Modelling the Interpretation of Novel Compounds

Dermot Lynott and Mark T. Keane [1]

Abstract. The understanding of novel compounds is a special case in which we can explore the deep generativity of natural language understanding. We report a model, PUNC, which captures the comprehension of novel noun-noun compounds. The model constructs multiple interpretations for a given compound, ranking these by their overall acceptability, using the key constraints of diagnosticity, plausibility and informativeness. We present a sensitivity analysis of the model's key variables, demonstrating a graceful degradation in performance as the weightings to these variables are altered, thus vindicating the model's underlying theoretical principles.

1 INTRODUCTION

Conceptual combination refers to the process of combining of existing concepts to create new meanings. This phenomenon is manifested particularly in novel (i.e., new) noun-noun compounds e.g., *trash cookies*, *daisy cup*, [1, 2, 3]. Part of the reason for examining the comprehension of novel, noun-noun compounds lies in the window they open onto the deep, generativity in natural language understanding. The PUNC (Producing and Understanding Novel Compounds) model of conceptual combination [3] is a cognitive model of conceptual combination, which has been shown to successfully reflect not only the range of interpretations that people produce, but also the efficiency displayed by people in understanding word combinations they've never heard before. The workings of the model have been described in detail elsewhere [3], so in this paper we present a sensitivity analysis of the model's key parameters.

2 PUNC: Producing and Understanding Novel Compounds

The PUNC model has been inspired mainly by the Constraint Theory [2] of conceptual combination (CT). CT explains how people construct interpretations for a given compound by combining the modifier (i.e., the first element) and head (i.e., the second element) concepts of a compound, using the constraints of diagnosticity, plausibility and informativeness. For example, in the compound *cactus beetle*, "cactus" is the modifier, while "beetle" is the head and possible interpretations could be a) a beetle that eats cacti, b) a prickly beetle, or c) a beetle that can conserve water.

Diagnosticity refers to the features of a concept that best distinguish it from other concepts. For example, the prickly nature of cacti is more diagnostic than the fact that they are green in colour. Plausibility means that interpretations must display a link between concepts that is consistent with our prior knowledge of those concepts. For example, take two candidate interpretations for the compound *angel pig* – (i) An *angel pig* is a pig with wings on its torso, (ii) An *angel pig* is a pig with wings on its tail. In this case, interpretation (i) is considered more plausible because we have previously encountered entities with wings on their torsos, whereas we have not encountered entities with wings on their tails. Finally, informativeness means that a valid interpretation must provide some new information relative to the original compounds. For example, the interpretation "a spiky hedgehog" would not be considered an informative one for the compound *cactus beetle*. This is because hedgehogs are spiky, so the description provides no new information about the head or modifier concepts.

PUNC is a computational implementation of the constraints specified in CT. The model is the first to incorporate both comprehension and production components [4], but here we focus exclusively on the comprehension side of the model. As input, the

$$(1)\ Acceptability = \log_{j}(kp \times k[ch + cm])$$

model takes a noun-noun compound and outputs a set of candidate interpretations, with each interpretation assigned an acceptability score.

The acceptability of an interpretation is given by the formula in (1), where p represents the plausibility of the interpretation, h is the diagnosticity of the head concept's feature and m is diagnosticity of the modifier's feature. The diagnosticity of the head and modifier concepts is weighted equally, by c, while the overall diagnosticity and the plausibility are also weighted equally by k. This value is then logged (to the base j = 0.15) to allow direct comparison to human performance.

It has been shown that PUNC can closely model both the interpretations that people produce and also the relative frequency with which they are produced, with correlations greater than 0.6 for large sets of novel compounds [3].

3 SENSITIVITY ANALYSIS

For any cognitive model there are a number of variables that contribute to its performance. Some of these variables may be theoretically motivated (e.g., diagnosticity), while some may simply be implementational details of the model (e.g., constants). In PUNC, the two variables used in calculating the acceptability of interpretations - diagnosticity and plausibility - have been theoretically motivated. We focus on these factors and the extent to which they contribute to the performance of the model. If these factors are central to the calculation of interpretation acceptability, then varying the weight of their contribution to the acceptability function should result in a graceful degradation in the model's ability to simulate human performance. If the sensitivity analysis reveals no such degradation, then we can conclude that the model's power is stemming from the specific implementation, rather than these theoretically motivated factors.

3.1 Relative Importance of Variables

Before examining the model's sensitivity to changes in the key

[1] Department of Computer Science, University College Dublin, Dublin 4, Ireland, email: dermot.lynott@ucd.ie, mark.keane@ucd.ie

variables, it is important to ascertain the extent to which our variables contribute to the acceptability function. The relative importance of the variables can be established in two steps. First, we must generate a space of all possible values for both variables, calculating the acceptability for each combination of values. Second, we perform a linear regression on these values to calculate the beta coefficients of each. The beta coefficients indicate the extent to which a variable contributes to the model, and the regression analysis also tells us whether this contribution is significant.

We constructed a space of all possible values and generated the acceptability score for each combination using the formula given in (1). A regression analysis revealed that diagnosticity contributed slightly more to the model than plausibility, with beta coefficients of 0.589 and 0.538 respectively, and with both variables' contributions being significant (p's < 0.001). This result suggests that both variables are central to the model's performance, but a more rigorous analysis is required to test the robustness of the model in response to changes in these variables.

3.2 Model Robustness

We have seen previously [3, 4] that PUNC's acceptability scores correlate well with people's frequency of production of interpretations. In this section, we examine the model's sensitivity to changes in the contribution of both diagnosticity and plausibility in the acceptability function. A systematic variation of the weights applied to each variable will give us a picture of their relative importance in the system. If the model is robust, we should see a graceful degradation of the model's performance (when compared to human data), rather than no change whatsoever or a catastrophic drop in performance.

The results of the sensitivity analysis are presented in Table 1. The table shows the weights applied to plausibility on the horizontal axis and diagnosticity on the vertical axis. Although diagnosticity is calculated by combining the head and modifier diagnosticity information, both are equally weighted in the formula. Therefore, for simplicity of presentation we compare only plausibility to the overall Diagnosticity. Weightings of 1-75% indicate a lighter weighting, 100% representing no change, and 125-200% indicates an increase in the variable weighting. Each entry in the table indicates the correlation between the model scores for interpretations and the human frequency of production for the same interpretations. For example, the entry ">0.6" means the *r* value for the correlation is greater than 0.6.

The sensitivity analysis reveals a key region where the model best reflects the human data, indicated by the shaded area in Table 1. We can see that when the Plausibility and Diagnosticity factors are given a weighting around 75%, performance is most stable. The central square of this area is the only one where all adjacent squares have correlation scores of at least 0.6; moving away from this central area reveals a graceful degradation in the model's performance. As the plausibility weighting is increased, performance degrades because this has the effect of making good plausibility scores look poor, thus decreasing the correlation. Similarly, as we decrease the plausibility weighting performance degrades as the model is relying solely on the diagnosticity values to calculate interpretation acceptability. This shows that both variables must contribute for the model to accurately reflect human performance.

The sensitivity analysis reveals that the model's performance is best when the weightings of both diagnosticity and plausibility calibrated between 50% and 150%. Within this area, represented by the grey shading in Table 1, the correlation with human responses does not fall below 0.6. Given that there is a range of weightings where the model can perform optimally, we can conclude that the model is not reliant on a fixed set of parameters, but displays a graceful degradation in performance, in keeping with a separation of theoretical and implementational details [5].

5 CONCLUSIONS

In this paper we have presented a sensitivity analysis of the PUNC model of conceptual combination [3, 4], based on the constraints originally proposed by Costello and Keane's Constraint Theory [2]; diagnosticity and plausibility. The analysis reveals that these factors are equally important in the models ability to simulate human responses. Moreover, in altering the model's weightings, a graceful degradation in performance was evident, validating the model's theoretical motivations.

ACKNOWLEDGMENTS

This work has been funded by a grant from the Irish Research Council for Science, Engineering and Technology under the Embark Initiative under Grant No.03/IN.3/I361 to the first author.

REFERENCES

[1] G. Cannon. *Historical change and English word-formation*. New York: Lang, (1987).

[2] F. J. Costello & Keane, M. T. Efficient creativity: Constraints on conceptual combination. *Cognitive Science*, **24**, 2, 299-349, (2000).

[3] D. Lynott, G. Tagalakis & M. T. Keane. Conceptual Combination with PUNC. *Artificial Intelligence Review*. In press.

[4] D. Lynott (2004). *Comprehension and Production in Conceptual Combination*. Ph.D. Thesis, University College Dublin, Ireland.

[5] R. Cooper, J. Fox, J. Farringdon, and T. Shallice. A systematic methodology for cognitive modelling. *Artificial Intelligence*, **85**, 3-44, (1996).

Table 1 Sensitivity analysis

Weight for Diagnosticity of Interpretation	Weight for Plausibility of Interpretation								
	1%	25%	50%	75%	100%	125%	150%	175%	200%
1%	>.45	>.45	>.45	>.45	>.45	>.45	>.5	>.5	>.5
25%	>.45	>.5	>.55	>.55	>.6	>.6	>.6	>.6	>.6
50%	>.45	>.55	>.6	>.6	>.6	>.6	>.6	>.6	>.6
75%	>.45	>.55	>.6	>.6	>.6	>.6	>.6	>.55	>.55
100%	>.45	>.6	>.6	>.6	>.6	>.55	>.55	>.5	>.5
125%	>.45	>.6	>.6	>.6	>.55	>.55	>.5	>.5	>.45
150%	>.5	>.6	>.6	>.6	>.55	>.5	>.45	>.45	>.4
175%	>.5	>.6	>.6	>.55	>.5	>.5	>.45	>.4	>.35
200%	>.5	>.6	>.6	>.55	>.5	>.45	>.4	>.35	>.35