Multi-label Annotation in Scientific Articles - The Multi-label Cancer Risk Assessment Corpus

James Ravenscroft[1], **Anika Oellrich**[2], **Shyamasree Saha**[3], **Maria Liakata**[4]
[1] University of Warwick, Coventry, CV47AL
[2] King's College London, Denmark Hill, London, SE5 8AF, UK
[3] Queen Mary University London, Mile End Road, London E1 4NS
[4] University of Warwick, Coventry, CV47AL, m.liakata@warwick.ac.uk

Abstract

With the constant growth of the scientific literature, automated processes to enable access to its contents are increasingly in demand. Several functional discourse annotation schemes have been proposed to facilitate information extraction and summarisation from scientific articles, the most well known being argumentative zoning. Core Scientific concepts (CoreSC) is a three layered fine-grained annotation scheme providing content-based annotations at the sentence level and has been used to index, extract and summarise scientific publications in the biomedical literature. A previously developed CoreSC corpus on which existing automated tools have been trained contains a single annotation for each sentence. However, it is the case that more than one CoreSC concept can appear in the same sentence. Here, we present the Multi-CoreSC CRA corpus, a text corpus specific to the domain of cancer risk assessment (CRA), consisting of 50 full text papers, each of which contains sentences annotated with one or more CoreSCs. The full text papers have been annotated by three biology experts. We present several inter-annotator agreement measures appropriate for multi-label annotation assessment. Employing several inter-annotator agreement measures, we were able to identify the most reliable annotator and we built a harmonised consensus (gold standard) from the three different annotators, while also taking concept priority (as specified in the guidelines) into account. We also show that the new Multi-CoreSC CRA corpus allows us to improve performance in the recognition of CoreSCs. The updated guidelines, the multi-label CoreSC CRA corpus and other relevant, related materials are available at the time of publication at `http://www.sapientaproject.com/`.

Keywords: Multi-label annotations, functional discourse, scientific discourse, Core Scientific Concepts, Cancer Risk Assessment

1. Background

1.1. The CoreSC scheme, corpus and applications

In order to extract semantic information from textone needs to pay attention to different aspects of the discourse such as how different sentences or clauses interconnect, alternative mentions of the same entities or concepts, change of theme or topic, communication roles served by different discourse segments (Webber et al., 2012). The largest resource for discourse annotation remains the Penn Discourse TreeBank which contains over 18k explicitly signalled relations (Prasad et al., 2014; Prasad et al., 2011). Core Scientific Concepts (CoreSC)(Liakata et al., 2010; Liakata and others, 2012) is a three-layer functional discourse scheme used to annotate scientific publications at the sentence level. The first layer corresponds to 11 categories (*Hypothesis, Motivation, Background, Goal, Object, Method, Experiment, Model, Observation, Result, Conclusion*), deemed suitable in expressing the structure of a scientific investigation while the second layer provides for the annotation of properties of the concepts (e.g. "New", "Old"). A depiction of the first and second layer in a single flattened representation can be found in Table 1. The third layer of the scheme caters for identifiers (conceptID), which link together instances of the same concept, e.g. all the sentences pertaining to the same method will be linked together with the same conceptID (e.g. "Met1").

CoreSC concepts have been used to manually annotate the discourse structure, at the sentence level, of 265 full text publications in the domains of chemistry and biochemistry (ART corpus) (Liakata and Soldatova, 2009; Liakata et al., 2010) (Liakata and Soldatova, 2009). The annotation was conducted by following a set of 45 page guidelines[1] allocating a single CoreSC to each sentence.

To date the first layer of the scheme (11 CoreSC concepts) has been used to train automated classifiers (Liakata and others, 2012) and the automatically produced annotations have been utilised to create extractive summaries, as alternatives to abstracts (Liakata et al., 2013), to allow fine-grained searches of papers and recognise paper types (e.g. reviews, research papers etc.) (Ravenscroft et al., 2013) and identify drug-drug interactions (Boyce and others, 2013).

A limitation of the previous guidelines and annotation was the provision for a single CoreSC concept per sentence. However, it may be the case that more than one CoreSC concept (e.g. *Goal* and *Method*) are expressed within the same sentence. Here we address this shortcoming by adapting the annotation guidelines to make it possible to assign multiple CoreSC anno-

[1] {`http://repository.jisc.ac.uk/88/`}

Category	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Goal	A target state of the investigation where intended discoveries are made
Motivation	The reasons behind an investigation
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Conclusion	statements inferred from observations & results relating to research hypothesis
Result	factual statements about the outputs of an investigation
Model	A statement about a theoretical model or framework
Observation	the data/phenomena recorded in an investigation
Experiment	An experimental method
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Background	Generally accepted background knowledge and previous work

Table 1: Overview of the CoreSC Annotation scheme

tations. The smallest unit to be annotated are still sentences, even though in principle individual clauses may fit better with CoreSCs and reduce the overlap in annotations. However, we decided against the annotation of clauses, as recently employed clause recognition algorithms seem to be purpose-built for specific application areas (e.g (Del Corro and Gemulla, 2013) or (Kanayama and Nasukawa, 2012)) and prior built clause detection mechanisms performed with F-measures up to 78.63% (Tjong et al., 2001), which in itself could introduce noise to the task of automatically identifying CoreSC concepts.

The new guidelines (49 pages) contain: a decision tree to guide annotators along possible annotation paths by means of questions; detailed description of the semantics of the categories; a revised section on the priority of concepts, which helps reviewers resolve pairwise conflicts between concepts during annotation and also help us assign concepts in the consensus in the case of disagreement between all three annotators; extensive examples from chemistry and biology papers. For more details we refer the reader to the guidelines, released along with the corpus at `http://www.sapientaproject.com`. These guidelines have been used by three domain experts to annotate the Multi-CoreSC CRA corpus, consisting of 50 full papers related to cancer risk assessment. Another three papers were used to train the annotators on the annotation scheme, but these have not been included in the released version of the corpus as the final annotations are the results of discussion rather than being independently annotated.

1.2. Inter annotator agreement for multiple labels

Inter-annotator agreement (IAA) is commonly calculated to assess the quality of annotations in corpus linguistics as well as highlight potential outliers within a document set. Standard IAA scoring mechanisms (such as Cohen's pairwise κ (Cohen, 1960)) assume the assignment of one category label per unit of annotation, which is not directly applicable in this case due to allowing multiple CoreSC categories per sentence. (Rosenberg and Binkowski, 2004) have suggested an extension to the pairwise κ to allow for multiple labels in a setting which assumes an ordering between multiple chosen labels, so that different weights are assigned to agreement on the first, second etc. chosen category, where all weights per annotation unit sum up to 1. We have followed this approach in one of the IAA metrics we report in the following section together with other alternatives for computing kappa.

Other IAA metrics suitable for a multi-label annotation scenario include Krippendorff's α (Krippendorff, 1970; Krippendorff, 2004), which considers difference/distance in annotation on all possible annotation units, irrespective of the number of annotators or labels and the type of annotation (categorical, numeric, ordinal). More recently (Dou and others, 2007) introduced fuzzy Kappa, a version of Cohen's Kappa to allow for annotations with a high degree of subjectivity and which defines an agreement between fuzzy classifiers, incorporating a user-dependent probability distribution on selected values. (Bhowmick et al., 2010) extend this to multiple annotations and change the composition function within the agreement function to make it more suitable for including information on annotator confidence. We report on simple, variants on the

pairwise kappa suitable for multi-label annotation and demonstrating a different degree of strictness.
The rest of the paper is structured as follows. Section 2. describes the papers that were chosen for this corpus and the guidelines which have been employed by the annotators to assign CoreSC concepts. Section 3. describes various measures for inter-annotator agreement and how the annotations from three different annotators were merged to build the gold standard multi-label CRA corpus based on the different inter-annotator agreement measures. Finally, Section 4. shows the results of an evaluation of the corpus using machine learning for automatically assigning CoreSC concepts. The overall experimental setup is illustrated in Figure 1.

2. The Multi-CoreSC CRA Corpus

2.1. Data

The corpus consists of 50 journal papers from the discipline of cancer risk assessment, selected by a domain expert (21 papers from Environmental Health Perspectives, 15 from Carcinogenesis, 9 from Toxicological Sciences, 3 from the Journal of Biological Chemistry, 1 from Occupational and Environmental Medicine, 1 from PlosOne). Each of the 50 papers is annotated at the sentence level with at least one CoreSC (to a maximum of three). The 50 papers correspond to 8,501 unique sentences with sentence count per paper ranging between 85 and 432. Note that the title counts as one sentence and abstracts are included into the sentences count as well as acknowledgments, but the latter, in contrast to the former, are not annotated.

2.2. Annotation Methodology

The previous CoreSC annotation guidelines [2] assumed the assignment of a single CoreSC per sentence. However, this may be rather restrictive when a sentence reports about several different aspects of the scientific discourse. For example, consider the sentence:

"Bone marrow stromal cells were treated with AhR agonists and bacterial lipopolysaccharide (LPS) to mimic innate inflammatory cytokine responses and to study the effects of AhR ligands on those responses."(Jensen and others, 2003)

Here we have both cases of *Method* "Bone marrow stromal cells were treated with AhR agonists and bacterial lipopolysaccharide (LPS)"as well as *Goal* "to mimic innate inflammatory cytokine responses" and "to study the effects of AhR ligands on those responses". According to the previous guidelines that allowed only one CoreSC per sentence, and in case of conflict between several candidates supplied a list of concept priorities, this sentence would have been annotated as *Goal* and we would have missed the *Method* annotation. The categories as described in Table 1 are sorted according to their priorities with the one being listed first possessing the highest priority, and conversely, the one with the lowest priority as last. Priority of concepts was determined based on the observed frequency of the concept in scientific publications, with low frequency concepts such as *Hypothesis, Goal, Motivation* receiving high priority. A revised set of guidelines[3] allow for the annotation of up to three CoreSC concepts and the priority of CoreSC concepts is expressed in terms of annotation ranking (CoreSC1, CoreSC2, CoreSC3). Three concepts were chosen as the maximum, so as to focus annotators and reduce potential sparsity in multi-label annotation. Indeed less than 1% of the total number of sentences were assigned three labels by annotators as can be seen in Table 2 while only 3.25% of sentences in the consensus have more than one annotation. The disparity of the percentage of multiple annotations is a result of the strict manner in which the consensus is obtained and can clearly be improved for multi-label annotation.

# Labels Assigned	C	A1	A2	A3
1	8171	7440	6869	7636
2	274	937	1306	769
3	2	70	53	37

Table 2: # of CoreSCs assigned to sentences according to the annotators (A1-A3) and the consensus C.

Returning to the above example, with the new guidelines the example is annotated as both *Goal* and *Method* (see Figure 2), with Goal taking priority as the rarer category of the two. The guidelines stipulate that "multi-annotation should not be used as a solution for annotator uncertainty with respect to CoreSC assignments; multiple CoreSC annotation should reflect that there are indeed more than one CoreSC present in a sentence". This ensures that we avoid the unnecessary addition of multiple annotations.
Annotation was performed according to the revised set of guidelines by three biology curators, after a first round of annotation of another set of three papers, to calibrate annotations and align the annotators' understanding of the guidelines. Each of the three annotators then proceeded to annotate the same 50 CRA papers and details about this corpus are reported in the next section. To apply the annotations, they used the SAPIENT annotation tool (Liakata et al., 2009). Each of the 50 papers is annotated at the sentence level with at least one CoreSC (to a maximum of three).

[2] http://repository.jisc.ac.uk/88/

[3] see http://www.sapientaproject.com/

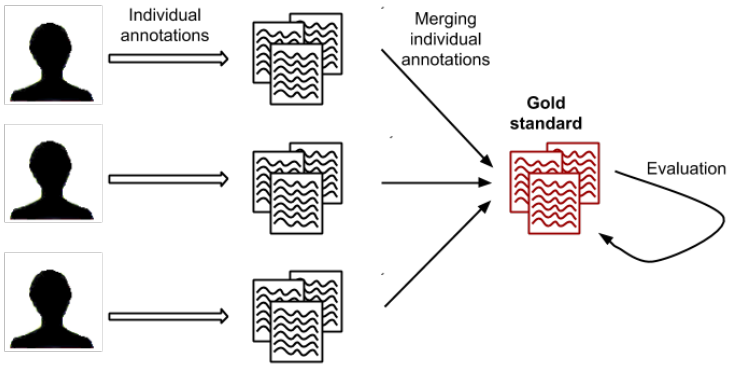

Figure 1: Experimental design for this study. Three annotators independently annotate the corpus, annotations are merged into a gold standard, which is then further evaluated through building a automated classifier from this corpus.

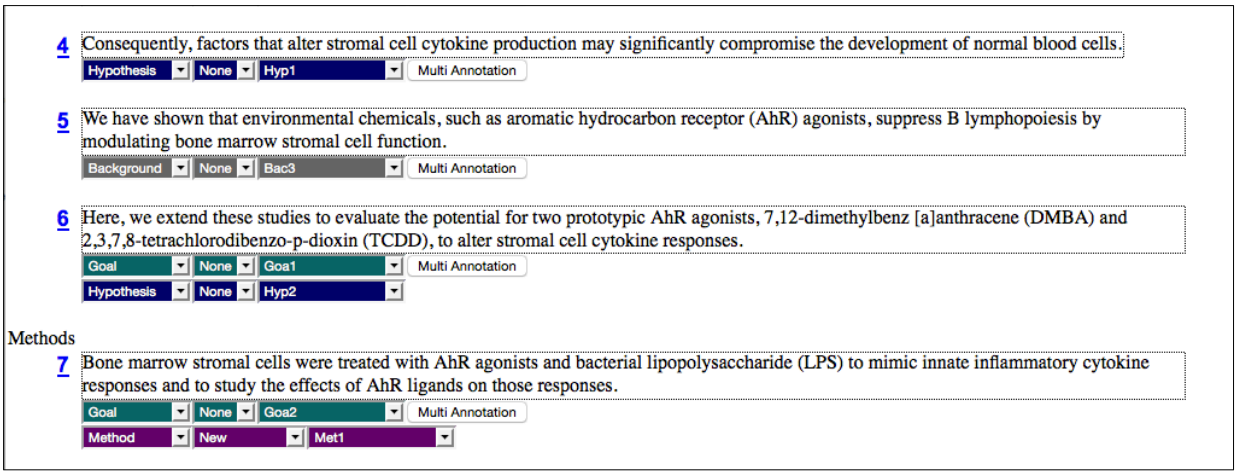

Figure 2: Example of the annotation of multiple CoreSC concepts per sentence using the SAPIENT Tool.

3. Building the multi-CoreSC CRA corpus

3.1. Corpus and Annotator characteristics

We first computed statistics on the annotations to identify the likelihood of multiple CoreSC assignments and to create an annotator profile showing tendencies and preferences for each of the annotators.

Table 3 shows how many sentences were annotated by each annotator with one (SA) or more categories (MA). There was a small percentage of missing annotations, marked as (WA). The table also shows how likely each of the annotators were to use multiple annotations for one sentence. On average, 12.5% of sentences obtained a multi-CoreSC label. Annotator A3, who also had the highest percentage of missing sentences, was the most likely to assign more than one annotation, at 16%, whereas annotator A2 was the least likely at 9.52%. As mentioned in 2.1., the corpus papers contained acknowledgement statements, which were not to be annotated. However, this was not followed strictly by all annotators leading to some variation in the number of annotated sentences across the different annotators.

	SA (%)	MA (%)	WA (%)
A1	7,440 (87.51)	1,007 (11.84)	54 (0.64)
A2	7,636 (89.82)	806 (9.48)	59 (0.69)
A3	6,869 (80.80)	1,359 (15.98)	273 (3.21)
C	8169 (96.09)	276 (3.25)	54 (0.63)

Table 3: The number of sentences that have been annotated with one annotation (SA), multiple annotations (MA) and no annotations (WA), according to each annotator.

Table 4 shows the percentage of annotations for each CoreSC category assigned by each of the three annotators. When we compare these to the the distribution of the same concepts in the CoreSC Corpus (Liakata et al., 2010), we find that there are some domain differences: Background features more prominently here (20-22 vs 17%) and there are more Experiment and Method sentences (20 and 13% here vs 17 and 10 in the ART CoreSC corpus). Goal, Hypothesis and Object are stable at around 2,2 and 4% respectively, Results are sta-

ble around 18% while both Observations and Conclusions are down from (14 and 11% to 9 and 7% respectively). There are virtually no Model sentences in the CRA corpus, showing that this is a practical rather than theoretical discipline. The potential increase in Background and Methodology possibly reflect the need for more justification and grounding in this research, while the decrease in Conclusions and Observations may suggest more speculation in reporting findings, given the severity of erroneous reports.
The results also show that there is variation in perception of the CoreSC concepts by individual annotators, even after alignment during the training phase. For example, compared to their colleagues, annotator A1 tends to over-assign Background, Method and Result and under-assign Object and Observation, A2 over-assigns Observations and under-assigns Conclusions whereas A3 over-assigns Experiments. It is not possible to infer agreement at the sentence level from this table as concept rankings are not included.

3.2. Inter-annotator agreement (IAA)

To calculate inter-annotator agreement (IAA) we used five different variants on the pairwise Cohen's κ, adapted to suit multi-label annotation and the task at hand. There is no widely accepted metric for multi-label annotation of scientific corpora. Our experimental setup involved three annotators curating the same 50 papers, which is why we introduced variants on the pairwise kappa to better determine differences between individual annotators. An assumption we employed here is that if one of the annotators is consistently performing well in all five measures, this annotator is the most reliable. Note that, missing sentences are not taken into account when calculating inter-annotator agreement. The description of the different kappa scores calculated is presented below:

1. **Weighted kappa:** agreement where the order of the CoreSCs matters, following (Rosenberg and Binkowski, 2004). If there is a single category assignment per sentence by an annotator, this category receives a score of 1, when two categories are assigned then the first category receives a score of 0.6 and the second one of 0.4. If three categories are assigned the scores become, 0.6, 0.3 and 0.1 respectively for the first, second and third categories. The formulas for expected and observed agreement are adjusted following (Rosenberg and Binkowski, 2004). This mechanism to score agreement is stricter than other methods employed here as annotators do not only have to agree on whether multiple annotations need to be assigned, but also on the CoreSC category assigned based on the sentence. The weighted kappa is calculated based on

   $K' = \frac{p(A)-p(E)}{1-p(E)}$

   with p(A) the probability of this CoreSC concepts and p(E) the by chance expected probability of the CoreSC concept. The expected probability by chance is calculated with

   $p(E) = \sum_{y=1}^{M} Freq_A[y] * Freq_B[y]$

   where $Freq_A[y]$ and $Freq_B[y]$ are the assigned probabilities of this CoreSC concepts for the two annotators A and B (see 3. for more information on assigned probabilities); M represents the different CoreSC concepts.

2. **Loose kappa:** Agreement on at least one CoreSC, where the order doesn't matter. All assignments receive a score of 1. This is a more relaxed way of scoring as the annotators can have overlap in the annotations, even if they do not agree on how many annotations need to be assigned and what CoreSC categories have to be assigned as first, second or third annotation to a sentence.

3. **All-but-one kappa:** Agreement on all but one, where the order doesn't matter. In most cases this will be equivalent to the Loose kappa but will differ in the case where three annotations have been assigned.

4. **CoreSC1 kappa:** Agreement on the first category, CoreSC1. This measures agreement on the first chosen category only. It simplifies the multi-label problem into a single label problem, where only agreement on the CoreSC1 annotation (first annotation) is taken into account. It means, however, that if there is disagreement on the first label potential agreements on the second and third labels do not count. On the other hand if two annotators agree on the first label but not on the rest, this can return a higher value than weighted kappa. See Tables 5 and 6 that show cases when weighted kappa can be higher than CoreSC1 kappa and vice versa.

5. **Strict kappa:** This measures exact match. This is the strictest type of match, where order and number of agreements matter and we only count exact matches as being correct. The equivalent of a logical AND, where each annotation is given a score of 1.

We present the results for each of the kappa measures by aggregating all agreements for all papers together and then computing the kappa scores (micro-averaging, see Table 5) and by computing the kappa scores per paper and then averaging over all papers (macro-averaging, see Table 6). The results show an agreement of $\kappa > 0.55$ in the case of the weighted, loose and CoreSC1 kappa and $kappa > 0.5$ in the case

	Bac (%)	Con (%)	Exp (%)	Goa (%)	Hyp (%)	Met (%)	Mod (%)	Mot (%)	Obj (%)	Obs (%)	Res (%)
A1	22.2	7	19.1	2.2	2	14.1	0.01	3.6	3.6	4.6	21.7
A2	19.1	5.8	19.8	3	2.3	12.5	0.26	4.5	4.8	10.7	17.2
A3	19.8	7.6	21.1	2	2.2	12.1	0.18	3.2	5.8	8.8	17.2
C	21.3	7.0	21.4	2.1	1.7	11.7	0.0	3.7	3.6	7.4	20.1

Table 4: Percentage of annotations assigned per category in the Multi-CoreSC CRA corpus. Sentences falling into multiple categories were counted multiple times. Rows A1, A2 and A3 illustrate the results obtained by the annotators while row C shows the results of the consensus built from the individual annotators.

of the strict matching kappa, making the quality of annotations satisfactory for training automated classifiers. Table 6 (macro-averaging) shows slightly lower results than Table 5 suggesting that some papers were harder to annotate than others, resulting in more disagreements between annotators.

	weighted	loose	CoreSC1	abo†	strict
A1 - A2	0.633	0.750	0.607	0.750	0.587
A1 - A3	0.593	0.747	0.579	0.747	0.523
A2 - A3	0.571	0.696	0.653	0.696	0.496

Table 5: Table reporting all micro pairwise Kappa measures for the three annotators. Here the kappa measures are computed on the aggregation of all pairwise agreements. † all-but-one kappa measure.

	weighted	loose	CoreSC1	abo	strict
A1 - A2	0.610	0.697	0.631	0.685	0.563
A1 - A3	0.572	0.694	0.586	0.679	0.501
A2 - A3	0.550	0.658	0.558	0.649	0.475

Table 6: Table reporting all macro pairwise Kappa measures for the three annotators. Here pairwsie kappa measures are computed for each individual paper and then the average is computed over all papers.

3.3. Generating consensus from annotations assigned by multiple annotators

As described in the previous section, multiple measures were used to identify the reliability of annotators to guide the step of building a consensus and consequently a gold standard from the assigned annotations. To form this consensus, we followed a conservative methodology, which is explained in more detail in the following. Given the complexity of the task, there are a number of cases that need to be distinguished. The simplest case is that all annotators assign only one annotation and they agree, in which case this annotation is propagated to the gold standard. If a majority vote can be cast on the curator annotations, this majority vote is propagated to the gold standard (see example 1, Table 7). If a majority vote cannot be found because all annotators used a different CoreSC category, the annotation with the highest priority (as discussed in Section) according to the guidelines is propagated to the consensus.For details see example 2, Table 7).

A more complex case is when at least two annotators have chosen a multi-label annotation, by using at least two (maximum three) different CoreSC categories for one sentence. In this case, a multi-label annotation will be propagated to the gold-standard. The gold standard labels are chosen by ranking the CoreSC labels by popularity amongst the annotators whilst also taking into account the label priority. In the case of two distinct labels, the first and second most popular label selections are given scores of 0.6 and 0.4 respectively (see example 3, Table 7). In the case of three distinct labels the top three most popular labels are assigned scores of 0.6, 0.3 and 0.1 respectively. If disagreement about the types of annotation exist, the multi-label annotation of the highest priority CoreSC label will be propagated.

We note here that this procedure has three shortcomings: that we potentially bias the gold-standard towards (i) certain CoreSC labels through the use of our prioritisation table, (ii) the most reliable annotator and (iii) that the measure is conservative and favours single-label annotations. However, applying a conservative measure means also that we only propagate annotations we are confident in, which means that the overall quality of the corpus is expected to be higher. We also expect the bias towards certain CoreSC concepts to be low as the priority ranking was chosen according to their occurrence in text, meaning that rarely occurring CoreSC concepts possess a high priority. Thus, using the label priority in the case of disagreement, favours the selection of rarely occurring CoreSC concept, but there will only be a small number where this rule needs to be applied due to the low occurrence of these concepts. The potential bias toward the most reliable annotator cannot be eliminated, but as we applied a diverse set of scores to determine reliability, we assume higher quality from annotations assigned by this curator. Both decisions were taken to favour a high quality of the resulting gold standard.

3.4. The multi-CoreSC CRA gold standard corpus formed by consensus

An overview of the annotations contained in the gold standard based on the consensus rules described before

ID	Sentence	A1	A2	A3	GS
1	Breast cancer is the most common cancer and the second most frequent cause of cancer death in women.	Bac	Mot	Bac	Bac
2	p38 MAP kinase has been implicated in the regulation of TNF-α production in monocytes and other cell types (28).	Bac	Hyp	Mot	Hyp
3	We have developed an animal model of mammary gland carcinogenesis using a combination of oestradiol and testosterone, and succeeded in inducing a high percentage of female Noble rats to develop mammary cancer in a relatively short time (6 months).	Obj Res Met	Obj Res	Res Met	Res Obj
4	More recently, a number of researchers have demonstrated that, among all plasma steroids, the evidence for association of testosterone levels with breast cancer is strongest, although they could not determine whether this association is a cause or an effect of malignancy (20).	Mot Bac	Bac	Bac	Bac
5	We believe this is a far better model for in depth study of the mechanisms of mammary carcinogenesis, as it closely mimics the natural human breast cancer.	Res Met	Obj	Met	Met
6	Although the amount of testosterone implanted that eventually got converted into oestrogen is not known in this case, a scenario of coexistence of oestrogens (natural and converted) and testosterone can be established.	Res	Hyp	Mot Obs	Res

Table 7: Examples illustrating the algorithm by which consensus among individual annotations has been achieved to form gold standard corpus.

are provided in Tables 3 and 4. Table 3 shows that the annotators tend to disagree on when multiple annotations need to be assigned to a sentence which is why the resulting consensus largely (96%) consists of sentences with single annotations. However, these single annotations per sentence may be more accurate overall as opposed to the previously published corpus as the final decision on the gold standard annotation has been derived from a conservative consensus, taking the priority of concepts and the most reliable annotator into account.

Table 4 shows that distributions of the different CoreSC categories are more or less the harmonised values from the different annotators. However, the Model category ('Mod') stands out from the others in that it almost disappears from the gold standard. This is due to following the most reliable annotator (A1) and the priority weightings in case of disagreement between annotators (see Section 3.2.).

4. Evaluation of the multi-label CRA corpus through training a machine learning classifier for CoreSC annotation

As described in the previous sections, annotations in the CRA corpus were assigned by three curators according to a multi-label extension of the guidelines used to create the ART/CoreSC corpus (Liakata et al., 2010). The annotations from the three curators are then harmonised into one gold standard as described in Section 3.2.. We wanted to verify to what extend this new gold standard corpus can help us train a classifier for the automatic assignment of CoreSC concepts to sentences. For this purpose, we used the trained Conditional Random Field (CRF) model (see (Liakata and others, 2012)) to automatically assign CoreSC concepts to the CRA corpus and evaluated the results on the gold standard. Using the old model on this novel corpus will elucidate how robustly the automatic assignment of CoreSC can be learnt independently of the domain that the papers belong to. In the case where CoreSC categories are heavily domain-dependent, we expect the old model to have a poor performance for those categories.

In a second evaluation step, we trained a set of new CRF models by using three-fold cross-validation based on the gold standard of the new CRA corpus. However, in this step, we only included the highest ranking CoreSC concept for each annotated sentence in the corpus, which automatically possesses the highest priority with respect to the annotator consensus and priorities defined in the guidelines. After training the new model, the performance measures between old and new models can be compared, to distinguish between categories that can be well distinguished across domains and those that seem to be possess domain-specific rules. We also obtain an indicator of the quality of the new corpus (CRA) in training new classifiers.

4.1. Evaluation results

The original CRF classifier presented in Liakata et al 2012 (Liakata and others, 2012) achieved 50.4% accuracy during 9-fold cross validation on the ART corpus, upon which it was trained. The same model was re-run on the complete CRA corpus and used to identify the most relevant label for each sentence, achieving 51.9%

accuracy without any modification or feature adaptation. This result, run on the multi-label CoreSC CRA corpus, demonstrates that the original SAPIENTA classifier, trained on the original bio-chemistry corpus did learn a good model for the annotation of scientific discourse that can port well across these two related but distinct disciplines. This result also suggests that the multi-label CoreSC CRA corpus may have been annotated more consistently, which could be either due to the domain, the selection of papers incorporated or the fact that annotators can assign now multiple labels, which we than harmonise to obtain the final corpus.

Label	Precision		Recall		F-measure	
	Old	New	Old	New	Old	New
Bac	56.9	**69.9**	48.9	**84.7**	52.6	**76.7**
Con	30.1	**58.0**	**66.0**	50.3	41.3	**53.9**
Exp	**80.1**	78.0	81.7	**89.5**	80.9	**83.4**
Goa	55.0	**60.1**	35.7	**49.4**	43.3	**54.2**
Hyp	**25.6**	25.0	07.5	**95.9**	11.6	**13.9**
Met	55.0	**65.0**	42.9	**52.9**	48.2	**58.3**
Mod	00.0	0.00	00.0	0.00	00.0	0.00
Mot	**63.2**	52.0	03.9	**38.0**	07.4	**43.9**
Obj	30.1	**37.0**	**23.0**	21.2	26.1	**27.0**
Obs	30.1	**50.3**	**42.0**	37.6	35.1	**43.0**
Res	47.6	**66.8**	47.3	**69.9**	47.5	**68.3**

Table 8: Classifier results for CRF model trained on ART corpus run on CRA corpus (Old) and CRF model trained on CRA corpus (New) and run on CRA corpus. Best outcome for each measurement in bold. F-measure improves for all CoreSC categories suggesting that most influential features for CoreSC annotation are domain specific.

Table 8 shows Recall, Precision and F-measure for both the 2012 CRF model on the CRA corpus as well as the measures of the newly trained CRF models based on the CRA corpus. Both models fail to allocate the 'Mod' label to any sentences, producing a score of zero for Recall, Precision and F-measure. This may be attributed to a combination of a dependency upon domain specific features describing the 'Model' class and there being very few 'Model' sentences in the consensus corpus for the new model to learn from. As shown in Table 4, due to the variation in perception of what constitutes a Model sentence among the annotators, only four sentences obtain the label 'Mod' in the gold standard, based on which the new model was trained.

With respect to the label 'Mot', the old CRF model trained on the biochemistry corpus achieves good Precision but conversely poor Recall for the respective sentences. It is quite likely that the n-grams contributing to the recognition of these sentences in the old corpus do not appear with regularity in the new corpus since the recall for 'Mot' improves by a factor of 10 when retrained on the new corpus. The high recall for 'Hyp' is perhaps due to the priority bias of the consensus towards this rarest of categories. The 'Exp' label for Experiment was most successfully identified with no domain adaptation required, with an initial (Old) F-measure of 80.9 on the CRA corpus.

Table 8 also shows that the F-measure improves for all categories for the newly trained model. The best category recognised is still Experiment, which is probably unsurprising as this is the largest group of annotations in the CRA corpus (see Table 4). While the categories 'Con', 'Goa', 'Mot', and 'Obs' are all comparatively small subsets of the overall corpus (between 2.1% and 7.4%), the model achieves respectable performance measures for this group: F-measures all in the range of 43% to 54.2%. Especially, the Goal category seems to be sufficiently concise as it constitutes only 2.1% of the gold standard but still yields an F-measure of 54.2%. These categories all mark an improvement compared with their recognition in the ART Corpus. The Object category labeled with 'Obj' seems to cause the biggest problems for the newly trained classifier with a performance of 27% F-measure while accounting for 3.6% of annotations in the gold standard. Improvements on either annotation guidelines or learning features need to be explored for this category in future work.

5. Conclusions

We have described the creation of the Multi-CoreSC CRA corpus, based on an updated set of annotation guidelines originally published in (Liakata et al., 2010). The corpus was annotated by three independent annotators and we presented results on several IAA measures which show that agreement levels are sufficient (weighted kappa > 0.55) for training machine learning methods on this corpus. Furthermore, we have described a conservative algorithm to derive a consensus gold standard annotation set from the three individual annotation sets. Based on this multi-CoreSC gold standard, we trained a CRF model and can show that this novel model outperforms the previously published model. All materials generated within this study are available from the project web pages at `http://www.sapientaproject.com/`.

Further work needs to address model performance on low scoring categories such as Hypothesis and Object. Furthermore, we plan to investigate domain adaption techniques, to investigate whether models trained in one domain can be transfer to another and whether this can positively influence the performance of the employed machine learning techniques. The ability to train annotation models to predict secondary and tertiary CoreSc labels was also not explored in this paper, nor was the direct impact of these additional annotations as a feature to inform a classifier about what primary label to assign to a sentence. We plan to work on these latter

topics as future work stemming from the Multi-CoreSC CRA corpus.

6. Bibliographical References

Bhowmick, P. K., Basu, A., and Mitra, P. (2010). Determining reliability of subjective and multi-label emotion annotation through novel fuzzy agreement measure. In *LREC*.

Boyce, R. D. et al. (2013). Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness. *Journal of Biomedical Semantics*, 4:5.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. International World Wide Web Conferences Steering Committee.

Dou, W. et al. (2007). Fuzzy kappa for the agreement measure of fuzzy classifications. *Neurocomputing*, 70(4â6):726 – 734. Advanced Neurocomputing Theory and MethodologySelected papers from the International Conference on Intelligent Computing 2005 (ICIC 2005)International Conference on Intelligent Computing 2005.

Jensen, B. et al. (2003). Aryl hydrocarbon receptor (ahr) agonists suppress interleukin-6 expression by bone marrow stromal cells: an immunotoxicology study. *Environmental Health: A Global Access Science Source*, 2(1):16.

Kanayama, H. and Nasukawa, T. (2012). Unsupervised lexicon induction for clause-level detection of evaluations. *Natural Language Engineering*, 18(01):83–107.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological methodology*, pages 139–150.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality & quantity*, 38:787–800.

Liakata, M. et al. (2012). Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*.

Liakata, M. and Soldatova, L. (2009). The art corpus. Technical report, Aberystwyth University.

Liakata, M., Q, C., and Soldatova, L. N. (2009). Semantic annotation of papers: Interface & enrichment tool (sapient). In *Proceedings of the BioNLP 2009 Workshop*, pages 193–200, Boulder, Colorado, June. Association for Computational Linguistics.

Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C. R., et al. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *LREC*.

Liakata, M., Dobnik, S., Saha, S., Batchelor, C. R., and Rebholz-Schuhmann, D. (2013). A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *EMNLP*.

Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. (2011). The biomedical discourse relation bank. *BMC Bioinformatics*, 12(1):1–18.

Prasad, R., Webber, B., and Joshi, A. (2014). Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*.

Ravenscroft, J., Liakata, M., and Clare, A. (2013). Partridge: An effective system for the automatic cassification of the types of academic papers. In *Research and Development in Intelligent Systems XXX*, pages 351–358. Springer.

Rosenberg, A. and Binkowski, E. (2004). Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 77–80. Association for Computational Linguistics.

Tjong, E. F., Sang, K., and Déjean, H. (2001). Introduction to the conll-2001 shared task: clause identification. In *Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7*, page 8. Association for Computational Linguistics.

Webber, B., Egg, M., and Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, 18(04):437–490.

7. Language Resource References

Maria Liakata and Larisa Soldatova. (2009). *ART Corpus*. ART Project, 1.0.