On the Discoursive Structure of Computer Graphics Research Papers

Beatriz Fisas, Francesco Ronzano and Horacio Saggion
Natural Language Processing Group, Pompeu Fabra University, Barcelona, Spain
{beatriz.fisas,francesco.ronzano,horacio.saggion}@upf.edu

Abstract

Understanding the structure of scientific discourse is of paramount importance for the development of appropriate Natural Language Processing tools able to extract and summarize information from research articles. In this paper we present an annotated corpus of scientific discourse in the domain of Computer Graphics. We describe the way we built our corpus by designing an annotation schema and relying on three annotators for manually classifying all sentences into the defined categories. Our corpus constitutes a semantically rich resource for scientific text mining. In this respect, we also present the results of our initial experiments of automatic classification of sentences into the 5 main categories in our corpus.

1 Introduction

Understanding the internal organization of text documents is important for many content assessment tasks such as summarization or information extraction. Several studies have investigated the structure and peculiarities of scientific discourse across distinct domains, such as biology (Mizuta and Collier, 2004), chemistry and computational linguistics (Teufel et al., 2009), or astrophysics (Grover et al., 2004). The coherence of the argumentative flow that authors adopt to expose scientific contents is essential to properly contextualize these contents, to characterize their connections with related pieces of research as well as to discover relevant aspects, novelties and future directions.

Because of both the huge, growing amount of scientific literature that is accessible online and the complexity that often characterizes scientific discourse, currently researchers and professionals are experimenting more and more difficulties when trying to keep themselves up to date.

The analysis of the internal organization of the scientific discourse and the identification of which role each piece of text plays in the scientific argument contribute to structure and thus ease the interpretation of scientific information flow. In addition, the explicit characterization of scientific discourse provides useful meta-information to support tasks like targeted information extraction, content retrieval and summarization.

Although several studies have characterized scientific domains, the area of Computer Graphics, a sub-field of Computer Science, has not been studied in previous work. We have developed an annotation scheme and produced an annotated corpus of scientific discourse in this domain.

The rest of the paper is structured as follows: After a review of previous work in the next section, we present and motivate the annotation scheme in section 3, describing the corpus dataset in section 4. We provide details of annotation process in section 5, followed by the values of the attained inter-annotator agreement and an analysis of the structure of the resulting corpus in section 6. Finally, before closing the paper with conclusions and future work, we explain our first experiments in automatic sentence classification in section 7.

2 Scientific discourse characterization: related work

The analysis and annotation of scientific discourse has been approached from different points of view in previous works.

Although the focus of the analysis is manifold and spans along different linguistic concepts, the scientific discourse annotation schema we propose in this paper builds upon the proposals of Teufel (1999; 2009; 2010) and Liakata (2010) hence the following subsections describe in more detail their contributions.

Simone Teufel's model (Teufel, 1999; Teufel and Moens, 2002; Teufel et al., 2009), which was named Argumentative Zoning, focuses on knowledge claims and is based on previous schemes for

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 42–51,
Denver, Colorado, June 5, 2015.

classifying the citation functions (Garfield, 1965; Melvin Weinstock, 1971; Spiegel-Rösing, 1977).

Liakata (2010) analyses the content and conceptual structure of scientific articles with an ontology-based annotation scheme, the Core Scientific Concepts scheme (CoreSc). Closely related to this approach is the multidimensional scheme of Nawaz (2010), tailored to bioevents, and the works of De Waard (2009) in classifying sentences in 5 epistemic types and White (2011), who concentrates on identifying hypothesis, explanations and evidence in the biomedical domain.

In terms of scope, abstracts, considered to be a brief summary of the whole article, have been the object of research in the works of Guo (2010), Lin (2006), Ruch (2007), Hirohata (2008) and Thompson (2009).

Among researchers who explore full articles, Lin (2006) and Hirohata (2008) have based their analysis on section names, offering a coarse-grained annotation, while Liakata (2010; 2012), Teufel (2009) and Shatkay (2008) adopt a finer-grained approach.

The annotation unit is also a controversial matter. While most researchers agree to classify sentences into categories (Liakata and Soldatova, 2008; Liakata et al., 2010; Teufel and Moens, 2002; Teufel et al., 2009; Lin et al., 2006; Hirohata et al., 2008), others segment sentences into smaller discourse units (Shatkay et al., 2008; DeWaard, 2009).

Bioscience is by far the most studied domain and acts as a motor for research in information extraction from scientific publications (Mizuta et al., 2006; Wilbur et al., 2006; Liakata et al., 2010). Nevertheless, some work has also been done in the Computational Linguistics and Chemistry domains, where Teufel (2009) has implemented her AZ-II extended annotation scheme.

2.1 Argumentative Zoning - AZ

Teufel's main assumptions are that scientific discourse contains descriptions of positive and negative states, refers to other's contributions, and is the result of a rhetorical game intended to promote the authors contribution to the scientific field. In fact, Teufel argues that scientific texts should make clear what the new contribution is, as opposed to previous work and background material.

From a theoretical point of view she develops the Knowledge Claim Discourse Model (KCDM) which she adapts into three annotation schemes: Knowledge Claim Attribution (KCA), Citation Function Classification (CFC) and Argumentative Zoning (AZ).

Teufel annotates a corpus of Computational Linguistics papers with the first version of Argumentative Zoning (AZ) (Teufel and Moens, 2002). She later extends the AZ scheme for annotating chemistry papers, thus creating a new version, the AZ-II, with 15 categories (Teufel et al., 2009) instead of the first 7 in AZ.

The AZ-II annotated corpus consists of 61 articles from the Royal Society of Chemistry.

2.2 Core Scientific Concepts - CoreSc

Liakata (2010) believes that a scientific paper is a human-readable representation of a scientific investigation and she therefore seeks to identify how and where the components of a scientific research are expressed in the text.

As Teufel, Liakata also proposes a sentence-based annotation for scientific papers, but unlike Teufel, who proposes a domain independent annotation scheme based on argumentative steps, Liakata's scheme supports ontology motivated categories representing the core information about a scientific paper.

It was constructed with 11 general scientific concepts based on the EXPO ontology (Soldatova and King, 2006), which constitute the first layer of the annotation. The second layer allows the annotation of properties (New/Old, Advantage/Disadvantage) of certain sentences labeled in the first layer. Finally, in the third layer, several instances of a concept can be identified.

With the CoreSC annotation scheme and guidelines, Liakata's team produced the CoreSC corpus, constituted by 265 annotated papers from the domains of physical chemistry and biochemistry.

Liakata (2010) compares her approach to Teufel's and concludes that they are complementary and that combining the two schemes would be beneficial. They are both computational-oriented as the annotated corpora are intended to serve as a basis for linguistic innovative technologies such as summarisation, information extraction and sentiment analysis. CoreSC is more fine-grained in content-related

categories while AZ-II covers aspects of knowledge claims that permeate across several CoreSC concepts.

Corpora annotated with Argumentative Zoning-II (Teufel et al., 2009) and Core Scientific Concepts (Liakata et al., 2010) have been exploited to build automatic rhetorical sentence classifiers.

3 Scientific Discourse Annotation Scheme

3.1 The domain: Computer Graphics

Computer Graphics is a vast field which includes almost anything related to the generation, manipulation and use of visual content in the computer. It is a relatively young discipline which has not been yet described in terms of its discourse, which differs mainly from the Bioscience's discourse in its much more mathematical content.

Research in Computer Graphics is based on multiple technical backgrounds, (mainly Physics, Mechanics, Fluid Dynamics, Geometry, Mathematics) and its results are the development of practical applications for their exploitation in several industries.

Scientific publications in Computer Graphics reflect the characteristics of this domain. It is expected that they include a section where a theoretical model is presented in detail - with algorithms, equations, algebra and mathematical reasoning - and a section where a computational experiment demonstrates an application that contributes to the knowledge in the area or to enhance techniques already in use in the mentioned industries. Experiments in computational sciences are basically algorithmical and do not include materials nor physical processes in laboratories.

3.2 The annotation scheme design

We defined our Scientific Discourse Annotation Schema by relying on both Teufel's and Liakata's annotation schemas and contributions. In particular, we extended and enriched Liakata's CoreSc scheme at this first stage, leaving the knowledge claim approach for a second stage.

A thorough review of the previous work in annotation of scientific publications as well as the analysis of the contents of papers in our domain, lead us to select 9 categories from Liakata's annotation scheme and the Discourse Elements Ontology (DEO), which were later increased to a total of 16, in order to cover the scientific concepts that might appear in an article.

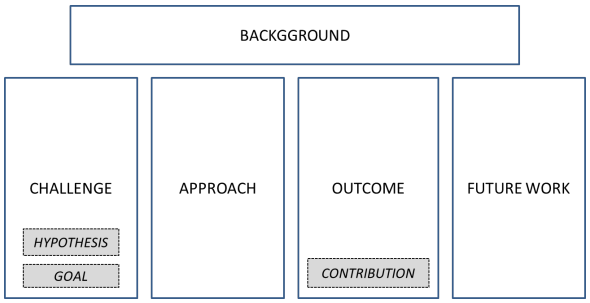

Figure 1: Simplified Annotation Scheme: 5 categories and 3 subcategories

However, this first scheme proved to be too complex, and we agreed to follow an annotation workflow characterized by subsequent steps with different levels of granularity. Thus, the corpus annotation process should go through a first coarse-grained phase and later increase the level of details with a finer-grained annotation scheme.

The 16 categories of the extended scheme were grouped into 5 main categories (Fig. 2).

Nevertheless and in order to provide the annotated corpus with more detailed information, we decided to leave annotators the possibility to specify three especially significant sub-categories: Hypothesis, Goal and Contribution.

Fig.1 shows the final version of our scientific discourse simplified annotation scheme.

4 Corpus Dataset for annotation: Data collection and Annotation unit

To populate our corpus we randomly selected a set of 40 documents, available in PDF format, among a bigger collection provided by experts in the domain, who pre-selected a representative sample of articles in Computer Graphics. Articles were classified into four important subjects in this area: Skinning, Motion Capture, Fluid Simulation and Cloth Simulation. We included in the corpus 10 highly representative articles for each subject.

The annotation is sentence based as we have considered sentences to be the most meaningful minimal unit for the analysis of scientific discourse, in agreement with earlier work.

CHALLENGE: The current situation faced by the researcher: it will normally include a Problem Statement, the Motivation, a Hypothesis and/or a Goal.
BACKGROUND:This section presents all the information which is helpful for understanding the situation or problem that is the subject of the publication. It will include sentences that state widely accepted knowledge in the domain (Common Ground) as well as previous related work (Related Work).
APPROACH: In this section the author explains HOW he intends to carry out the investigation. He may refer to a theoretical model or framework (Model), give some or many details of the experimental setup (Experiment), point to some data/phenomena observed during the experimentation (Observations) or comment on his decisions for choosing this methodology (Method).
OUTCOME: Here the author offers the study findings: measurable data without discussion (Results), an interpretation or analysis of the results in support of the conclusion (Discussion), how the research will contribute to the current knowledge in the field (Contribution) and an overall conclusion that should reject or support the research hypothesis (Conclusion). Any comments on the limitations of the authors work will also be included in the OUTCOME section.
FUTURE WORK: In most articles, the author will suggest or recommend further research to improve or extend his own work.

Figure 2: Description of the 5 categories of our Simplified Discourse Annotation Scheme

5 The Annotation Process

5.1 Annotators

The annotators are not domain experts. Two of them are computationally oriented linguists and the third is both a linguist and the developer of the annotation scheme. Each of them has annotated the whole set of documents. Therefore, the annotation outcome is a collection of 40 papers whose sentences have been annotated by the three annotators. The categories associated to each sentence by each annotator are then merged to create the Gold Standard version of the corpus.

5.2 Annotation Task

The 40 documents selected for our corpus were provided to each annotator so as to start the sentence annotation process. All the annotators use GATE v.7.1 as annotation tool, with a customized view where they have a window with the ready-to-annotate documents, segmented into sentences. Their task is to select a sentence and choose the appropriate category from a pop-up list.

Each sentence of each document of the Corpus is classified as belonging to a category among: *Approach*, *Background*, *Challenge*, *Challenge_Goal*, *Challenge_Hypothesis*, *FutureWork*, *Outcome* or *Outcome_Contribution*. Sentences were classified as *Unspecified* when the identification of the category was not possible (for example, metadiscourse or acknowledgements) or as *Sentence* when the selected text was characterized by segmentation or character encoding problems (for example, when a footnote appears incorrectly in the text flow).

5.3 Annotation Support

In order to ensure the quality of the annotation, the annotators were provided with the following support: an introductory training session, a visual schema of the proposed discourse structure, guidelines for the annotation, a series of conflict resolution criteria and recommendations. Moreover, two follow-up conflict-resolution meetings were scheduled to perform inter-annotator agreement checks along the first stages of the annotation process.

5.4 Annotation Workflow

After the training session, the annotators were encouraged to test the tool, and try the schema with a couple of documents before the annotation task really started. Once the process was triggered, two conflict resolution meetings were scheduled after the annotation of the first 5 papers, and after the subsequent 10 papers. Agreement was measured in these two milestones in order to detect deviations in an early stage. The articles were sorted by subject, to facilitate the better comprehension of the text for the annotators, as articles concerning the same subject

Category	Annotated Sent.	%
Approach	5,038	46.70
Background	1,760	16.32
Challenge	351	3.25
Challenge_Goal	91	0.84
Challenge_Hypotesis	7	0.06
FutureWork	136	1.26
Outcome	1,175	10.89
Outcome_Contribution	219	2.03
Unspecified	759	7.04
Sentence	1253	11.61
Total	10,789	100

Table 1: Number/Percentage of sentences per category

deal with similar concepts and terminology.

6 Annotation Results

6.1 Annotated corpus description

The Corpus includes 10,789 sentences, with an average of 269.7 sentences per document.

We are currently defining the best approach to make Corpus annotations available to the research community, since most of its 40 documents are protected by copyright.

The Gold Standard was built with the following criteria for each sentence: If all annotators or two of them assigned the same category to the sentence, it was included in the Gold Standard version with such category; otherwise, the category selected by the annotator who designed the scheme was preferred and used in the Gold Standard. Table 1 details the number of sentences of each category in the Gold Standard version of the annotated corpus and its percentage in reference to the total number of annotated sentences in the whole corpus.

6.2 Inter-annotator Agreement Values

We used Cohen κ (Cohen et al., 1960) to measure the inter-annotator agreement. Cohen κ is an extensively adopted measure to quantify the inter-annotator agreement, previously exploited in several other annotation efforts, including the corpora created by Liakata and Teufel, previously introduced.

Depending on how documents are combined, there are several options for calculating the agreement measures over a corpus. Micro averaging essentially treats the corpus as one large document, whereas macro averaging calculates on a per document basis, and then averages the results. Macro averaging tends to increase the importance of shorter documents.

	κ	N	n	k	domain
Liakata	0.57	255	11	9	Biochem.
Liakata	0.50	5022	11	9	Biochem.
Teufel	0.71	3745	15	3	Chemistry
Teufel	0.65	1629	15	3	Comp.Ling.
Teufel	0.71	3420	7	3	Comp.Ling.

Table 2: Summary of κ values in previous works: N=#sentences, n=#categories, k=#annotators

In our corpus, the κ value of inter-annotator agreement (Cohen's κ), averaged among all annotators' pairs, considering the 5 categories and the 3 subcategories of our Simplified Annotation Schema (see Figure 1) is equal to 0.6567 for the macro average and 0.6667 if the micro average is computed. If we consider only the 5 top categories of our Simplified Annotation Schema the inter-annotator agreement grows: the macro average becomes 0.674 and the micro average 0.6823. In both cases, the micro average is slightly greater than the macro average since there are documents with a number of sentences below the mean (269.7 sentences per document) that are characterized by low κ values, thus negatively affecting the macro-averaged computation of κ.

These κ values are comparable to those achieved by Teufel for 1,629 sentences in the domain of Computational Linguistics, with an annotation scheme of 15 categories and 3 annotators (see Table 2). The micro average κ achieves the cut-off point of 0.67, over which agreement is considered difficult to reach in linguistic annotation tasks (Teufel, 2010).

The agreement measures in the 2 milestones, showed evolution of the inter-annotator agreement throughout the annotation process: Cohen's κ is substantially stable between two of the annotators, while the third annotator sensibly improves his agreement with the other two very quickly in the first 5 documents and remains stable after the second milestone. In particular, the annotator with the lowest agreement in the initial stage increased his

agreement with the other two annotators respecively from 0.59 for the first 5 documents to 0.68 for the last 25 documents and from 0.56 for the first 5 documents to 0.66 for the last 25 documents.

An analysis of the sentence distribution according to their agreement degree results in the following values: totally agreed sentences (65.09%), partially agreed sentences (31.24%) and totally disagreed sentences (3.66%).

Not all the categories are equally distributed, as each one of them has its own characteristics in terms of number of sentences, ambiguity or conflicts with other categories.

Background and *Approach*, the most highly represented categories, are highly reliable. In fact, more than 45% of the sentences of the corpus were tagged with agreement by the three annotators pairs as *Approach* or *Background*. If we also take into account the sentences with partial agreement (2 annotators agreed), then sentences classified as *Approach* and *Background* are more than 60% in the Gold Standard version of our annotated corpus.

FutureWork and *Outcome* are quite reliable, although the difference between them is that the ratio of totally agreed/partially agreed is considerably higher in *FutureWork* compared to the same ratio in *Outcome* (3.3 vs 0.9). This is due to the fact that although *FutureWork* sentences (1.3%) are much fewer than *Outcome* sentences (10.9%), those are much more easily recognized, as they include specific lexical clues (*for further research, in future investigation, more research is needed in, it could be interesting to, a better understanding,etc.*).

Clearly, *Challenge* is the category where the proportion of total disagreement is higher. This category which tends to appear at the beginning of a scientific paper shows more than any other the author's skills in writing, synthesis and ability to communicate the scope of the challenge they are presenting. Authors must be able to provide a context and outline the situation in order to attract the attention of the reader, who must understand the goal and complexity of the research.

When studying the relation between the number of sentences of a category and the annotation match between annotators, data reveal that the observed agreement among annotator pairs varies considerably according to the relative frequency of the annotation classes in the Corpus.

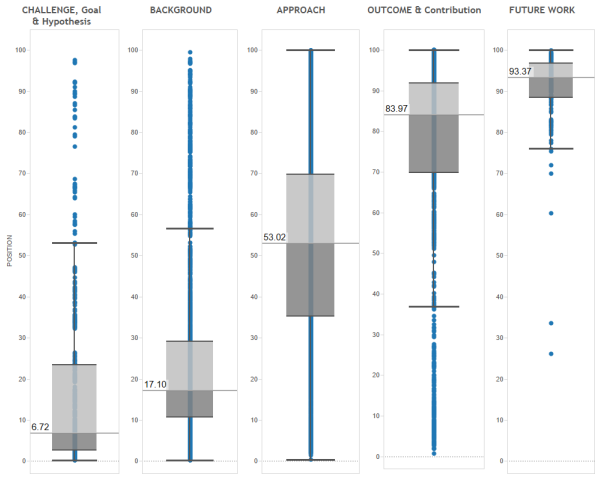

Figure 3: Box plots that show the distribution of the sentences of the 5 main categories of the Scientific Discourse Annotated Corpus

Agreement improves as the number of sentences of the category increases, getting close to 0.80 for the most frequent categories.

6.3 Discoursive Structure Analysis

The box plots of the 5 main categories (Fig. 3) give a clear picture of the discoursive structure of an average scientific paper in the Computer Graphics domain. In fact, the 5 main categories show a neat layout of the main zones (inside the box) in the argumentative structure distributed along the article. Even if one can find all types of sentences along the whole document, the central 50% of each category seems clearly limited to a zone with little overlapping of one another. When searching for information about one of these categories, a reader or researcher will find the central 50% of the sentences of each category in the following article length ranges: *Challenge* in between the 3% and 23%, *Background* in between the 11% and 29%, *Approach* in between the 35% and 70%, *Outcome* in between the 70% and 92%, *FutureWork* in between the 88% and 97%.

The identification of these ranges will allow readers, scientists, search engines, etc. to focus the exploring effort in a specific area of the article.

7 Automatic sentence classification: initial experiments

Recently several approaches to the automatic classification of the discursive function of textual excerpts from research papers have been proposed (Merity et al., 2009; Liakata et al., 2012; Guo et al., 2013). We present our initial experiments of automatic sentence classification with our Corpus. We describe the set of features we use to model and thus to characterize the contents of each sentence in order to enable the execution of proper classification algorithms. In particular, by relying on these features, we compare the performances of two classifiers: Logistic Regression (Wright, 1995) and Support Vector Machine (Suykens and Joos, 1999).

7.1 Description of sentence features

In order to support the extraction of the features that should characterize each sentence, we mine its contents by means of a pipeline of natural language processing tools, properly customized so as to deal with several peculiarities of scientific texts. As a consequence we are able to automatically extract from each sentence:

- **inline citation markers** - like *(AuthorA et al., 2010)* or *[11]*;
- **inline citation spans** that are text spans made of one or more contiguous inline citation markers. Examples of inline citation spans including one incline citation marker are: *(ALL2011)* or *[11]*. Examples of text spans including more than one inline citation marker are: *[10, 12]* or *(AuthorA. and AuthorB, 2010; AuthorC, 2014)*;
- for each inline citation span, if it has or not a **syntactic role**. For instance, in the sentence *[11, 12] demonstrate the theorem*, the inline citation span *[11, 12]* has a syntactic role since it is the subject of the sentence. In the sentence *We exploited the ABA method [14]*, the inline citation span *[14]* has no syntactic role.

We process each sentence by a MATE dependency parser (Bohnet, 2010) to determine its syntactic structure. A customized version of the parser is exploited to properly deal with the presence of inline citations. In particular, inline citations spans are excluded from the dependency tree if they have no syntactic functions in the sentence where they are present. After dependency parsing is performed, it is possible to identify the token of each sentence together with their Part-Of-Speech and syntactic relations.

- **unigrams**, **bigrams** and **trigrams** built from the lemmas of each sentence, lowercased and without considering stop-words. We included only unigram, bigrams and trigrams with corpus-frequency equal or greater than 4;
- **depth** and **number of edges by edge type** of the dependency tree;
- **dependency tree tokens** with corpus-frequency equal or greater than 4. Each dependency tree token is the result of the concatenation of three parts: kind of dependency relation, lowercased lemma of the source and lowercased lemma of the target of the dependency relation. For instance, one of the dependency tree tokens of the sentence *We demonstrate the theorem* is: *SBJ_we_demonstrate*, because "we" is the subject (SBJ) of the verb "demonstrate";
- **number of inline citation markers**;
- **number of inline citation spans that include two or more contiguous inline citation markers**;
- **number of citations with a syntactic role**;
- **position of the sentence in the document**, by dividing the document in 10 unequal segments (referred to as Loc. feature in (Teufel, 1999));
- **position of the sentence in the section**, by dividing the section into 7 unequal slices (referred to as Struct-1 feature in (Teufel, 1999));
- **category of the previous sentence**. We use gold standard previous sentence categories in our experiments.

7.2 Classification experiments

By relying on the features just described, we compare the sentence classification performances of two

Category	**Logistic Regression**	**SVM**
Approach	0.876	0.851
Background	0.778	0.735
Challenge	0.466	0.430
Future Work	0.675	0.496
Outcome	0.679	0.623
Avg. F1:	**0.801**	**0.764**

Table 3: F1 score of 10-fold cross validation of Logistic Regression and SVM - 10 fold cross validation over 8,777 manually classified sentences.

classifiers: Logistic Regression and Support Vector Machine with linear kernel. From our corpus we consider the set of 8,777 sentences that have been manually associated to one of the 5 high level classes of our scientific discourse annotation schema (see Figure 1): *Background*, *Challenge*, *Approach*, *Outcome*, and *Future Work*. We collapse the subcategories *Hypothesis* and *Goal* into the parent category *Challenge* and the sub-category *Contribution* into the parent category *Outcome*. We perform a 10-fold cross validation of the two classification algorithms, over the collection of 8,777 sentences. The results are shown in the Table 3.

The Logistic Regression classifier outperforms the SVM one both globally and by considering each single category. We can note that in general the F1 score obtained in each category decreases as the number of training instances does. This trend is not confirmed by the category *Future Work*. The corpus includes 136 sentences that belong to the category *Future Work*. This number is considerably lower than the 449 examples of *Challenge* sentences and the 1,175 examples of *Outcome* sentences. Anyway, the Logistic Regression F1 score of the category *Future Work* (0.675) is almost equal to the one of the category *Outcome* (0.679) and considerably higher than the F1 score of the category *Challenge* (0.446). This happens because some linguistic features that characterize *Future Work* sentences are strongly distinctive with respect to the elements of this class. For instance, the use of the future as verb tense as well words like *plan*, *future*, *venue*, etc. consistently contribute to automatically distinguish *Future Work* sentences, even if we have few training examples in our corpus.

8 Conclusions and Future Work

We have developed an annotation scheme for scientific discourse, adapted to a non-explored domain, Computer Graphics. We relied on the 5 categories and 3 subcategories of our annotation schema to manually annotate the sentences of a scientific discourse corpus made of 40 papers.

We have observed that the larger categories (in terms of number of sentences) - *Approach*, *Background* and *Outcome* - are highly predictable, while *Challenge*, which corresponds mainly with the introductory part of the scientific discourse is more heterogeneous and highly dependable of the author's style. Sentences classified as *FutureWork* have special lexical characteristics as confirmed by the results of our automatic classification experiments. We have also characterized specific zones for each of the 5 categories, thus contributing to a deeper knowledge of the internal structure of the scientific discourse in Computer Graphics.

In future we plan to focus on the characterization of other peculiarities of scientific text, including citations, thus properly extending our annotation schema. We are also confident that our Simplified Annotation Scheme will be suitable in other domains, and are therefore planning to verify it. A two-layered annotation scheme could then be applicable to most domains, the first layer being coarse-grained and general, and a second layer being finer-grained and domain-dependent for certain categories.

As future venues of research concerning automatic sentence classification, we are planning to carry out more extensive experiments and evaluations by increasing the set of features that describe each sentence, evaluating the contributions of single features and considering new classification algorithms.

Acknowledgments

The research leading to these results has received funding from the European Project Dr. Inventor (FP7-ICT-2013.8.1 - grant agreement no 611383).

References

Bernd Bohnet. 1999. *Very high accuracy and fast dependency parsing is not a contradiction*. Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.

Paolo Ciccarese, Elizabeth Wu, Gwen Wong, Marco Ocana, June Kinoshita, Alan Ruttenberg, and Tim Clark. 2008. *The SWAN biomedical discourse ontology*. Journal of Biomedical Informatics, 41, (5):739–751.

J. Cohen. 1960. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement, 20(1):(37).

Anita de Waard, Paul Buitelaar and Thomas Eigner 2009 *Identifying the Epistemic Value of Discourse Segments in Biology Texts* Proceedings of the Eighth International Conference on Computational Semantics, IWCS-8 '09, 351–354,Stroudsburg,PA, USA, Association for Computational Linguistics.

Eugene Garfield 1965. *Can Citation Indexing Be Automated?*, Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, National Bureau of Standards Miscellaneous Publication volume 269, 189–192 Prentice-Hall, Englewood Cliffs, NJ.

Claire Grover, Ben Hachey, and Ian Hughson. 2004. *The HOLJ Corpus: supporting summarisation of legal texts.*. Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC-04) Geneva, Switzerland.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun and Ulla Stenius. 2010. *Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes*. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing:99–107, Uppsala, Sweden. Association for Computational Linguistics.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen 2013. *Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review*. (Bioinformatics 29.11): 1440-1447

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. *Identifying sections in scientific abstracts using conditional random fields*. In Proceedings of the IJCNLP 2008, p.381–388.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. *Automatic recognition of conceptualization zones in scientific articles and two life science applications*. Bioinformatics, 28,(7):991–1000.

Maria Liakata and Larisa Soldatova. 2008. *Guidelines for the annotation of general scientific concepts*. Aberystwyth University, JISC Project Report, http://ie-repository. jisc. ac. uk/88.

Maria Liakata, Simone Teufel, Advaith Siddharthan and Colin Batchelor. 2010. *Corpora for the Conceptualisation and Zoning of Scientific Papers*. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta, may 2010 Nicoletta Calzolari et al., European Language Resources Association (ELRA).

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. *Generative Content Models for Structural Analysis of Medical Abstracts*. In Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis, BioNLP '06, p.65–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen Merity, Tara Murphy, and James R. Curran 2009. *Accurate argumentative zoning with maximum entropy models*. Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries Association for Computational Linguistics

Yoko Mizuta and Nigel Collier. 2004. *Annotation scheme for a rhetorical analysis of biology articles*. Proceedings of the Fourth International Conference on Language and Evaluation (LREC2004),1737–1740 , Lisbon, Portugal. European Language Resources Association (ELRA).

Yoko Mizuta, Anna Korhonen, Tony Mullen and Nigel Collier. 2006. *Zone analysis in biology articles as a basis for information extraction*. International Journal of Medical Informatics, 75(6):468–487.

Raheel Nawaz, Paul Thompson, John McNaught, and Sophia Ananiadou 2010 *Meta-Knowledge Annotation of Bio-Events*, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, May, 2010. European Language Resources Association (ELRA).

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbhler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis and Anne-Lise Veuthey. 2007. *Using argumentation to extract key sentences from biomedical abstracts*. International Journal of Medical Informatics,76,(2-3):195–200.

Hagit Shatkay, Fengxia Pan, and Andrey Rzhetsky, and W. John Wilbur. 2008. *Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users*. Bioinformatics, 24, 18:2086–2093.

Larisa N. Soldatova and Ross King. 2006. *An ontology*

of scientific experiments. Journal of the Royal Society Interface, 3 (11):795–803.

Ina Spiegel-Rösing. 1977. *Science Studies: Bibliometric and Content Analysis*, 7 (1). Social Studies of Science, 97–113.

Johan AK Suykens and Vandewalle Joos. 1999. *Least squares support vector machine classifiers*. Neural processing letters 9.3 (1999): 293-300.

Simone Teufel 1999. *Argumentative Zoning: Information Extraction from Scientific Text*, School of Cognitive Science, University of Edinburgh, UK.

Simone Teufel, 2010 *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization, CSLI Publications (CSLI Studies in Computational Linguistics), Stanford, CA*.

Simone Teufel and Marc Moens 2002 *Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status*. Computational Linguistics, 28, (4), 409–445.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. *Towards Discipline-independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3,EMNLP '09, Singapore, 1493–1502, Association for Computational Linguistics, Stroudsburg, PA, USA.

Paul Thompson, Syed A. Iqbal, John McNaught, and Sophia Ananiadou. 2009. *Construction of an annotated corpus to support biomedical information extraction*. BMC Bioinformatics, 10:349.

Melvin Weinstock 1971. *Citation indexes*, Encyclopedia of Library and Information Science, 5,16–40. Marcel Dekker, Inc., New York.

Elizabeth White, K. Bretonnel Cohen, and Larry Hunter. 2011. *Hypothesis and Evidence Extraction from Full-Text Scientific Journal Articles*. Proceedings of BioNLP 2011 Workshop:134–135, Portland, Oregon, USA, Association for Computational Linguistics.

W.John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. *New directions in biomedical text annotation: definitions, guidelines and corpus construction*. BMC Bioinformatics, 7:356.

Raymond E. Wright, 1995. *Logistic regression*.