Automatic generation of inter-passage links based on semantic similarity

Conference or Workshop Item

Automatic generation of inter-passage links based on semantic similarity

Petr Knoth, Jakub Novotny, Zdenek Zdrahal
Knowledge Media Institute
The Open University
p.knoth@open.ac.uk

Abstract

This paper investigates the use and the prediction potential of semantic similarity measures for automatic generation of links across different documents and passages. First, the correlation between the way people link content and the results produced by standard semantic similarity measures is investigated. The relation between semantic similarity and the length of the documents is then also analysed. Based on these findings a new method for link generation is formulated and tested.

1 Introduction

Text retrieval methods are typically designed to find documents relevant to a query based on some criterion, such as BM25 or cosine similarity (Manning et al., 2008). Similar criteria have also been used to identify documents relevant to the given reference document, thus in principle linking the reference document to the related documents (Wilkinson and Smeaton, 1999). This paper studies the correspondence between the results of this approach and the way linking is performed by people. The study confirms that the length of documents is an important factor usually causing the quality of current link generation approaches to deteriorate. As a result, methods working at a finer granularity than documents should be investigated. This will also improve the speed of access to information. For example, when users read through a long document, they should be able to quickly access a passage in another possibly long document related to the discussed topic. The automatic detection of document pairs containing highly related passages is the task addressed in this paper.

A number of approaches for automatic link generation have used measures of semantic similarity. While these measures were widely used for the discovery of related documents in practise, their correspondence to the way people link content has not been sufficiently investigated (see Section 2). As our contribution to this topic, we present in this paper an approach which tries to first investigate this correspondence on a large text corpus. The resulting method is then motivated by the outcomes of this analysis.

It has been recognised in information retrieval that when a collection contains long documents, better performance is often achieved by breaking each document into subparts or passages and comparing these rather than the whole documents to a query (Manning et al., 2008). A suitable granularity of the breakdown is dependent on a number of circumstances, such as the type of the document collection or the information need. In this work, we have decided to work at the level of documents and paragraphs. Our task can be formalized as a two-step process:

1. Given a collection of documents, our goal is to identify candidate pairs of documents between which a link may be induced.

2. Given each candidate pair of documents, our task is to identify pairs of passages, such that the topics in the passages are related in both documents.

The method presented in this paper has many

This work has been partially supported by Eurogene - Contract no. ECP-2006-EDU-410018)

potential applications. First, it may be used for the interlinking of resources that were not originally created as hypertext documents and for the maintenance or the discovery of new links as the collection grows. Second, the method can be applied to improve navigation in collections with long texts, such as books or newspaper articles. A link may be identified by the system automatically and the user can be pointed immediately to the part of the text which is relevant to the block of text currently being read. Similar application has been developed by (Kolak and Schilit, 2008) who provided a method for mining repeated word sequences (quotations) from very large text collections and integrated it with the Google Books archive. Other application areas may involve text summarization and information retrieval.

The paper makes the following contributions:

- It provides a new interpretation and insight in the use of semantic similarity measures for the automatic generation of links.
- It develops a novel two-step approach for the discovery of passage-passage links across potentially long documents and it identifies and discusses the selection of the parameters.

The rest of the paper is organized as follows. Section 2 presents the related work in the field. Section 3 discusses the data selected for our experiment and Section 4 describes how the data were processed in order to perform our investigation. In Section 5, the analysis in which we compared the results produced by semantic similarity measures with respect to the way people link content is presented. Section 6 then draws on this analysis and introduces the method for automatic generation of links which is finally evaluated in Section 7.

2 Related Work

In the 1990s, the main application area for link generation methods were hypertext construction systems. A survey of these methods is provided by (Wilkinson and Smeaton, 1999). In the last decade, methods for finding related documents became the de-facto standard in large digital repositories, such as PubMed or the ACM Digital Library. Search engines including Google also generate links to related pages or research articles.

Generating links pointing to units of a smaller granularity than a document, which can be considered as a task of *passage* or *focused* retrieval, has also been addressed recently. In this task, the system locates the relevant information inside the document instead of only providing a link to the document. The Initiative for the Evaluation of XML retrieval (INEX) started to play an essential role in link generation by providing tracks for the evaluation of link generation systems (Huang et al., 2008; Huang et al., 2009) using the Wikipedia collection at both the document and the passage level.

Current approaches can be divided into three groups: (1) *link-based* approaches discover new links by exploiting an existing link graph (Itakura and Clarke, 2008; Jenkinson et al., 2008; Lu et al., 2008). (2) *semi-structured* approaches try to discover new links using semi-structured information, such as the anchor texts or document titles (Geva, 2007; Dopichaj et al., 2008; Granitzer et al., 2008). (3) *purely content-based* approaches use as an input plain text only. They typically discover related resources by calculating semantic similarity based on document vectors (Allan, 1997; Green, 1998; Zeng and Bloniarz, 2004; Zhang and Kamps, 2008; He, 2008). Some of the mentioned approaches, such as (Lu et al., 2008), combine multiple approaches.

Although link generation methods are widely used in practise, more work is needed to understand which features contribute to the quality of the generated links. Work in this area includes the study of (Green, 1999) who investigated how lexical chaining based on ontologies can contribute to the quality of the generated links, or the experiments of (Zeng and Bloniarz, 2004) who compared the impact of the manually and automatically extracted keywords. There has also been effort in developing methods that can in addition to link generation assign a certain semantic type to the extracted links and thus describe the relationship between documents (Allan, 1997).

The method presented in this paper is purely content-based and therefore is applicable in any text collection. Its use in combination with link-based or semi-structured approaches is also possible. The rationale for the method comes from

the analysis of the prediction potential of semantic similarity for automatic link generation presented in Section 5. Related analysis is presented in (He, 2008) which claims that linked articles are more likely to be semantically similar[1], however, the study does not provide sufficient evidence to confirm and describe this relationship. In link generation, we are more interested in asking the opposite question, i.e. whether articles with higher semantic similarity are more likely to be linked. Our study provides a new insight into this relationship and indicates that the relationship is in fact more complex than originally foreseen by He.

3 Data selection

This section introduces the document collection used for the analysis and the experiments. The following properties were required for the document collection to be selected for the experiments. First, in order to be able to measure the correlation between the way people link content and the results produced by semantic similarity measures, it was necessary to select a document collection which can be considered as relatively well interlinked. Second, it was important for us to work with a collection containing a diverse set of topics. Third, we required the collection to contain articles of varied length. We were mostly interested in long documents, which create conditions for the testing of passage retrieval methods. We decided to use the Wikipedia collection, because it satisfies all our requirements and has also been used in the INEX Link-The-Wiki-Track.

Wikipedia consists of more than four million pages spread across five hundred thousands categories. As it would be for our calculation unnecessarily expensive to work with the whole encyclopedia, a smaller, but still a sufficiently large subset of Wikipedia, which satisfies our requirements of topic diversity and document length, was selected. Our document collection was generated from articles in categories containing the words United Kingdom. This includes categories, such as United Kingdom, Geography of United Kingdom or History of the United Kingdom. There are about 3,000 such categories and 57,000 distinct articles associated to them. As longer articles provide better test conditions for passage retrieval methods, we selected the 5,000 longest articles out of these 57,000. This corresponds to a set where each article has the length of at least 1,280 words.

[1] With respect to the cosine similarity measure.

4 Data preprocessing

Before discussing the analysis performed on the document collection, let us briefly describe how the documents were processed and the semantic similarity calculated.

First, the N articles/documents $D = \{d_1, d_2, \ldots, d_N\}$ in our collection were preprocessed to extract plain text by removing the Wiki markup. The documents were then tokenized and a dictionary of terms $T = \{t_1, t_2, \ldots, t_M\}$ was created. Assuming that the order of words can be neglected (the bag-of-words assumption) the document collection can be represented using a $N \times M$ term-document matrix. In this way, each document is modelled as a vector corresponding to a particular row of the matrix. As it is inefficient to represent such a sparse vector in memory (most of the values are zeros), only the non-zero values were stored. *Term frequency - inverse document frequency (tfidf)* weighting was used to calculate the values of the matrix. Term frequency tf_{t_i,d_j} is a normalized frequency of term t_i in document d_j:

$$tf_{t_i,d_j} = \frac{f(t_i, d_j)}{\sum_k f(t_k, d_j)}$$

Inverse document frequency idf_{t_i} measures the general importance of term t_i in the collection of documents D by counting the number of documents which contain term t_i:

$$idf_{t_i} = \log \frac{|D|}{|d_j : t_i \in d_j|}$$

$$tfidf_{t_i,d_j} = tf_{t_i,d_j}.idf_{t_i}$$

Similarity is then defined as the function $sim(\overrightarrow{x}, \overrightarrow{y})$ of the document vectors $\overrightarrow{x}$ and $\overrightarrow{y}$. There exists a number of similarity measures used for the calculation of similarity between two vectors (Manning and Schuetze, 1999), such as *cosine*, *overlap*, *dice* or *Jaccard* measures. Some studies employ algorithms for the reduction of dimensions of the vectors prior to the calculation

of similarity to improve the results. These approaches may involve techniques, such as lexical chaining (Green, 1999), Latent Semantic Indexing (Deerwester et al., 1990), random indexing (Widdows and Ferraro, 2008) and Latent Dirichlet Allocation (Blei et al., 2003). In this work we intentionally adopted perhaps the most standard similarity measure - cosine similarity calculated on the *tfidf* vectors and no dimensionality reduction technique was used. The formula is provided for completeness:

$$sim_{cosine}(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x}.\overrightarrow{y}}{|x|.|y|}$$

Cosine similarity with *tfidf* vectors has been previously used in automatic link generation systems producing state-of-the-art results when compared to other similarity measures (Chen et al., 2004). This allows us to report on the effectiveness of the most widely used measure with respect to the way the task is completed by people. While more advanced techniques might be in some cases better predictors for link generation, we did not experiment with them as we preferred to focus on the investigation of the correlation between the most widely used measure and manually created links. Such study has to our knowledge never been done before, but it is necessary for the justification of automatic link generation methods.

5 Semantic similarity as a predictor for link generation

The document collection described in Section 3 has been analysed as follows. First, pair-wise similarities using the formulas described in Section 4 were calculated. Cosine similarity is a symmetric function and, therefore, the calculation of all inter-document similarities in the dataset of $5,000$ documents requires the evaluation of $\frac{5{,}000^2}{2} - 5,000 = 12,495,000$ combinations. Figure 1 shows the distribution of the document pairs (on a log_{10} scale) with respect to their similarity value. The frequency follows a power law distribution. In our case, 99% of the pairs have similarity lower than 0.1.

To compare the semantic similarity measures with the links created by Wikipedia authors, all inter-document intra-collection links, i.e. links

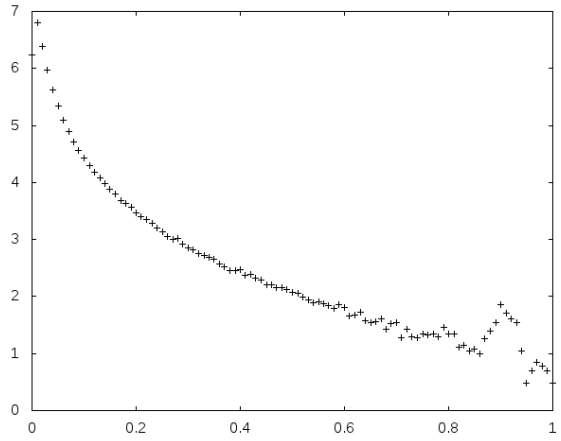

Figure 1: The histogram shows the number of document pairs on a $\log_{10}$ scale (y-axis) with respect to their cosine similarity (x-axis).

created by users of Wikipedia commencing from and pointing to a document within our collection, were extracted. These links represent the connections as seen by the users regardless of their direction. Each of these links can be associated with a similarity value calculated in the previous step. Documents with similarity lower than 0.1 were ignored. Out of the $120,602$ document pairs with inter-document similarity higher than 0.1, $17,657$ pairs were also connected by a user-created link.

For the evaluation, interval with cosine similarity $[0.1, 1]$ was divided evenly into 100 buckets and all 120,602 document pairs were assigned to the buckets according their similarity values. From the distribution shown in Figure 1, buckets corresponding to higher similarity values contain fewer document pairs than buckets corresponding to smaller similarity values. Therefore, for each bucket, the number of user created links within the bucket was normalized by the number of document pairs in the bucket. This number is the likelihood of the document pair being linked and will be called *linked-pair likelihood*. The relation between semantic similarity and linked-pair likelihood is shown in Figure 2.

As reported in Section 2, semantic similarity has been previously used as a predictor for the automatic generation of links. The typical scenario was that the similarity between pairs of documents was calculated and the links between the

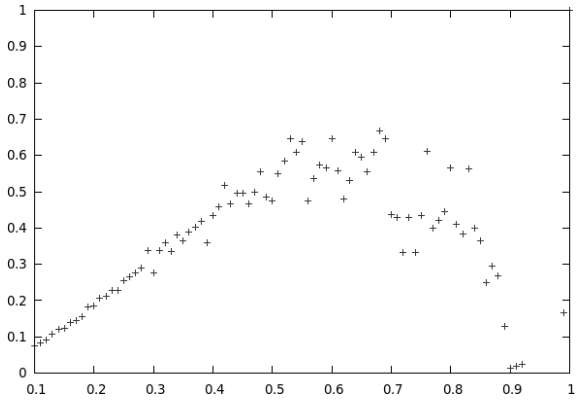

Figure 2: The linked-pair likelihood (y-axis) with respect to the cosine similarity (x-axis).

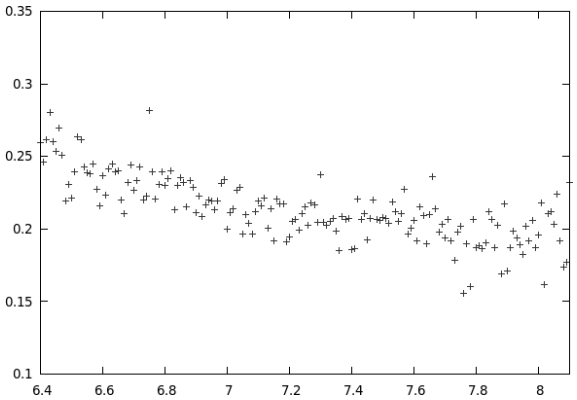

Figure 3: The average cosine similarity (y-axis) of document pairs of various length (x-axis) between which there exists a link. The x-axis is calculated as a $\log_{10}(l_1.l_2)$

most similar documents were generated (Wilkinson and Smeaton, 1999). If this approach was correct, we would expect the curve shown in Figure 2 to be monotonically increasing. However, the relation shown in Figure 2 is in accordance with our expectations only up to the point 0.55. For higher values of inter-document similarity the linked-pair likelihood does not rise or it even decreases.

Spearman's rank correlation and Pearson correlation were applied to estimate the correlation coefficients and to test the statistical significance of our observation. This was performed in two intervals: $[0, 0.55]$ and $[0.55, 1]$. A very strong positive correlation 0.986 and 0.987 have been received in the first interval for the Spearman's and Pearson coefficients respectively. A negative correlation -0.640 and -0.509 have been acquired for the second interval again for the Spearman's and Pearson coefficients respectively. All the measured correlations are significant for p-value well beyond $p < 0.001$. Very similar results have been achieved using different collections of documents.

The results indicate that high similarity value is not necessarily a good predictor for automatic link generation. A possible explanation for this phenomenon is that people create links between related documents that provide new information and therefore do not link nearly identical content. However, as content can be in general linked for various purposes, more research is needed to investigate if document pairs at different similarity levels also exhibit different qualitative properties. More specifically, can the value of semantic similarity be used as a predictor for relationship typing?

An important property of semantic similarity as a measure for automatic generation of links is the robustness with respect to the length of documents. As mentioned in Section 4, cosine similarity is by definition normalized by the product of the documents length. Ideally the cosine similarity should be independent of the documents length. To verify this in our dataset, we have taken pairs of documents between which Wikipedia users assigned links and divided them into buckets with respect to the function $\log_{10}(l_1.l_2)$, where l_1 and l_2 are the lengths of the two documents in the document pair and the logarithm is used for scaling. The value of each bucket was calculated as an average similarity of the bucket members. The results are shown in Figure 3. The graph shows that the average similarity value is slightly decreasing with respect to the length of the articles. Values -0.484 and -0.231 were obtained for Spearman's and Pearson correlation coefficients respectively. Both correlations are statistically significant for $p < 0.001$. A much stronger correlation was measured for Spearman's than for Pearson which can be explained by the fact that Spearman's correlation is calculated based on ranks rather than real values and is thus less sensitive to outliers.

Our experience from repeating the same experiment on another Wikipedia subset generated from categories containing the word Geography tells us that the decrease is even more noticeable when short and long articles are combined. The decrease in average similarity suggests that if cosine similarity is used for the automatic generation of links then document pairs with higher value of $l_1.l_2$ have a higher linked-pair likelihood than pairs with a smaller value of this quantity. In other words, links created between documents with small $l_1.l_2$ typically exhibit a larger value of semantic similarity than links created between documents with high value of $l_1.l_2$. Although the decrease may seem relatively small, we believe that this knowledge may be used for improving automatic link generation methods by adaptively modifying the thresholds with respect to the $l_1.l_2$ length.

6 Link generation method

In this section we introduce the method for the automatic generation of links. The method can be divided into two parts (1) Identification of candidate link pairs (i.e. the generation of document-to-document links) (2) Recognition of passages sharing a topic between the two documents (i.e. the generation of passage-to-passage links).

6.1 Document-to-document links

The algorithm for link generation at the granularity of a document is motivated by the findings reported in Section 5.

Algorithm 1: Generate document links
Input: A set of document vectors D,
min. sim. α, max. sim. $\beta \in [0,1], C = \emptyset$
Output: A set C of candidate links
of form $\langle d_i, d_j, sim\rangle \in C$ where d_i and d_j are documents and $sim \in [0,1]$ is their similarity
1.**for each** $\{\langle d_i, d_j\rangle | i,j \in \aleph_0 \wedge i < j < |D|\}$ **do**
2. $sim_{d_i,d_j} := similarity(d_i, d_j)$
3. **if** $sim_{d_i,d_j} > \alpha \wedge sim_{d_i,d_j} < \beta$ **then**
4. $C := C \cup \langle d_i, d_j, sim_{d_i,d_j}\rangle$

The algorithm takes as the input a set of document vectors and two constants - the minimum and maximum similarity thresholds - and iterates over all pairs of document vectors. It outputs all document vector pairs, such that their similarity is higher than α and smaller than β. For well chosen β, the algorithm does not generate links between nearly duplicate pairs. If we liked to rank the discovered links according to the confidence of the system, we would suggest to assign each pair a value using the following function.

$$rank_{d_i,d_j} = |sim_{d_i,d_j} - (\alpha + \frac{\beta - \alpha}{2})|$$

The ranking function makes use of the fact that the system is most confident in the middle of the similarity region defined by constants α and β, under the assumption that suitable values for these constants are used. The higher the rank of a document pair, the better the system's confidence.

6.2 Passage-to-passage links

Due to a high number of combinations, it is typically infeasible even for relatively small collections to generate passage-to-passage links across documents directly. However, the complexity of this task is substantially reduced when passage-to-passage links are discovered in a two-step process.

Algorithm 2: Generate passage links
Input: Sets P_i, P_j of paragraph document vectors for each pair in C
min. sim. γ, max. sim. $\delta \in [0,1]$ such that $\alpha < \gamma \wedge \beta < \delta, , L = \emptyset$
Output: A set L of passage links
of form $\langle p_{k_i}, p_{l_j}, sim\rangle \in L$ where p_{k_i} and p_{l_j} are paragraphs in documents d_i, d_j and $sim \in [0,1]$ is their similarity
1.**for each** $\{\langle p_{k_i}, p_{l_j}\rangle | p_{k_i} \in P_i, p_{l_j} \in P_j\}$ **do**
2. $sim_{p_{k_i},p_{l_j}} := similarity(p_{k_i}, p_{l_j})$
3. **if** $sim_{p_{k_i},p_{l_j}} > \gamma \wedge sim_{p_{k_i},p_{l_j}} < \delta$ **then**
4. $L := L \cup \langle p_{k_i}, p_{l_j}, sim_{p_{k_i},p_{l_j}}\rangle$

As Section 5 suggests, the results of Algorithm 1 may be improved by adaptive changing of the thresholds α and β based on the length of the document vectors. More precisely, in the case of cosine similarity, this is the quantity $lr = l_1.l_2$. The

value α should be higher (β lower) for pairs with low lr than for pairs with high lr and vice versa. Although the relative quantification of this ratio is left for future work, we believe that we can exploit these findings for the generation of passage-to-passage links.

More specifically, we know that the length of passages (paragraphs in our case) is lower than the length of the whole documents. Hence, the similarity of a linked passage-to-passage pair should be on average higher than the similarity of a linked document-to-document pair, as revealed by the results of our analysis. This knowledge is used within Algorithm 2 to set the parameters γ and δ. The algorithm shows, how passage-to-passage links are calculated for a single document pair previously identified by Algorithm 1. Applying the two-step process allows the discovery of document pairs, which are likely to contain strongly linked passages, at lower similarity levels and to recognize the related passages at higher similarity levels while still avoiding duplicate content.

7 Results

The experimental evaluation of the methods presented in Section 6 is divided into two parts: (1) the evaluation of document-to-document links (Algorithm 1) and (2) the evaluation of passage-to-passage links (Algorithm 2).

7.1 Evaluation of document-to-document links

As identified in Section 5 (and shown in Figure 2), the highest linked-pair likelihood does not occur at high similarity values, but rather somewhere between similarity 0.5 and 0.7. According to Figure 2, the linked-pair likelihood in this similarity region ranges from 60% to 70%. This value is in our view relatively high and we think that it can be explained by the fact that Wikipedia articles are under constant scrutiny by users who eventually discover most of the useful connections. However, how many document pairs that could be linked in this similarity region have been missed by the users? That is, how much can our system help in the discovery of possible connections?

Suppose that our task would be to find document pairs about linking of which the system is most certain. In that case we would set the thresholds α and β somewhere around these values depending on how many links we would like to obtain. In our evaluation, we have extracted pairs of documents from the region between $\alpha = 0.65$ and $\beta = 0.70$ regardless of whether there originally was a link assigned by Wikipedia users. An evaluation tool which allowed a subject to display the pair of Wiki documents next to each other and to decide whether there should or should not be a link between the documents was then developed. We did not inform the subject about the existence or non-existence of links between the pages. More specifically, the subject was asked to decide yes (link generated correctly) if and only if they found it beneficial for a reader of the first or the second article to link them together regardless of the link direction. The subject was asked to decide no (link generated incorrectly) if and only if they felt that navigating the user from or to the other document does not provide additional value. For example, in cases where the relatedness of the documents is based on their lexical rather than their semantic similarity.

The study revealed that 91% of the generated links were judged by the subject as correct and 9% as incorrect. Table 1 shows the results of the experiment with respect to the links originally assigned by the users of Wikipedia. It is interesting to notice that in 3% of the cases the subject decided not to link the articles even though they were in fact linked on Wikipedia. Overall, the algorithm discovered in 30% of the cases a useful connection which was missing in Wikipedia. This is in line with the findings of (Huang et al., 2008) who claims that the validity of existing links in Wikipedia is sometimes questionable and useful links may be missing.

An interesting situation in the evaluation occurred when the subject discovered a pair of articles with titles *Battle of Jutland* and *Night Action at the Battle of Jutland*. The Wikipedia page indicated that it is an orphan and asked users of Wikipedia to link it to other Wikipedia articles. Our method would suggest the first article as a good choice.

		Wikipedia link	
		yes	no
Subject's decision	yes	0.61	0.30
	no	0.03	0.06

Table 1: Document-to-document links from the $[0.65, 0.7]$ similarity region. The subject's decision in comparison to the Wikipedia links.

		Wikipedia link	
		yes	no
Subject's decision at page level	yes	0.16	0.10
	no	0.18	0.56

Table 2: Document-to-document candidate links generation from the $[0.2, 0.21]$ similarity region and document pairs with high lr ($lr \in [7.8 - 8]$).

7.2 Evaluation of passage-to-passage linking

The previous section provided evidence that the document-to-document linking algorithm is capable of achieving high performance when parameters α, β are well selected. However, Section 5 indicated that it is more difficult to discover links across long document pairs. Thereby, we have evaluated the passage-to-passage linking on document pairs with quite low value of similarity $[0.2, 0.21]$. According to Figure 2, this region has only 15% linked-pair likelihood.

Clearly, our goal was not to evaluate the approach in the best possible environment, but rather to check whether the method is able to discover valuable passage-to-passage links from very long articles with low similarity. Articles with this value of similarity would be typically ranked very poorly by link generation methods working at the document level.

Table 2 shows the results after the first step of the approach, described in Section 6, with respect to the links assigned by Wikipedia users. As in the previous experiment, the subject was given pairs of documents and decided whether they should or should not be linked. Parameters α and β were set to 0.2, 0.21 respectively. Table 2 indicates that that the accuracy $(16\% + 10\% = 26\%)$ is at this similarity region much lower than the one reported in Table 1, which is exactly in line with our expectations. It should be noticed that 34% of the document pairs were linked by Wikipedia users, even though only 15% would be predicted by linked-pair likelihood shown in Figure 2. This confirms that long document pairs exhibit a higher probability of being linked in the same similarity region than shorter document pairs.

		System's decision	
		yes	no
Subject's decision	yes (correct)	0.14	0.46
	no (incorrect)	0.24	0.16

Table 3: Passage-to-passage links generation for very long documents. Passages extracted from the $[0.4, 0.8]$ similarity region.

If our approach for passage-to-passage link generation (Algorithm 2) is correct, we should be able to process the documents paragraphs and detect possible passage-to-passage links. The selection of the parameters γ and δ influences the willingness of the system to generate links. For this experiment, we set the parameters γ, δ to 0.4, 0.8 respectively. The subject was asked to decide: (1) if the connection discovered by the link generation method at the granularity of passages was useful (when the system generated a link) (2) whether the decision not to generate link is correct (when the system did not generate a link). The results of this evaluation are reported in Table 3. It can be seen that the system made in 60% $(14\% + 46\%)$ of the cases the correct decision. Most mistakes were made by generating links that were not sufficiently related (24%). This might be improved by using a higher value of γ (lower value of δ).

8 Conclusions

This paper provided a new insight into the use of semantic similarity as a predictor for automatic link generation by performing an investigation in the way people link content. This motivated us in the development of a novel purely content-based approach for automatic generation of links at the granularity of both documents and paragraphs which does not expect semantic similarity and linked-pair likelihood to be directly proportional.

References

Allan, James. 1997. Building hypertext using information retrieval. *Inf. Process. Manage.*, 33:145–159, March.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *JOURNAL OF MACHINE LEARNING RESEARCH*, 3:993–1022.

Chen, Francine, Ayman Farahat, and Thorsten Brants. 2004. Multiple similarity measures and source-pair information in story link detection. In *In HLT-NAACL 2004*, pages 2–7.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Dopichaj, Philipp, Andre Skusa, and Andreas Heß. 2008. Stealing anchors to link the wiki. In Geva et al. (Geva et al., 2009), pages 343–353.

Geva, Shlomo, Jaap Kamps, and Andrew Trotman, editors. 2009. *Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and Selected Papers*, volume 5631 of *Lecture Notes in Computer Science*. Springer.

Geva, Shlomo. 2007. Gpx: Ad-hoc queries and automated link discovery in the wikipedia. In Fuhr, Norbert, Jaap Kamps, Mounia Lalmas, and Andrew Trotman, editors, *INEX*, volume 4862 of *Lecture Notes in Computer Science*, pages 404–416. Springer.

Granitzer, Michael, Christin Seifert, and Mario Zechner. 2008. Context based wikipedia linking. In Geva et al. (Geva et al., 2009), pages 354–365.

Green, Stephen J. 1998. Automated link generation: can we do better than term repetition? *Comput. Netw. ISDN Syst.*, 30(1-7):75–84.

Green, Stephen J. 1999. Building hypertext links by computing semantic similarity. *IEEE Trans. on Knowl. and Data Eng.*, 11(5):713–730.

He, Jiyin. 2008. Link detection with wikipedia. In Geva et al. (Geva et al., 2009), pages 366–373.

Huang, Wei Che, Andrew Trotman, and Shlomo Geva. 2008. Experiments and evaluation of link discovery in the wikipedia.

Huang, Wei Che, Shlomo Geva, and Andrew Trotman. 2009. Overview of the inex 2009 link the wiki track.

Itakura, Kelly Y. and Charles L. A. Clarke. 2008. University of waterloo at inex 2008: Adhoc, book, and link-the-wiki tracks. In Geva et al. (Geva et al., 2009), pages 132–139.

Jenkinson, Dylan, Kai-Cheung Leung, and Andrew Trotman. 2008. Wikisearching and wikilinking. In Geva et al. (Geva et al., 2009), pages 374–388.

Kolak, Okan and Bill N. Schilit. 2008. Generating links by mining quotations. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 117–126, New York, NY, USA. ACM.

Lu, Wei, Dan Liu, and Zhenzhen Fu. 2008. Csir at inex 2008 link-the-wiki track. In Geva et al. (Geva et al., 2009), pages 389–394.

Manning, Christopher D. and Hinrich Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition, June.

Manning, Ch. D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge, July.

Widdows, Dominic and Kathleen Ferraro. 2008. Semantic vectors: a scalable open source package and online technology management application. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias, editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Wilkinson, Ross and Alan F. Smeaton. 1999. Automatic link generation. *ACM Computing Surveys*, 31.

Zeng, Jihong and Peter A. Bloniarz. 2004. From keywords to links: an automatic approach. *Information Technology: Coding and Computing, International Conference on*, 1:283.

Zhang, Junte and Jaap Kamps. 2008. A content-based link detection approach using the vector space model. In Geva et al. (Geva et al., 2009), pages 395–400.