WHAT IS THE RELATIONSHIP ABOUT?
Extracting Information about Relationships from Wikipedia

Brigitte Mathiak[1], Víctor Manuel Martínez Peña[2] and Andias Wira-Alam[1]
[1]*GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Köln, Germany*
[2]*Institute for Web Science and Technologies, University of Koblenz-Landau, Koblenz-Landau, Germany*

Keywords: Relationship Extraction, Wikipedia.

Abstract: What is the relationship between terms? Document analysis tells us that ”Crime” is close to ”Victim” and not so close to ”Banana”. While for common terms like Sun and Light the nature of the relationship is clear, the measure becomes more fuzzy when dealing with more uncommonly used terms and concepts and partial information. Semantic relatedness is typically calculated from an encyclopedia like Wikipedia, but Wikipedia contains a lot of information that is not common knowledge. So, when a computer calculates that Belarus and Ukraine are closely related, what does it mean to me as a human? In this paper, we take a look at perceived relationship and qualify it in a human-readable way. The result is a search engine, designed to take two terms and explain how they relate to each other. We evaluate this through a user study which gauges how useful this extra information is to humans when making a judgment about relationships.

1 INTRODUCTION

The purpose of semantic relatedness measures is to allow computers to reason about written text. They have many applications in natural language processing and artificial intelligence (Budanitsky, 1999), and have consequently received a lot of attention from the research community.

However, the pure measure of relatedness in numbers is not very helpful to normal users. These people are not so much interested in the quantity of relatedness, but the quality. We use a modified standard method to measure relatedness between Wikipedia entries based on a combination of link analysis and text analysis, which evaluate comparably to other similar measures. We leverage information used by these methods to find text snippets on Wikipedia, which are significant for the relationship and describe it in a human-readable way.

The goal of these snippets is to inform the user of the quality of the relationship, especially in the case of an information gap. For evaluation, we have made a user study using Mechanical Turk. We have first asked the participants what they know about the relation between two concepts, such as Barack Obama and Chicago and then present them with a number of snippets extracted with our method. The general feedback was very positive, with most participants finding most of the snippets helpful. The learning effect was also quite visible, while in one group 58 % of the participants felt there was a relationship between both, only 22 % could specify that relationship precisely, while the others were either very general (”both in America”) or plainly wrong (”He was the governor the State of Illinois”). Yet, they quickly accepted the connections made by the application as meaningful.

2 RELATED WORK

Current research explores two fundamentally different ways to compute semantic relatedness between two terms. The first is *link-based*. In a hierarchical structure, usually a taxonomy, this typically applies to the shortest path between the two concepts. This is often modified with other parameters, such as the depth of the term in the taxonomy, weights derived from the semantics of the taxonomy and so on (Leacock et al., 1998; Slimani et al., 2006). This can also be applied to Wikipedia, through the use of categories, like is done with WikiRelate (Strube and Ponzetto, 2006). More accuracy is gained by exploiting the link structure between the articles, such as in (Islam and Inkpen, 2008), simply because there are many more links than categories per page. Beyond the very simple distance of counting the shortest path, in (Milne

Mathiak B., Manuel Martínez Peña V. and Wira-Alam A..
WHAT IS THE RELATIONSHIP ABOUT? - Extracting Information about Relationships from Wikipedia.
DOI: 10.5220/0003936506250632
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 625-632
ISBN: 978-989-8565-08-2

and Witten, 2008) the anchor texts of links and link structure itself is used to find. They use link counts weighted by the probability of the link occurring on the page (inspired by tf-idf) as a vector representation of the article while calculating the cosine similarity on the vectors for the similarity measure. This may look very similar to our approach, but we use tf-idf on the terms not the links as well as a directly computable measure for the link structure, so we can calculate our measure online with only two requests to the Wikipedia API. Thus, we combine a link-based measure with the second category of text based measures.

Text based measures take an example corpus of documents that are known to relate to the two terms and then calculate the semantic distance between the two document sets, thereby splitting the problem of relatedness between terms into two problems: choosing a suitable data set and calculating the semantic distance between the documents. There are large numbers of semantic distances to choose from: Lee distance, tf-idf cosine similarity (Ramos, 2003), Jaro-Winkler distance, and Approximate string matching (Navarro, 2001), just to name a few. In (Islam and Inkpen, 2008), the Semantic Text Similarity (STS) has been developed as a variety of the Longest Common Subsequence (LCS) algorithm and a combination of other methods. It is optimized on very short texts, such as single sentences and phrases. This method was evaluated by using definitions from a dictionary.

The *Explicit Semantic Analysis* ESA (Gabrilovich and Markovitch, 2007) uses Wikipedia, just like our approach, and calculates a complete matrix of term to concept relatedness, which can be further refined by introducing human judgments. Unlike our approach, however, it requires the processing of the whole of Wikipedia in a non-linear process, which is very expensive and has not been replicated on the scale since.

There are other approaches to mix both link and text analysis, such as (Nakayama et al., 2008) which extracts explicit relationships such as Apple is Fruit, Computer is Machine, Colorado is U.S. state. The goal of this paper, however, is not to use Wikipedia to find relationships which conform to established standards and semantics, but quite the opposite, to produce explanatory text suited for unusual relationships.

3 RELATIONSHIP EXTRACTION

3.1 Architecture

The RelationWik Extractor was built as a web information system. From a user's point of view, it's function it quite simple. The articles for which a relationship is sought and a few parameters are input over a web site and the system will show the results as both a score and snippets illustrating the connection from both sides.

The Wikipedia articles are then downloaded directly via the Wikipedia API. The text is then scanned for additional information such as links, templates, etc. and stripped of its Wikipedia syntax. Both text and meta-information is stored in a database cache. The results of the algorithms are visualized with PHP and the Google Chart API.

3.2 Calculating Relatedness

For the actual calculation of the relationship, two differently approaches are used. One is based on the link structure and the other on the textual closeness of the texts. A third approach is a mixture of both.

The first algorithm measures the connectedness of the terms, by studying inlinks and outlinks. When looking at connections that go over several hops, it becomes clear that the connection can be quite thin. For example, Banana and Berlin are connected by an enzyme that occurs in the Banana and was identified in Berlin. This gets worse when looking at connections with even more connectors in between. Therefore, we decided to ignore all connections involving more than one intermediary. The connections with one intermediary that made the most sense occurred in the scenario where both articles link to the same article. This occurs, for example, when both of the given articles A and B are connected to a category or another larger super-concept by linking to it. Also, in terms of computation time, it is the fastest possible link analysis since outlinks are the easiest to extract from.

Following that argumentation, we only look at articles that either have intersecting outcoming links or have a link from A to B or from B to A. Any other pairs are given a relatedness of zero. Connected articles A and B receive a base relatedness b of 0.5 and are given a boost of 0.1 for additional connections (e.g. 0.7 when A and B link to each other and link to at least one third article). This base value is further modified by the number of backlinks of the linked-to article in relation to the links from the other article.

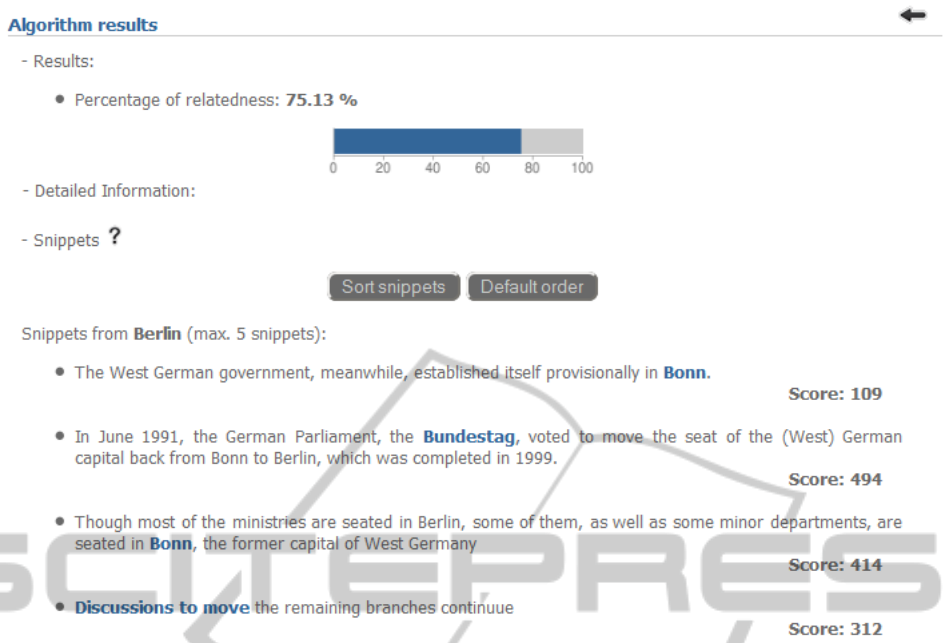

Figure 1: The result page of the Relationship Extractor with the terms Bonn and Berlin. The algorithm is set to sentence-sized snippets. The Score given next to the snippets is a relevance measure based on the terms used in both documents.

And, if applicable, by the ratio between common outlinks c and total outlinks $l_{A\rightarrow}$ respective of $l_{B\rightarrow}$.

$$Rel_{AB} = b - \frac{l_{A\rightarrow B} + l_{B\rightarrow A}}{l_{total}} + \frac{c}{2l_{A\rightarrow}} + \frac{c}{2l_{B\rightarrow}} \qquad (1)$$

It was originally planned to optimize the choice of b and introduce weighting factors, but the initial choices performed quite well in the evaluation and so no further optimization was necessary and might have introduced overfitting.

For the second algorithm, we use a standard cosine similarity between the articles. The articles are preprocessed by stripping wikisyntax, punctuation and symbols, removing stop words and unique terms and using only basic stemming by removing plurals. The term vectors A and B are calculated with tf idf.

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \qquad (2)$$

As our third method, we average the results of the two similarities above. Again weighing was considered, but since our goal is not to optimize the relationship score, but to present human-readable snippets, the exact optimal ratio, was of no consequence to us.

4 EVALUATION

For the evaluation of the relatedness score, we use the WordSimilarity-353 Test Collection (Finkelstein et al., 2002). It contains 353 English word pairs with human-assigned similarity judgments. It contains antonyms, synonyms and similar words and connected terms, such as Film and Popcorn. Not all terms from the dataset can be used directly, e.g. keyboard was mapped to the Wikipedia article Computer keyboard, Plane to Fixed-wing aircraft, etc. The mappings were constructed by using the Wikipedia search engine and choosing the first entry. A few terms had to be removed, such as Diego Maradona, because of errors on the page. The similarity scores from all three algorithms were tested on the data set by using a Pearson linear correlation coefficient (Rodgers and Nicewander, 1988). The results are shown in Table 1.

On closer inspection, we can observe that the cosine similarity tends to judge too low on somewhat similar articles, such as Radio, Computer or Internet. The links similarity on the other hand is vulnerable to over judging relatedness due to singular rogue links and has problems in general with articles containing only few links. On average, both effects seem to dampen each other.

The combined score is competitive with other methods such as described in (Strube and Ponzetto, 2006), (Gabrilovich and Markovitch, 2007) and

Table 1: Pearson-correlation.

Algorithm	r-Pearson coefficient
links	0.65
cosine	0.55
combined	0.69
(Strube and Ponzetto, 2006)	0.49
(Gabrilovich and Markovitch, 2007)	0.75
(Milne and Witten, 2008)	0.69

(Milne and Witten, 2008), but it is very fast, without needing any pre-processing and is able to work online. The wait time mostly depends on the speed of downloading both articles. All three other methods work on a Wikipedia dump, which is more or less extensively pre-processed. With the pre-processed dataset however, they achieve much faster response times. We rarely go beyond 10 seconds for any given pairs of terms, though this depends on the current traffic on the Wikipedia server. A caching mechanism has been implemented to alleviate the effect, lowering response times to much lower numbers.

We have not addressed some of the serious questions in the field, such as how to match search terms from the evaluation set to the Wikipedia articles. Since we expect user interaction, the disambiguation can be done with Wikipedia-specific means. Alternatively, methods, as outlined in the above-mentioned publications can be employed.

The terms, we are most interested in, are not generally found in evaluation data sets. And with good reason: they are terms, which have hidden or not commonly known relationships. Human judges would give varying degrees of relationship, depending on whether they happen to know the details about the relationship or not. Terms, such as Belarus and Ukraine[1] are blindly judged as completely unrelated by 28.8 % of the participants, yet our algorithms judge the relationship as high. We believe the algorithm is right and that the information gap is something to be closed.

5 SNIPPETS

We assume that a common user is not so much interested in how much two terms are related, but rather how they are related. As we have shown above, the connections that we find are positively correlated to the human-perceived relation between two terms. This offers us a connection point between the terms. For the linking algorithm it is rather simple. The links themselves serve as a direct connection. For intersecting outlinks the links to the intersecting articles are a connection point. The cosine algorithm also gives us a measure of which terms are most highly relevant for the similarity and we can use those a connector. Those connection points are then transformed into snippets and shown to the user. All three can be mapped to one or more specific text positions inside of Wikipedia articles. The corresponding snippet is generated from there, choosing either paragraph, sentence or a fixed-size window. Since there can be more than one link on a page and terms can be mentioned several times, we receive a large number of snippets. The remaining questions are: What is the optimal window? What is best way to rank? And most importantly, is the method beneficial in the first place?

5.1 Methodology of the User Study

We offered .25 US$ at Amazon Mechanical Turk to 40 participants to answer a short survey on our snippets. We chose 10 pairs of terms (ref. 2). The last 4 term pairs are taken from the WordSimilarity-353 test set as a control group. The first 6 were chosen based on a lesser known connection. We were striving to take into account a variety of connections, such as biographical events, historical similarity, recent news, spatial closeness, same super-category and part-of. Also, we were trying to find term pairs in which at least one term should be known by the participants, and preferably both.

For the study, we first asked the participants to rate the relationship between the terms between 0 and 10, without using any secondary information sources, such as the Internet. We then presented up to five snippets for each pair and asked for an evaluation of each individual snippet on a scale ranging from not good to very good (description of the relationship). Then we asked again for a rating of the relationship between the two terms.

As a control mechanism, we asked the participants to give us a catchword description of the relationship from their point of view, in order to understand why they would rate in a certain way. This was checked both on the initial rating and after the snippets had been given.

Final methodological note: As one of the participants pointed out to us (a self-proclaimed MS in Resource Economics), we did not offer an ”opt-out” button on our rating scales. This introduces bias towards

[1]Two neighboring countries that used to be part of the USSR.

Table 2: Term pairs used for the user study.

Term 1	Term 2	connection
Barack Obama	Chicago	Where he went to law school
Bonn	Berlin	Both were capital of Germany at some point
Google	Apple	Recent law suit concerning Motorola
Belarus	Ukraine	Neighboring countries; ex-USSR
German language	English language	Both indogermanic languages
Dave Mustaine	Metallica	Founding member; guitarist
Radio	Television	
Cat	Tiger	
Sex	Love	
Student	Professor	

stating an opinion even if there is no informational basis for this opinion. The participants have to answer the question to gain the monetary incentive, even if they do not in fact know anything about the subject matter, thus (in economical theory) they are prone to answer randomly[2]. We have pondered this issue, but since our aim is to measure how much they know in the first place, giving them an easy way out seemed like losing too much information and thus introducing a bias against the uninformed. We assume they will not answer randomly, but choose something on the low end of the scale. The data seems to corroborate our assumption. There is a strong gap between the relationship rating in the first 6 pairs between before the snippets and after, although with a large spread, which could be a result of the random choices.

5.2 Experiences from the User Study

There were no technical difficulties with the Amazon Mechanical Turk platform. However, setting up a suitable survey was a bit tricky for non-psychologists (see above for some pitfalls). We were forced to change the procedures quite a bit, before finding a method to adequately measure what we were interested in. Still, some participants simply did not play by the rules, e.g. one participant wrote in the general comments "(...)we can complete this survey easily using search engines like Google(...)", although we stated twice and in large letters that the Internet was not to be used. Overall the general comments were very helpful in designing better versions of the survey.

There were some complaints concerning for example money or a lack of understanding about the purpose of the survey. However, we decided it was not wise to explain what we were looking for in order to avoid the interviewer-compliance bias as much as possible. Some treated it as a game and wondered whether they had won.

Quite a lot (37%) of the participants did not complete the survey, for unknown reasons. We did not raise the incentive to test for lack of incentive. Some of the drop-outs can probably be explained by participants being annoyed with the forced-choice answers. Many of the participants that eventually dropped-out gave mocking answers to the open-ended questions.

We did award the incentive to everyone who answered more than a few questions and claimed to be finished and used answers from unfinished questionnaires for the analysis.

5.3 Learning Effect

When looking at the median or mean differences in the relationship rating the difference seems slight, comparable to the variance in the control group (cf. table 3). Now, when looking at the ratio of zero relationship votes (cf. table 4), we can see a pronounced change between before and after. What is surprising, though, is that the values also drop significantly for the control group.

One part of this effect is that the snippets show new information between well known concepts such as Cat and Tiger, two concepts that both belong to the same animal class. A look at the catchwords that were provided by the participants to explain their relationship rating confirms this view. New information from the snippets was incorporated there allowing the participants to waver from their belief that there was no connection at all. However, while some new informa-

[2]The literature on this is in fact extensive and not as clear-cut as that. While the general opinion, such as (Dillman and Bowker, 2001) seems to be that it is better to avoid forced-choice answer sets as it puts extra strain on the participants, they are common practices for special purposes like memory tasks (Martin et al., 1993). In (Smyth et al., 2006), it has been shown that forced-choice increases both the time spent on answering the question and the quality of the data in a web scenario similar to ours.

Table 3: Ratings before the snippets were shown and after. Numbers are (in order): median, simple mean.

Term 1	Term 2	before	after
Barack Obama	Chicago	7, 7.0	8, 7.2
Bonn	Berlin	5, 5.4	7, 6.7
Google	Apple	6, 5.3	6, 5.2
Belarus	Ukraine	6, 5.9	7, 6.1
German language	English language	5, 5.0	6, 5.7
Dave Mustaine	Metallica	5, 5.5	7, 6.7
Radio	Television	6, 6.2	7, 6.8
Cat	Tiger	7, 6.7	8, 7.2
Sex	Love	7, 6.7	7, 6.5
Student	Professor	8, 7.1	7, 6.9

Table 4: Ratio of relationships rated as zero.

Term 1	Term 2	before	after
Barack Obama	Chicago	18.2	3
Bonn	Berlin	28.8	1.5
Google	Apple	18.2	1.5
Belarus	Ukraine	28.8	0.0
German language	English language	27.3	0.0
Dave Mustaine	Metallica	36.4	7,8
Radio	Television	13.6	0.0
Cat	Tiger	9.1	0.0
Sex	Love	13.6	1.5
Student	Professor	13.6	1.5

tion led to upgrades in the relationship rating, it often led to downgrades as well.

One reason for this was misinformation. A number of participants wrongly stated that Barack Obama used to be the governor/senator of Illinois. After they read the snippets, they revised this opinion and accordingly downgraded the relationship level. On the other hand, most of the participants rating the relationship between the President and Chicago as zero, gave only general catchwords, such as ”America” and later upgraded their rating when they learned more. For other term pairs (Apple, Google), they digested the new information, but did not see any reason to adjust their rating. For some pairings, such as Bonn, Berlin and Dave Mustaine, Metallica, there was quite a shift in the mean rating, but mostly, it seemed, to account for the fact that many did not know of Bonn or Dave Mustaine beforehand.

What is interesting, though, is that the knowledge of the participants concerning the numerical relationship rating was somewhat stable, regardless of additional information. This is a good sign that relationship ratings tend to become clearer with more information; the variance gets lower, and especially the extreme statement of ”not related” gets rarer. This trend for humans to rate more gradual with more information is especially visible when looking at which rating category received the most ”votes” (cf. table 5). For the pairs with hidden information this jumped from 0 to the mean, while for the control group it remained stable.

5.4 Sentences vs. Paragraphs

Apart from the general learning effect gained from the snippets, we also investigated which type of snippets (paragraph or sentence) were favored and why. For each set of snippets we asked the participants, which they found most useful. They were split into two control groups, each alternating sentence and paragraph. Overall, 40 participants chose 112 sentences and 89 paragraphs as the best description, almost an equal number. The slight bias does not allow a conclusive choice. We therefore decided to integrate a user choice into the web interface.

Still, distribution was not equal. For ”Barack Obama” and ”Chicago”, the top choices were a paragraph with 25% and a sentence with 17.5% of the votes. However, both were from the same connection point, telling the story about the career of Barack Obama with some timeline. For ”Dave Mustaine” and ”Metallica”, the participants again chose the same

Table 5: Rating that most participants agreed upon.

Term 1	Term 2	before	after
Barack Obama	Chicago	0	10
Bonn	Berlin	0	8
Google	Apple	0	6
Belarus	Ukraine	0	7
German language	English language	0	7
Dave Mustaine	Metallica	0	8 &10
Radio	Television	8	8
Cat	Tiger	7	8
Sex	Love	10	10
Student	Professor	10	7

connection point about the career history, regardless of paragraphs or sentences. The connection itself was the criteria for voting the best snippets. In the paragraph scenario, they just happened to be more interesting than the alternatives. For the other pairs, we had a consistent bias towards sentences.

Curiously, we found that the characteristics of a ”best” snippet is not so much tied to length, but to containing interesting bits of knowledge, regardless of the information content to the relationship. For example, in the snippets for ”German language” ”English language”, the snippet ”English replaced German as the dominant language of science Nobel Prize laureates during the second half of the 20th century.” was chosen as best by the majority. However, the snippet does not so much add to the knowledge of the relationship, as it is simply an interesting piece of trivia. In a similar vain, the sentence ”Television sends the picture as AM and the sound as AM or FM, with the sound carrier a fixed frequency (4.5 MHz in the NTSC system) away from the video carrier.” won the majority vote for ”Radio” and ”Television”.

We conclude that users prefer interesting snippets over relevant ones and that length does not seem to be important. Therefore, the vote of which snippet is best cannot be used as a direct measure to which snippet educated the user best about the relationship. This can only be determined indirectly (e.g. through the analysis of catchwords).

6 CONCLUSIONS AND FUTURE WORK

By showing text snippets around the connectors, the user gets a good overview on the nature of the relationship between any two given terms, especially in those cases in which the relationship is usually not well known (e.g. Berlin and Bonn). Please try yourself on http://multiweb.gesis.org/RelationShipExtractor/.

Apart from the obvious use of educating the user about the relationship between two terms, the snippets are also an adhoc way to produce natural language about the relationship, as can be leveraged by Text Mining systems. In ontologies, such as DBPedia, there are often only a limited number of different properties defined between two concepts, making it difficult to properly understand the semantics of the property, as sometimes only a word is given. There are too many properties defined to encode the semantics manually, yet many relationships go unnoticed, because they do not fit the scheme. The snippets offer a way of solving both problems. The semantics of an existing property can be described by using examples from the database and generating snippets for them. Unknown relationships can be found by using the scores and indirectly defined over the snippets.

REFERENCES

Budanitsky, A. (1999). Lexical semantic relatedness and its application in natural language processing. Technical report, Department of Computer Science, University of Toronto.

Dillman, D. A. and Bowker, D. K. (2001). *The Web Questionnaire Challenge to Survey Methodologists*, pages 159–178. Pabst Science Publishers.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20:116–131.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Islam, A. and Inkpen, D. (2008). Semantic text similarity

using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, 2:10:1–10:25.

Leacock, C., Miller, G. A., and Chodorow, M. (1998). Using corpus statistics and wordnet relations for sense identification. *Comput. Linguist.*, 24:147–165.

Martin, R. C., Bolter, J. F., Todd, M. E., Gouvier, W. D., and Niccolls, R. (1993). Effects of sophistication and motivation on the detection of malingered memory performance using a computerized forced-choice task. *Journal of Clinical and Experimental Neuropsychology*, 15(6):867–880.

Milne, D. and Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence WIKIAI08*, pages 25–30.

Nakayama, K., Hara, T., and Nishio, S. (2008). Wikipedia link structure and text mining for semantic relation extraction. *Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008)*, pages 59–73.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Comput. Surv.*, 33:31–88.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855*.

Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Slimani, T., Ben Yaghlane, B., and Mellouli, K. (2006). A new similarity measure based on edge counting. *World Academy of Science, Engineering and Technology*, 23(8):34–38.

Smyth, J. D., Dillman, D. A., Christian, L. M., and Stern, M. J. (Spring 2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, 70(1):66–77.

Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1419–1424. AAAI Press.