A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality

Ziqi Zhang[1], Sam Chapman[2], Fabio Ciravegna[1,2]

[1] Department of Computer Science, University of Sheffield, UK
[2] K-Now, UK
{z.zhang@dcs.shef.ac.uk, sam@k-now.co.uk, f.ciravegna@dcs.shef.ac.uk}

Abstract. Manual document annotation is an essential technique for knowledge acquisition and capture. Creating high-quality annotations is a difficult task due to inter-annotator discrepancy, the problem that annotators can never agree completely on what and exactly how to annotate. To address this, traditional document annotation involves multiple domain experts working on the same annotation task in an iterative and collaborative manner to identify and resolve discrepancies progressively. However, such a detailed process is often ineffective despite taking significant time and effort; unfortunately, discrepancies remain high in many cases. This paper proposes an alternative approach to document annotation. The approach tackles the problem by firstly studying annotators' suitability based on the types of information to be annotated; then identifying and isolating the most inconsistent annotators who tend to cause the majority of discrepancies in a task; finally distributing annotation workload among the most suitable annotators. Tested in a named entity annotation task in the domain of archaeology, we show that compared to the traditional approach to document annotation, it produces larger amounts of better quality annotations that result in higher machine learning accuracy while requires significantly less time and effort.

Keywords: inter annotator disagreement, annotator discrepancy, document annotation, knowledge acquisition, machine learning, named entity recognition.

1 Introduction

Manual document annotation is the basis to provide data for training and evaluating a supervised machine learning system. It has been recognised that the annotation process is often laborious and costly, and has been the major bottleneck to the development and adaptation of knowledge acquisition systems [6][22]. Crucial to the annotation process is resolving annotator discrepancies and achieving reasonable inter-annotator agreement, the problem stems from annotators behaving differently and inconsistently for the same annotation task. This is due to the differences in their skills, knowledge and experiences, and issues such as workload and tiredness. The problem affects the quality of annotation and therefore, the learning accuracy of a system [2][29]. For this reason, the typical annotation process requires a number of

domain experts to work in an iterative and collaborative manner in order to discover and resolve discrepancies progressively. Usually in each iteration, a set of documents are duplicated across all annotators, who are required to annotate the same documents for the same types of information (e.g., person and place names) independently. Outputs from different experts are then cross-checked, discussed, validated, consolidated and a sophisticated annotation guideline is documented, followed, and refined [10][17][24] in following iterations. The process is repeated as much as possible until the level of discrepancies is reduced to a satisfactory level. Such a repetitive process often requires months and even up to years of work from experienced researchers [2][29], yet discrepancies can never be eliminated [14] and the resulting annotations and guidelines are often application-specific and non-generalisable. Given such an ineffective and inefficient process, the tremendous cost from key experts means that it is inapplicable in many practical situations such as industries, due to resource limitations (e.g., finance, time and personnel) [15]. A more effective and efficient approach is required.

Essentially, the majority of discrepancies among annotators are caused by the differences in their knowledge and experiences [14]. The traditional annotation process identifies these differences and aims to minimise them iteratively, eventually producing an output that best matches the subtly varying viewpoints across a community. We argue that, these differences result in different levels of annotators' suitability for an annotation task or sub-tasks. In most cases, no candidates are perfectly suitable for all tasks; however, one can be more suitable for particular tasks than others (e.g., based on the classification scheme in an entity classification task). Therefore, the key to improving annotation quality is not correcting the differences revealed by the repetitive checking process at the maximum effort, but rather identifying annotators' suitability and suitability-based task assignment. Inconsistent annotators unsuitable for a task should be identified and isolated such that the annotation quality is not compromised.

This paper details an alternative approach to document annotation based on the analysis of annotators' suitability for annotation tasks and annotator selection based on their suitability. We illustrate the approach using a typical annotation task; named entity annotation and classification, in which annotators' suitability analysis and annotator selection are carried out on the per-entity-class basis, namely, the type of information to be annotated. The studies reveal different levels of discrepancy and individual inconsistency for different classes of entities, suggesting the task be split and treated differently in the annotation process. The nature of task specialisation allows one to assign specific sub-tasks to the most suitable, mutually consistent annotators, among whom workload may be distributed. A set of experiments are designed and performed to show improved quality of annotations compared to those obtained by the traditional approach; whilst the time required for annotation is significantly reduced, but the total amount of annotations produced is largely increased.

The rest of this paper is organised as the follows: Section 2 gives an insight to the difficulties of document annotation and the prevailing problems of annotator discrepancy. It then reviews the typical document annotation process adopted in most scientific research and summarises important lessons learnt from these work. Section 3 proposes our alternative method to manual document annotation, and describes the

study carried out in a large archaeology named entity annotation task, with details on our experiment design, results and findings. Section 4 is a further discussion of the results and other problems noted in this study, and finally concludes the paper.

2 Isn't It Easy to Annotate?

In the past few decades extensive amount of research has been dedicated to automate the process of knowledge acquisition, in which creating manual annotations for training or testing an automated system is an essential step. Typically in the field of Machine Learning, high quality and sufficient quantity of annotations are required for an automated system to be able to learn accurately. Unfortunately, the process of creating the necessary annotations has never been an easy or rewarding experience.

2.1 Annotator Discrepancy

Research has shown human annotators can never agree completely with each other on what and how to annotate [14], and they even tend to disagree with themselves in some situations [7]. The first case is often referred to as *inter*-annotator "agreement", "consistency" or "discrepancy". The second case is referred to as *intra*-annotator "agreement", "consistency" or "discrepancy". Inter-annotator discrepancies are often caused by the differences in annotators' knowledge and experiences, their understanding and reasoning of the corpora [17]. Intra-annotator discrepancies exist because annotators' level of interest and motivation may drop and level of fatigue rises as the annotation process continues [12], as a result, annotators make mistakes. This paper focuses on the *inter*-annotator issue only.

Inter-annotator discrepancy is a prevailing issue in the research of knowledge acquisition. It has been noted in many relevant fields, such as Word Sense Disambiguation (WSD) [23], Speech Recognition [12], Information Retrieval [1], event recognition [17] and Named Entity Recognition (NER) [10][28][29]. Depending on the difficulty of the annotation task, the inter-annotator agreement can vary significantly. For example, [27] indicated that the agreement between human annotators varied between 40% and 75% for different tasks. Most reports of inter-annotator discrepancy are found in the field of NER, which concerns recognising and classifying atomic texts into pre-defined categories. Research by [10] and [8] has shown that in NER, discrepancies typically arise due to three types of difficulties in annotating entities. Firstly, it is difficult to choose the right category (e.g., Ben Nevis can refer to person or a mountain in the UK); secondly, it is difficult to select the candidate texts and delimitation boundaries (e.g., should we annotate proper nouns only, or also pronouns and definitional descriptions); thirdly, how to annotate homonyms, e.g., "England" may refer to a location or a football team. These problems become even harder to resolve within specialised domains such as bioinformatics and engineering, due to the intrinsic complexity of terms in these domains including multi-word expressions, complex noun phrase compositions, acronyms, ambiguities and so on [28]. Typically, the inter-annotator agreement in NER found in these domains is between 60% and 80% [29][5][21].

From all these studies, it is evident that perfect agreement between annotators is difficult to reach, and it is also difficult to obtain a high level of inter-annotator consistency, especially in specialised domains. However, researchers advocate that consistency highly increases the usefulness of a corpus for training or evaluation purposes, and it is crucial to the success of machine learning algorithms [2][29]. Therefore, studies have been conducted to research scientific methodology for creating high quality annotations, addressing inter-annotator consistency. In the following, a list of these literature is described.

2.2 How to Annotate Properly: What Have We Learnt?

The typical process of annotating a corpus often involves a group consisting of a number of domain experts and ideally also linguists working on a same range of annotation tasks in an iterative and collaborative approach aimed at resolving discrepancies. For example, in NER, multiple domain experts are required to annotate a corpus for the same sets of entity classes. In each iteration, a duplicated set of documents are annotated by each domain expert independently. Then, their output is cross-checked; discrepancies are discussed and resolved as much as possible. The entire process and decision making logic is documented to form a guideline for the annotation task, which is to be followed in future exercises. Due to the nature of the work, it is always a lengthy and costly process. The guidelines are often subject to the specialised domain and not generalisable to other problems.

For example, Brants [2] reports their work on creating syntactic annotations (part-of-speech and structural information) on a German newspaper corpus. The activity involves trained annotators performing the annotation tasks at sentence level independently, then cross-checking and discussing together to resolve discrepancies. They report that a trained annotator needs on average 50 seconds per sentence, with average of 17.5 tokens; however, the total annotation effort including the consolidation activity increases to 10 minutes per sentence. Pyysalo et al. [26] annotates a corpus of 1100 sentences from abstracts of biomedical research articles for biomedical named entities, relationships between entities and syntactic dependencies. They also adopt a repetitive process, which took 15 man-months of effort. Wilbur et al. [29] conduct experiments to investigate inter-annotator agreement in a text annotation task in biomedical domain and identify factors that can help improve consensus. Their experiment involves twelve annotators annotating the same set of 101 sentences. Multiple iterations were conducted in a period of over one year, during which they develop and refine a guideline considered applicable for similar annotation problems. The resulting inter-annotator agreement stayed between 70% and 80%. They conclude that annotators must have a good understanding of the language and experience in reading scientific literature, and must be properly trained in order to deliver high quality annotations. Also, they indicate the presence of a clear, well developed annotation guideline as critical.

Other researchers have also recognised the necessity for a clear annotation guideline. Kim et al. [17] show by experiments that high level of discrepancy will form without annotation guidelines even if the task is carried out by well-educated domain experts. Their studies on event annotation on the Genia corpus [24] took 1.5

years of effort of five graduate students and two coordinators. Whenever new annotators joined the project, they had to be trained using previously annotated examples and follow the guideline. Colosimo et al. [5] and Tanabe et al. [28] also conduct corpus annotation in the biology domain and conclude that clear annotation guidelines are important, and the annotations should be validated by proper inter-annotator-agreement experiments.

Even if well-prepared guidelines are available for annotation problems, they are not the ultimate answer to the problem. Firstly, most guidelines are lengthy documents and are difficult to read. For example, Ferro et al. [9] design guidelines for annotating temporal information, which has 57 pages. The entity recognition task defined by ACE [18] is accompanied with a guideline of over 70 pages for annotating only five classes of entities [25]. Secondly, interpretation of the guideline documents differs from annotator to annotator; as a result, some annotation criteria remain problematic and can cause discrepancies [10]. For example, the event annotation on the Genia corpus by [17] only achieves 56% inter-annotator agreement with strict match [20] even though all annotators have been trained and educated using example annotations and guidelines.

2.3 The Reality Check

The conclusion of previous research advocates for clear definition of annotation guidelines to be followed, well-educated domain experts with proper training in document annotation, careful study of inter-annotator agreement and iterative attempts to address the issues revealed by the study and to resolve discrepancies, all of which demand costly investment. Many scientific research tracks such as MUC [11] present a scenario in which the cost of such effort is not considered important [6]. However, the scenario breaks as the technology is to be adopted by various specialised domains, in which the cost is a serious issue [22]. Industries and businesses are not willing to invest resources (personnel, finance and time) into lengthy document annotation exercises [15]; annotators feel overwhelmed by the scale of monotonous annotation tasks expressing a strong reluctance to doing them. They want a shortcut.

One exception to this is the domain of bio-informatics, where well-curated resources are richly available and users are more familiar with the benefits that can follow from annotation. Unfortunately, these resources are hardly re-usable across domains because they address specific issues in bio-informatics; and demands for similar resources in other specialised domains such as aerospace engineering, astronomy and arts and humanity are equally high, these however are scarcely addressed [15][21][16].

Recognising the urgency of this issue, in the last decade there has been an enormous amount of research dedicated to weakly-supervised learning methods [22] and domain-adaptation for Machine Learning [19] in order to reduce a learning system's dependence on manually annotated data. Unfortunately, evidence of these methods applied to specialised domains is scarce. Their applicability in these areas is questionable given the intrinsic complexity of language and decreased availability of knowledge resources in these areas.

Given the complexity of these problems and the inadequacy of existing technologies, this work has identified a strong demand for more effective, efficient and practical approaches to manual document annotation. In the following, we propose a new method to manual document annotation to achieve this goal.

3 Towards a New Document Annotation Approach

In this section, we propose a new approach towards effective and efficient manual document annotation. We present the details using a case study of named entity annotation in the domain of archaeology; however, the methodology is generic and can be applied to other annotation problems. Following the traditional named entity annotation process, in each iteration, domain experts are required to annotate a set of documents for the same set of entity classes. Then for each entity class, all annotations are cross-checked, and discrepancies are identified, discussed and resolved as much as possible. In contrast, our method is based on the hypothesis that the different levels of knowledge and experiences of annotators lead to different levels of suitability for an annotation task, or sub-tasks. This is reflected by different levels of discrepancies they demonstrate in annotating different classes of entities. Therefore, annotator discrepancies and suitability must be studied on per-entity-class basis, and only the most suitable annotators should be selected for annotating specific entity classes other than all classes.

The method contains three phases. In the first phase, we follow the traditional approach to manual document annotation to create sufficient amount of annotations that sample the level of discrepancy in this task. The size of this corpus is properly controlled such that efforts required from annotators are minimised to an acceptable level and the annotations created are just adequate for studying the inter-annotator agreement. In the second phase, a set of experiments are carried out to evaluate machine learning accuracy using these annotations. The results together with the inter-annotator agreement studies in phase one are used to evaluate annotators' suitability of annotating the documents for a particular class of entity, and then specific annotators (*best-fit-annotators*) are chosen to annotate the classes of entities (*best-fit-class*) for which they are most suitable. In the third phase, the final set of documents to be annotated is selected. Then for each class of entity, the documents are split equally between each member of the *best-fit-annotators* to annotate just for that class, i.e., their *best-fit-class*. This ensures all documents are annotated by the most consistent annotators for all entity classes, while no annotators perform redundant work. Compared to the traditional approach, this is a desirable feature since the distributional nature of work in the final phase allows workload to be reduced and total output to be increased. A set of experiments are then carried out to evaluate the machine learning accuracy obtainable on this corpus.

3.1 The Archaeology Domain

The domain of modern archaeology is a discipline that has a long history of active fieldwork and a significant amount of legacy data dating back to the nineteenth

century and earlier. Despite fast-growing large corpora existence, little has been done to develop high quality meta-data for efficient access to the contained information in these datasets, and there is a pressing need for knowledge acquisition technologies to bridge the gap [16]. Manual document annotation in archaeology is a challenging task because of the complexity of language characterised by ambiguities, uncertainties, long and composite terms, changing language use over the extended timeframe of the corpora, acronyms and so on. As a result, low inter-annotator agreement has been noted in related work [3].

Our work deals with archaeological entity extraction from un-structured legacy data, which mostly consist of full-length archaeological reports archived by the Arts and Humanities Data Service (AHDS[1]). The reports vary from five to over a hundred pages. According to [16], three classes of entities are most useful;

- Subject - topics that reports refer to, such as findings of artifacts and monuments. It is the most ambiguous class because it covers various specialised domains such as warfare, architecture, agriculture, and machinery. For example "Roman pottery", "spearhead", and "courtyard".
- Temporal terms - archaeological dates of interest, which are written in a number of ways, such as years "1066 - 1211", "circa 800AD"; centuries "C11", "the 1st century"; concepts "Bronze Age", "Medieval"; and acronyms such as "BA" (Bronze Age), "MED" (Medieval).
- Location of interest - place names of interest, such as site addresses and site types related to a finding or excavation. In our study, these typically refer to UK-specific places.

3.2 The First Phase – Sampling Annotator Discrepancy in NER for Archaeology

Overview of the Procedure In this phase, five documents were randomly selected from the AHDS archive. Each document varied from five to thirty pages, containing much more content than standard datasets used in MUC and abstracts used in bioinformatics NER. The size of the corpus was decided by the domain experts, who considered the workload to be acceptable. Meanwhile, the selection of documents was ensured such that there were sufficient contents for annotation (as indicated by the *tag density* and number of annotations revealed in the post-annotation statistical analysis). The total number of words in this corpus was 47,101, and the average *tag density* (the percentage of words tagged as entities) by all annotators was 8.7%, compared to MUC7 11.8% and Genia 33.8% [21]. The average total number of annotations for all three classes was approximately 2,100. This corpus is referred to as "*trial corpus*". It was then to be annotated by five full-time archaeology researchers in three iterations following the traditional document annotation approach.

Throughout phase one, two annotators were constantly involved in all meetings with knowledge acquisition (KA) experts to provide feedback from all annotators and design simple annotation guidelines and ensure they are followed. The annotation process consisted of four mini-iterations. In the first iteration, two annotators made trial attempts at annotating two medium sized documents from the *trial corpus*.

[1] http://ahds.ac.uk/

Discrepancies were identified at this early stage and were discussed and resolved in the meeting with the KA experts. The output of this process were some guidelines for annotation, which were then provided to all five annotators in the second iteration, during which each annotated 1 ~ 2 documents. The purpose of this exercise is again to identify as many discrepancies as possible at low costs. By studying these annotations, the guideline for annotation was further refined and enriched. In the third iteration, all five annotators were required to follow the guideline to re-annotate the *trial corpus* independently and fully in a series of intensive workshops. In the final iteration, one annotator undertook final validation by checking 10% of all annotations to correct obvious mistakes that violated the guidelines. These corpora are used to study inter-annotator consistency and machine learning accuracy.

Cost of the Process Thanks to the size of the sample corpus, according to the annotators' estimation, the first iteration of phase one took 2 person-days of work; the second iteration took 5 person-days of work; the third iteration took 5 person-days of work; and the final iteration took 2 person-days of work. The total estimated cost in terms of person-days work is 14.

Inter-Annotator Agreement Many different measures are available for computing inter-annotator agreement, and the most popular is the κ-statistics [4]. However, it is not suitable for entity recognition tasks [26]. We adopt the *F-measure* proposed by [13], which allows computing pair-wise inter-annotator agreement using the standard *Precision*, *Recall* and the harmonic *F-measure* in information studies by treating one annotator as gold standard and the other as predictions. Table 1 shows the pair-wise agreement for each entity class.

Table 1 Pair-wise inter-annotator-agreement *F-measure*. A, B, C, D, E are identifiers of domain-expert annotators

Location					
	A	*B*	*C*	*D*	*E*
A	1	0.8	0.69	0.77	0.66
B	0.8	1	0.72	0.75	0.75
C	0.69	0.72	1	0.69	0.7
D	0.77	0.75	0.69	1	0.69
E	0.66	0.75	0.7	0.69	1

Temporal					
	A	*B*	*C*	*D*	*E*
A	1	0.83	0.77	0.79	0.77
B	0.83	1	0.67	0.77	0.83
C	0.77	0.67	1	0.78	0.71
D	0.79	0.77	0.78	1	0.77
E	0.77	0.83	0.71	0.77	1

Subject					
	A	*B*	*C*	*D*	*E*
A	1	0.55	0.65	0.63	0.62
B	0.55	1	0.51	0.53	0.49
C	0.65	0.51	1	0.51	0.51
D	0.63	0.53	0.51	1	0.5
E	0.62	0.49	0.51	0.5	1

Comparing the figures, it is evident that even with reasonable effort from well-trained and skilled archaeology professionals devoted to developing annotation guidelines and resolving discrepancies in several iterations, the task of annotating domain specific entities remained difficult and the level of discrepancy remained high. Annotating *Subject* is a much harder task than the other two classes of entities. This is expected because *Subject* spans across multiple specialised domains and terms are characterised by a high level of ambiguity and heterogeneity. Most discrepancies were due to identifying the boundaries of composite noun phrase entities, acronyms and identifiers (object codes, ID's). Also for every class of entities, we can always identify sub-groups of annotators that are more mutually consistent than with other annotators. This raised the issue of annotator suitability and the question that it is beneficial to eliminate in-consistent annotators from an annotation task to reduce discrepancies.

3.3 The Second Phase - Evaluating Machine Learning Accuracy and Annotator Selection

In order to gain a different view of the quality of the annotations produced in such an iterative way, two sets of experiments were conducted to evaluate how well a machine can learn from these annotations. In the first set of experiments, we created a corpus including annotations from all annotators to reflect the high level of discrepancy in the annotations. Annotations produced by the five annotators were selected randomly, whilst ensuring the five documents are covered in full and roughly proportional annotations were selected from each annotator. This corpus is referred to as *consolidated-trial-corpus*. In the second set of experiments, we used each individual annotator's corpus separately, thus there were five corpora for testing and they are referred as *individual-trial-corpus*. On each of these six corpora, an SVM[2]-based named entity tagger was trained and evaluated in a five-fold cross validation experiment, in which annotations are randomly split to 5 complementary subsets[3], and the learning algorithm learns from four subsets and is then validated on the other one subset. The process is repeated for 5 iterations, where each time different subsets are used for training and validation and the final performance is the average of the performance figures obtained in all iterations. Throughout the experiment we kept consistent settings (parameters, features, etc.) for the learning algorithm in order to fairly compare the effect of corpus quality.

Firstly, we applied the experiment on the *consolidated-trial-corpus*, which had inter-annotator inconsistency as discussed in the previous section. We refer to this as *collective-annotator-learning*. Next, we applied the experiment on each *individual-trial-corpus* that was annotated by a single annotator. Since there was only one annotator for each corpus, this is equivalent to perfect inter-annotator agreement. We refer to this as *intra-annotator-learning*. Results of these are shown in Table 2.

[2] http://www.support-vector-machines.org/

[3] To cope with varied document lengths we split documents into sections of sentences.

Results of this set of experiments show interesting findings. Given no inter-annotator issues in each individually annotated corpus, one would expect higher levels of consistency and better annotation quality, which translate to better machine learning accuracy. This was mostly true compared to results obtained on the *consolidated-trial-corpus*. However, exceptions were noticed for annotator A on *Location* (2 percent lower), and E on *Temporal* (1 percent lower). Also, comparing across different entity types for each annotator, the entity tagger had the lowest

Table 2 F-measure of entity taggers obtained from individual-trial-corpus.

Annotator	Subject	Temporal	Location
A	0.73	0.78	0.62
B	0.66	0.78	0.65
C	0.76	0.74	0.69
D	0.78	0.84	0.7
E	0.79	0.67	0.75
Consolidated-trial-corpus	0.53	0.68	0.64

performance on *Location* among four annotators (A, B, C, D), possibly indicating the lower quality of annotations and that it was the hardest task among all three classes. Whereas, for the annotations created by person E, the learning algorithm performed badly for *Temporal*, possibly indicating person E had more inconsistency at annotating *Temporal*. Comparing across different annotators for each entity class, we noticed that most annotators produced fairly good annotations for *Temporal* but person E, of whom the result in *F-measure* was even lower than that obtained from the *consolidated-trial-corpus*; and similar exception of person B was noted for *Subject*, and person A for *Location*. We believe that the results so far have revealed several conclusions that are useful for document annotation. Firstly, inter-annotator discrepancy has a major impact on the quality of corpus and therefore, machine learning accuracy. High level of discrepancy damages the quality of annotations, and decreases obtainable machine learning accuracy on a corpus. On the other hand, given uniform settings for a learning algorithm, different accuracies obtained from similar corpora may indicate different levels of quality of the corpora; secondly, annotators may have different skill levels for annotating different classes of entities, possibly due to the difference in the focus of their knowledge. This has caused varying levels of inconsistencies in an annotator's annotations, depending on the specific entity class. Therefore, there is the need for considering annotator's suitability for a task and isolating inconsistent annotators from a task. In line with the conclusion from Table 1, these results foster the motivation of identifying and selecting most suitable annotators (mutually consistent) for each entity-class annotation task.

Annotator Selection Using these analyses, we split the document annotation task by entity-class and select *best-fit-annotators* for specific *best-fit-class* annotations. For each class of entity, we selected three most consistent annotators based on the experiment results. However, depending on the availability of annotators, the

workload and inter-annotator consistency analysis, one can select more or fewer annotators if needed. In the simplistic form, we can select annotators with the highest average agreement in *F-measure* for each entity type. To do so, we simply add up the scores for each row in Table 1 (excluding him/herself) and divide the total by four, results of which are shown in Table 3. However, as concluded from Table 2, certain annotators had high levels of in-consistency in annotating a particular class of entity as indicated by the machine learning accuracy (*F-measure*) tested on their annotations, possibly due to gaps in their knowledge. Therefore, we believe it is important to exclude these annotators and their contributions to the calculation of inter-annotator agreement. As a result, for each class of entity, we eliminated those annotations on which the learner obtained the lowest *F-measure*, particularly those below that from the *consolidated-trial-corpus*. This caused person A eliminated from *Location*, person B eliminated from *Subject* and person E eliminated from *Temporal*. Re-calculating the average agreement using figures in Table 1, we obtained "revised" scores, as indicated in the columns of "Revised" in Table 3.

Table 3 Average agreement in F-measure for each entity type.

Annotator	Subject	*Revised Subject*	Temporal	*Revised Temporal*	Location	*Revised Location*
A	0.61	*0.63*	0.79	*0.8*	0.73	-
B	0.52	-	0.78	*0.76*	0.76	*0.74*
C	0.55	*0.56*	0.73	*0.74*	0.7	*0.7*
D	0.54	*0.55*	0.78	*0.78*	0.73	*0.71*
E	0.53	*0.54*	0.77	-	0.7	*0.713*

With these figures, we simply selected three annotators that have the highest scores, that is, persons A, C, D for annotating *Subject*; persons A, D, B for annotating *Temporal* and persons B, E, D for annotating *Location*, as shown in Table 4.

Table 4 Selected annotators and annotation task.

Annotator	Subject	Temporal	Location
A	O	O	
B		O	O
C	O		
D	O	O	O
E			O

3.4 The Third Phase – Final Corpus Annotation

The annotation exercise continued next by selecting the final corpus of 25 full-length documents from the AHDS archive, and giving them to the selected annotators for

annotation. However, unlike in the first phase, no duplicate documents were given to different annotators, and annotators were only required to annotate entities that they were chosen for, as indicated by the "O" in Table 4. For each class of entities, the documents were split into equal portions among different *best-fit-annotators*. For example, the 25 documents are split into three sets and each set was given to an annotator (A, C, or D) for annotating *Subject* entities. In the end, all annotations were merged into a single collection of 25 documents. This is based on the assumption that mutually consistent annotators will continue annotating consistently for the same annotation problem and the same type of corpus even without the process of consolidation and discrepancy resolution. Therefore, we can distribute the workload among different but consistent annotators for a particular entity-class annotation task, expecting equal level of consistency in the annotations they jointly create. The annotation activity was performed in a series of intensive workshops, during which annotators were free to raise questions and discuss about discrepancies. However, generally speaking, this kind of process is much more cost-saving and workload for each annotator is much lighter than if done in the traditional way as in phase one.

Cost of the Process and Quality of Annotation The annotation process of phase two took roughly 10 - 15 person-days of work, although in practice it was spread across a couple of weeks to minimise fatigue to ensure annotators have the highest level of concentration during the work. The final annotated corpus (*final-corpus*) was also used for a 5-fold cross validation experiment. The final experiment results in *F-measure* are shown in Table 5.

Table 5 Results on the *final-corpus* in *F-measure*.

	SUB	TEM	LOC
Final-corpus	*0.68*	*0.83*	*0.71*
Consolidated-trial-corpus	0.53	0.68	0.64
Best result on individual-trial-corpus	0.79	0.84	0.75

As shown in Table 5, compared against results obtained on the *consolidated-trial-corpus*, the machine learning algorithm produced much better results on the *final-corpus*, which can be attributed to fewer discrepancies and therefore high quality of the annotations. Compared against the best results obtained on the *individual-trial-corpora*, which we consider the top ceiling performance under zero inter-annotator discrepancy, the machine learning system achieved very good results. The relatively smaller improvement on *Subject* is believed due to the heterogeneity of information encompassed by the entity class, which would have increased the difficulty of reaching agreement, as indicated by the inter-annotator agreement studies before. We believe these results are strong evidence supporting the applicability and technical soundness of our methods for annotator selection and task assignment in document annotation and yet producing high quality annotations in a much more effective and efficient way.

4 Discussion and Conclusion

This paper addresses manual document annotation in knowledge acquisition. Document annotations are crucial resources for knowledge acquisition and capture applications. However, creating high-quality annotations is a difficult task due to the inter-annotator discrepancies caused by differences in annotators' knowledge and experiences. Consequently, the process of document annotation typically requires significant amount of effort and time from multiple domain experts to work iteratively and collaboratively to identify and resolve discrepancies. The process is often expensive and time-consuming, preventing its application to practical scenarios.

To address this issue, this paper has proposed an effective and efficient alternative approach to document annotation based on the idea of identification of annotator suitability, task specialisation and annotator selection for specific annotation tasks. Illustrated using a typical named entity annotation scenario, the method starts by sampling the annotator discrepancy problem using the traditional document annotation process on a small corpus; the annotations are then used to evaluate machine learning accuracy to gain an insight to the annotator discrepancies in the task. Results of these experiments show that even with reasonable effort following the traditional annotation approach, high-level discrepancy may still remain, and can lead to low machine learning accuracy. Further analysis reveals that annotators may have different skill levels for annotating different classes of entities, suggesting the need for considering annotators' suitability in specialised annotation tasks. Using this information, the annotation task is split according to entity-classes and sub-tasks are treated differently where the most suitable candidates are chosen for specific annotation tasks. Essentially, matching *best-fit-annotators* to *best-fit-classes* allows distribution of workload, which reduces workload per annotator, but increases the potential amount of annotations that can be produced whilst retaining high quality of annotations. Shown by the experiments, the approach produced a final annotated corpus of five times of the size of the corpus created using the traditional approach (phase one). The machine learning accuracy obtained on these annotations is far better than that obtained from the annotations created in the traditional way, and is very close to the best result obtained under zero inter-annotator discrepancy in the *intra-annotator-learning* experiments.

In terms of the time required for this annotation process, the method has significantly shortened the process required in the traditional document annotation approach. The first phase of the experiment that follows the traditional approach was estimated to cost 14 person-days to annotate 5 documents; whereas, the last phase of the experiment that follows our method was estimated to cost only 10-15 person-days to annotate 25 documents. In total, the annotation exercise undertook less than 1 person month, yet produced high quality annotations for machine learning purposes.

Although applied to the named entity annotation problem, the method can be generalised and applied to other document annotation tasks. Essentially, the key is to sample the discrepancy issue based on which the task can be specialised, annotators' suitability can be evaluated and annotators selected. For example, in document classification, the problem may be analysed based on the topics of documents (e.g., science, entertainment) since some annotators maybe more suitable for dealing with certain kinds of topics than others, especially when they have different academic

backgrounds; in WSD, the analysis may be performed from the angle of word classes, or contexts (e.g., different documents) in which words appear; and likewise the different types of events in event recognition.

However, several inadequacies can be further investigated in future research. Firstly, intra-annotator agreement has been isolated from this study due to unavailability of resources. Studying intra-annotator agreement will reveal valuable details of annotators' skills in an annotation task, and evidence should be combined with results from *intra-annotator-learning* to make stronger support for annotator selection. Secondly, our annotator selection criteria can be improved. Ideally, figures from the experiments should be combined in a mathematical formula to transform the numbers into an appropriate measure of the annotator suitability. Lastly, the method proposed splits an annotation task from the angle of entity-class and base the studies of annotator suitability and selection on this type of task specialisation. On the other hand, an annotation task could also be specialised according to the characteristics of documents, such as structured and un-structured documents. Although these characteristics were not evident in our testing corpora, it may prove useful in other scenarios. In the future, our work will concentrate on these areas.

Acknowledgement This work was funded by the Archaeotools project that is carried out by Archaeology Data Service, University of York, UK and the Organisations, Information and Knowledge Group (OAK) of the Department of Computer Science, University of Sheffield, UK. This work was further supported by Knowledge Now Limited, Sheffield, UK.

References

1. Bermingham, A., Smeaton, A.: A Study of Inter-Annotator Agreement for Opinion Retrieval. In Proceedings of SIGIR'09 (2009)
2. Brants, T.: Inter-annotator agreement for a German newspaper corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation, LREC (2000).
3. Byrne, K.: Nested Named Entity Recognition in Historical Archive Text. In Proceedings of International Conference on Semantic Computing, (2007).
4. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. In Computational Linguistics, 22:2 (1996) 249-254
5. Colosimo, M., Morgan, A., Yeh, A., Colombe, J., Hirschman L.: Data preparation and internannotator agreement: BioCreAtIvE Task 1B. In BMC Bioinformatics (2005)
6. Ciravegna, F., Lavelli, A., Satta, G.: Bringing information extraction out of the labs: the Pinocchio Environment. In Proceedings of the 14th European Conference on Artificial Intelligence (2000)
7. Cucchiarini, C., Strik, H.: Automatic transcription agreement: An overview (2003) 347 – 350.
8. Ehrmann, M.: Les entites nommees, de la linguistique au TAL: statut theorique et methods de desambiguisation. Ph.D. thesis, Univ. Paris (2008).
9. Ferro, L., Mani, I., Sundheim, B., Wilson, G.: TIDES Temporal Annotation Guidelines. Draft Version 1.0. MITRE Technical Report MTR 00W0000094, October (2000).
10. Fort, K., Ehrmann, M., Nazarenko, A.: Towards a methodology for named entities anntoation. In Proceedings of the Third Linguistic Annotation Workshop, ACL-IJNLP (2009) 142-145

11. Grishman, R., Sundheim, B.: Message understanding conference - 6: A brief history. In Proceedings of International Conference on Computational Linguistics. (1996)
12. Gut, U., Bayerl, P. S.: Measuring the Reliability of Manual Annotations of Speech Corpora. Proceedings of Speech Prosody (2004), Nara, 565-568
13. Hripcsak, G., Rothschild, A.: Agreement, the F-measure and Reliability in Information Retrieval: In Journal of the American Medical Informatics Association, 296-298. (2005)
14. Hripcsak, G., Wilcox, A.: Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc (2002) 1–15.
15. Iria, J.: Automating Knowledge Capture in the Aerospace Domain. In Proceedings of the fifth international conference on Knowledge capture. 97-104. (2009)
16. Jeffrey, S., Richards, J., Ciravegna, F., Chapman, S., Zhang, Z.: The Archaeotools project: Faceted Classification and Natural Language Processing in an Archaeological Context, In Special Theme Issues of the Philosophical Transactions of the Royal Society A,"Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructures" (2009)
17. Kim, J., Ohta, T., Tsujii, J.: Corpus annotations for mining biomedical events from literature. In BMC Bioinformatics (2008)
18. Linguistic Data Consortium (2008). Automatic Content Extraction (ACE). URL <http://projects.ldc.upenn.edu/ace/>.
19. Minkov, E., Wang, R., Cohen, W.: Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. In Proceedings of HLT/EMNLP-05 (2005)
20. Morante, R., Asch, V., Daelemans, W.: A memory–based learning approach to event extraction in biomedical texts. Proceedings of the Workshop on BioNLP: Shared Task (2009) 59-67
21. Murphy, T., McIntosh, T., Curran, J.: Named entity recognition for astronomy literature. Australian Language Technology Workshop (2006)
22. Nadeau, D.: PhD Thesis: Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision (2007).
23. Ng, H., Lim, C., Foo, S.: A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. In Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources SIGLEX99 (1999), 9-13
24. Ohta, T., Tateisi, Y., Kim, J.: The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In Proceedings of the second international conference on Human Language Technology Research (2002) 82-86
25. Olsson, F.: PhD thesis: Bootstrapping Named Entity Annotation by Means of Active Machine Learning: A Method for Creating Corpora (2008)
26. Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., Salakoski, T.: BioInfer: a corpus for information extraction in the biomedical domain. In BMC Bioinformatics (2007)
27. Saracevic T.: Individual differences in organizing, searching, and retrieving information. In Proceedings of the 54th Annual ASIS Meeting (1991), 82-86
28. Tanabe, L., Xie, N., Thom, L., Matten, W., Wilbur, W.: GENETAG: a tagged corpus for gene/protein named entity recognition. In BMC Bioinformatics (2005)
29. Wilbur, W., Rzhetsky, A., Shatkay, H.: New directions in biomedical text annotation: definitions, guidelines and corpus construction. In Bioinformatics (2006)