is inner to scientific practice. Tracing causal knowledge is one of the most relevant jobs of the natural sciences, as Physics. Two of the most important empirical sciences tasks are explanation and prediction. Explanation involves a general statement, usually a physical law and a singular sentence, both configuring the *explanans*, base of the explanation process. Physical laws are paradigmatically causal statements. Thus, causality is largely involved in the explanation activity. Prediction demands anticipating the future. To predict the effect of changes, naturally or artificially implemented, is a desideratum of science. In this task, inductive or probabilistic logic plays a relevant role. Conditional probabilities and the Markov Principle permits to anticipate the behavior of a causal net performing interventions on

* Corresponding author.
E-mail addresses: cristina.puente@comillas.edu (C. Puente), alejandro.sobrino@usc.es (A. Sobrino), joseangel.olivas@uclm.es (J.A. Olivas), eduardo.garrido@uam.es (E. Garrido).

http://dx.doi.org/10.1016/j.jal.2016.11.020
1570-8683/

it, analyzing the dependence or independence of the involved variables and performing a causal inference according to that [16].

In an empirical, but most pragmatic scenario, as the engineering one, causality is usually related to the man-machine interaction and the qualities that a system description should have. A system should be a) clear for customers, b) clear for the system developers and c) the system itself should be clearly expressed and analyzed if required. The language of causation seems to be a natural tool for reaching clarity in the specification of the requirement of complex systems. Causal analysis should answer a question as *is a the cause of b given the system description*? [11]. Causal logic offers a vocabulary and rules in order to explain and predict complex processes in terms of cause-effect links, validating the Popper dictum that the most important fact in science is not precision, but clarity.

Since the beginning, Artificial Intelligence deals with the imitation of reasoning patterns from scenarios with impact in human knowledge and life. Physics, Engineering or Medicine have a history that started in the ancient times, and have been contributing to the human wisdom and to welfare. If the knowledge provided by those disciplines is largely causal, analyzing and extracting causal mechanisms from texts seems to be a sensible and rewarding task. Medicine shows causality as a complex process involving mechanisms. Mechanisms provide an illustration of how the prior cause evolves in intermediate causes over time before the final effect is reached. Although there are several systems retrieving causal sentences from texts, there is not any one organizing those sentences in causal mechanisms. Frequently, summaries are made from single sentences, appropriately arranging them. But there is not any approach to get summaries from causal mechanisms, focusing on the content, – not the structure – of the texts. Our paper attempts to deal with both improvements.

2. Related work

The unstoppable growth of the Internet leads to an increasingly access to more and more texts. These texts often lack of a clear and uniform structure and machines have difficulties for extracting relevant pieces of information. In texts we can found different kind of relationships: conditional, causation, correlation, a part of, etc. As we said, in this work we will focus on causal relations. Causation is important for question-answering if questions ask not only for a reply but for an explanation. On the other hand, causes are relevant in decision making because the effects of a decision can be determined by its causes. Causation permits to identify factors that make a choice rational. In IA, causation is an influential topic. It is related with planning, generation of explanations and natural language processing. Due to its importance, causality has been addressed since the 90s by diverse and interesting studies. Next, we will briefly describe some of them.

Kaplan and Berry-Bogge [7] approached a knowledge-based inference system to detect causal knowledge in scientific texts. They used linguistic templates to match causal relations even if the lexicon and grammar is handcrafted for a particular domain. The Achilles heel of this approach is its lack of scalability in real applications.

Khoo et al. developed in [9] an automatic system to extract cause-effect information in newspapers using linguistic marks. Performing a manual analysis of the documents, a set of linguistic clues indicating causal relations were isolated. Then, a pattern-matching tool extracted causal sentences. The system did not use parsing of sentences or knowledge-based inference. In [8], the authors dealt with a knowledge extraction system that retrieved causal information from texts using graphical patterns based on syntactical parsing, showing 68 causal patterns that leaded to the graphs.

In [6], Girju and Moldovan presented a system to automatically identify lexicon-syntactic patterns expressing causal relations. Syntactic patterns were identified with noun phrases linked by causative verbs. The system validated those patterns in a semi-automatically way. Based on an inductive learning approach, in [5]

Girju provided a method for automatically discover lexical and semantic constraints for the disambiguation of causal relations used in Q/A systems.

Rink et al. in [14] dealt with a method for detecting causal relations between events related in a text. The method was able to find if two events from the same sentence present a causal relation by building a graph representation of the sentence, automatically extracting graph patterns from that graph representation and training a binary classifier that decides if an event is causal or not based on the extracted graph patterns.

In the field of summarizing information, the work of linguistics, logic and statistics are the most popular to take into consideration. A summary is defined as a text that meets the main ideas of the original one but with a shorter length. Another possible definition of an abstract may be a brief statement of the essential on an issue or subject, as Basagic proposes in [2]. Summarizing texts can be considered a merely subjective task in which some people would consider an extract precise and others may throw it away. Although there are mechanisms to determine how effective a summary is, and they can be used to make an assessment of the outcome. This problem can lead to a discussion of whether an automatic procedure can deal with this subjective feat as proposed by Bawakid [3]. Still, there are certain patterns and connections between entities that can determine the most important pieces of information over other information [1].

Connecting causality and summarization, Endres-Niggemeyer [4] suggests that if events belong to a causal chain, the procedure to read and order the sequence from the beginning to the end of the chain will produce a good quality summary. Particular events or isolated ones are more difficult to connect, as they would be meaningless, or have to be set up into a context; on the other hand, if these events are ordered in a causal chain, the context is already given, and the quality of the resultant summary will be higher.

Taking these premises into account, in this paper we present a system 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 part of the system 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 part 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. This part of the system generates an answer in natural language thanks to another procedure able to build phrases using a generative grammar.

3. Extraction and representation of causal sentences

In [13], Puente, Sobrino, Olivas and Merlo described a procedure to create and display a causal graph from medical knowledge included in texts from several webs, like the Mayo Clinic, Mount Sinai Medical Center, etc.

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. This program was able to locate 20 patters previously selected of the forms that causality could have in texts, like *due to*, *owing to*, *cause*, *is provoked*, and some others until 20.

Once selected a database about a topic (like cancer, as in the example), the user has to introduce a question to select only those sentences related to the concepts involved in the question. To do so, we analyzed the way that questions and causal questions are proposed.

The way we ask a question is relevant to broad 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

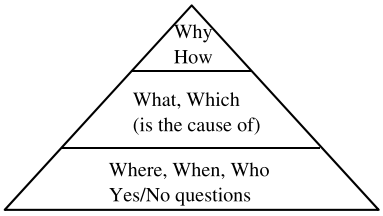

Fig. 1. Pyramid of questions' complexity.

What	WP
provokes	
lung	NN
cancer	NN
?	SENT

Fig. 2. Extraction and representation of causal sentences.

others, which, who, when, where, what, how and why. The pyramid in Fig. 1 arranges those particles depending on the potential complexity of their answers [18].

So, 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. 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.

With this theoretical analysis, we used the Stuttgart tree tagger POST [15] to select a concept from an input query and to know whether the user is asking for causes or consequences. For example, if the user asks *What provokes lung cancer?*, the POST tagger would return the following information (see Fig. 2).

POST output shows that the nominal clause is lung cancer. Processing this clause with the morphological analyzer, the program detecting the word *provokes* (associated to causes), plus the interrogative pronoun *what* would assume that the user is asking for the cause of *lung cancer*.

Once the nominal clause has been selected and isolated, another program extracts the sentences in which these concepts are contained. The search set is the file created with the conditional and causal sentences. This set of sentences will be the input for the sentence summary process.

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 in Fig. 3.

Once the system was developed, an experiment was performed to answer the question *What provokes lung cancer?* (*question introduced by the user about a given topic*), obtaining a set of 15 causal sentences related to this topic which served as input for a causal graph representation. The whole system 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 Fig. 4.

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

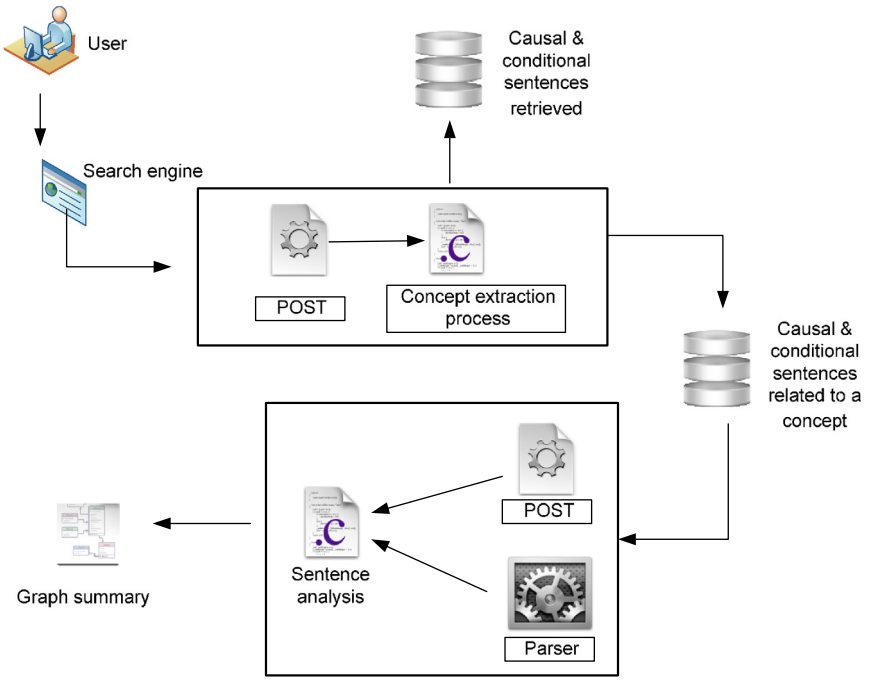

Fig. 3. Extraction and representation of causal sentences.

4. Summarizing the content of a causal graph

One of the possible applications of a causal graph, as the one presented in Fig. 4, can be a summary, generated by reading the nodes and the links among them. This part of the article proposes a design of a possible approach to do so. The size of the graph could be smaller than the presented one as not all the causal sentences are critical to appear in the final summary. Some sentences may contain redundant information with similar words, so that we have created a filtering process to detect if two sentences have basically the same meaning and remove the less important. The summary created by the graph has to be readable by a human as if it was a text created by other human, or as closer as possible.

The filtering process is in charge of removing similar concepts. For example, “smoking” and “tobacco use” have a similar meaning in the graph so one of these concepts could be redundant. In addition to synonymy, other semantic relations such as hyperonymy or meronymy are relevant 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. This permits to obtain similarity degrees according to each relation and evaluate how close the two concepts are [17].

To produce a summary, we need several computational steps to read the graph, reduce the redundancy, and generate the summary. The diagram in Fig. 5 shows the design of the summary system that is created to solve this issue, including the leading processes and the main tools needed.

To reduce the graph to useful information we have to minimize the redundancy problem. A redundancy analysis process is created to solve this trouble 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.

We have used Wordnet synsets from Java using 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 compared entities, declaring the type and intensity of these relationships.

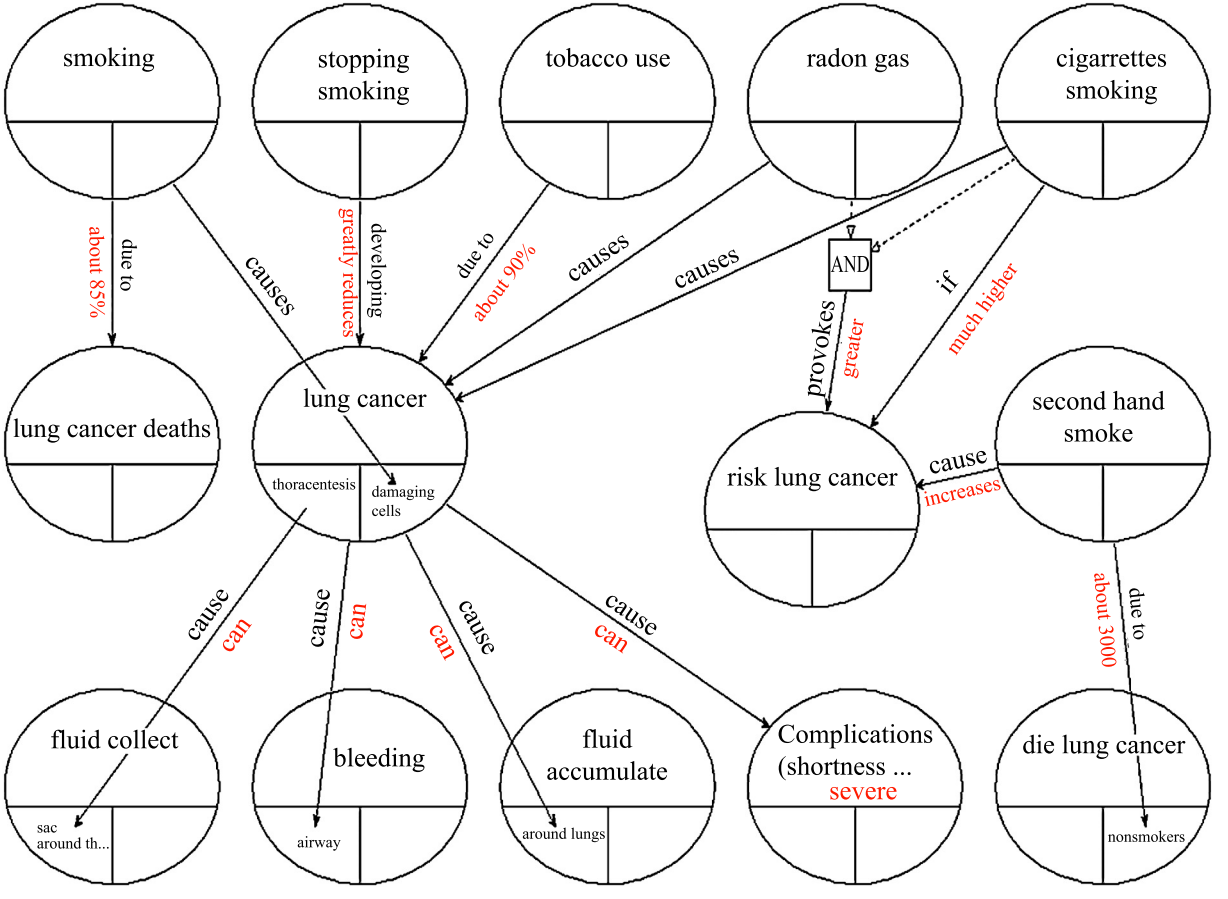

Fig. 4. Causal representation related to the question *What provokes lung cancer?*

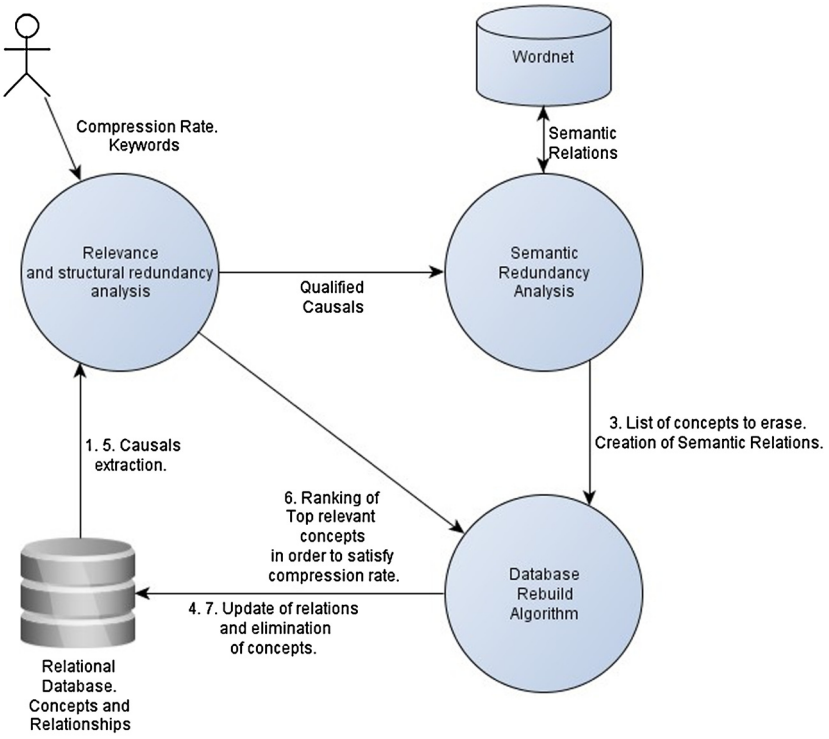

Fig. 5. Design of the summarization process.

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

Fig. 6. Comparison matrix built by the semantic redundancy algorithm.

The degree of their similarity with other concepts is computed as well, being a measure used in the relevance analysis. Different algorithms of similarity between concepts such as Path Length, Leacock and Chodorow [10] or Wu Palmer [19], are executed through platforms like Wordnet::Similarities.

A comparison matrix is then generated with all this information, showing the similarity between terms according to different semantic relations. Concepts with higher similarity degrees with others are considered as the redundant ones (see Fig. 6).

As a result of the execution, we obtain a list of semantic relations between entities, providing the information about them as well as the sequence of entities which are going to be deleted. This is the entry for the graph reconstruction algorithm. Additionally, a report on this first version is obtained (see Fig. 7).

Once a relation has been found, the next challenge 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 [12] we proposed an attempt to analyze the relevance of each concept. To do so, classical measures that analyses the frequency of concepts in the text like TF-Algorithm are used.

After these analyses, the information of the graph was summarized obtaining this new graph (see Fig. 8).

So basically, these four nodes condense the important information to be taken into account to create a summary, without losing relevant aspects.

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 modules and measures can be parameterized by means of

```
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. 7. Final results.

a weight-value algorithm. In order to have a readable graph, the information should be expressed using natural language sentences. Then, the last process consists of an algorithm that generates natural language from the top ranked causal sentences by the semantic redundancy and relevance analysis. We have performed three 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, and 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, removing all the redundant and irrelevant information. This new summary represents a 10% of the original text, obtaining the causal graph represented in Fig. 9, with the following resulting summary:

"*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. The length of the summary is of 118 characters, what represents less than a 10% of the length amount of the original text. Therefore the system is able to modify its behavior considering different configurations of the weights of redundancy and relevance algorithms and the compression rate.

We made experiments with other texts which passed essential quality tests such as measuring the syntax of the texts or assuring the compression ratio. Precision and recall were above 70% in these experiments. Reading this original text we can see the logic of the summary:

"*This is a text inspired by the famous case discussed in Ethic class Ford Pinto. When a CEO introduces a new product in the industry it has several options, new product options rarely are doing no quality test and unfrequently a fast manufacturing test is done. New product options normally imply doing a normal manufacturing process. If the organization is not meeting standards then incidents are occasionally caused by this behavior. No quality tests may produce incidents but no quality tests often imply being the first in the market. A fast manufacturing process rarely produces incidents but there are cases. Fast manufacturing*

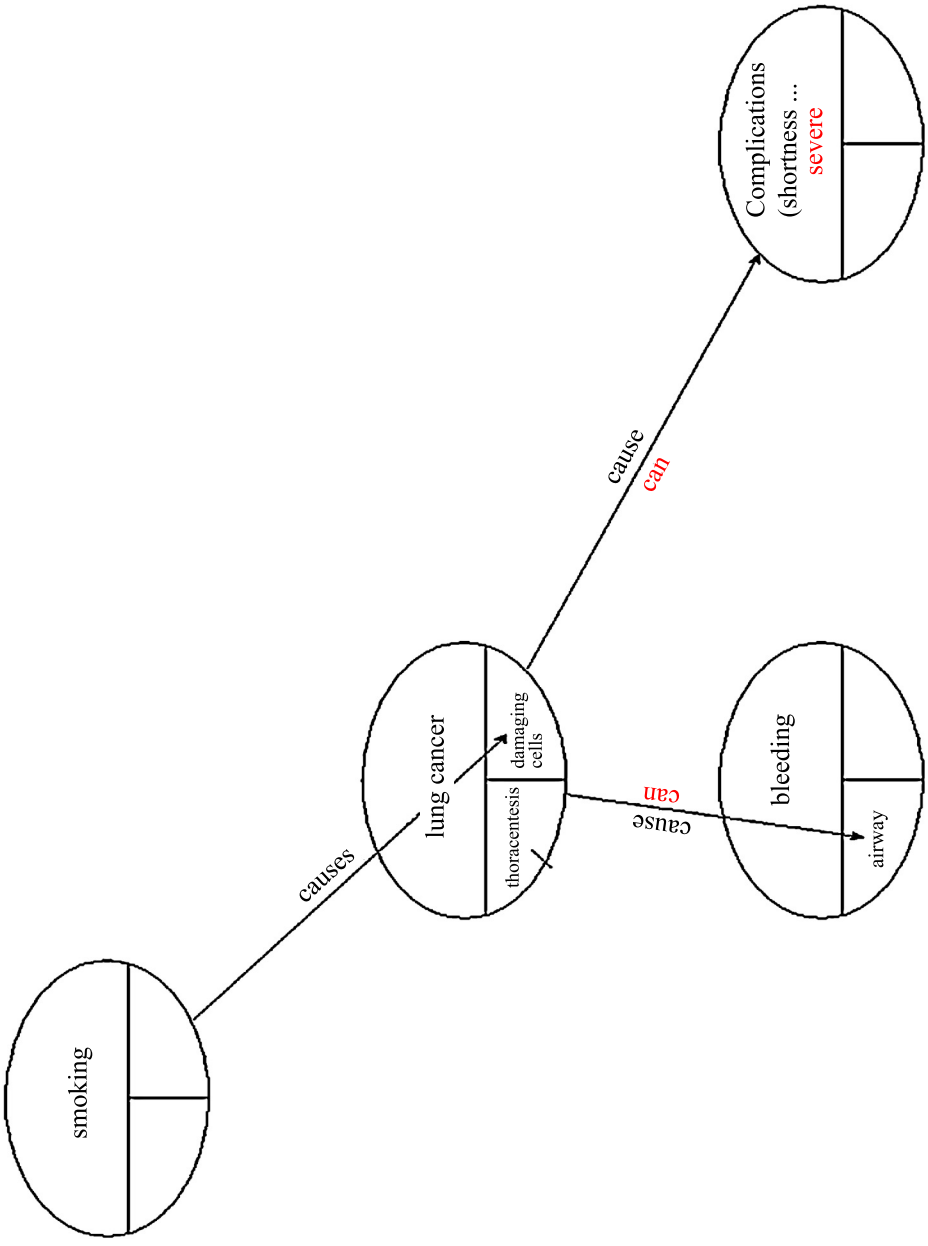

Fig. 8. Causal graph summarized.

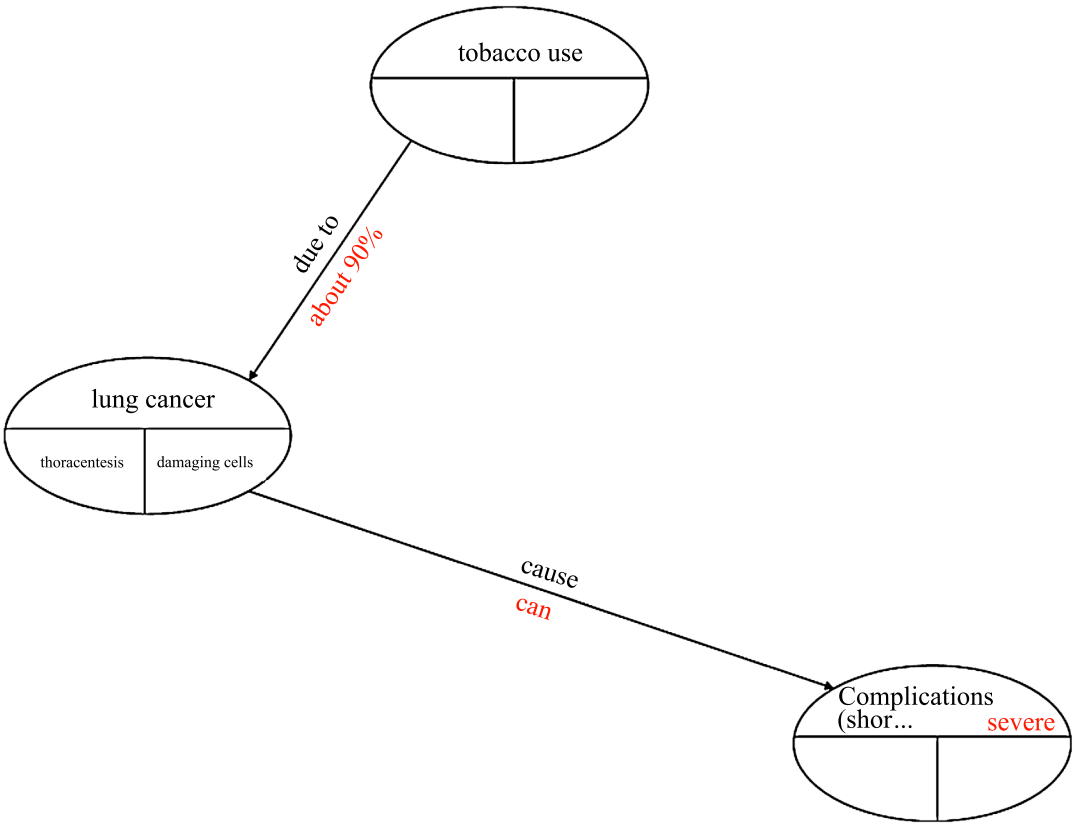

Fig. 9. Causal graph compressed with threshold 10%.

processes often implies being first in the market. A fast manufacturing process is hardly ever the cause of losses because of things that are not done. Incidents are constantly the cause of jail or prison by the CEO"s that do not follow the security standards, but the temptation is the following condition: If company is being the first in the market then it will always earn profits for a short time. But jail or prison is always the cause of losses and human lifes, CEO"s have to be aware of the ethics of not following the security standards."

Having a compression rate of 0.2 and using the configuration by default the obtained summary is the following one:

"*What is discussed is that being first in the market is implied hardly ever by new product options. Loss is never caused by being first in the market. Prison is constantly implied by incident. Eventually, prison always produces loss.*"

The original length of this text was of 1180 characters, and using a compression rate of 0.2, the length of the obtained summary has been 231 characters, which is actually a 19,58% of the original text. The degrees are those expected according to the original text and the most important semantic content of the text is contained in the summary.

5. Conclusions and future works

Causality is a powerful way to generate knowledge and to create summaries as we have presented in this work. Extracted causal knowledge varies from single causal sentences to causative rules linking, in a general way, principles and effects. Templates are used to match syntactic and semantic evidences of causality. But as we largely evidence is this paper, causality is not a single process from prior_cause-to-final_effect. Areas as engineering, biology or medicine, show causality as a complex and evolutionary process with multiple causes evolving over time. In most disciplines, instead of causality, it is frequently speak of causal

mechanisms, denoting how a phenomenon comes about or how some meaningful processes work. This is particularly true in the case of medicine. Our proposal deals not preferentially with the extraction of single causal sentences from texts, but with the organization of them in a casual mechanism, showing how a prior cause is transformed in effects that, in turn, become intermediate causes attempting the final effect.

Summarizing texts is another main task of the A.I. business. There are a lot of systems shortening the contents from the most varied sources. Medicine, law or email threads are some of the investigated areas. Most proposals are focused more on the text structure than on the displayed content. For example, it is frequently conjectured that main content usually appears in conclusions, or that an underlined text must be considered as relevant. Our approach puts forward content as the main source to obtain good summaries. Causal content is largely considered as relevant, as it answers to *how* or *why-questions*, both at the top of the significance ranking of questions. Our proposal gets summaries with different contraction levels from graphs illustrating mechanisms and thus, providing an essential *précis* of a causal mechanism which is itself also essential in content.

Our approach for summarizing conceptual content from causal mechanisms is novel. It shares with the previous proposals the extraction of causal sentences, but moves away of them organizing the retrieved sentences in a causal graph or mechanism, showing causal dependences and timing causal influences in a graph. This graph is used to generate in a new fashion way a summary that supports different degrees of contraction.

These aims are a small step in the approach of causal information retrieval, and we are aware that a lot of challenges remain to be inquired if substantial causal questions, as *how* or *why* -questions are approached and, specially, if we use them as a basis for generating deeper summaries.

References

[1] L. Alonso, I. Castellón, S. Climent, M. Fuentes, L. Padró, H. Rodríguez, Comparative Study of Automated Text Summarization, IMA 02-02-RR, Universitat de Girona, 2002.
[2] R. Basagic, D. Krupic, B. Suzic, Automatic text summarization, in: Information Search and Retrieval, WS 2009, Institute for Information Systems and Computer Media, Graz University of Technology, Graz, 2009.
[3] A. Bawakid, M. Oussalah, A Semantic Summarization System, in: Proceedings of the First Text Analysis Conference (TAC), University of Birmingham at TAC, United Kingdom, 2008.
[4] B. Endres-Niggemeyer, Summarizing Information, Springer, 1998.
[5] R. Girju, Automatic Detection of causal relations for question/answering, in: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, ACL 2003, in: Workshop in Multilingual Summarization and Q/A Machine Learning and Beyond, vol. 12, 2003, pp. 76–83.
[6] R. Girju, D. Moldovan, Text mining for causal relations, in: Proc. Florida Artificial Intelligence Research Society, FLAIRS 2002, Florida, May 2002, pp. 60–364.
[7] R.M. Kaplan, G. Berry-Bogghe, Knowledge-based acquisition of causal relationships in text, Knowl. Acquis. 3 (1991) 317–337.
[8] C.S.G. Khoo, S. Chan, Y. Niu, Extracting causal knowledge from a medical database using graphical pattern, in: Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, October 1–8, 2000, pp. 336–344.
[9] C.S.G. Khoo, S. Chan, Y. Niu, A. Ang, A method for extracting causal knowledge from textual database, Singap. J. Libr. Inf. Manag. 29 (1999) 48–63.
[10] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: Fellbaum, 1998, pp. 265–283.
[11] J. Moffett, et al., A model for a causal logic for requirements engineering, Requir. Eng. 1 (1996) 27–46.
[12] C. Puente, E. Garrido, J.A. Olivas, R. Seisdedos, Creating a natural language summary from a compressed causal graph, in: Proc. of the Ifsa-Nafips'2013, Edmonton, Canada, 2013.
[13] C. Puente, A. Sobrino, J.A. Olivas, R. Merlo, Extraction, analysis and representation of imperfect conditional and causal sentences by means of a semi-automatic process, in: Proceedings IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2010, Barcelona, Spain, 2010, pp. 1423–1430.
[14] B. Rink, C.A. Bejan, S. Harabagiu, Learning textual graph patterns to detect event relations, in: Proceedings of the Twenty-Third Int. Florida Artificial Intelligence Research Society Conference, FLAIRS, 2010, pp. 265–270.
[15] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of International Conference on New Methods in Language Processing, 1994.
[16] P. Spirtes, et al., Causation, Prediction and Search, The MIT Press, 2001.
[17] G. Varelas, E. Voutsakis, P. Raftopoulou, G.M.E. Petrakis, E. Milios, Semantic similarity methods in WordNet and their application to information retrieval on the Web, in: WIDM'05, Bremen, Germany, November 5, 2005.

[18] E. Vogt, et al., The Art of Powerful Questions: Catalyzing Insight, Innovation and Action, Whole Systems Associates, 2003.
[19] Z. Wu, M. Palmer, Verb semantics and lexical selection, in: 32nd Annual Meeting of the Association for Computational Linguistics, Resnik, 1994, pp. 133–138.