Automatic sentence classifier using sentence ordering features for Event Based Medicine: Shared task system description

Spandana Gella
University of Melbourne
sgella@student.unimelb.edu.au

Duong Thanh Long
University of Melbourne
lduong@student.unimelb.edu.au

Abstract

In this paper, we propose an automatic sentence classification model that can map sentences of a given biomedical abstract into set of pre-defined categories which are used for Evidence Based Medicine (EBM). In our model we explored the use of various lexical, structural and sequential features and worked with Conditional Random Fields (CRF) for classification. Results obtained with our proposed method show improvement with respect to current state-of-the-art systems. We have participated in the ALTA shared task 2012 and our best performing model is ranked among top 5 systems.

1 Introduction

Evidence Based Medicine (EBM) or Evidence based practice is "systematically locating, appraising, and using contemporaneous research findings as the basis for clinical decisions" (Rosenberg and Donald, 1995). Considering the huge amounts of literature and millions of clinical articles currently available and continuously being added to databases like PubMed[1], automating the information access or searching scientific evidence for EBM is a crucial task. Currently evidence based practitioners use the PICO criterion which was proposed by Armstrong (1999) to construct queries and search information in EBM tasks. The PICO concepts or tag-sets are: *Population* (P), *Intervention* (I), *Comparison* (C) and *Outcome* (O).

In this paper, we present a method that classifies sentences in the abstract of a clinical article according to PIBOSO criteria which is an extension of PICO. PIBOSO has six tags: *Population* (P), *Intervention* (I), *Background* (B), *Outcome* (O), *Study Design* (SD) and *Other* (Oth). This information could be used in constructing queries or searching relevant articles in the EBM task. A clear description of the PIBOSO tag-set is available in (Kim et al., 2011), who proposed the tag-set. Our system is based on the CRF algorithm which was earlier used by Kim et al. (2011) for a similar task and proven to be useful.

The major contribution of this paper is that we use a simple and large set of features such as lexical, structural and sequential features and show major improvements on the task of sentence classification over earlier attempts. Our classification techniques have shown clear improvement over existing state-of-the art systems especially for unstructured abstracts.

The paper is organised as follows: We present our related work in Section 2, describe the dataset for training and evaluation in Section 3, and our method and experimental setup in Section 4. We present the analysis of our results in Section 5 and conclude in Section 6.

2 Related work

The first attempt to classify abstract sentences based on the PIBOSO schema is made by Kim et al. (2011). They used the Conditional Random Field (CRF) classifier for learning, and their feature set included lexical features (unigram and bigram with part-of-speech), semantic features (using metathesaurus), structural features (sentence positional features) and sequential features (features from previous sentences). They found out that the best features are **unigrams, sentence po-**

[1] http://en.wikipedia.org/wiki/PubMed

Spandana Gella and Duong Thanh Long. 2012. Automatic sentence classifier using sentence ordering features for Event Based Medicine: Shared task system description. In *Proceedings of Australasian Language Technology Association Workshop*, pages 130−133.

sitional attributes, and sequential information. Using this best configuration of features and the same data set as in our experiment, they did 10 fold cross validation. The best microaverage F-score for each class or label for both Structured (S) and Unstructured (U) data are summarised in Table 3.

The other attempt of same 6 way PIBOSO classification on the same dataset is presented by (Verbeke et al., 2012). In this method, the input sentences are pre-processed with a named-entity tagger and dependency parser. They used a statistical relational learning approach in which features are constructed declaratively using intentional relation. Unlike us and Kim et al. (2011) they have used SVM-HMM[2] for learning. Similar to Kim et al. (2011) they did 10 fold cross validation and the best microaverage F-score of their system is also summarised in Table 3.

3 Dataset

To build the EBM classifier we used the 800 expert annotated training abstracts and 200 test abstracts which were given as part of the shared task. Kim et al(2011) annotated this data using abstracts retrieved from MEDLINE. Both the training and test abstracts have two types of abstracts, structured (S) and unstructured (S). In structured abstracts sentences are organised (and labelled) in an orderly fashion such as *Aim, Method, Results, Conclusions and Other* whereas these labels are absent in unstructured abstracts.

Please note that the way we categorised an abstract as structured or unstructured might be a bit different from previous approaches by Kim et al. (2011) and Verbeke et al. 2012. If the first sentence in an abstract is a sentence ordering label then we considered the abstract as structured or else unstructured. There are 1000 abstracts containing 11616 sentences in total. Statistics of the dataset used are presented in Table 1 and Table 2

	All	**S**	**U**
Abstracts	1000	37.4%	62.6%
Sentences	11616	54.4%	45.6%

Table 1: Dataset statistics

[2]http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html

	All	**S**	**U**
Labels	12211	6553	5658%
-Background	22%	10.5%	35.7%
-Intervention	5.9%	4.9%	7.1%
-Outcome	38.9%	35.2%	43.3%
-Population	6.9%	5.8%	8.4%
-Study Design	2.0%	2.36%	1.6%
-Other	29.2%	44.7%	10.8%

Table 2: Dataset statistics

4 System Description

In this section we present the details of our feature set, the training (classification) algorithm, the tools used and assumptions made in executing the experiments.

4.1 Features

We have trained our classifier with different set of features which include lexical features, structural features, sequential features and dependency features [3].

- Lexical features include lemmatized bag-of-words, their part-of-speech, collocational information, the number of content words, verbs and nouns in the sentence (we have used the MedPost (Smith et al., 2004) part-of-speech tagger).
- Structural features include position of the sentence in the abstract, normalised sentence position, reverse sentence position (Kim et al., 2011).
- Sequential features include previous sentence label, similar to Kim et al. (2011).

Additionally, for structured abstracts, we use the sentence ordering labels as features: Heading, Aim, Method, Results, Conclusions. These are provided in the data. Since unstructured abstracts do not have these ordering labels, we automatically annotate the training and testing data with ordering labels using simple heuristics. In the unstructured training data, sentences are classified into an ordering label based on its PIBOSO label: *Background* –> Aim, (*Population or Intervention or Study Design*) –> Method, *Outcome* –> Results and *Other* –> Other. In the unstructured testing data, we have divided sentences into four equal groups based on their position and mapped

[3]We have tried using dependency relations as features but found they did not improve the results. The reason for this could be data sparsity.

them to Aim, Method, Results and Conclusions in this order. **Using *sentence ordering labels* for *unstructured abstracts* is the main difference compared to earlier methods (Kim et al., 2011; Verbeke et al., 2012)**.

We tried 6 combinations of features which will be discussed in Results section.

	Kim et al.		Verbeke et al.		Our System	
Class	**S**	**U**	**S**	**U**	**S**	**U**
Background	81.84	68.46	86.19	76.90	**95.55**	**95.02**
Intervention	20.25	12.68	26.05	16.14	23.71	**50.79**
Outcome	92.32	72.94	92.99	77.69	**95.24**	**99.04**
Population	56.25	39.8	35.62	21.58	42.11	**60.36**
Study Design	43.95	4.40	45.5	6.67	0.67	3.57
Other	69.98	24.28	87.98	24.42	83.71	**91.76**
Overall	80.9	66.9	84.29	67.14	81.7	**89.2**

Table 3: F-score per class for structured (S) and unstructured (U) abstracts (bold states improvement over other systems)

4.2 Algorithm

Our sentence classifier uses CRF learning algorithm[4]. We have also executed few experiments using SVM and observed CRF performed better over this dataset with our choice of features. Due to space constraints in this paper we are not comparing CRF versus SVM results.

For feature selection, we used Fselector[5] package from R-system[6]. From the pool of features, we select the "meaningful" features based on the selecting criteria. We have tested several criteria including (1) information gain (2) oneR (3) chi-square test (4) spearman test. Among them, information gain outperformed the others. We select the 700 best features from our pool of features based on information gain score.

Other technique we used for this shared task is "bootstrapping". Our system performed very well on training data but did not perform well on test data, perhaps it suffered over-fitting. To overcome this, we ran our current best model on test data (without using gold-standard labels) and then merge the result with train data to get the new train. In that way, under ROC evaluation, we improved our final scores **by 3%**. In addition, we also pre-process the data. Since the heading such as "AIM,OUTCOME,INTRODUCTION etc." are always classified as "other" in train data, when we find sentence which has less than 20 characters and all in upper case (our notion of heading), we directly classify it as "other" in test data.

[4]We used open-source CRF++ tool. http://crfpp.googlecode.com

[5]http://cran.r-project.org/web/packages/FSelector/index.html

[6]http://www.r-project.org/

5 Results

Features	**B**	**I**	**O**	**P**	**SD**	**Oth**	**All**
BOW	9.1	3.2	68.8	2.9	0	31.7	38.4
+lexical	18.2	7.0	71.6	11.1	0	65.2	55.3
+struct	60.7	8.3	87.7	17.1	0.6	57.4	62.2
+ordering	93.7	**23.7**	**96.6**	41.0	**1.3**	80.9	80.8
All	95.2	23.7	95.2	42.1	0.6	**83.7**	**81.7**
All+seq	**95.5**	23.7	94.9	**44.2**	0.6	82.9	81.4

Table 4: Analysis of structured abstracts: microaverage f-score, best performance per column is given in bold

Features	**B**	**I**	**O**	**P**	**SD**	**Oth**	**All**
BOW	13.0	0.7	79.1	1.8	0	14.3	38
+lexical	34.2	1.5	68.0	2.2	0	13.3	40.0
+struct	58.1	5.0	72.1	12.3	1.2	26.9	52.6
+ordering	93.7	40.2	**99.2**	52.4	1.2	**96.6**	88.0
All	**95.0**	**50.7**	99.0	**60.3**	**3.5**	91.7	**89.2**
All+seq	94.9	50.7	98.7	60.1	3.5	90.8	89.0

Table 5: Analysis of unstructured abstracts: microaverage f-score, best performance per column is given in bold

In this section we present the analysis of results on structured and unstructured abstracts separately. In all our experiments, we performed 10-fold cross validation on the given dataset. Shared task organisers have used Receiver operating characteristic (ROC) to evaluate the scores. According to ROC our best system scored **93.78%** (public board) and **92.16%** (private board). However, we compare our results with (Kim et al., 2011) and (Verbeke et al., 2012) using the micro-averaged F-scores as in Table 3. Our system **outperformed previous works in unstructured** abstracts (22% higher than state-of-the-art). Our system performed well in classifying Background, Outcome and Other for both structured and un-structured data. However, our system performed poor in classifying study design as very few instances of it is available in both test and train.

We present the results of 6 systems learned using different feature sets: Table 4 for structured abstracts and Table 5 for unstructured abstracts. We choose bag-of-words (BOW) as the base features, *+lexical* includes BOW and lexical features, *+struct* include BOW and structural features, *+ordering* includes BOW and sentence

ordering labels, *All* includes BOW, lexical, struct and ordering features. *All+seq* includes all these features and sequential features.

In previous works, F-scores for unstructured data are low (compared to structured data). However, adding the automatic **sentence ordering label** to the unstructured data improved the performance drastically. This is the main difference compared to earlier models. Overall, our system outperformed existing systems in both structured and unstructured in many labels, which are highlighted in Table 3 under our system section.

Finally, combining **BOW, lexical, structure and sentence ordering** features showed the highest performance for both structured and unstructured data. It also showed that adding the sequential feature (i.e. the PIBOSO label of the previous sentence) do not help in our system, in fact the result slightly reduced. (81.7 –> 81.4 for structured and 89.2 –> 89.0 for unstructured).

6 Conclusions

In this paper, we have presented a brief overview of our method to classify sentences to support EBM. We showed that structural and lexical features coupled with a CRF classifier is an effective method for dealing with sentence classification tasks. The best features in our setting are found to be words, lexical features such as part-of-speech information, sentence positional features, collocations and **sentence ordering labels**. Our system outperformed earlier existing state-of-art systems (Kim et al., 2011; Verbeke et al., 2012).

Acknowledgements

This work is funded by Erasmus Mundus European Masters Program in Language and Communication Technologies. We thank Dr. Paul Cook for his valuable suggestions and helpful comments. We also thank the shared task organisers for providing the gold standard data for evaluations.

References

Kim, S. N., Martinez, D., Cavedon, L., and Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(Suppl 2):S5.

Rosenberg, W. and Donald, A. (1995). Evidence based medicine: an approach to clinical problem-solving. *Bmj*, 310(6987):1122–1126.

Smith, L., Rindflesch, T., Wilbur, W., et al. (2004). Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.

Verbeke, M., Asch, V. V., Morante, R., Frasconi, P., Daelemans, W., and Raedt, L. D. (2012). A statistical relational learning approach to identifying evidence based medicine categories. In *EMNLP-CoNLL*, pages 579–589.