Department of Computer Science and Engineering
Indian Institute of Technology, Madras
CS6370: Natural Language Processing

Improving Wikipedia hyperlink structure using Wikispeedia

Team members:
Pranav Murali
(CS18M041)
Vedansh Gurunathan
(CS18M058)

2019

1 Problem Statement

Computing the semantic distance between realworld concepts is crucial for many intelligent applications. We implement a method that leverages data from 'Wikispeedia', an online game played on Wikipedia; players have to reach an article from another, unrelated article, only by clicking links in the articles encountered. In order to automatically infer semantic distances between everyday concepts, our method effectively extracts the common sense displayed by humans during play, and is thus more desirable, from a cognitive point of view, than purely corpus-based methods.
We use this human common sense from knowledge extracted from Wikispeedia game to predict the missing links or alternatively filter the irrelevant ones.

2 Methodology

We used a semantic based distance measure based on information theory. It quantifies how many bits are needed to encode a common-sense Wikipedia path between two concepts. The fewer bits are needed, the more strongly the two concepts are related.
Let A, A' and G be random variables representing the current Wikipedia page, the next Wikipedia page and the goal page of a game respectively. For any Wikipedia article a and any Wikipedia goal (or target) article g, one can consider the probability distribution $P(A'|A = a, G = g)$ over $a's$ out links.This distribution is multinomial and specifies, for each article a' that can be reached in one hop from a, the probability that a player continues to a', if he/she is currently on a and is trying to find goal article g. This can be estimated from the observed games using standard Bayesian methods, as the mean of the Dirichlet distribution which is the conjugate prior of $P(A'|A = a, G = g)$. We use P^* to denote the posterior click probability estimated after seeing all the data.

$$P^*(A' = a'|A = a, G = g) = \frac{N(A' = a', A = a, G = g) + \alpha}{N(A = a, G = g) + \alpha L_\alpha} \tag{1}$$

where α is the Dirichlet parameter representing the initial confidence in the uniform prior distribution, L_α is $a's$ out-degree (i.e. the number of articles linked from a), $N(A = a, G = g)$ is the number of times a was encountered on paths for which g was the goal, and $N(A = a', A = a, G = g)$ counts how often the link to a was chosen in this situation.
Prior click probability is given by:-

$$P^0(A' = a'|A = a, G = g) = \frac{1}{L_\alpha} \tag{2}$$

Consider one particular path $p =< a_1, a_2, a_3, ..a_n >$. We can compute a $path - specific$ distance from every article a_i along p to the goal g i.e, every i with $1 \leq i < n$ we get

$$d_p(a_i, g) = \frac{-\sum_{j=1}^{n-1} \log P^*(A^{'} = a_{j+1} | A = a_j, G = g)}{\log PageRank(g)} \tag{3}$$

The numerator is $\log P^*(A^{'} = a_{j+1} | A = a_j, G = g)$ is the information content of the link from a_j to a_{j+1}. So far, we have described distances that are derived from single paths. To get a path-independent distance from a to g, we simply average over all paths running through a and reaching goal g.
It has been observed that players try to reach, as quickly as possible, a general concept, whose article has a lot of outgoing links. From such hubs it is easy to reach many parts of the Wikipedia graph. After this initial 'getting-away' phase, the 'homing in' phase starts: the search narrows down again towards more specific articles that get more and more related to the goal. So basically the information gain first kind of decreases(the getting-away phase) and then increases(homing-in phase).
So we used the information gain to guess where the homing-in phase, and thus the relevant part of a single game path, starts. An article will then be erased from lists which has all the paths to a goal from a source, if it never occurred in the homing-in phase of a game with the given goal. Doing supervised learning, we trained a neural net to predict where the relevant part starts. Using the dataset available from Wikispeedia website where the human raters mark the split position of 500 game paths. So we use it to train our Neural network using cross validation method with 80(Training Data):20(Test Data) split. We use two input features, the number of links between the input article and the article with minimum information gain along the path, and the number of links between the input article and the goal. Once the network is trained, we plan it to split unseen paths as follows. For every article along the path, we feed its two features into the network and will compute the prediction. We predict the relevant part of the path to start with the article for which the net outputs the highest value.

3 Datasets

- For Neural network training we are using the data set available online on Wikispeedia website maintained by West.
- For human paths we are planning to use Wikispeedia Navigation paths, available on SNAP by Stanford university.
- For evaluation we need related concepts from user. For that we are planning to use Google forms and circulate it within the class.

4 Observations

There are three main observations in the process of finding Wiki similarity which helps us in giving intuition of the results achieved and expressed in the following section.

- **Pagerank**
  Pagerank is used in the denominator of the distance calculation. We implemented the PageRank algorithm and ran it locally on the Wikipedia graph to get these numbers. One can think of PageRank(g) as the prior probability of being in article g, and of the entire denominator as g's information content, or the number of bits needed to code article g independently of any game. This serves the purpose of normalization: intuitively, a concept that is hard to reach (hard to 'explain') is allowed to be related to concepts that are farther from it on Wikipedia paths. For instance, UNITED STATES has PageRank 0.010 (1% a random walk will be spent on the UNITED STATES article), while TURQUOISE has a PageRank of $5.6 * 10^{-5}$. Since $-\log(PG(UnitedStates)) = 5.88$ and $-\log(PG(Turquoise)) = 13.62$ a path from an article a to goal TURQUOISE may take twice as many bits to code as a path from some article b to goal UNITED STATES, and still we will have d(a,TURQUOISE) approximately equal to d(b,UNITED STATES)
- **Asymmetric distance**
  Distances between two word or concepts need not be equal or symmetric. If we consider two words Noam chomsky and Linguistics, then intuitively, the distances between them need not be symmetric. We observed that d(Linguistics,Noam chomsky)=0.03 and d(Noam Chomsky,Linguistics)=0.02.

5 Results

Following are the actual and predicted split points of some paths evaluated on test set where 'AS' stands for Actual split point and 'PS' stands for Predicted split point:-

Path	**AS**	**PS**
$14^{th} century \rightarrow$ time $\rightarrow$ light $\rightarrow$ color $\rightarrow$ rainbow	3	3
acceleration $\rightarrow$ albert eintein $\rightarrow$ germany $\rightarrow$ dresden	2	2
achilles tendon $\rightarrow$ achilles $\rightarrow$ black sea $\rightarrow$ sea $\rightarrow$ ocean	3	3
$14^{th} century \rightarrow$ europe $\rightarrow$ africa $\rightarrow$ atlantic slavetrade $\rightarrow$ africa slavetrade	3	3
acceleration $\rightarrow$ sea $\rightarrow$ mars $\rightarrow$ water $\rightarrow$ hyrdogen	3	3
achilles tendon $\rightarrow$ achilles $\rightarrow$ danube $\rightarrow$ north sea $\rightarrow$ atlantic ocean $\rightarrow$ ocean	3	4

Table 1: Actual and Predicted Split points.

On an average, the difference between predicted split point and actual split point is around **0.98**

6 Empirical Evaluation of the Distance measure

In order to test the quality and psychological validity of our distance measure, we compare it to Latent Semantic Analysis(LSA) method, respectively. For LSA, we used the same corpus as Huettig et al., 2006: 'General Reading up to 1st year college' (300 factors). Since the above is a web interface Landauer and Kintsch, 1998 we used an automated program that would fetch us results. We chose top 5 words fro our method(WikiSimilarity) and top 5 words from LSA method and interpreted which words were closer.
For eg:

- For the word juice Wiki Similar suggestions were lemon, citrus, scurvy, coconut, florida and for LSA we got gastric, lennin, semiliquid, stomach, lumen. Clearly in this case WikiSimilar words are better
- For the word ocean Wiki Similar suggestions were atlantic ocean, pacific ocean, water france earth and for LSA we got shallowest, abyssal, glomar, rica, bolivia, orinoco, bolivar. Again in this case WikiSimilar suggestions are really close to the actual human meaning.
- But in some cases like economics Wiki similar suggestions were china, mathematics, technology, civilization, science and LSA words are macroeconomics, microeconomics, economists, scarcity. Here it seems the LSA suggestions are better but the important thing to note is it was better because of the words have a huge overlap with the key word in vector representation. The 'economy' root is almost common in all words
- For the word set Wiki similar suggestions were algebra, boolean logic, arabic language, china, mathematics and LSA gave suggestions like aside, foundry, undercontrolled, overcontrolled, clocks. Here too clearly LSA made weaker suggestions than Wiki similarity

We had observed nearly 77 words and in majority of them our Wiki similar suggestions were better than the standard LSA comparison. Out of total 77 word similarities we calculated the percentage of the similarities given by both the measures.

Method	**Votes**	**Percentage**
WikiSpeedia	46	0.61
LSA	27	0.35
Both	4	0.04

Table 2: Results of the comparison to LSA.

7 References

[1] Robert West, Joelle Pineau, and Doina Precup. Wikispeedia : An Online Game for Inferring Semantic Distances between Concepts. In 21st International Joint Conference on Artificial Intelligence (IJCAI'09), pp. 1598–1603, Pasadena, Calif., 2009.

[2] Wikispeedia dataset references : https://www.cs.mcgill.ca/ rwest/wikispeedia/data/

[3] Snap dataset references : http://snap.stanford.edu/data/wikispeedia.html

[4] F. Huettig, P. T. Quinlan, S. A. McDonald, and G. T. M. Altmann. Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. Acta Psychologica, 121(1), 2006.

[5] T. Landauer and W. Kintsch. LSA. Website, 1998. http://lsa.colorado.edu.