Designing an Electronic Reverse Dictionary Based on Two Word Association Norms of English Language

Jorge Reyes-Magaña, Gemma Bel-Enguix, Gerardo Sierra, Helena Gómez-Adorno
jorge.reyes@correo.uady.mx

ELEX 2019: SMART LEXICOGRAPHY. October 1 to 3, 2019, Sintra, Portugal.
Content

• Introduction.
• Word Association Words.
• Graph.
  • Algorithms. Betweenness Centrality.
  • Algorithms. Page Rank.
• Search Model.
• Evaluation corpus.
• Experiments.
• Results
• Evaluation
• Conclusion
Two types of dictionaries

- Semasiological. Provides meanings, ie. given a word, the user obtains the meaning of such word.
- Onomasiological. Works in the opposite way, given the description of a word, the user obtains the related concept (Baldinger, 1970)
Introduction

• Onomasiological search nowadays
Introduction

• This work perform a lexical search over a knowledge graph in a similar way onomasiological dictionaries.

A small rodent living in trees with a long bushy tail

Small rodent live tree bushy tail

Squirrel
Word Association Words

- Free word associations (WA) are commonly collected by presenting a stimulus word (SW) to the participant and asking him to produce in a verbal or written form the first word that comes to his mind. The answer generated by the participant is called response word (RW).

- Compilations of WA are called Word Association Norms
Word Association Norms

- Edinburgh Associative Thesaurus (EAT) (Kiss et al., 1973a)
  - 8,211 stimulus words, and 20,445 different words including both, stimuli and responses.

- Collection of the University of South Florida (USF) (Nelson et al., 1998)
  - 6,000 participants that produced nearly three-quarters of a million responses to 5,019 stimulus words.
Graph

- The graph representing the WAN has been elaborated with lemmatized lexical items.
- The graph is undirected, so that every stimulus is connected to every associated word without any precedence order.
- For the weight of the edges there are two different functions:
  - Frequency. Counts the number of occurrences of every associated to its stimulus in the whole dataset.
  - Association Strength. Establishes a relation between the frequency (F) and the number of associations for every stimulus.
- For the system to work in the shortest paths, we need to calculate the IF and the IAS.
Algorithms. Betweenness centrality

• Given a definition, we search in the graph the word that better match with it.

• For this purpose, we used a variation of the betweenness centrality (BT) algorithm (Freeman, 1977)

• The traditional betweenness algorithm assumes that important nodes connect other nodes.
Algorithms. Betweenness centrality

• For a given node (v) in a graph (G), the BT is calculated as the relation between the number of shortest paths between nodes i and j that pass through node v and the number of shortest paths between nodes i and j.

\[ C_{btw}(v) = \sum_{i,j \in N} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}} \]

• N = the total number of nodes in the graph.
Algorithms. Betweenness centrality

All the edges have a weight of 1

\[ C_{BT}(B) = \frac{\sigma_{A,B}(B)}{\sigma_{A,B}} + \frac{\sigma_{A,C}(B)}{\sigma_{A,C}} + \frac{\sigma_{A,D}(B)}{\sigma_{A,D}} + \frac{\sigma_{B,C}(B)}{\sigma_{B,C}} + \frac{\sigma_{B,D}(B)}{\sigma_{B,D}} + \frac{\sigma_{C,D}(B)}{\sigma_{C,D}} \]

\[ = \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{0}{1} \]

\[ = 5 \]
Our hypothesis is that, if we use a subset, the nodes of the WAN graph (WG) that represent the words of a definition as initial and final nodes in the BT algorithm, and calculate the centrality of the other nodes in WN taking these nodes as pairs, then the more central nodes will be the concept of such definition.
• PageRank computes a ranking nodes in a graph $G$ based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages. It was developed by Page et al. (1999).
In our case, the pages described above are the words in the WAN datasets, the web page links correspond to all the relations given by the stimuli-response between words.

The hypothesis driven here is that the target word tested with a definition to be searched corresponds to the higher scores returned by the PageRank algorithm.
Search Model

Algorithm 1: Reverse dictionary

Data: WAN datasets, definitions to search
Result: list of ranked concepts

pre-process(WAN datasets);
pre-process(definitions to search);
GraphWAN = build-graph(WAN datasets);
GraphWAN = prune-graph(GraphWAN);
for each definition do
  definition = remove-StopWords(definition);
  definition = filter-WordsInWAN(definition);
  build_subgraph(definition);
  ranking_nodes_BT = BT(GraphWAN, subgraph);
  ranking_nodes_PR = PR(GraphWAN);
  ascending_order(ranking_nodes_BT);
  ascending_order(ranking_nodes_PR);
Evaluation corpus

• We used an evaluation corpus consisting of 7 concepts.
  • 10 definitions were provided.
  • Human native speakers. In most cases, the definitions are very different from the ones found in dictionaries; they lack specialized terms and include cultural references and connotations.
  • Selected words: water, squirrel, bench, hurricane, lemon, bucket and clothes.
Evaluation corpus

<table>
<thead>
<tr>
<th>Definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s a little rodent and can be red or grey, it has a big bushy tail</td>
<td>A small rodent living in trees with a long bushy tail</td>
</tr>
<tr>
<td>A small rodent living in trees with a long bushy tail</td>
<td>A small rodent which lives in trees, collects nuts and has a bushy tail</td>
</tr>
<tr>
<td>A small rodent which lives in trees, collects nuts and has a bushy tail</td>
<td>Animal, grey/red, bushy tail, lives in trees, buries nuts</td>
</tr>
<tr>
<td>Small animal, lives in trees, eats acorns, has a bushy tail</td>
<td>Animal, bushy tail, eats nuts, builds nests in trees called dreys</td>
</tr>
<tr>
<td>Small funny animal with big, bushy tail, likes nuts, likes trees</td>
<td>Small grey mammal, relative to the rodent, found in both countryside and town</td>
</tr>
<tr>
<td>Animal that lives in trees and collects acorns, has a long tail</td>
<td>A small-sized animal, habitat in trees</td>
</tr>
</tbody>
</table>

Definitions of squirrel given by the students.
Experiments

• For the evaluation of the inference process, we used the technique of precision at k $p(k)$ (Manning et al., 2009).

• $P(1)$ stands that the concept associated to a definition given was ranked correctly in the first place, in $p(3)$ the concept was in the first three results, and the same applies to $p(5)$. 
# Results

## Results in terms of precision of our model with EAT dataset

<table>
<thead>
<tr>
<th>Weighting function</th>
<th>Graph Algorithm</th>
<th>p@1</th>
<th>p@3</th>
<th>p@5</th>
<th>p@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverse Frequency (IF)</td>
<td>Betweenness Centrality (BT)</td>
<td><strong>0.152</strong></td>
<td>0.186</td>
<td>0.220</td>
<td>0.237</td>
</tr>
<tr>
<td>Inverse Association Strength (IAS)</td>
<td>Betweenness Centrality (BT)</td>
<td><strong>0.152</strong></td>
<td><strong>0.220</strong></td>
<td><strong>0.237</strong></td>
<td><strong>0.254</strong></td>
</tr>
<tr>
<td>Inverse Frequency (IF)</td>
<td>PageRank (PR)</td>
<td>0.000</td>
<td>0.074</td>
<td>0.129</td>
<td>0.129</td>
</tr>
<tr>
<td>Inverse Association Strength (IAS)</td>
<td>PageRank (PR)</td>
<td>0.000</td>
<td>0.0740</td>
<td>0.129</td>
<td>0.129</td>
</tr>
</tbody>
</table>

## Results in terms of precision of our model with USF dataset

<table>
<thead>
<tr>
<th>Weighting function</th>
<th>Graph Algorithm</th>
<th>p@1</th>
<th>p@3</th>
<th>p@5</th>
<th>p@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverse Frequency (IF)</td>
<td>Betweenness Centrality (BT)</td>
<td>0.236</td>
<td>0.309</td>
<td><strong>0.418</strong></td>
<td>0.436</td>
</tr>
<tr>
<td>Inverse Association Strength (IAS)</td>
<td>Betweenness Centrality (BT)</td>
<td><strong>0.290</strong></td>
<td><strong>0.363</strong></td>
<td><strong>0.418</strong></td>
<td><strong>0.5272</strong></td>
</tr>
<tr>
<td>Inverse Frequency (IF)</td>
<td>PageRank (PR)</td>
<td>0.037</td>
<td>0.074</td>
<td>0.129</td>
<td>0.222</td>
</tr>
<tr>
<td>Inverse Association Strength (IAS)</td>
<td>PageRank (PR)</td>
<td>0.037</td>
<td>0.074</td>
<td>0.148</td>
<td>0.222</td>
</tr>
</tbody>
</table>
Evaluation

• Comparison to other IR models
  • OneLook Thesaurus. Allows to describe a concept and returns a list of words and phrases related to that concept.
  • Okapi BM25. Based on probabilistic models with a bag of words implementation (Robertson & Zaragoza, 2009).
Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>P@1</th>
<th>P@3</th>
<th>P@5</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>OneLook</td>
<td>0.202</td>
<td>0.347</td>
<td>0.376</td>
<td>0.434</td>
</tr>
<tr>
<td>Reverse Dictionary with USF (IAS)</td>
<td><strong>0.290</strong></td>
<td>0.363</td>
<td>0.418</td>
<td><strong>0.5272</strong></td>
</tr>
<tr>
<td>BM25 with EAT</td>
<td>0.257</td>
<td>0.357</td>
<td>0.414</td>
<td>0.471</td>
</tr>
<tr>
<td>BM25 with USF</td>
<td>0.257</td>
<td><strong>0.400</strong></td>
<td><strong>0.457</strong></td>
<td>0.514</td>
</tr>
</tbody>
</table>

- The BM25 algorithm showed better performance than the Onelook reverse dictionary when the search is performed over the WAN datasets.
- The higher results are consistent with the ones seen in the reverse dictionary, USF norms show the best performance.
Conclusions

• This paper introduces a model for onomasiological searches that has some novelties, among them the simplicity, the use of graph-based techniques.

• We observed that the graph built with all the nodes and edges contained in the datasets tends to be not so good due to the number of paths that outcome on wrong results.
Conclusions

• We have shown how descriptions of concepts that are made by common people with nonscientific specifications can retrieve accurate results using our method.

• The success of the system with non-scientific input can drive new lines of applied research, and the implementation of different assistant writing systems especially oriented to people with a range of aphasias, like dysnomia and Alzheimer’s disease.
References

Thank you!