Judging Whether a Document Changes in Subject

Colin Nicholson
Institute for Artificial Intelligence
University of Georgia
dooz@uga.edu

Abstract

This paper describes a method for determining whether a document is composed of text related to a single subject or text that changes subjects. The algorithm involves dividing the document into five equal parts and measuring the text similarity of the different sections with one another. Documents that drift in subject are shown to have a higher standard deviation of similarity values than documents that remain on one subject. This method requires a threshold value that is specific to the domain to work properly.

1. Introduction

Coherence in a text is semantic unity. Discourse made of parts that seem connected in subject matter has a high level of coherence; text that jumps around in subject matter has a low level of coherence. No universally accepted method for measuring coherence exists since coherence relates to a subjective interpretation of how well subjects connect with one another. In fact, individuals in different fields have defined different types of coherence before devising algorithms to measure them [10] [12].

Knowledge of a document's coherence level can help with various tasks. Computerized coherence measurement provides assistance to teachers with grading of essays [13]. Research groups have found coherence measurement to be an important step for developing systems that can locate topically-related material in streams of broadcast speech [14]. Also, determining whether a website is on one or multiple topics can assist search engines return the most relevant pages for queries. Coherence level is information that can assist computational linguists to discover different styles of writing which can help accomplish tasks such as determining authorship of text. Along with many other measurements of text, coherence level can give computers a better indication of the nature of the text they are working with.

Medical researchers also benefit from computerized coherence level measurement. Some cite speech abnormalities as an indicator of certain mental disorders [3]. Currently, software is being developed to advance medical knowledge by finding connections in seemingly unrelated texts [6]. In recent times, researchers have developed software [1] [7] to measure distortions in language that can indicate schizophrenia [11].

The method proposed in this paper has a big advantage over most of the previous work on this subject: simplicity. For example, Todd's work on coherence measurement involves a complicated multi-step process that identifies key concepts in a text, maps relationships between these concepts, and builds a hierarchy to be used with the text to try to measure coherence [13]. Potential for error can be high in multi-step processes, especially when later steps hinge on tasks such as relationship-mapping which can be difficult for computers to accomplish with a high level of accuracy. Elvevaag's Latent Semantic Analysis project to measure coherence [4] is difficult to replicate due to the size of the training corpus; LSA projects usually require hundreds of thousands to millions of documents to train on in order to be effective in distinguishing between subjects [8]. It also suffers from an issue that most coherence-measuring algorithms are stuck with: it does not know what to do with text on a subject that was not trained by the system beforehand.

Brown's work on link detection [2] provided a start in a direction that could be quite worthwhile. The concept is simple: compute the text similarity of two documents and determine if the similarity resides above a predetermined threshold. This system for computing coherence is fast, simple, and works on any domain; it needs no knowledgebase of subject documents to map the current text with. Unfortunately the method produced unreliable results when tested and the research ended with a report on the failed experiment and no usable algorithm. It seems as if one change needs to be made: classification needs to be based on a value other than the text similarity score. The method proposed in this paper involves a measurement for classification that is more stable, even under conditions where the writing style might skew similarity values one way or another.

2. Document Division and Similarity Calculation

In order to measure text similarity, a query document must be compared with other documents to return some kind of value. With one document to work

978-1-4244-3978-2/09/$25.00 ©2009 IEEE

with, the document must be divided up into sections, or subdocuments, which can then be treated as separate documents with different similarity values from a query.

Let's consider the following small document which consists of five sets of sentences, all on the same subject: *1) Dogs are nice pets. Many people own dogs. 2) They usually like people. Dogs are usually loyal. 3) Dogs are good companions. People really love dogs. 4) I own two dogs. They are quite energetic. 5) Dogs eat many things. They are always hungry.*

The four groups were formed by their placement; they represent the beginning, end, and three middle fifths of the document. These will be our five subdocuments. First, we must assign values to each word so that our comparisons will produce numerical results. A popular value to use for text comparison is the inverse document frequency (IDF).

The IDF value aims to give high values to "important" words while minimizing the score for trivial words [5]. This is done by assigning a value to the word that is inversely proportional to the amount of documents the word appears in. The desired effect is to drive common words such as "the," which can appear in any given document, down in value and words that are unique to a given document, perhaps indicating subject matter, up in value. The formula for calculating IDF is to take the logarithm of the number of documents over the number of documents the term appears in. If we assign IDF values to each unique word in the example document we get: *Dogs: 0, Are: 0, Nice: 0.6989, Pets: 0.6989, Many: 0.3979, People: 0.2218, Own: 0.3979, They: 0.0969, Usually: 0.6989, Like: 0.6989, Loyal: 0.6989, Good: 0.6989, Companions: 0.6989, Really: 0.6989, Love: 0.6989, I: 0.6989, Two: 0.6989, Quite: 0.6989, Energetic: 0.6989, Eat: 0.6989, Things: 0.6989 , Hungry: 0.6989.*

Now that each term has a value we can compare the five subdocuments to each other. First we must construct document vectors that give the value of each word with respect to each document. For most terms, the value will be the same as the IDF value; however, every time a term repeats in a document, its value is increased by that term's IDF value.

To compare similarity, we multiply each term value in one document by another and then add the results to get the vector dot product value [6]. Let's take the first subdocument and compute its similarity value to all of the other subdocuments.

S(D1, D2) = 0.0586 *S(D1, D3) = 0.0492*

S(D1, D4) = 0.0586 *S(D1, D5) = 0.0586*

Now let's calculate the similarity for a similar document, only one that drifts in subject. *1) Dogs are nice pets. Many people own dogs. 2) They usually like people. Dogs are usually loyal. 3) Dogs are good companions. People really love dogs. 4) Some people teach math. Everyone must learn math. 5) Math can be difficult. Kids usually dislike math.*

Here are the IDF values for each unique term in the document: *Dogs: 0.2218, Are: 0.2218, Nice: 0.6989, Pets: 0.6989, Many: 0.6989, People: 0.0969, Own: 0.6989, They: 0.3979, Usually: 0.3979, Like: 0.6989, Loyal: 0.6989, Good: 0.6989, Companions: 0.6989, Really: 0.6989, Love: 0.6989, Some: 0.6989, Teach: 0.6989, Math: 0.3979, Everyone: 0.6989, Must: 0.6989, Learn: 0.6989, Can: 0.6989, Be: 0.6989, Difficult: 0.6989, Kids: 0.6989, Dislike: 0.6989.*

Again, we will take the first subdocument as the query and compare it to the next four subdocuments.

S(D1, D2) = 0.1078 *S(D1, D3) = 0.1570*

S(D1, D4) = 0.0094 *S(D1, D5) = 0*

Now it is our job to try and distinguish between the document that is all on one subject and the document that drifts in subject based on the similarity scores. If we look at the averages of these similarity values, the document that was all on one subject had a mean similarity score of 0.0562 while the document that drifted in subject had a mean similarity score of 0.0685. Here we run into a similar problem that Brown faced: the values of the documents are not distinct enough for us to feel comfortable with setting a threshold that would be able to accurately distinguish between documents based on coherence. In fact, if a few key words had been repeated more in the first document, its similarity value would be the same as or more than the second document's similarity score.

Beyond the values themselves, one can find an interesting trend among the data: the distribution of the values. On the document that is all on one subject, the distribution of the values is constant; all similarity scores are fairly close in value. However, the document that drifts in subject shows a lowering in similarity values in the subdocuments that are on a different subject than the input.

The idea of using changes in term frequency to extract information from text is not new; in his work on document cohesion, Hearst looked at the sharpest boundaries where changes of words occur in a document to draw lines on where different sections of a document begin and end [9]. The best way to measure changes in word patterns is with the standard deviation of similarity values. For the document that is all on one subject, the standard deviation of similarity values is 0.0047 while the document that drifts has a standard deviation of 0.0765 for the similarity values, sixteen times as much. Remember, the second document's similarity score is only 1.2 times as much as the first. Clearly, the standard deviation values of similarity scores are more distinct than the average of the similarity scores themselves.

2.1. The Method

The first step in the algorithm for determining a document's coherence is to divide the document into five equally-sized subdocuments; this was shown on a small scale in the example document discussed earlier. We can not have too few subdocuments or we risk the possibility of combining sections that are on different subjects and making sections have a higher similarity value than they should. Also, we can not have too many subdocuments in one document because the average standard deviation value will be weighed down by several low standard deviation values.

In this paper, we will use five subdocuments because that is an ideal number for documents that change in subject around halfway through the document; using five subdocuments makes the first two subdocuments on one subject and the last two on a different subject; when one of these four subdocuments in used as the query, it provides with high values (when it is compared to the nearest subdocument) and low values (when compared to the other two); this provides a great distinction. Now that we have the sections, we assign IDF values to all the unique terms in the document, depending on how many subdocuments each term is in.

At this point, we will find the similarity scores of the first subdocument with the other four subdocuments. When comparing different documents to determine which has a higher trend of coherence, we must normalize the values so that all similarity scores can be compared evenly on a scale from 0 (no similarity) to 1 (perfect similarity) by dividing the four other subdocument similarity scores by the input subdocument's similarity score with itself. Now we have four similarity scores for the first subdocument and we take the standard deviation of these values. The standard deviation tells us how far apart the similarity scores are; if the values have a great range of high and low values we will receive a high standard deviation.

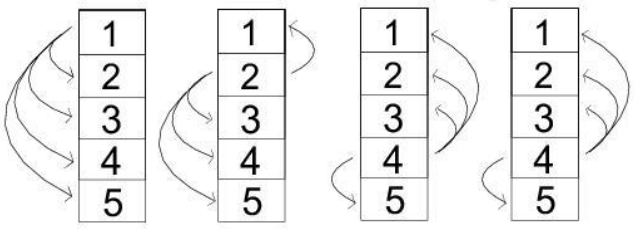

Figure 1

Now we have a value that represents the diversity of the similarity of the first subdocument with the other four subdocuments. We will repeat this same process, only with the second, forth, and fifth subdocuments as the inputs. Figure 1 shows the comparison steps in order. We do not consider the middle subdocument because in documents that change at the halfway mark, it will contain terms from both topics and will give high similarity values to all other subdocuments. Now that we have four standard deviation values, we take the average of these four values to attain our final value. Testing has shown documents that drift in subject generally have higher average standard deviation value than documents that remain on one subject.

2.2. Example Document

Let's run a 3,000 word document on the subject of Mars (retrieved from the website Wikipedia) through the algorithm to illustrate the process. Initially, we make five groups of 600 words each: the beginning, three middle fifths, and end of the document; then we calculate the IDF scores for each unique term depending on how many sections that term appears in.

Now we can calculate the similarity values using the dot product of the vectors and the normalized similarities by dividing the similarities by the query document's similarity with itself. The standard deviation values are calculated using the four normalized similarity values of each subdocument. Table 1 shows the similarity of each subdocument with the others as it is used as a query; Table 2 shows the same results normalized by dividing each result by the similarity of the query subdocument with itself.

Table 1

query	match with 1	match with 2	match with 3	match with 4	match with 5
1	57.7	8.9	3.5	3.1	2.4
2	8.9	52.2	9.3	3.8	3.5
4	3.1	3.8	5.4	66.8	10.8
5	2.4	3.5	2.4	10.8	155.7

Table 2

query	match with 1	match with 2	match with 3	match with 4	match with 5
1	1	0.1551	0.0619	0.0541	0.0423
2	0.1713	1	0.1795	0.0728	0.0680
4	0.0467	0.0569	0.0813	1	0.1627
5	0.0156	0.0228	0.0160	0.0698	1

The average of these four standard deviation values is 0.04779. It should be noted that subdocuments

nearest the query document usually give higher similarity scores than the other subdocuments. This occurs in documents that are all on the same subject, because nearby sections usually share a more specific topic than the subdocuments that are not as close; however, in this example, the size difference in similarity is slight. In a document that drifts, the similarity scores for nearby subdocuments would be much greater and lead to standard deviation values that are, on average, at least twice as high as the one in this document.

2.3. Testing

In this test, 156 pages from the website Wikipedia (http://www.Wikipedia.org), taken on June 27, 2008, were used. Wikipedia is a website that contains pages on various subjects, much like an online dictionary. Pages on each subject are usually divided into sections related to subtopics of the general topic; the text on a given Wikipedia page is often times written by more than one author.

In this test, each page represented the entry page for a different country, from Abkhazia to Zimbabwe. Each page on a certain country is composed of different sections relating to a different aspect of the country. For these pages, the length of the documents ranged from 1,449 words (Andorra) to 16,074 (Cuba). Most documents were in the middle of that range. The goal of this test is to determine if the algorithm proposed in this paper is able to categorize which files are composed of text from one article on one country and which files are composed of text from two articles on two different countries.

2.4. Results

The first step in such a process is to find a threshold value that is appropriate for the given domain. Pages on the first 80 articles (Abkhazia to Japan) were used to represent the average standard deviation value for a document all on one subject while documents that drift in subject were represented by documents that were made of the first half of an article on one country and the second half of the article on the next country, starting with the first article.

The average standard deviation values for single subject documents had an average value of 0.04834. The average standard deviation values for the mixed articles had an average value of 0.11498. To get the threshold, we compute the midpoint of the average standard deviation values for single subject and the average standard deviation values for mixed subject documents: 0.08166.

The threshold value can be used to classify the remaining articles. Two groups are used here: the first is the remaining 74 documents (from Jordan to Zimbabwe) one subject. The other group is made up of the first half of each article combined with the second half of each article below it, starting with the second article. The average standard deviation values for each document is computed, and each document is classified according to whether it is above or below the threshold; documents with values above the threshold are said to drift in subject while documents with values below the threshold are said to remain on subject.

Table 3 shows the values for one-subject documents. Table 4 shows the values for two-subject documents. In the one-subject documents, three documents out of 74 were misclassified (in bold); in the two-subject documents, 8 documents out of 78 were misclassified (in bold).

Table 3

Subject	STD Dev	Subject	STD Dev
Jordan	0.0279	Poland	0.0543
Kazakhstan	0.0585	Portugal	0.0340
Kenya	0.0434	Puerto Rico	0.0590
Korea, North	0.0399	Qatar	0.0534
Korea, South	0.0381	Romania	0.0440
Kosovo	0.0459	Russia	0.0437
Kuwait	0.0371	Rwanda	0.0580
Kyrgyzstan	0.0332	Saudi Arabia	0.0444
Laos	0.0540	Senegal	0.0671
Latvia	0.0249	Serbia	0.0609
Lebanon	0.0517	Sierra Leone	0.0518
Macedonia	0.0442	Singapore	0.0408
Madagascar	0.0538	Slovakia	0.0672
Malaysia	0.0725	Slovenia	0.0753
Mali	**0.0870**	Somalia	0.0521
Malta	0.0532	South Africa	0.0662
Mauritania	0.0539	**Spain**	**0.0820**
Mexico	0.0651	Sri Lanka	0.0231
Moldova	0.0726	Sudan	0.0677
Mongolia	0.0495	Sweden	0.0494
Montenegro	0.0310	Switzerland	0.0680
Morocco	0.0345	Syria	0.0423
Mozambique	0.0376	Taiwan	0.0280
Nambia	0.0534	Tanzania	0.0369
Nepal	**0.1209**	Thailand	0.0618
Netherlands	0.0577	Tunisia	0.0551
New Zealand	0.0517	Turkey	0.0454
Nicaragua	0.0422	Uganda	0.0397
Niger	0.0594	United Arab Emirates	0.0368
Nigeria	0.0451	United Kingdom	0.0393

Norway	0.0552	United States	0.0497
Oman	0.0519	Uruguay	0.0485
Pakistan	0.0535	Uzbekistan	0.0399
Panama	0.0246	Venezuela	0.0275
Papua New Guinea	0.0570	Vietnam	0.0362
Paraguay	0.0552	Yemen	0.0402
Peru	0.0519	Zambia	0.0747
Philippines	0.0596	Zimbabwe	0.0366

Table 4

Subject	STD Dev	Subject	STD Dev
Afghanistan/ Albania	0.1243	Japan/Jordan	0.1232
Algeria/American Samoa	0.1252	Kazakhstan/Kenya	0.1434
Andorra/Angola	0.1069	**Korea, North/Korea, South**	**0.0675**
Argentina/Armenia	0.1685	Kosovo/Kuwait	0.2033
Aruba/Australia	0.1390	Kyrgyzstan/Laos	0.1125
Austria/Azerbaijan	0.1431	Latvia/Lebanon	0.1370
Bahamas/Bahrain	0.0894	Macedonia/ Madagascar	0.1406
Bangladesh/ Barbados	0.0910	Malaysia/Mali	0.1216
Belarus/Belgium	0.1104	Malta/Mauritania	0.0911
Belize/Benin	0.0946	Mexico/Moldova	0.0960
Bermuda/Bhutan	0.1308	Mongolia/ Montenegro	0.1111
Bolivia/Bosnia and Herzegovina	0.1224	Morocco/ Mozambique	0.1030
Botswana/Brazil	0.1407	**Nambia/Nepal**	**0.0814**
Bulgaria/Burma	0.1912	Netherlands/New Zealand	0.1364
Cambodia/ Cameroon	0.1102	Nicaragua/Niger	0.1235
Canada/Central African Republic	0.1162	Nigeria/Norway	0.1068
Chad/Chile	0.0927	Oman/Pakistan	0.1093
China/Colombia	0.1678	Panama/Papua New Guinea	0.1792
Congo DR/Costa Rica	0.1329	**Paraguay/Peru**	**0.0707**
Côte d'Ivoire/Croatia	0.1015	Philippines/Poland	0.1314
Cuba/Cyprus	0.1567	Portugal/Puerto Rico	0.1715
Czech Republic/Denmark	0.1260	**Qatar/Romania**	**0.0628**
Dominican Republic/East Timor	0.1233	Russia/Rwanda	0.1160
Ecuador/Egypt	0.0966	Saudi Arabia/Senegal	0.0914
El Salvador/Estonia	0.1533	Serbia/Sierra Leone	0.2067
Ethiopia/Fiji	0.1221	Singapore/Slovakia	0.0954
Finland/France	0.1145	Slovenia/Somalia	0.1167
Gambia/Georgia	**0.0273**	South Africa/Spain	0.1793
Germany/Ghana	**0.0648**	Sri Lanka/Sudan	0.1610
Greece/Greenland	0.1635	Sweden/Switzerland	0.1791
Grenada/Guam	0.1206	Syria/Taiwan	0.1211
Guatemala/Guinea	0.0884	Tanzania/Thailand	0.1297
Guyana/Haiti	**0.0791**	Tunisia/Turkey	0.0984
Honduras/Hong Kong	0.1938	Uganda/U.A.E.	0.1014
Hungary/Iceland	0.1829	United Kingdom/United States	0.1222
India/Indonesia	0.0946	Uruguay/Uzbekistan	0.1435
Iran/Iraq	**0.0660**	Venezuela/Vietnam	0.1149
Ireland/Israel	0.1930	Yemen/Zambia	0.0918
Italy/Jamaica	0.1127	Zimbabwe/Abkhazia	0.1362

2.5. Why Were Some Misclassified?

2.5.1. Iran/Iraq The document composed of half of Iran and half of Iraq's articles was given a low standard deviation value for similarity. This is understandable if one looks through the articles and sees that both contain sections for the Iran-Iraq war and both articles make references to the other article's country.

2.5.2. North Korea/South Korea The document composed of half of North Korea and half of South Korea's articles was also given a low standard deviation value for similarity. This is understandable because both countries have the same key term, "Korea," and reference the other article's country frequently.

2.6.3. Paraguay/Peru The average standard deviation value of the Paraguay/Peru document was 0.0707; this was not too far below our threshold, but the document was misclassified nonetheless. Each document was on a South American country close in distance to the other one and the two documents contained many similar terms.

2.5.4. The Others Of the five remaining two-subject documents that were misclassified, Guyana/Haiti and Nambia/Nepal just barely missed our threshold value. There is a trend in the remaining three misclassified two-subject documents that may explain what happened with them: one subject is represented by much more text than the other. For instance, the Gambia/Georgia document gave a very low value; however, the Gambia section was less than 20% of the document. Remember: the documents are first divided into five parts. When one subject is less

than one fifth of the total document, it can not produce high similarity values when it is compared with the other sections. This causes a low standard deviation.

Perhaps for these three documents, dividing into five subdocuments was not the best choice and a higher number should have been used. Additionally, three of the one-subject documents were misclassified. Spain just barely missed our threshold value. Nepal and Mali were further away, but still below the threshold. This can happen when there are distinct sections in the article that use many terms unique to that section.

3. Conclusions and Future Work

Out of 152 documents, eleven were misclassified; for an accuracy of 92.7%. However, the previous section of this paper explains how a few of the documents contained overlapping subjects; a test on more distinct subjects would lead to a higher accuracy rating.

Many documents that were misclassified were close to the threshold value. Perhaps instead of a standard classification, the project should take more of a fuzzy logic approach where documents are given scores on a scale of zero to one of how much drift is measured. This would show a distinction between articles that just barely miss the threshold and documents that were far away from the value. Also, there may be other information in the documents that can be combined with the average standard deviation value to lead to more accurate classification. Future work can examine different values from the data to see if they can assist with showing greater distinction between documents that drift and documents that remain on subject.

4. References

[1] Brown, C.; Snodgrass, T.; Kemper, S. J.; Herman, R.; and Covington, M. A. (2008) Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods, 40* (2), 540-545.

[2] Brown, R. D.; Pierce, T.; Yang, Y; and Carbonell, J. G. (2000) Link Detection Results and Analysis. *1999 National Institute of Standards and Technology Topic Detection and Tracking Workshop*. Web: http://citeseer.ist.psu.edu/brown00link.html.

[3] Chaika, E. (1974) A linguist looks at "schizophrenic" language. *Brain and Language. 1*, 257–276, 1974.

[4] Elvevaag, B.; Foltz, P.W.; Weinberger, D.R.; and Goldberg, T.E. (2006) Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research, 93* (1-3), 304 - 316, 2007.

[5] Grossman, D., and Frieder, O. (2004) *Information Retrieval.* Springer: Dordrecht, 2004.

[6] Guernsey, L. (2003) Digging For Nuggets Of Wisdom. *The New York Times.* 16 Oct. 2003.

[7] He, C.; Weinstein, S.; and Covington, M. A. (2006) Speech analysis software for psychiatric research: the case of D-Level Rater, poster presented at the First annual Georgia/South Carolina Neuroscience Consortium, Charleston, April 2006.

[8] Hearst, M. (1997) TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics, 23* (1): 33 - 64, March 1997.

[9] Landauer, T.K.; Foltz P.W.; and Laham, D. (1998) Introduction to Latent Semantic Analysis. *Discourse Processes. 25*: 259-284.

[10] Mann, W. C.; Thompson, S. A. (1988) Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8* (3), 243–281.

[11] Snowdon, D.A.; Kemper, S.I.; Mortimer, J.A.; Greiner, L.H.; Wekstein, D. R; and Markesbery, W. R. (1996) Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. *Journal of the American Medical Association.* 1996; *275* (7):528-532.

[12] Stubbs, M. (1983) *Discourse analysis: The sociolinguistic analysis of natural language.* Oxford: Blackwell, 1983.

[13] Todd, R. W.; Thienpermpool, P.; and Keyuravong, S. (2004) Measuring the coherence of writing using topic-based analysis. *Assessing Writing, 9*, 85-104, 2004.

[14] Wayne, Charles, 2000, Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation In *Proceedings of the Second International Language Resources and Evaluation Conference*, Athens, May 2000.