Abs-224

Carmen Dayrell[1], Arnaldo Candido Jr.[2,3], Stella Tagnin[1], Gabriel Lima[2,3], Valéria Delisandra Feltrim[4] and Sandra Aluisio[2,3]

[1] Departamento de Letras Modernas, Universidade de São Paulo, Brazil
[2] Núcleo Interinstitucional de Linguística Computacional (NILC), Brazil
[3] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil
[4] Universidade Estadual de Maringá, Brazil

Towards a multi-label sentence classifier for automatic identification of rhetorical moves in English abstracts

The relevance of automatically identifying rhetorical moves has been widely acknowledged due to its various applications to the development of Natural Language Processing tools [1,2,4,5,8,9,13,14]. A "move" refers to "a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse" [11].

Two approaches have been proposed to automatically detect moves in scientific texts: (i) Argumentative Zoner [12] is language-dependent and makes use of lexical, syntactical and structural features; (ii) the Text Categorization approach is language-independent and uses a bag of clusters with n-grams from 1 to 5 [1,8]. Although rhetorical moves can be realized by a clause, a sentence, or several sentences [11], most current machine-learning classifiers have established a one-to-one relationship between sentences and moves. Another drawback is the limited size of their training corpora.

The present study intends to overcome these limitations and builds on our previous work [5], which introduced AZEA (Argumentative Zoning for English Abstracts), a high-accuracy system which uses a robust set of linguistic features to automatically detect moves in abstracts from pharmaceutical sciences. Our primary aim is to develop a multi-label sentence classifier, using AZEA´s set of features and a list of formulaic-expressions created automatically [7] and two large training corpora from two broad research fields: (i) physical sciences and engineering (PE) and (ii) life and health sciences (LH). The former is made up of 845 abstracts (144,683 tokens) and the latter consists of 690 texts (150,248 tokens), taken from research papers written in English and published by various leading academic journals. Seven moves are considered [10,6]: background, gap, purpose, methodology, results, conclusion, and outline, which is used to classify instances making reference to the structure of the paper.

The Kappa Statistics [3] indicated that the multi-label sentence classification is reproducible, although some disagreements should be settled. Three experienced annotators tagged 38 abstracts from the PE and 34 from the LH corpus, previously parsed with a full syntactic parsing (OpenNLP project) and a scripting code to identify clauses and prepositional phrases. The kappa values were 0.69 (N=529, k=3, n=21) and 0.60 (N=453, k=3, n=22) for the LH and the PE corpora, respectively. The overall kappa was 0.65. The LH corpus was then automatically tagged by AZEA's current version and manually validated by one single annotator. Full annotation of the PE corpus is in progress.

Since AZEA´s accuracy drops considerably when used in a corpus with research areas different than those from the training phase, we propose to develop two classifiers (for LH and PE). Another critical issue is that multi-labeled sentences represented only 5% of sentences from the manually annotated corpus. This paper discusses the various challenges involved in automatically assigning multi-labels to a given sentence and works towards satisfactory solutions. In addition, we also intend to make the two corpora publicly available so that they may serve as benchmark for the task.

References

1. Anthony, L.; Lashkia, G. (2003) Mover: A machine learning tool to assist in the reading and writing of technical papers. IEEE Transactions on Professional Communication, 46(3):185–193.

2. Burstein, J.; Marcu, D.; Knight, K. (2003) Finding the WRITE stuff: automatic identification of discourse structure in student essays. IEEE Intelligent Systems, 18(1):32–39.

3. Carletta, J. (1996): Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics, vol. 22, n. 2, pp. 249-254.

4. Feltrim, V. D.; Teufel, S.; Nunes, M. G. V.; Aluísio, S.M. (2005) Argumentative Zoning Applied to Critiquing Novices' Scientific Abstracts. Computing Attitude and Affect in Text: Theory and Applications.1st ed. Dordrecht, The Netherlands: Springer, 1: 159-170.

5. Genovês Jr.; L., Feltrim; V.D., Dayrell C.; Aluísio, S. (2007) Automatically detecting schematic structure components of English abstracts. Proceedings of the RANLP´2007, Workshop on Natural Language Processing for Educational Resources, Borovets, Bulgaria, pp. 23-29.

6. Hyland, K. (2000). Disciplinary Discourses. Harlow, UK: Longman.

7. Machado Jr., D.; Feltrim, V. D. (2009) Extração Automática de Expressões Indicativas para a Classificação de Textos Científicos. Proceedings of The 7th Brazilian Symposium in Information and Human Language Technology, I TILIC, 2009 (Poster Presentation in Portuguese).

8. Pendar, N. ; Cotos, E. (2008) Automatic Identification of Discourse Moves in Scientific Article Introductions. Proceedings of The Third Workshop on Innovative Use of NLP for Building Educational Applications, Columbus, Ohio, USA, pp. 62-70.

9. Siddharthan, A.; Teufel, S. (2007) Whose idea was this and why does it matter? Attributing scientific work to citations. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 316-323.

10. Swales J. M.; Feak, C. B. (2009) Abstracts and the Writing of Abstracts, Michigan: University of Michigan Press.

11. Swales, J. (2004) Research Genres: Exploration and applications. Cambridge University Press, Cambridge.

12. Teufel, S. (1999) Argumentative Zoning: Information Extraction from Scientific Text, unpublished PhD Thesis, School of Cognitive Science, University of Edinburg, Edinburg, UK.

13. Teufel, S.; Moens, M. (2002) Summarising scientific articles experiments with relevance and rhetorical status, Computational Linguistics 28 (4), pp. 409-446.

14. Teufel, S. (2005) Argumentative Zoning for improved citation indexing. In James G. Shanahan, Yan Qu, and Janyce Wiebe (Eds.), Computing Attitude and Affect in Text: Theory and Applications, Springer, Dordrecht, The Netherlands, 2005, pp. 159-170.