Tell Me What You Read: Automatic Expertise-Based Annotator Assignment for Text Annotation in Expert Domains

Hiyori Yoshikawa, [1] **Tomoya Iwakura,** [1] **Kimi Kaneko,** [2*] **Hiroaki Yoshida,** [1]
Yasutaka Kumano, [3] **Kazutaka Shimada,** [4] **Rafal Rzepka,** [5] **Patrycja Swieczkowska** [5]
[1] Fujitsu Limited, Tokyo, Japan [2] Toyota Motor Corporation, Aichi, Japan
[3] G-Search Limited, Tokyo, Japan [4] Kyushu Institute of Technology, Fukuoka, Japan
[5] Hokkaido University, Hokkaido, Japan
[1,3] {y.hiyori,iwakura.tomoya,yoshida.hiro-15,kumano.yasutaka}@fujitsu.com
[2] kimi_kaneko@mail.toyota.co.jp [4] shimada@ai.kyutech.ac.jp
[5] {rzepka,swieczkowska}@ist.hokudai.ac.jp

Abstract

This paper investigates the effectiveness of automatic annotator assignment for text annotation in expert domains. In the task of creating high-quality annotated corpora, expert domains often cover multiple sub-domains (e.g. organic and inorganic chemistry in the chemistry domain) either explicitly or implicitly. Therefore, it is crucial to assign annotators to documents relevant with their fine-grained domain expertise. However, most of existing methods for crowdsoucing estimate reliability of each annotator or annotated instance only *after* the annotation process. To address the issue, we propose a method to estimate the domain expertise of each annotator *before* the annotation process using information easily available from the annotators beforehand. We propose two measures to estimate the annotator expertise: an explicit measure using the predefined categories of sub-domains, and an implicit measure using distributed representations of the documents. The experimental results on chemical name annotation tasks show that the annotation accuracy improves when both explicit and implicit measures for annotator assignment are combined.

1 Introduction

Preparation of training data has been a critical issue in applying the supervised and semi-supervised methods to real world problems. Training data construction is often quite costly and the quality is hard to assure, especially when the task requires expert knowledge of the target domain such as chemistry and medicine. A possible solution to alleviate the lack of annotated data is the use of a crowdsourcing platform. Crowdsourced annotation has proven to be successful for lowering the annotation costs in tasks that require general knowledge, such as POS tagging (Hovy et al., 2014), textual entailment, word sense disambiguation (Snow et al., 2008), or recognition of general named entities like PERSON and ORGANIZATION (Finin et al., 2010; Nguyen et al., 2017). Meanwhile, there is limited success in using crowdsourcing for annotation in expert domains mainly because of the poor annotation quality made by annotators with low domain expertise. In the medical domain, for example, Nye et al. (2018) combine both *non-expert* crowdsourced annotators and expert annotators for text annotation.

We hypothesize that a key challenge of crowdsourced annotation in expert domains lies in the difficulty of estimating the candidate annotators' expertise in terms of the relevance to the documents to be annotated. Some crowdsourcing platforms such as Amazon Mechanical Turk [1] provide the feature to specify workers that have the expertise in particular domains. However, it is still not optimal for annotation tasks of documents such as scientific papers, as they tend to require expertise in various specific sub-domains. For example, inorganic chemistry and drug discovery fields belong to a broader category of the chemical domain. Annotation on drug discovery papers would be difficult for experts in inorganic chemistry because it substantially differs from the drug discovery field. However, limiting the candidate annotators to the experts in drug discovery would result in insufficient number of annotators that is not enough for obtaining a large size annotated corpus within limited time. Moreover, it is even hard to identify the required domain knowledge to understand each document with such high granularity.

A practical solution to this issue would be to find a way to estimate the relevance between an annotator and a document and to assign annotators to documents based on the relevance between

[*] Work done while the author was at Fujitsu Laboratories Ltd.

[1] https://www.mturk.com/

Proceedings of Recent Advances in Natural Language Processing, pages 1575–1585
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_177

them. It is expected that even if we can not recruit a sufficient number of annotators for each particular sub-domain, we can assign annotators with expertise in different but relevant sub-domains and obtain better quality annotations.

This motivated us to investigate the annotator assignment problem for text annotation in expert domains. The main question is whether we can improve the overall text annotation quality by using not only explicit information such as the annotators' major fields of study but also implicit information estimated by other kinds of information available before the annotation process. To answer the question, we conduct a series of experiments on chemical name annotation tasks in academic paper abstracts from various sub-fields of chemistry. The annotators are graduate students in chemistry departments, so they have expertise in different chemistry sub-domains. As the *explicit* knowledge of the annotators' expertise, we asked the annotators their fields of expertise from predefined categories. We also asked them to submit information about the literature strongly related to their study. We use the literature information to estimate more fine-grained *implicit* relevance between the annotated document and the annotators that can not be captured by the explicit information. The experimental results indicate that combining both the explicit and implicit information helps estimate the document-annotator relevance and results in better annotation quality.

Our approach is orthogonal to the annotation aggregation methods (Nguyen et al., 2017; Li et al., 2014) commonly used in the crowdsourced annotation. While the annotation aggregation methods improve the annotation quality by aggregating the annotations from different annotators *after* the annotation process, our approach improve the annotation by assigning more relevant domain experts to each text *before* the annotation process. The difference comes from the fact that most of the studies on crowdsouced annotation focus on annotation tasks that require general knowledge or a single expert domain, where the assignment of annotators has relatively small effect on the final annotation quality. Thus, it is possible that our approach and the annotation aggregation methods complement each other. Our approach is also different from the existing methods to select the best annotators which requires training based on the previous annotation results by the candidate annotators (Donmez et al., 2010; Kamar et al., 2013; Kamar and Horvitz, 2015; Tran-Thanh et al., 2014). We believe that estimation of the relevance between documents and annotators can also be helpful for such methods as prior knowledge.

2 Related Work

Preserving the quality of the crowdsourced work is difficult as annotators possess very diverse abilities, skills, or interests (Daniel et al., 2018). For this reason, many researchers have proposed their methods for improving the quality of annotations or maintaining it without increasing the cost. Some strategies deal with the problem of the ground truth absence – from increasing the number of annotators (Sheng et al., 2008), through intelligently weighting them based on their inferred expertise (Donmez et al., 2009; Raykar et al., 2010; Welinder et al., 2010) to recruiting a small group of top quality workers (Zhao et al., 2013; Li et al., 2014; Li and Liu, 2015; Carvalho et al., 2016) or seeing how they learn over time (Pan et al., 2016). When it comes to methods used for selecting the best annotators, Donmez et al. (2010) utilize a sequential Bayesian estimation algorithm for continuous tracking and selecting the best annotators over time. Markov Decision Process can be used to model agent conduct during the consensus tasks and to predict a candidate annotator's work (Kamar et al., 2013; Kamar and Horvitz, 2015). Tran-Thanh et al. (2014) introduce a recruiting algorithm based on a variation of the multi-armed bandit model (MAB) outperforming previous methods by up to a remarkable 300%. Recommendation of tasks to workers has also been studied with use of a worker's task browsing history (Yuen et al., 2015), or taking into account implicit negative feedback (Lin et al., 2014).

However, all these approaches focus on annotators, not the target data. Unlike these methods, our approach is to assign documents to experts in potentially different sub-domains so that their expert domains are as close as possible to those of the assigned documents. An example of annotation improvement in a scenario requiring specialized expertise (clinical NLP) is given in Dumitrache et al. (2015), where authors pay attention to documents being annotated. The main problem they aim to solve is not only the lack of ground truth for training and benchmarking but also ambiguity in the target documents. They have shown that, with proper processing, the crowd performs just as well as medical experts in terms of the quality and efficacy of annotations, while being cheaper and more readily available. However, their results indicate that at least ten workers per sentence are needed to get the highest quality annotations, while we aim at acquiring high-quality results with fewer workers by assuring assignments of targets well-fit to the annotators.

Assignment of experts has been studied in the context of paper-reviewer assignment problem at academic conferences (Dumais and Nielsen, 1992; Charlin and Zeme, 2013; Dumais and Nielsen, 2016). These methods use previous publications of reviewers and those publications are used to build reviewers' profiles. Methods for building profiles include not only use of words in publications, but also Latent Semantic Indexing (Salton and McGill, 1983) and Latent Dirichlet Allocation (Blei et al., 2003). Submitted papers are also encoded with the same methods for building reviewer profiles. Then, the similarity between submitted papers and each reviewer's profile is calculated based on the encoded representations and assignment is done just with the calculated similarity or by casting assignment as an Integer Linear Programming problem (Karimzadehgan and Zhai, 2009). Our work is in the similar spirit as this line of work. The experimental results indicate that such an approach is also promising in selecting annotators for expert domain annotation tasks.

To sum up the related work: current crowdsourcing-based annotation methods integrate several annotation results into one final annotation, utilize machine learning with several annotation results, or target experts where the most skillful annotators are automatically discovered. To the best of the authors' knowledge, the proposed task of automatic assignment of documents for annotation is the first of its kind. Besides, this is the first research that evaluates the effectiveness of consideration of annotator's expertise in quality.

3 Expertise-Based Annotator Assignment

3.1 Problem Description

We consider the annotator assignment problem for text annotation, where the goal is to assign the most relevant annotators to the given document for better annotation quality. In particular, we consider annotation tasks on documents that require special knowledge for full understanding such as scientific papers and clinical records. We also assume that the required sub-domain of knowledge differs depending on the document to be annotated. For the chemistry domain, the sub-domains may include inorganic chemistry, drug discovery, and so on. As clues of annotator's expertise, we consider two types of information: *explicit expertise* such as predefined sub-domains and *implicit expertise* that is estimated from the information available before the annotation process and potentially represents more fine-grained relevance between documents and annotators. The implicit expertise can be helpful not only for complementing explicit expertise, which might not be available or might be insufficient but also for capturing the similarity between different sub-domains to estimate the relevance of the papers to the annotators in different sub-domains. We describe these two types of expertise in more details in the following section.

In this paper, we evaluate our approach on the chemical name annotation task using chemistry paper abstracts. In what follows, we describe our proposed method based on this specific task. However, the method can be generalized to different domains and tasks where similar type of prior information about the target documents and the candidate annotators is available. It is worth noting that our approach depends only on the type of information about the documents and the annotators available before the annotation process, and is independent of the type of the annotation task.

3.2 Prior Information of Annotator Expertise

3.2.1 Explicit Expertise

For explicit expertise, we assume existence of predefined categorical labels representing expert sub-domains. Both the documents to be annotated and the candidate annotators should be associated with categorical labels representing the sub-domains of the documents' or the annotators' expertise. In practice, the labels can be the conference venues of papers, IPC classification of patents, and so on. Multiple labels can be associated with each document and annotator. We use the binary indicator $I_{\text{cat}}(i, j) \in \{0, 1\}$ for the explicit expertise-based relevance score between the annotator i and the document j, which indicates whether the annotator i's areas of expertise include at least one of the categories of the document j or not.

3.2.2 Implicit Expertise

In addition to the explicit expertise measure, we propose document-annotator relevance based on implicit expertise that captures more fine-grained and multi-dimensional aspects of the domain expertise. We assume that we have access to a small subset of documents that each annotator is the most familiar with from the same domain as those to be annotated (hereinafter called *annotators' documents*). This requires the candidate annotators to present additional information, but it is not hard for annotators to present. For example, it can be a few papers relevant to their fields of study.

We encode both the annotators' documents and the documents to be annotated into a low-dimensional vector in some way. These document representations include semantic knowledge about the specific sub-domains of the documents and the annotators that can not be captured by the categorical information alone. Suppose that we have a set of candidate annotators I and a set of documents to be annotated J. Let $d_{il}^{(s)} (l \in \{1, \dots, L\})$ be the representation of the i-th annotator's documents (assuming that each annotator presents L relevant documents), and $d_j^{(t)}$ be the representation of the j-th document to be annotated. For annotators that have expertise in multiple sub-areas, each of the annotators' documents may represent different aspects of their expertise. Therefore, we compare two different ways to aggregate these paper representations and compute the final relevance score. The first one is to compare each document to be annotated with the average of the document representations of the annotator's documents:

$$s_{\mathrm{av}}(i,j) = \mathrm{sim}(\frac{1}{L}\sum_{l=1}^{L} d_{il}^{(\mathrm{s})}, d_j^{(\mathrm{t})}), \qquad (1)$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity. The second one is to compute the similarity between each of the annotator's documents and the document to be annotated, and take the score from the most relevant one:

$$s_{\mathrm{nearest}}(i,j) = \max_{l} \mathrm{sim}(d_{il}^{(\mathrm{s})}, d_j^{(\mathrm{t})}). \qquad (2)$$

3.2.3 Relevance Scores

In our experiments, we evaluate the effect of combining these explicit and implicit relevance measures. We employ the following combinations:

- *cat*: $I_{\mathrm{cat}}(i,j)$,
- *catsim-av*: $I_{\mathrm{cat}}(i,j) + s_{\mathrm{av}}(i,j)$, and
- *catsim-nearest*: $I_{\mathrm{cat}}(i,j) + s_{\mathrm{nearset}}(i,j)$.

Note that as $I_{\mathrm{cat}}(i,j)$ takes the value 0 or 1 and $s_{\mathrm{av}}(i,j)$ and $s_{\mathrm{nearest}}(i,j)$ is between $[0,1]$, the agreement based on the explicit expertise has priority over that of implicit expertise in terms of the final relevance score. Therefore, the implicit expertise-based relevance is expected to work as the auxiliary information that provides the knowledge that can not be captured by the explicit expertise-based measure.

3.3 Calculating Annotator Assignment

The document-annotator assignment is calculated by solving the following Integer Linear Programming (ILP) problem:

$$\begin{aligned} \text{Maximize} \quad & \textstyle\sum_{i\in I, j\in J} s_{i,j} \cdot x_{i,j} & (3)\\ \text{s.t.} \quad & k_{\min} \le \textstyle\sum_{j\in J} x_{i,j} \le k_{\max} \quad \forall i \in I, & (4)\\ & \textstyle\sum_{i\in I} x_{i,j} = n_{\mathrm{ann}} \quad \forall j \in J, & (5)\\ & x_{i,j} \in \{0,1\} \quad \forall i \in I, \forall j \in J, & (6) \end{aligned}$$

where $s_{i,j}$ is one of the relevance scores between the i-th annotator and the j-th document introduced above. The problem is to find the assignment $x_{i,j}$ that maximizes the total relevance scores of the assigned pairs of documents and annotators. $x_{i,j}$ is the indicator variable that becomes 1 if the i-th annotator is assigned to the j-th document, and 0 otherwise. The first constraint represents the maximum $k_{\max}$ and the minimum $k_{\min}$ numbers of documents that can be assigned to a single annotator. In order to evenly assign documents to annotators, we use $k_{\min} = \lfloor |J|/|I| \rfloor$ and $k_{\max} = k_{\min} + 1$ in our experiment. The second constraint means that each document should be assigned to n_{ann} annotators. n_{ann} is set to 1 in our experiment.

4 Experimental Settings

We evaluated the proposed annotator assignment method on a chemical name annotation task using the abstracts of chemistry papers.

4.1 Datasets and Guidelines

We used chemistry papers in two different languages: English and Japanese. The papers are selected randomly from the predefined categories to construct a dataset of approximately 500 abstracts for each language. The sub-domain categorization used for each language is shown in Table 1.

	English		Japanese	
Category	Docs	Anno	Docs	Anno
(1) Physical chemistry	✓	9	✓	2
(2) Analytical chemistry		7	✓	2
(3) Inorganic chemistry	✓*	6	✓	3
(4) Complex chemistry	✓*	1	✓	2
(5) Organic chemistry	✓	14	✓	2
(6) Polymer chemistry	✓	6	✓	3
(7) Medical care	✓	4	✓	3
(8) Drug discovery	✓	4	✓	5
(9) Biochemistry	✓	10		6
(10) Applied chemistry	✓	7		2
(11) Toxicology	✓	4		0
Others		14		3
Total number of annotators		40		20

Table 1: Expertise categories. Ticked ones in the Docs columns indicate categories used for each language. The Anno columns indicate the number of annotators worked on each language that have expertise in the subdomain. The total number of annotators differs from the sum of the annotators in individual categories as we allow declaration of multiple sub-domains. (*) We integrated categories (3) Inorganic chemistry and (4) Complex chemistry of English into a single category because we could not distinguish them in the journal names of target documents.

4.1.1 English Papers

We used CHEMDNER corpus (Krallinger et al., 2015) for the English annotation task. It consists of 10,000 abstracts of chemistry-related publications that were chosen according to journal titles. These titles helped us assign categories to English texts (listed in Table 1; multiple categories for each paper were allowed). Each abstract in the CHEMDNER corpus was annotated with seven types of chemical term classes defined by the CHEMDNER task organizer. To evaluate our annotation, we used paper abstracts from the CHEMDNER test set which consists of 3,000 abstracts. We randomly sampled 504 abstracts while maintaining an approximately even number of papers in every category.

For annotation evaluation, we followed the annotation guidelines of the CHEMDNER corpus[2]. The labels in the CHEMDNER test set were used as the gold standard. The original paper reports an inter-annotator agreement of 85.26%.

4.1.2 Japanese Papers

We obtained Japanese chemistry paper abstracts from JDREAM III[3] repository of scientific publications based on their categories as of September, 2017. JDREAM III provides predefined categories corresponding to the first eight of those in Table 1. We retrieved 2,500 papers for each category, and the total number of papers was 20,000. We applied to them an in-house chemical named entity recognizer to select the paper abstracts that were likely to include some chemical compound names or chemistry-related terms. Finally, we randomly sampled 520 abstracts from the selected abstracts so that the number of abstracts in each category was 65.

For annotation of Japanese texts, we prepared a new annotation guideline defining 20 mention types for chemical entities. We defined twelve entity types for chemical substances including organic and inorganic molecules and eight entity types for chemistry-related concepts such as drug names, products, properties, and numerical expressions. An example of annotated document is shown in Appendix A.

To prepare the gold standard data, we employed 17 in-house expert annotators to annotate chemistry-related terms in these paper abstracts. In order to create a reliable gold standard, each abstract was checked by three annotators. First, two annotators annotated each abstract, and then the remaining annotator, who had not annotated the abstract, checked and integrated the annotations. The inter-annotator agreement between the two annotators was 0.449 in terms of exact match entity-level F_1 score.

4.2 Annotators

We employed 49 graduate students who major in chemistry at their universities. 40 and 20 students performed annotation for English and Japanese, respectively. Some students annotated both English and Japanese texts.

As prior information for estimating the subdomain expertise of each student, we asked the students to provide the following information:

- Expertise category as listed in Table 1. Multiple choice was allowed.
- Titles and abstracts of five scientific papers each that were relevant to their study.

[2] To be precise, we used the annotation guidelines for *CHEMDNER patent* corpus as CHEMDNER server was not accessible during our experiment. However, the guidelines of the patent corpus defined the same seven mention types and in our preliminary evaluation we confirmed that the differences between them are small.

[3] https://jdream3.com/

The distribution of their expertise categories is shown in Table 1.

All annotations were conducted using the BRAT annotation tool (Stenetorp et al., 2012). For both English and Japanese papers, we provided the annotators with the same annotation guidelines that had been used to construct the gold-standard dataset (i.e. the CHEMDNER patent guideline for English and our new guideline for Japanese). In order to alleviate the problem of lower quality caused by inexperience in annotation and to assure proper understanding of the annotation guidelines, we first trained all the annotators by asking them to annotate the same five paper abstracts in a trial stage. After that, we asked each annotator to annotate the assigned abstracts.

In addition to *cat*, *catsim-av* and *catsim-nearest*, we also evaluated the performance of *non-expert*s who have no expertise in chemistry. The *non-expert*s' fields of expertise were other than chemistry, such as architectonics, physics and electricity. Abstracts assignment to the *non-expert*s was random.

4.3 Annotator Assignment

As the measure of the implicit expertise, we calculated semantic representations of submitted paper abstracts using a simple sentence embedding method. For each paper, we averaged the sum of embeddings of all content words [4] in the title and abstract of the paper after normalizing the Euclidean norm of each word vector to 1. We used 200-dimensional word embeddings trained with `word2vec` (Mikolov et al., 2013) for both English and Japanese. Training data for the English word embeddings was the whole CHEMDNER corpus in addition to the titles and abstracts of papers obtained from MEDLINE 2017 version[5] that appear in the same set of journals used in the CHEMDNER corpus. Training data for the Japanese word embeddings were the titles and abstracts of 20,000 papers from JDREAM III.

The ILP problem for the assignment was solved with the COIN CBC solver of PuLP[6] and the optimum solutions were obtained for all the assignments. The optimization result was obtained within a few seconds. Each annotator was asked to annotate the combined set of abstracts assigned by the three assignment methods. The order of the abstracts given to each annotator was random and the annotators were not informed which method was used to assign each abstract. As some of the annotation jobs were canceled after the assignment, we ended up doing the evaluation with 323 paper abstracts for English and 375 paper abstracts for Japanese for which we obtained results from all the assignment methods.

5 Results

5.1 Main Results

Table 2 shows the annotation performance with different assignment methods. We use the standard evaluation metrics for named entity tagging, i.e. recall, precision and F-measure (F_1) based on the exact match of the tagged entities compared with the gold standard.

The performance by *non-expert*s is significantly worse than those of domain experts, indicating the importance of domain expertise for our evaluation task. On the other hand, the performance of expertise-based assignments is higher than the agreement between experts (0.449) for the Japanese documents. [7] Compared to (*cat*), which only uses the explicit expertise for assignment, *catsim-nearest* showed higher F_1 for both English and Japanese data set. *catsim-av* demonstrated higher F_1 for English but lower F_1 for annotations in Japanese. A possible reason why *catsim-nearest* measure generally performs better than *catsim-av* is that the former is better at capturing expertise over multiple sub-domains. When an annotator has expertise in multiple sub-domains, *catsim-av* represents their expertise with a single vector by averaging the representations of all the annotator's documents. On the other hand, *catsim-nearest* keeps the individual representations of all the annotator's documents and uses the one that is the most relevant to each target document for the relevance score. The improvement is relatively higher in recall than in precision for both English and Japanese, indicating that sub-domain expertise helps annotators in detecting more technical terms in the documents.

[4] We used NLTK (Bird, 2006) and Mecab (Kudo et al., 2004) tokenizers for English and Japanese, respectively, to identify content word using part-of-speech tags.

[5] https://www.nlm.nih.gov/bsd/medline.html

[6] https://github.com/coin-or/pulp

[7] Although the reported inter-annotator agreement of the English corpus is 85.26%, it is hard to directly compare the result with ours as no details are provided for the calculation of the agreement. In addition, the annotators of the corpus were also involved in revising the annotation guideline, indicating higher proficiency in the guideline.

	English			Japanese		
	R	P	F_1	R	P	F_1
non-expert	0.313	0.495	0.384	0.434	0.420	0.427
cat	0.525	0.481	0.502	0.455	**0.510**	0.481
catsim-av	0.550	0.496	0.522	0.458	0.502	0.479
catsim-nearest	**0.569**	**0.507**	**0.536**	**0.465**	**0.510**	**0.486**

Table 2: The annotation quality for each assignment measured with recall (R), precision (P) and F_1 scores.

To compare the results of annotations, we employed a McNemar paired test on the labeling disagreements following the procedure introduced in Sha and Pereira (2003). As in Kudo et al. (2004) for morphological analysis, we compared the results based on character-based IOB2 format (Tjong Kim Sang and Veenstra, 1999) instead of the usual word-based version, because the endings or beginnings of chemical terms do not always correspond to word boundaries. The results confirmed statistical significance ($p < 0.01$) of the differences between *cat* and *catsim-nearest* for both English and Japanese.

5.2 Error Analysis

Table 3 shows the statistics of the annotation results with respect to the types of errors. We classified the annotation errors into four types: incorrect tags assigned to characters other than chemical terms (Ex), chemical terms which were not annotated (Miss), annotations whose spans are different from the correct annotation (SD), and annotations where spans are correct, but the type of the chemical terms is incorrect (TD). An example of each error type is shown on the right-hand side of the table. The results show that assigning more relevant domain experts to the task helps to reduce the number of missed annotations while slightly increasing the number of excessive annotations. It is understandable though that annotators are likely to give more labels to documents relevant to their expertise.

We also evaluated proposed methods by mention-wise comparison of annotation results. The results show that the annotations corresponding to implicit expertise-based assignments are better than explicit expertise-based ones in terms of the number of correctly annotated mentions. McNemar paired test showed statistical significance ($p < 0.05$) of the improvement in English results, while no significance was observed in Japanese results. We also computed the correlation between the relevance scores and the annotation accuracy as shown in Table 5. Similarly to the mention-wise comparison, we observed significant correlation between implicit expertise-based relevance scores and the corresponding F_1 scores for the English task, while no significant correlation was observed for the Japanese task. Possible reasons why the proposed method is less efficient in the Japanese task is discussed in the following section.

6 Discussion and Conclusion

Good quality corpora of specialized scientific texts could become important not only for specialists from various fields, but also for the promising AI subfield of automatic scientific discovery. This paper proposes a novel method for automatic annotation task assignment based on the expertise of the annotator estimated with scientific paper abstracts that the annotator has presented as relevant to their own research interests. The experimental annotation results of English and Japanese chemistry-related paper abstracts showed that our method contributes to higher accuracy compared to the annotation by *non-expert*s and use of predefined technical field categories.

We could not cover all possible experimental settings (e.g. assignment using only implicit expertise) due to the time and budget restrictions. In our experimental design it is required that all assignments to be compared are calculated on the same set of annotators and then annotated together, which makes adding other experimental settings afterwards difficult. Improvement in experimental design is needed in order to compare different factors more flexibly.

An interesting observation is that the cause of annotation errors is not always the lack of annotator expertise. For example, we found that errors which result from poor understanding of annotation guidelines are not negligible. For the Japanese task,

		# Corr	# Ex	# Miss	# SD	# TD
English	*non-expert*	860	336	1316	280	306
	cat	1441	787	519	395	430
	catsim-av	1511	798	476	325	459
	catsim-nearest	1563	806	467	409	357
Japanese	*non-expert*	3484	1940	1652	2080	1174
	cat	3655	1165	1912	1717	961
	catsim-av	3682	1229	1818	1746	1016
	catsim-nearest	3734	1211	1789	1649	1062

	Gold	Annotation
Ex	solvent	⟨C⟩solvent⟨/C⟩
Miss	⟨C⟩acetylene⟨/C⟩	acetylene
SD	⟨C⟩tert-butyl⟨/C⟩	tert-⟨C⟩butyl⟨/C⟩ tert-⟨G⟩butyl⟨/G⟩
TD	⟨C⟩acetylene⟨/C⟩	⟨G⟩acetylene⟨/G⟩

Table 3: (Left) The number of correct annotations (Corr), incorrect tags assigned to characters other than chemical terms (Ex), chemical terms which were not annotated (Miss), annotations whose spans are different from the correct annotation (SD), and annotations where spans are correct, but the type of the chemical terms is incorrect (TD). (Right) Examples of errors for each error type.

		(English)							(Japanese)				
		catsim-av			*catsim-nearest*				*catsim-av*			*catsim-nearest*	
		T	F		T	F			T	F		T	F
cat	T	1029	412	T	1060	381	*cat*	T	2777	878	T	2736	919
	F	482	824	F	503	803		F	905	3474	F	998	3381

Table 4: Correlation matrices comparing two of three annotation results by mentions. The upper-left cell (T–T) of each result indicates the number of mentions which were correctly annotated by both methods, the upper-right cell (T–F) corresponds to the mentions which were correctly annotated by *cat* but were not by the other method, and so on.

	Corr. (En)	Corr. (Ja)
cat	0.02	0.00
catsim-av	0.12*	0.09
catsim-nearest	0.14*	-0.01

Table 5: Spearman rank correlation coefficients between task assignment scores and F_1 scores on the corresponding annotation results. (En) and (Ja) correspond to experiments on English and Japanese abstracts, respectively. $*$ means that the relationship is statistically significant at $p < 0.05$.

in which we failed to find any significant improvement, as many as 20 of mention types might have made the annotation task too complicated for annotators that are domain experts but not experts in linguistic annotation. As shown in Table 3, a large number of annotation erros in the Japanese task are span difference (SD). This type of errors is often less relevant to domain expertise: an example is ”Sb(CN)3(2,2’-bipy)” vs. ”[Sb(CN)3(2,2’-bipy)]”.

Another future work topic is designing an annotation framework for domains requiring expertise knowledge. For experiments described in this paper, we recruited students who major in chemistry with the help of university faculties. However, in order to continue building corpora for domains requiring expertise knowledge, this procedure is not optimal. In fact, it took about three weeks just to hire annotators for the experiments. In the future, it is necessary to develop not only accurate assignment methods, but also an annotation framework for expert domains. In order to apply our method to documents other than scientific papers, it is also necessary to find alternatives for the “relevant scientific papers” that represent the annotators’ domain expertise. The application fields include Q&A texts and blogs that feature domain specific topics. It is reported that annotation accuracy on such texts gets worse when the task requires specific knowledge of cartoons and TV programs and so on (Komiya et al., 2016). For such cases, it might be helpful to use cartoon titles, TV-show titles, news articles, blog posts, etc., recently read or watched by annotators.

Acknowledgments

A part of this work was carried out in the RIKEN Center for Advanced Intelligence Project.

References

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006*

Interactive Presentation Sessions, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Arthur Carvalho, Stanko Dimitrov, and Kate Larson. 2016. How many crowdsourced workers should a requester hire? *Annals of Mathematics and Artificial Intelligence*, 78(1):45–72.

Laurent Charlin and Richard Zeme. 2013. The Toronto paper matching system: An automated paper-reviewer assignment system. In *Proc. of Workshop on Peer Reviewing and Publishing Models*.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):7.

Pinar Donmez, Jaime Carbonell, and Jeff Schneider. 2010. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 826–837. SIAM.

Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268. ACM.

Susan T. Dumais and Jakob Nielsen. 1992. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 233–244.

Susan T. Dumais and Jakob Nielsen. 2016. The new automated IEEE INFOCOM review assignment system. In *IEEE Network 30, 5*, pages 18–24.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015. Achieving expert-level annotation quality with crowdtruth. In *Proc. of BDM2I Workshop, ISWC*.

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, Los Angeles. Association for Computational Linguistics.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382, Baltimore, Maryland. Association for Computational Linguistics.

Ece Kamar and Eric Horvitz. 2015. Planning for crowdsourcing hierarchical tasks. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1191–1199. International Foundation for Autonomous Agents and Multiagent Systems.

Ece Kamar, Ashish Kapoor, and Eric Horvitz. 2013. Lifelong learning for acquiring the wisdom of the crowd. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Maryam Karimzadehgan and ChengXiang Zhai. 2009. Constrained multi-aspect expertise matching for committee review assignment. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1697–1700.

Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki, and Hiroyuki Shinnou. 2016. Comparison of annotating methods for named entity corpora. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 59–67, Berlin, Germany. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):S2.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Hongwei Li and Qiang Liu. 2015. Cheaper and better: Selecting good workers for crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Hongwei Li, Bo Zhao, and Ariel Fuxman. 2014. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proc. of WWW'14*, pages 165–176.

Christopher H. Lin, Ece Kamar, and Eric Horvitz. 2014. Signals in the silence: Models of implicit feedback in a recommendation system for crowdsourcing. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 908–914.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 299–309, Vancouver, Canada. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Shengying Pan, Kate Larson, Josh Bradshaw, and Edith Law. 2016. Dynamic task allocation algorithm for hiring workers that learn. In *IJCAI*, pages 3825–3831.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

Gerard Salton and Michael J. McGill. 1983. *Comment Introduction to modern information retrieval*. McGraw-Hill.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 213–220.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179, Bergen, Norway. Association for Computational Linguistics.

Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 214:89–111.

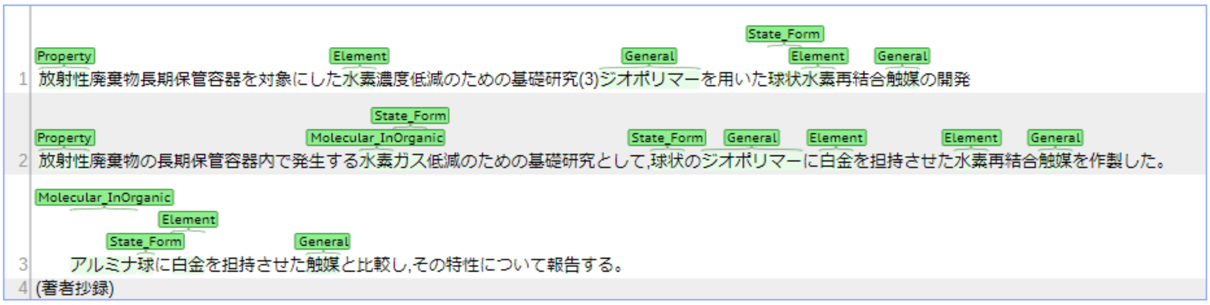
Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432.

Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2015. Taskrec: A task recommendation framework in crowdsourcing systems. *Neural Processing Letters*, 41(2):223–238.

Zhou Zhao, Da Yan, Wilfred Ng, and Shi Gao. 2013. A transfer learning based framework of crowdselection on twitter. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1514–1517. ACM.

A Annotation of Japanese Documents

Figure 1 shows an example of an annotated part of a Japanese paper.



放射性廃棄物長期保管容器を対象にした水素濃度低減のための基礎研究(3)ジオポリマーを用いた球状水素再結合触媒の開発

放射性廃棄物の長期保管容器内で発生する水素ガス低減のための基礎研究として,球状のジオポリマーに白金を担持させた水素再結合触媒を作製した。

アルミナ球に白金を担持させた触媒と比較し,その特性について報告する。

(著者抄録)

Figure 1: An example of annotated Japanese text.