A New Approach to Zone Identification Based on Considering Features with High Semantic Richness

Kambiz Badie
Knowledge Management & e-Organization Group
IT Research Faculty, ICT Research Institute
Tehran, Iran
K_badie@itrc.ac.ir

Nasrin Asadi
IT Research Faculty, ICT Research Institute
Tehran, Iran
asadi@itrc.ac.ir

Maryam Tayefeh Mahmoudi
Multimedia Research Group
IT Research Faculty, ICT Research Institute
Tehran, Iran
mahmodi@itrc.ac.ir

***Abstract*— In this paper, we propose a new approach to zone identification based on considering features with high semantic richness such as specialized names and mode of verbs belonging to a text's domain of interest and besides that mode of verbs, while taking into account features with less computational cost compared to those of conventional methods. Out of the scenarios of selecting features for identifying a zone based on classifying the sentences in a text, we came to notice that in the scenario where specialized names and mode of verbs are taken into account together with reduced versions of conventional features including history, an accuracy rate of 61% (resp. 81%) is obtained which is higher than those belonging to both Liakata's and Fisas's approach. Also, to have a genuine comparison, both Liakata's and Fisas's corpuses are used in our experiments. Such accuracy is obtained at the place where less computational cost is taken for extracting the features.**

Keywords*— *Text summarization; zone identification; scientific paper (text); linguistic features; semantic richness; domain of interest; classification accuracy.

I. INTRODUCTION

Within the past years, zone identification has been elaborated as a major research concern within the areas of text mining in general and text summarization in particular [1,2,3]. The major purpose behind this issue is to identify those zones in a text which tackle a certain concept issue, topic or subject from the reader's point of view. Examples can be mentioned for the approaches trying to find out which parts (comprising a number of sentences) in a text refer to "background", "the proposed approach", "experimentation", "approach validation" or "conclusion" as the important perspectives which are to be followed (pursued) in a paper, or for instance the approaches which try to figure out whether or not a certain scientific concept, subject or issue has been addressed somewhere in a scientific paper. Obviously, the more concrete the concept of a zone class with well–defined elements, a higher possibility would exist to identify the desired zone meaningfully with less emphasis on complicated features. Also, the higher the abstraction level of a zone class, the more effort would be necessary to take into account higher order linguistic features to identify the desired zones. This is mainly because, in comparison with a simple phrase, a subject or an issue usually calls for sophisticated relations between a variety of simple concepts to have itself characterized meaningfully. In this paper, we will demonstrate how through considering features with high semantic richness such as specialized names (belong specifically to a text's domain of interest) and besides that mode of verbs, a higher classification accuracy can be attained for zone identification at the place where features with less computational cost compared to the conventional features are being used.

II. RELATED WORKS

Recently, automatic identification of zone categories existing within the scope of articles has become quite important and many researchers have tried to analyze the content of scientific texts from various points of view. Some of them focus on categorizing the sentences in the abstracts of articles [4,5], while many others have worked on full-text articles [6-8]. It should be mentioned that these zone identification approaches may be different in classification method, selected features, annotation scheme and domain of the dataset to be used. Most of the existing zone identification approaches make use of classification techniques such as Support Vector Machine (SVM), Naïve Bayes, Logistic Regression and Conditional Random Fields (CRF) to classify the sentences [9,10]. Seaghdha et.al. [11] proposed BOILERPLATE-LDA, an unsupervised model, that elicits some aspects of rhetorical structure from unannotated text and uses them as features to classify the sentences in zone categories. In another study, using association rule mining on the dependency structure of the sentences, Groza detected structural differences between zone categories [12]. The features used to classify the sentences take into account different aspects of a sentence, ranging from its location in the article and the headline of the corresponding section within the document, to features which relate to the components of a sentence such as verbs, n-grams and the relation between them, like Grammatical triples [9,10].

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

With regard to annotation schemes, several alternatives have also been created for various fields of sciences, some of which include only a few number of categories [13] while some others are finer-grained [3] that capture the content and the conceptual structure of scientific articles. These schemes have been applied to articles in various domains such as biochemistry, chemistry, graphic computer, etc.

Soldatova et.al. [14] introduced a sentence-based three-layer scheme, called CoreSC, which recognizes the key points of articles and consists 11 categories. Liakata et.al. made use of SVM and CRF techniques to classify the related sentences in the biochemistry texts. In this regard, a classification rate of 51.6% was obtained by applying multi-class classification through using SVM [3]. Fisas et.al. used both SVM and logistic regression to make classification on sentences within the area of computer graphics, and a rate of 80% was obtained at most by using regression [10].

The current paper comes up from an effort to improve the aforementioned results of Liakata and Fisas. The advantage of our work is that we exploit features with high semantic richness, and in the meantime features with less computational cost. Taking this point into account we were able to obtain a higher accuracy in results.

III. THE PROPOSED APPROACH

A. Basic idea

The main point in our approach is to see how far, through considering features with semantic richness such as mode of verbs in a sentence, one can attain a better perception toward the zone class to which a sentence belongs to. In the meantime a right perception toward the status of specialized nouns (either general or specific) in a sentence may have the potential to help zone identification be performed in a more meaningful way with less amount of computational cost. Status of verbs is important since the identity of a zone class in many cases depends on the way its specialized nouns are verbalized. Meanwhile, relative's position of a sentence in the text for which a variety of parameters are to be considered, is to be characterized with reasonable amount of information to avoid extra computational cost.

Another point is that mapping correctly a sentence onto zone classes which share some similarities in is a difficult issue, which calls for further features preferably with deeper linguistic sense is an issue which is quite hard to be tackled from natural language processing viewpoint. With regard to Liakata's approach, examples can be mentioned for a sentence belonging to zones such as "result", "observation" and "conclusion", or those belonging to zones like "experiment" and "validation". Zones such as "model" and "method" and also "goal" and "objective" have equally such a characteristic as well. Also, with regard to Fisas's approach, zones such as "Background", "Challenge" and "Approach" as the main sections in a paper or article have been taken into account.

Taking into account the aforementioned points, in this paper we propose a structure for mapping from a sentence onto a zone class using SVM as the classifier. The motive for using SVM is that, in comparison with other classifiers, it has the ability to remarkably separate similar classes due to the feasibility of selecting a proper kernel function. Features like "position of a sentence in text", "tense of verbs", "class of previous sentence" "both general and specific specialized names", "highly - frequent verbs", "particular modes of verbs", … are taken into account in this respect.

B. Feature Used in the Suggested Approach

As discussed before, our main objective in this paper is to show how a fine classification accuracy can be obtained for zone identification through considering more significant features with less amount of computational cost. Here, features used by Liakata in her approach to zone identification, are considered as the ground for our trial. Our intention is to see whether we can replace some of these features by some other features with less computational cost but meaningful enough from some other perspectives. In the meantime we are curious to see how through adding features with semantic richness we may compensate for a possible drop in accuracy which is resulted due to this replacement. Below, we present some details regarding these features.

- Location: Dividing the whole paper into 11 unequal parts and deciding to which part the given sentence belongs to. (This resembles the so-called "Loc" applied by Teufel [2]; however we refine it here by dividing the fifth part into two equal parts)
- Heading types: The heading of the section within which a particular sentence exists. There are 8 types of heading called Introduction, Related Works, Proposed Approach, Experimental Results, Conclusion, Abstract, Specific and None.
- Citations, figures and tables: The number of citations, figures and tables in a sentence.
- Verb tense: The tense of the main verbs in the sentence, including present, past, present perfect, past perfect and future.
- Passiveness or activeness: Status of passiveness or activeness of the main verbs in the sentence.
- Adjective: The number of superlative or comparative adjectives in a sentence.
- First-person pronoun: Presence of first-person pronouns in the sentence.
- History: The zone class of the sentence previous to a current sentence [3].
- Frequent verb class: The ratio of the number of verbs in each zone to the number of sentences in this zone. Fifty highly frequent verbs in each zone have been considered in this respect.
- Mode of verbs: Verbs which are frequently used in the whole corpus which are manually divided into two main classes: Description verbs like "describe", "explain", "introduce" and "suggest", and Evaluation verbs like "evaluate", "measure", "increase" and "test".

- Specialized Names: A noun phrase is said to be a general specialized name (GSN) if it addresses a general aspect in that domain. It is called a specific specialized name (SSN) in case it addresses some specific aspect of an issue or a subject, such as the name of tools and methods [15]. For instance, “hydrogen” and “temperature” are GSN and “fluorescence spectrum” and “hydrogen bonding” are SSN in chemical field. More specifically, we extract the noun phrases from the training data and classify them into 3 categories by using both domain-specific and domain-general ontologies. Here, we exploit ChEBI [16] and Gene [17] ontology as chemistry ontologies and WordNet [18] as a general ontology. This is subject to the following rules in Fig. 1.

```
If the NP[a] is in acronym form, then NP class= SSN;
If the number of words of the NP==1{
        If the NP is found in the domain ontology then NP class = GSN;
        else if the NP is found in the general ontology then NP class= general;
}else {
        If at least one word in the NP is found in the domain ontology then NP class= SSN;
        else NP class = GSN;
}
```

Fig. 1. Main rules for classifying noun phrases(NP)

IV. Experimental Results

A. Dataset used in simulations

In an attempt, to show the effectiveness of the suggested approach, we decided to compare it with Liakata's approach for zone identification which has for the first time been applied to identify a variety of significant zones such as Motivation, Observation, Method and Conclusion in scientific papers like chemical tests [14]. Regarding this, Art Corpus [19] was decided to be a ground for such a comparison.

The ART corpus consists of 225 papers in the field of chemistry and biochemistry and has become annotated by 20 expert chemists. It is based on CoreSC scheme that comprises the following categories: Background (Bac), Goal (Goa), Object (Obj), Motivation (Mot), Hypothesis (Hyp), Method (Met), Model (Mod), Experiment (Exp), Observation (Obs), Result (Res) and Conclusion (Con). Table I illustrates the statistics of the ART corpus.

TABLE I. Statistics of the ART corpus

Zone class	Number of sentences	Percentage
Bac	6656	19%
Goa	507	1.4%
Obj	1022	2.9%
Mot	466	1.3%
Hyp	654	1.8%
Met	3747	10.8%
Mod	3456	10%
Exp	2841	8.2%
Obs	4659	13.5%
Res	7370	21.3%
Con	3077	8.9%
Total	34455	

In another attempt, we decided to have our approach compared with Fisas’s approach in the scope of computer graphics. The related corpus which is called Dr. Invertor corpus (DRI corpus) [20] consists of 40 papers in the area of computer graphics and has been annotated by 3 computationally oriented linguists. The whole dataset has been divided into four subgroups each of which contains 10 papers and concerns a specific field in computer Graphics; these include Skinning, Motion, Fluid simulation and Cloth simulation.

The scientific annotation schema includes five top level categories and three sub-categories. Namely, the former includes Background, Challenge, Approach, Outcome and Future Work while Contribution is served as a sub-category of Outcome; moreover Hypothesis and Goal are referred to as sub-categories of Challenge [10, 21]. Table II illustrates the statistics of DRI corpus.

B. Analysis of Simulation Results

Our main goal in simulation was to show how tending to features with semantic richness as well as considering highly–frequent verbs for each zone class (instead of co–occurrence and status of grammatical triple between verbs) can lead to a reasonable separation between the related zone classes. With regard to semantic richness, specialized noun phrases (both general and specific) which take part in different types of zone, as well as verbs which have a particular mode like those standing for "description" and "evaluation", are taken into account. Following scenarios were considered for simulation:

Scenario 1: In order to show how far the modified features, which are more cost–effective compared to those in Liakata's approach, can behave successfully, classification was performed with these feature but excluding "history" as a feature and considering "specialized names" (instead of "n-gram" in Liakata's approach) instead. The motive for such a simulation was that, extracting "history" calls for a pre – tagging on the papers, which in turn is in need of intensive experience.

TABLE II. STATISTICS OF THE DRI CORPUS

Zone class	Number of sentences	Percentage
Background	1591	20%
Challenge	405	5%
Approach	4477	56%
Outcome	1259	16%
Future Work	127	1.6%
Total	7859	

In this scenario, a classification accuracy of 48.8% was obtained for Art corpus; this is about 3% lower than the one obtained by Liakata. Furthermore, the classification rate for DRI corpus turned out to be 72.3% which, compared to the results of Fisas, indicates nearly 4% decrease in the accuracy.

Scenario 2: to show how features with semantic richness such as "specialized names" and "mode of verbs" can increase the classification accuracy, simulations were done with these features but not considering "history" as a feature, and a classification accuracy of 50.5% was obtained. This rate is quite close to the one obtained by Liakata and its message is that features with semantic nature is good alternatives for replacing "history" as a feature.

Scenario 3: To show to what extent "history" is significant, simulations were done taking into account this feature, but avoiding "mode of verbs" as a feature. A classification accuracy of 60.4% was then considering the sentences in a near neighborhood of a particular sentence may result in a more precise identification of zones, due to the fact that consecutive sentences lie very often in a neighborhood of a particular sentence.

Scenario 4: Results of previous scenarios show that both "history" and features with semantic richness play a high role in increasing the classification accuracy. This persuades us to see how co–presence of these features may lead us to a higher classification accuracy. According to this, simulations were performed and, concerning ART corpus, a classification rate of 61% was obtained which is quite remarkable compared to Liakata's. For DRI corpus, the classification rate is in the meantime higher than that of Fisas which does not exceed 76% using SVM.

We performed a 9-fold cross-validation of LibSVM [22] (a library for SVM) with a linear kernel. Table III demonstrates the corresponding results together. Perhaps Fig. 2. Provides a better understanding of the results.

The precision, recall and F-measure of each zone class of ART corpus and DRI corpus associated to scenario 4 are respectively shown in Tables IV and V.

TABLE III. CLASSIFICATION ACCURACY RESULTS

	Features List	Accuracy for ART corpus	Accuracy for DRI corpus
Scenario 1	Syntax features + Specialized Names	48.8%	72.4%
Scenario 2	Syntax features + Specialized Names+ Mode of verbs	50.5%	73.1%
Scenario 3	Syntax features + Specialized Names + History	60.4%	81.2%
Scenario 4	Syntax features + Specialized Names + History + Mode of verbs	61 %	81.4%

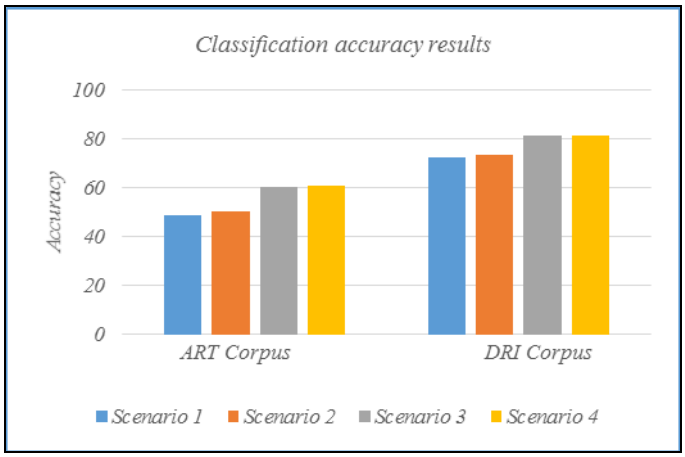

Fig. 2. Classification accuracy results

TABLE IV. PRECISION, RECALL AND F-MEASURE OF SCENARIO 4 FOR ART CORPUS, LIBSVM, 9-FOLD CROSS-VALIDATION

Zone class	precision	recall	F-measure
Obj	0.38	0.28	0.32
Met	0.54	0.55	0.55
Mod	0.68	0.69	0.68
Mot	0.27	0.16	0.16
Hyp	0.4	0.26	0.32
Obs	0.58	0.58	0.58
Res	0.55	0.61	0.58
Exp	**0.81**	**0.8**	**0.8**
Con	0.6	0.52	0.56
Bac	0.66	0.71	0.69
Goa	0.32	0.19	0.24
Avg	**0.6**	**0.61**	**0.6**

Analyzing the results of Table IV stands for the fact that Exp (0.8), Bac (0.69) and Mod (0.68) are of the highest F-measure, while the lowest F-measure is assigned to Mot (0.16) and Goa (0.24). Moreover, the large gap between the F-measures has possibly been raised by the unbalanced distribution of the train data assigned to every zone, besides the noises imposed by manual annotation. Despite suitable number of instances of the training data for the particular categories Obs, Res, and Con, they do not correspond to a high F-measure. This may be construed in view of the fact that the sentences belonging to these three categories are close in the meaning and probably some more features of high semantic richness are needed for a better result.

As it comes up from Table V, Approach (0.88) and Background (0.84) are of the highest F-measure, whereas Challenge (0.48) and Future Work (0.49) are of the lowest F-measure. This happens because Approach and Background categories have high percentage of training instances (more than 76% together) while Challenge and Future Work have a very small portion of training instances (less than 7% together). However, the number of instances in Future Work category is less than that of Challenge and, in the meantime, it attains a higher F-measure than that of Challenge.

This exception is due to some features like verb tense and frequent verb class which are powerfully distinguishing between Future Work category and the other zone categories. This brings us to the fact that more strong semantic features could help characterize the zone classes significantly even though the number of instances in the training data is not enough.

In Fisas's and Liakata's approaches, all unigrams, bigrams and trigrams with frequency greater than or equal to 4 have been included in the feature vector. It should however be noticed that the number of afore-mentioned features in ART corpus is 10515, 42438, and 11854 respectively. Thus the length of the feature vector has increased substantially and this, in turn, has led to a high computational cost. However, in the suggested approach, instead of working with all these n-grams, we focused only on specialized noun phrases (GSN and SSN) which particularly causes a sharp decrease in the length of the feature vector. Thus reducing significantly the computational cost essential to extract the features. Let say we just used 1300 features thus requiring only 28 minutes to train ART corpus, and 16 minutes to test a single fold.

TABLE V. PRECISION, RECALL AND F-MEASURE OF SCENARIO 4 FOR DRI CORPUS, LIBSVM, 10-FOLD CROSS-VALIDATION

Zone class	precision	Recall	F-measure
Background	0.84	0.76	0.8
Challenge	0.54	0.43	0.48
Approach	0.84	0.92	0.88
Outcome	0.72	0.65	0.68
Future Work	0.75	0.49	0.59
Avg	**0.81**	**0.81**	**0.81**

V. CONCLUDING REMARKS & FUTURE PROSPECTS

In this paper, we demonstrated that how, through considering features with high semantic richness such as specialized names and mode of verbs, one may attain a higher classification accuracy (with regard to zone identification) for the sentences in a text at the place where features with less computational cost are being used. This seems to be mainly because features with high semantic richness have principally the ability to participate effectively in classification with less need for involving highly syntactical features which in turns call for high computational cost. Taking this point into account a deeper investigation on the linguistic features with high semantic richness, is expected to eventually lead to higher performance. Since a zone identity manifests highly in a set of neighboring sentences, it would therefore more reasonable to perform classification on the ground of fusion between the local decisions belonging to the neighboring sentences. Realizing such an objective can be viewed as an essential research work in future.

REFERENCES

[1] Y. Hong, et al, "Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians," *Journal of biomedical informatics, vol.* 40, no. 3, pp. 236-251, 2007.

[2] S. Teufel and M. Moens. "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computational linguistics, vol.* 28, no. 4, pp. 409-445, 2002.

[3] M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and D. Rebholz-Schuhmann, "Automatic recognition of conceptualization zones in scientific articles and two life science applications," in Bioinformatics, vol. 28, no. 7, pp. 991-1000, 2012.

[4] L. McKnight and P. Srinivasan, "Categorization of sentence types in medical abstracts," AMIA AnnuSymp Proc. pp. 440-444, 2003

[5] S. V. Bonn, and J.M. Swales, "English and French journal abstracts in the language sciences: Three exploratory studies," *Journal of English for Academic Purposes,* vol. 6, no. 2, pp. 93-108, 2007.

[6] S. Teufel and M. Y. Kan,"Robust argumentative zoning for sensemaking in scholarly documents, " Springer Berlin Heidelberg, 2011, pp. 154-170.

[7] Y. Mizuta, A. Korhonen, T. Mullen and N. Collier, "Zone analysis in biology articles as a basis for information extraction," *International journal of medical informatics,* vol. 75, no.6, pp. 468-487, 2006.

[8] T. Groza, H. Hassanzadeh, and J. Hunter, "Recognizing scientific artifacts in biomedical literature," *Biomedical informatics insights, vol.* 6, pp. 15-27, 2013.

[9] H. Kilicoglu, "Biomedical Text Mining for Research Rigor and Integrity: Tasks, Challenges, Directions," *bioRxiv,* 2017.

[10] B. Fisas, R. Francesco, and S. Horacio, "On the discoursive structure of computer graphics research papers," *The 9th Linguistic Annotation Workshop held in conjunction with NAACL June 2015*, Denver, Colorado, pp. 42-51, 2015.

[11] D. O. Séaghdha, S. Teufel, "Unsupervised learning of rhetorical structure with un-topic models," Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 2–13, Dublin, Ireland, August 23-29, 2014.

[12] T. Groza, "Using typed dependencies to study and recognise conceptualisation zones in biomedical literature," PLoS ONE, vol. 8, no. 11, e79570, 2013.

[13] S. Agarwal and Y. Hong, "Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion," *Bioinformatics,* vol. 25, no. 23, pp. 3174-3180, 2009.

[14] L. Soldatova and M. Liakata, "An ontology methodology and cisp-the proposed core information about scientific papers," *Technical Report*. JISC Project Report 137, Aberystwyth University, 2007.

[15] N. Asadi, K. Badie and M. T. Mahmoudi, “Identifying categories of zones in scientific papers based on lexical and syntactical features”, In Web Research (ICWR), 2016 Second International Conference on, pp. 177-182, 2016.

[16] https://www.ebi.ac.uk/chebi/

[17] http://www.geneontology.org/

[18] https://wordnet.princeton.edu/

[19] https://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

[20] http://sempub.taln.upf.edu/dricorpus

[21] F. Ronzano, S. Horacio, "Knowledge extraction and modeling from scientific publication," In the Proceedings of the Workshop Semantics, Analytics, Visualisation: Enhancing Scholarly Data co-located with the 25th International World Wide Web Conference, Montreal, Canada, 2016.

[22] C.C. Chang, C.J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, 2011.