Published in Lecture Notes in Computer Science 3408, 112-126, 2005
which should be used for any reference to this work

Features Combination for Extracting Gene Functions from MEDLINE

Patrick Ruch[1], Laura Perret[2], and Jacques Savoy[2]

[1] University Hospital of Geneva
Patrick.Ruch@sim.hcuge.ch
[2] University of Neuchâtel
{Laura.Perret, Jacques.Savoy}@unine.ch

Abstract. This paper describes and evaluates a summarization system that extracts the gene function textual descriptions (called GeneRIF) based on a MedLine record. Inputs for this task include both a locus (a gene in the LocusLink database), and a pointer to a MedLine record supporting the GeneRIF. In the suggested approach we merge two independent phrase extraction strategies. The first proposed strategy (LASt) uses argumentative, positional and structural features in order to suggest a GeneRIF. The second extraction scheme (LogReg) incorporates statistical properties to select the most appropriate sentence as the GeneRIF. Based on the TREC-2003 genomic collection, the basic extraction strategies are already competitive (52.78% for LASt and 52.28% for LogReg, respectively). When used in a combined approach, the extraction task clearly shows improvement, achieving a Dice score of over 55%.

1 Introduction

As an increasing amount of information becomes available in the form of electronic documents, the increasing need for intelligent text processing makes shallow text understanding methods such as the Information Extraction (IE) particularly useful [19]. Until now, IE has been defined in a restricted manner by DARPA's MUC (Message Understanding Conference) program [4], as a task involving the extraction of specific, well-defined types of information from natural language texts in restricted domains, with the specific objective of filling pre-defined template slots and databases. Examples of such classical information extraction tasks are given by the BioCreative[1] named-entity recognition task or the JNLPBA shared task (e.g., [14]). Recently, the TREC-2003 Genomics Track proposed that the IE task be extended by extracting entities that were less strictly defined. As such, the 2003 Genomics Track suggested extracting gene functions as defined in the LocusLink database. In this repository, records (called *locus*, which refer to a gene or a protein) are provided with a short fragment of text to explain their biological function together with a link to the corresponding scientific article. These so-called Gene Reference Into Functions (GeneRIFs) are usually short extracts taken from MedLine articles. As with classical Named-Entities (NE) such as the names of persons, locations or genes, GeneRIFs are

[1] www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04

too extensive to be comprehensively listed, but their major difference is that gene functions are usually expressed by a sentence rather than a single word, an expression or a short phrase. GeneRIF variations thus contrast with those of other related tasks. Just as in the context of the BioCreative challenge, the automatic text categorization task also attempts to predict the function of proteomic entities (based on the SwissProt repository) using literature excepts [7]. In that task however, the set of available functions is strictly defined in the Gene Ontology[2]. Moreover, the application of machine learning techniques [22] to filter relevant fragments has received little attention in IE ([5] is an exception) compared to other tasks such as named entity recognition. This lack of interest is due to the type of texts that are generally handled by IE, which are those proposed in the MUC competition. These texts are often short newswires and the information to be extracted is generally dense, so there is much less if any need for prefiltering. For example, the type of information to be extracted may be company names or seminar starting times, often only requiring a shallow analysis. The computational cost is thus fairly low and prefiltering can be avoided. This is clearly not the case in other IE tasks such as those for identifying gene functions in genomics, the application that we describe here.

The remainder of the paper is organized as follows: Section 2 provides an overview of the state of the art. Section 3 describes the different methods and their combinations, as well as the metrics defined for the task. Section 4 reports on the results.

2 Background and Applications

Historically, seminal studies dedicated to the selection of textual fragments were done for automatic summarization purposes, but recently, due to developments in life sciences, more attention has been focused on sentence filtering.

2.1 Summarization

In automatic summarization, the sentence is the most common type of text-span[3] used because in most general cases, it is too difficult to understand, interpret, abstract, and generate a new document or a short summary. By choosing sentences as generation units, many co-reference issues [28] are partially avoided. Although more knowledge intensive approaches have been investigated, it seems simpler and more effective to view the summarization problem as a sentence extraction problem.

In this vein, Goldstein *et al.* [8] distinguish between two summary types: generic and query-driven. This distinction is useful relative to our information extraction task, since it involves question answering (with the query-driven type) or summarization (generic). In our study, other summarization criteria such as the length and style of the generated abstract can be ignored. Feature selection and their weighting, often based on term frequency and inverse document frequency factors (*tf idf*) have been reported. Conclusions reached are however not always consistent [25] about *tf idf*. Among other

[2] www.geneontology.org/

[3] Berger & Mittal [1] define a summarization task, called *gisting* which aims at reducing the sentences by modeling content-bearing words. The suggested strategy seems effective for summarizing non-argumentatively structured documents such as Web pages.

interesting features, both sentence location as well as sentence length seem important [15]. In addition these authors rely on a set of frequent phrases and keywords. Finally, to extract important sentences from documents, a document's titles and uppercase words such as named-entities are reported to be good predictors. Of particular interest for our approach, Teufel & Moens [31] define a large list of manually weighted triggers (using both words and expressions such as *we argued*, *in this article*, *the paper is an attempt to, etc.*) to automatically structure scientific articles into seven argumentative moves, namely: BACKGROUND, TOPIC, RELATED WORK, PURPOSE, METHOD, RESULT, and CONCLUSION.

2.2 A Genomics Perspective on Information Extraction

To date and as with gene functions, descriptions of most of the biological knowledge about these interactions cannot be found in databanks, but only in the form of scientific summaries and articles. Making use of these represents a major milestone towards building models of the various interactions between entities in molecular biology; and sentence filtering has therefore been greatly studied for its potential in mining literature on functional genomics. For example, sentence filtering for protein interactions was previously described in [23] and [3]. In these studies, sentence filtering is viewed as a prerequisite step towards deeper understanding of texts.

Input

Locus - ABCA1: ATP-binding cassette, sub-family A (ABC1), member 1

MedLine record - PMID - 12804586

TI - Dynamic regulation of alternative ATP-binding cassette transporter A1 transcripts.

AB – […] The longest (class 1) transcripts were abundant in adult brain and fetal tissues. Class 2 transcripts predominated in most other tissues. The shortest (class 3) transcripts were present mainly in adult liver and lung. To study the biochemical significance of changes in transcript distribution, two cell models were compared. In primary human fibroblasts, upregulation of mRNA levels by oxysterols and retinoic acid increased the relative proportion of class 2 transcript compared to class 1. Phorbol ester stimulated human macrophage-derived THP-1 cells increased the abundance of class 1 transcripts relative to class 2. In both cell lines class 3 transcript levels were minimal and unchanged. It is shown here for the first time that the **regulation of ABCA1 mRNA levels exploits the use of alternative transcription start sites**.

Output

GeneRIF - **regulation of ABCA1 mRNA levels exploits the use of alterntive transcription start sites**

Fig. 1. Example of a LocusLink record and the corresponding GeneRIF (bold added)

2.3 TREC-2003 Corpus

To provide a general view of the problems underlying the generation of the most appropriate GeneRIF during the TREC-2003 Genomics Track [10], a simple example is provided in Fig. 1. In this figure we can see the locus (“ABCA1”) and the MedLine

record identifier ("PMID – 12804586"). Under the label "TI" is found the article's title and under "AB" its abstract (from which the GeneRIF is extracted).

A preliminary study [21] showed that around 95% of the GeneRIF snippets were extracted from the title or from the abstract of the corresponding scientific paper. Moreover, from this set, around 42% were a direct "cut & paste" from either the title or the abstract (Fig. 1 is such an example) while another 25% contained significant portions of the title or abstract.

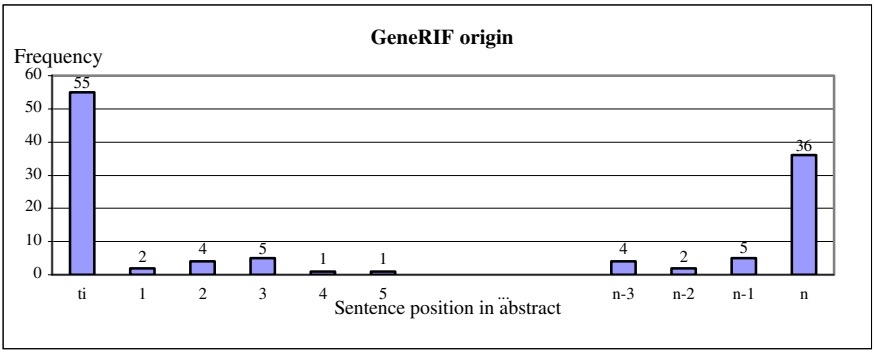

Fig. 2. GeneRIF distribution in titles ("ti") and in abstracts (from 1 to n)

In the TREC evaluation data, we analyzed the sentence location distribution used to produce the GeneRIF. In this case, we considered the title (see Fig. 2, the first column labeled "ti") and the abstract's sentence sequence. From the 139 GeneRIFs used in our experiments, 55 were mainly extracted from the article's title, as depicted in Fig. 2. The second most frequent source of GeneRIF was the abstract's last sentence (see the last column in Fig. 2, following the label "n"), showing the source of 36 GeneRIFs. Between these two extreme positions, the GeneRIF location distribution is rather flat.

3 Methods

As for automatic abstracting, evaluating sentence classifiers is difficult. First of all establishing the benchmark notion is clearly a more complex task. Secondly it is less universally defined, as compared to other automatic text classification tasks such as spelling correction, document routing or in information retrieval systems evaluation.

3.1 Metrics

In general, for each input text the classification techniques yield a ranked list of candidates. Thus, sentence filtering like information extraction and text categorization may be formally evaluated by the usual recall and precision measures. However, we must recognize that it is hard to obtain complete agreement regarding the appropriate measure that should be used. It has been argued [18] that in evaluating a binary classi-

fication system, one should use effectiveness measures based on estimates of class membership rather than measures based on rankings. On the other hand, a precision oriented-metric such as 11-point average precision has been suggested [17]. In the TREC-2003 genomic evaluation campaign, a third type of measure was used to evaluate information extraction: the Dice coefficient as shown in Eq. 1. In this formula, the numerator indicates the number of common words between the candidate sentence and the exact GeneRIF, while the denominator represents the total number of words in the GeneRIF and in the candidate. Thus, this similarity coefficient measures the lexical overlap between a candidate and the corresponding correct GeneRIF.

$$\text{Dice} = \frac{2 \cdot |X \cap Y|}{|X| + |Y|} \tag{1}$$

More precisely, four Dice coefficients variants were suggested, and all were found to be highly correlated. Thus, in our experiments the Dice metrics shown in Eq. 1 will be used. This measure assumes that a binary decision was made prior to computing the Dice: a unique candidate GeneRIF must be selected.

3.2 Common Pre- and Post-processing Strategies

We started designing the task as though a form of ranking task. Sentences were the entities to be classified, so we assumed that GeneRIFs were sentences or significant sentence fragments. This has its limitations since some examples in the training and test data showed the opposite effect: GeneRIFs were sometimes the synthesis of more than one sentence. Such examples were however not the norm and also the generation of a well-formed sentence required the resolution of complex linguistic phenomena (e.g., anaphora, pronoun generations), and this was beyond the scope of our study. For sentence splitting, we developed a robust tool based on manually crafted regular expressions. This served to detect sentence boundaries with more than 97% precision on MedLine abstracts, and was deemed competitive with more elaborate methods [26]. In order to avoid applying our classifiers on erroneously segmented sentences, segments with less than 20 characters were simply removed from the list of candidate sentences.

Next, both strategies ranked the candidate sentences separately. From these two rankings, our aim was to identify a confidence estimator and then choose the final candidate when both our schemes disagreed on the best choice. This last step transformed the two ranking tools into a binary classifier, thus finally deciding whether a candidate sentence was relevant or not. The relevant sentence (the one that was unique in each [locus, abstract] pair) is post-processed by a syntactic module, in an attempt to eliminate irrelevant phrases from the selected sentence.

This sentence reduction step used a part-of-speech tagger [27] and a standard list of 369 stopwords (e.g., *so*, *therefore*, *however*, *then*, etc.) together with a set of stop phrases (e.g., *in contrast to other studies*, *in this paper*, etc.). When these stop phrases occurred they were removed from the beginning of the selected GeneRIF candidate. Part-of-speech information was used to augment the list of stopwords, thus any ad-

verb (e.g., *finally*, *surprisingly*, etc.) located at the beginning of a sentence was removed. In the same manner, this procedure removed non-content bearing introductory syntagms when they were located at the beginning of the sentence: any fragment of text containing a verb and ending with *that*, as in *we show that*, *the paper provides the first evidence that*, were deleted. The stopword and stop phrase removal steps were applied sequentially, but we arbitrarily limited the length of the deleted segment at a maximum of 60 characters. Moreover, text removal was blocked when clauses contained gene and protein names (GPN). Our GPN tagger is based on a very simple heuristic: any non-recognized token was considered as a GPN. We used the UMLS SPECIALIST Lexicon and a frequency list of English words (totaling more than 400,000 items) to separate between known and unknown words.

INTRODUCTION: Chromophobe renal cell carcinoma (CCRC) comprises 5% of neoplasms of renal tubular epithelium. CCRC may have a slightly better prognosis than clear cell carcinoma, but outcome data are limited. *PURPOSE:* In this study, we analyzed 250 renal cell carcinomas to a) determine frequency of CCRC at our Hospital and b) analyze clinical and pathologic features of CCRCs. *METHODS:* A total of 250 renal carcinomas were analyzed between March 1990 and March 1999. Tumors were classified according to well-established histologic criteria to determine stage of disease; the system proposed by Robson was used. *RESULTS:* Of 250 renal cell carcinomas analyzed, 36 were classified as chromophobe renal cell carcinoma, representing 14% of the group studied. The tumors had an average diameter of 14 cm. Robson staging was possible in all cases, and 10 patients were stage 1) 11 stage II; 10 stage III, and five stage IV. The average follow-up period was 4 years and 18 (53%) patients were alive without disease. *CONCLUSION:* The highly favorable pathologic stage (RI-RII, 58%) and the fact that the majority of patients were alive and disease-free suggested a more favorable prognosis for this type of renal cell carcinoma.

Fig. 3. Example of an explicitly structured abstract in MedLine

3.3 Latent Argumentative Structuring

The first classifier (called LASt) started ranking abstract sentences as to their argumentative classes (as proposed in [20, 30]). Four argumentative categories defined four moves: PURPOSE, METHODS, RESULTS and CONCLUSION. These moves were chosen because in scientific literature they have been found to be quite stable [24], [29] and were also recommended by ANSI/ISO guidelines for professionals. We obtained 19,555 explicitly structured abstracts from MedLine in order to train our Latent[4] Argumentative Structuring classifier (LASt) (this set does not contain the MedLine records used during the evaluation). A conjunctive query was used to combine the four following strings: "PURPOSE," "METHODS," "RESULTS," "CONCLUSION". From the original set, we retained 12,000 abstracts (an example is given in Fig. 3) used for training our LASt system, and 1,200 were used for fine-tuning and evaluating the tool, following removal of explicit argumentative markers.

[4] We assume that documents to be classified contain at least a *latent* argumentative structure.

CONCLUSION |00160116| The highly favorable pathologic stage (RI-RII, 58%) and the fact that the majority of patients were alive and disease-free suggested a more favorable prognosis for this type of renal cell carcinoma.
PURPOSE |00156456| In this study, we analyzed 250 renal cell carcinomas to a) determine frequency of CCRC at our Hospital and b) analyze clinical and pathologic features of CCRCs.
PURPOSE |00167817| Chromophobe renal cell carcinoma (CCRC) comprises 5% of neoplasms of renal tubular epithelium. CCRC may have a slightly better prognosis than clear cell carcinoma, but outcome data are limited.
METHODS |00160119| Tumors were classified according to well-established histologic criteria to determine stage of disease; the system proposed by Robson was used.
METHODS |00162303| Of 250 renal cell carcinomas analyzed, 36 were classified as chromophobe renal cell carcinoma, representing 14% of the group studied.
RESULTS |00155338| Robson staging was possible in all cases, and 10 patients were stage 1) 11 stage II; 10 stage III, and five stage IV.

Fig. 4. Classification example for abstract shown in Fig. 3. The attributed class comes first, then the score obtained by the class, and finally the text segment

3.3.1 Features and Heuristics

Our system relied on four Bayesian classifiers [16], one binary classifier for each argumentative category. Each binary classifier combined three types of features: words, word bigrams and trigrams. The log of the class frequency represented the weight of each feature, but for every category, DF thresholding [33] was applied so that rare features were not selected. Finally, the class estimate provided by each binary classifier was used to attribute the final class (an example is shown in Fig. 3 and 4): for each sentence the classifier with the highest score assigns the argumentative category. Optionally, we also investigated the sentence position's impact on classification effectiveness through assigning a relative position to each sentence. Thus, if there were ten sentences in an abstract: the first sentence had a relative position of 0.1, while the sentence in position 5 received a relative position of 0.5, and the last sentence has a relative position of 1. The following heuristics were then applied: 1) if a sentence has a relative score strictly inferior to 0.4 and is classified as CONCLUSION, then its class becomes PURPOSE; 2) if a sentence has a relative score strictly superior to 0.6 and is classified as PURPOSE, then its class is rewritten as CONCLUSION.

Table 1 shows the results of argumentative classification system based on the evaluation set. This table indicates the confusion matrices between the four classes, with and without the use of relative position heuristics. When the sentence position was not taken into account, 80.65% of PURPOSE sentences were correctly classified, while 16% were wrongly classified as CONCLUSION, and 3.23% as RESULTS. On the other hand, when the sentence position was taken into account, 93.55% of PURPOSE sentences were correctly classified. The data depicted in this table demonstrates that position can be useful for separating between the PURPOSE and CONCLUSION classes. However, the percentages of correct classified sentences in the METHODS or RESULTS classes did not vary when the sentence position was taken into account. In both cases,

the percentage of correct answers was similar, 78% and around 50% respectively, for the METHODS and RESULTS classes.

Table 1. Confusion matrices for the argumentative classifier: the columns denote the manual classification and the rows indicate the automatic ones; percentages on the diagonal give the proportion of sentences, which are appropriately categorized by the argumentative classifier (evaluation done on 17,612 sentences)

Without sentence positions				
	PURP	METH	RESU	CONC
PURP	80.65%	0%	3.23%	16%
METH	8%	78%	8%	6%
RESU	18.58%	5.31%	52.21%	23.89%
CONC	18.18%	0%	2.27%	79.55%
With sentence positions				
	PURP	METH	RESU	CONC
PURP	93.55%	0%	3.23%	3%
METH	8%	78%	8%	6%
RESU	12.43%	5.31%	74.25%	13.01%
CONC	2.27%	0%	2.27%	95.45%

3.3.2 Argumentation and GeneRIF

In another preliminary experiment we tried to establish a relation between GeneRIF and argumentative moves. We selected two sets of 1000 GeneRIFs from our training data and submitted them to the argumentative classifier. Set A was a random set and set B was also a random set, but we imposed the condition that the extract describing the GeneRIF had to be found in the abstract (as exemplified in Fig. 1). We wanted to verify if the argumentative distribution of GeneRIF originating from sentences is similar to the distribution of GeneRIF originating from both titles and abstracts. Results of the argumentative classification are given in Table 2 for these two sets. These proportions indicate that GeneRIFs are mainly classified as PURPOSE and CONCLUSION sentences (respectively 41% and 55% in Set A). The significance of these observations was accentuated for the GeneRIFs coming from the abstract sentences (see Set B in Table 2). In this case, two thirds of the sentence-based GeneRIFs came from the CONCLUSION, and around a quarter from the PURPOSE section. Together, these two moves concentrated between 88% (Set B) and 96% (Set A) for the GeneRIFs in LocusLink. Fortunately, as shown in Table 1, the discriminative power of the argumentative classifier was better for these two classes than for the RESULTS and METHODS classes.

Based on these findings, the sentence ranking order would be: CONCLUSION, PURPOSE, RESULTS, METHODS, and thus our classifier would return to the first position, when available, a sentence classified as CONCLUSION. However, selecting the best conclusion sentence is not sufficient (such a strategy exhibits a Dice performance of 35.2%), due to the fact that 45% of GeneRIFs in the TREC evaluation set were

strictly “cut & paste” from the article's title. In our argumentation-based ranking we clearly needed to take the title into account. To do so, we simply computed the Dice distance between each candidate and the title, so that among sentences classified as CONCLUSION and PURPOSES, those lexically similar to the title would move to the top of the list. In a complementary manner, a negative filter was also used; meaning sentences without GPNs were simply eliminated. Finally, to select between the title and the best-ranked sentence from the abstract, the Dice score was again used. If the sentence score was above a given threshold, then the sentence was selected, otherwise the title was returned. From our training data, the best threshold was 0.5. This threshold value gives the best results on the test set: the classifier choses 14 sentences from the abstract vs. 125 from the title, from a total of 139 queries (see Table 6).

Table 2. Class distribution in 1000 GeneRIFs after argumentative classification. Sets A and B are samples of GeneRIFs as in LocusLink, but Set B contains only GeneRIFs originating from the abstract

	Set A (%)	**Set B (%)**
PURPOSE	41%	22%
METHODS	2%	4%
RESULTS	2%	8%
CONCLUSION	55%	66%

3.4 Logistic Regression

The second suggested extraction strategy (called LogReg) is based on logistic regression and works in two stages. During the first step, the system computes a score for each sentence in order to define the best possible candidate sentence. During the second step and as was done is our LASt scheme, the selected candidate was compared to the paper’s title in order to define whether the title or the candidate should be returned as the suggested GeneRIF. In the first step, we removed all stopwords (we used the SMART stopword list) appearing in the title or in the abstract sentences. We then applied the S stemmer [9] in an attempt to remove the English plural form (mainly the final « -s »). After removing stopwords and applying the S stemmer, we computed a score for each sentence and for the title, using the following formula:

$$\text{score} = \frac{1}{\text{len}} \sum_{j=1}^{\text{len}} w(tf_j) \qquad (2)$$

where tf_j was the term frequency in GeneRIF vocabulary, len was the sentence length measured by word count and $w(tf_j)$ was a weight function as defined in Table 3, returning an integer that depended on the term frequency tf_j. To define each term frequency in the GeneRIFs vocabulary, we simply counted the number of occurrences of the corresponding term in all GeneRIFs. For example, we were able to observe that the term “cell” appeared 36 times, “role” 25 or “protein” 21.

Table 3. Ad hoc weight according to term frequency in GeneRIF

tf_j	$w(tf_j)$
$9 < tf_j$	4
$4 < tf_j \leq 9$	3
$2 < tf_j \leq 4$	2
$1 < tf_j \leq 2$	1
$tf_j \leq 1$	0

Finally, we ranked the sentences (including the title) according to their scores in decreasing order and then selected the desired candidate: the sentence with the highest score (this candidate could be the title). Such a weighting scheme thus promoted the sentence having the most terms in common with the vocabulary found in the GeneRIFs. Moreover, if these common terms were also frequent words (e.g., like "cell" or "protein"), the underlying score would increase.

Table 4. Title and candidate sentence for Query #30

Original title	Comparative surface accessibility of a pore-lining threonine residue (T6') in the glycine and GABA(A) receptors.
After stopword removal and stemming	Comparative surface accessibility pore-lining threonine residue (T6') glycine GABA(A) receptor
Original candidate	This action was not induced by oxidizing agents in either receptor.
After stopword removal and stemming	action induced oxidizing agent either receptor

Just as in the LASt approach, knowing that the title is often an appropriate GeneRIF source we wanted to account for this by suggesting a selection model. This selection scheme had to choose between either the paper's title or the best candidate sentence. The following example illustrates how this selection scheme worked. As shown in Table 4, for Query #30 we reported the paper's title and the best candidate sentence. Note that the table shows both the original form and the resulting expression once stopwords were removed and the stemming procedure applied.

For each candidate sentence, we computed statistics such as length (denoted "Len"), number of indexed terms (or number of words appearing in GeneRIF vocabulary, denoted "Terms"). We also added statistics related to the *idf* value, based on the work of Cronen-Townsend *et al.* [6], who demonstrated that the *idf* could be, under certain conditions, a good estimator for predicting query performance.

Table 5. Variables used in our logistic model

Variable	Estimate	Meaning
Len	0.4512	candidate sentence length
Unknown	0.2823	number of unknown terms in WordNet
Terms	-0.5538	number of indexed terms
Max2Idf	-0.3638	2nd max idf of candidate sentence
MinIdf	-0.5611	min idf of candidate sentence
d.Len	-0.3560	length difference between candidate and title
d.Terms	0.4465	indexed term number difference between candidate and title
d.Max2Idf	0.4084	2nd max idf difference between candidate and title
d.MinIdf	1.0351	min idf difference between candidate and title

Moreover, since we compared the title and a given sentence, we were able to compute statistics on the differences between this sentence and the paper's title. For example, we included the length difference (d.Len) or idf minimum difference (d.MinIdf), between the candidate and the title. Once a set of potential useful explanatory variables was obtained, we selected the most important ones using the `stepAIC` procedure [32]. Table 5 describes all retained variables used in our selection model.

To implement this selection procedure we chose the logistic regression model [11] in order to predict the probability of a binary outcome variable according to a set of explanatory variables. In this case, our logistic regression model returned a probability estimate that the candidate sentence was a good GeneRIF, based on the explanatory variables depicted in Table 5. For example, if the candidate sentence length was greater than the title length (variable "d.Len" would be positive), then the probability that the candidate sentence was a good GeneRIF would decrease (because, as shown in Table 5, the estimate for "d.Len" is negative). Finally, if the estimated probability was greater than 0.5, then the sentence was returned as the proposed GeneRIF, otherwise the article title was returned as the GeneRIF. Using this method, we returned the paper's title 97 times and the candidate sentence 42 times (see Table 6).

3.5 Fusion of Extraction Strategies

This last step attempts to combine our two extraction schemes. To achieve this goal, we used the following rules: 1) Agreement - if the sentence selected by LASt is also chosen by the logistic regression strategy (LogReg), then we keep it; 2) Disagreement - if both strategies do not agree, then we look at the probability estimate returned by LogReg: if this probability is below a given threshold (0.5), then the candidate sentence provided by LASt is selected, otherwise the LogReg candidate is returned. Finally, if a unique candidate GeneRIF is selected and if this segment does not come from the title, then the sentence is processed by the reduction procedure (see Section 3.2). The output segment is used for comparison to the correct GeneRIF provided by LocusLink's annotators, as explained in the next section.

4 Results and Related Works

In this section, we first evaluated each isolated extraction strategy. Second, we evaluated our suggested combined approach. Table 6 depicts the overall performance measure using the Dice coefficient (last column). The table's middle column shows how the proposed GeneRIF may have originated from the article's title or from an abstract sentence. Our baseline approach was very simple. For each of the 139 queries (composed of a locus and a MedLine article), we returned the article's title. Such a naïve selection procedure achieved a relatively high performance of 50.47%, due to the fact that 45% of GeneRIFs were extracted from the article's title. On the other hand, if for each query we had an oracle that always selected the title or the sentence achieving the highest Dice score, we could obtain a performance of 70.96%, one that represents our upper bound. In this optimal run, we had to extract 59 titles and 80 sentences from the abstract. We could not however obtain a better performance level due to the fact that LocusLink's annotators may have used words that did not appear in the article's title or in the abstract. Moreover, correct GeneRIFs may paraphrase a sentence or the article's title, revealing the same gene function with different words or expressions. Finally, GeneRIFs may be expressed using more than one sentence. In this case, the human annotator chose to combine different segments, taken from various sentences or in part from the article's title.

Table 6. Performance of each basic strategy and their combination

	Origin of proposed GeneRIF		
	Title	**Abstract**	**Dice**
Baseline	139	0	50.47%
LASt	125	14	51.98%
LogReg	97	42	52.28%
Combination	106	33	54.44%
Combination & shortening	106	33	55.08%

As shown in Table 6, the LASt extraction approach produced an overall performance of 51.98%, and in this case, 125 GeneRIFs came from the article's title and 14 from the article's abstract. Our second extraction scheme (run labeled "LogReg") performed at similar levels (52.28%). However, in this case, it was seen that a greater number of proposed GeneRIFs came from the abstract (42 vs. 14 in the LASt scheme). The last two rows of Table 6 indicate the performance of our combined approach (54.44%), clearly showing better overall results than those for each extraction scheme run separately. When we applied our sentence reduction procedure, the Dice score increased slightly (55.08% vs. 54.44%). When analyzing the origin of each proposed GeneRIF in this combined approach, we could see that 106 come from the title and 33 from the abstract. Moreover, when applying another point of view, we found that 48 suggested GeneRIFs were provided by LASt, 22 came from LogReg, and the two extraction strategies agreed 69 times.

While these results reveal attractive performance levels when compared to other runs in the TREC-2003 genomic evaluation campaign [10], several teams were faced with the same extraction problem yet suggested other interesting approaches. For example, Bhalotia *et al.* [2], ranked second at TREC (Dice = 53%) suggested a scheme that selected between the article's title and the last sentence of the article's abstract (as shown in Fig. 2, 91 out of the 139 GeneRIFs were extracted from either the title or the abstract's last sentence). These authors suggested basing this selection on a Naive Bayes [22] machine learning approach. The relevant variables were the verbs, the MeSH and the genes, all weighted by *tf idf*, as well as a Boolean value representing the presence of the target gene in the abstract. Although we were not able to reproduce their results based on their TREC report, Jelier *et al.* [12] report a Dice score close to 57%, using similar classifiers, but trained on the sentence position in the abstract. Another interesting approach proposed by Kayaalp *et al.* [13] separates the articles, abstracts and titles into sentences in order to combine their various characteristics, such as the number of words, number of figures and number of uppercase letters. The first model applied a linear combination on a set of characteristics so as to extract the best candidate sentence, whereas the second model was based on the predicate calculus, using another set of characteristics.

5 Conclusion

This research focuses on the extraction of gene functions (a GeneRIF) from a MedLine record given a gene name, as was proposed in the TREC Genomics Track in 2003 [10]. Because almost half of the human-provided GeneRIFs were simply "cut & paste" from the title, the method focused on deciding whether a sentence from the abstract would likely express the GeneRIF or if the title would be a better choice. The investigated method combines two independent extraction strategies. The first relies on argumentative criteria and considers that apart from the title, the best GeneRIF candidate should appear in the article's conclusion or purpose sections. The second extraction approach is based on logistic regression which returns a probability estimate that the selected sentence provides a better GeneRIF than does the title. The probabilistic estimates are based on the lexical usage in the sentence and on various statistical properties (together with their differences) shared between the candidate sentence and the title (e.g., the length difference, the minimal *idf* value, etc.). Each extraction strategy operates on the same basic unit: the sentences and/or the article's title. Moreover, each suggested approach shows a preference for the sentences in which genes and protein names occur.

When examined separately, each method (argumentative filtering and logistic regression) yielded effective results during the TREC-2003 challenge [10]. However combining achieved a highly competitive score: the lexical overlap – measured by Dice metrics – was improved by about 9% compared to the baseline (55.08% vs. 50.47%). In conclusion, the methods used in these experiments provide a general view of the gene function extraction task within the TREC genomic evaluation campaign. As with summarization and sentence selection, these methods clearly show that a variety of feature sets must be considered when performing such information extraction tasks.

Acknowledgments

This research was supported in part by the SNSF (Grants 21-66 742.01 and 3200-065228) and in part by the EU/OFES (Grant 507505/03.0399).

References

[1] Berger, A.L., Mittal, V.O.: OCELOT: A System for Summarizing Web Pages. In Proceedings ACM-SIGIR'2000, ACM Press, New York, 144-151
[2] Bhalotia, G., Nakov, P.I., Schwartz, A.S., Hearst, M.A.: BioText Team Report for the TREC 2003 Genomics Track. In Proceedings TREC-2003, NIST, 612-621
[3] Blaschke, C., Andrade, M.A., Ouzounis, C.A., Valencia, A.: Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. In Proceedings Intelligent Systems for Molecular Biology, 1999, 60-67
[4] Chinchor, N., Hirschman, L.: MUC-7 Named-Entity Task Definition. In MUC-7 Proceedings, 1998.
[5] Chuang W.T., Yang, J.: Extracting Sentence Segments for Text Summarization. In Proceedings ACM-SIGIR'2000, ACM Press, New York, 152-159
[6] Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting Query Performance. In Proceedings ACM-SIGIR'2002, ACM Press, New York, 299-306
[7] Ehrler F., Jimeno Yepes A., Ruch P.: Data-Poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot. BMC Bioinformatics, 2005, to appear
[8] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.: Summarizing Text Documents. In Proceedings ACM-SIGIR'99, ACM Press, New York, 121-128
[9] Harman, D.: How Effective is Suffixing? J Am Soc Infor Scien. 42 (1991) 7-15
[10] Hersh, W., Bhupatiraju, R.T.: TREC Genomics Track Overview. In Proceedings TREC-2003, NIST, 14-23
[11] Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. 2nd edn. John Wiley & Sons, New York (2000)
[12] Jelier, R., Schuemie, M., van der Eijk, C., Weeber, M., van Mulligen, E., Schijvenaars, B., Mons, B., Kors, J.: Searching for GeneRIFs: Concept-Based Query Expansion and Bayes Classification. In Proceedings TREC-2003, NIST, 225-233
[13] Kayaalp, M., Aronson, A.R., Humphrey, S.M., Ide, N.C., Tanabe, L.K., Smith, L.H., Demner, D., Loane, R.R., Mork, J.G., Bodenreider, O.: Methods for Accurate Retrieval of MEDLINE Citations in Functional Genomics. In Proceedings TREC-2003, 441-450
[14] Kirsch, H., Rebholz-Schuhmann D.: Distributed Modules for Text Annotation and IE applied to the Biomedical Domain. COLING Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004, 50-53
[15] Kupiec, J., Pedersen, J., Chen, F.: A Trainable Document Summarizer. In Proceedings ACM-SIGIR'95, ACM Press, New York, 68-73
[16] Langley, P., Iba, W., Thompson, K.: An Analysis of Bayesian Classifiers. In Proceedings AAAI, Menlo Park, 1992, 223-228
[17] Larkey, L.S., Croft, W.B.: Combining Classifiers in Text Categorization. In Proceedings ACM-SIGIR'96, ACM Press, New York, 289-297
[18] Lewis, D.D.: Evaluating and Optimizing Autonomous Text Classification Systems. In Proceedings ACM-SIGIR'95, ACM Press, New York, 246-254
[19] Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. The MIT Press, Cambridge (1999)

[20] McKnight, L. Srinivasan, P.: Categorization of Sentence Types in Medical Abstracts. In Proceedings AMIA 2003, 440-444
[21] Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M., Ward, J.M.: Gene Indexing: Characterization and Analysis of NLM's GeneRIFs. In Proceedings AMIA 2003, 460-464
[22] Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
[23] Nédellec, C., Vetah, M., Bessières, P.: Sentence Filtering for Information Extraction in Genomics. In Proceedings PKDD, Springer-Verlag, Berlin, 2001, 326-237
[24] Orasan, C.: Patterns in Scientific Abstracts. In Proceedings Corpus Linguistics, 2001, 433-445
[25] Paice, C.D.: Constructing Literature Abstracts by Computer: Techniques and Prospects. Inform Proc & Manag. 26 (1990) 171-86
[26] Reynar, J.C., Ratnaparkhi, A.: A Maximum Entropy Approach to Identifying Sentence Boundaries. In Proceedings Applied NLP, 1997, 16-19
[27] Ruch, P., Baud, R., Bouillon, P., Robert, G.: Minimal Commitment and Full Lexical Disambiguation. In Proceedings of CoNLL, 2000, 111-116
[28] Strube, M., Hahn, U.: Functional Centering. In Proceedings of ACL, Morgan Kaufmann, 1996, 270-277
[29] Swales, J.: Genre Analysis: English in Academic and Research Settings. Cambridge University Press, Cambridge (1990)
[30] Tbahriti, I., Chichester, C., Lisacek, F., Ruch P.: Using Argumention to Retrieve Articles with Similar Citations from MEDLINE. COLING Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004, 8-14
[31] Teufel S., Moens, M.: Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In: I. Mani, M. Maybury (eds.): Advances in Automatic Text Summarization. MIT Press, Cambridge (1999) 155-171
[32] Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S-Plus. 3rd edn. Spinger-Verlag, New York (2000)
[33] Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In Proceedings Machine Learning, Morgan Kaufmann, 1997, 412-420