Extracting Connected Concepts from Biomedical Texts using Fog Index

Rushdi Shams∗, Robert E. Mercer

Department of Computer Science, University of Western Ontario, London, ON N6A 5B7, Canada

Abstract

In this paper, we establish Fog Index (FI) as a text filter to locate the sentences in texts that contain connected biomedical concepts of interest. To do so, we have used 24 random papers each containing any of the four pairs of connected concepts. For each pair, we categorize sentences based on whether they contain both, any or none of the concepts. We then use FI to measure the difficulty of the sentences of each category and find that sentences containing both of the concepts have low readability. We rank sentences of a text according to their FI and select 30 percent of the most difficult sentences. We use an association matrix to track the most frequent pairs of concepts in them. This matrix reports that the first filter produces some pairs that hold almost no connections. To remove these unwanted pairs, we use the Equally Weighted Harmonic Mean of their Positive Predictive Value (PPV) and Sensitivity as a second filter. Experimental results demonstrate the effectiveness of our method.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of 12th Conference of the Pacific Association for Computational Linguistics (PACLING 2011)

Keywords: Computational Linguistics, Text mining, Connected Concepts, Fog Index, Text Readability.

1. Introduction

In recent years, extraction of connected biomedical concepts (i.e., disease, treatment, and genes) from texts has drawn the attention of scientists interested in finding functional similarity (i.e., identification of genes involved in human diseases) [1]. Although benchmark research has reported successful methods to extract biomedical concepts [2, 3], they have rarely followed simple procedures. For example, Perez-Iratxeta et al. [4] could not relate diseases with gene functions from biomedical texts forthwith – they needed to apply a twofold intermediary process of connecting disease with chemical components and chemical components with gene functions. The key reason for not applying simple methods to extract connected concepts from biomedical texts is manifold. While some researchers concentrated on the number of co-occurrence of concepts in the abstract of a paper [4, 5], others preferred to comb through either the full text [6] or pre-specified segments (i.e., Introduction, Methods, Results, and Discussion) [7]. Moreover, the connections can be either very general (i.e., biochemical connections) or very specific (i.e., regulatory connections). Therefore, the demand of developing simple methods to identify and extract biomedical concepts from a
scientific literature that maintain connections, general or specific, with one another is not met till to date. This situation suggests to use simple yet improved computational method to identify and extract important, explicit, and implicit connections from biomedical texts.

As text is highly structured by syntax and semantics of natural language, it is believed that any relation extraction method should involve these two features. However, several reports asserted their complexity [8, 9]. Apart from this, Sherman [10] proposed that scientific literature is subject to statistical analysis and zeroed in on the importance of average sentence length. Gunning [11] practically demonstrated this important measure along with the number of complex words (i.e., words with three or more syllables) to assess the readability of text known as the Fog Index (hereinafter, FI). It is now considered a yardstick for readability assessment of books, scientific literature and newspapers, and even to detect online chatting bots [12]. Besides, an interesting ascertainment that texts become relatively difficult to read when contain ideas and relations [18] can be motivational for using FI to find hidden relations in papers.

In this paper, we report a simple novel statistical method to extract connected biomedical concepts from biomedical texts using FI. We statistically established FI as a text filter, experimenting on 24 random papers that describe four pairs of concepts: Ischemia-Glutamate, Ataxia-Dehydrogenase, Hypogonadism-Gonadotropin, and Epilepsy-GABA. Besides FI, our method also uses the equally weighted harmonic mean of the connections’ Positive Predictive Value (hereinafter, PPV) and sensitivity as a second filter. While the prior concentrates on the important part of the text where the connections are stated, the latter assesses their representativeness. We selected the sentences of a paper that are difficult to read and ranked the most frequent connected concepts present in them. With careful observations, we noticed that the first filter produces some noisy pairs of concepts that hold almost no connection. To exclude them, we re-ranked every pair of concepts based on the equally weighted harmonic mean of their PPV and sensitivity, and filtered them.

In the remainder of this paper, we describe related work, illustrate the methodology, report and discuss experimental results, and draw conclusions.

2. Related Work

New research trends in the biomedical field include the discovery of hidden connections in texts to form new hypotheses that can be explored further by conventional experimentation [6]. A series of investigations by Swanson [13, 14] showed that these hidden connections can lead us to new discoveries. He reported that fish oil leads to change in blood viscosity and red blood cell rigidity that helps prevent Raynauds syndrome [13]. Later, investigative reports started to discover suggestions for clinical therapies and basic physiological linkages from bibliographically isolated texts. However, the working principles of Swanson’s empirical research include the computational burden of full-text syntactic analysis and involve large literature databases like MEDLINE. Our work, though it does not generate hypotheses, can be a good means of finding implicit connections in texts using fewer computations (as it filters out texts according to their readability and does not operate syntactically) without involving literature repositories.

A handful of research work in semantic relation classification or extraction from bioscience texts depends on the proper identification of connections. Rosario and Hearst [15] concentrated on discovering connections between treatment and disease. They reported 79.6 percent accuracy in blindly identifying concepts that fall into either of the categories and are somehow connected with one another. They used a MEDLINE-based neural network that addresses it to be intriguing yet complicated. A similar machine learning technique was applied by Franza and Inkpen [8] to extract disease-treatment connections from texts. Their reported accuracy surpassed the results of Rosario and Hearst although their interest was limited to MEDLINE 2001 titles and abstracts. Their paper, like many other prominent work [16, 17], has a significant use of PPV and sensitivity to evaluate the mining technique. Contrary, we used these measures to evaluate the representativeness of the connected concepts.

Perez-Iratxeta et al. [4] proposed a massive framework to prioritize disease associated genes. Instead of looking into literature, they combined several isolated pieces of biomedical repositories like Medical Subject Heading (MeSH), Gene Ontology (GO), RefSeq database, and MEDLINE. They used both databases and
ontology that have lack in communication with one another and thus experienced tedious and complex scoring methods and formidable number of intermediate stages. In our work, we decided to stick with texts only to remain simple yet capable of producing improved results.

Robert Gunning [11] first introduced Fog Index (FI) to measure semantic difficulties using average sentence length and polysyllabic words in his 1952 book *The Technique of Clear Writing*. We were motivated to apply FI as our text filter when we came across the results of an experiment carried out by Duffy and Kabance [18]. They converted a passage with no more than two phrases into primer prose and applied FI to test its readability. They found the score well below the readability index (i.e., it was excessively easy to read). Their investigation on this phenomenon suggested that easy articles (in this case the primer prose) obscure the relationships and ideas as they emphasize each of them equally. In other words, difficult articles possess relationships and ideas and emphasize them in particular that yields low readability. We believe that if biomedical texts display a similar attribute, then FI can be an appropriate measure to filter texts that bear associations of scientific interest.

### 3. Methodology

The work of Perez-Iratxeta *et al.* [4] lists pairs of connected concepts like disease-chemical components, chemical component-genes, and disease-genes. Among them, we considered four disease-chemical component pairs, namely *Ischemia-Glutamate, Ataxia-Dehydrogenase, Hypogonadism-Gonadotropin,* and *Epilepsy-GABA*. We collected 24 scientific papers (six for each pair of concepts) at random from several biomedical literature repositories. To work with the text only, we removed the title, affiliations, keywords, footnotes, figures, tables, acknowledgements, and references from the paper.

We considered each pair of concepts and a paper related to them. We classified its sentences into three sets: sentences containing both of the concepts, none of the concepts and any of the concepts. For example, the sentence *Glutamate, which is potentially excitotoxic to brain neurons, is released excessively during ischemia*, will be put into the set of sentences containing both of the concepts *Ischemia* and *Glutamate*, as *Ischemia* and *Glutamate* are both present. Then, we applied Gunning’s formula for FI (Eq. 1) to score the sentences of every set. According to this formula, the lower the score of a sentence, the easier it is to read.

$$FI = 0.4 \times \left( \frac{Words}{Sentences} \right) + 100 \times \left( \frac{Complex \; Words}{Words} \right)$$

It can be noted that according to Gunning, words that are polysyllabic (i.e., contain three or more syllables) are called Complex Words. Also, as we applied FI on every sentence, the value of Sentences is always 1. We normalized this score by the paper’s average number of syllables per word because readability score of long and short sentences varies due to the total number of syllables [11]. Eq. 2 provides the normalized FI ($FI'$) of the sentences in every set.

$$FI' = \frac{FI}{Average \; Number \; of \; Syllables \; per \; Word}$$

The $FI'$ calculated for the three sets of sentences for the 24 papers in groups related to the four pairs of connected concepts are shown in Table 1. Table 1 shows that, for every pair of connected concepts, while the set of sentences containing any or both of the concepts displays either **Low** or **Medium** readability, the set of sentences containing none of them consistently has **High** readability (i.e., **Low** $FI'$). This observation leads us to decide that those sentences that are easier to read contain fewer connected concepts and therefore, we should look into low-readable sentences for hidden connections.

Now that we have FI as a functioning text filter, we need to define a means to determine the number of low-readable sentences to be considered for concept extraction. To do this, we ranked every sentence in a paper based on their FI score and sorted them in descending order (i.e., the most difficult sentences are at the top of the list). From this sorted list, in five chunks, we selected the top 50 percent, 40 percent, 30 percent, 20 percent, and 10 percent of the sentences. For every chunk, we tagged these sentences with Genia Biomedical POS tagger [19], identified the nouns in them and used an association matrix to record
the frequency of their co-occurrences (i.e., number of occurrences of one noun with the other). For instance, the connected concepts in the sentence Glutamate, which is potentially excito-toxic to brain neurons, is released excessively during ischemia are glutamate-brain, glutamate-neurons, glutamate-ischemia, brain-neurons, brain-ischemia, and neurons-ischemia. From the output of the association matrix, we kept the 20 most frequent connected concepts for our experiment. As we observed, some chunk \( i \) contains new connections that are absent in chunk \( i-1 \) and vice versa. To find a threshold, we tracked the number of connections revealed and missed by every chunk \( i \) with respect to its previous chunk \( i-1 \). From Figure 1, we see that for the first chunk (50 percent of the sentences), all of the 20 most frequent connections are new. The number of new connections remains steady up to the third chunk (30 percent of the sentences) but then reaches the extremes in the fourth and fifth. The results in Figure 1 are shown for six papers related to Ischemia and Glutamate. Similar experiments with the other connected concepts showed that if we take less than 30 percent of the ranked sentences, the number of new concepts reaches the extremes.

We recorded a similar behavior for the number of connections dropped by every chunk. Figure 2 shows that as we start with it, the first chunk (10 percent of the sentences) does not miss any connection but the number of dropped connections suddenly starts to reach the extremes in the fourth and fifth. Again, the results in Figure 2 are produced by six papers on Ischemia and Glutamate. Similar experiments carried out with the other connected concepts showed that if we take less than 30 percent of the ranked sentences, the number of dropped connections reach the extremes.

These two observations indicate that the degree of concepts connected with each other is conserved if we take 30 percent of the low-readable sentences. Similar results are obtained for the three other pairs of concepts.

Provided this threshold, Table 2 shows the 20 most frequent connected concepts found in a paper on Ischemia-Glutamate where the connections are ranked according to their frequency. For each pair shown in Table 2, we extracted those sentences from the paper that contain both of the concepts. These sentences are fed to the Unified Medical Language System (UMLS) semantic relation network [20] to find out if the concepts have any semantic connection. Surprisingly, we found that among the 20 connected concepts, only nine have textual semantic connections (Levels-Glutamate, Ischemia-Glutamate, Levels-Increase, Increase-Glutamate, 10min-Ischemia, Glutamate-Experiment, Glutamate-Neurons, Glutamate-CA4, and Ischemia-5min).

| Category | Ischemia-Glutamate | Readability | Ataxia-Dehydrogenase | Readability | Epilepsy-GABA | Readability | Dehydrogenase-Gonadotropin | Readability |
|----------|---------------------|-------------|-----------------------|-------------|---------------|-------------|-----------------------------|-------------|
| \( P_{\text{none}} \) | 5.99                | High        | 5.77                  | High        | 6.50          | High        | 5.58                        | High        |
| \( P_{\text{both}} \) | 8.26                | Low         | 7.23                  | Medium      | 7.24          | Medium      | 10.29                       | Low         |
| \( P_{\text{any}} \) | 6.83                | Medium      | 7.33                  | Low         | 7.58          | Low         | 7.62                        | Medium      |

Table 1. Normalized FI for three sets of sentences from 24 papers
So, FI, as a text filter, brings in some text that contains most frequent connected concepts, some of which lack representativeness (i.e., they do not hold any connection). It urged us to provide a means to filter out these noisy pairs of concepts. As we collected texts at random, we observed that it is possible for the pairs to never co-occur in a sentence which indicates that our data set is imbalanced. So, we used the equally weighted harmonic mean of the PPV and sensitivity of the pairs of concepts provided by FI to evaluate their representativeness as it is a great evaluation metric for imbalanced dataset [8].

PPV\(^1\) is the percentage of correctly predicted connections and sensitivity represents the percentage of connections identified as relevant by our method. To measure the PPV and sensitivity of every pair of concepts, we first considered the set of sentences filtered by FI and counted the number. This is the total number of results returned by our system \((R)\) that comprises the number of True Positives \((TP)\) and False Positives \((FP)\). Then, we take a pair depicted in Table 2, searched the paper, and developed a second set of sentences that contain both of its concepts. The number of sentences in this set is the number of results that should have been returned by our system \((S)\) and comprises the number of True Positives \((TP)\) and False Negatives \((FN)\). Finally, we counted the number of sentences that are present in both sets – which is the number of TPs by our system. Afterwards, \(FP\) is obtained by subtracting \(TP\) from \(R\) and \(FN\) is obtained by subtracting \(TP\) from \(S\). So, the PPV of every pair of connected concepts is \(\frac{TP}{TP+FP}\) and the sensitivity of every pair of connected concepts is \(\frac{TP}{TP+FN}\). We then applied the formula in Eq. 3 to determine the equally weighted harmonic mean for the given pair of concepts. In this way, we measured this mean for every pair of concepts in Table 2.

\[
\text{Harmonic Mean of PPV and Sensitivity} = 2 \times \left( \frac{\text{PPV} \times \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}} \right)
\] (3)

We re-ranked the pairs of concepts in Table 2 according to their individual Harmonic Mean of PPV and Sensitivity, and considered the first 10 pairs of concepts. These 10 pairs of connected concepts are said to be the representative connected concepts of the paper. Similar procedure is followed to evaluate the representativeness of the pairs of concepts for the rest of the connected concepts: Ataxia-Dehydrogenase, Hypogonadism-Gonadotropin, and Epilepsy-GABA.

We also measured the accuracy of every connected pair by using Eq. 4 –

\[
\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}
\] (4)

where \(TN\) is the number of True Negatives and can be found by subtracting \(TP+FP+FN\) from the total number of sentences in a text, and ranked them accordingly. However, we found that in the case of accuracy, the ranked connections are not well distinguished.

\(^1\)Similar to F-Score, Precision, and Recall but their use in Information Retrieval and Classification is different. The terms PPV and Sensitivity have been used to avoid confusion with the evaluation terminology.
4. Results and Discussions

In this section, the re-ranked connected concepts according to their individual Harmonic Mean of PPV and Sensitivity are reported. The results show that most of the biomedical connected concepts extracted by the proposed method are reported to be semantically connected by UMLS. This indicates that the use of the Harmonic Mean substantially decreased the number of noisy relations extracted by the FI.

Table 3 lists the 10 connected concepts for a paper on Ischemia and Glutamate among which seven pairs of concepts are reported as semantically connected by UMLS. It can be seen that the pairs of concepts in Table 2 that hold merely no relation between them are decreased significantly. However, the three pairs of concepts, not having any semantic connection, could not be filtered because of their high frequency of co-occurrence in the text.

Table 4 shows the 10 connected concepts for a paper on Ataxia and Dehydrogenase, seven of which are semantically connected in the UMLS semantic relation network. Our observation of this domain reveals that PDHC (Pyruvate Dehydrogenase Complex) is manifested in Ataxia patients, especially those suffering from Friedreich’s Ataxia. So, the relations among Friedreich, Ataxia, and PDHC are vividly represented in the list. Pyruvate-Ataxia is extracted as connected concepts because in the text, the elaboration of PDHC co-occurred with Ataxia many times.

Table 5 lists the 10 connected concepts for a paper on Hypogonadism and Gonadotropin. According to the UMLS semantic relation network, eight of these pairs are semantically connected. Steroids have significant effects on diseases like Hypogonadism, where the release of testosterone plays an important role. Therefore, the connection between AAS (a shorthand for Anabolic Androgenic Steroid) that induces Hypogonadism and Testosterone is present in the list. The pairs of concepts not semantically connected are still reported by our method for their extremely high co-occurrences in the text.

Table 6 displays the connected concepts present in a paper on Epilepsy and GABA. Epilepsy is a neuronal disease that causes inhibition, significantly affects neuronal structure like the Hippocampus, and is

| Rank | Connected Concepts | Harmonic Mean | Semantic Connection |
|------|-------------------|---------------|---------------------|
| 1    | Ischemia-Glutamate | 51.85         | Yes                 |
| 2    | Levels-Ischemia   | 43.47         | No                  |
| 3    | Levels-Glutamate  | 41.66         | Yes                 |
| 4    | Glutamate-Neurons | 39.02         | Yes                 |
| 5    | 10min-Ischemia    | 37.50         | Yes                 |
| 6    | Glutamate-CA4     | 35.89         | Yes                 |
| 7    | Increase-Glutamate| 32.55         | Yes                 |
| 8    | 10min-Glutamate   | 31.81         | No                  |
| 9    | Ischemia-5min     | 31.57         | Yes                 |
| 9    | Glutamate-5min    | 31.57         | No                  |

Table 3. Extracted Connected concepts for a paper on Ischemia and Glutamate

| Rank | Connected Concepts | Harmonic Mean | Semantic Connection |
|------|-------------------|---------------|---------------------|
| 1    | Friedreich-Ataxia | 59.25         | Yes                 |
| 2    | PDHC-Ataxia       | 56.00         | Yes                 |
| 3    | Activity-Friedreich| 43.47         | Yes                 |
| 3    | Patients-Ataxia   | 43.47         | Yes                 |
| 3    | Activity-Ataxia   | 43.47         | Yes                 |
| 3    | PDHC-Friedreich   | 43.47         | Yes                 |
| 4    | Preparations-Ataxia| 40.00        | No                  |
| 4    | Preparations-Friedreich| 40.00     | No                  |
| 5    | Pyruvate-Ataxia   | 38.09         | No                  |
| 6    | Patients-Friedreich| 36.36         | Yes                 |

Table 4. Extracted Connected concepts for a paper on Ataxia and Dehydrogenase

| Rank | Connected Concepts | Harmonic Mean | Semantic Connection |
|------|-------------------|---------------|---------------------|
| 1    | Inhibition-GABA   | 26.08         | Yes                 |
| 2    | GABA-Synapse      | 20.25         | Yes                 |
| 3    | Neurons-Synapse   | 14.70         | Yes                 |
| 4    | Inhibition-Hippocampus| 12.30        | Yes                 |
| 5    | Synapse-Change    | 9.37          | No                  |
| 6    | Neurons-GABA      | 8.00          | Yes                 |
| 7    | Properties-GABA   | 6.45          | Yes                 |
| 7    | GABA-Change       | 6.45          | No                  |
| 8    | GABA-Number       | 6.34          | No                  |
| 9    | Cl-Gradient       | 3.33          | Yes                 |

Table 5. Extracted Connected concepts for a paper on Hypogonadism and Gonadotropin

| Rank | Connected Concepts | Harmonic Mean | Semantic Connection |
|------|-------------------|---------------|---------------------|
| 1    | Treatment-Therapy | 12.90         | Yes                 |
| 2    | Use-AAS           | 21.62         | No                  |
| 3    | AAS-Testosterone  | 18.46         | Yes                 |
| 4    | Gonadotropin-Treatment| 18.18       | Yes                 |
| 5    | Testosterone-Treatment| 14.92       | Yes                 |
| 6    | Levels-Testosterone| 14.49         | Yes                 |
| 7    | AAS-Conditions   | 12.90         | Yes                 |
| 7    | Treatment-HCG    | 12.90         | Yes                 |
| 8    | Replacement-Therapy| 12.90        | No                  |

Table 6. Extracted Connected concepts for a paper on Epilepsy and GABA
caused by low levels of GABA. In the list, we find seven concepts that are semantically related according to UMLS. It should also be noted that the Harmonic Means of the pairs of concepts are significantly lower than that found from other papers, meaning either the paper is short or the concepts co-occurred infrequently.

5. Conclusions

In this paper, we report on the extraction of connected concepts from biomedical texts by assessing text readability. The readability of text is determined by a metric called Fog Index (FI). We curated 24 random papers by using four pairs of connected concepts as keywords and applied FI on them. Experimental results showed that sentences display low readability if they contain connected concepts. We selected 30 percent of the most difficult-to-read sentences, and used an association matrix to track the most frequent pairs of concepts in them. To remove those pairs of concepts that have a rather weak connection, we used the equally weighted harmonic mean of their positive predictive value and sensitivity as a second ranking filter. The results are supported by finding almost all of the extracted concepts semantically connected by the UMLS semantic relation network.

Acknowledgements

This work was partially funded through a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to Robert E. Mercer.

References

[1] Yuan F, Wang R, Guan M, He G. A Novel Computational Method for Predicting Disease Genes Based on Functional Similarity. Lecture Notes in Computer Science 2010; 6216:42-51.
[2] Zhao S. Named Entity Recognition in Biomedical Texts using an HMM Model. International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications 2004, Geneva, Switzerland; 84-87.
[3] Kazama J, Makino T, Ohta Y, Tsujii J. Tuning Support Vector Machines for Biomedical Named Entity Recognition. ACL-02 Workshop on Natural Language Processing in the Biomedical Domain 2002, Philadelphia, USA; 3:1-8.
[4] Perez-Irurzua C, Bork P, Andrade M. Literature and Genome Data Mining for Prioritizing Disease-Associated Genes. Discovering Biomolecular Mechanisms with Computational Biology (Molecular Biology Intelligence Unit II) 2006; 74-81.
[5] Dai H, Chang Y, Tsai RT, Hsu W. New Challenges for Biological Text-Mining in the Next Decade. Journal of Computer Science and Technology 2010; 25:1:169-179.
[6] Lindsay R, Gordon M. Literature-Based Discovery by Lexical Statistics. Journal of the American Society for Information Science 1999; 50:7:574-587.
[7] Agarwal S, Yu H. Automatically Classifying Sentences in Full-text Biomedical Articles into Introduction, Methods, Results and Discussion. Bioinformatics 2009; 25:23:3174-3180.
[8] Frunza O, Inken D. Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences. 2010 Workshop on Biomedical Natural Language Processing, ACL 2010 2010, Uppsala, Sweden; 91-98.
[9] Leroy G, Chen H, Martinez JD. A Shallow Parser based on Closed-class Words to Capture Relations in Biomedical Text. Journal of Biomedical Informatics 2003; 36:145-158.
[10] Sherman AL. Analytics of literature: A Manual for the Objective Study of English Prose and Poetry. Boston: Ginn and Co; 1893.
[11] Gunning R. Fog Index after Twenty Years. Journal of Business Communication 1969; 6:3:3-13.
[12] Goh OS, Fung CC, Depickere A, Wong KW. Using Gunning-Fog Index to Assess Instant Message Readability from ECAs. 3rd International Conference on Natural Computation (ICNC 2007) 2007, Hainan, China; 480-486.
[13] Swanson DR. Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge. Perspectives in Biology and Medicine 1986; 30:7-18.
[14] Swanson DR. A Second Example of Mutually Isolated Medical Literatures related by Implicit, Unnoticed Connections. Journal of the American Society for Information Science 1989; 40:432-435.
[15] Rosario B, Hearst MA. Classifying Semantic Relations in Biocreative II Gene Mention Recognition, Genome Biology 2008; 9:2.
[16] Unified Medical Language System (UMLS). UMLS Terminology Services (UTS); Available at: https://uts.nlm.nih.gov/home.html [August 3, 2011].