A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task

Maria Liakata
University of Warwick/
EMBL-EBI, UK
M.Liakata@warwick.ac.uk

Simon Dobnik
University of Gothenburg, Sweden
simon.dobnik@gu.se

Shyamasree Saha
EMBL-EBI, UK
saha@ebi.ac.uk

Colin Batchelor
Royal Society of Chemistry, UK
batchelorc@rsc.org

Dietrich Rebholz-Schuhmann
University of Zurich, Switzerland/
EMBL-EBI, UK
rebholz@ebi.ac.uk

Abstract

We present a method which exploits automatically generated scientific discourse annotations to create a content model for the summarisation of scientific articles. Full papers are first automatically annotated using the CoreSC scheme, which captures 11 content-based concepts such as Hypothesis, Result, Conclusion etc at the sentence level. A content model which follows the sequence of CoreSC categories observed in abstracts is used to provide the skeleton of the summary, making a distinction between dependent and independent categories. Summary creation is also guided by the distribution of CoreSC categories found in the full articles, in order to adequately represent the article content. Finally, we demonstrate the usefulness of the summaries by evaluating them in a complex question answering task. Results are very encouraging as summaries of papers from automatically obtained CoreSCs enable experts to answer 66% of complex content-related questions designed on the basis of paper abstracts. The questions were answered with a precision of 75%, where the upper bound for human summaries (abstracts) was 95%.

1 Introduction

The publication boom of the last few years, especially in the life sciences, has highlighted the need to facilitate automatic access to the information content of articles. Researchers, curators, reviewers all need to process a continuously expanding flow of articles whether the purpose is to follow the state of the art, curate large knowledge bases or have a good working knowledge of their own and related disciplines to assess progress in research. While a lot of effort has concentrated on information extraction of particular types of entities and relations from the scientific literature (Cohen and Hersh, 2005; Kim et al., 2009; Ananiadou et al., 2010; Kim et al., 2011), with a view to support scientists in obtaining relevant information from scientific articles and abstracts, less work has focussed on automatically combining such information in the form of a cohesive summary which preserves the context. Researchers rely to a great extent on author-written abstracts, but the latter suffer from a number of problems; they are less structured, vary significantly in terms of length, are often not self-contained and have been written independently of the main document (Teufel, 2010, p.83).

Teufel (2001; 2010), (Teufel and Moens, 2002) identify argumentative zones within scientific articles and use them to create use-targeted extractive summaries. Argumentative zones are annotations which designate the type of knowledge claim and rhetorical status for a sentence and how these relate to the communicative function of the entire paper. A selection of various combinations of argumentative zones are chosen for the use-targeted extractive summaries (rhetorical extracts), each of which fulfills a different role. For instance, purpose-oriented extracts less than 10 sentences long are generated containing a predetermined number of AIM, SOLUTION and BACKGROUND zones. As the emphasis of this approach was the identification of the argumentative zones, less attention was given to the sentence selection criteria for the extractive summaries.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 747–757,
Seattle, Washington, USA, 18-21 October 2013.

The sentences chosen for the rhetorical extracts were either all sentences of a particular category (in the case of rare categories) (Teufel and Moens, 2002), selected according to a classifier trained on a relevance gold standard (Teufel and Moens, 2002), manually or randomly selected (Teufel, 2010, p.60).

More recently Contractor et al. (2012) have used automatically annotated argumentative zones (Guo et al., 2011) to guide the creation of extractive summaries of scientific articles. Here argumentative zones are used as features for the summariser, along with verbs, tf-idf values and sentence location. They use a standard approach to summarisation, with a binary classification recognising candidate sentences which are then fed into a clustering mechanism. Extracts can be created to summarise the entire paper or focus on specific user-specified aspects. The number of sentences to include in the summary is prespecified (either directly or using a compression ratio).

Our approach also makes use of the scientific discourse for summarisation purposes. We use the scientific discourse to create a content model for extractive summarisation, with a focus on representing the content of the full paper, while keeping the cohesion of the narrative. We first automatically annotate the articles with a scheme which captures fine-grained aspects of the content and conceptual structure of the papers, namely the Core Scientific Concepts (CoreSC) scheme (Liakata et al., 2010; Liakata et al., 2012). The CoreSC scheme is "uniquely suited to recovering common types of scientific arguments about hypotheses, explanations, and evidence" (White et al., 2011), which are not readily identifiable by other annotation schemes. Also, when compared to argumentative zoning and more specifically its extension for chemistry papers, AZ-II (Teufel et al., 2009), it was shown to provide a greater level of detail in terms of categories denoting objectives, methods and outcomes whereas AZ-II focusses on the attribution of knowledge claims and the relation with previous work (Liakata et al., 2010).

We then use the distribution of CoreSC categories observed in abstracts to create a content model which provides a skeleton for extractive summaries. The reasoning behind this is to try to preserve cohesion within the summaries and we hypothesise that the sequence of CoreSC categories is a good proxy for cohesion (see section 3.1). In creating the summary, instantiating the content model, we identify independent categories and dependent categories, and we argue that in order to preserve the cohesion of the text the independent categories should be determined first (see section 3.2). We also preserve in the summary the distribution of CoreSC categories found in the corresponding full paper.

Finally, we evaluate the extractive summaries in a complex real world question-answering task, in which we assess the usefulness of the summaries as well as to what extent the generated CoreSC summaries represent the content of the original article. Experts are presented with different types of summaries and are asked to answer article-specific questions on the basis of the summaries (see section 4.1). Our results show that automatically generated CoreSC summaries can answer 66% of complex questions with 75% precision, outperforming a baseline of microsoft autosummarise summaries (See section 4.2).

We have also peformed an intrinsic evaluation of the summaries using ROUGE and automatic measures for summary informativeness, such as the Jensen-Shannon divergence, yielding positive results (See section 4.2). However, as such measures have not yet reached maturity and are harder to interpret, we consider the user-based evaluation to be a more reliable measure of summary quality.

Code for generating the summaries can be obtained by contacting the first author and/or visiting http://www.sapientaproject.com/software.

2 Related work

The Core Scientific Concepts (CoreSC) Scheme: The CoreSC scheme consists of three layers; the first layer corresponds to eleven concepts (Background (BAC), Hypothesis (HYP), Motivation (MOT), Goal (GOA), Object (OBJ), Method (MET), Model (MOD), Experiment (EXP), Observation (OBS), Result (RES) and Conclusion (CON)); the second layer corresponds to properties of the concepts (e.g. New/Old) and the third layer provides identifiers which link instances of the same category. Liakata et al. (2010) created a corpus of 265 full scientific articles from chemistry and biochemistry annotated

with this scheme and trained classifiers using SVMs and CRFs in (Liakata et al., 2012), with an accuracy of >51% across the 11 concepts. Their data and CoreSC classification system are available online and can provide a good benchmark for comparison. Louis & Nenkova (2012) have successfully used the CoreSC corpus for evaluating syntax-based coherence models, which indicates the strong connection between coherence and discourse structure.

Summarisation for scientific articles: A lot of the work on summarising scientific articles has focussed on citation-based summaries. Qazvinian & Radev (2008) use sentences from papers citing the article to be summarised. Sentences are clustered together creating a topic, with the combination of clusters forming a citation summary network. Qazvinian & Radev (2010), (Qazvinian et al., 2010) also make use of citation sentences in other scientific papers to summarize the contributions of a paper. The drawback of citation summaries is that a paper must be already cited, so this type of summary will not be useful to a paper reviewer. Also, citations of articles will have been influenced by other citations rather than the paper itself.

Document models for summarisation: Our content model has some similarities with content modelling using global sentence ordering (Barzilay and Lee, 2004; Chen et al., 2009). In (Barzilay and Lee, 2004) unsupervised methods are used to create HMM topic sequence models for newswire text articles. Topics are assigned to texts according to the content model and extracts of fixed length are created by selecting the topics most likely to occur in summaries. While we use supervised methods to annotate papers with a fixed set of topics (CoreSCs) in scientific papers, our summary content model for extracts shares similar principles such as global ordering of sentences and non-recurrence. However, their evaluation involved newspaper articles and extracts which are a lot shorter (15 and 6 sentences, respectively).

It is not clear whether unsupervised topic modelling such as (Chen et al., 2009) can be applied to scientific articles (over 100 sentences long), which by nature include repetition of topics. It would be interesting to make comparisons with summaries using content models learnt from our data automatically, following a similar approach to (Sauper et al., 2010) which learns a content model jointly with a particular supervised task in web-based documents.

3 Extractive Summarisation using CoreSCs

In this section we describe how we use CoreSC discourse categories annotated at the sentence level to create extractive summaries of full papers, which we subsequently evaluate in a question answering task in section 4.

To generate summaries we follow classic text extraction techniques while making use of a document content model based on CoreSCs. Our aim is for the content model to reflect both the distribution of CoreSCs in the paper as well as the discourse model of human summaries, as the latter is indicated by the generic ordering of CoreSC categories in abstracts encountered in a corpus of 265 annotated full papers (Liakata and Soldatova, 2009; Liakata et al., 2012). While we do not consider abstracts to be adequate summaries, we at least consider them to be coherent summaries, which is why the content model reflects the distribution of CoreSCs in the abstracts.

To create our summaries, we employed automatically generated CoreSC annotations, which are the output of the classifiers described in (Liakata et al., 2012). These classifiers assign CoreSC categories to sentences on the basis of features local to a sentence, such as significant n-grams, verbs and word triples, as well as global features such as the position of the sentence within the document and within a paragraph and section headers. The following subsections give details about the creation of extractive summaries from CoreSC categories.

3.1 A content model for CoreSC extractive summaries

Building an extractive summary using a computational model of document structure is an idea shared by many previous approaches, whether the model is hand-crafted, based on rhetorical elements (McKeown, 1985; Teufel and Moens, 2002) or rhetorical relations (Marcu, 1998b; Marcu, 1998a) or whether it is a content model, learnt automatically from text as in (Barzilay and Lee, 2004), focussing on the local content or a combination of the local content and global structure (Sauper et al., 2010).

Our document content model is primarily based on the global discourse of the article as provided by the type and number of CoreSC categories. However, unlike (Teufel and Moens, 2002), who take a fixed number of AZ categories of specific type to create rhetorical extracts, the number of categories used from each CoreSC category depends on their distribution in the original article. Any and all types of CoreSC category could potentially appear in a summary, as our summaries are meant to be representative of the entire content of the paper. Also, the ordering of the categories in the summary is learnt to reflect the ordering of categories observed in abstracts of papers from the same domain.

Our model also caters for local discourse dependencies. For example, the selection of a particular 'Method' sentence for inclusion in the summary should influence the choice of 'Experiment' sentences, which refers to particular experimental procedures performed. This is not an issue of concern to (Teufel and Moens, 2002), but relates to the notion of NUCLEUS and SATELLITE clauses, which form the foundation of Rhetorical Structure Theory (Mann and Thompson, 1998), and guides the summarisation paradigm of (Marcu, 1998a; Marcu, 1998b). However, the difference here is that we define a-priori certain categories to be independent (have the property of playing the role of nucleus in the discourse) and specify their relation with particular types of dependent categories. Thus, nuclearity becomes a property of the CoreSC category, which is indirectly inherited by the sentence.

Therefore, when creating the CoreSC content model for summaries we addressed the following issues: (i) summary length; (ii) number of sentences from each CoreSC, (iii) the ordering in which sentences from each CoreSC category should appear and (iv) the extraction of sentences according to independent and dependent categories.

- **Summary length:** While the literature (Teufel, 2010, p.45) suggests that 20–30% of the original document is required for an adequately informative summary, (Teufel, 2010, p.55) assumes this is too long for scientific papers. For this reason and to allow better comparison between papers of varying lengths, we fixed our summary length to 20 sentences. This is reasonable considering we have 11 CoreSCs, any and all of which can appear in both abstracts and full papers.
- **Number of sentences from each category:** To reflect the content of the paper, the distribution of the CoreSC categories in the extract follows the distribution of CoreSCs in the full paper.

  For each CoreSC we determine the number of sentences to be selected ($n(selected(C))$) by multiplying the ratio of that category in the paper by 20. A difficulty arises if the ratio of a particular concept in the paper is very low (≤ 0.05) in which case we prefer to include one sentence. If a particular concept is not at all present in the paper, the number of selected sentences for that category will be 0.
- **Ordering of CoreSC categories in the summary:** According to a study of empirical summaries (Liddy, 1991), sentences of a particular textual type appear in a particular order. Since paper abstracts were the closest approximation of human summaries available to us, CoreSC category transitions found in abstracts have been adopted in our content model for extracts. The transitions were derived semi-empirically. First, we extracted initial, medium and final bi-grams of categories from paper abstracts together with transition probabilities.

  Using this information we manually constructed transitions of the CoreSC categories that best fit the observed frequencies and our own intuitions. This gave us the following sequence: MOT $>$ (HYP) $>$ OBJ $>$ GOA $>$ BAC $>$ MOD $>$ MET $>$ EXP $>$ OBS $>$ (HYP) $>$ RES $>$ CON. HYP appears twice in the sequence as annotators had distinguished two types of hypotheses, global hypotheses (stated together with other objectives) and hypotheses about particular observations. The model provides an amalgamated representation of CoreSC concepts in abstracts. Interestingly, our semi-empirically derived model closely follows the content model for abstracts described in (Liddy, 1991). It would be interesting to see how this compares to a Markov model of CoreSC categories learnt from the annotated abstracts.

3.2 Sentence extraction based on independent and dependent categories

Sentence extraction involves selecting the most relevant sentences to include in a summary. Typically, this entails ranking the sentences according to some measure of salience and selecting the top n-best sentences. For example, a sentence will be represented by a number of features associated with it, such as whether it contains certain high frequency words or cue phrases, its location in the document, location in a paragraph (Brandow et al., 1995; Kupiec et al., 1995). Other methods include clustering based on sentence similarity and choosing the centroids (Erkan and Radev, 2004) or choosing the best connected sentences (Mihalcea and Tarau, 2004).

When sentences are classified according to CoreSC categories features such as the ones described above for text extraction are taken into account. Liakata et al. (2012) report that the most salient features for classifying CoreSC categories are overall n-grams, verbs and direct objects whereas other features such as the location of the sentence, the neighbouring section headings and whether a sentence contains citations play an important role for some of the categories. Thus, classification into CoreSC categories already provides a selection bias for sentence extraction.

As explained in section 3.1, the number of CoreSC categories in the summaries is determined according to their distribution in the paper and the order of the categories is specified in the content model. Salience for sentence extraction in this case is determined by the need to select the most representative sentences for a category. There isn't much point, for example, in identifying that we need to include a Method sentence (MET) and that this should be followed by an Experiment sentence (EXP), if we are not sure that those are indeed the categories of the sentences we are about to select.

We therefore rank sentences according to the classifier confidence score (probability) with which they were assigned a CoreSC category in (Liakata et al., 2012). The intuition behind this is that sentences with high classifier confidence will be less noisy, high precision cases and more representative of a particular category. Indeed, (Liakata et al., 2012) report statistical significance for the correlation between high classifier confidence and agreement between manual and automatic classification

However, as mentioned in section 3.1, there is inter-dependence between sentences in the text, which is in turn inherited by the categories assigned to them. For example, the highest ranking MET sentence will be related to an Experiment (EXP) or Background (BAC) sentence, which may not be the ones with the highest confidence score in their category.

In order to preserve discourse cohesion it is important to select related sentences from different categories. We resolve this by distinguising the CoreSCs into independent categories, which by definition are expected to show nucleus behaviour, and dependent categories. We also specify the relation between independent and dependent categories. The independent categories include the categories with the lowest percentage of sentences in scientific articles as reported in (Liakata et al., 2012), namely: Motivation (MOT) (1%), Goal (GOA)(1%), Hypothesis (HYP)(2%), Object (OBJ)(3%), Model (MOD)(9%), Conclusion (CON)(9%) and Method (MET)(11%). Categories whose sentence selection semantically depends on the former are Experiment (EXP)(10%), Background (BAC)(19%), Result (RES)(21%) and Observation (OBS)(14%). The independent categories also have higher precision than recall, in contrast to the dependent categories. While MET and EXP are almost equally represented in the CoreSC corpus, EXP by definition provides the detailed steps of an experimental method and thus it is semantically dependent on some MET category. More specifically, the dependencies are considered to be as follows: EXP, BAC depend on MET, RES depends on CON and OBS depends on RES (OBS is double-dependent).

Sentence extraction is driven by first identifying the independent categories based on classifier confidence scores and then choosing the corresponding dependent categories on the basis of both relatedness to the independent categories and classifier confidence. We use sentence proximity (defined below) as a measure for relatedness and combine it with classifier confidence during sentence extraction.

The mechanism to select sentences for inclusion in the summary, which considers category dependencies, proceeds as follows:

- For an independent category CatI, order sentences by decreasing order of confidence score. The confidence score is the average confidence score of the SVM and CRF classifiers reported in (Liakata et al., 2012) for a sentence.
- For a dependent category Cat, for which we need n sentences, given the selected sentences m from the corresponding independent category CatI we do the following:
  - If $m = 0$, then treat Cat as independent category for this case.
  - Otherwise, for each selected sentence t_i in CatI, calculate its proximity score to every sentence c_j of the dependent category Cat. $Proximity$ is defined as $1 - Distance$ where $Distance$ is an absolute difference in sentence ids between c_j and t_i normalised by the maximum absolute distance found between all c_j and t_i pairs.
  - The classifier prediction score for each c_j is multiplied by the $Proximity(c_j, t_i)$ score and the sentences are re-ranked according to the new scores, where only the n highest ranking c_js are kept. The last two steps result in an $m \times n$ matrix.
  - If $m = 1$, then the choice for the n sentences for Cat is straightforward.
  - Otherwise, we pick the n highest ranking c_js, proceeding row-wise. Thus, the highest ranking c_js for the highest ranking independent sentences t_i are given priority and any c_j is chosen at most once.

Once the sentence ids are selected for each independent and each dependent category we plug them into the content model. Sentence order is preserved within each CoreSC category. For example, if two Result sentences are selected, the order in which they appear in the paper will be preserved in the summary.

4 Summary evaluation via question answering

4.1 Task Description and experimental setup

We evaluate the extractive CoreSC summaries in terms of how well they enable 12 chemistry experts/evaluators (with at least a Masters degree in chemistry) to answer complex questions about the papers. Our test corpus consists of 28 papers held out from the ART/CoreSC corpus, roughly 1/9, which were annotated automatically with the SVM and CRF classifiers described in (Liakata et al., 2012) trained on the remaining 8/9 of the corpus. For each of the 28 papers in the test corpus, we generated CoreSC summaries automatically using the method described in section 3. We compare the performance of the experts on a question answering (Q-A) task when given the CoreSC summaries and two other types of summary, amounting to a total of three experimental conditions (A,B,C). The other two types of summary are the original paper abstracts (summaries A), in the absence of human summaries, and summaries generated by Microsoft Office Word 2007 AutoSummarize (summaries B).

Microsoft Office Word 2007 AutoSummarize (MA) is a widely available commercial system with reportedly good results (Garcia-Hernandez et al., 2009) and performance equivalent to TextRank (Mihalcea and Tarau, 2004). MA works by assigning a score to each word in a sentence depending on its frequency in the document and sentences are ranked and extracted according to the combination of scores of the words they contain. MA therefore follows classic lexicalised text extraction techniques, is domain independent and is completely agnostic of the discourse. For the latter reason, we considered MA to be a suitable baseline the comparison with which would illustrate the effect of using CoreSC categories on the summary and the merits of having a discourse based model for summarisation.

Neither the paper title nor section headings were available to any of the summarising systems as our extractive system does not make direct use of them and we were not sure how they would influence MA.

To ensure that each evaluator considered only one type of summary per paper, so as to avoid bias from previous stimuli, and to make sure all experts were exposed to all papers and all types of summary, the 12 experts were assigned to four groups (G1-G4) and were allocated 28 summaries each according to the Latin Square design in Table 1.[1].

The experimental setup follows the paradigm of (Teufel, 2001). However, while (Teufel, 2001) developed a Q-A task to evaluate summaries showing the contribution of a scientific article in relation to previous work, the purpose of the Q-A task at hand

[1]Initially we had four experimental conditions but one was dropped, so is not presented in this context

is to show the usefulness of the extracted summaries in answering questions on the paper, and how they compare to a discourse-agnostic baseline. In the case of (Teufel, 2001) the task consists of a fixed set of five questions, the same for all articles tuned particularly to the relation of current and previous work. By contrast, the current Q-A task aims to show how well the summaries represent the content of the entire paper, which means that questions are individual to each paper and required domain knowledge to create.

Each of the 12 experts answered three content-based questions per summary, where the questions were individual to each paper. An example of the questions and the corresponding answers for a given paper can be found below.

Example 4.1.1

- **Q:What do DNJ imino sugars inhibit the action of?**
  A: They inhibit glycosidases and ceramide glucosyltransferases.
- **Q:What methods do the authors use to study the conformation of N-benzyl-DNJ?**
  A: They use resonant two-photon ionization (R2PI), ultraviolet–ultraviolet (UV–UV) hole burning, and infrared (IR) ion-dip spectroscopies in conjunction with electronic structure theory calculations.
- **Q:What is the conformation of the exocyclic hydroxymethyl group?**
  A: The exocyclic hydroxymethyl group is axial to the piperidine ring (gauche- to the ring nitrogen).

As one can see, the questions are complex wh-questions and correspond to answers with multiple components. Questions were complex, to minimise the likelihood of correct random answers. They were designed by a senior chemistry expert with knowledge of linguistics, so that they could be answered based on the abstracts (A). For this purpose, the senior expert chose abstracts that were at least three sentences long. Ideally, the questions and answers should have been set on the basis of the entire paper, but this was not possible given our timeframe for the experiment.The underlying assumption is that a good summary should cover most of the main points of the paper. One of the merits of setting the questions on the basis of the abstracts was that the answers to be identified were deemed sufficiently important to be expressed in the humanly created abstract. However, automatic summaries created in the way proposed here could potentially answer questions beyond the scope of the abstract and in cases of very short abstracts be much more informative.

Experts were told that summaries were automatically generated with no details about different types of summary; it is assumed that none of them is completely familiar with the work mentioned in the 28 papers.

On average, it took experts less than 10 minutes to read a summary and answer the three content-based questions.

	Papers (28)			
Evaluator groups	1–7	8–14	15–21	22-28
G1	A	B	-	C
G2	C	A	B	-
G3	-	C	A	B
G4	B	-	C	A

Table 1: Distribution of summaries to evaluators

4.2 Results and Discussion

We compared each evaluator's answers obtained after reading a summary against the model answers set by the senior expert, the author of the questions, based on the abstract (A) of the corresponding paper. If an evaluator's answer is identical to a model answer, then this counts as "matched".

For instance in example 4.1.1 above, "axial to the piperidine ring", "gauche- to the ring nitrogen" and "The OH6 group is axial (Gauche) to the ring nitrogen" were all considered correct, fully matched answers to the question "What is the conformation of the exocyclic hydroxymethyl group?". In the case of the second question in the same example all of the following were considered correct and fully matched: "Resonant two-photon ionization (R2PI), UV/UV hole-burn, and IR ion-dip spectroscopies in conjunction with electronic structure theory calculations", "R2PI UV/UV hole-burn IR ion-dip e- structure theory calculations" and "a combination of resonant two-photon ionization (R2PI), UV/UV hole-burn, and IR ion-dip spectroscopies in conjunction with electronic structure theory calculations".

If the answer requires listing more than one item (as is the case with questions one and two of example 4.1.1), all of the items have to be matched. Partially matched answers are counted as "partially matched". Non-matching answers can be of two

types. If an un-matched answer coincided with the answer the senior expert would have given after reading that particular summary, then it was marked as "un-matched:justified": Such answers were correct given the particular summary, but are not necessarily correct with respect to the paper and do not count as alternative answers. If the answer was un-matched and also unjustified given the content of the summary, then it was marked as "un-matched:unjustified" . These are cases of evaluator error. Similarly, cases where the evaluator gave "N/A" as an answer were marked as "justified" or "unjustified" according to whether the senior expert could find the answer in the summary or not. The results from marking answers are shown in Table 2.

Number of	A	B	C
Matched	240	126	135
Partially matched	0	4	3
Un-matched:justified	0	25	15
N/A:justified	0	71	71
Un-matched:unjustified	5	11	17
N/A:unjustified	7	15	11
All answers	252	252	252

Table 2: Matches between summary-based answers and model answers

	Micro-AVG			Macro-AVG		
S. types	R	P	F	R	P	F
A	1	0.95	0.98	1	0.95	0.97
B	0.64	0.70	0.67	**0.64**	0.64	0.60
C	**0.66**	**0.75**	**0.70**	**0.64**	**0.70**	**0.65**

Table 3: Precision, Recall and F-score for answering questions using the four types of summary. A: abstracts, B: autosummarize, C:automatic CoreSC summaries.

We report Precision, Recall and F-score (P-R-F) for answering questions given each type of summary (Table 3). To calculate these we define TP as matched answers, FN as N/A:justified and FP everything else (partially matched + un-matched:justified + un-matched:unjustified + N/A:unjustified). Here, the standard definition of recall (TP/(TP+FN)) demonstrates how many questions can be answered using the summary (summary coverage) and Precision (TP/(TP+FP)) how well the questions are answered (summary clarity).

We consider the F-measure to be an overall indicator of the summary usefulness. Micro-averaging is obtained by adding all answers from all papers to calculate TP, FN and FP whereas macro-averaging calculates P-R-F first per paper and then averages over all papers.

The rankings remain consistent regardless of the averaging method. Condition A (abstracts) shows perfect Recall (the evaluators are able to answer all the questions) whereas Precision is affected by unjustified failed matches (Table 2). The perfect recall is hardly surprising as the questions are designed on the basis of the abstract but provides a sanity check for the experiment. The precision sets an upper bound for precision with automatic summaries. Summaries of condition C provide answers to more questions (Recall) and with greater accuracy (Precision) than summaries B. When macro-averaging, the Recall score of summaries C is tied with that for summaries B but Precision is 6% higher.

To verify the statistical significance for the difference in precision and recall for summaries B and C respectively, we performed Monte Carlo sampling 10000 times, for the populations of answers for summaries B and C. During each iteration of sampling, precision and recall were calculated, creating populations of 10000 recalls and 10000 precisions propagated to be representative of the original population of answers. A t-test performed on the population of precision and the population of recalls showed statistical significance at 95% in both cases, with summaries C having a precision of 5% higher and a recall of 1.4-1.6% higher than summaries B (see Table 4). Therefore, we can say that CoreSC summaries C are overall better for answering questions than summaries B.

Comparison between B and C (B-C)	
precision	recall
t = -105.90 df = 19959.79 p-value $<$ 2.2e-16	t = -32.52 df = 19994.40 p-value $<$ 2.2e-16
alternative hypothesis: true difference in means $\neq$ 0	
95% confidence interval: -0.051 -0.049 sample estimates: mean of x mean of y 0.696 0.746	95% confidence interval: -0.016 -0.014 sample estimates: mean of x mean of y 0.639 0.655

Table 4: Test for statistical significance betwen summaries B (microsoft) and C (CoreSC)

The difference in precision between summaries B and C shows the advantage of having a con-

tent model: summaries C are significantly clearer. We had also expected CoreSC summaries to have a much higher coverage than summaries B, and therefore significantly higher recall. However, this difference was less pronounced perhaps because autosummarize favours shorter sentences, which are more likely to be found in the abstracts. We expect that a refinement in the sentence selection criterion, which would also take sentence length into account, will help to showcase further the benefits of using a CoreSC-based content model.

Analysis using ROUGE showed that while summaries C had a slightly higher ROUGE-1 measure than summaries B (0.75 vs 0.73), with respect to abstracts, ROUGE-L was the same for the two (0.70).

In table 5 we also report measurements on summary informativeness based on divergence (Kullback Leibler (KL) divergence and Jensen Shannon (JS) divergence), as in (Louis and Nenkova, 2013). KL divergence is asymmetric and reflects the average number of bits wasted by coding samples of a distribution P using another distribution Q. JS divergence is an information-theoretic measure, reflecting the average distance of the KL divergence between summary and input (the full paper in our case) from the mean vocabulary distributions. Compared to other measures, JS divergence has been found to produce the best predictions of summary quality (Louis and Nenkova, 2013). In practice, what JS divergence tells us is how 'different'/divergent the summary is from the original paper. Low divergence scores are indicative of greater overlap between the summaries and the original paper and are considered positive in terms of the summary information content.

type	KLI-S	KLS-I	UnJSD	SJSD
B	1.66	0.70	0.21	0.19
C	1.40	0.62	0.18	0.17
random	1.61	0.79	0.21	0.19

Table 5: Macro-averaged divergence scores for the 28 test summaries. B: Autosummarize, C: CoreSC, random: random summaries each 20 sentences long for each paper. KLI-S: Average Kullback Leibler divergence between input and summary. KLS-I: Kullback Leibler divergence between summary and input, since KL divergence is not symmetric. UnJSD: Jensen Shannon divergence between input and summary. No smoothing. SJSD:A version with smoothing.

One can see the that CoreSC summaries have consistently lower divergence (both KL and JS) than microsoft autosummarise summaries and random summaries of the same length. This is a positive outcome but since such automatic measures of summary quality have not yet reached maturity and are harder to interpret, we consider the manual evaluation a more reliable indicator of summary informativeness and usefulness. Note that it is not appropriate to use divergence to assess the abstracts as this measure is influenced by the length of a text, which varies dramatically in the case of abstracts.

5 Conclusions and future work

We have shown how a content model based on the scientific discourse as annotated by the CoreSC scheme can be used to produce extractive summaries. These summaries can be generated as alternatives to abstracts. Since they preserve the distribution of CoreSCs in the paper and are not produced independently of it, as is the case with many abstracts, they are potentially more representative of abstracts than the full article. We have tested the usefulness CoreSC based summaries in answering complex questions relating to the content of scientific papers. Extracts from automated CoreSCs are informative, outperform microsoft autosummarise summaries, in both intrinsic and extrinsic evaluation, and enable experts to answer 66% of complex questions with a precision of 75%.

In the future we would like to experiment further with refining the sentence selection method so as to consider criteria for local cohesion, such as lexical chains. We would also like to perform comparisons with automatically induced content models and check their viability for scientific articles. We also would like to perform a human based evaluation of coherence and explore the full potential of these summaries as alternatives to author-written abstracts. This work constitutes a very important step in producing automatic summaries of scientific papers and enabling experts to extract information from the papers, a major requirement for resource curation, which is dependent on constant reviewing of the literature.

Acknowledgements

This work has been funded by an Early Career Leverhulme Trust Fellowship to Dr Liakata and by EMBL-EBI, UK. The authors would like to thank Annie Louis, Yufan Guo, Simone Teufel, Stephen Clark and the anonymous reviewers for their valuable comments. We would also like to thank Mo Abrahams for the python version of the summarisation code and the cafe summary toolkit.

References

S. Ananiadou, Pyysalo S., and J. Tsujii. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004: Proceedings of the Main Conference*, pages 113–120. Best paper award.

R. Brandow, K. Mitze, and L. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31:675–685.

Harr Chen, S. R. K. Branavan, Regina Barzilay, and David R. Karger. 2009. Content modeling using latent permutations. *J. Artif. Int. Res.*, 36:129–163, September.

A. M. Cohen and W. R. Hersh. 2005. A survey of current work in biomedical text a survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6:57–71.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using argumentative zones for extractive summarization of scientific articles. In *COLING*, pages 663–678.

Güneş Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:2004.

R.A. Garcia-Hernandez, Y. Ledeneva, G.M. Mendoza, A.H. Dominguez, J. Chavez, A. Gelbukh, and J.L.T. Fabela. 2009. Comparing commercial tools and state-of-the-art methods for generating text summaries. In *Artificial Intelligence, 2009. MICAI 2009. Eighth Mexican International Conference on*, pages 92–96, November.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 273–283. Association for Computational Linguistics.

T. Kim, J.and Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on BioNLP: Shared Task*, pages 1–9, Boulder, Colorado.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA, June. Association for Computational Linguistics.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 68–73, New York, NY, USA. ACM.

M. Liakata and L.N. Soldatova. 2009. The ART Corpus. Technical report, Aberystwyth University.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valetta,Malta.

M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and Rebholz-Schuhmann D. 2012. Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*, 28:991–1000.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: an exploratory study. *Inf. Process. Manage.*, 27:55–81, February.

Annie Louis and Ani Nenkova. 2012. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1157–1168. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

W. C. Mann and S. A. Thompson. 1998. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1998a. Improving summarization through rhetorical parsing tuning. In *Proceedings of The Sixth Workshop on Very Large Corpora*, pages 206–215, Montreal,Canada.

Daniel C. Marcu. 1998b. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, Toronto, Ont., Canada, Canada. AAINQ35238.

Kathleen R. McKeown. 1985. *Text generation: using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press, New York, NY, USA.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696, Morristown, NJ, USA. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir R Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 555–564. Association for Computational Linguistics.

Vahed Qazvinian, Dragomir R Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 895–903. Association for Computational Linguistics.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2010. Incorporating content structure into text analysis applications. In *EMNLP'10*, pages 377–387.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Comput. Linguist.*, 28:409–445, December.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP-09*, Singapore.

Simone Teufel. 2001. Task-based evaluation of summary quality: Describing relationships between scientific papersworkshop "automatic summarization", naacl-2001. In *NAACL-01 Workshop "Automatic Text Summarisation"*, Pittsburgh, PA.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. Center for the Study of Language and Information, Stanford, California.

Elizabeth White, K. Bretonnel Cohen, and Larry Hunter. 2011. Hypothesis and evidence extraction from full-text scientific journal articles. In *Proceedings of BioNLP 2011 Workshop*, pages 134–135, Portland, Oregon, USA, June. Association for Computational Linguistics.