Linguistic Graph Similarity for News Sentence Searching

Kim Schouten and Flavius Frasincar
{schouten, frasincar}@ese.eur.nl

Econometric Institute, Erasmus University Rotterdam
PO Box 1738, NL-3000 DR, Rotterdam, the Netherlands

With the amount of available news being too much to handle for any individual, efficient ways of searching through news items that are mostly unstructured text are needed. However, most text search algorithms are based on the bag-of-words approach and do not take more advanced linguistic features into account. In this abstract we propose a system that effectively leverages the grammatical relations found within sentences to get a better search performance. An extended version can be found in [1].

The main idea is to use a parser to extract the dependency graph, consisting of the words in a sentence as nodes, and the grammatical relations between those words as labeled, directed edges. Every news sentence in our database is processed and stored in this graph format. When a user searches for a sentence in a news corpus, that sentence query is also processed into a dependency graph. Instead of comparing the presence or absence of individual words that are in the user query, we can now compare the graph that describes the user query with the graphs in the database for all the news sentences. The news sentences graphs that are most similar to the query graph are then returned as the top ranking results.

This process is essentially a form of approximate sub-graph isomorphism, and we have developed an algorithm that iteratively traverses both graphs in parallel, keeping a score of all matching elements. The algorithm will stop traversing when there are either no more nodes to compare or when the nodes that are available are too dissimilar to be compared. When stopped, the value returned by the scoring function is the sum of all similarity comparisons made between nodes and edges from the query graph and the news sentence graph.

Edges are compared by looking at the edge label that denotes the type of involved grammatical relation. Nodes (i.e., the words) are compared based on five characteristics: stem, lemma, literal word, basic POS category (e.g., noun, verb, adjective, etc.), and detailed POS category (plural noun, proper noun, verb in past tense, etc.). Furthermore, we check for synonyms and hypernyms using the WordNet dictionary and assign scores for those as well. Last, the node scores are adjusted for frequency so finding a rare word yields a higher score than finding a common word.

Since certain grammatical relations, like subject and object, might be more important than others, we assign a weight to each relation type. These weights are optimized using a basic genetic algorithm optimization.

To compare the proposed method with the TF-IDF baseline, we created a small database of 19 news items that together consist of 1019 sentences in total, as well as 10 query sentences. All possible combinations of query sentence and news sentence were annotated by at least three different persons and given a score between 0 (no similarity) and 3 (very similar). The results are compared using the normalized Discounted Cumulative Gain (nDCG) over the first 30 results, Spearman's Rho, and Mean Average Precision (MAP) and are shown in Table 1.

Table 1: Evaluation results

	TF-IDF baseline	our method	improv.	t-test p-value
nDCG	0.238	0.253	11.2%	<0.001
MAP	0.376	0.424	12.8%	<0.001
Sp. Rho	0.215	0.282	31.6%	<0.001

Our implementation of the proposed method shows the feasibility of searching news sentences in a linguistic fashion, as opposed to using a simple bag-of-words approach. Because the graph-representation preserves much of the original semantic relatedness between words, the search engine is able to utilize this information, resulting in a higher performance for all three considered metrics.

References

[1] K. Schouten and F. Frasincar. Web News Sentence Searching Using Linguistic Graph Similarity. In *Proceedings of the 12th International Baltic Conference on Databases and Information Systems (DB&IS 2016)*, pages 319–333. Springer, 2016.