Using the argumentative structure of scientific literature to improve information access

Antonio Jimeno Yepes
National ICT Australia
Victoria Research Laboratory
Melbourne, Australia
antonio.jimeno@gmail.com

James G. Mork
National Library of Medicine
8600 Rockville Pike
Bethesda, 20894, MD, USA
mork@nlm.nih.gov

Alan R. Aronson
National Library of Medicine
8600 Rockville Pike
Bethesda, 20894, MD, USA
alan@nlm.nih.gov

Abstract

MEDLINE/PubMed contains structured abstracts that can provide argumentative labels. Selection of abstract sentences based on the argumentative label has shown to improve the performance of information retrieval tasks. These abstracts make up less than one quarter of all the abstracts in MEDLINE/PubMed, so it is worthwhile to learn how to automatically label the non-structured ones.

We have compared several machine learning algorithms trained on structured abstracts to identify argumentative labels. We have performed an intrinsic evaluation on predicting argumentative labels for non-structured abstracts and an extrinsic evaluation to predict argumentative labels on abstracts relevant to Gene Reference Into Function (GeneRIF) indexing.

Intrinsic evaluation shows that argumentative labels can be assigned effectively to structured abstracts. Algorithms that model the argumentative structure seem to perform better than other algorithms. Extrinsic results show that assigning argumentative labels to non-structured abstracts improves the performance on GeneRIF indexing. On the other hand, the algorithms that model the argumentative structure of the abstracts obtain lower performance in the extrinsic evaluation.

1 Introduction

MEDLINE®/PubMed® is the largest repository of biomedical abstracts. The large quantity of unstructured information available from MEDLINE/PubMed prevents finding information efficiently. Reducing the information that users need to process could improve information access and support database curation. It has been suggested that identifying the argumentative label of the abstract sentences could provide better information through information retrieval (Ruch et al., 2003; Jonnalagadda et al., 2012) and/or information extraction (Mizuta et al., 2006).

Some journals indexed in MEDLINE/PubMed already provide the abstracts in a structured format (Ripple et al., 2012). A structured abstract[1] is an abstract with distinct labeled sections (e.g., Introduction, Background, or Results). In the MEDLINE/PubMed data, these labels usually appear in all uppercase letters and are followed by a colon (e.g., *MATERIALS AND METHODS:*). Structured abstracts are becoming an increasingly larger segment of the MEDLINE/PubMed database with almost a quarter of all abstracts added to the MEDLINE/PubMed database each year being structured abstracts. A recent PubMed query (April 22, 2013) shows 1,050,748 citations from 2012, and 249,196 (23.72%)[2] of these are considered structured abstracts.

On August 16, 2010, PubMed began displaying structured abstracts formatted to highlight the various sections within the structured abstracts to help readers identify areas of interest[3]. The XML formatted abstract from MEDLINE/PubMed separates each label in the structured abstract and includes a mapping to one of five U.S. National Library of Medicine (NLM) assigned categories as shown in the example below:

<AbstractText Label="MATERIALS AND METHODS" NlmCategory="METHODS">

The five NLM categories that all labels are mapped to are OBJECTIVE, CONCLUSIONS, RESULTS, METHODS, and BACKGROUND (Ripple et al., 2011). If a label is new

[1] http://www.nlm.nih.gov/bsd/policy/structured_abstracts.html
[2] hasstructuredabstract AND 2012[pdat]
[3] http://www.nlm.nih.gov/pubs/techbull/ja10/ja10_structured_abstracts.html

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 102–110,
Sofia, Bulgaria, August 4-9 2013.

or not in the list of reviewed structured abstract labels, it will receive a category of *UNASSIGNED*. There are multiple criteria for deciding what abstracts are considered structured abstracts or not. One simple definition would be that an abstract contains one or more author defined labels. A more rigid criterion which is followed by NLM[4] is that an abstract must contain three or more unique valid labels (previously identified and categorized), and one of the labels must be an ending type label (e.g., *CONCLUSIONS*). The five NLM categories are normally manually reviewed and assigned once a year to as many new labels as possible. Currently, NLM has identified 1,949 (August 31, 2012) unique labels and categorized them into one of the five categories. These 1,949 labels make up approximately 98% of all labels and label variations found in the structured abstracts in MEDLINE/PubMed[3]. An example of structured abstract is presented in Table 1.

Several studies have shown that the labels of the structured abstracts can be reassigned effectively based on a Conditional Random Field (CRF) models (Hirohata et al., 2008). On the other hand, it is unclear if these models are as effective on non-structured abstracts (Agarwal and Yu, 2009).

In this paper, we compare several learning algorithms trained on structured abstract data to assign argumentative labels to non-structured abstracts. We performed comparison tests of the trained models both intrinsically on a held out set of the structured abstracts and extrinsically on a set of non-structured abstracts.

The intrinsic evaluation is performed on a data set of held out structured abstracts that have had their label identification removed to model non-structured abstracts. Argumentative labels are assigned to the sentences based on the trained models and used to identify label categorization.

The extrinsic evaluation is performed on a data set of non-structured abstracts on the task of identifying GeneRIF (Gene Into Function) sentences. Argumentative labels are assigned to the sentences based on the trained models and used to perform the selection of relevant GeneRIF sentences.

Intrinsic evaluation shows that argumentative labels can be assigned effectively to structured abstracts. Algorithms that model the argumentative structure, like Conditional Random Field (CRF), seem to perform better than other algorithms. Results show that using the argumentative labels assigned by the learning algorithms improves the performance in GeneRIF sentence selection. On the other hand, models like CRF, which better model the argumentative structure of the structured abstracts, tend to perform below other learning algorithms on the extrinsic evaluation. This shows that non-structured abstracts do not have the same layout compared to structured ones.

2 Related work

As presented in the introduction, one of the objectives of our work is to assign structured abstract labels to abstracts without these labels. The idea is to help in the curation process of existing databases and to improve the efficiency of information access. Previous work on MEDLINE/PubMed abstracts has focused on learning to identify these labels mainly in the Randomized Control Trials (RCT) domain. (McKnight and Srinivasan, 2003) used a Support Vector Machine (SVM) and a linear classifier and tried to predict the labels of MEDLINE structured abstracts. Their work finds that it is possible to learn a model to label the abstract with modest results. Further studies have been conducted by (Ruch et al., 2003; Tbahriti et al., 2005; Ruch et al., 2007) to use the argumentative model of the abstracts. They have used this to improve retrieval and indexing of MEDLINE citations, respectively. In their work, they have used a multi-class Naïve Bayes classifier.

(Hirohata et al., 2008) have shown that the labels in structured abstracts follow a certain argumentative structure. Using the current set of labels used at the NLM, a typical argumentative structure consists of OBJECTIVE, METHODS, RESULTS and CONCLUSION. This notion is somewhat already explored by (McKnight and Srinivasan, 2003) by using the position of the sentence.

More advanced approaches have been used that train a model that considers the sequence of labels in the structured abstracts. (Lin et al., 2006) used a generative model, comparing them to discriminative ones. More recent work has been dealing with Conditional Random Fields (Hirohata et al., 2008) with good performance.

(Agarwal and Yu, 2009) used similar approaches and evaluated the labeling of full text articles with the trained model on structured abstracts. Their evaluation included as well a set of

[4]http://structuredabstracts.nlm.nih.gov/Implementation.shtml

<Abstract> <AbstractText Label="PURPOSE" NlmCategory="OBJECTIVE">To explore the effects of cervical loop electrosurgical excision procedure (LEEP) or cold knife conization (CKC) on pregnancy outcomes.</AbstractText>
<AbstractText Label="MATERIALS AND METHODS" NlmCategory="METHODS">Patients with cervical intraepithelial neoplasia (CIN) who wanted to become pregnant and received LEEP or CKC were considered as the treatment groups. Women who wanted to become pregnant and only underwent colposcopic biopsy without any treatments were considered as the control group. The pregnancy outcomes were observed and compared in the three groups.</AbstractText>
<AbstractText Label="RESULTS" NlmCategory="RESULTS">Premature delivery rate was higher (p = 0.048) in the CKC group (14/36, 38.88%) than in control group (14/68, 20.5%) with a odds ratio (OR) of 2.455 (1.007 - 5.985); and premature delivery was related to cone depth, OR was significantly increased when the cone depth was more than 15 mm. There was no significant difference in premature delivery between LEEP (10 / 48, 20.83%) and the control groups. The average gestational weeks were shorter (p = 0.049) in the CKC group (36.9 +/- 2.4) than in the control group (37.8 +/- 2.6), but similar in LEEP (38.1 +/- 2.4) and control groups. There were no significant differences in cesarean sections between the three groups. The ratio of neonatal birth weight less than 2,500 g was significantly higher (p = 0.005) in the CKC group (15/36) than in the control group (10/68), but similar in the LEEP and control groups.</AbstractText>
<AbstractText Label="CONCLUSION" NlmCategory="CONCLUSIONS">Compared with CKC, LEEP is relatively safe. LEEP should be a priority in the treatment of patients with CIN who want to become pregnant.</AbstractText>
</Abstract>

Table 1: XML example for PMID 23590007

abstracts manually annotated. They found that the performance on full-text was below what was expected. A similar result was found in the manually annotated set. They found, as well, that the abstract sentences are noisy and sometimes the sentences from structured abstracts did not belong with the label they were assigned to.

A large number of abstracts in MEDLINE are not structured; thus intrinsic evaluation of the algorithms trained to predict the argumentative labels on structured abstracts is not completely realistic. Extrinsic evaluation has been previously performed by (Ruch et al., 2003; Tbahriti et al., 2005; Ruch et al., 2007) in information retrieval results evaluating a Naïve Bayes classifier. We have extended this work by evaluating a larger set of algorithms and heuristics on a data set developed to tune and evaluate a system for GeneRIF indexing on a data set containing mostly non-structured abstracts. The idea is that GeneRIF relevant sentences will be assigned distinctive argumentative labels.

A Gene Reference Into Function (GeneRIF) describes novel functionality of genes. The creation of GeneRIF entries involves the identification of the genes mentioned in MEDLINE citations and the citation sentences describing a novel function. GeneRIFs are available from the NCBI (National Center for Biotechnology Information) Gene database[5]. An example sentence is shown below linked to the BRCA1 gene with gene id 672 from the citation with PubMed® identifier (PMID) 22093627:

FISH-positive EGFR expression is associated with gender and smoking status, but not correlated with the expression of ERCC1 and BRCA1 proteins in non-small cell lung cancer.

There is limited previous work related to GeneRIF span extraction. Most of the available publications are related to the TREC Genomics Track in 2003 (Hersh and Bhupatiraju, 2003). There were two main tasks in this track, the first one consisted of identifying relevant citations to be considered for GeneRIF annotation.

In the second task, the participants had to provide spans of text that would correspond to relevant GeneRIF annotations for a set of citations. Considering this second task, the participants were not provided with a training data set. The Dice coefficient was used to measure the similarity between the submitted span of text from the title and abstract of the citation and the official GeneRIF text in the test set.

Surprisingly, one of the main conclusions was that a very competitive system could be obtained by simply delivering the title of the citation as the best GeneRIF span of text. Few teams (EMC (Jelier et al., 2003) and Berkley (Bhalotia et al., 2003) being exceptions), achieved results better than that simple strategy. Another conclusion of the Genomics Track was that the sentence position in the citation is a good indicator for GeneRIF sentence identification: either the title or sentences close to the end of the citation were found to be the best candidates.

Subsequent to the 2003 Genomics Track, there has been some further work related to GeneRIF

[5]http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

sentence selection. (Lu et al., 2006; Lu et al., 2007) sought to reproduce the results already available from Entrez Gene (former name for the NCBI Gene database). In their approach, a set of features is identified from the sentences and used in the algorithm: Gene Ontology (GO) token matches, cue words and sentence position in the abstract. (Gobeill et al., 2008) combined argumentative features using discourse-analysis models (LASt) and an automatic text categorizer to estimate the density of Gene Ontology categories (GOEx). The combination of these two feature sets produced results comparable to the best 2003 Genomics Track system.

3 Methods

As in previous work, we approach the problem of learning to label sentences in abstracts using machine learning methods on structured abstracts. We have compared a large range of machine learning algorithms, including Conditional Random Field. The evaluation is performed intrinsically on a held out set of structured abstracts and then evaluated extrinsically on a dataset developed for the evaluation of algorithms for GeneRIF indexing.

3.1 Structured abstracts data set

This data set is used to train the machine learning algorithms and to peform the intrinsic evaluation of structured abstracts. The abstracts have been collected from PubMed using the query *hasstructuredabstract*, selecting the top 100k citations satifying the query.

The abstract defined within the Abstract attribute is split into several AbstractText tags. Each AbstractText tag has the label *Label* that shows the original label as provided by the journal while the *NlmCategory* represents the category as added by the NLM.

From this set, 2/3 of the citations (66,666) are considered for training the machine learning algorithms while 1/3 of the citations (33,334) are reserved for testing. The abstract paragraphs have been split into sentences and the structured abstract label has been transferred to them. For instance, all the sentences in the INTRODUCTION section are labeled as INTRODUCTION.

An analysis of the abstracts has shown that there are cases in which the article keywords were included as part of the abstract in a *BACKGROUND* section. These were easily recognized by the original label *KEYWORD*. We have removed these paragraphs since they are not typical sentences in MEDLINE but a list of keywords. We find that there are sections like OBJECTIVE where the number of sentences is very low, with less than 2 sentences on average, while RESULTS is the section with the largest number of sentences on average with over 4.5 sentences.

There are five candidate labels identified from the structured abstracts, presented in Table 2. The distribution of labels shows that some labels like CONCLUSIONS, METHODS and RESULTS are very frequent. CONCLUSIONS and METHODS are assigned to more than one paragraph since the number is bigger compared to the number of citations in each set. This seems to happen when more than one journal label in the same citation map to METHODS or CONCLUSION, e.g. PMID: 23538919.

Label	Paragraphs	Sentences
BACKGROUND	53,348	132,890
CONCLUSIONS	101,830	205,394
METHODS	107,227	304,487
OBJECTIVE	60,846	95,547
RESULTS	95,824	436,653

Table 2: Structured abstracts data set statistics

We have compared the performance of several learning algorithms. Among other classifiers, we use Naïve Bayes and Linear Regression, which might be seen as a generative learner versus discriminative (Jordan, 2002) learner. We have used the implementation available from the Mallet package (McCallum, 2002).

In addition to these two classifiers, we have used AdaBoostM1 and SVM. SVM has been trained using stochastic gradient descent (Zhang, 2004), which is very efficient for linear kernels. Table 2 shows a large imbalance between the labels, so we have used the modified Huber Loss (Zhang, 2004), which has already been used in the context of MeSH indexing (Yeganova et al., 2011). Both algorithms were trained based on the one-versus-all approach. We have turned the algorithms into multi-class classifiers by selecting the prediction with the highest confidence by the classifiers (Tsoumakas and Katakis, 2007). We have used the implementation of these algorithms avail-

able from the MTI ML package[6], previously used in the task of MeSH indexing (Jimeno-Yepes et al., 2012).

The learning algorithms have been trained on the text of the paragraph or sentences from the data set presented above. The text is lowercased and tokenized. In addition to the textual features, the position of the sentence or paragraph from the beginning of the abstract is used as well.

As we have seen, argumentative structure of the abstract labels has been previously modeled using a linear chain CRF (Lafferty et al., 2001). CRF is trained using the text features from sentences or paragraphs in conjunction of the abstract labels to perform the label assignment. In our experiments, we have used the implementation available from the Mallet package, using only an order 1 model.

3.2 GeneRIF data set

We have developed a data set to compare and evaluate GeneRIF indexing approaches (Jimeno-Yepes et al., 2013) as part of the Gene Indexing Assistant project at the NLM[7]. The current scope of our work is limited to the human species. The development is performed in two steps described below. The first step consists of selecting citations from journals typically associated with human species. During the second step, we apply Index Section rules for citation filtering plus additional rules to further focus the set of selected citations. Since there was no GeneRIF indexing before 2002, only articles from 2002 through 2011 from the 2011 MEDLINE Baseline [8] (11/19/2010) were used to build the data set.

A subset of the filtered citations was collected for annotation. The annotations were performed by two annotators. Guidelines were prepared and tested on a small set by the two annotators and refined before annotating the entire set.

The data set has been annotated with GeneRIF categories of the sentences. The categories are: Expression, Function, Isolation, Non-GeneRIF, Other, Reference, and Structure. We assigned the GeneRIF category to all the categories that did not belong to Non-GeneRIF. The indexing task is then to categorize the sentences into GeneRIF sentences and Non-GeneRIF ones. Based on their annotation work on the data set, the F-measure for the annotators is 0.81. We have used this annotation for the extrinsic evaluation of GeneRIF indexing.

This data set has been further split into training and testing subsets. Table 3 shows the distribution between GeneRIF and Non-GeneRIF sentences.

Set	Total	GeneRIF	Non-GeneRIF
Training	1987	829 (42%)	1158 (58%)
Testing	999	433 (43%)	566 (57%)

Table 3: GeneRIF sentence distribution

In previous work, the indexing of GeneRIF sentences, on our data set, was performed based on a trained classifier on a set of features that performed well on the GeneRIF testing set (Jimeno-Yepes et al., 2013). Naïve Bayes was the learning algorithm that performed the best compared to the other methods and has been selected in this work as the method to be used to combine the features of the argumentative labeling algorithms.

The set of features in the baseline experiments include the position of the sentence from the beginning of the abstract, the position of the sentence counting from the end of the abstract, the sentence text, the annotation of disease terms, based on MetaMap (Aronson and Lang, 2010), and gene terms, based on a dictionary approach, and the Gene Ontology term density (Gobeill et al., 2008).

4 Results

As mentioned before, we have performed the evaluation of the algorithms intrinsically, given a set of structured abstracts, and extrinsically based on their performance on GeneRIF sentence indexing.

4.1 Intrinsic evaluation (structured abstracts)

Tables 4 and 5 show the results of the intrinsic evaluation for paragraph and sentence experiments respectively. The algorithms are trained to label the paragraphs or sentences from the structured abstracts. The precision (P), recall (R) and F_1 (F) values are presented for each argumentative label. The methods evaluated include Naïve Bayes (NB), Logistic Regression (LR), SVM based on modified Huber Loss (Huber) and AdaBoostM1 (ADA). These methods have been trained on the text of either the sentence or the paragraph, and might include their position feature, indicated with the letter P (e.g. NB P for Naïve Bayes trained

[6] http://ii.nlm.nih.gov/MTI_ML/index.shtml

[7] http://www.lhncbc.nlm.nih.gov/project/automated-indexing-research

[8] http://mbr.nlm.nih.gov

Label		NB	NB P	LR	LR P	ADA	ADA P	Huber	HuberP	CRF
BACKGROUND	P	0.6047	0.6853	0.6374	**0.7369**	0.6098	0.7308	0.5862	0.7166	0.7357
	R	0.5672	0.7190	0.5868	**0.7207**	0.3676	0.7337	0.4984	0.6694	0.7093
	F	0.5854	0.7017	0.6110	**0.7287**	0.4587	0.7323	0.5387	0.6922	0.7223
CONCLUSIONS	P	0.7532	0.8626	0.8365	0.9413	0.6975	0.8862	0.7578	0.9051	**0.9769**
	R	0.8606	0.9366	0.8675	0.9552	0.8246	0.9404	0.7987	0.9340	**0.9784**
	F	0.8033	0.8981	0.8517	0.9482	0.7557	0.9125	0.7777	0.9193	**0.9776**
METHODS	P	0.9002	0.9278	0.9113	0.9396	0.8256	0.9041	0.8668	0.9116	**0.9684**
	R	0.9040	0.9126	0.9294	0.9493	0.8955	0.9250	0.9012	0.9237	**0.9675**
	F	0.9021	0.9201	0.9203	0.9444	0.8591	0.9144	0.8837	0.9176	**0.9680**
OBJECTIVE	P	0.7294	0.7650	0.7167	0.7531	0.6763	0.7565	0.6788	0.7160	**0.7608**
	R	0.6453	0.7190	0.7255	0.7549	0.6937	0.7228	0.6733	0.7365	**0.7759**
	F	0.6848	0.7413	0.7210	0.7540	0.6849	0.7393	0.6761	0.7261	**0.7683**
RESULTS	P	0.8841	0.9106	0.9086	0.9372	0.8554	0.9157	0.8560	0.9122	**0.9692**
	R	0.8414	0.8542	0.8857	0.9216	0.7842	0.8564	0.8447	0.8846	**0.9758**
	F	0.8622	0.8815	0.8970	0.9294	0.8182	0.8851	0.8503	0.8981	**0.9725**
Average	P	0.7743	0.8303	0.8021	0.8616	0.7329	0.8387	0.7491	0.8323	**0.8822**
	R	0.7637	0.8283	0.7990	0.8604	0.7131	0.8357	0.7433	0.8296	**0.8814**
	F	0.7690	0.8293	0.8005	0.8610	0.7229	0.8372	0.7462	0.8310	**0.8818**

Table 4: Intrinsic evaluation of paragraph based labeling

Label		NB	NB P	LR	LR P	ADA	ADA P	Huber	HuberP	CRF
BACKGROUND	P	0.4983	0.6313	0.5558	**0.6862**	0.4779	0.6417	0.5153	0.6495	0.6738
	R	0.4980	0.6921	0.5084	**0.7139**	0.3207	0.6993	0.3372	0.6554	0.7104
	F	0.4981	0.6603	0.5311	**0.6998**	0.3838	0.6693	0.4076	0.6524	0.6916
CONCLUSIONS	P	0.5876	0.7270	0.6794	0.8431	0.5672	0.7651	0.6153	0.7767	**0.8977**
	R	0.7103	0.8388	0.6788	0.8187	0.4998	0.6816	0.5163	0.7213	**0.8671**
	F	0.6431	0.7789	0.6791	0.8307	0.5314	0.7209	0.5615	0.7480	**0.8821**
METHODS	P	0.7857	0.8206	0.8193	0.8549	0.7224	0.7793	0.7343	0.7894	**0.8931**
	R	0.8084	0.8366	0.8427	0.8696	0.7789	0.8152	0.7828	0.8250	**0.8988**
	F	0.7969	0.8285	0.8308	0.8622	0.7496	0.7968	0.7578	0.8068	**0.8960**
OBJECTIVE	P	0.5522	0.6237	0.6032	**0.6696**	0.5497	0.6671	0.5525	0.6259	0.6258
	R	0.4894	0.5530	0.4995	**0.5534**	0.4082	0.4518	0.4479	0.5036	0.5779
	F	0.5189	0.5862	0.5465	**0.6060**	0.4685	0.5388	0.4947	0.5581	0.6009
RESULTS	P	0.8294	0.8517	0.8071	0.8449	0.6903	0.7665	0.6957	0.7877	**0.8892**
	R	0.7517	0.7743	0.8429	0.8679	0.7998	0.8143	0.6957	0.8208	**0.8995**
	F	0.7886	0.8112	0.8246	0.8563	0.7410	0.7897	0.6957	0.8039	**0.8943**
Average	P	0.6506	0.7309	0.6930	0.7797	0.6015	0.7239	0.6226	0.7258	**0.7959**
	R	0.6516	0.7390	0.6745	0.7647	0.5615	0.6924	0.5560	0.7052	**0.7907**
	F	0.6511	0.7349	0.6836	0.7721	0.5808	0.7078	0.5874	0.7154	**0.7933**

Table 5: Intrinsic evaluation of sentence based labeling

with the features from text and the position). The results include those based on CRF trained on the text of either the sentence or the paragraph taking into account the labeling sequence.

CRF has the best performance in both tables, with the differences being more dramatic on the paragraph results. These results are comparable to (Hirohata et al., 2008), even though we are working with a different set of labels. Comparing the remaining learning algorithms, LR performs better than the other classifiers. Both AdaBoostM1 and SVM perform not as well as NB and LR; this could be due to the noise referred to by (Agarwal and Yu, 2009) that appears in the structured abstract sentences. Considering either the paragraph or the sentence text, the position information helps improve their performance.

CONCLUSIONS, METHODS and RESULTS labels have the best performance, which matches the most frequent labels in the dataset (see Table 2). BACKGROUND and OBJECTIVE have worse performance compared to the other labels. These two labels have the largest imbalance compared to the other labels, which seems to negatively impact the classifiers performance.

The results based on the paragraphs outperform the ones based on the sentences. Argumentative structure of the paragraphs seems to be easier, probably due to the fact that individual sentences have been shown to be noisy (Agarwal and Yu, 2009), and this could explain this behaviour.

4.2 Extrinsic evaluation (GeneRIFs)

Extrinsic evaluation is performed on the GeneRIF data set presented in the Methods section. The idea of the evaluation is to assign one of the argumentative labels to the sentences, based on the models trained on structured abstracts, and evaluate the impact of this assignment in the selection of GeneRIF sentences. From the set of machine learning algorithms intrinsically evaluated, we have selected the LR models trained with and without position information (Pos) and the CRF model. The LR and CRF models are used to label the GeneRIF training and testing data with the argumentative labels.

Table 6 shows the results of the extrinsic evaluation. Results obtained with the argumentative label feature and with or without the set of features used in the baseline are compared to the baseline model, i.e. NB and the set of features presented in the Methods section. In all the cases, precision (P), recall (R) and F_1 using the argumentative features improve over the baseline.

The intrinsic evaluation was performed either on sentences or paragraphs. The sentence models perform better than the paragraph based models. We find as well that LR with sentence position performs slightly better than when combined with the baseline features, with higher recall but lower precision. Contrary to the intrinsic results, LR performs better than CRF, even though both outperform the baseline. This means that non-structured sentences do not necessarily follow the same argumentative structure as the structured abstracts.

Label	P	R	F
Baseline	0.6210	0.6605	0.6405
LR Par	**0.7235**	0.6767	0.6993
LR Par + Base	0.7184	0.8014	**0.7576**
LR Par Pos	0.5978	**0.8891**	0.7149
LR Par Pos + Base	0.6883	0.8060	0.7426
LR Sen	0.7039	0.7852	0.7424
LR Sen + Base	**0.7325**	0.7968	0.7633
LR Sen Pos	0.7014	**0.9007**	**0.7887**
LR Sen Pos + Base	0.7222	0.8406	0.7769
CRF Par	0.6682	0.6744	0.6713
CRF Par + Base	**0.7036**	**0.8060**	**0.7513**
CRF Sen	0.6536	**0.8499**	0.7390
CRF Sen + Base	**0.7134**	0.7875	**0.7486**

Table 6: GeneRIF extrinsic evaluation

5 Discussion

Results show that it is possible to automatically predict the argumentative label of the structured abstracts and to improve the performance for GeneRIF annotation. Intrinsic evaluation shows that paragraph labeling is easier compared to sentence labeling, which might be partly due to the noise in the sentences as identified by (Agarwal and Yu, 2009). The excellent performance for paragraph labeling was already shown by previous work (Hirohata et al., 2008) while sentence labeling issues for structured abstracts was previously introduced by (Agarwal and Yu, 2009). In both intrinsic tasks, adding the position of the paragraph or sentence improves the performance of the learning algorithms.

Extrinsic evaluation shows that, compared to the baseline features for GeneRIF annotation, adding argumentative labeling using the trained models improves its performance, which is close to the human performance reported in the Methods section. On the other hand, we find that the CRF models show lower performance compared to the LR models. From the LR models, the position of the sentence or paragraph seems to have better performance.

In addition, the LR model trained on the sentences performs better compared to the model trained on the paragraphs. This might be partly due to the fact that sentence based models seem to be better suited than the paragraph based ones as might have been expected. The fact that the CRF models performance is below the LR models denotes that the structured abstracts seem to follow a pattern that is different in the case of non-structured abstracts. Looking closer at the assigned labels, the LR models tend to assign more CONCLUSIONS and RESULTS labels to the GeneRIF sentences compared to the CRF ones.

6 Conclusions and Future Work

We have presented an evaluation of several learning algorithms to label abstract text in MEDLINE/PubMed with argumentative labels, based on MEDLINE/PubMed structured abstracts. The results show that this task can be achieved with high performance in the case of labeling the paragraphs but this is not the same in the case of sentences. This intrinsic evaluation was performed on structured abstracts, and in this set the CRF models seem to perform much better compared to the

other models that do not use the labeling sequence.

On the other hand, when applying the trained models to MEDLINE/PubMed non-structured abstracts, we find that the extrinsic evaluation of these labeling on the GeneRIF task shows lower performance for the CRF models. This indicates that the structured abstracts follow a pattern that non-structured ones do not follow. The extrinsic evaluation shows that labeling the sentences with argumentative labels improves the indexing of GeneRIF sentences. The argumentative labels help identifying target sentences for the GeneRIF indexing, but more refined labels learned from non-structured abstracts could provide better performance. An idea to extend this research would be evaluating the latent discovery of section labels and to apply this labeling to the proposed GeneRIF task and to other tasks, e.g. MeSH indexing. Latent labels might accommodate better the argumentative structure of non-structured abstracts.

As shown in this work, the argumentative layout of non-structured abstracts and structured abstracts is not the same. There is still the open question if there is any layout regularity in the non-structured abstracts that could be exploited to improve information access.

7 Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

This work was also supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

References

S Agarwal and H Yu. 2009. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*, 25(23):3174–3180.

A R Aronson and F M Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

G. Bhalotia, PI Nakov, A S Schwartz, and M A Hearst. 2003. BioText team report for the TREC 2003 genomics track. In *Proceedings of TREC*. Citeseer.

J Gobeill, I Tbahriti, F Ehrler, A Mottaz, A Veuthey, and P Ruch. 2008. Gene Ontology density estimation and discourse analysis for automatic GeneRiF extraction. *BMC Bioinformatics*, 9(Suppl 3):S9.

W Hersh and R T Bhupatiraju. 2003. TREC genomics track overview. In *TREC 2003*, pages 14–23.

K Hirohata, Naoaki Okazaki, Sophia Ananiadou, Mitsuru Ishizuka, and Manchester Interdisciplinary Biocentre. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proc. of 3rd International Joint Conference on Natural Language Processing*, pages 381–388.

R Jelier, M Schuemie, C Eijk, M Weeber, E Mulligen, B Schijvenaars, B Mons, and J Kors. 2003. Searching for GeneRIFs: concept-based query expansion and Bayes classification. In *Proceedings of TREC*, pages 167–174.

A Jimeno-Yepes, J G Mork, D Demner-Fushman, and A R Aronson. 2012. A One-Size-Fits-All Indexing Method Does Not Exist: Automatic Selection Based on Meta-Learning. *Journal of Computing Science and Engineering*, 6(2):151–160.

A Jimeno-Yepes, J C Sticco, J G Mork, and A R Aronson. 2013. GeneRIF indexing: sentence selection based on machine learning. *BMC Bioinformatics*, 14(1):147.

S Jonnalagadda, G D Fiol, R Medlin, C Weir, M Fiszman, J Mostafa, and H Liu. 2012. Automatically extracting sentences from medline citations to support clinicians' information needs. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pages 72–72. IEEE.

A Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841.

J D Lafferty, A McCallum, and F Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

J Lin, D Karakos, D Demner-Fushman, and S Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 65–72. Association for Computational Linguistics.

Z Lu, K B Cohen, and L Hunter. 2006. Finding GeneRIFs via gene ontology annotations. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 52. NIH Public Access.

Z Lu, K B Cohen, and L Hunter. 2007. GeneRIF quality assurance as summary revision. In *Pacific Symposium on Biocomputing*, page 269. NIH Public Access.

A McCallum. 2002. Mallet: A machine learning for language toolkit. *URL http://mallet.cs.umass.edu*.

L McKnight and P Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *AMIA Annual Symposium Proceedings*, volume 2003, page 440. American Medical Informatics Association.

Y Mizuta, A Korhonen, T Mullen, and N Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487.

A M Ripple, J G Mork, L S Knecht, and B L Humphreys. 2011. A retrospective cohort study of structured abstracts in MEDLINE, 1992–2006. *Journal of the Medical Library Association: JMLA*, 99(2):160.

A M Ripple, J G Mork, J M Rozier, and L S Knecht. 2012. Structured Abstracts in MEDLINE: Twenty-Five Years Later.

P Ruch, C Chichester, G Cohen, G Coray, F Ehrler, H Ghorbel, and V Müller, Hand Pallotta. 2003. Report on the TREC 2003 experiment: Genomic track. *TREC-03*.

P Ruch, A Geissbuhler, J Gobeill, F Lisacek, I Tbahriti, A Veuthey, and A R Aronson. 2007. Using discourse analysis to improve text categorization in MEDLINE. *Studies in health technology and informatics*, 129(1):710.

I Tbahriti, C Chichester, F Lisacek, and P Ruch. 2005. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the MEDLINE digital library. In *International Journal of Medical Informatics*. Citeseer.

G Tsoumakas and I Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

L Yeganova, Donald C Comeau, W Kim, and J Wilbur. 2011. Text mining techniques for leveraging positively labeled data. In *Proceedings of BioNLP 2011 Workshop*, pages 155–163. Association for Computational Linguistics.

T Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM.