Discourse Connective - A Marker for Identifying Featured Articles in Biological Wikipedia

Sindhuja Gopalan[1], Paolo Rosso[2] and Sobha Lalitha Devi[1]

[1]AU-KBC Research Centre, MIT, Anna University, Chromepet, Chennai, India
[2] PRHLT Research Centre, Universitat Politècnica de València, Spain
{sindhujagopalan, sobha}@au-kbc.org, prosso@dsic.upv.es

Abstract. Wikipedia is a free-content Internet encyclopedia that can be edited by anyone who accesses it. As a result, Wikipedia contains both featured and non-featured articles. Featured articles are high-quality articles and non-featured articles are poor quality articles. Since there is an exponential growth of Wikipedia articles, the need to identify the featured Wikipedia articles has become indispensable so as to provide quality information to the users. As very few attempts have been carried out in the biology domain of English Wikipedia articles, we present our study to automatically measure the information quality in biological Wikipedia articles. Since the coherence shows representational information quality of a text, we have used the discourse connective count measure for our study. We compare this novel measure with two other popular approaches word count measure and explicit document model method that have been successfully applied to the task of quality measurement in Wikipedia articles. We organized the Wikipedia articles into balanced and unbalanced set. The balanced set contains featured and non-featured articles of equal length and the unbalanced set contains randomly selected featured and non-featured articles. The best result for the balanced set is obtained with F-measure of 83.2%, while using Support Vector Machine classifier with 4-gram representation and Term Frequency-Inverse Document Frequency weighting scheme. Meanwhile, the best result for unbalanced corpus is obtained using the discourse connective count measure with an F - measure of 98.06%.

Keywords: Wikipedia articles quality, Document classification, Featured article, Non-featured article, Word count measure, Discourse connective count measure.

1 Introduction

Wikipedia is a web based, free content encyclopedia with openly editable content. Anyone can write or edit these articles. It was created in 2001 and is a multilingual project in 290 languages. Wikipedia articles are classified into various categories based on their quality. The quality of information includes traditional dimensions such as accuracy, consistency, timeliness, completeness, accessibility, objectiveness and relevancy. Over 4500 articles have been designated as featured articles and 22000

articles as good articles by the Wikipedia community. Featured articles are considered to be the best articles. Wikipedia's strength and weakness is that it is open to anyone. Hence it may also contain low quality content. Non-featured articles are low quality articles that are not of good standard. As Wikipedia articles are increasing enormously in size, it is important to classify these articles as featured and non-featured to provide quality information to the users. The document classification task is to assign a document to one or more classes or categories. Currently there are various document classification works being done on Wikipedia articles of general domain. But, very few works are available for biological domain. Hence, in this study, we have focused on automatically identifying the featured biological Wikipedia articles. In order to increase the participation of computational biology researchers in English Wikipedia and other Wikimedia projects, Wiki Project Computational Biology aims at improving and organizing articles on computational biology, bioinformatics, systems biology and related topics. The overall goal of the project is to improve the article quality of Wikipedia articles within the biology domain [16]. We developed our data set by collecting these articles. Previous works, describes various approaches used to automatically assess the quality of Wikipedia articles.

The paper [2] describes two classification approaches categorization and clustering. This paper discussed the advantages of document classification methods for organizing explicit knowledge. [12] demonstrated the feasibility of building an automatic system to assign quality ratings to Wikipedia articles. Their model obtained an accuracy of 74.6%. [7] investigated four different methods for text classification tasks that include Naive Bayes classifier, the nearest neighbor classifier, decision trees and subspace method. They applied these machine learning techniques to seven class yahoo news groups. The best classification accuracy achieved on seven class problem is approximately 83%. [5] in their work has used inductive learning to categorize natural language documents into predefined content categories.

A variety of simple approaches have been used in the past like number of edits, word count measure, fact count, etc., for automatic classification of featured and non-featured Wikipedia articles. [3] in their paper have analyzed a novel set of features for the task of automatic edit category classification. Using a supervised machine learning experiment, they achieved a micro average F-measure of 62% on a corpus of edits from English Wikipedia. [1] proposed a simple metric word count for measuring article quality. They measured the length of the articles in words. [17] offered new metrics for an efficient quality measurement. Their metrics are based on the life cycles of low and high quality articles. The metrics refer to the changes of the persistent and transient contribution throughout the entire life span. These two metrics are used to measure the editing intensity. [6] in their work demonstrated a simple statistical measure, factual density based on facts extracted from web content using open information extraction. They obtained an F-measure of 90.4% on unbalanced corpus. On balanced corpus, they used relational features for categorizing Wikipedia articles into featured/good and non-featured articles. They obtained an F-measure of 86.7%.

The paper [8] presents the authors' first study to automatically assess information quality in Spanish Wikipedia articles. They evaluated the featured article identification as a binary classification task. Their results show that featured article identifica-

tion for Spanish Wikipedia articles can be performed with an F-measure of 81% when the Support Vector Machines (SVMs) algorithm is used. We have used a novel measure the discourse connective count measure for identifying the featured articles and have compared this approach with other two approaches, word count measure and explicit document model method.

Coherence shows the representational informational quality. Discourse analysis is concerned with measurements of cohesion and coherence. Discourse connective connects the overall text and establishes coherence between the sentences and coherence shows how well the information hangs together. It gives completeness and relevance among the text [12]. Since the quality of the featured article lies on these traditional dimensions, connectives can be used as a measure to distinguish featured and non-featured articles. In past works various algorithms have been proposed to measure cohesion. However, coherence is more difficult to quantify. We have used discourse connectives as a measure of coherence. We have collected the featured and non-featured Wikipedia articles from the index of biology articles that belongs to general biology, molecular biology and evolutionary biology. Then it is organized into balanced and unbalanced set. First, we used word count measure to classify the articles. Then, we used the discourse connective count measure. The results obtained after using connectives showed that this feature outperforms the word count measure for unbalanced corpus. On balanced corpus, the explicit language model method performed better than the word count measure and discourse connective count measure. The results are comparable with state-of-art systems.

In the next section, we describe the datasets used to develop the balanced and unbalanced corpus. In Section 3 experiments performed are explained in detail and in Section 4 results obtained are discussed. We conclude our paper in Section 5.

2 Corpus Used

Our data set consists of totally 2028 featured and non-featured biological Wikipedia articles. The Wikipedia article quality grading scheme classifies articles into different classes. In our work we have considered the Wikipedia articles belonging to Featured article, A, Good Article, B and B plus categories as featured articles and articles belonging to C, start and stub class as non-featured articles. Featured articles are well written, accurate, and stable and images are well illustrated. These articles are well organized and complete. The non-featured articles miss important content and contain irrelevant information. These featured and non-featured articles are organized as the balanced corpus and the unbalanced corpus. The balanced corpus contains the featured and non-featured articles of similar length. The unbalanced corpus contains randomly selected featured and non-featured articles without considering the document size. The balanced and unbalanced corpora contain 518 featured and 496 non-featured articles each. For the experiments on the balanced and the unbalanced corpora, we have used 811 articles for training and 203 articles for testing in the ratio 80:20. Figure 1 shows the corpus statistics.

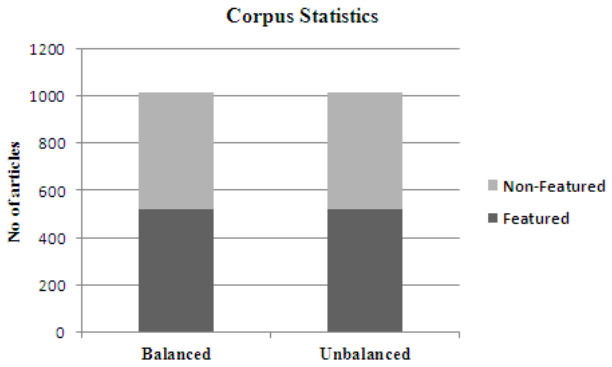

Fig. 1. Corpus statistics

3 Experiments

In our work, we automatically classify the biological Wikipedia articles as featured and non-featured articles. To identify the biological featured articles, we used word count measure, connective count measure and explicit document model method. These approaches are described in detail in the following sections.

3.1 Word Count Measure

First, we evaluated the word count measure as a baseline on the balanced and unbalanced corpora. The word count measure is a simple and effective measure of quality for Wikipedia articles [1]. It simply means the number of words in a document. Based on the number of words, each article is classified as featured and non-featured. To evaluate using the word count measure, first the documents were tokenized and the numbers of words were counted. While, [1] in his work have classified the articles with more than 2000 words as featured and those with fewer than 2000 words as random, [6] have used word count of 200 for lower quality articles and 1400 for high quality articles. [8] used a threshold of 3070 words for the unbalanced set and 955 words for the balanced set to classify featured and non-featured articles for Spanish Wikipedia articles. In this study, we found that the word count of non-featured articles on unbalanced corpus is lesser than 1800 words. Meanwhile, on the balanced corpus the word count of non-featured articles is between 800 to 3600 words and word count of featured articles is between 2700 to 15000 words. Hence, on an average the unbalanced corpus articles having more than 2000 words and balanced corpus articles having more than 3000 words are categorized as featured articles.

3.2 Discourse Connectives Count Measure

The quality of information lies in how well the provided information is useful to the users. [10], has described ten dimensions of information quality. The quality infor-

mation has to meet certain criteria like accuracy, timeliness, relevancy, etc. coherence is one among the ten dimensions that plays a significant role in defining information quality. Coherence makes a text semantically meaningful. It can be achieved through syntactic features such as deitic, anaphoric and cataphoric elements, presuppositions etc. [13]. Discourse connectives are one such syntactic feature that establishes coherence between two units in a text/discourse. They connect two discourse units that include single clauses or multiple clauses and in some cases it may include whole sentences and even multiple sentences. The units the discourse connectives connect are called as arguments. The relation can be established explicitly or implicitly [11].

Example 1
Some DNA sequences are transcribed into RNA **but** they are not translated into protein products.

In Example 1, "but" is the explicit discourse connective that connect two clauses. Here, connective "but" establishes coherence between two clauses.

Example 2
In the absence of SOX2, there is no equivalent rapidly proliferating cell population, the only surviving cells being trophoblast giant cells **and** ExEn.

In Example 2, "and" acts as a connective that connect two entities "trophoblast giant cells" and "ExEn". Here "and" is not a discourse connective because the minimal unit required for a connective to act as discourse connective is a clause that is tensed or non-tensed [11].

Example 3
Further studies found that L-PHP was expressed in pancreas. **<IMPLICIT: However>** The biological role of pancreatic L-PHP is still not clear.

In Example 3, the two sentences are related but there is no discourse marker that explicitly shows the relation. Hence the relation can be established implicitly by inserting a discourse connective "however".

In our work, we have considered explicit discourse connective count measure to identify the featured articles in biology domain. The discourse connectives are first identified from the text and the connective count is obtained. Since all the connectives in a text are not discourse connectives as in Example 2, it is necessary to develop a system to automatically identify the discourse connectives. We followed a similar method used by [4], to develop a system for automatic identification of discourse connectives. They have used a hybrid approach using linguistic rules and machine learning approach to identify the discourse relations. Likewise, we used the CRF++ tool [14], an open source implementation of Conditional Random Fields (CRFs) and linguistic rules to develop the system. PubMed abstracts were tagged with discourse connectives. Then the documents were tokenized and the features were extracted. Word, Part of Speech (PoS), Chunk, Combination of word, PoS and chunk were used as features. This corpus is trained using CRF++ tool and language models are created. Further, we also used linguistic rules to identify the connectives. The accuracy of the system is 97.3%.

Using this system the discourse connectives are identified automatically from the Wikipedia articles. The word count directly influences the discourse connective count, i.e. if the number of words in a document is higher, then number of connectives will be higher. In this dataset, the non-featured articles on the unbalanced corpus contain less than 50 connectives on an average, while the connective count of non-featured articles on the balanced corpus is between 50-270 connectives. Therefore, we performed our experiments by setting an average threshold of 50 connectives on the unbalanced corpus and 150 connectives on the balanced corpus for featured articles. Figure 2 shows the discourse connective count in the unbalanced corpus and Figure 3 shows the discourse connective count in the balanced corpus.

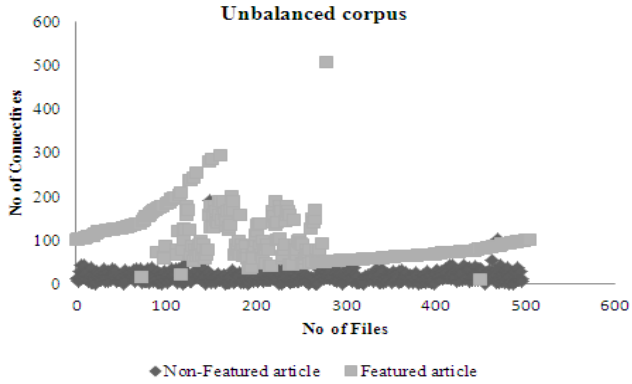

Fig. 2. Discourse connective count in the Unbalanced Corpus

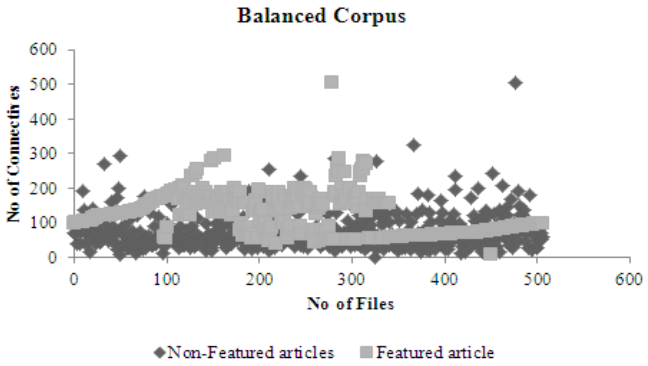

Fig. 3. Discourse connective count in the Balanced Corpus

We finally used explicit document model method to classify featured and non-featured articles for both the balanced and unbalanced corpora, which is described in detail in the next section.

3.3 Explicit Document Model

Finally, we employed explicit document model method [7] to identify featured articles from biological Wikipedia articles. [8] used this method for Spanish Wikipedia document classification. Their corpus includes the articles that belong to the snapshot of the Spanish Wikipedia from 8th July, 2013. They used two classifiers, Naive Bayes (NB) and SVMs with Term Frequency-Inverse Document Frequency (TF-IDF) and binary document models for the balanced and unbalanced corpora. We applied this method for classification of biology Wikipedia articles into featured and non-featured articles. We performed the experiments using WEKA data mining software [15] and used NB and LIBSVM classifiers. Explicit document model representation includes n-gram vectors and bag-of-words.

We extracted 3, 4 and 5 grams from plain text. Bag-of-words is a simple representation used in Natural Language Processing and Information Retrieval. It is commonly used in the methods of document classification, where the occurrence of each word is used as feature for training a classifier. In our work, n-gram vector and bag-of words are used as features with TF-IDF and binary weighting schemes.

TF-IDF is a numerical statistics. It shows the importance of a word in a document or corpus. It is the product of term frequency (TF) and inverse document frequency (IDF). It is a way to score the importance of words in a document based on how frequently they appear across multiple documents. TF is the number of times a word appears in a document normalized by dividing the total number of words in a document.

TF(t) = Number of times term t appears in a document / Total number of terms in the document. *(1)*

IDF measures how common a word is among all documents. An inverse document frequency factor diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

IDF(t) = log_e (Total number of documents / Number of documents with term t in it). *(2)*

TF-IDF is simply the product of TF(t) and IDF (t).

TF-IDF= TF(t) * IDF(t) *(3)*

Words with high TF-IDF suggest that if the word appears in a query, the document could be of interest to the user. We used the information gain method to rank the words to be used as features. For bag of words representation, both binary document model and TF-IDF weighting scheme had words connective, fossil, gel, flowers, ecology, heart, mRNA, intermediates and host ranked in first 30 positions. The results obtained are discussed in the next section.

4 Results and Discussion

First, we used the word count measure to classify the featured and non-featured articles. We obtained an F-measure of 92.47% on the unbalanced corpus. For the balanced corpus, we obtained an F-measure of 56.1%. This method showed good performance on the unbalanced corpus, while on the balanced corpus word count measure failed to give good results. Then, using the novel measure discourse connective count measure we performed our experiments on the balanced and unbalanced corpora. For the unbalanced set, we obtained an F-measure of 98.06%, when the threshold was set to 50 connectives. For the balanced set, we obtained 65.61% F-measure, when the threshold was set to 150 connectives. The discourse connective count measure outperformed the word count measure on unbalanced corpus. For the balanced corpus, the discourse connective count measure performed better than the word count measure. Finally, we performed our experiments using the explicit document model method.

The results for NB and SVMs classifiers with TF-IDF and binary document models on the balanced and unbalanced corpora are presented in the Table 1 and 2.

Table 1. F-measure in % for featured and non-featured articles classification on the balanced corpus

Features	Balanced Corpus			
	Binary		TF-IDF	
	NB	SVM	NB	SVM
Bag of words	79.6	81	79.1	79
3-grams	75.7	81	73.6	79.5
4-grams	78.6	**82.9**	81.5	**83.2**
5-grams	76.6	80.5	79.1	79.5

Table 2. F-measure in % for featured and non-featured articles classification on the unbalanced corpus

Features	Unbalanced corpus			
	Binary		TF-IDF	
	NB	SVM	NB	SVM
Bag of words	94.6	95.1	94.6	95.4
3-grams	96	**97**	92.6	96
4-grams	95.1	95.6	91.7	**97.1**
5-grams	96	96.5	94.1	95

The above results show that SVM performs slightly better than the NB method. The best result on the balanced and the unbalanced corpora is obtained using SVM classifier with 4-gram representation and TF-IDF weighting scheme. We obtained F-

measure of 83.2% on the balanced corpus and 97.1% on the unbalanced corpus. The system shows better performance on the unbalanced corpus than the balanced corpus. However, this result on unbalanced corpus shows that the discourse connective count measure performs better than explicit document model method. [6] used fact frequency based features and obtained an F-measure of 90.4% on unbalanced corpus. Fact frequency based features require direct information about the number of facts obtained by an information extraction process from a text. The facts are computed using Reverb Open Information extraction framework.

Using the fact count factual density is calculated. This feature worked well for identification of featured articles on unbalanced corpus. [8] used the word count discrimination rule and obtained an F-measure of 96% for unbalanced corpus for Spanish Wikipedia articles. [1] achieved 96.31% on an unbalanced corpus of English Wikipedia articles using the word count measure. [9] have identified featured articles from English Wikipedia domains biology and history. Their unbalanced set contained featured and non-featured articles in the ratio 1:6 respectively. They used word discrimination rule and obtained accuracy of 96% on unbalanced corpus. From the results of previous works, our works show that the discourse connective count measure outperformed word discrimination rule and factual density measure for unbalanced corpus. On the balanced corpus, the discourse connective count measure showed better performance than word count measure, but the explicit document model approach outperformed both the approaches.

The best result on balanced corpus is obtained using SVM method with F-measure of 83.2%. [8] achieved highest F-measure of 80% for 4-grams features using SVM classifiers when applied to binary representation. [9] obtained an accuracy of 96% within Biology and 92% within History when the binarized character trigram vector representation combined with an SVM was used. [6] used relational features to classify Wikipedia articles into featured/good and non-featured ones. For articles of similar lengths, they achieve an F-measure of 86.7% and 84% otherwise. The results show that the explicit document model method showed “comparable” state-of-art results on balanced corpus.

5 Conclusion

In our work we carried out various experiments to automatically classify the featured and non-featured biological Wikipedia articles. We created two corpora, balanced and unbalanced. The balanced corpus contains featured and non-featured articles of equal length, while the unbalanced corpus contains articles of dissimilar length. The word count measure, discourse connective count measure and explicit document model approach were used to identify the featured articles. The word count measure is a simple method used in the past for identifying the featured articles. Hence we used this method as a baseline to categorize the biological Wikipedia articles. Then, we used a novel approach, the discourse connective count measure. This measure outperformed other approaches used in the past work on the unbalanced corpus. Finally, we used machine learning classifier NB and SVM on the balanced and unbalanced corpo-

ra. We used bag of words and n-gram features with TF-IDF and binary weighting schemes. N-gram features include 3-gram, 4-gram and 5-gram representation. We obtained best results using SVM classifier and 4-gram feature. The results obtained on the balanced and unbalanced corpora are "comparable" to state-of-art systems.

Acknowledgement. This work is the result of the collaboration between AU-KBC Research Centre, Chennai, India and the Universitat Politècnica de València (UPV), Spain in the framework of the WIQ-EI IRSES research project (grant no. 269180) within the EC FP7 Marie Curie. The work of the second author is also in the framework of the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAPATER (PrometeoII/2014/030).

References

1. Blumenstock, J.: Size Matters: Word Count as a Measure of Quality on Wikipedia. In: 17th International Conference on World Wide Web, pp. 1095–1096.Beijing, China (2008)
2. Brucher, H., Knolmayer, G., AndreMittermayer, A.: Document Classification Methods for Organizing Explicit Knowledge. In:Third European Conference on Organizational Knowledge, Learning, and Capabilities, Athens, Greece(2002)
3. Daxenberger, J., Gurevych, I.: Automatically Classifying Edit Categories in Wikipedia Revisions. In:Conference on Empirical Methods in Natural Language Processing,pp. 537-547.Seattle, Washington, USA (2013)
4. Lalitha Devi, S.,Gopalan, S., Sreedhar. L., Rao, P.R.K., Ram, R.V.S., and Malarkodi C.S.: A Hybrid Discourse Relation Parser in CoNLL 2015. In: Nineteenth Conference on Computational Natural Language Learning: Shared Task, pp. 50-55. Beijing, China (2015)
5. Lewis D David., Ringuette, M.: A Comparison of Two Learning Algorithms for Text Categorization. In: Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93. ISRI, Las Vegas(1994)
6. Lex, E., V¨olske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Stein, B., Granitzer, M.: Measuring the Quality of Web Content Using Factual Information. In: 2nd Joint WICOW/AIRWeb Workshop on Web Quality, pp. 7-10. New York, USA (2012)
7. Li, Y.H., Jain, A.K.: Classification of Text Documents. The Computer Journal. 41(8), 537-547 (1998)
8. Lian, P., Edgaro, F., Errecalde, M.: Identifying Featured Articles in Spanish Wikipedia. In: Feierherd, G.E., Pesado, P.M., Spositto, O.M. (eds.) XX Argentine Congress of Computer Science Selected Papers. Computer Science & Technology Series, pp. 171-182. (2015)
9. Lipka, N., Stein, B.: Identifying Featured Articles in Wikipedia: Writing Style Matters. In: 19th International Conference on World Wide Web, pp. 1147–1148. Raleigh, USA (2010)
10. Miller, H.:The Multiple Dimensions of Information Quality. Information Systems Management. 13(2), 79-82 (1996)
11. Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B.: The Penn Discourse Treebank 2.0. In: Sixth International Conference on Language Resources and Evaluation (LREC'08), pp. 2961-2968. Marrakech, Morocco.(2008)
12. Rassbach, L., Pincock, T., Mingus, B.: Exploring the Feasibility of Automatically Rating Online Article Quality. In: 9th Joint Conference on Digital Libraries, (2007)

13. Rouchota, V.: Discourse Connectives: What do They Link?. UCL Working Papers in Linguistics. 8, 199-214 (1996)
14. Taku, K., CRF++, an Open Source Toolkit for CRF, http://crfpp.sourceforge.net,(2005)
15. Waikato Environment for Knowledge Analysis, http://www.iasri.res.in/ebook/win_school_aa/notes/WEKA.pdf
16. Wikipedia: WikiProject Computational Biology, en.wikipedia.org/wiki/Wikipedia:WikiProject_Computational_Biology
17. Wohner, T., Peters, R.: Assessing the Quality of Wikipedia Articles with Lifecycle Based Metrics. In:5th International Symposium on Wikis and Open Collaboration, pp. 1-10. Orlando, Florida, USA(2009)