This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal, 62(4/5), 2010.

Semantic metadata annotation. Tagging Medline abstracts for enhanced information access.

Ibekwe-SanJuan Fidelia
ELICO - University of Lyon 3
4 cours Albert Thomas, 69008 Lyon.
France
ibekwe@univ-lyon3.fr

Biography

Fidelia Ibekwe-SanJuan is currently a professor at the Information-Communication Department of the Jean Moulin University in Lyon, France. Her research revolves around the use of language models and terminology engineering for knowledge organization and for information retrieval. Of particular interest to her is the design of methodologies to assist the creation of advanced forms of knowledge organization (KO) systems from texts (topic maps, ontologies, SKOS, thesauri). She is also interested in the epistemological foundations of KO and LIS (Library and Information Science) and in particular, how research in KO intersects with other related fields such as LIS, Natural Language Processing (NLP) and Information Retrieval (IR).

Abstract.
Purpose. The object of this study is to develop methods for automatically annotating the argumentative role of sentences in scientific abstracts. Working from Medline abstracts, we classified sentences into four major argumentative roles: objective, method, result, conclusion. The idea is that if the role of each sentence can be marked up, then this metadata can be used during information retrieval to seek for particular types of information such as novelty, conclusions, methodologies, aims/goals of a scientific piece of work.

Methodology. Two approaches were tested: linguistic cues and positional heuristics. Linguistic cues are lexico-syntactic patterns modeled as regular expressions implemented in a linguistic parser. Positional heuristics make use of the relative position of a sentence in the abstract to deduce its argumentative class.
Findings. Our experiments showed that positional heuristics attained a much higher degree of accuracy on Medline abstracts with an F-score of 64% whereas the linguistic cues only attained an F-score of 12%. This is mostly because sentences from different argumentative roles are not always announced by surface linguistic cues.
Research limitations/implications. A limitation to this study is that we were not able to test other methods to perform this task such as machine learning techniques which have been reported to perform better on Medline abstracts. Also, to compare the results of our study to earlier studies using Medline abstracts, the different argumentative roles present in Medline had to be mapped onto four major argumentative roles. This may have favorably biased the performance of the sentence classification by positional heuristics.
Originality/value. To the best of our knowledge, our study presents the first instance of evaluating linguistic cues and positional heuristics on the same corpus.

Keywords: semantic metadata annotation, discourse analysis, biomedical abstracts, argumentative role, sentence classification
Classification: research paper

1. Introduction

The importance of semantically annotated content for effective information retrieval (IR) in corporate as well as institutional organizations has long been recognised. Despite the success of keyword and *bag-of-word* oriented approach to IR, corporate organizations and industries have recognized the need for a more fine-grained access to electronic content. Semantic retrieval solutions are flourishing with companies such as iSeek, SmartLogic, Cogito who offer solutions sometimes embedding semantic components into popular search tools such as Google and MicroSoft desktop solutions. Also, the scientific community is organizing more conferences on automated methods for semantic metadata annotation for information retrieval (cf the ESAIR workshops[1] or the Semantic Analysis Technology seminar recently organised by ISKO-UK[2]).

There is a growing awareness that retrieval of content will be boosted if semantic metadata are explicitly added to the content. Such semantic metadata may be broad thematic categories such as sports, news, politics, travel, religion or named entities (persons, places, brand names), implemented as taxonomies. Alternatively, semantic metadata can be designed as structural roles of text components such as actors, agents, instruments, etc. The SKOS (Simple Knowledge Organization System) initiative is in line with this research concern for enhancing interoperability of semantic content across heterogeneous data repositories. However, in the case of SKOS, the annotation process is designed to be manually effected.

The availability of semantically annotated data is also useful in advancing research in IR in general and in particular in tasks such as automatic summarization, text categorization, entity named recognition. Recently, the New York Times has released an annotated corpus of articles with their corresponding indexing taxonomy. Because the manually annotated corpus constitutes a *gold standard*, it will indeed be easy to have a ground truth against which systems can be evaluated for accuracy in semantic annotation.

Ideally, metadata modeling and annotation of raw texts will be best performed manually, by information professionals and knowledge organizers. However, we acknowledge that human resources cannot face up to this task given the staggering volume of contents to be tagged and their exponential growth. Hence some research effort have been expended lately into automating these tasks. Information professionals still play an important place in the chain of events. They are often needed at the beginning of the process, to manually identify a set of seed categories or patterns on which automated methods can be trained or from which these methods can new patterns in order to tag new incoming data. The specific task we have been working on is the automatic identification of the discourse role of a sentence in a scientific abstract. The idea is that if the role of each sentence can be marked up, then it can be used during content retrieval to highlight particular types of information such as novelty, conclusions, methodologies, aims/goals of a scientific piece of work. The user in such a case will be able not only to specify the content of his/her search via the usual query words but will also be able to specify in what role or capacity s/he is seeking such content. In the next section (§2), we will review earlier work on discourse annotation with argumentative roles. In section §3, we will present our methodology for sentence annotation using argumentative roles. This methodology tests the effectiveness of two methods - linguistic cues and positional heuristics on annotating Medline abstracts. Section §4 is devoted to evaluation of both methods by comparison with the original Medline annotated abstracts. Finally, we conclude with a discussion of lessons learned from this research study and on future directions.

2. Studies on sentence classification by argumentative role

Previous studies (Swales 1990, Salager-Meyer, Tbahriti et al. 2005) have characterized scientific research as a problem solving activity. This is apparent in the way scientific texts are structured. In many domains, this problem—solution structure is materialized by a fixed

[1] http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=esair_2009

[2] http://www.iskouk.org/semantic_nov2008.htm

presentation with the presence of Introduction, Method, Result and Conclusion sections and in that order. Also, it has been established that abstracts share this problem solving property and being more condensed, are even more dense in information than full texts. Much research has been done on discourse analysis of scientific texts to identify thematic roles (Swales 1980, Salager-Meyer 1990, Orasan 2001, Teufel & Moens 2002, Ibekwe-SanJuan *et al*., 2008). Various applications have been targeted by these studies among which the most common is automatic summarization. Recently, some studies have viewed sentence tagging as a means to detect novelty in published research and as a means to perform information extraction (IE), especially in the biomedical field (Ruch *et al*. 2007, Mizuta et al. 2005, McNight & Srinivasan 2003, Lisacek et al. 2005). These studies have found that researchers usually focused on two types of sentences in scientific texts to decide rapidly if a paper is worth reading: the purpose and conclusion sections. Hence using features from sentence in certain argumentative roles can enable IE and IR systems respond to very specific domain queries. For instance, a clinical researcher may require the system to get '*all the documents using type X method to address type Y problem*'. This can also be formulated as "*get me records where method X is used to cure symptom Y*'.

The studies which worked on texts in the biomedical field used machine learning techniques to learn the roles from a manually annotated set of texts. Ruch *et al*. (2007) explored the use of machine learning techniques to learn sentence roles from a subset of Medline abstracts, then applied the learned features to tag unseen abstracts from the same corpus. Their objective was formulated as a key sentence selection task in which a key sentence is a conclusion sentence. Ultimately, the application targeted is the identification of novelty from scientific abstracts by selecting sentences bearing new or putative facts. In their study, sentences were classified into four major argumentative roles found in Medline abstracts: purpose/aim, method, result, conclusion. From a corpus of 12 000 Medline abstracts, 10 000 were set aside for training the classifier (90%) and 1200 (10%) abstracts were used for testing the accuracy of the classifier. Another 100 abstracts without any discourse marker were manually tagged by indexers and also used for testing. The two sets were of course stripped of the argumentative role markers for the classifiers. Four classifiers were built, each had to learn the features of sentences from one of these classes.

The classifiers learned the feature of each sentence type from the training set. The authors also tested the influence of adding positional heuristics in correctly classifying sentences, i.e; the hypothesis that the first ¼ of the sentences will belong to the "Purpose/aim" class, the second quarter to "Method", the third quarter to "Results" and the last quarter to "Conclusion".

While the machine learning approach implemented in Ruch *et al* (2007) does not require manually crafted rules as in linguistically based approaches (Teufel & Moens 1999), it does require the existence of large training sets: more than 90% of the available data was needed to train the classifiers. While this may be possible with Medline abstracts, the method will be difficult to generalize to other fields that do not have such large training datasets. Moreover, there is always the risk of over-fitting when the majority of the solution is used to train the classifier.

McKnight & Srinivasan (2003) evaluated the use of machine learning techniques in the classification of sentence type. A collection of 7253 structured abstracts and 204 unstructured abstracts representing Randomized Controlled Trials from Medline were fed into classifiers built using Support Vector Machine (SVM) and Linear Classifier models. The sentences were represented as a simple "*bag-of-words*". Each sentence was labeled as one of four types (Introduction, Method, Result, or Conclusion).

Mizuta *et al.* (2005) studied the argumentative structure of sentences in biology articles. Based on the extensive study carried out by Teufel & Moens (2002) on the discourse structure of scientific papers, Mizuta *et al.* (2005) determined seven major argumentative roles in abstracts. They particularly focused on modeling the argumentative role structure, including building nested roles (a role embedded in another). In this scheme, "Result" is seen as a sub-class of "Own" work. This allows for nested annotation, i.e. tagging a sentence with a single role, then tagging phrases

within that sentence with a sub-role. This is a more fine-grained classification scheme than in Ruch *et al.* (2007) and in McNight & Srinivasan (2003). The authors presented several classes and linguistic patterns found in these classes. This was done manually. In particular, no automatic detection of these patterns nor of sentence belonging to these classes was performed.

Lisacek *et al.* (2005) focused on the novelty detection aspect of sentence classification for IR. The problem addressed in this study is how to search a large collection of documents in order to extract the few documents that contain breakthrough results and new discoveries. The authors rightly argued that current IR techniques used by classical IR systems would not meet this need. They proposed a method based on computational linguistics for distinguishing novel facts (extraction) that may be important in future developments (prediction). They sought in the texts for linguistic cues that may indicate actual or potential breakthroughs (indicating that the authors of biological articles have identified a problem representing a break from conventional knowledge).
This is also a way of capturing "paradigm shifts". Some of the surface linguistic cues would be indicators of first appearance of evidence *(the data provide the first evidence of...),* emerging trend ("*Growing evidence indicates that..."), c*ontradiction of conventional knowledge (*In contrast with previous hypotheses...)* or *c*ontradiction, controversy, debate.
Although the goal is akin to novelty detection, novelty here is sought for within a single document and not by comparison of multiple documents.
A prototype of their method was implemented on abstracts on neurodegenerative diseases. Relevant records on this topic were extracted from PubMed. The abstracts were analysed for sentences with paradigm shifts. Out of 3300 abstracts, their system detected only one putative paradigm shift sentence in 175 of these. The authors then manually analysed 131 abstracts. The authors claimed that their method was able to propose protein lists that were overlooked in previous articles but later became known.
Although this approach is interesting in its use of linguistic cues, the particular way of implementing them is not tractable, particularly the use of a syntactic parser to identify dependency relations in sentences. Syntactic parsers often fail to parse a sentence or are able to parse it only partially. Also the idea of constructing a word list that can reflect paradigm shift is problematic because such a list is potentially infinite. However, the idea of imposing constraints on the relationship between the constituent words that might express paradigm shift is a good one and should reduce noise if it can be correctly implemented. Also, the objective of the study – to detect novelty via sentence tagging is one which we are particularly interested in.

3. Our approach to sentence classification

In previous studies (Ibekwe-SanJuan 2005, Ibekwe-SanJuan et al. 2008), we tested the usefulness of semantic annotation of abstracts for retrieval effectiveness. However, we could not evaluate the performance of the sentence annotation component as the text collection used had no gold standard (no set of manually annotated abstracts). In the current study, we chose to work on Medline abstracts as they come marked up with argumentative roles. A nice regularity in Medline abstracts is that each argumentative role name is followed by a colon (**:**), which uniquely identifies the sentences from this zone from the others. Figure 1 shows an example of such an abstract. We have highlighted the argumentative roles in bold.
Although in principle, scientific abstracts contain many more argumentative roles, we focused here on the four major roles "Objective, Method, Result, Conclusion" to ensure ease of comparison with previous studies, especially Ruch *et al*. (2007). We searched PubMed[3] for articles published during the year 2008 that contained any of the four argumentative roles in the abstract field. This search yielded 8134 abstracts. We wanted to first carry out a feasibility study and determine the specific problems that tagging Medline sentences might pose before automation. For that purpose, we

[3] The web interface for searching Medline database, available at http://www.ncbi.nlm.nih.gov/pubmed/.

narrowed down our survey to the first 200 abstracts out of the 8134 abstracts upon which we will focus in the rest of the study. Although our query required that returned abstracts had one of the four argumentative roles, 38 abstracts out of the 200 had no role markers at all. Thus we will be working on 162 abstracts. We explored two approaches to sentence role classification: surface linguistic cues (lexico-syntactic patterns) and positional heuristics. Each approach is described in more details hereafter.

PMID 19150703.
Granulocyte-macrophage colony stimulating factor administered as prophylaxis for reduction of sepsis in extremely preterm, small for gestational age neonates (the PROGRAMS trial): a single-blind, multicentre, randomised controlled trial.
BACKGROUND: Systemic sepsis is a major cause of death in preterm neonates. There are compelling theoretical reasons why treatment with haemopoietic colony-stimulating factors might reduce sepsis and improve outcomes, and as a consequence these agents have entered into use in neonatal medicine without adequate evidence. We assessed whether granulocyte-macrophage colony stimulating factor (GM-CSF) administered as prophylaxis to preterm neonates at high risk of neutropenia would reduce sepsis, mortality, and morbidity.
METHODS: We undertook a single-blind, multicentre, randomised controlled trial in 26 centres between June, 2000, and June, 2006. 280 neonates of below or equal to 31 weeks' gestation and below the 10th centile for birthweight were randomised within 72 h of birth to receive GM-CSF 10 microg/kg per day subcutaneously for 5 days or standard management. From recruitment to day 28 a detailed daily clinical record form was completed by the treating clinicians. Primary outcome was sepsis-free survival to 14 days from trial entry. Analysis was by intention to treat. This study is registered as an International Standard Randomised Controlled Trial, number ISRCTN42553489.
FINDINGS: Neutrophil counts after trial entry rose significantly more rapidly in infants treated with GM-CSF than in control infants during the first 11 days (difference between neutrophil count slopes 0.34 x 10(9)/L/day; 95% CI 0.12-0.56). There was no significant difference in sepsis-free survival for all infants (93 of 139 treated infants, 105 of 141 control infants; difference -8%, 95% CI -18 to 3). A meta-analysis of this trial and previous published prophylactic trials showed no survival benefit.
INTERPRETATION: Early postnatal prophylactic GM-CSF corrects neutropenia but does not reduce sepsis or improve survival and short-term outcomes in extremely preterm neonates.

Figure 1. Example of a Medline abstract.

3.1 Sentence annotation using linguistic cues

Previous studies (Salage-Meyer 1992, Orasan 2001, Swales 1990, Teufel and Moens 2002, Ibekwe-Sanjuan 2005 & 2008, Ruch et al. 2007) have established that scientific discourse follow some fixed argumentative patterns. These patterns are announced by certain linguistic constructs (lexico-syntactic cues) which can be modeled as regular expressions. From a study of abstracts from different domains (IR, biology, astronomy, medicine), we modeled the linguistic cues announcing the different argumentative roles of sentences in scientific abstracts as regular expressions implemented in finite state automata with the Unitex[4] linguistic toolbox. Unitex relies on a wide coverage electronic dictionary which it uses to tag the corpus before applying the finite state grammars. We modeled seven argumentative classes "*objective*, *related_work, newthing*, *result*, *future_work, hypothesis, conclusion*".
We did not model the "method" class although this class of sentences is always present in scientific abstracts. This is because we focused particularly on roles which may be interesting in the context of novelty detection. Also, method sentences are more difficult to identify with lexico-syntactic cues because often, this section of the abstract is not announced by explicit argumentative cues. Hence, it will be possible to determine by elimination that all the sentences which are not detected in the other categories belong to the method class. We built seven finite state grammars to recognize sentences that are likely to belong to each of these classes. Table 1 give samples of some lexico-

[4] http://igm.univ-mlv.fr/~unitex

syntactic patterns from each class. Figure 2 shows the finite state grammar for the “Objective” class.

Some patterns can introduce two different categories of sentences. The pattern "*In this paper we show that...*" could either introduce "*objective"* or "*results*" sentences. This is in agreement with earlier studies. Teufel and Moens (2002) then Ruch *et al.* (2007) had observed that it was not easy for their different methods to distinguish between “results” and “conclusion” sentences. It would seem that patterns announcing these two categories of sentences are fundamentally ambiguous. In theory, the same lexico-syntactic pattern can be used to introduce sentences from multiple argumentative categories. Some patterns require the combined use of part-of-speech (POS[5]) and lexical information. For instance, in figure 2, the patterns in sharp angles <> contain grammatical functions (DET= determiner, V=verbs, N=nouns, <PRO>=pronoun, ADV=adverbs, <NB>=number). The grey boxes represent phrase structures that represent syntactic constructs such as noun phrases (NP) or verb phrases (VP). These phrase structures are themselves finite state grammars which can be embedded in one another. In the example above, we have two such embedded grammars which describe noun phrase structures. For instance, «SN-cc-max-enum1» is a local grammar that identifies complex NPs (NPs with embedded simpler NPs). This grammar in turn embeds another simpler NP grammar. The expressive power of such local grammars can be quite high as more simpler grammars can be embedded into more complex ones to achieve a considerable level of complexity.

Argumentative role	**Lexico-syntactic patterns**
OBJECTIVE	In this_{current|present} {article|paper|study|research|work}... We_{examine|investigate|describe|present|outline|introduce|consider}.... DET_{motivation:|aim|goal|objective|problem}...
NEWTHING	we propose a novel approach | This analysis reveals | Emerging evidence suggests that | Interestingly, our results indicate that | new evidence |novel
RELATED_WORK	{in contrast to|unlike|in common|n comparison to|in contrast to common belief|despite} our {work|study|hypothesis|observation|approach..} {<contradict.V>|<disagree.V>...}
RESULT	In this paper we show that | Our research suggests that | Results confirm that | It is shown here for the first time that | This approach may represent a step forward toward
HYPOTHESIS	DET_NP_{may|might}_{ADV|V_NP} | Our findings support the view that DET_NP_can_{V|NP.}..
FUTURE_WORK	{Further|Future|more}_{work|investigation|observation}_<verb>
CONCLUSION	This paper concludes... |Conclusion**:** |Finally, ...|As a conclusion

Table 1. Examples of patterns from the seven argumentative role classes.

[5] Part-Of-Speech

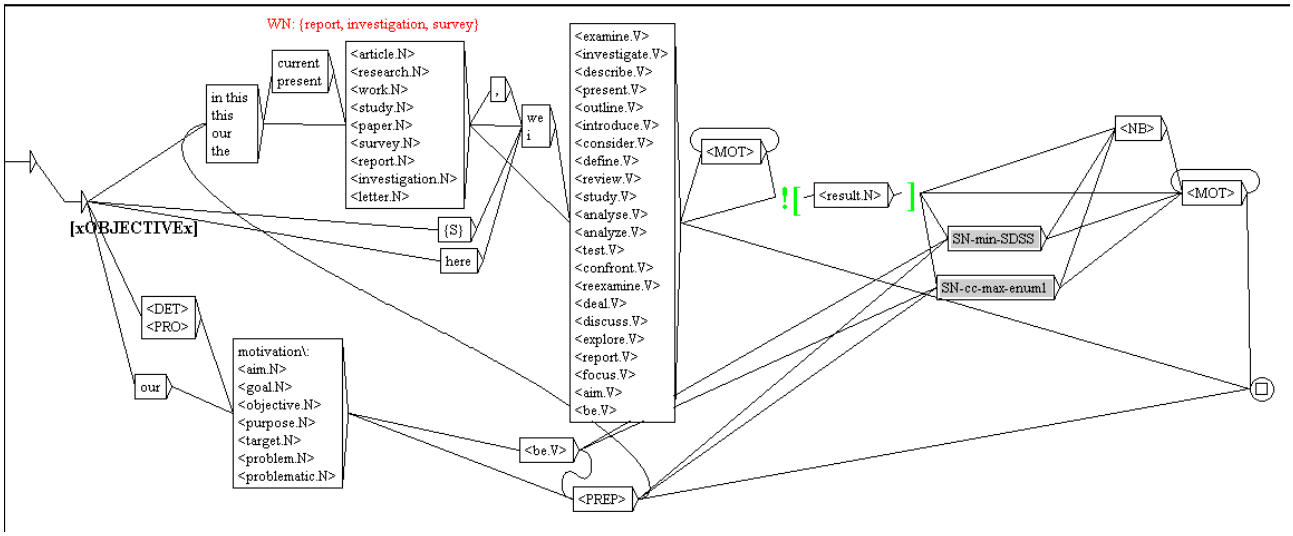

Figure 2. The finite state grammar for the "Objective" class.

The linguistic pattens we have modeled are domain independent but language dependent.
Figure 3 below shows the result of sentence annotation using our linguistic cues on a Medline abstract. Note that abstracts were stripped of their original Medline tags before subjecting them to sentence annotation by linguistic and positional cues.

{S}19130928.
{S} Notch signaling is involved in cell fate determination along with the development of the immune system.
{S} However, very little is known about the role for Notch signaling in mast cells.
{S} **[xOBJECTIVEx]** *We investigated* the role of Notch signaling in mast cell functions.{S} After mouse bone marrow-derived mast cells (BMMCs) or peritoneal mast cells (PMCs) were cocultured with mouse Notch ligand-expressing chinese hamster ovary cells for 5 days, we examined the mast cell surface expressions of MHC-II molecules and OX40 ligand (OX40L), Fc epsilon RI-mediated cytokine production, and the effects of the mast cells on proliferation and differentiation of naive CD4(+) T cells in vitro.
{S} **[xRESULTx]** *We showed that* BMMCs and PMCs constitutively expressed Notch1 and Notch2 proteins on the cell surface.
{S} **[xRESULTx]** *We also found* that Delta-like 1 (Dll1)/Notch signaling induced the expression of MHC-II and upregulated the expression level of OX40L on the surface of the mast cells.
{S} Dll1/Notch signaling augmented Fc epsilon RI-mediated IL-4, IL-6, IL-13, and TNF production by BMMCs.
{S} Dll1-stimulated MHC-II(+)OX40L(high) BMMCs promoted proliferation of naive CD4(+) T cells and their differentiation into T(H)2 cells producing IL-4, IL-5, IL-10, and IL-13.
{S} Dll1/Notch signaling confers the functions as an antigen-presenting cell on mast cells, which preferentially induce the differentiation of T(H)2.

Figure 3. An Medline abstract annotated using linguistic cues.

The tag name is inserted to the left of the sentence. An 'x' is added to tag names in order to differentiate them from similar words occurring in the abstracts. We have highlighted with italics the exact lexico-syntactic cues that triggered the sentence tag. As we can see, the percentage of sentences tagged in an abstract by linguistic cues can be quite low, in this case 37% only. This occurs because some sentences are not introduced by explicit cues that will trigger their annotation with any of the modeled argumentative role. In this example, the first sentence was not introduced by explicit patterns that would have triggered the insertion of the "objective" tag. The 3rd sentence is correctly tagged objective because it displays one of the cues that triggers this tag "*We investigated...*". Also, the "conclusions" section has no explicit marker and thus could not be

annotated. Obviously, this abstract does have sentences from the four argumentative divisions roles but does not make use of explicit cues to introduce them. This is normal in writing. Authors do not always explicitly announce the argumentative role of the section they are embarking on to the reader. This points to the insufficiency of linguistic cues alone and calls for the use of additional features.

3.2 Sentence annotation using positional heuristics

The hypothesis, empirically confirmed in previous studies, is that the argumentative divisions in scientific discourse, and especially in abstracts follow one another in an orderly sequence. For instance, all abstracts would normally begin with the *Objectives* (diversely called "Introduction, Goal, Aims, Background, Context") which may be split into two zones – an Introduction then an objective. This section would be logically followed by a *Method* paragraph (alternatively called "*design and method*, *design*, etc) stating the methods employed in the study. Next, we would have either *Findings* and/or *Results*, then the *Conclusion*. Empirically, we have observed that well-written scientific abstracts do conform to this orderly pattern. Hence, the natural assumption that the argumentative role of sentences can be detected by using positional cues. Ruch *et al*. (2007) applied machine learning techniques to train four classifiers for sentence classification into the four major roles. They then added positional heuristic to correct some of the tags of the classifier by assigning a score to each sentence based on its relation position in the abstract.

The heuristics we implemented for classifying sentences into the four argumentative roles (objective, method, result, conclusion), according to their position is as follows[6]:

Starting from the beginning:

- up to ¼ of the sentences = Objective
- after ¼ and before ½ of sentences = Method
- from ½ - ¾ of sentences = Results
- from ¾ to the last sentence = Conclusion.

With this simple rule, it was possible to annotate all the sentences in the 162 abstracts. Thus, by definition, positional heuristics have a 100% coverage of the corpus.

4. Evaluation against Medline's original annotation

To evaluate the effectiveness of linguistic and positional cues for correctly classifying sentences into one of the four classes, we compared the annotation obtained by both methods against the original annotation in the Medline abstract. However, in Medline, each sentence is not annotated, rather the abstract is divided into argumentative divisions and each one is preceded by the name of the argumentative role name followed by the punctuation colon (**:**) (see figure 1). In that case, we consider that all the sentences between an argumentative role name and the next one to belong to the one preceding it. Thus, to enable comparison of the three annotations, we first needed to propagate the argumentative role to every Medline sentence. Once each sentence is annotated, then the original class assigned by Medline can be compared to the ones assigned by linguistic cues or positional heuristics. Before we discuss the results, a few words need to be said about the variability of Medline's original argumentative roles.

4.1. Variability of Medline argumentative roles

[6] We thank Eric SanJuan, professor at the University of Avignon, for writing the programs used in this study, for applying the positional heuristics and for comparing the classifications obtained by both methods – linguistic and positional, against Medline's original annotation.

Upon perusal of the original abstracts, we observed a great number of variations in the names of the argumentative role and their types. In the first 200 abstracts, we found not less than 35 different role names (not counting the plural forms of the same role). Table 2 shows these role names and in the 2nd column the argumentative role to which we mapped them in order to bring them down to the four major argumentative roles.

Argumentative role names found in 200 Medline abstracts	**Mapped to**
OBJECTIVE, BACKGROUND, INTRODUCTION, AIM, AIMS, AIMS AND BACKGROUND, BACKGROUND/ AIMS, CONTEXT, PURPOSE	OBJECTIVE
METHOD, METHODS, METHODOLOGY, DESIGN, DESIGN, SETTING AND PARTICIPANTS, STUDY DESIGN, STUDY DESIGN AND METHODS, RESEARCH DESIGN, PATIENTS AND METHODS, MATERIAL AND METHOD, MATERIALS AND METHODS, DATA SOURCE, DATA SUMMARY, SETTING, PARTICIPANTS, SUBJECTS, PARTICIPANTS, INTERVENTION	METHOD
RESULTS, FINDINGS, METHODS AND RESULTS, METHODOLOGY/PRINCIPAL FINDINGS, OUTCOME MEASURES, MAIN OUTCOME MEASURE, MAIN OUTCOME MEASURES	RESULT
LIMITATIONS, INTERPRETATION, CONCLUSION, CONCLUSIONS, DISCUSSION, CONCLUSIONS/SIGNIFICANCE, SIGNIFICANCE AND IMPACT OF THE STUDY	CONCLUSION

Table 2. The original argumentative role names found in Medline abstracts.

As we can see from this table, the same basic role can be named by up to five synonyms. There are cases of strict synonymy such as "*method*" being also named "*methodology, design, study design and methods*" or "*Objective*" being named "*Purpose:, Aims:, Goals:, Background:, Introduction:*". "Method" is also referred to as "*Study design and method:, Research design:, Design:, Methods:, Design, Setting and Participants:*". "*Results*" is also alternatively called "*Findings:, Conclusions:, Interpretation*".

There are other cases of near synonymy or roles reflecting other facets of the study that may be considered necessary in the medical domain such as "*data source, subject, participants, outcome measures*".

For the purpose of comparison with previous works by Ruch *et al.,* (2007), we will not split hairs and will consider these different facets as reflecting information on one of the four major argumentative roles: *objective, method, result, conclusion*.

Although, we specified in our Medline search that we wanted only those abstracts that contained argumentative roles, 38 abstracts out of the first 200 were not annotated, thus did not carry any explicit argumentative role marker. The evaluation will be carried out against the 162 abstracts with explicit argumentative roles.

4.2. Performance of linguistic and positional cues for sentence classification

Recall that we described seven different argumentative roles using linguistic cues (cf. §3.1) These also had to be mapped onto the same four argumentative roles considered (Ruch *et al*., 2007) as follows:

- our *objective* class is mapped onto *"objective"*
- our "*results, related_work, newthing"* classes are mapped onto *" results"*
- our "*hypothesis, future work, conclusion"* classes are mapped onto *"conclusion"*

As we did not describe the "method" class, there was no need to map this class.

Using the traditional information retrieval metrics of precision and recall, we measured the effectiveness of each approach. Precision here refers to the proportion of correctly tagged sentences

out of all tagged sentences for a given role. Recall measures the proportion of tagged sentences compared to the total number of sentences for a given role. Precision and recall ratios were calculated for each of the four argumentative roles (Tables 3 & 4). We also calculated the F-score using the harmonic mean formula in which the same weighted is assigned to precision and recall scores. The F-score is calculated as follows:

$$F = \frac{2 . \ precision \, . \, recall}{(\ precision + recall \ )}$$

			Linguistic cues			
Medline tags	**nb**	**Agree**	**total**	**Prec**	**Recall**	**F-score**
Obj	417	34	53	0,64	0,08	0,14
Method	439	-	-	-	-	-
Results	713	8	218	0,04	0,01	0,02
Conclusion	301	92	157	0,59	0,31	0,40
Total	*1870*	*134*	*428*	***0,31***	***0,07***	***0,12***

Table 3. Precision, Recall and F-score measures for sentences tagged by linguistic cues.

			Positional cues			
Medline tags	**nb**	**Agree**	**total**	**Prec.**	**Recall**	**F-score**
Obj	417	333	409	0,81	0,80	0,81
Method	439	244	408	0,60	0,56	0,58
Results	713	336	486	0,69	0,47	0,56
Conclusion	301	288	567	0,51	0,96	0,66
Total	***1870***	***1201***	***1870***	***0,64***	***0,64***	***0,64***

Table 4. Precision, Recall and F-score measures for sentences tagged by positional cues.

The results leave us in no doubt as to the superiority of positional heuristics over linguistic cues on Medline abstracts. Overall the former had an F-score of 64% while linguistic cues could only attain 12%. The poor performance of linguistic cues can be explained by several factors:

- linguistic cues are based on surface patterns that introduce sentences. These patterns cannot be exhaustively collected, they vary and more importantly, some sentences (about 50%) do not have them, thus they have very low recall, i.e., they enable to tag on the average 2 sentences per abstract, because of the lack of explicit surface markers that introduce each argumentative role. It is therefore nearly impossible to tag the argumentative role of all sentences in scientific texts by relying on linguistic cues alone. They can be used as complementary evidence to post-process sentence classification using another method.
- in 18 cases, the linguistic cues made the right choice whereas the positional heuristics made the wrong choice. However, in 164 cases, positional correctly tagged a sentence whereas the linguistic cues made the wrong choice. Based on that, we conclude that linguistic cues cannot boost positional cues, at least not on Medline abstracts.

Positional heuristics were particularly successful in identifying sentences from the beginning and at the end of abstracts. Objective sentences were tagged with a precision and recall score of 81% and 80% respectively while the recall score of for conclusion sentences was the highest at 96%. Linguistic cues also attain a honourable performance for these two classes with 64% precision

for objective sentences and 59% for conclusion sentences. This is also in agreement with earlier studies (Teufel & Moens 2002; Ruch *et al*. 2007). However, we were surprised by the modest precision scores of classifying conclusion sentences (51%) by positional cues. Ruch *et al*. (2007) had reported a precision of 97% using machine learning techniques alone. This score was slightly increased to 98% when positional heuristics were added to correct the classifier's choice. Conforming to observations from earlier studies, positional heuristics are moderately successful for classifying method and results sentences with F-scores of 56% and 66% respectively. We had already observed that method sentences were difficult to model in terms of recurring patterns because they hardly had any. Results sentences were also difficult to separate from conclusion sentences.

Another factor which probably enhanced the performance of positional heuristics is the few number of target classes, here four. This is especially true for non symbolic methods (probabilistic) because the fewer classes to be predicted, the lower the possibility of making a wrong guess. This explains why the performance of classifiers reported in Ruch *et al.* (2007) are higher than the performances obtained by the positional heuristics. They reported a precision of 93% for purpose (objective) sentences, 93% for method sentences, 71% for result sentences and 97% for conclusion sentences

5. Discussion

Ruch et al. (2007)'s study had shown that result sentences were often misclassified as conclusion (in 15,56% of the instances), and that adding positional heuristics does not improve the classification of sentences in this category nor that of method sentences. Indeed, it is difficult even from a linguistic viewpoint to distinguish result and conclusion sentences. The two categories tend be contiguous (results sentences flowing into conclusions). Positional heuristics on the other hand improved the classification of conclusion and purpose sentences. This is to be expected as earlier studies have shown that these two categories are the most likely to be distinguished from the others. Indeed, earlier studies had established that sentences at the beginning and at the end of abstracts are likely to be the most informative, hence adding positional heuristics is likely to enhance their recognition. Also, one may wonder if using purely positional information would not have been sufficient to identify these two categories of sentences. Previous studies had noted that positional heuristics were most effective for the unstructured abstracts, i.e. abstracts that had no explicit argumentative role marker.

Basing on the results obtained from this study, we have to test these methods on more roles as a possible bias may have been introduced by collapsing the different argumentative roles in Medline's and the seven linguistic categories into the same four categories. We also plan to use paragraph separation as a transition indicator between rhetoric divisions if the original abstracts contain these marks.

We also intend to test machine learning methods to perform sentence classification. The next step will then be to run the best method on a much larger corpus and use the annotated sentences to boost information retrieval.

References

1. Ibekwe-SanJuan F., Chen C., Pinho R., (2008a), Identifying Strategic Information from Scientific Articles through Sentence Classification, *6th International Conference on Language Resources and Evaluation Conference (LREC-08)*, Marrakesh, Morocco, 26 May -1st June, 2008, 5p.
2. Ibekwe-SanJuan F., Fernandez S., SanJuan E., Charton E., (2008b) Annotation of Scientific Summaries for Information Retrieval, Workshop on "Exploiting Semantic Annotations in Information Retrieval", in *30th European Conference on Information Retrieval (ECIR-08)*, Glasgow, Scotland, 30th March 2008, pp.70-83.
3. McKnight L., Srinivasan P. (2003), Categorization of Sentence Types in Medical Abstracts,

Proceedings of AMIA Annual Symposium 2003, pp. 440-444.

4. Lisacek F., Chichester C., Kaplan A., Sandor A., (2005) Discovering Paradigm Shift Patterns in Biomedical Abstracts: Application to Neurodegenerative Diseases, *Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM)*, Morgan Kaufmann, pp. 212–217.
5. Mizuta Y., Korhonen A., Mullen T., Collier N., (2005). Zone Analysis in Biology Articles as a Basis for Information Extraction. *International Journal of Medical Informatics*, vol. 75, N° 6, pp.468-487.
6. Orasan C. (2001), Patterns in scientific abstracts, in *Proceedings of the Corpus Linguistics Conference*, Lancaster University, Lancaster, UK, 2001, pp. 433-443.
7. Paice C.D., Jones P.A. (1993), The identification of highly important concepts in highly structured technical papers, in *Proceedings of the ACM SIGIR'93*, pp. 123-135.
8. Ruch P., Boyer C., Chichester C., Tbahriti et al. (2007), Using argumentation to extract key sentences from biomedical abstracts, *International Journal of Medical Informatics*, vol. 76, pp. 195-200.
9. Ruch P., Tbahriti I., Gobeill J., Aronson A. R. (2006), Argumentative Feedback: A Linguistically-motivated Term Expansion for Information Retrieval, *Proceedings of the COLING/ACL 2006*, Sydney, July 2006, pp. 675–682.
10. Salager-Meyer F. (1990), Discoursal movements in medical English abstracts and their linguistic exponents: a genre analysis study, INTERFACE: *Journal of Applied Linguistics* 4(2), 1990, pp. 107 – 124.
11. Tbahriti I., Chichester C., Lisacek F., Ruch P., (2006) Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the MEDLINE digital library, *International Journal of Medical Informatics,* 2006, vol. 75, pp. 488—495.
12. Teufel S., Moens M. (2002), Summarizing scientific articles: Experiments with relevance and rhetorical status, *Computational Linguistics*, 2002, vol. 28, n° 4, pp. 409-445.
13. Swales J. (1990), *Genre Analysis: English in academic and research settings*, Cambridge University Press, 1990.