Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora

Carmen Dayrell[1], Arnaldo Candido Jr.[2], Gabriel Lima[2], Danilo Machado Jr.[3], Ann Copestake[4], Valéria D. Feltrim[3], Stella Tagnin[1], Sandra Aluisio[2]

[1] Department of Modern Languages, University of São Paulo

Rua Prof. Luciano Gualberto, 403 - Cidade Universitária - São Paulo, Brasil CEP: 05508-900

[2] Center of Computational Linguistics (NILC), University of São Paulo

Avenida Trabalhador São-Carlense, 400 – Centro - São Carlos, Brasil CEP: 13566-590

[3] Department of Informatics, State University of Maringa

Avenida Colombo, 5.790 – Jd. Universitário – Maringá, Brasil CEP: 87020-900

[4] Computer Laboratory, University of Cambridge

15 JJ Thomson Avenue – Cambridge, UK CB3 OFD

E-mail: dayrellc@gmail.com, arnaldoc@icmc.usp.br, tiopalada@gmail.com, danilo.mjr@gmail.com, Ann.Copestake@cl.cam.ac.uk, valeria.feltrim@din.uem.br, seotagni@usp.br, sandra@icmc.usp.br

Abstract

The relevance of automatically identifying rhetorical moves in scientific texts has been widely acknowledged in the literature. This study focuses on abstracts of standard research papers written in English and aims to tackle a fundamental limitation of current machine-learning classifiers: they are mono-labeled, that is, a sentence can only be assigned one single label. However, such approach does not adequately reflect actual language use since a move can be realized by a clause, a sentence, or even several sentences. Here, we present MAZEA (*Multi-label Argumentative Zoning for English Abstracts*), a multi-label classifier which automatically identifies rhetorical moves in abstracts but allows for a given sentence to be assigned as many labels as appropriate. We have resorted to various other NLP tools and used two large training corpora: (i) one corpus consists of 645 abstracts from physical sciences and engineering (PE) and (ii) the other corpus is made up of 690 from life and health sciences (LH). This paper presents our preliminary results and also discusses the various challenges involved in multi-label tagging and works towards satisfactory solutions. In addition, we also make our two training corpora publicly available so that they may serve as benchmark for this new task.

Keywords: English abstracts, rhetorical moves, multi-label sentence classifier

1. Introduction

The relevance of identifying rhetorical moves in scientific texts has been widely acknowledged in the literature. This is mainly because rhetorical moves are viewed as a crucial element in the organization and structure of texts and as such can play a key role in genre-based pedagogies and writing tools. By *move*, we refer to "a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse" (Swales, 2004: 228).

One of the major challenges in the investigation of rhetorical moves is that manual annotation tends to be rather subjective and time-consuming. Not surprisingly, much effort has been made to develop systems to automatically identify rhetorical moves in scientific texts. Some systems (for instance, Teufel & Moens, 2002; Feltrim *et al.*, 2006; Genovês *et al.*, 2007) adopt a linguistic approach, that is, they make use of a linguistically rich set of features which covers lexical, syntactical and structural aspects. Examples of such features include: verb tense/mood/voice, semantic profile of verbs, negation, and stance markers. Other systems are language independent and resort to a bag of clusters with n-grams and statistical methods (e.g. Anthony & Lashkia, 2003; Pendra & Cotos, 2008).

The present study follows the linguistic approach and sets out to identify rhetorical moves in abstracts of standard research papers written in English. The topic has in fact received increasing attention (among others, Mcknight & Arinivasan, 2003; Shimbo *et al.*, 2003; Ito *et al.*, 2004; Yamamoto & Takagi, 2005; Wu *et al.*, 2006; Lin *et al.*, 2006; Ruch *et al.*, 2007; Hirohata *et al.*, 2008). This also includes our previous work (Genovês *et al.*, 2007) which has proposed AZEA[1] (*Argumentative Zoning for English Abstracts)*, a high-accuracy machine learning classifier that automatically detects rhetorical moves in English abstracts. AZEA relies on a set of 22 features and was trained using a corpus of 74 abstracts from the pharmaceutical sciences. These abstracts had been manually annotated according to the following moves: (i) background, (ii) gap, (iii) purpose, (iv) method, (v) result, and (vi) conclusion. These categories were determined on the basis of moves proposed by Hyland (2004:67), Swales (2004:228-238) and Swales & Feak (2009:4). The highest level of accuracy (80.3%) was achieved using the Support Vector Machine (Cortes *et al.*, 1995) method for well-structured published abstracts.

Like other machine-learning abstract classifiers of its kind, AZEA is mono-labeled, that is to say, a given sentence can

[1] http://www.nilc.icmc.usp.br/azea-web/

only be assigned one single label which, at least in principle, should be the most salient one. Such an approach is fully justifiable given that, in multi-label processing, the system would first need to decide whether the sentence should be mono- or multi-labeled. For the latter, there would then be the additional decisions on where to break the sentence as well as which order labels should appear. However, mono-label approaches have a fundamental limitation. They do not adequately reflect actual language use. As Swales (2004:229) explains, a move is "is better seen as flexible in terms of linguistic realizations" since it is "a functional, not a formal, unit" that can be realized by a clause, a sentence, or even several sentences.

This paper intends to address this issue. Here, we present MAZEA (*Multi-label Argumentative Zoning for English Abstracts*), a multi-label classifier which automatically identifies rhetorical moves in English abstracts and allows for a given sentence to be assigned as many labels as appropriate. Although much work is still needed to overcome the various challenges involved in this type of work, our initial results are especially promising. As we shall see shortly, MAZEA achieves satisfactory results in terms of deciding whether multi-label categorization applies as well as, if that is the case, which labels should be assigned.

The system focuses on two broad fields: (i) physical sciences and engineering (PE) and life and health sciences (LH). It is made up of two independent classifiers which have been trained on two separate corpora, manually annotated according to the six abovementioned rhetorical moves. Our PE corpus consists of 645 abstracts (144,683 tokens) and the LH corpus comprises 690 abstracts (50,248 tokens). This is in fact another important contribution of our study given that existing, related classifiers have used training corpora of up to 100 texts (e.g., Teufel & Moens, 2002; Anthony & Lashkia, 2003; Feltrim *et al.*, 2006; Genovês *et al.*, 2007).Table 1 shows the composition of these two corpora in terms of discipline considered and number of texts from each.

Physical Sciences and Engineering (PE)		**Life and Health Sciences (LH)**	
Physics	325	Dentistry	235
Computing	230	Pharmaceutical Sciences	195
Engineering	90	Biology	105
		Biophysics	105
		Bioengineering	25
		Biomedical Sciences	25
TOTAL	**645**	**TOTAL**	**690**

Table 1: Composition of the PE and the LH corpora by discipline

All abstracts were taken from research papers published by leading international academic journals. The selection of individual texts proceeded on the basis of authors' affiliation: either the first author or most authors should be affiliated to a department of the disciplines in question. Abstracts cutting across two or more disciplines investigated here were discarded. In addition, preference was given to papers by authors affiliated to universities in English-speaking countries.

This paper presents the results of our initial attempt to build a multi-label classifier to automatically identify rhetorical moves in English abstracts. We also discuss the limitations of the current system and the various issues raised throughout the process. Another major contribution of our study is that we also make our training corpora publicly available[2] so as to serve as benchmark for the task.

The remainder of this paper is organized as follows. The next section describes the process of annotation rhetorical moves in all abstracts. Section 3 explains our working environment, enumerating the various tools and algorithms we have selected to perform the task. The results are presented in section 4. We conclude with a discussion of our main contributions and the various issues we intend to address in future studies.

2. Corpus Annotation

Since our primary purpose was to build a classifier that would assign as many labels as appropriate to a given sentence, our initial challenge was to decide on when and how to segment sentences. In order to do so, we randomly selected five abstracts from each corpus and used a parser to divide all sentences into either prepositional phrases or clauses. Three human annotators were then asked to independently assign one of the following rhetorical moves to each segment: (i) background, (ii) gap, (iii) purpose, (iv) method, (v) result, and (vi) conclusion.

However, such task turned out to be far more complex than expected. The pre-established segmentation proved rather inefficient given that, in many cases, it did not match with how a human annotator would wish to segment the sentence. This is therefore the main reason why we abandoned the idea of parsing sentences.

At the same time, this initial phase provided empirical grounds for establishing the guidelines to be adopted to manually annotate rhetorical moves in abstracts and hence ensure consistency throughout the process. These guidelines basically explain the main aspects and characteristics of each move. Very briefly, these can be summarized as follows:

(i) **Background:** the context of the study, including any reference to previous work on the topic, relevance of the topic and main motivations behind the study;

(ii) **Gap:** any indication that the researched topic has not been explored, that little is known about it, or that previous attempts to overcome a given problem or issue have not been successful;

(iii) **Purpose:** the intended aims of the paper or hypotheses put forward;

(iv) **Method:** the methodological procedures adopted as

[2] http://www.nilc.icmc.usp.br/mazea-web/

well as the description of the data/materials used in the study. Specifications of the structure of the paper are mostly categorized as methods, taking into consideration that it would refer to how the purpose was achieved;

(v) **Result:** main findings or, in some cases, indication that the findings will be described or discussed; discussion or interpretation of the findings, which includes any hypothesis raised on the basis of the findings presented in the paper;

(vi) **Conclusion:** general conclusion of the paper; subjective opinion about the results; suggestions and recommendations for future work.

As regards sentences reflecting more than one move, we have decided to follow Swales' (2004:229) approach closely and view rhetorical moves as functional rather than grammatical units. Thus, rather than imposing syntactical boundaries, we have opted for allowing annotators to decide whether (or not) and how to segment sentences on the basis of their own subjective judgment. In other words, no criteria were set to determine where and how to segment multi-labeled sentences. For moves cutting across several sentences, it was simply a matter of repeating the label over all sentences.

To facilitate the process of manual annotation, we have resorted to the AZEA system (Genovês *et al.*, 2007) to automatically tag all abstracts from both corpora according to the abovementioned moves. As mentioned earlier, AZEA works at the sentence level and assigns one label per sentence.

The next step was then to randomly select 38 abstracts from the PE corpus and 34 abstracts from the LH corpus so that AZEA's categorization could be independently validated by the same three human annotators on the basis of the abovementioned guidelines. These figures represent 5% of the overall number of texts included in each corpus. In this case, abstracts were chosen taking into consideration the proportion of texts from each discipline.

This validation process involved correcting labels mistakenly assigned as well as errors related to its misinterpretation of sentence boundaries. In addition, annotators could also assign more than one label to a given sentence whenever they found it appropriate. Here is an example of a multi-label sentence:

<method> Bioinformatic analysis **</method> <result>** demonstrated that 7 of 12 breakpoints combined among 3 complex cases aligned with repetitive sequences, as compared to 4 of 30 breakpoints for the 15 deletion cases. **</result>**

The level of agreement among human annotators was measured by applying the Kappa Statistics (Carletta, 1996). This calculation was done at sentence level. Here, we have assessed whether annotators agreed on the labels assigned as well as on the segmentation of the sentence, if any. For multi-label sentences, the order of labels within the sentence was also compared. However, for the sake of simplicity, no consideration was given to whether annotators segmented sentences at exactly the same point. The Kappa Statistics yielded the following values: 0.652 (N=306, k=3, n=20) and 0.535 (N=148, k=3, n=18) for the LH and the PE corpora, respectively. These figures indicate substantial agreement in the annotation of the LH corpus and moderate agreement in the PE corpus (Landis & Koch, 1977). All in all, we can conclude that the multi-label sentence classification is reproducible, although disagreements should be settled.

Thus, after discussing the various issues raised in annotation of these for 38 abstracts from the PE corpus and 34 abstracts from the LH corpus, the same three annotators revised the abovementioned guidelines accordingly. This basically consisted of further explaining some key aspects and characteristics of each rhetorical move and providing examples for debatable points.

The remaining abstracts of the two corpora were then divided among five annotators who have revised AZEA's automatic categorization of rhetorical moves according to the criteria provided. These are precisely the versions of the corpora we have used to test and train our multi-label classifiers and the same ones we have made publicly available (see footnote 2).

3. Working Environment

Our multi-label classifier categorization consists of a pipeline of several NLP tools and algorithms (Figure 1).

In the **preprocessing phase**, all abstracts were first segmented, tokenized and lemmatized. For text segmentation, tokenization and PoS tagging, we have resorted to OpeNLP (http://opennlp.apache.org/). Word lemmatization was done by means of the Wordnet API (http://sourceforge.net/projects/jwordnet/).

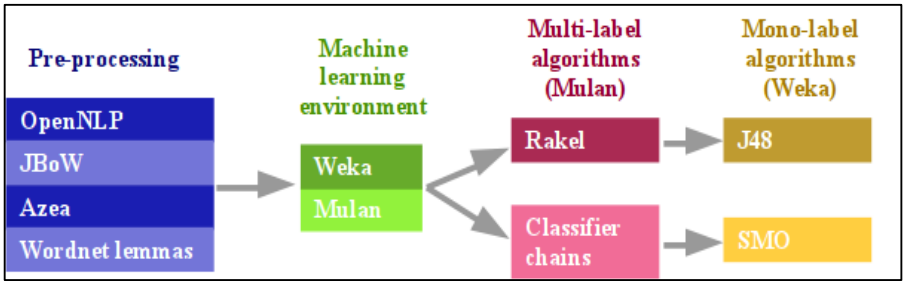

Figure 1: Processing Pipeline

The two corpora were each divided into two sets of similar size by randomly selecting the abstracts to be included in each. We then used one subcorpus from each area to automatically extract formulaic expressions (recurrent combinations of words), namely, PE-Exp and LH-Exp. The other two halves were used to train our classifiers in the machine learning phase and are referred to as PE-Training and LH-Training. This in other words means that the two subcorpora from which formulaic expressions were extracted are different from those used in the machine learning phase. Such division was made to avoid potential biases in the results.

To detect formulaic expressions, we have used the Jbow system (Machado Jr., 2009) which automatically identifies them on the basis of various statistical methods. The system then correlates the identified expressions with one (or more) of the above mentioned move categories: (i) background, (ii) gap, (iii) purpose, (iv) method, (v) result, and (vi) conclusion. This has therefore enabled us to generate six features, computed according to two main aspects: (i) the presence of a given formulaic expressions in the sentence under analysis and (ii) the statistical relation between such formulaic expressions and the labels.

In addition to these six lexical features, we have also selected six out of the 22 features used by Azea (Genovês *et al.*, 2007). These are: (i) position within the text; (ii) length; (iii) first formulaic expression (recurrent combination of words) related to the agent in the sentence; (iv) presence of a modal verb; (v) tense of the first finite verb; (vi) verb voice. Since many of these features are nominal, they had to be converted to a binary representation so that they could be adequately processed in the machine learning phase. The feature sentence length may serve as an example. If we assume that sentences can be of small, medium or big sizes, this leads us to consider three binary features: small_size, medium_size and large_size. However, only one would be deemed true for each sentence.

Thus, the initial six AZEA features have yielded a total of 46 numeric features. These features together with the other six generated by the JBow system have amounted to 52 numeric features, which were then used to train our classifiers.

In the **machine learning phase**, classifiers worked at the sentence level. Thus, when associated with more than one move, sentences would be assigned multiple labels. This phase was carried out in two stages. We first selected two learning algorithms from the Mulan library[3] (Tsoumakas *et al.*, 2010) which adopt various approaches for multi-label classification. These are: (i) Classifier Chain (Read *et al.* 2009), which processes each individual label separately; and (ii) Rakel (Tsoumakas *et al.* 2007), which groups correlated labels.

Since multi-label algorithms work on top of mono-label ones, we have also incorporated mono-label algorithms from Weka (Witten & Frank, 2005) into our classifiers. For doing so, we have selected two well established classifiers from both statistical (SMO) and symbolic (J48) fields. SMO was used together with Classifier Chain while J48 was combined with Rakel.

The two classifiers were trained and tested taking into consideration the following measurements:

- **LH:** in this case, only the LH corpus was used. Half of it (LH-Exp) provided the data from which formulaic expressions were retrieved, by means of the JBoW system. The other half (LH-Training) was used to test and train the two classifiers.
- **PE:** the same procedures were repeated for the PE corpus. Half of it (PE-Exp) was used to extract formulaic expressions and the other half (PE-Training) to test and train the two classifiers.
- **Mixed:** formulaic expressions were generated on the basis of half of the PE corpus only (PE-Exp) and the two classifiers were evaluated using the LH-Training subcorpus. The idea was to test whether formulaic expressions from a given area can be extended to another.

As baselines, we have considered the following:

(i) the expected accuracy of a random classifier applied to mono-label sentences: 16.66%, since the categorization takes into consideration six moves;

(ii) the most frequent move (method) in a mono-label classification for the two corpora altogether: 33.7%.

As gold standard, we have considered the kappa between human annotators.

4. Results

Within the framework of the present analysis, the first point to be made is that the vast majority of sentences from English abstracts reflect one single rhetorical move. Multi-label sentences accounted for 16.5% of all LH sentences (1,082 out of 6,544) and for 11.3% of all PE sentences (445 out of 3,933).

As regards the multi-label categorization task, the results from the Classifier Chain + SMO method were slightly better than those from the Rakel + J48. Figures 2 and 3 show the resulting values for the accuracy, micro-precision and hamming loss for the two methods employed in this study.

Here, we have opted for the Mulan example-based accuracy, which is an extension of classic accuracy and, in our view, more suitable for multi-label scenario. In this case, the number of labels correctly assigned is evaluated against the total number of predicted labels. This therefore allows us to estimate the chance that a given predicted label has to be accurate. For example, if the system suggests three labels for a given segment (an entire sentence or part of it) and only one is correct, the accuracy measure for that particular instance is 0.33. Thus, for the LH corpus, the Chain + SMO Classifier (Figure 2) has 69% of chance of assigning the correct label to a given sentence or part of it.

Under this perspective, the higher the number of labels within a given sentence, the lower the chance of automatically identifying all labels correctly. This is mainly because the chance of selecting the correct label (c) is estimated according to the number of labels associated with the sentence (c^n). If the chance of correctly selecting a single label is 60%, the chance of selecting two labels correctly is 36% (0.6 * 0.6). In our implementation,

[3] http://sourceforge.net/projects/mulan/

sentences can be assigned up to six labels, including "no label" for those cases when the classifier cannot decide which category the segment refers to ($0 \leq n \leq 6$).

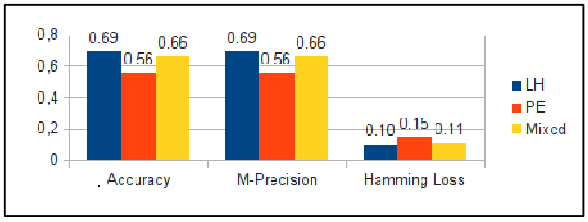

Figure 2: Classifier Chain + SMO evaluation

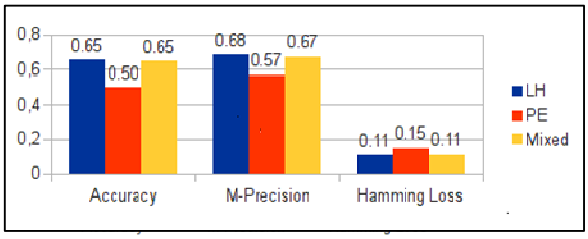

Figure 3: Rakel + J48 evaluation

As can be seen in Figures 2 and 3, the resulting accuracy was significantly higher than our reference baselines – 0.1666 for random classification and 0.337 when the most frequent move (method) is assigned –, irrespective of the method adopted. Ideal accuracy and micro-precision are both higher while hamming loss is lower.

We also found that, for both methods, performance was marginally worse for the PE in comparison with the LH corpus. This is in line with the measured kappa values for each corpus, since agreement among annotators was lower for the PE corpus. As expected, Mix results are not as good as LH ones. This is mainly because, for the former, the two classifiers were trained over the LH corpus using PE formulaic expressions.

The classifier which performed best (Chain + SMO) was also evaluated in relation to the kappa values obtained by comparing human annotation (0.652 (N=306, k=3, n=20) for the LH corpus and 0.535 (N=148, k=3, n=18) for the the PE corpus, see section 2 for details). For the Chain + SMO classifiers, the kappa analysis yielded the following values: 0.567 (N=306, k=4, n=20) and 0.409 (N=148, k=4, n=18) for the LH and PE corpus respectively. We consider these results as reasonably satisfactory, given that they are fairly close to the gold-standards.

An online demo Web application of our two Chain + SMO classifiers, one for each area (see footnote 2 for the website address). To obtain maximum performance, the PE-Exp and the PE-Training subcorpora served as the training corpora for developing the PE classifier while the LH-Exp and LH-Training collections were used to train the LH Classifier.

5. Conclusion

This paper has presented the results of our initial attempt to develop two machine learning systems to automatically identify rhetorical moves in English abstracts from (i) physical sciences and engineering and (ii) from life and health sciences. In addition, the systems were expected to assign as many labels as appropriate whenever a given sentence reflected more than one rhetorical move. Our a multi-label classifier – MAZEA – has produced encouraging results when deciding whether a given sentence was to be segmented and which labels should be assigned.

However, there is still room for improvement. To start with, for multi-label sentences, the systems are still unable to indicate their correct order. More importantly, our initial challenge of how to automatically segment sentences according to rhetorical moves remains unanswered. In future studies, we also intend to work towards enhancing our classifiers' performance by defining new features and exploring different approaches. It would be also particularly useful to further investigate the underlying regularities in the lexical and grammatical patterning of rhetorical moves in English abstracts. Additional benefits could also be obtained by improving human agreement on the task and hence refining our annotation guidelines.

Taking into account that this study is, to the best of our knowledge, the first attempt to build a multi-label sentence classifier, we conclude that our system's overall performance is particularly significant. This is in itself a major contribution given that such classifiers can effectively enhance the performance of natural language processing tools which are domain dependent. They are also an invaluable resource for linguists who wish to base their studies on large corpora by speeding up the arduous task of identifying and annotating moves.

Last but not least, it is also important to stress that our training corpora are much larger than those used by previous related studies and they have been made publicly available for other researchers.

6. Acknowledgements

We would like to thank FAPESP (Grant Reference numbers 2007/52405-3 and 2008/08963-4) and CNPq (201407/2010-8) for providing the financial support needed in this research.

7. References

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), pp.185--193.

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

Cortes, Corinna, Vapnik, Vladimir N. (1995). *Support-Vector Networks*, Machine Learning, 20.

Feltrim, V. D.; Teufel, S.; Nunes, M. G. V. and Aluísio, S.M. (2006). Argumentative zoning applied to critiquing novices' scientific abstracts. Computing Attitude and Affect in Text: Theory and Applications.

The Information Retrieval Series, 2006, Volume 20, pp. 233--246.

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

Hirohata, K.; Okazaki, N.; Ananiadou, S. and Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*. Asian Federation of Natural Language Processing, pp. 381--388.

Hyland, K. (2004). *Disciplinary Discourses*. Michigan: The University of Michigan Press.

Ito, T.; Simbo, M.; Yamasaki, T. and Matsumoto, Y. (2004). Semi-supervised sentence classification for medline documents. In *IPSJ SIG Technical Report*, v. 2004-ICS-138, pp. 141--146.

Lafferty, J.; McCallum, A. and Pereira, F (2001). Conditional random Fields: Probabilistic models for segmenting and labelling sequence data. In *Proceedings of the International Conference on Machine learning (ICML)*, Williams College, Williamstown, MA, USA, Morgan Kaufmann, pp. 282--289.

Landis, J. R. , Koch, G. G (1977). *The measurement of observer agreement for categorical data*. Biometrics, v. 33, pp. 159--174.

Lin, J.; Karakos, D.; Demner-fushman, D. and Khudanpur, S. (2006). Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BioNLP'06)*, Association for Computational Linguistics (ACL), pp. 65--72.

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

McKnight, L., Arinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *AMIA 2003 Symposium Proceedings*, American Medical Informatics Association, pp. 440--444.

Pendar, N. , Cotos, E. (2008). Automatic Identification of Discourse Moves in Scientific Article Introductions. In *Proceedings of The Third Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics (ACL), pp. 62-70.

Read, J.; Pfahringer, B.; Holmes, G. and Frank, E., (2009). Classifier Chains for Multi-label Classification. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases, Proceedings, Part II*, Lecture Notes in Computer Science, Vol. 5782, Subseries: Lecture Notes in Artificial Intelligence, Buntine, W.; Grobelnik, M.; Mladenic, D.; Shawe-Taylor, J. (Eds.), pp. 254--269.

Ruch, P.; Boyer, C.; Chichester, C.; Tbahriti, I.; Geissbuhler, A.; Fabry, P.; Gobeill, J.; Pillet, V.; Rebholz-Schuhmann, D.; Lovis, C. and Veuthey, A. L. (2007). *Using argumentation to extract key sentences from biomedical abstracts*. International Journal of Medical Informatics, v. 76, pp. 195--200.

Shimbo, M.; Yamasaki, T. and Matsumoto, Y. (2003). Using sectioning information for text retrieval: a case study with the medline abstracts. In *Proceedings of the 2nd International Workshop on Active Mining (AM'03)*, Maebashi, Japan; 2003 , pp. 32--41.

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

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

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

Tsoumakas, G., Vlahavas, I. (2007). Random k-Labelsets: An Ensemble Method for Multilabel Classification, In *Proceedings of the* 18th European Conference on Machine Learning (ECML 2007), Lecture Notes in Artificial Intelligence 4701, Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenic (Eds), pp. 406--417.

Tsoumakas, G.; Katakis, I. and Vlahavas, I. (2010). *Mining Multi-label Data*. Data Mining and Knowledge Discovery Handbook, O. Maimon, L. Rokach (Ed.), Springer, 2nd edition,.

Witten, I. H., Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. 2nd ed. San Francisco: Morgan Kaufmann.

Wu, J.; Chang, Y.; Liou, H. and Chang, J. S. (2006). Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 41--44.

Yamamoto, Y., Takagi, T. (2005). A sentence classification system for multi-document summarization in the biomedical domain. In *Proceedings of the 21st International Conference on Data Engineering Workshops (ICDEW'05), IEEE Computer Society Washington, DC, USA,* pp. 90--95.