Using Discourse Analysis to Improve Text Categorization in MEDLINE

Patrick Ruch[a], Antoine Geissbühler[a], Julien Gobeill, Frederic Lisacek[c], Imad Tbahriti[ab], Anne-Lise Veuthey[b], Alan R. Aronson[d]

[a]*Medical Informatics Service, University and Hospital of Geneva, Geneva, Switzerland*
[b]*Swiss-Prot, Swiss Institute of Bioinformatics, Geneva, Switzerland*
[c] *Protein Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland*
[d] *Lister Hill Center, National Library of Medicine, Bethesda, MD, USA*

Abstract

PROBLEM: Automatic keyword assignment has been largely studied in medical informatics in the context of the MEDLINE database, both for helping search in MEDLINE and in order to provide an indicative "gist" of the content of an article. Automatic assignment of Medical Subject Headings (MeSH), which is formally an automatic text categorization task, has been proposed using different methods or combination of methods, including machine learning (naïve Bayes, neural networks…), linguistically-motivated methods (syntactic parsing, semantic tagging, or information retrieval. METHODS: In the present study, we propose to evaluate the impact of the argumentative structures of scientific articles to improve the categorization effectiveness of a categorizer, which combines linguistically-motivated and information retrieval methods. Our argumentative categorizer, which uses representation levels inherited from the field of discourse analysis, is able to classify sentences of an abstract in four classes: PURPOSE; METHODS; RESULTS and CONCLUSION. For the evaluation, the OHSUMED collection, a sample of MEDLINE, is used as a benchmark. For each abstract in the collection, the result of the argumentative classifier, i.e. the labeling of each sentence with an argumentative class, is used to modify the original ranking of the MeSH categorizer. RESULTS: The most effective combination (+2%, p<0.003) strongly overweights the METHODS section and moderately the RESULTS and CONCLUSION section. CONCLUSION: Although modest, the improvement brought by argumentative features for text categorization confirms that discourse analysis methods could benefit text mining in scientific digital libraries.

Keywords: Text Mining; Abstracting and Indexing; Information Storage and Retrieval; Natural Language Processing; Libraries, Medical; Artificial Intelligence.

Introduction

Systems for text mining are becoming increasingly important in biomedicine because of the exponential growth of written information in this domain. The mass of scientific literature needs to be filtered and categorized to provide for the most efficient use of the data. The problem of accessing this increasing volume of data demands the development of systems that can extract pertinent information from unstructured texts, hence the importance of key words extraction, as well as key sentence extraction. While the former task has been largely addressed in text categorization studies [1], the latter has been more rarely studied. In this report, we propose to merge the two ideas to improve the first task. We intend to use sentence extraction and sentence ranking methods to improve text categorization in MEDLINE based on Medical Subject Headings (MeSH). Selecting a set of appropriate sentences likely to improve MeSH assignment is a complex task because, more than key words, the importance of a sentence is dependent on the point of view of every reader. However, as for key words, which can be more or less comprehensively listed in a controlled vocabulary, we believe it is possible to propose task-specific criteria, which can help to define such sentences. Our model is based on the implicit discourse-level information found in scientific reports; in particular, our model uses the argumentative sections as they often appear – sometimes explicitly but usually implicitly – in abstracts. The argumentative categories we are experimenting to rank the sentences originate from professional guidelines (ANSI Z39.14-1979). Indeed, articles in experimental sciences tend to respect strict argumentative patterns with at least four sections: purpose-methods-results-conclusion. These four moves –leaving aside minor variation of labels- are reported to be very stable across different scientific genres and experimental sciences in general (chemistry, anthropology, computer sciences, linguistics…) [4]. They are also confirmed in biomedical abstracts and articles [5][6][29]. Following recent trends, which show that argumentative criteria can be useful to improve various text mining and information retrieval applications such as related articles search [7], blind feedback [27] for ad hoc search, or feature and passage selection for automatic indexing in full-text articles [30], our objective is to evaluate the benefit of using discourse analysis representation levels for keyword assignment in MEDLINE.

Background: Argumentative Categorization

In this section, we present a set of background methods and already reported results, which are useful to understand the rationales supporting our approach. Our argumentative categorizer is formally defined as a mono class classifier: for each piece of text the system will have to decide whether it is a PURPOSE, or a METHOD, or a RESULT, or a CONCLUSION. Sentences are natural candidate segments for such a classification [8] because they are more self-contained than phrases; however anaphoric phenomena may demand larger segments.

Summarization has a related task

Modern summarization systems use annotated corpora in order to acquire appropriate knowledge; based on textual features summarization tools are able to conduct general summarization tasks. Thus, Kupiec et al. [10] report that in abstracts produced by professionals, 80% of sentences are also found in the source document. In such systems, the complex abstracting task is recast into a more modest sentence selection problem. To do so, experts identify a set of relevant sentences from large corpora; these sentences are then used to train the learning system. Complementary to machine learning approaches, Teufel et al. in [23], who design a task very similar to our one, combine a set of manually crafted triggered expressions, such as *finally, we have shown that, we conclude that…* to classify sentences into seven argumentative classes.

Basic classifiers, features selection and weighting

Choosing *a priori* an appropriate classifier for a given task is a fairly difficult task; therefore, empirical comparisons are often necessary. Among state-of-the-art classifiers for text categorization, such as k-NN [11], SVM [12][28], neural networks [13], and rule-induction systems [14], Bayesian classifiers [15] show a linear complexity, while most top performing algorithms have quadratic complexity; therefore they are often more adapted for rapid application development and exploratory studies [16] [24]. The basic features in text categorization are usually word-based. Possible variants are stems, which often implies stop-word removal, and/or sequences of stems, such as bi- or trigrams of stems. Examples of stemming algorithms are provided in Table 1.

Table 1: String Normalization Methods

Token	Lovins	Porter	S-stemmer
genetic	genet	genet	genetic
genetically	genet	genetically	genetically
genetics	genet	genet	genetic
gene	gene	gene	gene
genes	gene	gene	gene
homogeneous	gene	homogen	homogeneous
plaid	plai	plai	plai
play	plai	plai	plai

Our preliminary studies confirmed that elaborate string normalization and stop-word removal strategies such as Porter and Lovins did not outperform simpler approaches, such as plural normalization, which process morphological variations of plural forms (-s, -ies). This strategy appears sufficient to help the classifiers to generalize without removing interesting features, such as verb tense (as suggested in [4]), which is usually removed by stemming (cf. Table 1).

The last step concerns feature weighting. Indeed naive Bayes classifiers combine the log-likelihood of each feature in order to select the most probable category in the category space; however the real frequency observed in training corpora can follow different refinement and smoothing processes, known as feature weighting [11].

PURPOSE: While gemcitabine (GEM) is widely accepted for the treatment of advanced pancreatic cancer, capecitabine (CAP) has shown single agent activity and promising efficacy in combination with GEM. [...] METHODS: Patients had advanced pancreatic adenocarcinoma, no prior systemic chemotherapy other than that given concurrently with radiation therapy, at lease one measurable disease, and adequate organ functions. [...] RESULTS: The objective RR among 45 patients was 40.0% (95% CI; 25.1-54.9), including 1CR (2.2%). The median TTP and OS were 5.4 months (95% CI; 1.8-9.0) and 10.4 months (95% CI; 6.2-14.5), respectively. [...] CONCLUSIONS: The combination of GEM with dose escalated 14-day CAP is well tolerated and offers encouraging activity in the treatment of advanced pancreatic cancer. [...]

Figure 1: Partial example of an explicitly structured abstracts in MEDLINE.

We tested 3 weighting methods: tf-idf (term frequency-inverse document frequency), chi-2, and df-thresholding (only features appearing frequenty in each class are selected). Our conclusion is that chi-2 and df-thresholding perform similarly, while tf-idf weighting should be avoided. Indeed, tf-idf weighting is appropriate for weighting content-bearing features, while argumentative content is supported by functional words. Three types of features are linearly combined to get a final probability ranking per class: stems, stem bigrams and stem trigrams. As in [15], length normalization of sentences as been applied in order to overcome biases introduced by too long or too short sentences.

Training and Test Data

In text classification tasks, two types of strategies are competing: expert-driven and data-driven approaches. While the former, which rely on a domain expert, are often time and labour-intensive, the latter are directly dependent on the availability of large training sets. Fortunately, training data for our task can be acquired in a cheap way. Most abstracts in MEDLINE are unstructured (i.e. provided without explicit argumentative markers, such as METHODS, PURPOSES…); but fortunately, a significant fraction of these abstracts contain explicit argumentative markers. Using PubMed and its Boolean query interface, we collected a set of 12000 MEDLINE citations containing strings such as "PURPOSE:", "METHODS:", "RESULTS", "CONCLUSION:". This fully automatic data collection process introduces some argumentative noise since some of the explicit markers gather additional argumentative content. Thus, explicit markers such as

"BACKGROUND AND PURPOSES:" were also collected as pure "PURPOSES:" markers by this simple method. The collection was then split into two sets:

- set A (10800 abstracts) was used for training and validation purposed,
- set B (1200) was used for the final assessment.

In addition to sets A and B, we also collected a smaller set (C) of marker-free abstracts (100 items). Then, two human annotators were asked to manually annotate this set, using the four selected argumentative classes. In contrast with the automatically acquired sets, here we do not assume that argumentative segments and sentences are overlapping items. As shown in Figure 2, some sentences in set C receive more than one label, because they may express two different argumentative moves in the same sentence. In such cases, we do not attempt to identify segment boundaries (as explored in [18] and [19]) and instead ask the system to provide any of the relevant classes. The interannotator agreement on the C set for argumentative segments is 0.81, when measured by kappa statistics, which indicates that agreement is good.

As mentioned above, our goal is to extract conclusion sentences, but because the information is available in our training data, this binary task has been modified into a four-class problem: {PURPOSE, METHODS, RESULTS, CONCLUSION}. We expect that working with more classes will help the system to discriminate between classes that have been reported to be lexically similar [4][21], such as PURPOSE and CONCLUSION. In the data sets used for the evaluation (B and C), explicit argumentative markers have been removed.

<CONCLUSION> Skin surface proteolytic activity in the living animal was detected </CONCLUSION> <METHODS> by a sensitive, non-invasive methodol. Developed in our laboratory</METHODS>. <METHODS>A non-leaky well was constructed on the shaved back of an anesthetized guinea pig. The well contained the reaction mixture including the substrate 125I-S-carboxymethylated <GPN> insulin B-chain</GPN> (<GPN>ICMI</GPN>)</METHODS>. <RESULTS>The proteolytic activity was shown to be time-dependent. The activity was strongly inhibited by <ASP_GPN>pepstatin A</ASP_GPN>, indicating the involvement of aspartic proteinase(s) such as <ASP_GPN>cathepsin D</ASP_GPN> and/or <ASP_GPN>E</ASP_GPN>. Pretreatment of the skin with propylene glycol blocked the proteolytic ctiveity</RESULTS>. <CONCLUSION>The present study demonstrates the presence of proteolytic activity located on skin surface<CONCLUSION> <METHODS>using a unique, non-invasive method for in situ proteinase detn. In the living animal</METHODS>.

Figure 2: Example of an automatically structured abstracts. Four argumentative classes are annotated with XML tags; Gene and Protein Names (GPN and ASP_GPN), are also annotated.

Combining positions of segments

Optionally, we also investigated the sentence position's impact on the classification effectiveness through assigning a relative position to each sentence. Thus, if there were ten sentences in an abstract: the first sentence has a relative position of 0.1, while the sentence in position 5 receives a relative position of 0.5, and the last sentence has a relative position of 1. The following distributional heuristics are encoded in a distributional model: 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.

Categorization effectiveness

Table 2 indicates the categorization effectiveness of our argumentative categorizer. In Table 2, we evaluate the effect of positional information on the categorizer. We also evaluate the performance of the system on explicitly and non-explicitly structured abstracts. Finally, with an F-score (i.e the harmonic mean, with recall and precision having the same importance; cf [22]) above 85%, recall and precision measures are competitive with the trigger-based approach proposed by Teufel et al [23] (F-score ~ 68%), and the SVM learner used in McKnight and Srinivasan [25] (F-score ~ 80%). While recall exhibits excellent levels of performance, precision could still be improved. Thus, expected conclusion segments are well classified, but we observe that some non-conclusion sentences are also classified as conclusion (false positives).

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%

Table 2: Confusion matrices expressing the classification effectiveness of the argumentative categorizer, with and without using positional information.

Materials and Methods

To conduct our key word assignment experiments and evaluations, we used the OHSUGEN collection [25], which contains more than 300,000 MEDLINE records. From these records, we randomly selected 500 citations with abstracts and keywords to tune our system and 1000 citations with abstracts and keywords to provide the final results. All experiments were conducted with the MeSH 2000 version, which roughly corresponds to the time period covered by the collection. Our baseline system, which was originally developed for the Bio-Creative challenge 2003 to perform text categorization tasks with Gene Ontology categories, combine a vector space ranker and a pattern matcher. Both stems (Porter) and linguistically-motivated features were used by the system [35]. The system achieved competitive results in the context of the Bio-

Cretative challenge; cf. [34] for a comparative presentation and [35] for a comprehensive presentation and evaluation.

In our MeSH categorizer, the ranking was based on the categorization status value returned by the system. Thus, the best candidate categories (i.e. the relevance estimate) obtained the highest score, which directly expresses a similarity between a category and the input abstract. To overweight a given argumentative section, we simply modified the term frequency of the feature in the abstract. Thus, if we wanted to emphasize features appearing in the METHODS section, then every feature in a sentence classified as METHODS by the argumentative categorizer received a boosting factor. The fine-tuning of the optimal model, which includes the calculus of the optimal boosting factor, was based on direct search. We worked with integer values, ranging from 1 to 7 and we strove to maximize the mean average precision of our system, which is the best metric to express the full ordering skill of the system.

Results

Figure 3 provides the expected list of Medical Subject Headings for the abstract in Figure 1.

Assigned MeSH:
Adult; Cystic Fibrosis/genetics; DNA/analysis*; Genotype; HLA-DQ Antigens/genetics; Humans; Research Support, Non-U.S. Gov't; Reverse Transcriptase Polymerase Chain Reaction; Sequence Analysis, DNA/methods*; Spectrum Analysis, Raman

Top-12 categories proposed by the categorizer:
12. dna mutational analysis
11. dna
10. genetics
9. fibrosis [not in NP index]
8. alleles
7. sequence analysis, dna
6. genotype
5. mutation
4. fluorescence
3. oligonucleotides
2. polymerase chain reaction
1. cystic fibrosis

Figure 3: Expected and predicted Medical Subject Headings for abstracts in Figure 1 (PMID: 12404725). Major headings are listed with a star. We do not separate between major and minor headings. Qualifiers are ignored in our benchmark.

We observe that several headings, which relates to age, or population-related groups (*Adult*), and to grants (*Research Support, Non-U.S. Gov't*) cannot be inferred from the abstract; therefore, as is well known, both the average precision and the recall of MeSH categorizers are generally low. In contrast, some categories, in particular major headings, which are more important for annotators, can be relevantly inferred from abstracts. Table 3 provides results of the final evaluation. In particular, we observe that the best combination emphasizes the PURPOSE section (x5 times) and the METHODS and RESULTS sections (each with a multiplicative factor of 2). The overall resulting improvement (+2%) is statistically significant, although quite modest (p<0.003) [32]. We also see that another simpler yet quite effective combination can be

obtained by boosting just the PURPOSE section (x3 times). It is also interesting to observe that the best combination regarding the mean average precision (MAP), i.e. 0.217, is not the best one for precision at high ranks, as expressed by the Precision at recall = 0. In contrast, the best precision at high ranks (0.93) is achieved by trading recall for precision. Thus, in that setting, only 3064 relevant categories are proposed while 3068 were proposed with the baseline system.

Table 3: Results of the argumentative boosting on the effectiveness of the MeSH categorizer (C=CONCLUSION, P=PURPOSE; M=METHODS, R=RESULTS, MAP= Mean average precision).

Parameters Kc=C,P,M,R	Relevant Retrieved	Prec. at Rec.=0	MAP
Kc=1,5,2,2	3068	0.927 (+0.8%)	**0.217 (+2%)**
Kc=1,3,1,1	3064	0.93 (+1.1%)	0.216 (+1.4%)
Kc=1,1,1,1 (Baseline)	3068	0.92 (100%)	0.213 (100%)

Conclusion

Our results (precision at high ranks ~ 93%) suggest that argumentative contents as available in abstracts stored in digital libraries are helpful for text categorization tasks, such as automatic assignment of Medical Subject Headings (MeSH). The reported improvement is statistically significant. Although modest (+2%), the improvement confirms that discourse analysis methods are useful for a growing number of text mining applications. Indeed, while attempts to apply linguistically-motivated approaches based on syntactic tools (part-of-speech tagging, shallow or deep parsing) to information retrieval and text categorization were rather inconclusive, methods inherited from discourse analysis could provide a more scalable and dependable improvement. Finally, it would be interesting to evaluate the benefit of argumentative features using more elaborate approaches such as those working with novelty detection [33], full-text articles [30], or with more advanced learners [31].

Acknowledgements

References

[1]   F Sebastiani, Machine learning in automated text categorization. *ACM Computing Surveys* 34(1): 1-47, 2002.

[2]   A Aronson, O Bodenreider, H Chang, S Humphrey, J Mork, S Nelson, T Rindflesch, W Wilbur, The NLM Indexing Initiative. *Proc AMIA Symp*. 2000;:17-21.

[3]   P Ruch, R Baud, and A Geissbühler. Learning-free Text Categorization*, AIME* 2003, LNCS 2780.

[4]   C Orasan, Patterns in scientific abstracts, in Proceedings of Corpus Linguistics Conference, Lancaster University, Lancaster, UK, pp. 433 - 443, 2001.

[5]   J Swales, *Genre Analysis: English in academic and research settings*, Cambridge University Press, 1990.

[6]   F Salanger-Meyer, Discoursal movements in medical English abstracts and their linguistic exponents: a genre analysis study, *INTERFACE: Journal of Applied Linguistics* 4(2), 1990, pp. 107 - 124

[7]   Tbahriti, C Chichester, F Lisacek and P Ruch. Using Argumentation to Retrieve Articles with Similar Citations: an Inquiry into Improving Related Articles Search in the MEDLINE Digital Library. *Int J Med Inform*. 2006 Jun;75(6):488-95

[8]   Ruch, C Chichester, G Cohen, G Coray, F Ehrler, H Ghorbel, H Müller, V Pallotta. Report on the TREC 2003 Experiment: Genomic Track, *TREC* 2003.

[9]   J Kupiec, J Pedersen and F Chen: A Trainable Document Summarizer, *SIGIR* 1995, p. 55-60.

[10] Y Yang, An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval* 1 (1999) p. 67-88.

[11] S Dumais, J Platt, D Sahami: Inductive learning algorithms and representations for text categorization. In: CIKM, ACM (1998) p. 148-155.

[12] K Ming Adam Chai, Hai Leong Chieu, Hwee Tou Ng, Bayesian online classifiers for text classification and filtering, SIGIR 2002, p. 89-96.

[13] D Beeferman, A Berger, and J Lafferty, Statistical models for text segmentation. *Machine Learning*, (34):177-210, 1999. Special Issue on Natural Language Learning (C. Cardie and R. Mooney, eds).

[14] D Lewis and M Ringuette, M.: A comparison of two learning algorithms for text categorization. In: *SDAIR*. 1994, p. 81-93

[15] P Domingos and M Pazzani, On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, 29 (2-3), 1997, p. 103-130.

[16] Y Yang, J Pedersen, A Comparative Study on Feature Selection in Text Categorization, *ICML*, 14th International Conference on Machine Learning, 1997, p. 114-121.

[17] K Nigam, J Lafferty and A McCallum. Using maximum entropy for text classification. In IJCAI Workshop on Machine Learning for Information Filtering, pages 61-67, 1999.

[18] M Hearst, Multi-Paragraph Segmentation of Expository Text. Proceedings of the 32nd Meeting of the Association for Computational Linguistics, 1994, p. 79-94.

[19] J Reynar and A Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. *ANLP*, pages 16-19, 1997. ACL

[20] S Teufel and M Moens, Argumentative classification of extracted sentences as a first step towards flexible abstracting. Mani, M. Maybury (eds), *Advances in automatic text summarization*, MIT, 1999.

[21] C van Rijsbergen, Information Retrieval. Buttersworth, 1979.

[22] S Teufel and M Moens, Sentence Extraction and rhetorical classification for flexible abstracts, AAAI Spring Symposium on Intelligent Text summarization, 1998, p. 89-97.

[23] P Ruch, L Perret and J Savoy. Feature Combination for Extraction of Gene Functions. ECIR 2005. Springer LNCS, 112-126.

[24] L McKnight and P Srinivasan. Categorization of Sentence Types in Medical Abstracts. *AMIA* 2003.

[25] P Ruch, R Baud, P Bouillon, and G Robert. Minimal Commitment and Full Lexical Disambiguation: Balancing Rules and Hidden Markov Models. CoNLL-2000, pages 111-115.

[26] W Hersh. Report on the TREC 2004 Genomics track. SIGIR Forum (2005)

[27] P Ruch, I Tbahriti, J Gobeill, AR Aronson: Argumentative Feedback: A Linguistically-Motivated Term Expansion for Information Retrieval. *ACL* 2006

[28] Y Mizuta and N Collier: Zone Identification in Biology Articles as a Basis for Information Extraction, In Porceedings of JNLPBA (NLPBA/BioNLP), 28-29 August, Geneva, Switzerland.

[29] F Couto, M Silva and P Coutinho, P. FIGO: Findings GO Terms in Unstructured Text (2004). BioCreative Notebook Papers.

[30] C Gay, M Kayaalp, A Aronson. Semi-Automatic Indexing of Full Text Biomedical Articles. AMIA 2005 (2005).

[31] M Ruiz and P Srinivasan. Hierarchical text categorization using neural networks. Information Retrieval, 5(1), pp. 87-118, 2002.

[32] J Zobel. How reliable are large-scale information retrieval experiments? ACM-SIGIR∨, pages 307—314, 1998.

[33] F Lisacek, C Chichester, A Kaplan, S Agnes. Discovering paradigm shift patterns in biomedical abstracts: Application to neurodegenerative diseases. *SMBM* 2005.

[34] F Ehrler, A Jimeno Yepes, A Geissbühler and P Ruch. Data-poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot, *BMC Bioinformatics*, Special Issue on BioCreative: A Critical Assessment of Text Mining Methods in Molecular Biology, vol. 6 (suppl. 1), 2005.

[35] P Ruch: Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics* 22(6): 658-664 (2006).

Address for correspondence

Patrick Ruch, University Hospitals of Geneva
24 Micheli du Crest, CH-1211 Geneva, Switzerland
@: patrick.ruch@sim.hcuge.ch , tel: +41 22 372 61 64