Appears in Proceedings of the Critical Assessment of Information Extraction systems in Biology (BioCreative2004)

FiGO: Finding GO Terms in Unstructured Text

Francisco M. Couto
fcouto@di.fc.ul.pt
Phone: +351-918263676
Fax: +351-217500084

Mário J. Silva
mjs@di.fc.ul.pt
Phone: +351-217500128
Fax: +351-217500084

Departamento de Informática,
Faculdade de Ciências,
Universidade de Lisboa, Portugal

Pedro Coutinho
pedro@afmb.cnrs-mrs.fr
Phone: +33-491164515
Fax: +33-491164536
Architecture et Fonction des
Macromolécules Biologiques,
CNRS, Marseille, France

Abstract

The identification of biological entities is an important subject for biological text mining systems. More than identifying the gene and the protein names it is also important to identify their properties in the text. In this document, we introduce a novel method for identifying GO terms in unstructured text, involving the information content of their names. We have integrated this method with a functional semantic similarity measure, to test it on BioCreative tasks 2.1 and 2.2 to identify GO annotations and their evidences in literature. The results show that our approach has a large potential for this kind of application.

1 Introduction

We have developed a method to identify GO terms in unstructured text and named it FiGO (Finding GO). FiGO uses the information content of each word present in the terms' name. The information content is related to the number of times the word appears in all the names. Therefore, the information content of a word measures its importance to identify a GO term in the text. For instance, consider the GO term 'punt binding'. If the term name's word 'binding' occurs alone in the text, the probability of the term being referred is very low, because 'binding' is used in many other terms. On the other hand, if the word 'punt' occurs in the text, then we have a strong evidence that the term is referred in the text, because this word is not part of any other term's name.

2 Method

FiGO starts by identifying the set of all words present in the terms' names. FiGO removes from this set all the stop words, such as 'in' or 'on'.

Then, FiGO calculates the information content of each word. This value is inversely proportional to the number of occurrences, i.e., a word occurring very often has low information content. FiGO computes the information content (IC) of a word w using the following equation:

$$IC(w) = -log(\frac{\#w}{\#max}),$$

where $\#w$ is the number of GO terms whose name contains w, and $\#max$ is the maximum number of GO terms whose name contains a common word. This equation is based on Jiang and Conrath's definition of information content [3].

Since each term's name is composed by a set of words, we can define its information content as the sum of the information content of its words. Thus, FiGO computes the information content of a term's name n using the following equation:

$$IC(n) = \sum_{i=0}^{k}(IC(w_i)),$$

where the name n is composed by the words $w_0, \dots, w_k$.

A GO term may have multiple names. FiGO defines the information content of a term as the maximum IC of all its names. Therefore, FiGO computes the information content of a term t using the following equation:

$$IC(t) = max\{IC(n_i) : 0 \leq i \leq j\},$$

where $n_0, \dots, n_j$ represent all the names for t.

Given a piece of text, we can define the local information content of each term as the sum of the information content of its words that are also present in the text. Therefore, FiGO computes the local information content (LIC) of a term t in a piece of text p using the

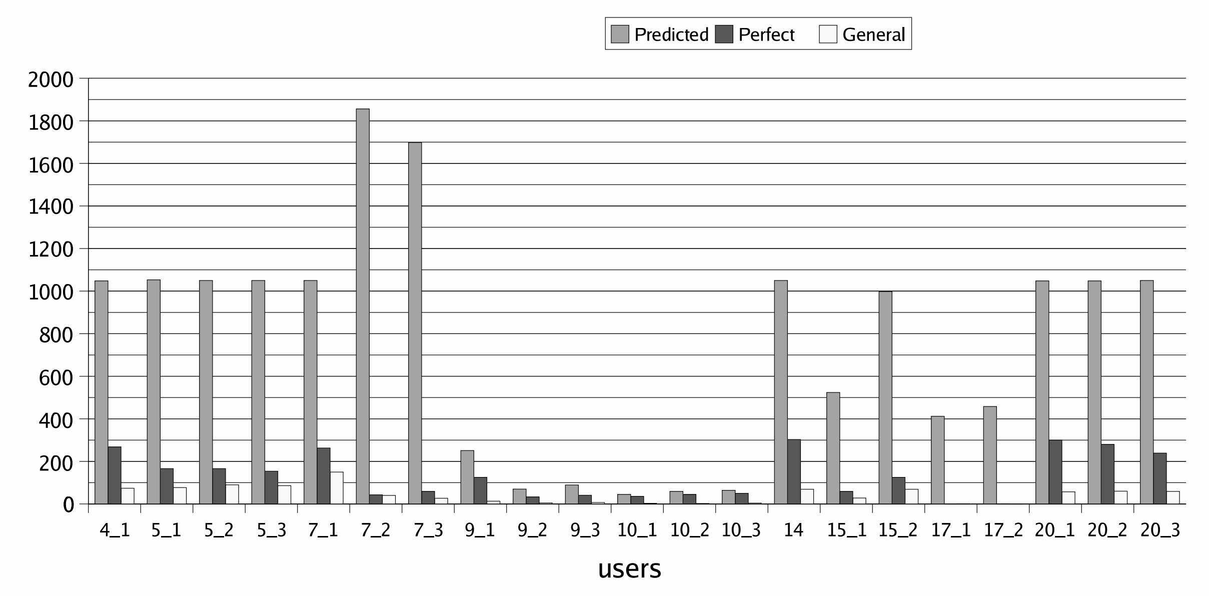

Figure 1: Results of all the submissions to BioCreative task 2.1. 20_1, 20_2 and 20_3 represent our three submissions. The figure shows, for each submission, the number of predictions made, and how many of them were evaluated as perfect and as general.

following equation:

$$LIC(t,p) = \sum_{i=0}^{l}(IC(w_i)),$$

where $w_0,\ldots,w_l$ represent all the words present in p and also in the name of t.

FiGO identifies a term in a piece of text, when its local information content is sufficiently close to its information content. Thus, given a piece of text p and a term t, FiGO only identifies the terms that satisfy the following equation:

$$LIC(t,p) \geq \alpha \times IC(t),$$

where $\alpha \in]0,1]$ representing how close LIC should be from IC to decide that t is referred in p. For instance, when $\alpha=1$, FiGO only identifies terms whose complete name appears in the text. Thus, the parameter α controls the recall and precision of FiGO. To increase precision we have to increase α, and to increase recall we have to decrease α.

3 BioCreative Application

We have applied FiGO to BioCreative tasks 2.1 and 2.2. Task 2.1 consisted in identifying the text from a given article that provides evidence for a given GO annotation. Task 2.2 consisted in identifying the GO annotations with an evidence text mentioned in a given article.

Before applying FiGO, we parsed the SGML file given for each document, and we structured the text in sentences. For FiGO each sentence represented a piece of text from where it identified GO terms.

In task 2.1, the GO term to identify was given, so we returned the sentence where FiGO identified the term. In case of having multiple sentences, we selected the one that contained at least one of the protein's names, and where the local information content was larger. In case of not having any sentence, we returned a sentence where FiGO identified the most similar term. To calculate the similarity between terms, we used FuSSiMeG [2]. In this task, we executed FiGO three times with the α parameter assigned to 0.3, 0.7 and 0.9, resulting in three different submissions.

In task 2.2, we selected the sentences where FiGO identified GO terms, which contained at least one of the protein's names. From these sentences, we selected those referring the most infrequently annotated terms. This selected the most meaningful annotations discarding common GO terms, such as 'protein', 'binding'. In this task, FiGO was executed three times with the α pa-

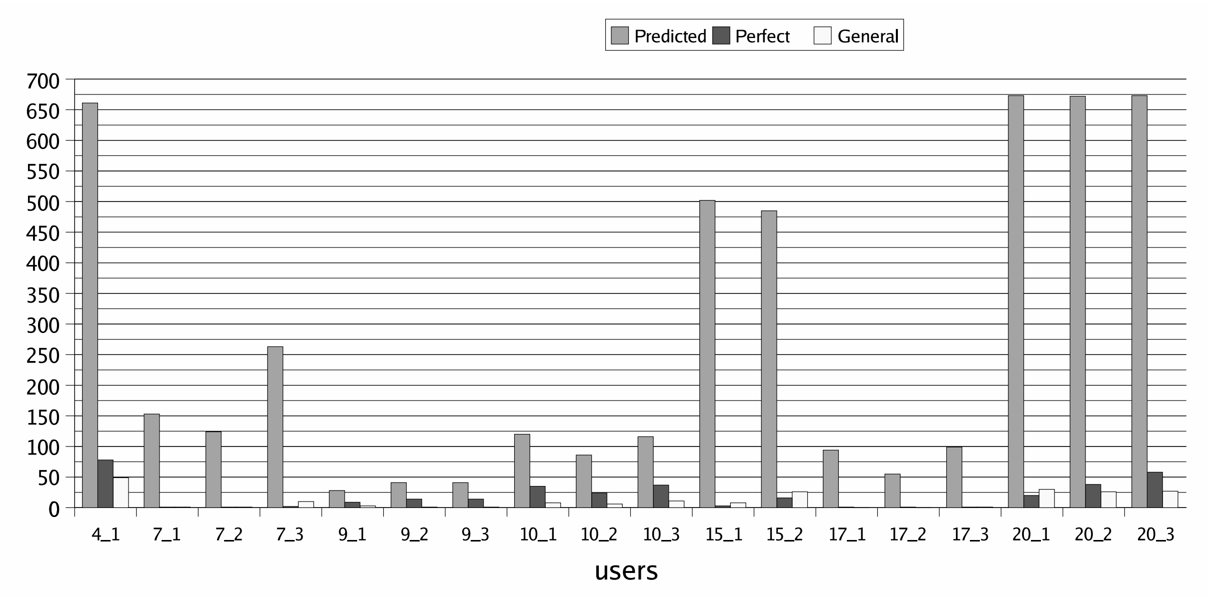

Figure 2: Results of all the submissions to BioCreative task 2.2. 20_1, 20_2 and 20_3 represents our three submissions. The figure shows, for each submission, the number of predictions made, and how many of them were evaluated as perfect and as general.

rameter assigned to 0.5, 0.7 and 0.9, resulting in three different submissions.

4 Results

Figure 1 shows the results of the 9 participants in BioCreative task 2.1. We were the participant number 20, thus 20_1, 20_2 and 20_3 represent our submissions with α equal to 0.3, 0.7 and 0.9, respectively. The "predicted" column represents the number of predictions made by each submission. The "perfect" column represents how many predictions were correct in terms of the GO term and in terms of the protein. The "general" column represents how many predictions were correct in terms of the protein, but predicting a generalization (a parent in GO) of the expected GO term. The results of the seven participants in task 2.2 are shown in figure 2. In this figure 20_1, 20_2 and 20_3 represent our submissions with α assigned to 0.5, 0.7 and 0.9, respectively.

The classification of each prediction was ”high” when the prediction was correct, ”generally” when the prediction is just a generalization of the correct prediction and ”low” when the prediction was incorrect. Figure 3 shows the evaluation of our approach in terms of the GO term for the different values of α. Figure 4 shows the evaluation of our approach in terms of the protein when the prediction was correct in terms of the GO term. Therefore, the high line in this figure represents our perfect predictions, since they are correct in terms of the GO term and protein. Figures 5 and 6 show analogous evaluations for task 2.2.

5 Discussion

In both tasks, our approach was very close to the largest number of perfect predictions achieved. However, in terms of accuracy, we were far from the best results, because we decided to submit the expected number of predictions even when a part of our predictions had a low confidence level. If we filtered the predictions according to their confidence level, we would certainly achieve a better accuracy without loosing a significant number of correct predictions.

We achieved a better performance in task 2.1 than in task 2.2, as the majority of the participants. This derives from the greater difficulty of task 2.2. The goal of task 2.1 was to identify where the evidence was, while in task 2.2 we had also to identify which evidences were mentioned.

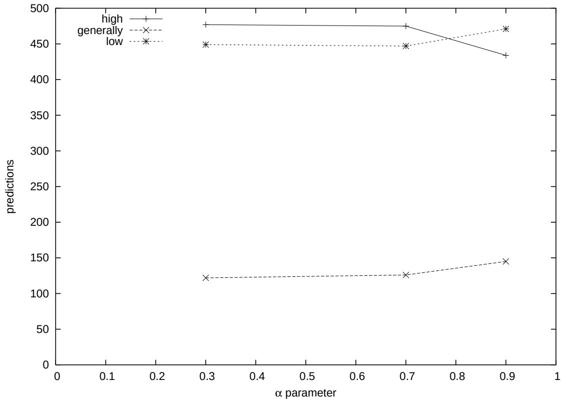

Figure 3: GO evaluation of our task 2.1 submissions. This figure shows the number of our predictions that provide a high, general and low evidence of the GO term for the values of α used.

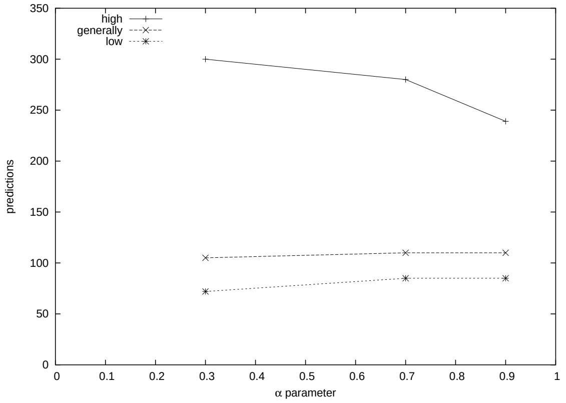

Figure 4: Protein evaluation of our task 2.1 submissions. For our predictions with a high evidence of the GO term, this figure shows how many of them provide a high, general and low evidence of the protein for the values of α used.

The manipulation of the α parameter had a different impact on the two tasks. In task 2.1, we obtained better results using a smaller α value, because there were a large number of terms not explicitly mentioned in the text. Some sentences were correctly selected when only less than 70% of the term's name was mentioned. Figure 4 shows that the protein evidences decrease when we increase α. Therefore, we achieved a better protein identification for smaller values of α. This was expected, because for smaller values of α FiGO provides a larger number of sentences where we could found the protein's name. On the other hand, in task 2.2 the increase of α implied a better performance of our approach. For smaller values of α, FiGO identified more terms that were not relevant in the given context. Thus, the selection of terms with a larger piece of its name in a sentence turned up to be an effective approach to identify the correct terms in some cases.

Figure 4 shows that in task 2.1 more than 150 predictions were not considered perfect just because they were incorrect in terms of the protein. We could increase the number of perfect predictions in more than 50% if we used a more effective protein identification method. On the other hand, in task 2.2 the protein identification was not so significant for the overall results, since there were a lower percentage of predictions not considered perfect because of the protein evaluation.

6 Conclusions

This document introduces FiGO, a novel approach for identifying GO terms in unstructured text involving the information content of their names. We integrated FiGO with a functional semantic similarity measure, to evaluate FiGO on the BioCreative tasks 2.1 and 2.2. Unlike other approaches that use domain knowledge, FiGO is fully automated, i.e. it does not rely on information introduced by human experts. Its domain knowledge comes from publicly available information, and not from specific training data. Thus, using FiGO represents little or no extra human intervention.

Despite of the good performance of our approach when compared to the performances obtained by other participants in BioCreative, it is still very far from being a perfect solution. To identify the protein evidences we applied a naïve method based on pattern matching. A more effective method would likely improve our results. Another limitation of our approach was the application of FiGO at the sentence level. If a term occurred in more than one sentence, we did not increase our confidence in the correctness of its identification. Frequently, the name of the protein and the GO term are not in the same sentence, but most of the times in the same paragraph. One possible solution is to make predictions based on the number of sentences that separate the protein from the term in the same paragraph. To improve performance on task 2.2, we need some do-

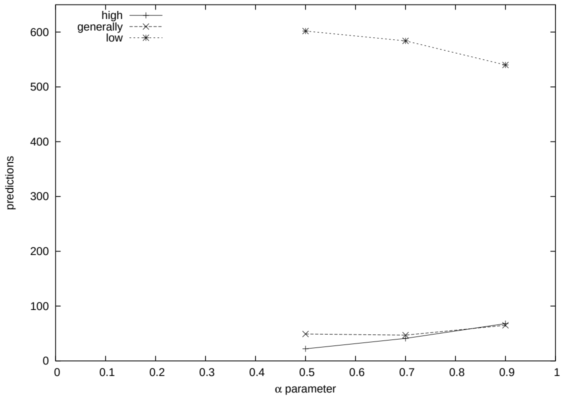

Figure 5: GO evaluation of our task 2.2 submissions. This figure shows the number of our predictions that provide a high, general and low evidence of the correct GO term for the values of α used.

main knowledge about the proteins and the articles to guide the filtering of terms out of context. The required domain knowledge could be obtained from various web resources could be an effective approach [1].

References

[1] F. Couto, B. Martins, , M. Silva, and P. Coutinho. Classifying biomedical articles using web resources. In *19th ACM Symposium on Applied Computing (SAC), Bioinformatics Track*. SAC, 2004.

[2] F. Couto, M. Silva, and P. Coutinho. Implementation of a functional semantic similarity measure between gene-products. DI/FCUL TR 03–29, Department of Informatics, University of Lisbon, November 2003.

[3] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *10th International Conference on Research on Computational Linguistics (ROCLING X)*, Taiwan, 1997.

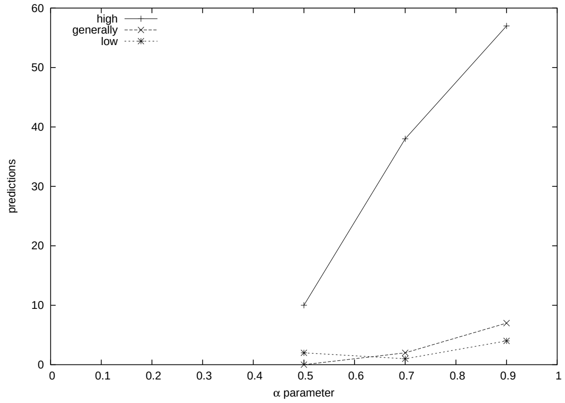

Figure 6: Protein evaluation of our task 2.2 submissions. For our predictions with a high evidence of the GO term, this figure shows how many of them provide a high, general and low evidence of the protein for the values of α used.