2007－NL－177　(7)
2007／1／26

MEDLINE 概要文の役割分類に対する信頼度の異なるデータからの学習の適用

ナタリー・アイゼンバーグ, 原 一夫, 新保 仁, 松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
[natali-a,kazuo-h,shimbo,matsu]@is.naist.jp

科学論文アブストラクトの各文は, 大きく分けて, 研究の背景, 目的, 実験手法, 結果および結論に分類できる. このような文の構造的な役割の推定は, 情報検索の際に, 絞り込みの手がかりとして用いることが可能である. 先行研究では, 文内および文脈の情報を表す種々の素性を用いて文の役割分類を行った結果, 高い精度を得ている. ただし, その際には十分な量の学習データを確保するために, 実際に運用されるデータとは異なる学習データ収集方法が取られており, このことに起因して, 運用データに対する精度が, (学習データと同じ分布のデータに対して適用した場合と比較して) 低い数値に留まったことが報告されている. 本論文では, この問題に対し事例の “データソース” に依存して, 異なるコストを割り当てることによって解決を試みた.

Learning from Data of Varying Quality for Sentence Role Identification Task in MEDLINE Abstracts

Natalia Aizenberg, Kazuo Hara, Masashi Shimbo and Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
[natali-a,kazuo-h,shimbo,matsu]@is.naist.jp

The abstract of a scientific paper, typically, consists of sentences describing the background and objective of the study as well as its experimental methods results and conclusions.
There has been an increasing interest in recent years in identifying such structural roles, with particular motivations from the information retrieval point of view. In previous research done with respect to MEDLINE abstract classification, various sentence feature combinations were used in order to achieve successful performance, but one important issue has not yet been addressed: the unrepresentativeness of the major part of learning data, as the learning set samples tend to originate from different sources baring many differences while the application data source distribution does not necessarily obey that of the learning set.
In this paper we solve this issue by applying “example source” sensitive costs in the training process.

1 Introduction

Identifying sentences functional roles in text is essential for proper retrieval of information. With the rapidly increasing amounts of scientific literature, in particular in the field of medicine, it is becoming imperative that sentences within scientific abstracts are classified into several relevant groups, to allow efficient retrieval for user queries. Sentences within a scientific abstract can typically be divided into five groups: BACKGROUND, OBJECTIVE, METHODS,RESULTS and CONCLUSIONS.

In our study, we focus on sentence role identification in MEDLINE[5] abstracts. MEDLINE, having over 14 million records of thousands of publications, constitutes one of the largest electronic collections of life science and biomedical information, hence providing highly challenging settings for information retrieval.

The MEDLINE abstract collection consists of mainly two groups of abstracts, *structured* abstracts and *unstructured* abstracts. Structured abstracts are abstracts which are divided into sections reflecting the structure of the abstract. Each section in such abstract is explicitly marked with a heading indicating the structural role of the sentences to follow. Unstructured sentences (outnumbering structured exceedingly), are plain abstracts with no further additions.

The structured and manually annotated unstructured sentences differ in several points: (1) Chronology of information. In structured abstracts sentence of same role are more likely to appear together in the same "chunk", rather than in different separated locations in the abstract. (2) Grammar. Having written a title over a sentence, authors often continue with a non-grammatical sentence. For instance: "OBJECTIVE: To assess the efficiency...", where the heading "OBJECTIVE" is followed by an infinitive. (3) Reliability. An author forced to title chunks of his abstract is more inclined to have a sentence of one class appear among sentences of a different class without the proper title. This often happens with CONCLUSIONS and RESULTS sentence.

2 Learning from Data of Varying Quality

Seeing as unstructured annotated sentences are expensive to acquire, and structured abstracts, while widely available, sometimes fail to represent properly the application data (as shown in Yamasaki et al.[6]), we wish to introduce a method that will allow the mutual contribution of both types of training data with preference to the more representative and trustworthy samples (i.e., unstructured abstracts).

In order to increase influence of better representing learning samples, we propose to allow "cheaper" violations of unrepresentative samples by the classifier.

2.1 SVM

Our choice of classifier is the SVM [7] [1]. From a given set $\{(x_i, y_i)\}_{i=1}^{M}$ of training data, where x_i is the input feature vector of the i-th example and $y_i \in \{+1, -1\}$ its label, the SVM learns a decision function

$$f(x) = \mathrm{sgn}(w \cdot x - b)$$

Where x is the input vector and f produces the class label +1 (positive) or −1 (negative) for each x. Generally, there can be infinitely many such decision functions that can correctly classify all training data, i.e., $f(x_i) = y_i$ for all $1 \leq i \leq N$. Among these, SVM chooses the one that maximizes the "margin" between the two classes. We demand that correct output of f for a positive example is over +1 and correct output for a negative example is lower than −1, hence defining two hyperplanes, H_{-1} and H_{+1} such that:

$$H_{+1} : w \cdot x - b = 1,$$

$$H_{-1} : w \cdot x - b = -1.$$

The margin is then defined as the distance between these two hyperplanes and can be expressed by:

$$2\frac{|w \cdot x - b|}{\| w \|} = \frac{2}{\| w \|}.$$

The task of maximizing the margin is then equivalent to the task of minimizing the value of $\| w \|$. If we allow a soft margin with violation slack ξ_i for each training vector, the task can be rewritten as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} w^T w + C(\sum_i \xi_i) \\ \text{s.t.} \quad & y_i(w \cdot x_i - b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

Where ξ_i is a violation slack and C is a constant set for punishing the violation.

[1] For our experiment in Section 4 we use libSVM by Lin[2]

2.2 SVM for Data of Varying Quality

In order to promote representative examples over unrepresentative examples, we introduce a new violation punishing scheme. Instead of using a constant C equal through all example as used by plain SVMs, we shell assign an example dependent value to C

$$C(i) : \mathbb{N} \mapsto \mathbb{R}.$$

Where:

$$C(\text{representative sample}) \gg C(\text{unrepresentative sample})$$

Where i is the examples index instead of $C(x_i)$. Let us rewrite (1) using the newly defined C:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}w^T w + \sum_i C(i)\xi_i \\ \text{s.t.} \quad & y_i(w \cdot x_i - b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{2}$$

By introducing Lagrangian multipliers $\alpha_i, \mu_i \geq 0$ to (2) we obtain the following primal Lagrangian:

$$\begin{aligned} L(w,b,\xi,\alpha,\mu) =& \frac{1}{2} \parallel w \parallel^2 + \sum_i C(i)\xi_i \\ & - \sum_i \alpha_i[y_i(w \cdot x_i + b) + \xi_i - 1] \\ & - \sum_i \mu_i \xi_i. \end{aligned} \tag{3}$$

The solution is given at he saddle point of L, at which the derivatives with respect to w,b and ξ vanish. Substituting these constraints into (3) yeilds the following Wolfe dual problem that should be maximized with respect to α

$$\begin{aligned} L_D(\alpha) =& \sum_i \alpha_i \\ & - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C(i). \end{aligned} \tag{4}$$

Since C does not actually depend on the example itself but only of its source (a structured or an unstructured abstract) we have

$$C(i) = \begin{cases} C_u & \text{if } x_i \in \text{unstructured abstracts}, \\ C_s & \text{if } x_i \in \text{structured abstracts}. \end{cases} \tag{5}$$

$$\text{with} \quad C_u \geq C_s$$

When $C_s = 0$ the problem is reduced to that of a plain soft-margin SVM using only representative (unstructured) examples with $C = C_u$.

3 Related Work

In previous research (Yamasaki et al. 2005)[6], the effect of intra-sentential features on quality of classification was studied. In this study, features such as words and word bi-grams were used, as well as some tense information. In this work, mainly structured abstracts were used to train a support vector machine, which was then tested on both structured and unstructured abstracts. Yamasaki et al. show that, although some intra-sentential information can contribute to the task, the accuracy is significantly lower when testing the machine on unstructured abstracts rather than structured. This is the motivation for our work.

Abe et al.[1] have suggested using cost-sensitive learning by manipulating the violation slack constant C. Their work however, only associated varying costs with different types of mis-classification hence manipulating the violation punishment depending on the class associated with the input and its correlation to the hypothesis output.

Geibel et al.[3] have proposed a theoretical framework for example dependent costs. In their work a similar approach to the one introduced in this paper is proposed, claiming that the violation slack punishment should be used for resolving the problem unrepresenatative data in training set. They suggest a general definition of both class and example dependent violation cost function. However, the training data used in their experiment is artificially constructed and contains a small dimensional feature space.

4 Experiment

4.1 Experiment Settings

Our training data consisted of 3342 annotated sentences, 2787 of which are from structured abstracts. The test data however will always be composed of unstructured abstracts only.

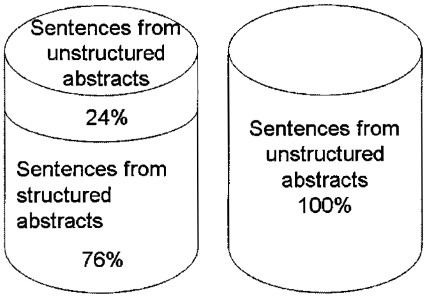
Sentences from unstructured abstracts
24%
Sentences from structured abstracts
76%
Sentences from unstructured abstracts
100%

Figure 1: The container on the left shows distribution of data sources in training set. The container on the right shows distribution of data sources in application set.

Each of these sentences is classified into one of the following 5 classes: BACKGROUND , OBJECTIVE, METHOD, RESULT and CONCLUSION. Distribution of these roles is shown in Table 1

To learn classification of these sentences we used the following features:
(1) surface words
(2) base words (lemmas)[2]
(3) part of speech information
(4)(5)(6)(7) all possible combinations of the previous three features
(8) bi-grams.

In this experiment we evaluate both the pairwise and the multi-class performance of the proposed method. From a total of 103813 abstract from the year 2002 we chose 4000 sentences belonging to fully annotated abstract only. We then randomly chose 20% from those sentences for testing data. From the chosen testing data we removed all sentences belonging to structured abstracts and were left with a test set of 126 sentences.

4.2 Baseline Construction

Since in our experiment we use different sets of data than those used in the previous study of MEDLINE sentence role identification, we start by reproducing two essential baselines, to prove addition of large quantities of somewhat noisy learning samples to a small amount of high quality samples can improve performance:

(1) plain SVM trained only on unstructured (i.e. representative) samples. This implies $C_s = 0$ since all structured examples are to be ignored.

(2) plain SVM trained on both representative and unrepresentative examples without discriminating between them. This implies $C_u = C_s$.

We have tested both of these baselines with different C values and for the optimal choice of C for each pairwise problem we have composed a multi-class model for each baseline. In the final multi-class test, models using representative training data only, outperform the models trained on all data, as shown in Table 2. However, in the in pairwise tests (Figure 2) we witness that in 3 of the 10 pairwise tests, mod-

Table 2: multi-class accuracy %, representative only vs all

	plain SVM with structured+unstructured $C_u = C_s$	plain SVM with structured only ($C_s = 0$)
Accuracy%	64.3	67.5

[2] For lemmatization and part of speech information we used GENIA Tagger[4] a part of speech tagger for biomedical text.

Table 1: Frequency of individual roles in structured and unstructured abstracts of training data.

source	BACKGROUND	OBJECTIVE	METHOD	RESULT	CONCLUSION
unstructured	11.5%	7.9%	18.9%	47.7%	13.8%
structured	13.3%	10.0%	24.5%	35%	16.6%
all	13.0%	9.6%	23.5%	37.5%	16.1%

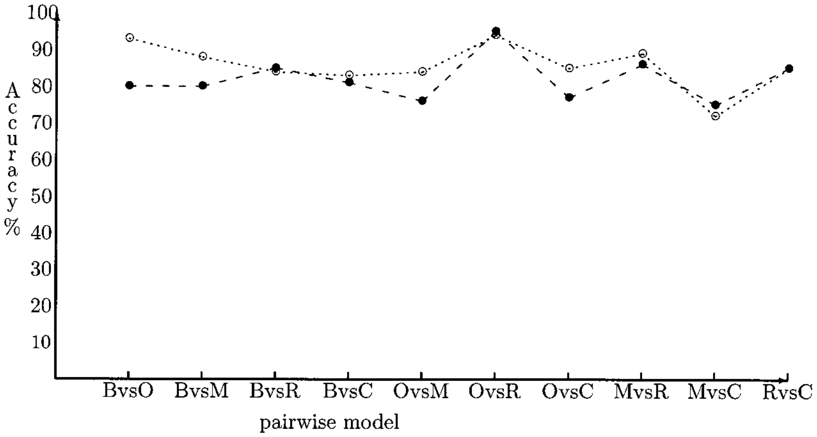

Figure 2: Pairwise Baselines. B=BACKGROUND, O=OBJECTIVE, M=METHOD, R=RESULTS and C=CONCLUSIONS. The dotted line plots the maximal accuracy achieved with unstructured sentences only. The broken line plots maximal accuracy achieved with equal C for all data.

els trained on all data, outperform models trained on representative data only (see Table 3).

pairwise case	all data	unstructured only
BACKGROUNDvsRESULT	85.5	84.3
OBJECTIVEvsRESULT	95.8	94.5
METHODvsCONCLUSION	75.7	72.7

Table 3: pairwise cases where models trained on all data $C_u = C_s$ outperform models trained on representative data only $C_s = 0$

4.3 Establishing C

Since in our training set samples come from exactly two sources, we set for each pairwise classification model two penalty constants, C_s for samples of structured abstracts and C_u for samples of unstructured abstracts. In our experiment we would like to establish the optimal (C_u, C_s) combinations for each pairwise model and test them against the previously mentioned baselines, both in the pairwise problem and in the multi-class problem.

To establish C values, we tested for each pair of classes, a penalty constant C_u, for unstructured examples that ranges between 10^{-10} and 10^{10}. For each C_u we then traversed over all structured penalty constants C_s, that will result in a 10^0 to 10^9 deviation from C_u.

For the case of METHODS vs OBJECTIVE for instance the maximal performance was chosen out of the values shown in figure 3

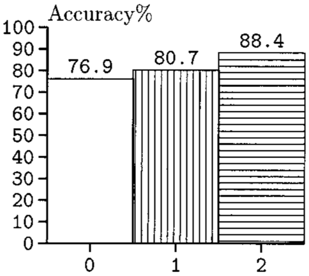

Figure 3: Methods vs Objectives C-search. The x axis stands for the $log_{10}(ratio)$ value $ratio = C_u/C_s$. Each column in the graph shows the maximal accuracy value for the given ratio. The C_u yielding accuracy 88.4 will be chosen. C_s of the maximal accuracy is hence $C_u/ratio$

Having found the performance for each of the stated C combinations for each pair, we choose the values that amount to the highest performance.

4.4 Testing Found Cs

In Table 4 we show that in the pairwise case our models can outperform both of the baselines defined earlier. In 3 out of the 10 cases our models outperform both baselines, in 6 of the left 7 cases, we repeat the accuracy of models trained on unstructured abstracts only, in 2 of those 7 cases we repeat the accuracy of the models trained with $C_s = C_u$. In 2 of the earlier mentioned 7 tie cases we use $C_s = 0$ in the source sensitive models since no higher C_s repeats (or exceeds) the performance of the "unstructured only" models.

4.5 Multi-Class Classification

To insure our implementation does indeed exceed the performance of both baselines, we performed a simple pairwise multi-class classification using the pairwise models we have already trained. In the multi-class case, yet again, (as figure 4 shows) models trained with optimal, source sensitive Cs seem to outdo, both the models trained on all data with equal (and optimal) C_u and C_s, as well as models trained exclusively on representative samples, unstructured data, also with optimal C values for each pair.

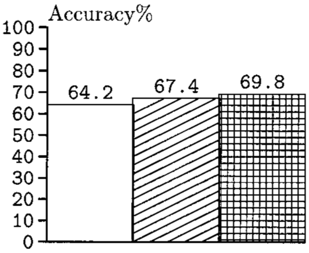

Figure 4: Multi-Class Accuracy %. In this figure the 1st column shows the optimal accuracy of models trained with $C_u = C_s$. The 2nd column shows the optimal accuracy of models trained only on unstructured abstracts (i.e. $C_s = 0$). The last column shows the maximum accuracy achieved by models trained with origin sensitive Cs.

5 Conclusions and Future Work

We have shown that using source sensitive violation costs is beneficial for both the pairwise problem and the multi-class problem. In future work, we will try extending this sensitivity to class dependent values as well. We also consider combining source sensitive violation costs with context dependent classification techniques, seeing as previous research (Yamasaki et al. [6]) shows that in sentence role identification task, context is important for achieving optimal performance.

Table 4: Pairwise Accuracy %. The bold rates are the highest rates for the given pairwise problem

Pairwise Problem	plain SVM with structured+unstructured ($C_s = C_u$)	plain SVM with unstructured only ($C_s = 0$)	source sensitive C
BACKGROUND vs OBJECTIVES	80.7	**83.7**	**83.7**
BACKGROUND vs METHODS	80.7	**88.4**	**88.4**
BACKGROUND vs RESULTS	85.5	84.3	**86.7**
BACKGROUND vs CONCLUSIONS	81.0	**83.7**	**83.7**
OBJECTIVES vs METHODS	76.9	84.3	**88.4**
OBJECTIVES vs RESULTS	95.8	94.5	**97.2**
OBJECTIVES vs CONCLUSIONS	77.7	**85.5**	**85.5**
METHODS vs RESULTS	86.7	**89.8**	**89.8**
METHODS vs CONCLUSIONS	**75.7**	72.7	**75.7**
RESULTS vs CONCLUSIONS	**85.5**	**85.5**	**85.5**

References

[1] Naoki Abe, Bianca Zadrozny, and John Langford. An iterative method for multi-class cost-sensitive learning. In *Proceedings of ACM KDD'04*, 2004.

[2] Chih-Chung chang and Chih-Jen Lin. A library for support vector machines. `https://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[3] Peter Geibel, Ulf Brefeld, and Fritz Wysotzki. Learning linear classifiers sensitive to example dependent and noisy costs. *Lecture notes in computer sceince*, 2810:167–178, 2003.

[4] GENIA Tagger. `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger`. Department of Information Science, Faculty of Science, University of Tokyo.

[5] MEDLINE. `http://nlm.nih.gov/databases/databases_medline.html`, 2003. U.S. National Library of Medicine.

[6] Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto. Sentence role identification in medline abstracts: Training classifier with structured abstracts. *Lecture Notes in Computer Science*, pages 236–254, 2005.

[7] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.