Towards a systematic structuring of MEDLINE abstracts

Mikaela Keller[a], Natalia GRABAR[b]
[a]*CNRS LIFL UMR8022, Mostrare INRIA, Université Lille 1&3, France*
[b]*CNRS STL UMR 8163, Université Lille 1&3, France*

Abstract. The bibliographical data are growing and their access becomes more difficult. A solution consists into providing structured abstracts in which their design is indicated explicitly. Our objective is to assign automatically the abstract sentences to relevant sections. We exploit machine learning algorithms and different sets of features (position of sentences, their forms and lemmas). The obtained f-measure is between 0.60 and 0.80, while one section (Objective) remains difficult to categorize. Our work indicates the feasibility of the abstract structuring on large amount of bibliographical data.

Keywords. Artificial Intelligence; Information Storage and Retrieval/methods; Libraries, Digital; MEDLINE; Natural Language Processing; Periodicals as Topic

1. Introduction

Several medical journals have been engaged on the way to provide more informative and structured abstracts [1], although the main amount of bibliographical data is still available within unstructured format. Within structured abstracts, the authors explicitly indicate their design. The standardized structure of the abstracts is known as the IMRAD structure [2]: introduction, methods, results, discussion. When available, structured abstracts facilitate several tasks usually performed by researchers and clinicians: find articles that are scientifically sound and relevant; ease the peer review before publication; allow more precise computerized literature searches [1]. The structured abstracts open new possibilities to the automatic systems from Natural Language Processing. Importance of the availability of the structured abstracts has been notified long ago [1,3-4], however little previous work exists. It is done within the RCT [5-6] and genomics [7] areas. We propose to decipher the structure of the MEDLINE abstracts with machine learning algorithms and while processing them independently from their clinical areas and journals.

2. Material and Methods

Material. Pubmed abstracts are the main material. We collected 67,303 different PMIDs which abstracts are already structured. This structure may be native or it may be generated during their integration within the MEDLINE. The five sections explicitly mentioned are: Background, Objective, Methods, Results, and Conclusions. The collected PMIDs are

provided by 4,254 different sources. This structuring is our reference data.

Methods. Our method consists of three main steps: (1) the preprocessing of the abstracts and the preparing of the features, (2) the training of the learning algorithms and (3) the test of the trained models and their evaluation against the reference data. During the preprocessing step, we extract from the XML MEDLINE files the raw data labeled by the section names (Background, Objective, Methods, Results, Conclusions). The abstract text is segmented into ordered sentences. The sentences are processed with a POS-tagger [8] and each word is assigned to a grammatical category and lemmatized. We exploit different types of data for the categorization: position of the sentences within abstracts, forms and lemmas of the words. The second step is dedicated to the training of the learning algorithm. 66% of the sentences are exploited as the training data and the remaining 33% (from which we remove the information on their structure) are exploited for tests. We exploit the SVM algorithm [9] with a linear kernel. Each sentence to be categorized is represented as a vector of features which can be: its position, its set of forms or of lemmas weighted with TFIDF. These features may be processed separately or combined. The aimed categories correspond to the five sections (Background, Objective, Methods, Results, and Conclusion). The models are optimized so that the number of sentences incorrectly included in the category and the number of sentences of the category that we are missing to identify on the training set reach the breakeven point for each section. The evaluation is performed within each category according to three classical measures: precision *P*, recall *R* and the f-measure *F*. The baseline is performed with only the position information.

3. Results and Discussion

Model	*Relative pos.*			*Absolute pos.*			*Forms*			*Lemmas*		
	F1	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*
Background	0.28	0.17	1	0.73	0.74	0.73	0.53	0.49	0.56	0.51	0.48	0.55
Background+pos.							0.80	0.78	0.83	0.80	0.78	0.83
Objective	0.06	0.04	1	0	0	0	0.07	0.07	0.07	0.03	0.05	0.02
Objective+pos.							0.17	0.13	0.25	0.16	0.13	0.22
Methods	0.39	0.24	1	0.55	0.55	0.56	0.62	0.61	0.61	0.61	0.60	0.62
Methods+pos.							0.65	0.65	0.65	0.64	0.64	0.64
Results	0.55	0.38	1	0.63	0.60	0.67	0.58	0.68	0.51	0.57	0.67	0.49
Results+pos.							0.58	0.67	0.51	0.57	0.67	0.50
Conclusion	0.31	0.18	1	0.54	0.56	0.53	0.37	0.26	0.61	0.32	0.19	0.94
Conclusion+pos.							0.81	0.79	0.83	0.81	0.79	0.83

Table 1: Evaluation results (pos. = position).

In table 1, we indicate the results obtained for each category (indicated in the first column of the table). The two first sets of the results correspond to the baseline obtained with absolute and relative positions. The relative position favors the recall (always perfect), while the precision is low. The strongest results are obtained for the Results section sentences. The performances with the absolute position of the sentences are higher: f-measure goes from 0.54 to 0.73. The Objective section shows null performance. Table 1 indicates then the performances for the categorization of the sentences when forms and lemmas are considered: the difference is small although the forms are more efficient than the lemmas. When only the forms and lemmas are processed, the f-measure can reach up to 0.60. The efficiency of these features is increased by 0.10 (Objectives), 0.30 (Background) and up to 0.50 (Conclusion) when they are combined with the position of the sentences. On the whole, our performances reach up to 0.60 and 0.80. The obtained results are slightly inferior to some previous work [6,7] (f-measure between 0.70 and 0.90) but are comparable to other works [6]. The comparison remains difficult because we work on a greater variety of abstracts, while in previous work the abstracts belong to a given area.

4. Conclusion and Perspectives

We presented an experience on structuring of the MEDLINE abstracts. We rely on the machine learning algorithms and on different kinds of features (position of the sentences, their forms and lemmas, and the combinations of them). The combination of features provides the best results: 0.60 to 0.80 for the main sections. One section (Objective) remains difficult to the categorization. Notice that this section is very rare in the collected data. To prepare further experiments and the improvement of the results, we plan to exploit other features such as terms and their semantic categories as well as information about the sequentiality of the sections.

References

[1] Haynes RB, Mulrow CD, Huth EJ, Altman DG & Gardner MJ. More informative abstracts revisited. Ann Intern Med. 1990 Jul 1;113(1):69-76.

[2] Sollaci LB & Pereira MG. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. J Med Libr Assoc. 2004 Jul;92(3):364-7.

[3] Harbourt AM , Knecht LS, and Humphreys BL. Structured abstracts in MEDLINE, 1989-1991. Bull M ed Libr Assoc. 1995 Apr; 83(2):190-5.

[4] Wilczynski NL, Walker CJ, McKibbon KA & Haynes RB. Preliminary assessment of the effect of more informative (structured) abstracts on citation retrieval from MEDLINE. Medinfo. 1995;8 Pt 2:1457-61.

[5] Chung GY. Sentence retrieval for abstracts of randomized controlled trials. BMC Med Inform Decis Mak. 2009 Feb 10;9:10.

[6] McKnight L & Srinivasan P. Categorization of sentence types in medical abstracts. AMIA Annu Symp Proc. 2003:440-4.

[7] Ruch P, Baud R, Chichester C et al. Extracting key sentences with latent argumentative structuring. Stud Health Technol Inform. 2005;116:835-40.

[8] Tsuruoka Y, Tateishi Y, Kim JD & al. (2005). Developing a robust part-of-speech tagger for biomedical text. In LNCS, p. 7746:382–92.

[9] Cristianini N & Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000