Finding Semantic Similarity in a Biological Domain: A Human-Centered Approach

Takashi Yamauchi and Na-Yung Yu

Mail Stop 4235, Department of Psychology
Texas A&M University College Station, TX 77843 USA
tya@psyc.tamu.edu, dbskdud40@tamu.edu

A behavioral study investigated how college students judge similarity between cell pictures. The study indicates that there is a strong tendency to rely on class-inclusion relations in judgments of similarity. This means that biological concepts are likely to be organized and conceptualized with respect to class-inclusion relations even for non-experts.

1 Introduction

Concepts are assemblies of knowledge that are developed, construed, modified, and constructed by people interacting in a particular domain. This means that the concepts that people form, which are the medium of ontology matching, are necessarily influenced by the way people process, represent, and retrieve information.

In this brief article, we will illustrate how college students who do have no special training in medicine judge semantic similarity among cell pictures, and show that there is a strong predisposition for lay people to rely on class-inclusion relations to determine conceptual similarity.

2 Study 1

Fig. 1 shows variations of two different animal tissues. The two pictures placed at the top of the two frames are original cell pictures (i.e., target pictures) and those at the bottom are morphed images of the two original pictures (i.e., base pictures). In this study, participants (undergraduate students, N=227) were presented with 60 triads of cell pictures similar to those shown in Fig. 1 and they judged which base picture, left or right, was more similar to the target picture placed on the top.

The question of interest was the effect of labeling. We hypothesized that class-inclusion relations are particularly important for the conceptualization in the biological domain; biological concepts are arranged and understood in the context of how entities relate to one another in their taxonomical relations rather than in their concrete appearance, attributes, or properties [1] [2].

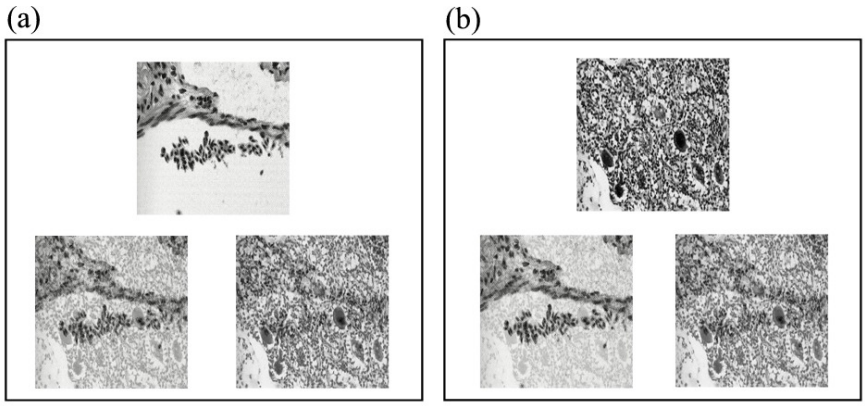

Fig. 1. Two samples of the stimulus frames used in Study 1. The base pictures (shown at the bottom) were produced by merging the two cell pictures shown at the top.

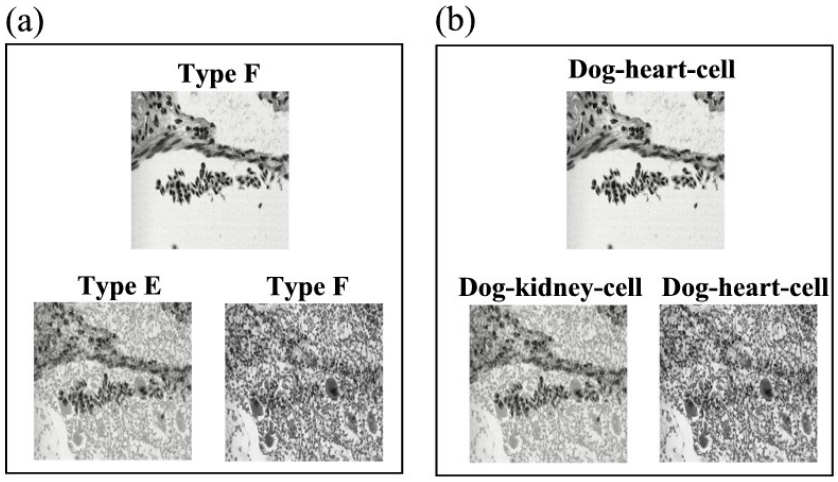

Fig. 2. Two samples of the stimulus frames in the alphabet-label and cell-label conditions in Study 1. Meaningless alphabetical labels are attached in (a) ("Type F" and "Type E"); meaningful verbal labels are attached in (c) ("Dog-heart-cell" and "Dog-kidney-cell").

To test this idea, we examined how the ontological labels attached to the cell pictures would influence participants' judgments of similarity. In one condition, no labels were attached to the pictures (control condition, Fig 1). In another condition, meaningless verbal labels were attached to the same cell pictures ("Type E" and "Type F," Fig 2a). In the other condition, fictitious yet conceptually meaningful labels were attached to the same cell pictures ("Dog-kidney-cell" and "Dog-heart-cell," Fig

2b). Given these three conditions, participants judged which cell pictures, left or right, were more similar to the target picture placed on the top.

2.1 Method

2.1.1 Materials

We produced a total of 60 triads from 5 pairs of original cell pictures. For each pair, one original picture was merged with the other original picture in different degrees, creating three groups of morphed pictures for each pair (low-, medium-, or high-level groups; see Fig. 3). These cell pictures carried different types of labels depending on the condition to which participants were assigned.

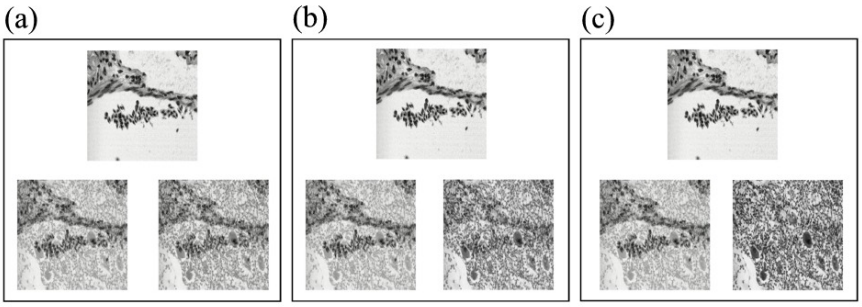

Fig. 3. Samples from three groups of morphed pictures.

2.1.2 Procedure

Sixty triads of cell pictures were presented to each participant one at a time at the center of the computer screen. Participants selected one base picture that was deemed similar to the target picture. The order of presenting the stimuli was determined randomly, and the location of placing base pictures (either left or right) was also determined randomly.

2.1.3 Design

The experiment had one between-subjects factor: (label condition; no-label, alphabet-label, cell-label). In the no-label condition, no pictures carried labels (Fig. 1). In the alphabet-label condition, the pictures carried meaningless alphabetical labels (Fig. 2a). In the cell-label condition, the pictures carried meaningful labels (Fig. 2b). Thus, the conceptual relations between the cell pictures were unclear in the no-label and alphabet-label conditions, but the conceptual relations were clear (e.g., heart vs. kidney) in the cell-label condition.

We employed two measures to assess the effects of labeling. First, we examined the proportion of participants selecting dissimilar base pictures as more similar as these cell pictures carried different kinds of labels. For example, we measured the proportion of participants selecting the base picture on the right in Fig. 2b; this base

picture was less similar to the target than the other base picture (the one on the left), so measuring the proportion of selecting dissimilar base picture would tell us the extent to which labels override perceived similarity. Second, we examined the impact of labeling in two situations – one in which the target and dissimilar base pictures had the same labels (i.e., the same-label condition, Fig. 2a and 2b), and the other in which the target and dissimilar base pictures had different labels (i.e., the different-label condition). These two conditions were produced by simply swapping the assignment of the labels to the base pictures. For example, in the different-label condition, "Type E" and "Type F" given to the two base pictures in Fig 2a were swapped.

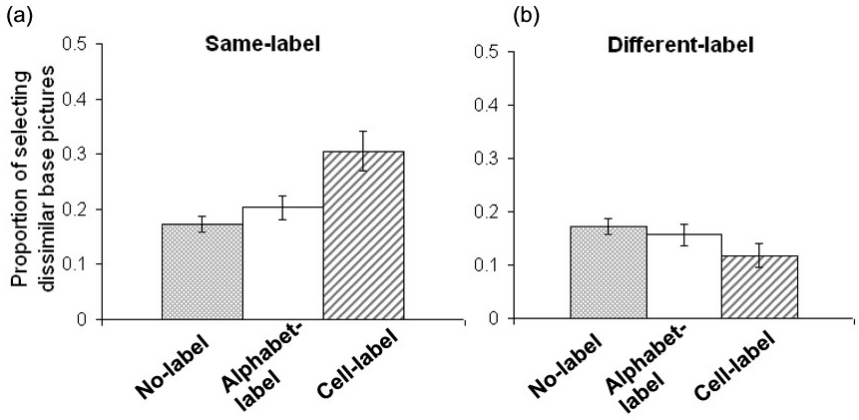

Fig. 4. A summary of the results from Study 1. The error bars represent two standard error units obtained from each condition.

2.2 Results

When the target and the dissimilar base pictures had the same label (Fig. 4a), the proportion of selecting the dissimilar base pictures increased significantly. When the target and the dissimilar base picture had different labels (Fig. 4b), the proportion of selecting the dissimilar base pictures declined substantially. This impact of labeling was present only when the pictures carried conceptually meaningful cell-labels, but not when they carried meaningless alphabetical labels.

In the same-label condition (Fig. 4a), the mean proportions of selecting the dissimilar base picture were significantly higher in the cell-label condition (M=0.31) than in the no-label (M=0.17) and alphabet-label (M=0.20) conditions; $F(2, 140)=7.84$, MSE=0.09, $p<0.001$; cell-label vs. no-label, $t(94)=3.54$, $p<0.001$, d=0.72; cell-label vs. alphabet label, $t(93)=2.47$, p=0.02, d=0.51. The proportions of selecting the dissimilar base pictures were not different between the no-label condition and the alphabetical label condition; $t(93)=1.23$, p=0.22, d=0.25.

Given different labels (Fig 4b), the mean proportion of selecting dissimilar base pictures was significantly lower in the cell-label condition (M=0.12) than in the no-

label (M=0.17) and alphabet-label (M=0.16) conditions; $F(2, 129)=4.77$, MSE=0.02, $p<0.05$; cell-label vs. no-label, $t(89)=2.87$, $p=0.005$, $d=0.60$; cell-label vs. alphabet-label, $t(80)=2.06$, $p=0.04$, $d=0.45$.

These results suggest that the labels influenced participants' similarity judgments only when the labels were conceptually meaningful, indicating that the conceptual links between cell pictures were crucial even in the perceptual judgment of similarity of the cell pictures.

3 Discussion

Ontologies are formal descriptions of concepts developed by people, so it appears natural to study how people acquire and use concepts in a given domain in order to develop viable mapping agents. In ontology matching, "similarity" is generally assessed in multiple levels (e.g., lexical, structural, and/or relational levels). The overall similarity between ontologies is specified as a weighted sum of individual similarity measures [3] [4] [5]. However, allocating appropriate weights to these similarity factors is not trivial. Because conceptualization arises from a highly interactive environment in which different sets of goals and constraints are required, the mapping agent needs to incorporate different heuristics and background knowledge to identify adequate weights for similarity factors. The present study suggests that conceptually meaningful class-inclusion relations are crucial even for lay people in determining perceptual similarity among cell pictures.

Acknowledgements. This research was supported by the Glasscock Center Faculty Fellow Award, and a Developmental Grant by Mexican American and U.S. Latino Research Center, Texas A&M University. We would like to thank Wookyoung Jung for her valuable comments.

References

1. Yamauchi, T., Yu, N.: Category Labels versus Feature Labels: Category Labels Polarize Inferential Predictions. Mem Cognition (in press)
2. Yamauchi, T., Markman, A. B.: Inference Using Categories. J Exp Psychol Learn 26 (2000) 776-795
3. Shvaiko, P., & Euzenat, J.: A survey of schema-based matching approaches. Lecture Notes in Computer Science Vol. 3730. Springer-Verlag, Berlin Heidelberg New York (2005) 146-171
4. Noy, N. F. Semantic Integration: A Survey of Ontology-based approaches. SIGMOD Record, 33 (2004) 65-70.
5. Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2004). Ontology matching: A machine learning approach. In S. Staab & R. Studer (eds.), Handbook on Ontologies (pp. 385-403). New York: Springer