Anna Ehrlemark
Embodied and Situated Language Processing
L2308 HT2015

Odd one out

Abstract

This paper introduces an experiment in collecting human judgements about semantic relatedness between objects in the fashion of an unsupervised game, Odd One Out. In contrast to existing human ranked similarity benchmarks, the Odd One Out game places semantic similarity in context by asking the user to single out the most different out of three objects. Initial data collection is showing promising results, but the application needs a little tuning before setting off.

1. Introduction

Odd One Out is a game that plays with similarity and difference between images. In the usual set up the player is presented with four different images of familiar things and asked to single out the one that is most different from the others. In my application I experiment with using the odd one out game as an unsupervised method to collect human judgements about semantic relatedness. The aim is to create an annotation tool that is fun, easy, and productive.

2. Background

An active field in computational linguistics concerns itself with word meaning, semantic similarity and relatedness (Baroni et al., 2014; Clark, 2015). The aim behind the research varies within the field, from exploring statistical data in NLP-applications to making theoretical claims about the cognitive origin of semantic categories in human perception (Lenci 2008) There are many different ways to represent semantic similarity, and an array of methods for extrinsic and intrinsic evaluation of the models (see i.e. Baroni et al. 2014). Here, I will just mention the two most prominent models for measuring semantic relatedness: hand-crafted taxonomies such as WordNet (Fellbaum, 2005) and corpus based distributional semantic models (DSMs).

In the taxonomic word sense database WordNet, words are organized in a hierarchical network with several specified relations, such as hyperonomy, hyponomy, and ISA-relations . Similarity between words is measured by a handful similarity functions that are all based on distance in the tree (Pedersen et al. 2004). Such hand crafted resources are time consuming to build and represent semantic relatedness in a rigid way, but manage to capture the hierarchical structure of word senses organized in classes and types.

Distributional semantic models, on the other hand, exploit the distributional hypothesis that words occurring in the same contexts tend to have similar meanings (Clark, 2015; Lenci, 2008). This method for harvesting semantic similarity scores is unsupervised (thus attractive) and only hungry for large amounts of raw text. In DSMs counting word co-occurrences results in vector representations of word meaning that can be used to generalize over topic clusters or to rank word similarity by cosine similarity between vectors. The results are depending on the corpora used and different methods of weighting the vectors. In addition, experiments have shown that the distributional hypothesis can be successfully transferred to model visual similarity by translating image features to so called 'bags-of-visual-words'; and models that

combine textual and visual relatedness outperform the purely text-based approach in the typical intrinsic evaluation tasks (Bruni et al. 2014).

All of this is getting really interesting when we consider the difficult task of evaluating the semantic models. Evaluation will inevitably mirror the claim of the research, either as a try to effectively model human semantic judgements, or to create useful tools for different NLP-tasks (such as semantic features in a language model). In the former case, that I will focus on here, the standard evaluation method is based on human ranked gold standard benchmarks, such as the WordSimilarity-353 Test Collection (Finkelsten et al. 2002). The WS-353 test set consists of word pairs and human ranked similarity scores from 0 (totally unrelated) to 10 (very closely related), such that for example the pair *food* and *fruit* is assigned a score of 7.52, while *forest* and *graveyard* get only 1.85. Using such a benchmark as evaluation is a simple and effective way to evaluate semantic models, but the underlying assumption is that humans are able to rank similarity in a consistent way, even between distant concepts.

Another important thing to consider before moving on to the next section is the distinction between word similarity and visual similarity. These are obviously different but related fields of interest. While taxonomic models make ontological relations between word senses explicit, DSMs will base semantic similarity on language context and syntactic use, and visual models pay attention to features like shape, texture or color. If we want to model human semantic judgements we will sooner or later be interested in combining these dimensions and expand our models to capture other multimodal senses as well. In the following section I will suggest a new approach to collecting human similarity judgements based on difference instead of similarity. My method is based on visual representation of concepts, but I believe that the method can be expanded to other modalities as well.

3. The game

Here is an idea, how about asking the question differently? The intuition behind the Odd One Out game is that human benchmarks represent an unstable source of information about semantic relatedness, simply because the question is too abstract. How similar is a *pry bar* to a roll of *duct tape* on a scale from 0 to 10? The question demands a context. Here is a better question: Out of a *pry bar*, a roll of *duct tape* and a *hammer*, which one is the most different? Odd one out is based on difference, so the objects are placed in context and helps the player to focus on different aspects of semantic relatedness.

The basic set up of the game is as simple as the instructions. Presented with three different images of things, chose the one that is most different from the others. Then repeat. For the first experiment I decided to build the game in two different versions, one with celebrities and one with tools. That is, one game that produces original output that cannot be evaluated by any of the existing models or benchmarks, and one that plays with a set of objects from a predefined class of objects (tools) that can be compared with for example WordNet taxonomic similarity scores. The images were selected by hand from open sources, and certain effort was made to pick images where the objects were presented in the same scale and fashion.

The game application was created with the python module Flask and set up on a local server as a simple webpage[1]. To play the game the user enters an alias and starts playing. On the screen is displayed a set of three different pictures. For every set of three images the task is to single out the one that is most different from the others, by any property, at the discretion of the user. The logfiles from different players are stored away under the user alias. The compiled logged data is then used to calculate co-occurrence counts in a way similar to

1 The basic set up of the web page in Flask was implemented by Mehdi Ghanimifard, see full credits in the documentation of the Odd One Out application.

semantic vector models of words in their context, but in my implementation the vector counts are collected using a scoring function that both awards similarity and accentuates difference. For each logged session the odd one out image gets downrated in relation to the others, and the images left in the context are awarded with co-occurrence scores while getting downrated in relation to the odd one out (see figure 1).

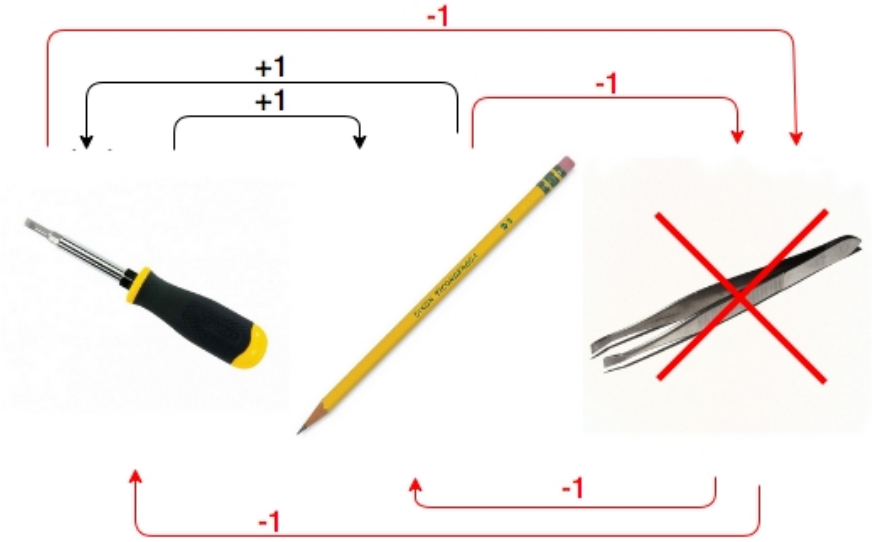

Figure 1. Scoring similarity and difference.

Since the game is entirely unsupervised, the users are encouraged to make judgements based on any salient property of the objects, such as visual similarity, functionality, domain or other. While data from a single user will be very subjective, collected judgements from many different people should capture the gradient nature of similarity and difference between objects, that give a better human benchmark than when simply comparing word senses out of context.

After harvesting the data and implementing the scoring function, image vectors can be compared using simple cosine similarity between image vectors. The final output function takes the name of an image and produces a similarity top list from most similar to least similar in relation to all the other images in the set.

4. Results and evaluation

In the first experiment round of Odd One Out we collected a total of 1080 rounds of judgements on the tools set and 1157 judgements on the celebrity set from five different users. To visualize the similarity function on the fly, we added a result page to the web interface where the user can select and image and see the images sorted by similarity from most similar to most different directly in the interface, based on all data collected up until that moment. Although the collected data is still quite sparse, the visual report gives a good indication that the human judgements are based not only on visual likeness, but also on function and domain.

In figure 2 we can inspect the results page for the image of a *pencil*. Among the tools ranked most similar to a *pencil* we find a *brush, eraser, duct tape* and a *drawing compass*. Although only the brush is similar both in shape and function, all the objects in the top similarity rank are drawing related tools. This is exactly the kind of gradient and complex similarity judgements we were hoping to capture with the Odd One Out. On the other hand,

the general impression of the results page is still a bit random. I will discuss this further in the discussion below.

Figure 2. Similarity report for the image of a pencil.

To further evaluate the results of the game poses an interesting challenge. Since the subsequent aim of the application is to collect human judgements, we are in fact proceeding to create a new human rated similarity benchmark. But since the application is still in an experimental face we need to ensure that the results are eligible. Even though it is highly questionable to simply treat images of things as equal to word senses, we tested to compare the collected odd one out similarity scores with the corresponding scores using WordNet wup-similarity. An example of the two resulting similarity rank lists for *pry bar* can be seen below in table 1.

PRY BAR Similarity rank – Odd One Out	PRY BAR Similarity rank – WN-similarity
1. saw -- 0.55	1. pliers -- 0.86
2. knife -- 0.40	2. scissors -- 0.86
3. clamps -- 0.39	3. eraser -- 0.78
4. protractor -- 0.37	4. needle -- 0.78
5. hammer -- 0.35	5. brush -- 0.74
...	...
34. hair dryer -- -0.15	34. duct tape -- 0.59
35. syringe -- -0.24	35. ruler -- 0.57
36. brush -- -0.26	36. gear -- 0.55
37. ruler -- -0.28	37. gloves -- 0.53
38. funnel -- -0.30	38. fork -- 0.50

Table 1. Similarity rank lists for WordNet similarity and Odd One Out similarity.

At first glance, the most evident conclusion is that the two lists are completely different. Indeed there is not a single correlation between top five and bottom five tools in these two lists. We proceeded to compute similar top lists for all the tools included in the test set and evaluated the general correlation between the two similarity functions using Spearman's rank correlation coefficient. The mean of all sets is a weak 0.13 coefficient, showing close to no correlation between scoring high in Odd One Out similarity and WordNet wup-similarity. However, from the example in table 1 I think it is safe to say that we may not automatically treat WordNet similarity as a gold standard for human judgements. Of the top five tools similar to *pry bar* in the WordNet list we find *eraser*, *needle* and *brush*, similar neither in visual nor functional comparison to *pry bars*.

6. Discussion and future work

The potential applications of gamified, crowd sourced data collection of human intuition is vast and very desirable. The key feature of such an application must be its attractiveness for users, in short – it must be fun to play. On the other hand, we must find a way to make sure that the accuracy of the data is reliable. Both these issues should be addressed in future development of Odd One Out. In this short and experimental case study we have shown that simply asking a different question can be a fruitful way to improve the quality of human similarity judgements based on difference.

Another issue that deserves further attention is to dig deeper into the semantic features hidden behind the unsupervised game. On what intuition do users base their decisions? Analyzing the data further could give a lot of information about different cuts in the set and further expand the notion of semantic relatedness. Since we wish to keep the game simple without asking users for clarifications and at the same time minimize the human effort behind game expansions, one possible way to proceed is to collect ontological data about set individuals via open source resources, such as Wikipedia. A known drawback to unsupervised semantic models is that however functional they are, they do not contain any information about *how* words are related. If we combine the Odd One Out game with ontological data there is promising future work in extracting this lost link.

A third thought concerns the already mentioned discrepancy between visual and word sense similarity. It comes down to what kind of semantic relations we wish to model and how big claims we wish to make about human judgement. It is a question too big to save for a final note, but also one that deserves undivided attention in future work on semantic relatedness.

7. References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238–247).

Bruni, E., N.-K. Tran, and M. Baroni. 2014. *Multimodal distributional semantics.* Journal Artificial Intelligence Research (JAIR) 49, 1–47.

Clark, S., 2015. *Vector Space Models of Lexical Meaning*, in Handbook of Contemporary Semantic Theory, The (eds S. Lappin and C. Fox), John Wiley & Sons, Ltd, Chichester, UK.

Fellbaum, Christian, 2005. *WordNet and wordnets*. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. 2002, *Placing Search in Context: The Concept Revisited*, ACM Transactions on Information Systems, 20(1):116-131

Pedersen, Patwardhan and Michelizzi, 2004. *WordNet::Similarity – Measuring the Relatedness of Concepts*. Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), pp. 1024-1025, July 25-29, 2004, San Jose, CA (Intelligent Systems Demonstration)