# Analysis, detection and classification of certain conditional sentences in text documents

**Cristina Puente**
Computer Science Dept.
Advanced Technical Faculty of Engineering - ICAI
Pontificia Comillas University, Madrid, Spain.
cpuente2@upcomillas.es

**José A. Olivas**
Information Technologies and Systems Dept.
University of Castilla-La Mancha.
Ciudad Real, Spain.
joseangel.olivas@uclm.es

## Abstract

This paper presents the main lines of investigation in the detection, classification and analysis of certain causal sentences in text documents. It also provides a few ideas to apply Soft-Computing techniques in order to analyze the semantic characteristics of the selected sentences.

**Keywords:** Web search, causal relations, deduction, Soft-computing.

## 1 Introduction

As we know, the most of search engines on the Internet are based on lexicographic indices, that is to say, they search for words that the user introduces in a query, which are possibly found in documents stored on the Web. The main problem is that the engine loses the capacity to recover documents conceptually related to the original request, or through the use of synonyms or antonyms.

There are many studies concerning "conceptual searches" (Crestani, [1], Olivas, [2], etc.). Also to this end, if techniques of Artificial Intelligence tolerant of imprecision and uncertainty (called Soft-computing by Professor Zadeh, the creator of Fuzzy Logic) more closely related to the form of reasoning and expression used by human beings are used, we have a new scenario in front of us, which shows promising in the area of retrieval information from the Internet in an intelligent way.

However, as Professor Zadeh proposed [6], we can go one step further and try to give search engines deductive capacities; in other words, they would be able to answer questions like " What is the third largest province on the Iberian peninsula?".

One way to attempt to do this could be through the analysis of causality and the relationship between concepts involved in the same sentence. This problem, the basis for this study, has been the subject of much discussion during the last few years.

Professor Zadeh maintains [7] that a generic definition of causality can not be established, it depends on multiple factors and it can not be lumped together with classical logic or with probability theories. Though it can be applied in an effective way without the need for a general, rigorous and operational formulation, since the factors to be dealt with are more complex than the situations in which the probability theory is applied.

There are two basic types of causality: first; the so-called "forward causality" expressed in the form: "What are the effects caused by a concrete event?" and the inverse causality, expressed in the form "What actions have been provoked by a certain event?". In context, the forward causality is easier to deal with than the inverse causality, because, the action involved is usually

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1097–1104
Torremolinos (Málaga), June 22–27, 2008

known. In the inverse causality there can be multiple factors that have provoked an action, therefore it is much more complex to deal with and analyze.

Since the objective of this project is to provide a contribution which could improve the efficiency of the Web search, the causality analysis is focused first on the extraction of conditional sentences within a specific document for subsequent analysis. If the analysis concentrates on the semantic aspects, some conditional sentences extracted can not be treated as causal, as can be seen in the following example: "If I pass, I'll stop calling myself John" where the fact that the person is named John has nothing to do with passing.

To catalogue a sentence as causal, it has to fulfill at least three conditions:

1. The cause must precede the effect.

2. Whenever the cause takes place, the effect must be produced.

3. Cause and effect must be closely related.

Not all conditional sentences meet these three premises, therefore, taking this into consideration and since the syntactic analysis is much less complex to deal with than the semantic, in this paper a preliminary analysis is presented on how to approach the detection of conditional phrases in texts from their syntactic units. For this reason, the grammatical analysis on which the study is based will be explained, followed by the detection process, the classification process, the fulfilled tests, and a brief introduction to separate causal sentences from the whole set of the conditional sentences extracted.

## 2 Syntactic analysis: Conditional structures based on a verbal form

In [3], a syntactic analysis of conditional sentences in the English language is presented. According to what we observed there were two types of structures within the conditional sentences. On one hand, there were sentences in which some verbs determined the conditional form of the sentence, on the other hand another type of structure could be found in which the conditional function was indicated by conjunctions in some cases and by adverbs in others.

This study showed a series of labeled structures defining the types of conditional sentences, indicating in each case the verbal tenses that must exist.

Since the number of analyzed structures was very high, we decided to implement a representative subset, as a first step in the creation of a prototype able to detect and classify them. Specifically we decided to analyze the structures from the English language belonging to the conditional in its classic form, "if x then y", corresponding to the first, second and third conditional, and some others that can form these types of sentences too. This produced as a result the 20 structures which are defined as follows and that served as the input and basis for the detection and classification processes:

Structure 1: if + present simple + future simple.

Structure 2 : if + present simple + may/might.

Structure 3 : if + present simple + must/should.

Structure 4 : if + past simple + would + infinitive.

Structure 5 : if + past simple + might/could.

Structure 6 : if + past continuous +would + infinitive.

Structure 7 : if + past perfect +would + infinitive.

Structure 8 : if + past perfect + would have+ past participle.

Structure 9 : if + past perfect + might/could have + past participle.

Structure 10 : if + past perfect + perfect conditional continuous.

Structure 11 : if + past perfect continuous + perfect conditional

Structure 12: if + past perfect + would + be + gerund

Structure 13: for this reason, as a result.

Structure 14: due to, owing to.

Structure 15: provided that.

Structure 16: have something to do, a lot to do.

Structure 17: so that, in order that.

Structure 18: although, even though.

Structure 19: in the case that, in order that.

Structure 20: on condition that, supposing that.

Figure 1: Set of conditional structures.

## 3 Detection process

The detection algorithm is in charge of filtering the conditional sentences whose make-up is found within the 20 previously mentioned. To perform this process the morphological analyzer *Flex,* created by the *GNU* project and freely distributed, was used.

The key reason for the selection of this morphological analyzer is its compatibility with the programming language *C*, because it allows for the definition of tokens and elements to be detected (in *Lex* language) and the processing of them using the *C* language.

In order to analyze the syntactic units which compose a sentence, a finite state automaton has been built which allows us to carry out this task. In order to do this, we first used a regular grammar (fig.2), from which it will be constructed a finite robot. *Flex* allows for simple definition of each one of the states. In order to develop the detection algorithm, only three states were used, the initial one and two more for conditionals.

1.- <I> → <id> <I>

2.- <I> → . <I>

3.- <I> → if <Conditional>

4.- <I> → <Other-conditionals> <Conditional2>

5.- <I> → EOF <Final>

6.- <Conditional> → <id> <Conditional>

7.- <Conditional> → . <I>

8.- <Conditional> → EOF <Final>

9.- <Conditional2> → if <Condicional>

10. < Conditional2> → <id> < Other-conditionals>

11. < Conditional2> → EOF <Final>

12. < Conditional2> → . <I>

13. <Final> → λ

14. < Other-conditionals> → {for this reason, as a result, due to, owing to , provided that, have something to do, a lot to do, so that, in order that, although, even though, in case, in case that, in order that, on condition that, provided that, on the condition that, supossing that}

Figure 2: Detection process regular grammar.

The process begins in the initial state, where syntactic units are processed. If a token processed matches the conditional conjunction, the automaton will move on to the *detect* state, labeling the sentence as conditional. On the other hand, if a token matches anything defined in the other structures as a conditional token, the automaton will pass to state *detect2* also labeling the sentence as conditional. On the contrary, the process will continue analyzing syntactic units, storing them in a buffer, waiting for a conditional conjunction to come up. If the process receives a syntactic unit with a full stop, it will reject the sentence, emptying the buffer where it had been stored.

If the analyzed syntactic unit matches a conditional conjunction, the automaton will have to determine the position of this token inside the sentence to establish if the sentence consequent has already been processed and is stored or if analysis is still pending, as much for the antecedent as the consequent. A problem related with the processing of this type of conditional sentences in which the consequent is previous to the antecedent is the elimination or detection of determined linguistic formulas which serve as a preamble to the conditional sentence but from a semantic point of view do not add anything to it, as can be seen in this example:

***CONSECUENT: By contrast, if*

***ANTECEDENT: the path integral were over non compact metrics one would have to specify the values of the module at infinity.*

*(Stephen Hawking, Quantum Cosmology, M-theory and the Anthropic Principle)*

In order to correctly manage these types of sentences we thought about establishing a minimum number of characters in the consequent, which would avoid mistakes such as the one shown in the previous sentence, but some sentences were detected in which the linguistic preamble was too long or some determined consequents with semantic significance were not handled correctly because they were shorter than the predetermined limit.

An alternative form of correcting part of this conflict could be to obtain a list of verbs which

meet the introductory functions in a determined context, but due to the problem of linguistic ambiguity already mentioned, we would be closing the doors to the appearance of these same verbs in contexts which could provide important information.

To solve this problem a vector of the most common verbal forms in the English language has been created, including present, past and future tenses.

According to experiments that have been done using different texts, we have observed that in all the detected consequents which provide semantic information to the sentence, appears some verbal form included in the vector, such as: *is*, *are*, *were*, *was*, *had*, *do*, *did*, *done*, *can*, *could*, etc., that is to say, forms which introduce action or movements within the sentence. In order to compile the final list, British grammar was analyzed as the basis, creating a vector of sixty positions.

Once the algorithm detects a verbal form contained in the vector, it activates the corresponding position with a flag, in such a way that when the conditional conjunction is detected, the vector is covered to see if a flag has been activated, thus establishing the sentence division between antecedent and consequent.

### 3.1 Detection process problems

When analyzing British grammar we found that some determined structures which contain the conditional conjunction do not conform to this type of sentences. To deal with the elimination of these structures, linguistic turns of phrase which used if were reviewed, obtaining three types as a result; So if, As if and What if.

So if : The conjunction so together with the conditional conjunction has an introductory function to the sentence which is presented . So in this case has the same function as other conjunctions which may also appear next to conjunction if, and that do not alter the meaning nor modify the conditional sentence type as can be seen in the following example:

*>Line number: 11: He pushed me hard to keep this project going, so if you don' t enjoy it blame him too .*

*(Devil wears Prada, Lauren Weisberger)*

*As if:* This clause generally shows up when the speaker is trying to give an explanation or establishes a comparison between two concepts, not having anything to do with this type of sentence with the conditionals, therefore they must be eliminated from the selected sentences.

*>Line number: 2710: As if Miranda had ever been attuned to other people' s happiness ! " Emily please .*

*(Devil wears Prada, Lauren Weisberger)*

*What if*: The pronoun *what* next to the conjunction *if* is usually employed to formulate conjectures about something in particular, giving rise to sentences which, as the same in the previous case, have little to do with the causal as may be seen in the following example:

*>Line number:1013 : "What if you were sleeping ? " I had stupidly asked.*

*(Devil wears Prada, Lauren Weisberger)*

In order to make the planned process function correctly and taking the analysis of the conjunctions which may appear together with the particle *if* into consideration, we decided to cast aside formulated structures such as *so if*, *as if* and *what if* which have little or nothing to do with conditionals.

## 4 Classification process

Once the detection algorithm has finished the analysis of all syntactic units which compose a conditional sentence, it proceeds to classify them. The generated algorithm for this prototype allows for the distinction of the conditional type detected as a function of the verbal form and the syntactic structure formation, using as a reference the 20 causal structures defined before (fig.1). To perform the classification process, the automaton has been improved by the addition of two new states, *cw* and *wd* which checks the composed verbal forms, since *Flex* defines the tokens by spaces. Therefore, if the verbal form to be processed dealt with *would have*, the

automaton would change to the status *wd* on detecting *would* and would remain there waiting for a verb that could form a composed tense such as *be* or *have*. If the following analyzed token corresponded with *have* the process would activate the position associated to *would have* in the verb vector instead of *would* and would return to the previous status. Once the automaton was enlarged, it proceeded to check the syntactic structures on which the foundations of this application are based. Studying the composition of each one, it was observed that there almost always appeared, whether in the sentence antecedent or in the consequent, a verbal form which marked the classification within one structure or the other. It was also perceived that the verbal form was generally composed of some of the verbs in the form vector implemented for the detection algorithm, therefore, in order to classify a sentence as a determinate type of structure it would be enough to run the aforementioned vector analyzing the verbal forms encountered within the sentence.

For example, if we examine the sentence: *"If the weather is fine, we'll go to the beach"*, the verbal form vector would register *is* in position 2 and *'ll* in position 21. When running through the vector and recognizing positions 2 and 21 as activated, the process would determine that we are dealing with a type 2 structure, with the form if + present simple + future simple as follows:

*>Line number:1 If the weather is fine we 'll go to the beach.*

***CLASSIFICATION: FIRST CONDITIONAL STRUCTURE 2*

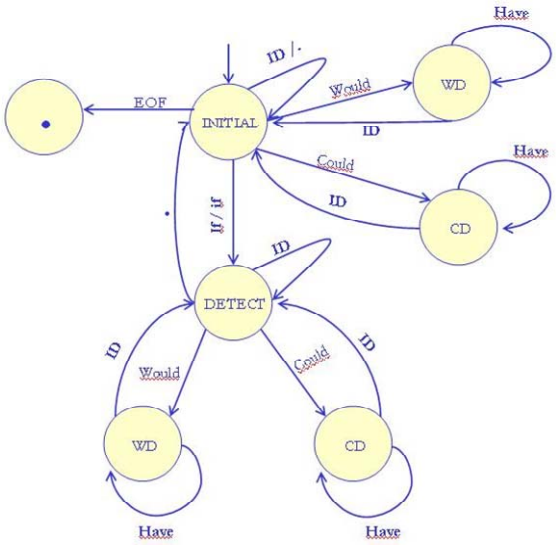


Figure 3:Detection and classification automaton

## 5 Experimental results

Measurement of the application's reliability was done using two types of tests. Firstly, the level of success was checked within the classified structures and secondly, the percentage of correctly classified structures within the control group. In order to do this, eleven texts from different linguistic genres were analyzed.

To evaluate the degree of success of the application within classified structures, the *Quantum Cosmology* speech delivered by *Stephen Hawking* was manually analyzed because it has a more direct and concise language. The results demonstrated that the structures which the process had been able to classify, were correctly labeled (18 classified structures out of 24 detected) except in the following case where the algorithm made a mistake.

*>Line number: 20 even if it were we are not concerned about measurements at infinity but in a finite region in the interior.*

***CLASSIFICATION:SECOND CONDITIONAL STRUCTURE 8*

*-Sentence analysis: Structure type 1*

*- Verb which causes the conflict: were*

Table 1: Texts comparative studio

| Document title | Page number | N°. of words | Detected | Rejects | Classified | Not classified | % Classified |
|---|---|---|---|---|---|---|---|
| Devil wears Prada | 198 | 139.883 | 358 | 41 | 218 | 140 | 60,9 |
| The Historian | 358 | 245.333 | 602 | 253 | 433 | 169 | 71,9 |
| Harry Potter | 288 | 150.666 | 345 | 136 | 251 | 94 | 72,7 |
| Lord of the Rings | 258 | 188.262 | 608 | 152 | 489 | 119 | 79,9 |
| Gospel | 145 | 88.232 | 216 | 1 | 161 | 55 | 74,5 |
| News Science | 7 | 3.069 | 5 | 0 | 5 | 0 | 100 |
| News International | 11 | 4.940 | 8 | 0 | 4 | 4 | 50 |
| Kill Bill script | 249 | 42.618 | 78 | 6 | 65 | 13 | 83,4 |
| Mother Teresa speech | 6 | 4.063 | 15 | 4 | 12 | 3 | 80 |
| Luther King speech | 15 | 8.628 | 22 | 5 | 21 | 1 | 95,4 |
| Quantum Cosmology | 8 | 4.509 | 24 | 0 | 18 | 6 | 70,8 |

The problem when this type of sentence turns up, is that it finds two verbs with different tenses, in past, present or future, and it is unable to distinguish, as in this case, which verb carries more weight in the sentence. Of the 24 detected sentences as conditionals in this text, the process was able to classify correctly 17, that is to say 70,8% of the total, mistaking one sentence and rejecting 6 (25%) as not classified.

In order to analyze efficiency and application results, eleven different texts were used to perform a study based on the results obtained and to be able to compare them. The documents selected were the following:

- Gospel according to Saint Matthew, Saint Mark, Saint Luke and Saint John.
- Harry Potter and the Half Blood Prince (R.K. Rowling).
- Lord of the Rings (J.R.R Tolkien).
- Devil Wears Prada (Lauren Weisberger).
- The Historian (Elisabeth Kostova).
- I Have a Dream (Martin Luther King).
- Nobel Peace Prize speech (Mother Teresa de Calcutta).
- Movie script Kill Bill (Miramax 2004).
- Technology News (The New York Post 5/4/2007).
- International News (Herald Tribune 5/4/2007).
- Quantum Cosmology (Stephen Hawking).

To compare the texts, several aspects such as text length, number of words, detected and rejected structures, in summary the factors shown in Table 1 which will serve as the basis for later analysis, and will have to be taken into account .

The detected structure column corresponds to the sentences which the process has selected from the total as conditionals. The numbers in the columns labelled *classified structures number, not classified and classified percentage,* are calculated using as reference the detected structures, without taking into account the number rejected. The rejected structure column shows those sentences which even though containing the conditional conjunction *if* do not have a conditional meaning. If we analyze Table 1 and calculate the mean percentage of classified structures, we see that the process is capable of dealing with 76.87% of the total number of detected structures (the rejected have not been taken into consideration).

If this calculation is combined with the percentage of reliability we obtained in the previous paragraph through manually checking the results (95%) we find that the application has a combined degree of reliability and success rate of 73,02%. As the results show, the lowest percentages when classifying a sentence correspond to the analyzed novels, no matter which genre they belong to. This is due to the type of language, which usually contains a larger number of metaphors, linguistic turns,

descriptive elements than a script or speech, which increases the complexity of the language as shown in the results. However texts with more precise language demonstrate a higher percentage of classified structures with the exception of international news and political articles extracted from the Herald Tribune, and the New York Times newspaper.

## 6 Causal sentences analysis

As we have mentioned before, a conditional sentence has to fulfill certain features to be consider as causal. Our next step then, is to extract these causal sentences within the whole group of possible conditionals selected.

To achieve this we have employed a stop-word list, in order to erase useless information from a sentence, and to leave the words that are more relevant such as verbs and nouns.

We found several stop-word lists to use in our program, but in the end, we had to create a new one based on the ones we found, due to the fact that there are some adverbs and conjunctions like *do, for*, and some others that may be relevant to our study.

So that, a sentence like "*If the weather is fine we 'll go to the beach."*, would become *" weather is fine 'll go beach"*. In this way it is easier to analyze the concepts which exist in a sentence, which in turn will make it easier to determine whether sentence is causal or not, for future projects.

## 7 Fuzzy techniques application

In [4], the bases are laid for what could constitute conditional sentence analysis from the point of view of fuzzy logic. In this project the problem is posed of the validity of a rationale based on the analysis of the conditional, which through an example of a very complex sentence of Vila-Matas, analyses the effect that it would have on the sentence, the analysis being based on diverse fuzzy implications. Using this job as a basis, it is being defined two investigation lines. The first deals with assigning fuzzy values to antecedents and consequents (in [4] line) in order to be able to trigger deduction mechanisms, representing the sentences identified through Protoforms [5], [8].

The second is trying to establish some type of fuzzy relation index between the concepts which underlay the antecedent and the consequent of the classified sentences, based on the grammatical type of the conditional and the fuzzy implication that could be associated with it, in order to provide another useful parameter (together with others such as synonyms, etc) in the relationship between concepts, with the aim of conceptually improving Web searches.

## 8 Conclusions

The tests performed and the results obtained lead us to two basic conclusions. The first is that the appearance of a major number of conditional structures has little to do with the volume of the analyzed text, but to its literary style and the manner of speech as has been shown in Table 1.

The second shows that the results obtained are coherent with the types of structures studied and adjust themselves to the defined standards, which explains the high percentage of classified structures compared to those identified. Finally, in part seven, we have suggested some lines of investigation applying fuzzy logic to the results obtained, which are intended to serve as a guide and orientation for subsequent projects.

## Acknowledgements

Partially supported by PAC06-0059 SCAIWEB project, JCCM, Spain, and TIN2007-67494 F-META project, MEC-FEDER, Spain.

## References

[1] Crestani, F.; Pasi, G. (2000). Soft computing in information retrieval: Techniques and applications, Physica Verlag, Series studies in fuzziness.

[2] Olivas, J. A.; Garcés, P.; Romero, F. P. (2003). An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents. International Journal of Approximate

Reasoning (Soft Computing in Recognition and Search) 34, 201 - 219.

[3] Olivas, J. A.; Puente, C.; Tejado, A. (2005). Searching for causal relations in text documents for ontological application. Proc. of the Int. Conf. on Artificial Intelligence IC-AI'2005, vol. II, CSREA Press, 463 – 468.

[4] Trillas, E. (2004). Lógica borrosa y narrativa: un párrafo de Vila–Matas, XII Congreso español sobre tecnologías y lógica fuzzy, Jaén, 2004, 23-34.

[5] Zadeh, L. A. (2001). A Prototype-Centered Approach to Adding Deduction Capability to Search Engines -- The Concept of Protoform, Berkeley Initiative in Soft Computing, Submitted to the BISC mailing list 19th Dec 2001.

[6] Zadeh, L. A. (2003). From search engines to Question-Answering System: The need for new tools, E. Menasalvas, J. Segovia, P.S. Szczepaniak (Eds.): Proceedings of the Atlantic Web Intelligence Conference - AWIC'2003. LNCS, Springer, 15 - 17.

[7] Zadeh, L. A. (2004). Precisiated Natural Language (PNL). AI Magazine, Vol. 25, No. 3, 74-91.

[8] Zadeh L. A.(2006). From Search Engines to Question Answering Systems: The Problems of World Knowledge, Relevance, Deduction and Precisiation. In Fuzzy Logic and the Semantic Web, edited by Elie Sanchez, Elsevier 2006.