Finding Instances of Deduction and Abduction in Clinical Experimental Transcripts

Maria Amalfi[1], **Katia Lo Presti**[2], **Alessandro Provetti**[3] and **Franco Salvetti**[4]

Abstract. This article describes the design and implementation of a prototype that analyzes and classifies transcripts of interviews collected during an experiment that involved lateral-brain damage patients. The patients' utterances are classified as instances of categorization, prediction and explanation (abduction) based on surface linguistic cues. The agreement between our automatic classifier and human annotators is measured. The agreement is statistically significant, thus showing that the classification can be performed in an automatic fashion.

1 Introduction

This article describes a software that we have designed and implemented in the context of our work on measuring utterances that are evidence of inferential ability, viz., prediction and explanation, trying to match a human annotator. The starting point was an experiment that involved patients with lateral-hemisphere brain damage (along with a control group) that was carried out at the Boston VA hospital and for which the results were available in writing. The experiment consisted in showing to the subject the picture in and asking them *"I have a picture here. As you see, there's a lot going on. Look it over and tell me about it."*

In this article we are not concerned with the cognitive interpretation of the data, nor with the validation of a particular hypothesis relating lateral brain damage to particular types of reasoning impairment. Rather, we would like to validate our approach to automated text classification by showing, via statistical analysis, that the results are comparable with those of human annotators. Our architecture is designed to avoid commitment to any particular model of rationality but could serve as a tool for validating Cognitive Science theories.

Indeed, one could say that the relatively simple software architecture described here is effective for textual analysis only when simple sentences having a limited lexicon are considered. However, the advantage of using an automated tool will be felt when similar experiments will be administered to large populations and human annotation will become uneven or even impossible. Although our software, which has SWI Prologat its core and is described next, is not suitable for large-scale activities *as is,* standard computational complexity analysis yields that our approach can indeed scale up to several hundred transcripts.

Let us now describe in detail the architecture and the data representation adopted in this project[5] The system consists of two main components that have been designed and implemented for the experiment described above.

1. program *patients-aggregator* takes as input data:
   - the patients' transcript;
   - local vocabulary [6] and
   - the general-purpose deduction rules for the Prolog inferential engine

   and produces:
   - rules data about categorizations, explanations and predictions;
   - patients' data

   Patients-aggregator scans the transcripts and creates a suitable Prolog representation of the phrases. Then, it also produces the schematic rules that the subsequent Prolog interpretation will use to discover the instances of prediction and explanation in the transcripts.
2. program *inference-finder* is written in Prolog; it takes as input the data generated by patients-aggregator and it produces:
   - statistics;
   - patient predictions and explanations and
   - firing rules.

2 Results and comparisons

All the interviews considered in this work were annotated by two independent panels of human annotators, here called B and M. Each panel was made of two graduate students of Computer Science, who received similar instructions and very precise instructions on how to annotate interviews.

2.1 Validation of the results

The overall No. of annotations obtained during our experiment is illustrated in Table 1.

In the following we described a statistical analysis of the results that supports a more sophisticated understanding of the results.

The Kappa index [1], introduced by Cohen [2], has been, proposed as a measure of the specific agreement for category among two observers. Kappa measures the accord among the answers of two observers (or the same observer in different moments), that appraises couples of objects or diagnostic categories.

[1] Grad. Student at University of Messina, Italy. *maria.amalfi@gmail.com*

[2] Grad. Student at University of Messina, Italy. *katia.lopresti@gmail.com*

[3] Dept. of Physics University of Messina, Italy. E-mail: *ale@unime.it*

[4] Umbria Inc. Boulder CO U.S.A. E-mail:*franco.salvetti@umbrialistens.com*

[5] The software (source and binary codes), the results and documentation is available from our group page: *http://mag.dsi.unimi.it/*

[6] The number of words of interest is finite and rather limited, i.e., 149 words as categorizations. So, it has been possible to represent all tokens of interest by means of Prolog facts.

Instances Found		
-	expl.	pred.
B panel	99	71
M panel	87	35
Program	71	32

Table 1. Overall no. of instances found

This index is captures and corrects the so-called *accidentals agreements,* An agreement is called accidental when two observers reach the same conclusion even though they did so by employing completely different sets of criteria to distinguish between the presence/absence of relevant conditions. In such cases the *raw* agreement index would not reflect a real agreement. The idea underlying the Kappa index is that the actual accord between two observers is as the difference between the raw agreement and the agreement we would have under the hypothesis that between the two there is no accord and thus their answers may coincide only by chance.

The value of K is given by the ratio between the excess agreement $(P_o - P_e)$ and the maximum obtainable agreement $(1 - P_e)$:

$$K = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where:

P_o it is the proportion of frequencies observed of accords among the two evaluation, and

P_e it is the proportion of accords obtained under the condition that the void hypothesis is true.

If there is complete agreement, then K will be equal to 1. If the observed agreement is greater or equal then the agreement attended only by chance obtained then the K index will result near zero or even slightly negative. Values of K above 0.75 suggest that there is an excellent agreement; values below to 0.40 represent a weak agreement, whereas values between 0.40 and 0.75 can represent a good agreement. To sum it up, k is the right type of index to assess the quality of our program vis-à-vis human analysis of some experimental results. The degree of agreement among the human panels and the program is as follows:

Explanations			
-	B	M	Program
B panel	1	0,331	0,335
M panel	-	1	0,433
Program	-	-	1

Table 2. The K degree of agreement on explanations

Between the M panel and the system there is a good enough agreement (values between 0.75 and 0.40); while group B has weak agreement both with group M and with the system (values below 0,40).

2.2 Interpretation of the results

The statistics described above show a good agreement between the program scores and those given by the M panel. Vice versa, the B panel results have a low agreement index K with both the program and the M panel. The B panel consistently finds more instances of reasoning (of any type) than the M panel and the system. These differences can be explained by the fact that the mental model of the M panel annotators is reflected in the program.

Predictions			
-	B	M	Program
B panel	1	0,244	0,260
M panel	-	1	0,447
Program	-	-	1

Table 3. The K degree of agreement on predictions

These results are very satisfying from an Artificial Intelligence perspective: they show that, e.g., from the point of view of the B panel, the classification expressed by the M panel and that of the system *are indistinguishable.*

3 Conclusions

We have described the the design and implementation of a prototype that analyzes and classifies transcripts of interviews collected during a cognitive science experiment that concerned assessing reasoning bias in lateral-brain damage patients. Our Prolog-based software takes a static description of reasoning rules and matches them on patients' transcripts. Hence, patients' utterances were classified as instances of categorization, prediction and explanation (abduction) based on surface linguistic cues. The agreement between our automatic classifier and human annotators is measured. The agreement is statistically significant, thus showing that the classification can be performed in an automatic fashion. The statistical results support our claim that our software can be safely applied to automate the analysis of experimental results of the type described earlier. Our program can be useful as a provider of second opinions to reveal possible overlooks or mistakes in the diagnostic analysis.

From a Cognitive science point of view, our project may be considered limited by the fact that is can analyze only verbal (transcripted) responses to experiments. Vice versa, from an A.I. point of view we can see that the pattern matching mechanism, though rather basic vis-à-vis current natural language processing techniques is implemented fairly elegantly and efficiently in Prolog.

We are currently working to incorporate such techniques, (e.g., regular expressions) into our program. It would be interesting to apply our classifier to the transcripts of the experiment in [3] since their experiment seems within the reach of the techniques we have employed. Another promising direction of research consist in attaching to the token words some *semantics* obtained by automated reference to Wordnet[7].

References

[1] Jean Carletta, 'Assessing agreement on classification tasks: the kappa statistic', *Computational Linguistics*, **22(2)**, 249–254, (1996).

[2] J. Cohen, 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, **20(1)**, 37–46, (1960).

[3] Katiuscia Sacco, Monica Bucciarelli, and Mauro Adenzato, 'Mental models and the meaning of connectives: A study on children,adolescents and adults', *Proc. of the XXIIIth Conf. of the Cognitive Science Society*, 875–880, (2001).

[7] http://wordnet.princeton.edu