Identifying Novel Information using Latent Semantic Analysis in the WiQA Task at CLEF 2006

Richard F. E. Sutcliffe*[1], Josef Steinberger#, Udo Kruschwitz*,
Mijail Alexandrov-Kabadjov*, Massimo Poesio*

Department of Computer Science*
University of Essex, Wivenhoe Park,
Colchester CO4 3SQ, UK

Department of Computer Science and Engineering#
University of West Bohemia, Univerzitni 8
306 14 Plzen, Czech Republic

rsutcl@essex.ac.uk jstein@kiv.zcu.cz udo@essex.ac.uk
malexa@essex.ac.uk poesio@essex.ac.uk

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; I.2 [Artificial Intelligence]: I.2.7 Natural Language Processing

General Terms

Question Answering, Latent Semantic Analysis, Information Filtering

Abstract

From the perspective of WiQA, the Wikipedia can be considered as a set of articles each having a unique title. In the WiQA corpus articles are divided into sentences (snippets) each with its own identifier. Given a title, the task is to find snippets which are Important and Novel relative to the article. We indexed the corpus by sentence using Terrier. In our two-stage system, snippets were first retrieved if they contained an exact match with the title. Candidates were then passed to the Latent Semantic Analysis component which judged them Novel if they did *not* match the text of the article. The test data was varied – some articles were long, some short and indeed some were empty! We prepared a training collection of twenty topics and used this for tuning the system. During evaluation on 65 topics divided into categories Person, Location, Organization and None we submitted two runs. In the first, the ten best snippets were returned and in the second the twenty best. Run 1 was best with Average Yield per Topic 2.46 and Precision 0.37. We also studied performance on six different topic types: Person, Location, Organization and None (all specified in the corpus), Empty (no text) and Long (a lot of text). Precision results in Run 1 for Person and Organization were good (0.46 and 0.44) and were worst for Long (0.24). Compared to other groups, our performance was in the middle of the range except for Precision where our system was equal to the best. We attribute this to our use of exact title matches in the IR stage. We found that judging snippets Novel when preparing training data was fairly easy but that Important was subjective. In future work we will vary the approach used depending on the topic type, exploit co-references in conjunction with exact matches and make use of the elaborate hyperlink structure which is a unique and most interesting aspect of Wikipedia.

1. Introduction

This article outlines an experiment in the use of Latent Semantic Analysis (LSA) for selecting information relevant to a topic. It was carried out within the Question Answering using Wikipedia (WiQA) Pilot Task which formed part of the Multiple Language Question Answering Track at the 2006 Cross Language Evaluation Forum (CLEF). We first define the WiQA task for this year. Following this is a brief outline of LSA and its previous application to Natural Language Processing (NLP) tasks. We then describe the development and tuning of our algorithm together with the system which implements it. The runs submitted and results obtained are then outlined. Finally, we draw conclusions for the project and present some directions for future work.

[1] On Sabbatical from University of Limerick, Ireland.

Topic	Carolyn Keene
Original Article	Carolyn Keene Carolyn Keene is the pseudonym of the authors of the Nancy Drew mystery series, published by the Stratemeyer Syndicate. Stratemeyer hired writers, including Mildred Benson, to write the novels in this series, who initially were paid only $125 for each book and were required by their contract to give up all rights to the work and to maintain confidentiality. Edward Stratemeyer's daughter, Harriet Adams, also wrote books in the Nancy Drew series under the pseudonym. Other ghostwriters who used this name to write Nancy Drew mysteries included Leslie McFarlane, James Duncan Lawrence, Nancy Axelrod, Priscilla Doll, Charles Strong, Alma Sasse, Wilhelmina Rankin, George Waller Jr., Margaret Scherf, and Susan Wittig Albert. " " by John Keeline lists the ghostwriters responsible for some individual Nancy Drew books.
Snippet 1	The name Carolyn Keene has also been used to author a shorter series of books entitled The Dana Girls.
Snippet 2	All Nancy Drew books are published under the pseudonym Carolyn Keene regardless of actual author.
Snippet 3	Harriet Adams (born Harriet Stratemeyer , pseudonyms Carolyn Keene and Franklin W. Dixon) (1893 - 1982), U.S. juvenile mystery novelist and publisher; wrote Nancy Drew and Hardy Boys books.

Training Data. A sample topic together with the corresponding article text and three candidate snippets. Topic titles are unique in Wikipedia. In our system, the title was added to the start of the article which is why the name appears twice. The existence of the double quotes is connected with the removal of hyperlinks – an imperfect process. Are the example snippets Important and Novel or not? See the text for a discussion.

2. The WiQA Task

The Wikipedia (Wiki, 2006) is a multilingual free-content encyclopaedia which is publicly accessible over the Internet. From the perspective of WiQA it can be viewed as a set of articles each with a unique title. During their preliminary work, the task organisers created an XML compliant corpus from the English Wikipedia articles (Denoyer and Gallinari, 2006). The title of each article was assigned automatically to one of four subject categories PERSON, LOCATION, ORGANIZATION and NONE. At the same time the text of each article was split into separate sentences each with its own identifier. The complex hyperlink structure of the original Wikipedia is faithfully preserved in the corpus, although we did not use this in the present project.

The general aim of WiQA this year was to investigate methods of identifying information on a topic which is present somewhere in the Wikipedia but not included in the article specifically devoted to that topic. For example, there is an article entitled 'Johann Sebastian Bach' in the Wikipedia. The question WiQA sought to answer is this: What information on Bach is there within the Wikipedia *other than* in this article? The task was formalised by providing participants with a list of articles and requiring their systems to return for each a list of up to twenty sentences (henceforth called snippets) from other articles which they considered relevant to the article and yet not already included in it. There were 65 titles in the test set, divided among the categories PERSON, LOCATION, ORGANIZATION and NONE. Evaluation of each snippet was on the basis of whether it was **supported** (in the corpus), **important** to the topic, **novel** (not in the original article) and **non-repeated** (not mentioned in previously returned snippets for this topic). Evaluation of systems was mainly in terms of the snippets judged supported and important and novel and non-repeated within the first ten snippets returned by a system for each topic. A detailed description of the task and its associated evaluation measures can be found in the general article on the WiQA task in this volume.

3. Latent Semantic Analysis

Latent Semantic Indexing (LSI) was originally developed as an information retrieval technique (Deerwester, Dumais, Furnas, Landauer and Harshman, 1990). A term-by-document matrix of dimensions *t*d* of the kind commonly used for inverted indexing in Information Retrieval (IR) is transformed by Singular Value Decomposition into a product of three matrices *t*r*, *r*r* and *r*d*. The *r*r* matrix is diagonal and contains the eponymous 'singular values' in such a way that the top left element is the most important and the bottom right element is the least important. Using the original *r*r* matrix and multiplying the three matrices together results exactly in the original *t*d* matrix. However, by using only the first *n* dimensions (*1 <= n <= r*) in the *r*r* matrix and setting the others to zero, it is possible to produce an approximation of the original which nevertheless captures the most important common aspects by giving them a common representation. In the original IR context, this meant that even if a word was not in a particular document it could be detected whether or not another word with similar meaning *was* present. Thus, LSI could be used to create representations of word senses automatically.

In abstract terms, LSI can be viewed as a method of identifying 'hidden' commonalities between documents on the basis of the terms they contain. Following on from the original work it was realised that this idea was applicable to a wide range of tasks including information filtering (Foltz and Dumais, 1992) and cross-language information retrieval (Littman, Dumais and Landauer, 1998). Outside IR, the technique is usually referred to as Latent Semantic Analysis. Within NLP, LSA has been applied to a variety of problems such as spelling correction (Jones and Martin, 1997), morphology induction (Schone and Jurafsky, 2000), text segmentation (Choi, Wiemer-Hastings and Moore, 2001), hyponymy extraction (Cederberg and Widdows, 2003), summarisation (Steinberger, Kabadjov, Poesio and Sanchez-Graillet, 2005), and noun compound disambiguation, prepositional phrase attachment and coordination ambiguity resolution (Buckeridge, 2005). It has also been applied to the problem of identifying given/new information (Hempelmann, Dufty, McCarthy, Graesser, Cai and McNamara, 2005).

4. Algorithm Development

4.1 Underlying Idea

We decided on a very simple form of algorithm for our experiments. In the first stage, possibly relevant snippets (i.e. sentences) would be retrieved from the corpus using an IR system. In the second, these would be subjected to the LSA technique in order to estimate their novelty. In previous work, LSA had been applied to summarisation by using it to decide which topics are most important in a document and which sentences are most related to those topics (Steinberger, Kabadjov, Poesio and Sanchez-Graillet, 2005). In this project, the idea was reversed by trying to establish that snippets were novel on the basis that they were **not** 'related' to the original topics.

4.2 Training Data

As this was the first year of the task, there was no training data. However, the organisers did supply 80 example topics. Twenty of these were selected. For each, the title was submitted as an exact IR search to a system indexed by sentence on the entire Wikipedia corpus. The 'documents' (i.e. snippets) returned were saved, as was the bare text of the original article. The snippets were then studied by hand and by reference to the original document were judged to be either 'relevant and novel' or not. This data was then used in subsequent tuning.

An example training topic (Carolyn Keene) can be seen in the figure. Below it are three sample snippets returned by the IR component because they contain the string 'Carolyn Keene'. The first one is clearly Important and Novel because it gives information about a whole series of books written under the same name and not mentioned in the article. The second one is Important because it states that all Nancy Drew books are written under this name. However, it is not Novel because this information is in the original article. Now consider the third example concerning Harriet Adams. The decision here is not quite so easy to make. Adams is mentioned in the article, so the fact that she wrote some of the books is not Novel. However, there is other information which is Novel, for example that she wrote books in the Hardy Boys series and also that she had another pseudonym, Franklin W. Dixon. The question is whether such information is Important. This is very hard to judge. Therefore we concluded that the task was not straightforward even for humans.

Run 1							
	All	Person	Location	Org	None	Empty	Long
No. Topics	65	16	18	16	15	6	15
No. Snippets	435	113	119	112	91	33	123
Supported Snippets	435	113	119	112	91	33	123
Important	226	74	54	61	37	13	49
Important & Novel	165	54	37	51	23	13	29
Important & Novel & Non-Repeated	161	52	37	49	23	12	29
Yield (Top 10)	160	52	36	49	23	12	29
Avg. Yield per Topic (Top 10)	2.46	3.25	2.00	3.06	1.53	2.00	1.93
Mean Reciprocal Rank	0.54	0.67	0.47	0.66	0.34	0.56	0.59
Precision (Top 10)	0.37	0.46	0.30	0.44	0.25	0.36	0.24

Run 2							
	All	Person	Location	Org	None	Empty	Long
No. Topics	65	16	18	16	15	6	15
No. Snippets	682	155	206	188	133	59	210
Supported Snippets	682	155	206	188	133	59	210
Important	310	87	93	84	46	26	81
Important & Novel	223	60	60	72	31	26	45
Important & Novel & Non-Repeated	194	52	53	61	28	19	36
Yield (Top 10)	152	44	37	47	24	15	24
Avg. Yield per Topic (Top 10)	2.34	2.75	2.06	2.94	1.60	2.50	1.60
Mean Reciprocal Rank	0.50	0.58	0.47	0.59	0.36	0.50	0.56
Precision (Top 10)	0.33	0.39	0.29	0.39	0.26	0.38	0.19

Summary of results. Both runs were identical except that in Run 1 the maximum number of snippets returned by the system was 10 while in Run 2 it was 20. The values under the column **All** are those returned by the organisers. All other columns are analyses on different subsets of the topics. The object is to see if performance of the system varies by topic type. The column **Person** shows results just for topics which were of type PERSON and similarly for the columns **Location**, **Organization** and **None**. **Empty** denotes just those topics which contain no text at all (!) while **Long** is restricted to 'long' topics which contain a lot of text.

5. System Architecture and Implementation

5.1 Pre-processing of the Corpus

Following our analysis of the training data, it was decided to adopt a similar approach in the final system. This involved retrieving snippets by exact phrase match. To facilitate this process, the corpus was re-formatted replacing sentence and document identifiers within attributes by the equivalent in elements and at the same time removing all hyperlink information which can occur within words and thus affects retrieval. The new version of the corpus was then indexed using Terrier (Ounis, Amati, Plachouras, He, Macdonald and Lioma, 2006; Terrier, 2006) with each individual snippet (sentence) being considered as a separate document. This meant that any matching of an input query was entirely within a snippet and never across snippets.

5.2 Stages of Processing

An input query consists of the title of a Wikipedia article. An exact search is performed using Terrier, resulting in an ordered list of matching snippets. Those coming from the original article are eliminated. In the actual runs, the number of snippets varies from 0 (where no matching snippets were found at all) to 947. The data is then formatted for LSA processing. The bare text of the original article with no formatting and one sentence per line, including the title which forms the first line, is placed in a text file. A second file is prepared for the topic which contains the snippets, one snippet per line. This file therefore contains between 0 and 947 lines, depending on the topic. LSA processing is then carried out. This assigns to each snippet a probability (between 0 and 1) as to whether it is novel with respect to the topic or not. The *n* snippets with highest probability are then returned, preserving the underlying order determined by Terrier. *n* is either 10 or 20 depending on the run. In the final stage, an xml document is created listing for each topic the selected snippets.

6. Runs and Results

6.1 Runs Submitted

Two runs were submitted, Run 1 in which the number of accepted snippets was at maximum 10, and Run 2 in which it was at maximum 20. In all other respects the runs were identical.

6.2 Evaluation Measures

The table summarises the overall results. In addition to the official figures returned by the organisers (denoted All) we also computed results on various different subsets of the topics. The objective was to see whether the system performed better on some types of topic than on others. Each topic in the corpus had been automatically assigned to one of four categories by the organisers using Named Entity recognition tools. The categories are Person, Location, Organization and None. We decided to make use of these in our analysis. Person denotes topics categorised as describing a person and similarly for Location and Organization. None is assigned to all topics not considered to be one of the first three. To these four were added two further ones of our own invention. Empty denotes an interesting class of topics which consist of a title and no text at all. These are something of an anomaly in the Wikipedia, presumably indicating work in progress. In this case the WiQA task is effectively to create an article from first principles. Finally, Long denotes lengthy articles, rather generally defined by us as ‘more than one page on the screen’.

Each snippet was judged by the evaluators along a number of dimensions. A Supported snippet is one which is indeed in the corpus. If it contains significant information relevant to the topic it is judged Important. This decision is made completely independently of the topic text. It is Novel if it is Important and in addition contains information not already present in the topic. Finally, it is judged Non-Repeated if it is Important and Novel and has not been included in an earlier Important and Novel snippet. Repetition is thus always judged between one snippet and the rest, not relative to the original topic text.

The main judgements in the evaluation are relative to the top ten snippets returned for a topic. The Yield is the count of Important, Novel and Non-Repeated snippets occurring in the first ten, taken across all topics in the group under consideration (e.g. All). The Average Yield is this figure divided by the number of topics in the group, i.e. it is the Yield per topic. Reciprocal Rank is the inverse of the position of the first Important, Novel and Non-Repeated snippet in the top ten (numbered from 1 up to ten) returned in response to a particular topic. The Mean Reciprocal Rank (MRR) is the average of these values over all the topics in the group. MRR is an attempt to measure ‘how high up the list’ useful information starts to appear in the output. Finally Precision is the mean precision (number of Important, Novel and Non-Repeated snippets returned in the first ten for a topic, divided by ten) computed over all the topics.

6.3 Run 1 vs. Run 2

Of the two runs we submitted, Run 1 gave better results than Run 2 in terms of overall Average Yield per Topic, MRR and Precision at top 10 snippets. Having said that, the second run did retrieve far more snippets than the first one (682 as opposed to 435); it also retrieved more Important, Important and Novel, and Important, Novel and Non-Repeated snippets than the first run. However, what counts is how many Important, Novel And Non-Repeated snippets were found in the ten snippets that were ranked highest for each topic (i.e. the Yield). When we look at all 65 topics, then the yield was lower for Run 2 (152 vs. 160). This is not just true for the overall figures, but also for the topics in categories Person, Organization and Long. For Location, None and Empty, yield in the second run was marginally better.

The difference in performance between the two runs can be accounted for by an anomaly in the architecture of our system. The underlying order of snippets is determined by their degree of match with the IR search query. Because this query is simply the title and nothing else, and since all returned snippets must contain the exact title, the degree of match will be related only to the length of the snippet – short snippets will match more highly than long ones. When this list of snippets is passed to the LSA component, a binary decision is made for each snippet (depending on its score) as to whether it is relevant or not. This depends on a snippet's LSA score but not on its position in the ranking. The difference between the runs lies in the number of snippets LSA is permitted to select. In Run 1 this consists of the best ten. In Run 2 when we come to select the best twenty this is a superset of the best ten, but it may well be that lower scoring snippets are now selected which are higher in the underlying Terrier ranking than the higher scoring snippets already chosen for Run 1. In other words, our strategy could result in high scoring snippets in Run 1 being inadvertently pushed down the ranking by low scoring ones in Run 2. This effect could explain why the amount of relevant information in Run 2 is higher but at the same time the Precision is lower.

6.4 Strengths and Weaknesses of the System

To find out where our approach works best we broke down the total number of topics into groups according to the categories identified earlier, i.e. Person (16 topics), Location (18), Organization (16), and None (15). A separate analysis was also performed for the two categories we identified, i.e. Empty (6) and Long (15). Instead of analysing individual topics we will concentrate on the aggregated figures that give us average values over all topics of a category.

We achieved the highest average Yield per Topic (3.25), highest MRR (0.67) as well as highest Precision at top 10 snippets (0.46) for topics of category Person in Run 1. In other words, for queries about persons a third of the top ten retrieved snippets were considered Important, Novel and Non-Repeated. However, the runs did not necessarily retrieve ten snippets for each query; usually we retrieved fewer than that. The Precision indicates that nearly half of all retrieved snippets were high quality matches. These values are better than what we obtained for any of the other categories (including Empty and Long) over both runs. This suggests that our methods work best at identifying Important, Novel and Non-Repeated information about persons.

We also observe that in Run 1 the Average Yield per Topic, MRR and Precision for topics of type Person or Organization are all better than those for type All. The same is true for Run 2.

On the other hand, both our runs are much worse on topics of type None. All the considered measures score much lower for topics of this category: Average Yield per Topic (1.53), MRR (0.34) as well as Precision at top 10 snippets (0.25). Interestingly, these lowest values were recorded in Run 1 which overall is better than Run 2. Nevertheless, Precision at top 10 snippets is even lower for Long topics in both runs (0.23 and 0.19, respectively). The last figure is interesting, because this lowest Precision for the long documents in Run 2 corresponds with the lowest Average Yield per Topic (1.6) but the highest average number of snippets per topic (14.0, i.e. 210 snippets divided by 15 topics). No other category retrieved that many responses per query on average. As a comparison, the lowest average number of returned snippets (5.5, i.e. 33 snippets divided by 6 topics) was recorded for Empty topics in Run 1. The conclusion is that the system (perhaps not surprisingly) seems to suggest few snippets for short documents and far more for long documents. More returned snippets do not, however, mean better quality.

Topics classified as Long or None are therefore a major weakness of our approach, and future work will need to address this. One possibility is that we could use the topic classifications at run time and then apply different methods for different categories.

7. Conclusions

This was our first attempt at this task and at the same time we used a very simple system based around LSA applied only to snippets containing exactly the topic's title. In this context the results are not bad. The system works best for topics of type Person and Organization. Our highest precision overall was 0.46 for Persons in Run

1. Compared with the overall results of all participants, we are broadly in the middle of the range. The exception is overall Precision for Run 1 where we appear to be equal to the highest overall in the task, with a value of 0.37. This result is probably due to our use of exact matches to titles in the IR stage of the system.

Concerning the general task, it was very interesting but in some ways it raised more questions than answers. While it is quite easy to judge Novel and Non-Repeated it is not easy to judge Important. This is a subjective matter and can not be decided upon without considering such factors as maximum article length, intended readership, the existence of other articles on related topics, hyperlink structure (e.g. direct links to other articles containing ‘missing’ information) and editorial policy.

We prepared our own training data and due to lack of time this only amounted to twenty topics with judged snippets. In future years there will be much more data to carry out tuning and this might well affect results.

Our policy of insisting on an exact match of a snippet with the title of a topic resulted in the vast majority of cases in the snippet being about the topic. (There are relatively few cases of topic ambiguity although a few were found.) In other words, the precision of the data passed on to the LSA component was high. On the other hand, we must have missed many important snippets. For example we used no co-reference resolution which might well have increased recall while not affecting precision, certainly in cases like substring co-reference within an article (e.g. ‘Carolyn Keene ... Mrs. Keene’).

The link structure of the Wikipedia is very complex and has been faithfully captured in the corpus. We did not use this at all. It would be very interesting to investigate snippet association measures based on the ‘reachability’ of a candidate snippet from the topic article and to compare the information they yield with that provided by LSA. However, the link structure is not all gain: In many cases only a substring of a token constitutes the link. When the markup is analysed it can be very difficult to recover the token accurately – either it is wrongly split into two or a pair of tokens are incorrectly joined. This must affect the performance of the IR and other components though the difference in performance caused may be slight.

8. References

Buckeridge, A. M. (2005). *Latent Semantic Indexing as a Measure of Conceptual Association for the Unsupervised Resolution of Attachment Ambiguities*. Ph.D. Thesis, University of Limerick.

Cederberg, S., & Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. *Proceedings of the Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*, 111-118.

Choi, F. Y. Y., Wiemer-Hastings, P., & Moore, J. D. (2001). Latent Semantic Analysis for Text Segmentation. *Proceedings of EMNLP, Pittsburgh*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**, 391-407.

Denoyer, L., & Gallinari, P. (2006). The Wikipedia XML Corpus. *SIGIR Forum*, **40**(1) 2006.

Foltz, P. W., & Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. *Communications of the Association for Computing Machinery*, **35**, 51-60.

Hempelmann, C. F., Dufty, D., McCarthy, P. M., Graesser, A. C., Cai, Z., & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 941-946). Mahwah, NJ: Erlbaum.

Jones, M. P., & Martin, J. H. (1997). Contextual spelling correction using Latent Semantic Analysis. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP ’97)*, 166-173.

Littman, M. L., Dumais, S. T., & Landauer, T. K. (1998). Automatic cross-language information retrieval using Latent Semantic Indexing. In G. Grefenstette (Ed.) Cross Language Information Retrieval (pp. 51-62). Norwell, MA: Kluwer Academic Publishers.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. *Proceedings of ACM SIGIR’06 Workshop on Open Source Information Retrieval (OSIR 2006), 10th August, 2006, Seattle, Washington, USA.*

Schone, P., & Jurafsky, D. (2000). Knowledge-free induction of morphology using Latent Semantic Analysis. *Proceedings of the Fourth Conference on Computational Natural Language Learning (CoNLL-2000) and the Second Learning Language in Logic Workshop (LLL-2000),* 67-72.

Steinberger, J., Kabadjov, M. A., Poesio, M., & Sanchez-Graillet, O. (2005). Improving LSA-based Summarization with Anaphora Resolution. *Proceedings of Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, October 2005*, 1–8.

Terrier (2006). http://ir.dcs.gla.ac.uk/terrier/

Wiki (2006). http://en.wikipedia.org