Active learning with applications in biomedical document annotation

Han, Xu

2017

Han, X. (2017). Active learning with applications in biomedical document annotation.
Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/71680>

<https://doi.org/10.32657/10356/71680>

ACTIVE LEARNING WITH APPLICATIONS IN BIOMEDICAL DOCUMENT ANNOTATION

HAN XU

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological
University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2017

Abstract

The rapidly increasing volume of published biomedical research literature is challenging individual biomedical researchers to keep up to date with all the latest development in their own research fields. With the advancement of natural language processing and Semantic Web technologies, more and more sophisticated biomedical natural language processing systems are becoming available to lessen the burden of information overload individual biomedical researchers face. However, the building and evaluation of such biomedical text mining systems need a critical requirement of the manually annotated corpora; yet the construction process of the corpora is time-consuming and expensive as it requires much yet tedious effort from human annotators.

Active learning is an approach to resolve this issue and aid annotators to reduce the time and effort needed for the corpus annotation process. In the widely used passive learning method, the documents are randomly and independently selected from the underlying distribution, while in active learning method, a selection module is allowed to, repeatedly, query un-annotated documents in order to single out the most informative document to be manually annotated and to update its learned rules to achieve the overall maximized efficiency.

In this study, we first propose a document scoring based active learning method for ontological event extraction. Our method can significantly reduce the amount of annotated corpora to saturate event extraction performance, compared to random selection of corpora for annotation, which is the common practice, and previous active learning methods for corpus selection. We evaluated the performance of all the active learning methods using the TEES event extraction system against the BioNLP Shared Tasks datasets, showing that our method can help the system achieve its previously reported performance only with 60%-70% of the original training data.

We then propose a committee-based active learning method for the event extraction and named entity recognition. The method is based on two systems as follows: We first employ an event extraction system to filter potential false negatives among unlabeled documents, where the system does not extract any event. We then adopt a statistical method to rank the potential false negatives of unlabeled documents 1) by using a language model that measures the probabilities of the expression of multiple events in documents and 2) by using a named entity recognition system that locates the named entities that can be event arguments (e.g., proteins). The proposed method further deals with unknown words in test data by using word similarity measures. We also apply our active learning method for the task of named entity recognition. We evaluate the proposed method against the BioNLP Shared Tasks datasets, and show that our method can achieve better performance than such previous methods as entropy and Gibbs error based methods and a conventional committee-based method. We also show that the incorporation of named entity recognition into the active learning for event extraction and the unknown word handling further improves the active learning method. In addition, the adaptation of the active learning method into named entity recognition tasks also improves the document selection for manual annotation of named entities.

Finally, we propose a novel clustering based active learning method for the biomedical NER task. We show that the underlying NER system using the proposed method outperforms those with other state of the art active learning methods, including density, Gibbs error and entropy based approaches, as well as the random selection. We compare variations of our proposed method and find the optimal design of the active learning method, which is to use the vector representation of named entities, and to select documents that are ‘representative’ and ‘informative’, as well as to use the Shared Nearest Neighbor (SNN) clustering approach. In particular, the optimal variant of the proposed method achieves a deficiency gain of 36.3% over random selection.

The proposed active learning method is a promising research direction and we will conduct further research to exploit the full potential of this method.

Acknowledgements

I heartily thank my supervisor, Dr. Kwoh Chee Keong, for his excellent guidance and invaluable suggestions. His kind support allowed me to overcome the difficulties encountered in my research and eventually accomplish this work, and his encouragement also helped me to overcome the difficulties encountered in my research.

I also would like to express my sincere gratitude to Dr. Kim Jung-jae, for introducing me to the field of natural language processing and investing huge amount of time in advising my doctoral research. I could not have become a capable researcher without his knowledge, suggestions, and discussions.

I would also like to pay my sincere gratitude towards the colleagues of Bioinformatics Research Center, for their valuable comments and generous help on my doctoral study.

Finally, I would like to pay my wholehearted appreciation to my wife, Gao Xin, and my parents for everything they have done for me. Without their constant encouragement and strong support, I would not have been able to make this achievement.

Contents

Abstract	iii
Acknowledgements	v
Contents	vii
List of Figures	xi
List of Tables	xiii
List of Abbreviations	xv
List of Notations	xvii
1 Introduction and Overview	1
1.1 Motivation	3
1.2 Report Outline	4
2 Literature Survey	7
2.1 Ontology	7
2.1.1 Biomedical ontologies	8
2.1.2 Summary and discussion	13
2.2 Text Mining tasks	13
2.2.1 Named entity recognition (NER) and synonym and abbreviation extraction	14
2.2.2 Event and relation extraction	18
2.2.3 Summary and discussion	22
2.3 Active Learning	22
2.3.1 Stream-based selective sampling	23
2.3.2 Pool-based selective sampling	25
2.3.3 Query strategy	27
2.3.4 Summary and discussion	30

3	Informativity score based active learning for biomedical event extraction	31
3.1	Introduction	31
3.2	Method	32
3.3	Experiment	36
3.3.1	Experimental settings	36
3.3.2	Experiment 1: Parameter optimization	36
3.3.3	Experiment 2: Evaluation of application of active learning	37
3.3.4	Incorporation of ontology hierarchy	40
3.4	Summary	41
4	Active learning for event extraction with named entity recognition and unknown word handling	43
4.1	Introduction	43
4.2	Method	44
4.2.1	Active learning method for the event extraction	45
4.2.2	Additional component for active learning method	49
4.2.3	Active learning method for named entity recognition task	54
4.2.4	Comparison with related works	55
4.3	Experiment	58
4.3.1	Experimental settings	58
4.3.2	Evaluation metrics	59
4.3.3	Parameter optimization	60
4.3.4	Evaluation of active learning method in event extraction task	62
4.3.5	Linear combination of n-gram and PAS	71
4.3.6	Evaluation of active learning for NER task	72
4.4	Summary	76
5	Clustering Based Active Learning for Biomedical Named Entity Recognition	77
5.1	Introduction	77
5.2	Method	80
5.2.1	Vector representation model for documents	82
5.2.2	Document selection strategy from clustering result	83
5.2.3	Clustering method for active learning	84

5.3	Experiment	84
5.3.1	Datasets and NER systems	84
5.3.2	Evaluation metrics	86
5.3.3	Results	86
5.4	Summary	92
6	Conclusions and Future Work	95
6.1	Conclusions	95
6.2	Future Work	96
6.2.1	Ensemble clustering based active learning	96
6.2.2	Real world evaluation for active learning	96
	Publication	99
	References	101

List of Figures

2.1	Hierarchy structure of the GRO classes	10
2.2	The structure of the GRO relations	11
2.3	Example of event and relation extraion.	18
3.1	Comparison of active learning with PAS, entropy-based method and random selection in GRO'13	39
3.2	Comparison of active learning with n-grams, entropy-based method and random selection in GRO'13	39
3.3	Comparison of active learning with PAS, entropy-based method and random selection in CG'13	40
3.4	Comparison of active learning with n-grams, entropy-based method and random selection in CG'13	41
3.5	Incorporation of ontology hierarchy into active learning	42
4.1	General steps of the active learning method.	46
4.2	Comparison of various active learning method, including the PAS based informativity, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GRO'13.	65
4.3	Comparison of various active learning method, including the unigram informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GRO'13.	65
4.4	Comparison of various active learning method, including the PAS informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in CG'13.	66
4.5	Comparison of various active learning method, including the unigram informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in CG'13	67

4.6	Comparison of various active learning method, including the PAS informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GE'13	68
4.7	Comparison of various active learning method, including the unigram informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GE'13.	68
4.8	Integration of NER into active learning with PAS and n-grams models.	69
4.9	Evaluation of integration of word vector into active learning.	71
4.10	Evaluation of linear combination for active learning components.	73
4.11	Comparison of active learning with informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection against BioCreative dataset.	74
4.12	Comparison of active learning methods and random selection.	75
5.1	Toy example demonstrating the selection strategies of density based and clustering based active learning method.	78
5.2	Comparison of active learning methods and random selection with BioCreative dataset	87
5.3	Comparison of the performance of active learning methods with different representation models for documents, as well as the baseline of random selection. The vector representation (VectorData) is used either with or without prediction of candidate NE (NE). The performance of active learning using classical bag-of-word model is also presented.	89
5.4	Comparison of the performance of active learning methods with different document selection strategies, as well as the baseline of random selection.	90
5.5	Comparison of the performance of active learning methods with different clustering methods, as well as the baseline of random selection. For the clustering methods, the HAC, GVM, DVSCAN and SNN are considered.	92

List of Tables

2.1	Example sentence for the CoNLL-2003 NER task	16
3.1	Example of co-occurrence table for n-gram (NG_l) and ontology event concept (E_i)	33
3.2	Proposed algorithm of informativity score based active learning	35
3.3	Summary of task datasets	36
3.4	Parameter optimization results	37
4.1	Co-occurrence analysis for calculation of correlation score between E_i and NG_j	49
4.2	Example of word cosine distance in word2vec	51
4.3	Proposed algorithm of committee based active learning method	53
4.4	Comparison of datasets and ontologies	59
4.5	Parameter optimization results.	61
4.6	Parameter optimization for incorporation of event participants.	61
4.7	Parameter optimization results for dealing with OOV issue.	62
4.8	Deficiencies of document selection methods for event extraction.	64
4.9	Deficiencies of active learning methods with and without integrating the prediction of named entities (NE) against GRO'13 dataset	70
4.10	Deficiencies of using word vector toolkit in the active learning	70
4.11	Deficiencies of linear combination for active learning components	72
4.12	Deficiencies of sample selection methods against the BioCreative and CoNLL datasets	74
5.1	Proposed clustering based active learning method . . .	81

5.2	Comparison of deficiencies of active learning methods and random selection	87
5.3	Comparison of deficiencies of active learning methods with different data representation models for documents	89
5.4	Comparison of deficiencies of active learning methods with different document selection strategies	91
5.5	Comparison of deficiencies of active learning methods with different clustering methods	92

List of Abbreviations

NER	Named Entity Recognition
SVM	Support Vector Machine
CRF	Conditional Random Field
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative

List of Notations

$\ \mathbf{x} - \mathbf{y}\ ^2$	Euclidian distance between $\mathbf{x}$ and $\mathbf{y}$
$ X $	Determinant of matrix X
$\mathbf{x} \bullet \mathbf{y}$	Element-wise multiplication of $\mathbf{x}$ and $\mathbf{y}$
$\mathbf{x} \bullet / \mathbf{y}$	Element-wise division of $\mathbf{x}$ by $\mathbf{y}$

Chapter 1

Introduction and Overview

The quantity of biomedical literature is expanding at an increasing rate. Up to 2015, over 23 million citations for biomedical literature are included in the MEDLINE database, and this number is growing at the rate of more than 500,000 each year [1]. With such high-speed growth, it becomes more and more challenging for the biomedical researchers to keep up to date with all the latest advancements within their own research fields.

However, such burden of information overload is lessened as more and more sophisticated biomedical text mining systems become available, which are based upon text mining and Semantic Web technologies. Text mining, in general, is about the facts extraction, implicit link discovery and hypotheses generation to meet users' need [2]. Research topics in text mining include named entity recognition (NER) and extraction of synonym and abbreviation, relation and event extraction [3]. On the other hand, paralleling with the growth of the biomedical literature are the contained biomedical terminologies, which call for standard representations as many biomedical entities have different names and expressions. The construction of ontologies, the fundamental technology of Semantic Web, is one way to address this issue of normalization. The ontology is a conceptual model that aims to support consistent and unambiguous formal representation of a specific knowledge domain [4]. The biomedical ontologies map biology concepts to their semantics, such as biological definitions and their relations to other biological concepts. By creating the commonly acceptable conceptual model, the adoption of ontology ensures that the data sets are machine readable and processable.

Many sophisticated biomedical text mining systems are based on machine learning. The study of machine learning is about the construction and analysis of systems that can learn from data. In biomedical text mining, machine learning algorithms can be designed and implemented to learn language patterns from the limited amount of ontologically annotated documents. However, the research and implementation of biomedical text mining systems often require the ‘gold-standard’ annotated corpora [5], which contain biomedical documents that have been manually annotated with terms and relations defined in ontologies. The annotated corpora can subsequently be applied to the training of text mining systems [6].

Usually, to construct the ‘gold-standard’ annotated corpus, a group of human annotators first randomly select a pool of biomedical documents to be annotated. With the domain expertise, the annotators then manually identify and label the instances of ontology concept with proper ontological concepts defined in the ontology. After the manual annotation phrase, the inter-annotator agreement (IAA) score is calculated for the annotated corpus, and the corpus in general should have an acceptable level of IAA score. For instance, in an corpus construction project, the IAA scores may fall within the range of 66% - 90%, depending on the difficulty of the annotation tasks and the annotator discrepancies [7]. The annotated corpus could then be used for the training and evaluation of machine learning models. Such methodology of using the pre-constructed training dataset is called *passive learning*. However, the whole process of constructing the annotated corpora is time-consuming and costly, as it requires much yet tedious effort from the human annotators. Nonetheless, the manual annotated corpus is often considered of high quality.

Therefore, it is particularly valuable to construct the annotated corpus as quickly and economically as possible, through a way of directed sampling process of *active learning*. Different from the passive learning, in active learning, the learner interactively queries the annotator for information of unannotated documents and chooses the most informative document from which it learns, in order to generate as good a learner as possible with the minimized number of annotated documents.

In applying the active learning method to the corpus construction,

various approaches have been proposed. For instance, one popular entropy based approach is to merely calculate the entropy of the unannotated documents, and then rank and select out the top informative documents [8] [9] [10]. Similarly, there are other approaches, including the pointwise mutual information, proposed in [11]. Such methods are useful in reducing the time and effort for corpus construction; however, few have utilized the backbone ontology in the ontological annotations from the documents, which is a promising approach to address the complicated issues of natural languages. Another seldom visited topic is whether the proposed active learning methods could be generalized and made compatible with different existing systems or frameworks.

1.1 Motivation

Biomedical text mining is a promising way to solve the information overload issue in the biomedical domain. Before any sophisticated biomedical text mining systems can be applied, one prerequisite for its implementation and evaluation is the ‘gold-standard’ annotated corpora whose documents are manually annotated. However, the construction of annotated corpora is a time-consuming and costly process.

Usually the documents contained in the corpora are randomly sampled from the underlying population distribution of biomedical literature. We intend to construct the annotated corpora as efficiently and economically as possible, through a way of directed sampling process of active learning. The measurement of uncertainty is critical to the ranking process in the active learning method. In this thesis, we propose a definition of the informativity scores function and compare it with the existing approaches. On the other hand, few, if any, existing approaches have incorporated the hierarchy structure of ontology in the active learning method. We adopt the novel idea of addressing this issue from the perspective of semantic information in the documents; specifically, through incorporation of the hierarchy structure of the underlying ontology into the active learning method.

In addition, we consider the problem improving the performance of active learning with the help of an existing supervised learning system. We propose a committee-based approach for active learning, by adopting both the underlying supervised learning system and the proposed informativity based system. The underlying supervised learning system is used to filter out potential false negative instances, so that the manual annotation work is further reduced.

In the next step, we consider the active learning problem from the perspective of unsupervised learning, and we propose a clustering based active learning method. We carry out these experiments and apply the active learning method into an event extraction system, to show the performance of active learning and compare its effect against other document selection strategies.

Finally, the question of generalizing active learning method emerges, as to whether such application of active learning can be generalized and made beneficial to existing machine learning frameworks and systems. We utilize different data sets from different tasks to prove the consistency of efficiency of the active learning method.

1.2 Report Outline

The organization of this dissertation is as follows.

In Chapter 2, the related work is presented. We first discuss the ontology, which is the foundation of corpus document annotation. With the annotated corpus, we extend our discussion to the text mining field. We only focus on relevant text mining tasks, including named entity recognition (NER) and event and relation extraction as they are targeted in this thesis. The supplementary information for these tasks is provided. Finally we conclude this chapter with the research progress in active learning, including the types of active learning applications, and the different strategies in active learning methods.

In Chapters 3 to 5, we discuss our experiments on active learning. In Chapter 3, we use a statistical model and propose a novel informativity based document scoring system. We follow the uncertainty based active learning framework and apply the document scoring

system to the biomedical event and relation extraction problem. The experiment results show the effectiveness of the informativity score.

Built upon the document scoring system, we follow the committee-based active learning framework and apply it to the event and relation task in Chapter 4. In this framework, we maintain a committee of classifier consisting of the aforementioned informativity scoring system and the state-of-the-art event extraction system. Furthermore, we integrate the information from NER classifier into the committee to improve the committee-based active learning framework.

In Chapters 3 and 4, the core ideas in our active learning methods, either uncertainty based or committee-based, are to select documents that are most ‘informative’. However, there is one limitation of not considering if the selected documents are ‘representative’ or not, where non-representative samples such as outliers may cause overfitting and undermine the training performance of the supervised learning framework. In Chapter 5, we deploy a clustering based active learning method, to explicitly select documents that are both ‘informative’ and ‘representative’.

In the last chapter, we conclude the dissertation with a summary of knowledge learned from the experiments, followed by some suggestions for future work.

Chapter 2

Literature Survey

In this chapter, the related works are introduced. The introduction starts with the topic of ontology, as it is the basic foundation for corpus document annotation. We then focus on the text mining tasks that are only relevant in this thesis, namely, the named entity recognition (NER) and event and relation extraction. We conclude this chapter with the discussion of basic idea of active learning, which is the main research question in this thesis.

2.1 Ontology

With the current high-speed growth in the biomedical literature, it is possible that the important links between individual concepts of biomedical knowledge are not paid attentions to. Therefore, there should be methods that can aid researchers in utilizing the existing tools to help them in their research, which in turn will promote the research in text mining. In order to store and represent effectively the rich knowledge in biomedical domain, experimental evidence should be linked to ontologies as the main repositories of the knowledge.

The ontology is a conceptual model that aims to support consistent and unambiguous representation for the knowledge, and it provides a framework for integration of knowledge [4]. An ontology can link concept labels to their semantics, such as the biomedical definition and their relations to other biomedical concepts [12]. In the biomedical research domain, the research areas have highly specialized sub-fields, and there is poor communication between disciplines. While

this may be due to the characteristics of the biomedical science and hence complex and detailed researches are required as a requirement in the research of biomedical science, this poor communication tends to impede the connections between discoveries found by different researchers. In the meantime, in the post genome era with the accomplishment of sequencing the human genome, the situation is becoming even more complicated by the incorporation of the genetic information to biomedical study. As the genetics play an important role in areas related to health and disease, many links between medical studies may be ignored. Hence the adoption of ontologies in the biomedical domain as commonly accepted conceptual models is one critical approach to connect the divided subfields in biology and biomedicine.

The relations in ontology are important and can be very different. They can either be relations such as 'is-a' and 'part-of', which are in almost any domain, or relations such as 'has-location', 'temporary-related' that are domain-specific for biomedicine. Therefore, the structure of the domain is reflected in the ontology, and the potential interpretations of terms are included in the ontology. As such, ontologies can be used to support automatic semantic interpretation of textual information [2], and thus provide a basis for sophisticated Text Mining (TM) applications.

The ontologies are typically organized as a hierarchy using the 'is-a' relation between concepts. Therefore, the similarity between the concepts can be quantified using this property; also the semantic similarity between the terms that are used to designate these concepts can be quantified [13]. This numerical information, which can be inferred from ontology, is particularly valuable for TM applications. For example, the measurement of semantic similarity can be either used in the case-based reasoning [14] or in a variety of TM tasks, such as clustering [15] and classification [16].

2.1.1 Biomedical ontologies

Similar to the ontology in the general domain, the biomedical ontology is, in principle, a controlled, structured vocabulary, which can

allow the available data, generated by today's high-throughput technologies, to be searched and algorithmically processed [17]. One of the most widely used ontologies is the Gene Ontology (GO) [18], which is included in the Open Biomedical Ontology (OBO) Foundry¹. The OBO Foundry is a collaborative effort to create reference ontologies that are inter-operable within the biomedical domain. In addition to the OBO ontologies, the OBO Foundry also provides documentations concerning the development of science-based ontology. The OBO Foundry now contains 6 main ontologies, namely, Gene Ontology (GO), Chemical entities of biological interest (CHEBI), Phenotypic quality (PATO), PRotein Ontology (PRO), *Xenopus* anatomy and development (XAO) and Zebrafish anatomy and development (ZFA). The OBO Foundry also contains a list of other candidate ontologies, in which the Gene Regulation Ontology (GRO) is included.

In addition, the BioPortal² also provides access to the 359 biomedical ontologies and more than 6 million biomedical terms, as well as tools for working with the ontologies. The most viewed ontologies in BioPortal include: SNOMED Clinical Terms, National Drug File, MedDRA, International Classification of Diseases and NCI Thesaurus.

2.1.1.1 Gene Regulation Ontology (GRO)

The Gene Regulation Ontology (GRO) is an ontology that covers biological processes that are on the intra-cellular level, for example, the binding of transcription factors to DNA binding sites, and existing entities that are involved in these processes, for instance, the genes and transcription factors, with ontology concepts linked by relations [19]. 507 ontological concepts relating the gene regulation process are included in the current version of GRO. Compared to the Gene Ontology, which is used to represent overly fine-grained classes for the database annotation purpose [18], the GRO is intended to express well-accepted knowledge related to the domain and the biological relations between the entities.

¹<http://obofoundry.org/>

²<http://bioportal.bioontology.org>

Basically, two branches are included in the GRO: the *continuant* and the *occurrent*. The *continuant* describes entities that exist physically or non-physically, and the *occurrent* describes biological events or processes, such as gene expression, and various regulatory processes [19]. Figure 2.1 shows the differences.

Figure 2.1: Hierarchy structure of the GRO classes

In all, the GRO contains around 500 classes, which contains both the general knowledge in the biology domain, and the specific knowledge focused on the representation of the gene regulation event that is related to the transcription.

Besides the hierarchy structure of the GRO classes, the GRO also includes the relations, as shown in Figure 2.2.

Figure 2.2: The structure of the GRO relations

The GRO classes are linked by pre-defined relations as shown in Figure 2.2. For instance, the relation ‘part-of’ is used to show the relationship for a part and the whole, for example a domain in the protein and the protein. The relation ‘from-species’ expresses the concept of the relation between an entity and the species it belongs to.

2.1.1.2 Gene Ontology (GO)

The Gene Ontology (GO), which includes both the ontology and the annotation, is used in many biological databases, automatic analysis

for annotating genomes [20], text mining [21] [22] , network modeling [23], clinical applications [24], and others [25]. GO contains formal representation for three areas, including the molecular function, biological process, and cellular component, which is different from the structure of the GENIA ontology. In the current version of GO, there are around 25,000 GO terms for biological process, 9,500 GO terms for molecular function, and 3,100 GO terms for cellular component.

The annotations for GO assign a gene with the terms in the ontologies; this association is performed by a curator or automatically. Therefore, an evidence code is included in all GO annotations to show the basis of the annotation. Broadly, the evidence code is of four categories: experimental, computational, indirectly derived, or unknown [25]. And the annotations that have the evidence code of direct experiment are generally of higher quality than those of computational or indirect evidence.

The annotation of genes is to associate with the specific terms to express the current knowledge on the gene. The terms in the GO are structured in the form of directed acyclic graph (DAG) structure, that is to say, a specific term is usually a child node of a more general term.

For example, if a gene is annotated with the “vesicle fusion”, it is also implicitly annotated with all of its parents’ term, such as “membrane fusion”, “transport” and so on, which are more general than the former specific term and are the parents’ term to the specific term. Thus, in the querying or making inferences about this gene, it is not only regarded as the “vesicle fusion”, but also can be labeled with all its parents’ terms.

The GO is mostly been used for the analysis of high-throughput research, where a set of gene names are produced and the researchers usually utilize the GO annotation to uncover what biological processes and functions are known in the gene names, and what new potential function of the gene can be inferred from the data, and the situation of how the genes can be categorized to different biological concepts.

2.1.2 Summary and discussion

Paralleling with a rapidly increasing volume of the biomedical literature are the biomedical terminologies. This causes the issue of the lack of standardization representations; for instances, many biomedical entities have multiple names and expressions. The ontology is a conceptual model that aims to support knowledge representation that is of high consistency, and that provides a framework for integration of knowledge [4]. Therefore, it is natural to adopt the biomedical ontology in the construction of document annotation, in order to ensure that the resulting documents are machine readable and processable.

2.2 Text Mining tasks

In this section, we introduce some of text mining tasks. While the amount of biomedical literatures is expanding at an increasing rate, it becomes extremely challenging for the biomedical researchers to keep up to date with all of the latest advancements within their own research fields. To aid the biomedical researchers in tackling the issue of information overload, more and more biomedical text mining systems have been made available.

Typically, a text mining system may consist of many individual modules, and each module may solve a pre-defined specific task in the text mining domain. In this section, we introduce the text mining tasks that are targeted in this thesis, i.e., named entity recognition (NER) and the relation and event extraction.

While introducing the respective tasks, we also discuss the relevant recent international competitions on these tasks. These community driven competitions are important in two folds: 1) Such shared tasks act as an important evaluation of the available text mining systems and the development of the domain. 2) The dataset prepared in the competitions are valuable resources to support the implementation and improvement of the biomedical text mining system, even after the competitions.

2.2.1 Named entity recognition (NER) and synonym and abbreviation extraction

One of the fundamental research areas in text mining is the named entity recognition (NER) task. Theoretically, the biomedical information extraction includes: 1) the recognition of terms expressing concepts for biomedical entities, for example, the gene names and protein names are to be recognized from the raw text, 2) the extraction of specific relations between biomedical entities, for instance, the discovery of the protein–protein interactions based on the raw text, and 3) the extraction of the more complex types of biomedical information. One example is the undergoing widely known GENIA project, which is to uncover the metabolic pathways of the protein NF-kappa B. The NER is a crucial step for other sophisticated applications such as information extraction, since the steps of the entities recognition and linking them to the standardized ontological concepts in the NER are the foundations for accessing the textually described domain-specific information in the information extraction. Even besides information extraction, the NER supports the solution of more complicated tasks in text mining, such as event and relation extraction [26] [27].

The linking of the terms in raw text with the concepts pre-defined in the ontology is a requirement for the text mining systems. This step is due to fact that the terms are more variable than concepts explicitly represented in the ontology. One example of the common variations is that “EGR – 1” and “EGR 1” and “EGR1” all express the same biological concept. Thus if the three variants all appear in the raw text, they all should be considered as the same concept. It is reported that such phenomenon of variance in raw text consists of one third of the total term occurrences [28], which suggests the identification of variance may help the recognition of terms in text.

On the other hand, the extraction of synonym and abbreviation comes from the inherent characteristics of terminologies in the biomedical domain. Paralleling the growth of the biomedical literature is the expanding volume of biomedical terminologies. In the current biomedical documents, many biomedical entities have more than one names

or abbreviations. The issue has been attracting more and more attention of research, since it would be advantageous to extract the synonyms and abbreviations, so as to aid users in document retrieval in an automatic way. In the current study of biomedical text mining, the synonym and abbreviation extraction is mainly about the extraction of gene names and abbreviations of biomedical terms.

One application of the extraction of synonym and abbreviation is that it can be incorporated into the search engines to improve the search results. For example, one may use the medical abbreviation in using the PubMed to search relevant documents; the search engine, however, can integrate a dictionary of medical abbreviations, based on the dictionary of abbreviations to augment search queries and improve the search results. One thing worth noting is that if an abbreviation, which is well-accepted in a particular domain yet is not included in the maintained dictionary, is encountered in an article, the aim of improving in research results is becoming more difficult to achieve, since the solution of this issue requires the expert knowledge for that particular domain, together with the interpretation of the context of such abbreviation.

The extraction for gene name and protein name synonym is a more challenging problem. One possible way is to construct a synonym list and keep updating it. However, the quality of extracting the list is not guaranteed because of the low precision in extraction systems. Cohen [29] has used a pattern extraction approach and co-occurrence analysis to generate a synonym extraction system. Without any NER module for the gene names, the system has a score of 22% in F-measure. Yet with the identification of logical relation, the recall for the system improved by 10%.

2.2.1.1 International competitions on NER

In the following sections, we will introduce some of the latest evaluation competition tasks on NER, including the CoNLL for the general domain, and the BioCreAtIvE for the biomedical domain.

CoNLL-2003 shared task: language-independent Named Entity Recognition

The CoNLL-2003 shared task is one of the most widely used tasks for NER in the general domain. The dataset for the English language is taken from the Reuters Corpus, which consists of Reuters news stories between August 1996 and August 1997 [30]. In the data file, one word is contained per line, and the empty lines represent the sentence boundaries. Within each line, four fields are contained: the word, the word's part-of-speech tag, the chunk tag for the word, and the named entity tag for the word. The named entity tag is defined as follows: the word with O tag means it is outside of named entity; while words with I-XXX tag indicates the opposite, i.e., inside a named entity that is typed XXX. In addition, where two named entities that are of the same type XXX and are adjacent to each other, the named entity tag for the first word is B-XXX. In the shared task, four types of named entities are defined: persons (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC). An example sentence of the CoNLL-2003 NER task is given in Table 2.1

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Table 2.1: Example sentence for the CoNLL-2003 NER task

The settings for the shared task are that, learning from the provided training dataset, the text mining system need to recognize the named entities that are contained in the separate evaluation dataset. In the shared task, 16 systems have employed a wide variety of machine learning techniques, including Maximum Entropy Model [31], Hidden Markov Model [32], AdaBoost.MH [33], memory-based learning [34], Transformation-based learning [32], Support Vector Machines [35] and Conditional Random Fields [36].

The evaluation results in the task are ranging from 69% to 88% in F-measure, while recently, the state-of-the-art Stanford NER system has reported the F-measure of around 90% in the task dataset [37].

BioCreAtIvE

Recently, more and more research work has been published on the application of text mining and information extraction to the biomedical literature. However, as Hirschman pointed out, each group had addressed different problems on different data sets ([38]). Hence it became important to assess the progress of biomedical text mining systematically. Similar to the CoNLL shared tasks, the aim of organizing BioCreAtIvE challenge (Critical Assessment of Information Extraction in Biology) is to support a systematic assessment of the state-of-the-art information extraction systems. However, comparing to the CoNLL shared tasks, the BioCreAtIvE challenge is information extraction in the biomedical domain, which has its own specialty comparing to the general domain.

Two tasks are defined in the BioCreAtIvE II [39]. The first task assessed the extraction of gene names or protein names, and the linking of gene or protein names into the standardized gene identifiers in the databases for model organisms in biology, including fly, mouse and yeast. The second evaluates the functional annotation, where the participating systems were given full-text articles and should recognize specific text that supports the annotation for specific proteins with the Gene Ontology ([21]).

One aim of BioCreAtIvE challenge is to support the curation process of biological databases. The BioCreAtIvE challenge was built on the KDD Challenge Cup, which is one of the earliest evaluations competitions in the biomedical text mining ([40]). Similar to the BioCreAtIvE, the KDD Cup also hosts tasks on the curation of biological literature, in particular, the recognition of articles that contain experimental evidence for the gene products of the database Flybase³.

27 groups participated in the BioCreAtIvE challenge, and the evaluation results in task 1 are ranging from 80% to 90% in F-measure. Task 2, on the other hand, created a baseline for future research on functional annotation.

From the BioCreAtIvE challenge, interesting questions relevant to biology were raised. The gene name tagging task in BioCreAtIvE is compatible for comparison of similar tasks in other domains. The

³<http://flybase.org/>

use the template-based syntactic patterns, which usually are regular expressions and generated by domain experts, to extract the specific relation from text [42]. Another method is to use the automatic template. In this method, the templates, which are similar to that in the manual generation of patterns, are automatically generalized patterns based upon the text close to the concept pairs, which have the relation of interest [43]. The third method is a statistical approach to identify the relationships through comparing to the random predication, the ratio of the possibility of finding the concept with each other to that by chance [44].

In the current post genomic era of the biology study, more and more studies have been focused on the relationships between genes and proteins. Through grouping genes by their functional relationships, both the analysis of the gene expression and database annotation can be aided. The grouping is currently divided into two ways. One way is based on how strong their associations are in the raw text. The other way is to use the relations between genes, proteins, or other biological entities and then to extract them from the text. Based on the previous accomplishment, it is suggested that the difficulty in extracting relations is different, as some can be extracted well but others cannot. The relationships that are very general, non-specific such as the gene groups' relation between genes are easy to extract; on the other hand, specific relations that depend on their particular context, for example, the process of the GO code assignment in the annotation, remain challenging.

While the task of relation extraction is the prediction of relationship between two arguments, the task of event extraction is a more general problem of prediction for arbitrary numbers of arguments. In the structure of an event, there is a *trigger word*, which expresses the main event, and the arguments for the trigger word, which describe the detailed information relevant to the trigger word. The event extraction is addressed in several approaches. The first is through the hand-craft patterns. The patterns are similar to those in the relation extraction, except that the patterns for events may involve a variable number of arguments. The semantic type constraint on the argument is specified, and is reflected in the composition of the ontology. One of the challenges in the hand-craft pattern is that a given event

may be expressed in a variety of forms, such as the morphological, syntactic and lexical variations. To improve the performance, some syntactic generalization is often introduced. After the application of hand-craft pattern to extract candidate events, there is often an additional step of consolidation, where task-specific rules are applied to the candidate events.

In addition, the event extraction task can also be addressed with the supervised extraction, and the general approach is to train multiple classifiers, first to identify the trigger word and then the arguments ([45] [46]). In such approach, the performance of the event traction is limited by the amount of available training data, yet it is also affected by the fact that the classifiers merely depend on the local context in the same sentence. To improve the performance of event extraction, strategies of expanding the context and searching the indicators of a relevant event with wider context, i.e., to search across the sentence or even the document, are often used. For instance, the sentence-level and local evidence can be combined to improve event extraction [47]. It is also shown that the document-level topic information is beneficial for the improvement of the event identification, either by building a model of a corpus ([48]) or by utilizing statistical data of event co-occurrence ([49]).

2.2.2.1 International competitions on biomedical event and relation extraction

BioNLP-Shared Task (BioNLP-ST)

Similar to the BioCreAtIvE challenge, the BioNLP shared task (BioNLP-ST) was also organized to evaluate the research progress of information extraction application in the biomedical text mining domain. However, their focuses were different. While the BioCreAtIvE challenge assess the simple representation of protein-protein interactions (PPI), the BioNLP task is more related to the detailed biological relation, such as biological events ([50]). In addition, while the BioCreAtIvE challenge is to support the curation of PPI databases, the BioNLP-ST was intended to support the more detailed and structured databases, for example, pathway ([51]) or Gene Ontology Annotation (GOA) ([52]) databases.

Different from the previous international competitions on biomedical event and relation extractions, which mainly deals with two or three biomedical event types of PPI, the BioNLP-ST focuses on more events, which is a notable advancement for the study of event extraction.

Event and relation extraction systems on biomedical documents

In this thesis, we adopted two event and relation extraction systems, the OSEE [53] and the TEES [54]. The OSEE is a pattern-matching system that is capable of first learning the language patterns from the data sets, and then applying the learned patterns to extract events from the data sets. The OSEE system uses the GRO to integrate the semantics at all levels. In the first step, ontological instances in text are recognized by the system. And the mentions are combined into “compositional structures”, with the help of language patterns contained in the system. In the next step, the OSEE system will deduce more “implicit” events from the previously identified events, based on the domain knowledge. This step is also called the inference step. Finally, the pre-defined event templates are used to extract the events, through the matching of candidate events against the event templates.

The TEES is a free and open source NLP system developed for the event and relation extraction from biomedical text. TEES has been evaluated in the following Shared Tasks, and has proven itself to be the state-of-the-art event extraction system in BioNLP.

- BioNLP 2009 Shared Task (1st place) [55]
- BioNLP 2011 Shared Task (1st place in 4/8 tasks, the only system to participate in all tasks) [56]
- DDI 2011 (Drug-drug interactions) Challenge (4th place, at 96% of the performance of the best system) [57]
- BioNLP 2013 Shared Task (1st place in 4/8 tasks) [58]

The TEES system is dependent on efficient machine learning approaches and features that are obtained from a full dependency analysis, generated by the applications of full dependency parsers into biomedical domain. For instance, many of the best-performing systems utilize the full dependency parses [59] [60] [61] [62]. The pipeline for the

system includes three steps: the first is *trigger recognition*, followed by the step of *argument detection*, the last is the step of *semantic post-processing*. For the trigger recognition step, the reason why the TEES separate the first step from the second step is that the TEES can use methods that are similar to the task of NER, with the exception of identifying the words for event trigger. With such approach, the problem of the detection of event arguments becomes the determination of whether the pairs, in the form of trigger-entity or trigger-trigger, are an instance of event arguments. Such steps are in essence classification tasks. The final step, semantic post-processing, is addressed through the rule-based approach by directly implementing the constraints defined the definition of each Shared Task.

2.2.3 Summary and discussion

In the previous sections, we have reviewed the research fields of text mining that are targeted in this thesis, including the named entity recognition, and the event and relation extraction. In addition, we have introduced the main international competitions in information extraction, including the *CoNLL* shared task for the general domain, and the *BioCreAtIvE* and *BioNLP-ST* for the biomedical research. We also described two event extraction systems in the biomedical domain, the *OSEE* system that is pattern matching based, and the *TEES* system, which is under the framework of machine learning.

In the next section, we will discuss the progress and issues of research work on active learning.

2.3 Active Learning

Active learning is a subfield of supervised machine learning. However, active learning has its specialty such that its learning algorithm can actively choose the data it learns, in order to generate a learner as well as possible with the minimized amount of annotated data provided by an oracle [63]. And the “ordinary” machine learning is often referred to as learning through random sampling from the available data sets or passive learning. Overall, the active learning

method is to keep the human annotation effort to a minimum. This is especially useful for many supervised learning tasks, where it is very difficult, and time-consuming to obtain the annotated data by an oracle. For instance, in the domain of information extraction, sophisticated information extraction systems should be trained using annotated documents. To produce such labeled documents, annotators who often have extensive knowledge in the specific domain markup entities and relations in text, which is a particularly time-consuming process. It is reported that even for simple newswire stories, it would take half an hour or more to locate entities and relations [64]. And for tasks in the other domains, such as annotating the gene and protein mentions for biomedical information extraction, annotators must possess additional expertise.

The active learning method tackles this bottleneck of labeling by asking queries to determine the unlabeled data to be annotated. In this way, the active learner minimizes the cost of obtaining labeled instance while still generating a classifier of good performance.

2.3.1 Stream-based selective sampling

Angluin [65] has first introduced a range of different types of queries the learner can ask, including the membership query (“Is this concept an instance of the target concept?”), equivalence query (“Is concept X equivalent to concept Y?”), and disjointness (“Are concept X and concept Y disjoint?”). Theoretically, the learner can generate nearly infinite queries by altering the attribute value of the instance so as to ensure the answers to the queries as informatively as possible. However, it is likely that the generated queries may be nonsense to the oracle. For instance in the task of text classification using bag-of-words model, if the learner generates new queries by replacing some of the words in the text and asks the oracle to which class the text belongs, it is likely that the generated queries do not make sense to the oracle at all. Hence, despite the fact that the generative queries are theoretically interesting, the current applications are much more based on example-driven means as the selective sampling.

Cohn [66] has introduced the selective sampling method, which roots its core concept in the membership query yet with an example-driven

perspective. Depending on the attribute of the original data, there are two different types of selective sampling, which are discussed in detail in the following sections.

The stream refers to the situation that a data point is sampled sequentially from the data set. Each time the learner considers one unlabeled example at a time in the Machine Learning tasks, and it needs to decide whether to query the human for its label or simply discard this example.

In stream-based selective sampling, Liere [67] and MaCallum [68] have reported that the unlabeled instances are provided one by one. In this scenario, after sampling from the underlying distribution, the learner should determine promptly whether to ask the oracle for its label or discard it. And there are several frameworks for the decision of whether such request should be made. For instance, Dagan [69] has proposed an approach of evaluating the samples using some ‘informative measurement’ in advance, then the learner can make a biased random decision. In this way, the instances, which are informative, are to be queried. On the other hand, Cohn [66] has introduced the concept of the region of uncertainty, and hence to only ask for queries for samples that fall into the particular region within the sample space.

The stream-based selective sampling has been studied in several tasks. In part-of-speech (POS) tagging task, Dagan [69] has efficiently trained a probabilistic classifier by training only the most informative instance in a stream of unlabeled data, and demonstrated the feasibility of incorporating the stream-based selective sampling with some sophisticated classifiers. In learning the ranking function for information retrieval, Yu [70] has proposed an SVM selective sampling technique (or active learning with SVM), which differs from learning classification functions due to the relative ordering of the labeled data in ranking. In applying the sampling technique to the data retrieval application, they have constructed a fuzzy search technique on relational databases by learning the users’ preferences, and showed that their technique is successful in significantly reducing the labeling effort in generating accurate ranking functions. In addition, in the task of word sense disambiguation (WSD), Fujii [71] has reported that the approach of using the selective sampling not only reduces

the needed annotation work, but also keeps the database used in machine learning as small as possible, which in turn expedites the classification algorithm.

2.3.2 Pool-based selective sampling

The pool refers that the learner has been given a large pool of unlabeled documents, and then it is allowed to rank the documents by richness of information. Then the learner can select the most informative document to query the human for its label.

The pool-based sampling is driven by this situation, in which there is a relative small set of labeled data and a large pool of unlabeled data. As the example provided by McCallum [68] shows, the learner can access and compare the unlabeled data, and then selects such instances without considering their individual ordering.

The pool-based selective sampling has been proven useful in many real-world problems. For instance, in text classification task, Lewis [72] has deployed an algorithm for pool-based uncertainty sampling with a single probabilistic classifier, and showed its superiority over other sampling techniques in that the pool-based uncertainty sampling can sharply reduce the amount of text, which must be manually labeled to create an effective classifier. Similarly based on a large pool of unlabeled documents, McCallum [68] has reduced the labeled documents needed by the text classifier, yet by combining Expectation-Maximization (EM) with active learning, the word co-occurrence information contained in unlabeled documents is also utilized. The core concept for such combination is that rather than randomly select document to label as the starting point for EM hill climbing, we can obtain a better starting point by using active learning. Besides through interleaving EM with active learning, not only does EM build on the results of active learning but also inform active learning. On the other hand, the research field of support vector machine (SVM) has solid theoretical foundation [73] and excellent success in numerous learning tasks, yet generally applied to randomly selected labeled instances as training data. In the work led by Tong [74], a new algorithm of combining the SVM with active learning is

proposed, and it turns out that such combination can significantly reduce the need for labeled training data in both the problem settings of SVM. In particular, they provide a theoretical motivation using the notion of version space and present three methods, i.e., Simple Margin, MaxMin Margin and Ratio Margin, for performing active learning with SVM. Previously, active learning studies have focused on selecting a single unlabeled document to query the oracle, yet Hoi [75] has presented a novel approach of selecting the unlabeled documents in a batch manner, and used the Fisher information matrix as the measurement for uncertainty and reduced the redundancy among the selected unlabeled documents. Experiment results based on three different datasets show the superiority of effectiveness of this approach against the state-of-the-art active learning techniques for text categorization.

In the non-classification tasks such as information extraction, pool-based selective sampling has also been studied by numerous researchers. For instance, Thompson [76] has applied active learning to their parser acquisition system and information extraction (IE) system, and shows that active sample selection by using only simple certainty measures can save as high as 44% for the amount of examples needed.

Besides on the research on text, active learning has also been used for other aspects. For instance, in image classification and retrieval, Tong [77] has proposed the use of a support vector machine (SVM) active learning algorithm for conducting effective relevance feedback for image retrieval. In this algorithm, the learner selects the most informative images to query a user for their labels and learns the boundary that could classify the images. Experimental results show that the algorithm can successfully achieve an accuracy higher than the traditional query refinement approach. In video classification and retrieval, Yan [78] has extended active learning from binary classes to multiple classes, and proposed a unified multi-class active learning approach for labeling video data automatically. Besides reducing the amount of labeled data needed, the approach can also achieve error reduction of more than fifty per cent over the random sample selection. In speech recognition, the state-of-the-art statistical classification systems are trained with a large amount of utterance labeled by human, which is labor intensive and time consuming. Tur [79]

has deployed an approach of combining active learning with semi-supervised learning, and ensured that all collected data is properly utilized. Such approach alleviated the data imbalance problems that are otherwise caused by applying only active or semi-supervised learning approach.

2.3.3 Query strategy

The concept of “query” is to describe the process from the perspective of the learner; it is to say that the learner can give some decisions about the example, such as whether this document’s label is very uncertain to the learner, which also suggests such example is very informative if the human annotator gives the document a proper label. Once the label for such informative example is provided for the learner, it can update the model. Therefore, the strategy of formulating the query is critical to the active learning process. There have been many proposed frameworks for formulating query strategies.

2.3.3.1 Query by Uncertainty

In the framework of query by uncertainty, the learner queries the unlabeled data point that has the least certainty (hence the most informative) to the oracle. This approach is often applicable to probabilistic learning models.

In the binary classification, the learner usually queries the unlabeled data whose posterior probability is close to 0.5 [80].

For classification tasks of three or more class labels, a more general form may query the unlabeled data point that has the lowest confidence. This approach has been useful with statistical sequence models in information extraction (IE) tasks. For instance, Culotta [81] has discussed that traditional active learning optimizes the order in which the examples are to be labeled so as to increase learning efficiency, yet such approach fails to consider the difficulty of labeling each instance. Hence a new active learning approach is proposed. Such approach reduces not only the number of instances the human

has to annotate, but also the difficulty to annotate. Overall, this reduces the total number of annotation actions by around twenty percent.

On the other hand, the least confident strategy fails to consider the information about the remaining label distribution. To overcome such shortcomings, some researchers deploy a different multi-class query by uncertainty approach, the margin sampling. The difference between the least confident strategy and the margin sampling is that the latter incorporates the posterior of the second most likely label to annotate. However, in the least confident strategy, it merely considers the most likely label.

A more general and popular approach of query by uncertainty uses entropy as the measurement for uncertainty. Shannon [82] has defined the entropy as an information-theoretic measurement that represents the amount of information needs to encode a distribution. In machine learning, entropy is often regarded as measurement of uncertainty.

In binary classification tasks, all the three strategies, including the least confident strategy, the margin sampling and entropy-based uncertainty sampling reduce to the same principle, as to query the instance with a class posterior close to 0.5. But for more complex structured instances, such as sequence or trees, the entropy-based uncertainty sampling generalizes to probabilistic multi-label classifier.

Besides using the probabilistic classifiers, the query by uncertainty can also be used with non-probabilistic classifiers. For instance, Lewis [80] has used the query by uncertainty with the decision tree classifier. Similarly, Lindenbaum [83] has studied the scenario of query by uncertainty with nearest-neighbor classifier, which is a non-parametric classification method and is especially useful when little information is known about the underlying distribution. In this approach, each neighbor of a particular instance is allowed to vote for the class of that instance, with the weight representing the posterior label probability.

The query by uncertainty can be applied not only to classification tasks, as discuss above, but also to regression problems, which have

a long history in statistic literature and is generally referred to as optimal experimental design [84].

2.3.3.2 Query by Committee

Another strategy of interest is query by committee [85], which maintains a committee of models, trained on current labeled data, representing competing hypotheses. Then each model is allowed to vote on the labeling of the candidate for querying. After voting, the most informative query is selected as the most disagreement between the committee members.

There are different ways to construct the committee. For example, Seung [85] has first implemented the idea of query by committee by sampling a committee of two random hypotheses that are all consistent with the labeled data. Abe [86] has combined the idea of query by committee with the boosting and bagging, which both are techniques to improve the performance of an existing learning algorithm. They test their two query learning methods – query by bagging and query by boosting – that select the next instance to annotate by picking a point on which the weighted majority voting has the least margin, and gain an increase in the data efficiency with two to four-folds. The boosting algorithm is to re-run the learning process on the re-sample data set, and updates the weight multiplicatively. The boosting algorithm does not require any prior knowledge of the weak performance algorithm [87]. On the other hand, bagging predictor is to generate multiple versions and use these to obtain an aggregated predictor [88]. As a result, the bagging algorithm can combine the output hypotheses to make a prediction by weighted majority voting.

Another ensemble-based approach to construct the committee is proposed by Melville [89]. To ensure the effectiveness of the query by committee, it is critical that the committee member should have consistent hypotheses and are very different from each other. Hence, they have introduced the DECORATE method that directly constructs diverse committee, and also incorporated the idea of active learning, in order to require far less labeled data and produce considerable reduction in error. The query by committee algorithm can also be

applied to regression tasks. One approach is to measure the variance among the committee members' prediction [90] .

2.3.4 Summary and discussion

In general, an active learning algorithm helps the learner obtain the desired outputs through interactively querying the oracle, in situations where unlabeled data are in abundance yet the human effort of manual labeling is expensive. In this way, the learner actively chooses data samples, and the number of data samples in need can be much lower than that in the classical passive learning. There are two general scenarios for the active learning; namely, the stream-based selective sampling, where the data points are provided sequentially and earlier points cannot be re-visited, and the pool-based selective sampling, where the data points can be re-used. In the task of annotating biomedical literature with biomedical ontologies, since there is no such restriction on the sequential order among biomedical documents, the pool-based selective sampling fits this scenario.

On the other hand, two query strategies have been proposed: 1) query by uncertainty, where the learner chooses the data point that is most uncertain; 2) query by committee, where multiple learners are maintained and the data point, whose classification results by each learner are most distant, is selected. In the specific circumstances of corpus construction, as the classifier is scarce, the strategy of query by uncertainty is preferred. Moreover, for the query by uncertainty, few existing approaches have tackled the problem from the perspective of ontological technology, which is a promising way to address the complicated problem of the natural language.

Chapter 3

Informativity score based active learning for biomedical event extraction

3.1 Introduction

In biomedical information extraction, some researchers have applied active learning to the extraction of PPIs. For instance, [91] proposed an uncertainty sampling-based active learning method, and [92] proposed maximum uncertainty based and density based sample selection strategies. In these works, the ‘informativity’ of a document is measured based on whether the document contains any expression of PPI or not. However, the identification and extraction of PPI is a simple task, while recent biomedical event and relation extraction tasks [93] are much more complicated.

In this work, we introduce the informativity based document scoring system. An ontology may have multiple event concepts, and a sentence may express multiple concepts in different parts. We thus consider the ‘informativity’ of a document as collective likelihood of containing individual concepts, ignoring the locations of the concept expressions in the document. The detailed definition is given in the Section 3.2.

Because of the ease of obtaining biomedical documents, we adapt pool-based active learning, where learning begins with a small portion of annotated documents and the learner is to determine whether

a specific unannotated document is informative or not for its subsequent learning process.

3.2 Method

The tasks of our interest are to annotate all relevant concepts and relations from given ontologies on specific spans of text [94]. To simplify the problem, our proposed method of using active learning does not locate the exact text spans that express ontology concepts/reasons, but checks if a document as a whole expresses any concept/reason. In the method, we collect every evidence of ontology concepts to measure the informativity of a document for a given ontology.

The method takes the following iteratively steps: We build a language model based on an initial training dataset. We measure the informativity of each unlabelled document for an ontology and select the most informative documents for manual annotation and for rebuilding the language model. We then update the informativity of unlabelled documents using the retrained systems and continue to increase the training data size until the system performance is saturated.

During each iteration of active learning, the informativity score of a document is measured at the sentence level, that is, the sum of the informativity scores for all the sentences is used. For each sentence (S_k), we define the informativity score $I(S_k)$ in two dimensions: the ontology event concepts (E_i) and relations (R_j), as expressed in Equation (3.1).

$$I(S_k) = \sum_{E_i \in S_k} P(E_i|S_k) + \sum_{R_j \in S_k} P(R_j|S_k) \quad (3.1)$$

The conditional probability in (3.1) is estimated using the Bayes' theorem as shown in formulas (3.2) and (3.3).

$$P(E_i|S_k) = \frac{P(E_i)P(S_k|E_i)}{P(S_k)} \quad (3.2)$$

	Event E_i exist	Event E_i NOT exist
N-gram NG_l exist	a	b
N-gram NG_l NOT exist	c	d

Table 3.1: Example of co-occurrence table for n-gram (NG_l) and ontology event concept (E_i)

$$P(R_j|S_k) = \frac{P(R_j)P(S_k|R_j)}{P(S_k)} \quad (3.3)$$

We decompose a sentence in two ways: n-grams (NG) and predicate-argument relations (PAS), using the Enju parser [95]. Equations (3.4, 3.5) show the conditional probabilities of event concept based on the two decomposition methods, respectively. The conditional probabilities of relation are calculated likewise.

$$P(S_k|E_i) \stackrel{\text{def}}{=} S(S_k|E_i) = \sum_{l; NG_l \in S_k} W(NG_l, E_i) \quad (3.4)$$

$$P(S_k|E_i) \stackrel{\text{def}}{=} S(S_k|E_i) = \sum_{l; PAS_l \in S_k} W(PAS_l, E_i) \quad (3.5)$$

The weight score $W(NG_l, E_i)$ (similarly for PAS_l and R_j) is dependent on co-occurrences between the n-gram (NG_l) and the ontology event concept (E_i).

Specifically, we first construct a co-occurrence table for the n-gram (NG_l) and the ontology event concept (E_i) based on the training dataset, which is shown in Table 3.1.

Based on the co-occurrence table, the score is thus calculated in three ways: chi-square test, relative risk and odds ratio [96]. For instance, by using the chi-square test, the score calculation is shown in Equation (3.6).

$$\begin{aligned}
W(NG_l, E_i) &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
&= \frac{(a - \frac{(a+b)(a+c)}{(a+b+c+d)})^2}{\frac{(a+b)(a+c)}{(a+b+c+d)}} + \frac{(b - \frac{(a+b)(b+d)}{(a+b+c+d)})^2}{\frac{(a+b)(b+c)}{(a+b+c+d)}} \\
&\quad + \frac{(c - \frac{(c+d)(a+c)}{(a+b+c+d)})^2}{\frac{(c+d)(a+c)}{(a+b+c+d)}} + \frac{(d - \frac{(b+d)(c+d)}{(a+b+c+d)})^2}{\frac{(b+d)(c+d)}{(a+b+c+d)}}
\end{aligned} \tag{3.6}$$

In addition, we incorporate the hierarchy structure of an ontology into the statistical active learning framework as follows: Given an event concept E_i and a sentence S_k , we replace $S(E_i|S_k)$ with the sum of $S(E_m|S_k)$ for all the ancestor concepts of E_i , as shown in Equation (3.7).

$$S(E_i|S_k) = \sum_{m; E_i \subseteq E_m} S(E_m|S_k) \tag{3.7}$$

Table 3.2 summarizes the pseudo codes for the proposed active learning method.

Input: labeled document pool L , unlabeled document pool U , batch size b

// Initialization
 ER_0 = the set of events/relations annotated on L
Learn a language model M_0 from ER_0

// Active Learning Loop
while U is **not** empty:
 for each document D_{ij} in U :
 Document informativity score $I(D_{ij}) = 0$
 for each sentence S_k in D_{ij} :
 for each ontological event/relation er :
 $I(D_{ij}) +=$ informativity score $I(S_k, er)$
 $I(D_{ij}) = I(D_{ij}) / \text{sizeOf}(D_{ij})$
 Rank D_{ij} in U based on $I(D_{ij})$ and select the top b documents, designated as B
 Remove B from U , add B to L , and add the annotations on B to ER_{i-1} , designated as ER_i
 Learn a new model M_i from ER_i

Table 3.2: Proposed algorithm of informativity score based active learning

3.3 Experiment

3.3.1 Experimental settings

The BioNLP shared tasks (BioNLP-ST) is a community task to evaluate the research progress for the information extraction area that is applied into the biomedical text mining domain. In this chapter, we used the datasets from two tasks, namely GRO'13 (Gene Regulation Ontology) and CG'13 (Cancer Genetics). Each corpus was manually annotated with an underlying ontology, whose number of concepts and hierarchy are different from the others. Those differences in each task's ontology bring about difference in the results of our experiments as shown in Section 4. A comparison between the two is given in Table 3.3.

Task	Number of ontology concepts	Ontology depth	Corpus size (abstracts)
GRO'13	507	7	300
CG'13	18	2	400

Table 3.3: Summary of task datasets

In this study, we propose a novel active learning method and test it against the two datasets above, using the state-of-the-art biomedical event extraction system, namely the TEES (introduced in 2.2.2.1). The TEES is based on SVM and was the only system that is applicable to all the tasks of BioNLP-ST'13, and has achieved the best performance in many tasks [56].

3.3.2 Experiment 1: Parameter optimization

We first take a separate parameter optimization step, in order to determine the optimal method for the calculation of the aforementioned weight score and the most effective n-gram size. A simulation of ontology concept prediction is carried out at the sentence level. In this task, for a sentence S_i that contains N manually annotated ontology concepts, only the original string of S_i and the number N are given to the informativity calculator. The calculator predicts top N candidate ontology concepts for this given sentence.

Using 10-fold cross validation, we calculate the average accuracy and report in Table 3.4. Each column stands for a particular n-gram size, and each row shows the co-occurrence calculation methods used for the prediction. When a larger size n-gram is used, the smaller size n-gram is not included. For instance, when $N=2$ (i.e. bi-grams), the unigrams are not included for the calculation. We carry out this parameter optimization using the GRO'13 dataset.

Calculation Method	N-gram				
	N=1	N=2	N=3	N=4	N=5
chi-square	0.507	0.413	0.159	0.036	0.009
relative ratio	0.341	0.395	0.307	0.128	0.038
odds	0.420	0.395	0.274	0.117	0.035

Table 3.4: Parameter optimization results

In Table 3.4, we note that for each weight score calculation method, the average accuracy drops when the size of N-grams increases. This may be due to the data sparseness issue when large N-grams are used. From the result of parameter optimization, we choose to use chi-square test and unigrams for the following experiments.

3.3.3 Experiment 2: Evaluation of application of active learning

In this section, we evaluate our proposed active learning method and compare it against other document selection strategies, including random selection (i.e. baseline) and entropy-based active learning [92]. Regarding the experiment setting, we set the selection round to ten, and in each round, 10% of the unlabeled documents are selected for annotation, and are added to training data once annotated. Before the first round, the initial model of the TEES is separately trained only on the development dataset. We compare the following document strategies for the 10% documents that are selected in each round:

- Random selection: We randomly split the training data into 10 bins in advance, and during the training phase in each round,

one bin is randomly chosen. We report the averaged performance of random selection for ten times (hereafter referred to as RS_Average).

- Entropy-based active learning: We calculate the entropy of each document regardless of the given ontology, sort documents by their entropy values and select the documents with top values to those with bottom values as training data. Note that the sorting will be done only once. (designated as Entropy)
- Proposed active learning: We vary the method in two orthogonal dimensions: 1) Using unigrams (Unigram) feature or predicate-argument relations (PAS) feature, and 2) the weight scores for events only (Event), relations only (Relation), and both events and relations (Event + Relation). In each round, we re-calculate the informativity of all remaining documents in the training data and feed those with top scores as training data.

Note that all the datasets were already annotated by the shared task organizers, and so there is no need to annotate them again. In the cases of real application to unlabelled data, after selecting a subset of the data as training data in a round, the documents in the subset should be annotated by domain experts. In the testing phase, the F-score will be measured as the performance of the methods.

We first applied those methods to the dataset of GRO'13 [97] and measured the performance change with incremental feed of the training data by using the TEES system. The evaluation results are plotted in Figures 3.1 and 3.2. As shown in the figures, the PAS models generally perform better than the unigram models. In Figure 3.1, the method using AL(Event_PAS + Relation_PAS) reaches the saturated performance only with 60% of the original training data and is stabilized as more training data are added for the system training. Saturated performance indicates the performance of a system when it is trained with 100% of the training data. Note that the proposed active learning models perform better than the random selection and the entropy-based active learning in most cases.

At the point of using 30% of training data in Figure 3.1, the performance of the active learning methods all fall significantly, though quickly restored afterwards. We examined the results at the point

Figure 3.1: Comparison of active learning with PAS, entropy-based method and random selection in GRO'13

and found that the performance of relation identification was relatively low. It may mean that the relation identification is unstable with small amount of training data.

Figure 3.2: Comparison of active learning with n-grams, entropy-based method and random selection in GRO'13

We then carried out a similar experiment using the CG'13 dataset. Figures 3.3 and 3.4 plot the results of the experiments. In these experiments, we can again confirm that the PAS models outperform

the unigram models and that the proposed active learning methods outperform the random selection and the entropy-based active learning. In Figure 3.3, the method using AL(Event_PAS + Relation_PAS) reaches the saturated performance only with 70% of the original training data.

However, the performance improvement is less significant compared to that of the GRO'13. Note that the ontology of CG'13 is much smaller than that of GRO'13, and thus that the CG'13 task would require less training data for performance saturation and so there is less room for contribution from active learning. This claim can be supported by comparing the performance changes for the random selection models against the two datasets. As shown in Figures 3.2 and 3.4, the performance of the random selection model for the CG'13 is getting saturated, while that for the GRO'13 is not.

Figure 3.3: Comparison of active learning with PAS, entropy-based method and random selection in CG'13

3.3.4 Incorporation of ontology hierarchy

Apart from the previous comparisons, we incorporate ontology hierarchies into the active learning method. We carried out experiments using the GRO'13 dataset, as its ontology depth is bigger than that of CG'13. Note that in this experiment, only the extraction of events but not relations is evaluated, since only event concepts have a hierarchy

Figure 3.4: Comparison of active learning with n-grams, entropy-based method and random selection in CG'13

structure in the GRO. We use PAS models for this experiment as they show better performance than unigram models in the previous comparisons.

The results of the ontology hierarchy incorporation is plotted in Figure 3.5. The incorporation helps system performance reach the reported performance of the TEES system from the point of using 60% of training data and not fall below the reported one, while the active learning method of AL(Event_PAS) without the ontology hierarchy information shows unstable performance. Additionally, we have also tried to use descendants, but the performance is worse than using ancestors.

3.4 Summary

In this study, we propose a novel active learning method for ontological event extraction, which is more complicated than the simple PPI extraction. Our method measures the collective 'informativity' for unrecognizable biological events expressed in documents. We found that the proposed method of using the predicate-argument structure and the ontology hierarchy is able to make the underlying event extraction system reach saturated performance with significantly fewer

Figure 3.5: Incorporation of ontology hierarchy into active learning

documents and hence reduce the amount of needed documents for manual annotation.

Chapter 4

Active learning for event extraction with named entity recognition and unknown word handling

4.1 Introduction

Previously in Chapter 3, we adopted a novel statistical model for informativity estimation, which predicts the possibility for the input text to express the event concept that is defined in the given ontology. The method uses statistical model in co-occurrences analysis. In this chapter, we use such informativity score system as an additional classifier, together with an existing event extraction system, to study the application of committee-based active learning into biomedical event extraction task.

By maintaining a committee of classifiers, the committee-based active learning approach is to select the documents so that the documents' label classification results mostly disagree from each other in the results of the committee of classifiers. The committee-based approach, however, has several issues in its application to the event extraction. First of all, the event extraction task is in essence different from the applications of active learning in the other domains, as they are mainly focused on document classification. However, in the definition of event extraction task, such as PPI extraction, gene regulation identification, the systems have to identify not only event keywords (also known as trigger word, e.g. regulation), but also

event participants (also known as event arguments, e.g. gene, protein) within the input text; in addition, the systems need to extract relations between them that are pre-defined in the given ontology. Hence, even if some systems may provide confidence scores for their recognized resultant events; however, these scores do not correspond to the probability for an particular event type is present semantically in a document, which is the probability to determine if a document belongs to that event type, as required in the committee-based active learning. Second, the previous classifiers adopted in the committee-based method may miss detailed information of events, such as the event participants. For instance, the presence of the keyword ‘expression’ in the text may mislead event extraction system to predict that the text are expressing gene expression event, while it may actually do not contain any gene name. This scenario is a false positive from the perspective of the active learning for event extraction.

We use an adjusted committee-based active learning method, and combine the statistical model with the underlying TEES system (introduced in the Section 2.2.2.1 in Chapter 2) as follows: we take the outputs from TEES as binary, which is the classification for the document of its presence of any instance of a concept. We capture the discordance or disagreement between TEES and the proposed statistical model only when 1) the TEES does not recognize any instance of concept in the text and 2) the probabilistic model estimates a high probability for the presence of the concept in the text. In this way, the TEES is used for removing false positives examples, while the probabilistic model is to rank the documents and select the documents.

In addition to the event extraction task, we further evaluate our active learning method in named entity recognition task, particularly the gene name recognition task. The experiment shows that our approach is also effective for named entity recognition.

4.2 Method

The general plan of active learning is as follows: At the start of round t , $\mathcal{U}^{(t-1)}$ denotes the set of unlabeled documents and $\mathcal{L}^{(t-1)}$ is the set of labeled documents, where t initiate from 1. In round t , the most

‘informative’ document $x^{(t)}$ from $\mathcal{U}$ is selected and annotated manually, and is added into $\mathcal{L}$. If the label $y^{(t)}$ is assigned to the document $x^{(t)}$ by the annotator, the labeled and unlabeled document are updated as follows:

$$\mathcal{L}^{(t)} = \mathcal{L}^{(t-1)} \cup \{(x^{(t)}, y^{(t)})\} \quad \mathcal{U}^{(t)} = \mathcal{U}^{(t-1)} \setminus x^{(t)} \quad (4.1)$$

Such process is repeated until the stopping condition is met, such as when $\mathcal{U} = \emptyset$ or the maximum number of iteration is reached. Rather than select one document a time in each round, the selection can be done in a batch mode, where in each round multiple documents are chosen for annotation.

4.2.1 Active learning method for the event extraction

Our active learning method follows the committee-based framework, and we deploy two classifiers: one based on the TEES system (introduced in the Section 2.2.2.1 in Chapter 2) and the other is the statistical classifier. Although the TEES system can produce the confidence score for each extracted event, such score is not used for active learning because it does not indicate the probability for the presence of event in the document. In this work, we consider that if TEES can extract an event (E) from a document (D), D is no longer informative for the annotation of E, because such true positive instances of extracted events are non-informative and the correction (i.e. annotation) of false positive instances may not be beneficial for the training of event extraction systems, as the text expressing the event is scarce, and the amount of negative instances are greater than that of positive instances. In other words, the main purpose in our active learning method is to correct more false negative instances, which suggests that we need to identify true events that are missed by the existing system. In Figure 4.1, we depict the plan of the proposed method.

Our method takes the following repeated steps: In round t , we train the TEES and the statistical model based on $\mathcal{L}^{(t-1)}$, then we evaluate the informativity for each unlabeled document in $\mathcal{U}^{(t-1)}$. Based on the informativity for each document, we select the top documents for the following manual annotation. The informativity score for the

Figure 4.1: General steps of the active learning method.

document is measured at sentence level, in which the average informativity value for all the single sentences is calculated, as illustrated in Equation (4.2).

$$x_{Informativity}^* = \underset{x}{argmax} I_{\theta^t}(x) I_{\theta^t}(x) = \frac{1}{||x||} \sum_{S_k \in x} I(S_k) \quad (4.2)$$

The θ^t indicates the present state of the TEES and the statistical classifier at round t , but we omit it hereafter for simplicity.

The sentence informativity (S_k) is calculated for event concept set

$\mathcal{E}$. Note that the event concept set $\mathcal{E}$ contains all the events that are defined in the ontology, as expressed in Equation (4.3). The $I(S_k, \mathcal{E})$ stands for the informativity score for the whole event concept set. In the BioNLP-ST'13 tasks, however, not only biomedical events but also relations are included. A fundamental difference between the two is that for event, it always involves an event keyword or trigger word (e.g. 'expression' for GeneExpression), but for relation there are no keywords or trigger words (e.g. hasAgent). For simplicity, we mention only events in this method introduction, while our method treat both events and relations in the same way.

$$I(S_k) = I(S_k, \mathcal{E}) \quad (4.3)$$

4.2.1.1 Informativity for event concept set

To calculate the informativity score for a sentence, we sum up the individual informativity score for the whole event concept set, which is as follows:

$$I(S_k, \mathcal{E}) = \sum_{E_i \in \mathcal{E}} I(S_k, E_i) \quad (4.4)$$

Here a sentence is treated as non-informative, for an particular event, only if any instance of the event can be identified by the TEES. If not, we calculate the informativity score by using the probability of the concept given the sentence, which is illustrated as follows:

$$I(S_k, E_i) = \begin{cases} 0 & \text{if } E_i \text{ is extracted in } S_k \text{ by TEES in round } t \\ p(E_i|S_k) & \text{otherwise} \end{cases} \quad (4.5)$$

The $p(E_i|S_k)$ is converted into (4.6) based on the Bayes' theorem.

$$p(E_i|S_k) = \frac{p(E_i)p(S_k|E_i)}{p(S_k)} \quad (4.6)$$

Based on the statistics of event annotations in the data, we measure the $P(E_i)$ using maximum-likelihood estimation (MLE) approach.

Regarding to $P(S_k|E_i)$, we calculate it by using the correlation information between sentence S_k and event E_i with a scoring function Z and we incorporate the scoring function into the softmax function to

represent it as a probabilistic value, shown in (4.7).

$$p(S_k|E_i) = \sigma(Z(S_k : E_i)) = \frac{1}{\sum_{E_j \in \mathcal{E}} \exp(Z(S_k : E_j))} \exp(Z(S_k : E_i)) \quad (4.7)$$

In this work, two types of features from the input sentence are used in the calculation: n-grams (NG) and predicate-argument relations (PAS), which are parsed by Enju parser [95]. We represent a sentence as a bag of elements of a unit.

A. Using N-gram feature for probability calculation

In this section, we calculate the probability using the bag of n-gram model. Particularly, the score $Z(S_k : E_i)$ is calculated with the average of the correlation score between the contained n-gram (NG) and the event. The formula is illustrated in (4.8), and the $\text{len}(S_k)$ is the normalization factor, which is the word count for sentence S_k .

$$Z(S_k : E_i) = \frac{1}{\text{len}(S_k)} \sum_{NG_j \in S_k} p(NG_j|E_i) \quad (4.8)$$

We measure the probability $p(NG_j|E_i)$ by using a correlation score $W(NG_j, E_i)$ between them, and then incorporate it into the softmax function, which is illustrated in (4.9).

$$\begin{aligned} p(NG_j|E_i) &= \sigma(W(NG_j, E_i)) \\ &= \frac{1}{\sum_{NG_l \in \mathcal{U}} \exp(W(NG_l, E_i))} \exp(W(NG_j, E_i)) \end{aligned} \quad (4.9)$$

We calculate correlation score $W(NG_j, E_i)$ by using the following statistical calculation approaches: 1) chi-square test, 2) relative risk, and 3) odds ratio [96]. In order to use the three methods for calculation, a separate 2×2 table is maintained for N-gram and event, as shown in Table 4.1. Each entry is the count at sentence level. For instance, the a is the statistics of sentences that include both N-gram NG_j and event E_i .

Based on the co-occurrence table, the three calculation method of chi-square test, relative risk, and odds ratio are shown in the formulas

	Express Event E_i	Not express Event E_i	Total
Include N-gram NG_j	a	b	N_A
Not include N-gram NG_j	c	d	N_B
Total	N_S	N_F	N

Table 4.1: Co-occurrence analysis for calculation of correlation score between E_i and NG_j

(4.10), (4.11), and (4.12), respectively.

$$W(NG_j, E_i) = \frac{N(|ad - bc| - N/2)^2}{N_S N_F N_A N_B} \quad (4.10)$$

$$W(NG_j, E_i) = \frac{a/(a+b)}{c/(c+d)} \quad (4.11)$$

$$W(NG_j, E_i) = \frac{a/b}{c/d} \quad (4.12)$$

B. Using PAS feature for probability estimation

We also use the bag of predicate-argument relation (PAS) model to calculate the probability. Similar to the previous section, the score $Z(S_k : E_i)$ is measured by using the average of the correlation scores between event and PAS, as in (4.13).

$$Z(S_k : E_i) = \frac{1}{len(S_k)} \sum_{PAS_j \in S_k} p(PAS_j | E_i) \quad (4.13)$$

4.2.2 Additional component for active learning method

In this work, we introduce two additional components in our method: we integrate the event participants in the method, and we solve the out-of-vocabulary issue.

4.2.2.1 Incorporation of event participants

In general, when there are predicted events in the text yet no existence of event participants are identified, the predicted events are potential false positive instances. To filter out the false positive instances, we integrate a gene name recognition system into the active learning system. The adopt gene name recognition system is

Gimli [98], as it is the state-of-the-art system in the gene name recognition domain. Note that we adopted Gimli because that most of the BioNLP shared tasks involve genes and proteins (e.g. protein phosphorylation, gene regulation) in the biomedical events. The integration of Gimli into the active learning method is illustrated as follows:

$$I(S_k) = I(S_k, \mathcal{E}, NG) + I(S_k, \mathcal{N}) \quad (4.14)$$

$$I(S_k, \mathcal{N}) = \delta \times T \quad (4.15)$$

T is the number of the recognized gene/protein names in S_k .

Do note that instead of the Gimli system, other named entity recognition systems can also be adopted when the event participants in the event extraction task are not mainly gene/protein. In this work, as we evaluate our method using the dataset from BioNLP shared tasks which are in the biomedical domain, we use Gimli system in the integration of event participants. However, as we also evaluate our active learning method in named entity recognition (NER) tasks, we have evaluated against two NER systems, which are Gimli and the Stanford NER.

4.2.2.2 Dealing with OOV issue with word similarity

When applying the n-gram model to calculate the informativity score, there is the Out-of-Vocabulary (OOV) issue, in which certain n-grams in the test dataset do not appear in the training dataset. To solve this problem, we also used the word2vec toolkit, which is a system to represent each word as a vector by using latent semantic model and for calculating word similarity [99].

The word2vec toolkit takes an unannotated corpus as input, and build the vector representation of each word as the output. With the vector representation of each word, one application is to calculate the word similarity using the cosine distances between the vectors of the words. One example of the word cosine distance is shown in Table 4.2, in which the input query word is **france**, and by using the word

vector representation, the words that have close cosine distance are shown, together with the computed similarity.

Input query word: france	
Word	Similarity
spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
switzerland	0.622323
luxembourg	0.610033
portugal	0.577154
russia	0.571507
germany	0.563291
catalonia	0.534176

Table 4.2: Example of word cosine distance in word2vec

We use the word2vec in the following way: For a given n-gram NG_{out} that is out-of-vocabulary, we use word2vec toolkit to identify the top- k n-grams NG_{in} that have greatest similarity to NG_{out} , where the similarity score between them is the $Sim(NG_{out}, NG_{in})$, provided by word2vec. We recalculate the correlation scoring function $W(NG_{out}, E_i)$ as shown in Formula (4.16). While the word2vec toolkit have the limitation that it can only find the similarity between single words or unigrams, yet in the previous parameter optimization experiments the unigrams show best performance, in our application of word2vec we do not consider its limitation. To obtain the similarity scores, we used the whole MEDLINE abstracts set released in April 2014 to train the word2vec toolkit.

$$W_{OOV}(NG_{out}, E_i) = \sum_{\substack{NG_{in} \in TrainingDataset \\ top-k}} W(NG_{in}, E_i) \times Sim_{NG_{out}, NG_{in}} \quad (4.16)$$

The revised correlation scoring function is shown in (4.16), and the informativity score in the the n-gram model is denoted as $I(S_k, NG_{OOV})$. In particular, when the correlation scoring function in (4.9) is used, the corresponding informativity in (4.4) is denoted as $I(S_k, \mathcal{E}, NG_{OOV})$.

4.2.2.3 Linear combination of n-gram and PAS features

Previously we use either n-grams or PAS as features, but we also evaluate the performance for the linear combination of the two, as follows:

$$\begin{aligned} I(S_k) &= \alpha \times I(S_k, NG_{OOV}) + \beta \times I(S_k, PAS) + \gamma \times I(S_k, \mathcal{N}) \\ &= \alpha \times I(S_k, \mathcal{E}, NG_{OOV}) + \beta \times I(S_k, \mathcal{E}, PAS) + \gamma \times I(S_k, \mathcal{N}) \end{aligned} \quad (4.17)$$

Table 4.3 illustrates the algorithm for the calculation of informativity scores.

Input: labeled document pool L , unlabeled document pool U , batch size b

// Initialization
 ER_0 = the set of events/relations annotated on L
Learn a TEES model M_0 from ER_0
 $i = 0$ // the index of the current round

// Active Learning Loop
while U is not empty:
 $i += 1$
 for each document D_{ij} in U :
 Document informativity score $I(D_{ij}) = 0$
 for each sentence S_k in D_{ij} :
 Apply M_{i-1} to S_k and collect the resultant events/relations set ER_{S_k}
 for each event/relation er s.t. $er \notin ER_{S_k}$:
 $I(D_{ij}) +=$ informativity score $I(S_k, er)$
 $I(D_{ij}) = I(D_{ij}) / \text{sizeOf}(D_{ij})$
 Rank D_{ij} in U based on $I(D_{ij})$ and select the top b documents, designated as B
 Remove B from U , add B to L , and add the annotations on B to ER_{i-1} , designated as ER_i
 Learn a new model M_i from ER_i

Table 4.3: Proposed algorithm of committee based active learning method

4.2.3 Active learning method for named entity recognition task

We also apply our proposed active learning method into the named entity recognition (NER) task, which is to recognize named entities that are defined in the ontology concept, such as Protein or DNA, instead of biological events, such as protein binding and gene expression. Similarly to our application of active learning method in event extraction task, in NER task we also estimate the informativity of a text based on its probability of expressing any named entities.

Similar to Equation (4.2), the informativity estimation in the NER task is expressed in (4.18).

$$x_{Informativity}^* = \underset{x}{argmax} I_{\theta^t}(x)I_{\theta^t}(x) = \frac{1}{||x||} \sum_{S_k \in x} I(S_k) \quad (4.18)$$

The θ^t indicates the current model, at round t , that are trained for the NER system and the statistical classifier, and we will omit it hereafter for simplicity. In this work, we show the performance of our method with two NER systems: Gimli for the NER task in the biomedical domain and Stanford NER system for NER task in the general domain.

We calculate the sentence informativity value using the sum of the informativity scores for all named entities, which is illustrated as follows:

$$I(S_k) = I(S_k, \mathcal{N}) = \sum_{N_i \in \mathcal{N}} I(S_k, N_i) \quad (4.19)$$

Similar to our proposed active learning method for event extraction, when any instances of the named entity is recognized by the NER system from the sentence, such sentence is not informative for that named entity Otherwise, we estimate the informativity score as the probability of the named entity given the sentence, which is illustrated as follows:

$$I(S_k, N_i) = \begin{cases} 0 & \text{if } N_i \text{ is recognized in } S_k \text{ by the NER system} \\ p(N_i|S_k) & \text{otherwise} \end{cases} \quad (4.20)$$

The probability $p(N_i|S_k)$ is calculated as follows:

$$p(N_i|S_k) = \frac{p(N_i)p(S_k|N_i)}{p(S_k)} \quad (4.21)$$

Using the maximum-likelihood estimation (MLE) based on the annotation of named entities in the training data, the $p(N_i)$ is calculated. For the $p(S_k|N_i)$, we follow similar steps in (4.7), yet we only use n-grams model (i.e. Formula (4.8)), not using PAS model (i.e. Formula (4.13)).

4.2.4 Comparison with related works

In this section, we introduce the previous methods that we compare with our proposed methods.

A. Conventional committee-based method

The committee based active learning maintains a committee of classifiers and then choose the documents in which the classifiers have greatest disagreements about its classification, which is expressed as follows:

$$x_{Committee}^* = \underset{x}{argmax} D_{\theta}(Y|x) \quad (4.22)$$

The $D_{\theta}(Y|x)$ measures such disagreement for a document x under the model θ , and the Y denotes the whole label set. In our work, the of disagreement over the sentence S_k contained in the document x is summed up.

$$D_{\theta}(Y|x) = \sum_{S_k \in x} D(Y|S_k) \quad (4.23)$$

In the sentence, the total disagreement over the whole event concept set $\mathcal{E}$, which is defined in the ontology, is measured by the summation of all disagreement for all event E_i .

$$D(Y|S_k) = \sum_{E_i \in \mathcal{E}} D(E_i|S_k) \quad (4.24)$$

The disagreement $D(E_i|S_k)$ is calculated using the absolute value for the abstraction result of the probability produced by the classifiers, which includes the informativity calculation system and the TEES system.

$$D(E_i|S_k) = |p_{Informativity}(E_i|S_k) - p_{TEES}(E_i|S_k)| \quad (4.25)$$

The $p_{TEES}(E_i|S_k)$ denotes the probability produced by TEES system, and $p_{Informativity}(E_i|S_k)$ is the informativity estimation using statistical method, which is calculated in Equation (4.6). Note that while $p(E_i|S_k)$ in Equation (4.5) is estimated as in Equation (4.6) only for sentences in which no E_i is recognized by TEES, the same informativity probability in Equation (4.25) is considered for all sentences within the unlabeled documents set.

Which the TEES is based on support vector machine (SVM) and do not output probabilistic values, we instead take the confidence value for the decision of event prediction into the calculation. This is illustrated as follows:

$$p_{TEES}(E_i|S_k) = \sigma(C(E_i|S_k)) = \frac{1}{\sum_{E_j \in \mathcal{E}} \exp(C(E_j|S_k))} \exp(C(E_i|S_k)) \quad (4.26)$$

The $C(E_i|S_k)$ is the confidence for the classifier.

The confidence is calculated using the **difference-2** distance produced by the classifier, as it is shown to have best performance in the active learning[100, 101] method. The estimation is illustrated as follows:

$$\begin{aligned} m_{max} &= \underset{m}{argmax} \, dist(m, S_k) \\ n &= \underset{n \neq m_{max}}{argmax} \, dist(n, S_k) \\ C(E_i|S_k) &= dist(m_{max}, S_k) - dist(n, S_k) \end{aligned} \quad (4.27)$$

The $dist(m, S_k)$ is the distance of the predicted label m in such sentence S_k .

In each sentence, we need to estimate the total disagreement over the whole named entity concept set $\mathcal{N}$, which is measured by using the

sum of all disagreement for all named entity N_i .

$$D(Y|S_k) = \sum_{N_i \in \mathcal{N}} D(N_i|S_k) \quad (4.28)$$

The disagreement $D(N_i|S_k)$ is similarly estimated using the absolute value for the difference of the probability of the classifiers, which are the NER system and the informativity score system.

$$D(N_i|S_k) = |p_{Informativity}(N_i|S_k) - p_{NER}(N_i|S_k)| \quad (4.29)$$

The $p_{NER}(N_i|S_k)$ stands for the marginal probability of the Conditional Random Field (CRF) model in NER system, and $p_{Informativity}(N_i|S_k)$ is the probability using the informativity estimation method.

B. Entropy based active learning method

Entropy is a most commonly used measurement for uncertainty, which estimates the average information content for a variable. The document selection strategy using the entropy-based approach is illustrated as follows:

$$x_{Entropy}^* = \underset{x}{argmax} H_{\theta}(Y|x) \quad (4.30)$$

The $H_{\theta}(Y|x)$ is document entropy for x under the model θ and the Y denotes the whole label set. We use the summation of entropy values over the sentence S_k that are included in the document x .

$$H_{\theta}(Y|x) = \sum_{S_k \in x} H(Y|S_k) \quad (4.31)$$

For each sentence S_k , we use the bag of n-gram model, and we then calculate the value of $H(Y|S_k)$ using the averaged entropy for each n-gram NG_j in S_k , as follows:

$$H(Y|S_k) = \frac{1}{len(S_k)} \sum_{NG_j \in S_k} H(Y|NG_j) \quad (4.32)$$

The total entropy for the whole event concept set $\mathcal{E}$, which are defined in the ontology, is measured using the sum of entropy value

for all event E_i contained in the set. This is illustrated as follows:

$$H(Y|NG_j) = \sum_{E_i \in \mathcal{E}} H(E_i|NG_j) \quad (4.33)$$

$H(E_i|NG_j)$ is estimated by using the Weka package [102].

C. Gibbs error based active learning method

In [103], the Gibbs error criterion is proposed and shown to be an effective approach for active learning method, which selects documents whose corresponding Gibbs error is the greatest. This is illustrated as follows:

$$x_{Gibbs}^* = \underset{x}{argmax} H_{Gibbs}(\theta) \quad (4.34)$$

Similarly to the previous section for the entropy-based active learning method, we calculate the total Gibbs error for a document x as follows:

$$H_{Gibbs}(\theta) = \sum_{S_k \in x} H_{Gibbs}(Y|S_k) = \sum_{S_k \in x} \frac{1}{len(S_k)} \sum_{NG_j \in S_k} \sum_{E_i \in \mathcal{E}} H_{Gibbs}(E_i|NG_j) \quad (4.35)$$

For the calculation of $H_{Gibbs}(E_i|NG_j)$, we use the conditional probability of $p(E_i|NG_j)$, as defined in [103], and $p(E_i|NG_j)$ is calculated as shown in (4.9):

$$H_{Gibbs}(E_i|NG_j) = 1 - p(E_i|NG_j)^2 \quad (4.36)$$

4.3 Experiment

4.3.1 Experimental settings

In this chapter, we used the datasets from three BioNLP shared tasks (BioNLP-ST) 2013, namely GRO'13 (Gene Regulation Ontology) [97], CG'13 (Cancer Genetics) [104] and GE'13 (Genia Event Extraction) [105]. In the dataset, each document was manually annotated with an ontology that are defined in the task, which have different number of concepts and hierarchy from each other. We compare the datasets

and corresponding ontologies in Table 4.4. Note that as the manual annotation for the official test datasets in the CG'13 and GE'13 tasks are hidden to public, we instead use certain documents in the official training dataset as the 'test' datasets for evaluation.

Task	Corpus size (Dev/Train/Test)	Document type	No. event concepts	No. rela- tions
GRO'13	3300 (50/150/100)	MEDLINE abstract	507	10
CG'13	400 (100/200/100)	MEDLINE abstract	58	1
GE'13	20 (5/10/5)	PubMed Cen- tral full text	13	20

Table 4.4: Comparison of datasets and ontologies

In addition, we also used the dataset from the CoNLL-2003 task [30] and BioCreative II Gene Mention task [39], which are both the named entity recognition (NER) tasks. In these tasks, we deploy the state-of-the-art Stanford NER [37] system and Gimli gene name recognition system [98], respectively.

4.3.2 Evaluation metrics

To compare the performance of different methods, we plot the learning curve of the underlying system under different document selection strategies in each iteration. While the difference between some curves may not be obvious, we mainly use the *deficiency* [106, 107] as the evaluation metric, which is defined as follows:

$$Def_n(AL, REF) = \frac{\sum_{t=1}^n (acc_n(REF) - acc_t(AL))}{\sum_{t=1}^n (acc_n(REF) - acc_t(REF))} \quad (4.37)$$

The acc_t stands for the performance of the underlying system at t^{th} round inside the active learning loop. The AL denotes the abbreviation of the active learning method, and REF refers to the baseline method. The n refers to the total number of rounds, in this work, we set $n = 10$. For the deficiency value, a value smaller than 1.0 indicates that such active learning method is performing better than

the baseline method. In general, the smaller the deficiency value the better the document selection method.

4.3.3 Parameter optimization

We first take a parameter optimization step to determine the most appropriate parameters for the proposed methods.

4.3.3.1 Correlation measure and n-gram size

As mentioned previously, we intend to use three correlation measures to calculate the correlation between n-gram and event. We also need to determine the optimal value for the size of n-grams. To solve this two issues, we simulate an ontology concept prediction task at the sentence level, which is designed as follows: Given the input sentence S_i and its manual annotation that have N_i ontology concepts manually annotated, we predict the same number of N_i ontology concepts in S_i using the previously proposed method, and we then compare the predicted concepts with the manually annotation, calculating the accuracy of the prediction. We determine the optimal parameter settings based on the prediction accuracy.

Using 10-fold cross validation, the average prediction accuracy is calculated and shown in Table 4.5. Each column corresponds to a particular size of n-gram, and each row is one of the co-occurrence calculation methods. Note that when a larger size of n-gram is used, it do not include the smaller size of the n-grams. For instance, when $N=2$ or the bi-grams is used, the unigram is not used in the calculation. However, we do consider the mixture of different size n-grams, as denoted in the N value. Such as when $N=1-2$, both the unigrams and bi-grams are considered in the calculation. This experiment use the dataset from GRO'13 task.

Based on the results in Table 4.5, for all co-occurrence calculation methods, the accuracy decreases when the size of n-gram increases. This may due to the data sparseness problem for large size n-grams. From the result of parameter optimization, we choose to use **chi-square test** and **unigrams** for the following experiments.

Calculation Method	N-gram					
	N=1	N=2	N=3	N=4	N=5	N=1-2
chi-square	0.507	0.413	0.159	0.036	0.009	0.436
relative ratio	0.341	0.395	0.307	0.128	0.038	0.361
odds	0.420	0.395	0.274	0.117	0.035	0.407

Table 4.5: Parameter optimization results.

4.3.3.2 Parameter of incorporation of event participants

To determine the optimal value of δ in Equation (4.15), which is used to quantify the effects of integrating event participants into event concept prediction. We evaluate the active learning method in Equation (4.14) using the dataset of GRO'13 task, and set the δ values differently from 0.15 to 0.35 with an increment of 0.1. The results of deficiency are summarized in Table 4.6. Based on the results, we set $\delta = 0.25$.

Method	GRO'13
RS_Average	1
$\delta = 0.15$	0.716
$\delta = 0.25$	0.706
$\delta = 0.35$	0.713

Table 4.6: Parameter optimization for incorporation of event participants.

4.3.3.3 Parameter for dealing with OOV issue

To tackle the OOV issue, we use the top- k similar words for an unknown word, as in Formula (4.16). To determine the optimal value for k , we use the linear combination method in Equation (4.17) with the other parameters set to $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 0.8$ and varying the k value from 5 to 25, and evaluate the active learning method in the GRO'13 dataset. The deficiencies for each method are summarized in Table 4.7. Based on the result, we set $k=25$ for the remaining experiments.

Method	GRO'13
RS_Average	1
LC_ $(\alpha=0.1, \beta=0.1, \gamma=0.8), k=5$	0.611
LC_ $(\alpha=0.1, \beta=0.1, \gamma=0.8), k=10$	0.600
LC_ $(\alpha=0.1, \beta=0.1, \gamma=0.8), k=15$	0.617
LC_ $(\alpha=0.1, \beta=0.1, \gamma=0.8), k=20$	0.628
LC_ $(\alpha=0.1, \beta=0.1, \gamma=0.8), k=25$	0.563

Table 4.7: Parameter optimization results for dealing with OOV issue.

4.3.4 Evaluation of active learning method in event extraction task

4.3.4.1 Informativity based Active learning method

In the following experiments, we plot the learning curves and summarize the deficiencies of the TEES with different document selection approaches using datasets from GRO'13, CG'13 and GE'13 task. Note that in this section, for the active learning methods we only use the informativity based method, but exclude the additional features such as integration of event participants and dealing with OOV issue, which is presented in next section.

We compare the informativity based method with other document selection strategies, such as random selection, entropy-based [92], and Gibbs error [103] based, as well as the classical committee based active learning methods. We use the random selection as the baseline in the comparison of deficiency. Regarding to the experiment setting, we set the selection round to ten, and in each round, 10% of the unlabeled documents are selected for annotation, and are added to training data once annotated. Before the first round, the initial model of the TEES is separately trained only on the development dataset. Note also that the test data for each dataset is held-out and fixed. We compare the followings document selection strategies for the 10% documents that are selected in each round:

- Random selection: The training data are first randomly split into 10 sets, and one set is randomly chosen in the training phase. Using the 10-fold cross validation, we show the averaged performance of the learning curve when the documents are randomly selected (referred as RS_Average).

- Entropy-based active learning: The entropy of the document is calculated using the Equation (4.30), and the documents are ranked according to their entropy values and the top ones are chosen. (referred as AL(Entropy))
- Gibbs error based active learning: The Gibbs error of the document is estimated using the Equation (4.34), and the documents are ranked based on their Gibbs error values. The documents with the top value are selected.(referred as AL(GibbsError))
- Proposed active learning: Using either unigram or PAS, we evaluate the informativity based active learning method. Using either model, the resultant method is designated as AL(Informativity_Unigram) and AL(Informativity_PAS).
- Conventional committee-based active learning: The committee based active learning method is evaluated using the Equation (4.22). The confidence score for the prediction of events in the classifier of TEES are used in this method. Similar to the informativity based active learning method, we use either unigram or PAS for the calculation of informativity. The resultant method is referred as AL(ConventionalCommittee_Unigram) and AL(ConventionalCommittee_PAS), respectively.

We first apply those document selection methods to the dataset from GRO'13 task [97]. We report the learning curve of the TEES system with the incremental training data in each round. The deficiency comparison is summarized in Table 4.8. From the Table 4.8, we notice that the proposed informativity based active learning methods and the conventional committee-based approach have deficiency values of less than 1, while for the entropy and Gibbs error method, their corresponding values are higher than 1, this means that they do not perform better than random selection.

In Particular, the AL(Informativity_Unigram) method has the best deficiency of 0.760, while the corresponding method in AL(ConventionalCommittee_Unigram) achieves the value of 0.832, which is 8.65% worse than the informativity based method. On the other hand, while using the PAS model, the AL(Informativity_PAS) has the value of 0.845, which is 1.78% inferior than the counterpart of committee-based method, which achieve the deficiency value of 0.830.

Method	GRO'13	CG'13	GE'13
RS_Average	1	1	1
AL(Entropy)	1.017	1.226	0.854
AL(GibbsError)	1.039	0.993	0.850
AL(ConventionalCommittee_PAS)	0.830	0.589	0.439
AL(ConventionalCommittee_Unigram)	0.832	0.788	0.263
AL(Informativity_PAS)	0.845	0.581	0.872
AL(Informativity_Unigram)	0.760	0.768	0.139

Table 4.8: Deficiencies of document selection methods for event extraction.

By comparing the performance of PAS and unigram model, it is obvious that using the unigram model the informativity based active learning method have an 10.1% gain in the deficiency over the PAS based method, yet such change is not obvious in the committee-based method. This may suggest that the unigram model is especially effective for the proposed informativity method in the GRO'13 dataset. We then report the learning curves for TEES when using the different document selection methods in Figures 4.2 and 4.3. In Figure 4.3, we notice that, after 50% of the documents are selected, there is a consistent better performance for the AL(Informativity_Unigram) method than the other methods. This also justifies the result in the comparison of deficiency values.

While comparing the average number of instances per ontological concept from the task of GRO'13 and GE'13, we notice that the GRO'13 dataset have 13 instances per concept, yet the GE'13 have 82 instances per concept, as provided in [108]. This statistics suggests that such datasets as GRO'13, where the concept annotation may not be sufficient, the unigram based active learning method may perform better than using PAS model.

Nonetheless, the results with the GRO'13 dataset prove that the informativity based active learning using the unigram model can obtain superior efficiency than the conventional committee-based, the entropy based and the Gibbs error based active learning methods.

We then implement a similar experiment using the dataset from the CG'13 task. The deficiency for each method is summarized in the Table 4.8. From the Table 4.8, we find that deficiency for the Gibbs error based approach is less than 1, while the entropy based method

Figure 4.2: Comparison of various active learning method, including the PAS based informativity, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GRO'13.

Figure 4.3: Comparison of various active learning method, including the unigram informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GRO'13.

achieves the value of 1.226. In addition, the PAS models have generally better deficiency values than those of unigram models, when

comparing the PAS model and unigram model. Particularly, for the PAS and unigram models in the committee-based method, the percentage of the change is 25.3%. In the proposed informativity method, we notice a similar effect, as there is change of 24.3% in the deficiency. This may indicate that the PAS model may be more suited to the CG'13 task. Comparing the informativity based and committee-based active learning, the informativity method, however, shows superior performance over the committee-based method. Based on deficiency change, the improvements are 0.020 and 0.008 for PAS and unigram model, respectively. However, this is less obvious for the informativity method, which also suggests that the PAS feature may be more responsive than that of unigram model in the CG'13 dataset. Note that only a single relation of *Equiv* is defined in CG'13 task. *Equiv* is a transitive binary and symmetric relation for annotating entity mentions as being equivalent in the sense of referring to the same real-world entity [109]. Such relation is not present in the GRO'13 or GE'13 dataset. The better performance in the PAS model may suggest that it is more effective for identification of equivalent entity mentions than that of the unigram model. The learning curves for those methods are charted in Figure 4.4 and 4.5.

Figure 4.4: Comparison of various active learning method, including the PAS informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in CG'13.

Figure 4.5: Comparison of various active learning method, including the unigram informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in CG'13 .

We also evaluate the active learning methods using the GE'13 dataset, and we analyze the deficiency of the methods in the Table 4.8. In Table 4.8, all the methods are performing better than that of the random selection. Concretely, the deficiency value for the Gibbs error based approach is 0.850, while the entropy method achieve deficiency value of 0.854. Note that using unigram model, the informativity based active learning methods performs superiorly over that of PAS model. For instance, in the committee-based method using unigram, there is an gain of 40.1% over committee-based method using the PAS model. This indicates that the unigram feature is more proper than that of the PAS feature for informativity based active learning in GE'13 dataset. Such effect is also evident in the informativity method. As the deficiency value for the best performing AL(Informativity_Unigram) is 0.139, while the corresponding committee-based active learning using the unigram model has the deficiency of 0.263. The learning curves are charted in Figures 4.6 and 4.7.

In the Figure 4.7, the unigram based active learning method shows superior performance over the baseline of random selection method,

Figure 4.6: Comparison of various active learning method, including the PAS informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GE'13 .

Figure 4.7: Comparison of various active learning method, including the unigram informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection in GE'13.

yet the PAS model show less evident gain over the baseline. One reason for this may be that the ontology defined in GE'13 is less complicated than the other two tasks. In addition, the document annotation in the GE'13 may be rich, as the average number of annotation instances per concept in GE'13 is 82, around six times more than

GRO'13 [108]. Given dataset with less complicated ontological concepts and rich document annotation, the unigram model may show evident enhancement in active learning.

4.3.4.2 Active learning with additional components

Incorporation of event participants

In this section, we evaluate how the integration of recognition of gene/protein names affects the active learning method, which is illustrated in Formula (4.14). We report the learning curve for the TEES system using the active learning method that is either integrated with the gene/protein names or not. We carry out the experiment with the GRO'13 dataset. The experiment results are charted in Figure 4.8 and the deficiency values are summarized in the Table 4.9.

Figure 4.8: Integration of NER into active learning with PAS and n-grams models.

In the Table 4.9, we notice that there is positive effects for the integration of recognition for gene/protein names towards the active learning method, and this effect holds for both of bag of n-gram or PAS model. Particularly, using the recognition for gene/protein names, the deficiency value for PAS based active learning is further

enhanced from 0.845 to 0.589, which is a 30.3% improvement. In the unigram model of the informativity method, the improvement is rather less noticeable of 7.1%, which may indicate that some recognized named entities are redundantly identified in the n-grams model.

Method	GRO'13
RS_Average	1
AL(ConventionalCommittee_PAS)	0.830
AL(ConventionalCommittee_PAS + NE)	0.693
AL(Informativity_PAS)	0.845
AL(Informativity_PAS + NE)	0.589
AL(Informativity_Unigram)	0.760
AL(Informativity_Unigram + NE)	0.706

Table 4.9: Deficiencies of active learning methods with and without integrating the prediction of named entities (NE) against GRO'13 dataset

In addition, there is similar improvement of 16.5% in the conventional committee-based method after the combination of event participants into the part of statistical informativity estimation, as the deficiency value is reduced from 0.830 to 0.693. However, this change is less noticeable than that in the informativity based method, which may suggest that the confidence scores provided in the TEES, which is adopted in the conventional committee-based method, hampers the effects of event participants.

Solving OOV issue with word similarity

In using the n-gram model, there is the issue of Out-of-Vocabulary (OOV) words that only exist in the testing data, not the training data. We solve this by using the word2vec system to identify top- k words that are close to a given OOV word in the test data, and then use their weights to determine the weight of the OOV word. The results of using the word vector are plotted in Figure 4.9 using the GRO'13 dataset, and the deficiency is summarized in Table 4.10.

Method	GRO'13
RS_Average	1
AL(Informativity_Unigram)	0.790
AL(Informativity_WordVec)	0.769

Table 4.10: Deficiencies of using word vector toolkit in the active learning

Note that such OOV word handling experiments are executed only for events but not for relations, as we observed that the relations of

Figure 4.9: Evaluation of integration of word vector into active learning.

the BioNLP-ST'13 tasks are little affected by the OOV issue, as the relations are not associated with trigger words, as explained in the introduction part. Using the word similarity, we notice that the performance for the n-gram model is further enhanced, as the deficiency is reduced from 0.790 to 0.769, an 2.66% change. Based on the rather less noticeable enhancement, we conclude that such OOV issue may be rather not ubiquitous in the GRO'13 dataset.

4.3.5 Linear combination of n-gram and PAS

Lastly, we evaluate the effects of linearly combination for the n-gram and PAS feature that are adopted in the active learning method, as expressed in Equation (4.17), in order to determine which method proposed in this work are more effective towards the overall performance.

In the experiment, we set four groups of parameter combinations, which is $(\alpha=0.8, \beta=0.1, \gamma=0.1)$, $(\alpha=0.1, \beta=0.8, \gamma=0.1)$, and $(\alpha=0.1, \beta=0.1, \gamma=0.8)$, and the equal weight $(\alpha=0.33, \beta=0.33, \gamma=0.33)$. As the AL(Informativity_PAS + NE) is the best performing method in previous experiments, it is used as the benchmark in the comparison here. Note that the

AL(Informativity_PAS + NE) is in essence the parameter combination for $(\alpha=0, \beta=1, \gamma=1)$. We also only using the named entity for the active learning as the benchmark, which is the parameter combination of $(\alpha=0, \beta=0, \gamma=1)$. This is to determine if merely using the total counts be effective enough for the active learning method.

We plot the results in Figure 4.10, and the deficiency values are summarized in Table 4.11. In the Table 4.11, the group of $(\alpha=0.1, \beta=0.1, \gamma=0.8)$ shows the best performance and achieve the deficiency value of 0.563. In addition, while comparing to PAS and unigram-based method, we notice that by incorporating the event participants, the performance of the active learning is mostly improved. However, merely incorporating the event participants, which is $(\alpha=0, \beta=0, \gamma=1)$, achieves the value of 0.583, which is worse than the minimum value. This suggests that the PAS or n-gram model are complementary to the additional component, i.e., the event participants model.

Method	GRO'13
RS_Average	1
AL(Informativity_PAS + NE)	0.589
LC_ $(\alpha=0.33, \beta=0.33, \gamma=0.33)$	0.740
LC_ $(\alpha=0.8, \beta=0.1, \gamma=0.1)$	0.772
LC_ $(\alpha=0.1, \beta=0.8, \gamma=0.1)$	0.752
LC_ $(\alpha=0.1, \beta=0.1, \gamma=0.8)$	0.563
LC_ $(\alpha=0, \beta=0, \gamma=1)$	0.583

Table 4.11: Deficiencies of linear combination for active learning components

4.3.6 Evaluation of active learning for NER task

We also evaluate the performance of the proposed active learning method in NER task, which is shown in Equation (4.18). We follow a similar experiment design as in the previous part. In the experiment, each document selection method starts with the same development dataset that are annotated for model initialization, and a pool of unlabeled training dataset that are used for selection. In each iteration round, 10% of the unlabeled documents are selected, guided by each document selection strategy. In the evaluation, we present the learning curve for the NER system that are trained with the selected document in each round, the held-out test dataset are the same as the

Figure 4.10: Evaluation of linear combination for active learning components.

official evaluation dataset and the evaluation steps follow that in official procedure .

The sample selection strategies are as follows:

- Random selection: The training data are first randomly split into 10 sets, and one set is randomly chosen in the training phase. Using the 10-fold cross validation(CV), we show the averaged performance of the learning curve when the documents are randomly selected. (hereafter referred to as RS_Average)
- Entropy-based active learning: The entropy of the document is calculated, and the documents are ranked according to their entropy values and the top ones are chosen . (designated as AL(Entropy))
- Gibbs Error based active learning: The Gibbs error, proposed in [103], of the document is estimated and the documents are ranked based on their Gibbs error values. The documents with the top value are selected.(designated as AL(GibbsError))
- Proposed active learning method using informativity scoring only: We use the informativity based active learning described

in Equation (4.18), and then selects the top documents based on their informativity scores. (designated as AL(Informativity))

- Conventional committee-based active learning: We adopt the committee based method that is illustrated in (4.22), particularly we use the confidence score in the NER system. This method is referred as AL(ConventionalCommittee).

We evaluate these document selection strategies using the BioCreative dataset and plotted the performance of Gimli in Figure 4.11, and and Table 4.12 summarize the deficiency values for each document selection method.

Method	BioCreative	CoNLL
RS_Average	1	1
AL(Entropy)	1.171	0.737
AL(GibbsError)	1.045	0.885
AL(ConventionalCommittee)	0.684	0.763
AL(Informativity)	0.514	0.575

Table 4.12: Deficiencies of sample selection methods against the BioCreative and CoNLL datasets

Figure 4.11: Comparison of active learning with informativity based, entropy-based, Gibbs error based, and conventional committee based method, and random selection against BioCreative dataset.

In Figure 4.11, we notice that the proposed active learning method perform stably better than other methods. Based on the deficiency value in Table 4.12, the deficiency value for informativity method is 0.514, while the conventional committee based method have the deficiency of 0.684.

We also evaluate the document selection methods using the CoNLL dataset, and plot the learning curves of the Stanford NER system in Figure 4.12, and compare the deficiency value for each method in Table 4.12.

Figure 4.12: Comparison of active learning methods and random selection.

We notice that the proposed informativity based active learning method outperforms the other methods in Figure 4.12. Note also that in the comparison of deficiency value, the proposed method achieves 0.575, which is around 42% better than the baseline of random selection. On the other hand, the Entropy and Gibbs error based methods also have deficiency value of less than 1, yet the change is 26% and 11%, respectively. Note that the conventional committee based method have deficiency value of 0.763. The results in these two datasets prove that the proposed informativity based active learning can show superior results than the conventional committee-based method, as well as the Entropy and Gibbs error based approaches.

4.4 Summary

In this work, we proposed a novel active learning method for event extraction task, which considers much more complex biomedical events than the previous PPI extraction task. Our active learning method selects documents based on the measurement of the collective ‘informativity’ for the unlabeled documents, which is also to estimate the potential possibility of events that are non-identifiable by event extraction systems. We evaluated the active learning method using datasets in the BioNLP Shared Tasks, and the experiment results proved that based on our method, the event extraction system can be more efficiently trained than other previous methods, including entropy based, Gibbs error based and conventional committee-based methods. In addition, we have incorporated named entity recognition into the active learning method for event extraction, and by handling the OOV word, we further improved the learning efficiency in the system. Finally, we evaluated the active learning method in NER tasks and the experiment results show that the proposed method is also effective in the document selection for the task.

Chapter 5

Clustering Based Active Learning for Biomedical Named Entity Recognition

5.1 Introduction

In Chapters 3 and 4, we have proposed the informativity score based active learning system, and have also explored the committee-based active learning approach in the biomedical event extraction. In these methods, the underlying principle are somehow similar: to select the samples that are most ‘informative’, according to certain measurement, to the underlying machine learning system. Despite the success of this principle, however, there is one limitation. It fail to consider if the selected samples are ‘representative’ or not, where non-representative samples such as outliers may cause overfitting and undermine the training performance of the supervised learning framework.

To select ‘representative’ examples and avoid outliers, there are other existing approaches. For instance, there is work of using the information density to guide the selection of documents [110]. This approach is to explicitly model the distribution of the data set with density weights, and is shown to outperform the uncertainty-based approaches. The informativeness of a sample is weighted by its average similarity to all other samples. This method was designed to avoid spending too much time on annotating outliers, which have low density. On the other hand, there are also active learning methods that integrate the unsupervised clustering techniques [111, 106, 112]. For

Figure 5.1: Toy example demonstrating the selection strategies of density based and clustering based active learning method.

Circles, triangles and squares denote the unlabeled samples, classified into different clusters. The crosses denote the supposed cluster centroid. Solid circles, triangles and squares denotes the selected samples for annotation. (a) Information Density approach selects documents that are within the dense areas. (b) Previous clustering method selects documents close to the cluster centroid. (c) Proposed method not only selects documents close to centroid, but also selects documents close to cluster boundary.

instance, in [111] they first cluster all samples before active learning and select the documents close to the cluster centroids as the initial set of annotation, instead of beginning with a random subset of samples. In [106] they further develop this idea by incorporating the information density to guide the document selection, through maintaining a k -nearest-neighbor-based density weight, in order to circumvent the risk of selecting outliers.

In this chapter, we propose a novel clustering based active learning method, where we integrate the two approaches by selecting a mixture of both ‘representative’ and ‘informative’ documents from clustering results. Rather than merely selecting the cluster centroids for annotation, we select examples that are either cluster centroid or on cluster boundary, which is illustrated in Figure 5.1. Different from

[106], we consider the distance difference with respect to the centroids, rather than using the density weight.

Previous work of text clustering often adopt bag-of-word model, in which a document is represented as the bag of its words, discarding other properties of text, such as word order and grammatical structure. Term frequency-inverse document frequency (TF-IDF) is then used to identify word weights (or features) in each document and represent a document as a vector based on the word features [113]. In this work of active learning for NER, we group documents that contain similar named entities, on the assumption that the documents that share similar named entities may have common features of the underlying classifier, thus redundant for the manual annotation work. In our active learning method, we use a matrix factorization method [114] for the vector representation of named entities in order to measure the similarity of named entities.

The cluster analysis is an unsupervised task which is to group a set of data in such a way that the individual data in the same cluster are more similar to each other than data that belong to different clusters. The clustering has a long history and has been developed in domains such as statistics, pattern recognition, data mining, and other fields [115]. A large number of clustering techniques have been proposed, such as hierarchical clustering, centroid-based clustering, as well as density-based clustering [116, 117].

The hierarchical clustering seeks to build a hierarchy of clusters, and the strategies may be divided into agglomerative methods, which are a series of successive fusions of individual objects into clusters, and divisive methods, which partition the set of objects successively into finer groups. We use hierarchical agglomerative clustering (HAC) in our analysis, as implemented in ELKI toolkit [118]. In the K-means clustering of centroid-based clustering, the general idea is to find the k cluster centers and assign the objects to their nearest cluster center, such that the distances between the objects and cluster centers are minimized. While the K-means clustering is an NP-hard problem, there is work to optimize the speed of execution [119]. While in density-based clustering, the clusters are defined as areas of higher density than the remainder of the data set [120]. The popular density-based clustering methods are DBSCAN [121] and Shared Nearest

Neighbor (SNN) clustering [122], as well as Greedy Variance Minimization (GVM) ¹. In this work, we also compare the variants of proposed method using different clustering methods.

5.2 Method

Our method measures the informativity of unlabeled documents using a state-of-the-art NER system for the dataset. It works iteratively as follows: 1) We train the NER system based on initial training data. 2) We apply the trained NER system to all the unlabeled documents to identify candidate named entities (NEs). 3) We obtain the vector representation of each candidate NE by adopting a matrix factorization method [114] and 4) cluster the unlabeled documents with the vector representation of candidate NEs. 5) We select the mixture of most ‘representative’ and ‘uncertain’ documents in each cluster for the manual annotation and 6) update the NER system with the annotated documents.

Table 5.1 illustrates the proposed method in pseudo codes. In the following subsections, we discuss three questions: 1) how to represent documents, 2) which documents to select for manual annotation among the results of document clustering, and 3) which clustering method to use. Note that the three questions are related to the steps (4) and (5) in the previous paragraph and that the implementation details of the other steps can be found in Section 5.3.1.

¹<http://www.tomgibara.com/clustering/fast-spatial/>

Input: labeled document pool L , unlabeled document pool U , batch size b

// Initialization

NE_0 = the set of named entities annotated on L

Learn an NER model M_0 from NE_0

$i = 0$ // the index of the current round

// Active Learning Loop

while U is not empty:

$i += 1$

// 1. Data Representation

for each document D_{ij} in U :

 Apply M_{i-1} to D_{ij} and collect the resultant named entities set $NE_{D_{ij}}$

 Construct the vector representation $\vec{D}_{ij}$ using $NE_{D_{ij}}$, (Equations (5.1,5.3))

// 2. Clustering

 Cluster all $\vec{D}_{ij}$ of U into $\frac{b}{2}$ clusters.

// 3. Document Selection

 Based on clustering results, select the b documents, (Equations (5.4,5.5)), designated as B , for annotation by annotator.

 Remove B from U , add B to L , and add the annotations on B to NE_{i-1} , designated as NE_i

 Learn a new model M_i from NE_i

Table 5.1: Proposed clustering based active learning method

5.2.1 Vector representation model for documents

We group documents that contain similar named entities into the same class to reduce the manual annotation work. We represent the documents in the perspective of its contained named entities, in the form of vectors. For this purpose, we utilize GloVe [114], which analyzes latent semantics of words by using global matrix factorization and local context window methods. We train a word vector model with the entire MEDLINE corpus for the BioCreative task [123].

We estimate the vector of a document as the average of vectors of named entities the document contains. Equation (5.1) shows how to represent a document (D) as a vector with the named entities (n_i) found in the document, where $|NE(D)|$ indicates the number of named entities found in document D . Note that $\vec{D}$ is the vector representation of document D .

$$\vec{D} = \frac{1}{|NE(D)|} \sum_{\forall n_i \in D} \vec{n}_i, \text{ if } |NE(D)| > 0 \quad (5.1)$$

For multi-word terms as named entities, we simplify the estimation of their vectors as follows, where $\text{len}(n_i)$ indicates the number of words in the named entity n_i , and w_k indicates the k th word in n_i , and $\vec{n}_i$ is the vector representation of named entity n_i .

$$\vec{n}_i = \frac{1}{\text{len}(n_i)} \sum_{\forall w_k \in n_i} \vec{w}_k \quad (5.2)$$

For documents that no NE is identified by the NER system (i.e., $|NE(D)|=0$), we use the average of vector of word contained in the document, as shown in Equation (5.3). The w_j indicates the j th word in document, and $w(D)$ indicates the number of words found in document D , the $\vec{w}_j$ is the vector representation of word w_j .

$$\vec{D} = \frac{1}{|w(D)|} \sum_{\forall w_j \in D} \vec{w}_j, \text{ if } |NE(D)| = 0 \quad (5.3)$$

5.2.2 Document selection strategy from clustering result

The main issue in active learning is the selection of documents for manual annotation. In clustering based active learning, there is a question of how to select the documents based on document clustering result. In this work, we select both ‘representative’ documents and ‘informative’ documents as mentioned.

By ‘representative’, we follow the previous work [106], which is to select documents that are close to the cluster centroid, that is, the document’s distance from the cluster centroid is the minimum among those distances of the other documents that are within the same cluster, as expressed in Equation (5.4). The d indicates the document that belongs to the cluster C , and x_C is the centroid for cluster C . $dist$ indicates the distance between the document and the centroid. The $\tilde{d}_{representative}$ stands for the resultant ‘representative’ document that is selected.

$$\tilde{d}_{representative} = \underset{d}{argmin} dist(d, x_C) \quad \forall d \in C \quad (5.4)$$

By ‘informative’, we mean the documents that are close to the boundary of the cluster, as its distance from its cluster centroid is maximized, comparing to other documents from the same cluster. Note that such documents could also be outliers. To circumvent of selecting outliers, we further require that such documents are close to the cluster’s boundary with other clusters, not to the boundary with no neighbor cluster. In this way, we choose the document whose distance to the cluster centroid is maximized, yet the difference between its distances to the cluster centroid and the neighbor cluster centroid is minimized. We consider that the annotations of such documents are ‘informative’ for the clustering, and also the training of the underlying system. These selection criteria are expressed in Equation (5.5). $\tilde{C}'$ denotes a neighboring cluster for C , and $x_{\tilde{C}'}$ is the centroid for cluster $\tilde{C}'$. The $\tilde{d}_{informative}$ stands for the chosen document that meets both of the requirements. In cases when no document is found to meet the conditions, the selection of ‘informative’ document is skipped.

$$\tilde{d}_{informative} = \begin{cases} \underset{d}{\operatorname{argmax}} \operatorname{dist}(d, x_C) & \forall d \in C \\ \underset{d}{\operatorname{argmin}} (\operatorname{dist}(d, x_{\tilde{C}'}) - \operatorname{dist}(d, x_C)) & \forall d \in C \end{cases} \quad (5.5)$$

To find the neighboring cluster $\tilde{C}'$, we calculate the distances between the cluster centroids, which is expressed in Equation (5.6).

$$\tilde{C}' = \underset{C'}{\operatorname{argmin}} \operatorname{dist}(x_{C'}, x_C) \quad (5.6)$$

We select a mixture of documents from each cluster that meet the requirements of Equations (5.4-5.5), one using Equation (5.4) and the other using Equation (5.5).

5.2.3 Clustering method for active learning

In this work, we propose a clustering based active learning method. Then there is the question of whether the different clustering results from the clustering methods would affect the performance of active learning methods, as well as to determine which clustering approaches is appropriate for the proposed method. We adopted different clustering methods, including the HAC [115], DBSCAN [121], SNN [122] and GVM for the analysis.

5.3 Experiment

5.3.1 Datasets and NER systems

In this work, we use the dataset from the gene mention recognition task of BioCreative II [123], and we utilize the state-of-the-art NER system, Gimli [98], for training and evaluating our method. We also use the clustering methods, implemented in ELKI toolkit [118], to compare their effects for active learning.

Methods for comparison

Each experiment starts with a held-out labeled development dataset for initialization and a pool of unlabeled training dataset for selection, and continues with ten rounds. In each round, 10% of the documents in the training dataset are selected by different sample selection strategies. For evaluation, we report the performance of the Gimli trained with the selected training document in each round, against the held-out test dataset in official evaluation procedure.

The sample selection strategies are as follows:

- Random selection: We randomly split the training dataset into 10 bins in advance, and one bin is randomly chosen in each round. Following 10-fold cross validation, we report the averaged performance. (hereafter referred to as Random Selection)
- Entropy-based active learning: We follow the uncertainty-based active learning method, and use the entropy as the selection criteria. We calculate the entropy of the documents, and select the documents that have top entropy values. The calculation of entropy follows the sequence entropy, proposed in [110], and is designated as AL(Entropy).
- Maximum Gibbs Error based active learning: Similar to the entropy-based method, we select documents that have top values in the Gibbs error, as introduced in [103], designated as AL(Gibbs).
- Cluster and Density based active learning: As proposed in [106], we use clustering to select the initial documents and then select documents based on document density for sub-sequential rounds. This is designated as AL(Cluster+Density).
- Density based active learning: We simulate the selection process that is purely based on document density, what is, without first selecting the initial document set using clustering approach. This is a variant of the work in [106], and is designated as AL(Density).
- Proposed clustering based active learning method: We use the proposed clustering based active learning method, introduced

in Section 5.2. Note that this includes the experiment design for the three subsequent questions, which is to use the vector representation model for document, and use the proposed document selection strategy from clustering results, as well as the SNN clustering method. This is designated as AL(proposed).

5.3.2 Evaluation metrics

To compare the performance of the different strategies of sample selection, we plot their performance after each of the ten rounds of iteration. Since the difference between some plots is not obvious, however, we also use the evaluation metric of *deficiency* [106, 124], defined as follows:

$$Def_n(AL, REF) = \frac{\sum_{t=1}^n (acc_n(REF) - acc_t(AL))}{\sum_{t=1}^n (acc_n(REF) - acc_t(REF))} \quad (5.7)$$

, where acc_t is the performance of the underlying classifier at t^{th} round of learning iteration. AL is an active learning method and REF is the baseline method (i.e. RS_Average). n refers to the total number of rounds (i.e. 10). Deficiency value smaller than 1.0 means that the active learning method is superior to the baseline method, and vice versa: In general, a smaller value indicates a better method.

5.3.3 Results

We applied these methods to the BioCreative dataset, plotted the learning curve of Gimli in Figure 5.2, and summarized their deficiencies in Table 5.2. As shown in Figure 5.2, our clustering method AL(proposed) outperforms the previous clustering method AL(Cluster+Density) of [106], always showing steady improvement. In contrast, AL(Cluster+Density) falters when 20% and 40% of samples are selected, which may mean that its density-based method requires significant amount of samples for further improvement of the underlying classifier. This is also shown in the AL(Density) when 10% and 40% of the documents are selected, where its learning curve is increasing slower than the rest of methods, which may be due to the lack of training data for this

method. The curves of the generic uncertainty based methods (e.g. entropy, Gibbs error) do not show significant improvement over the random selection.

Figure 5.2: Comparison of active learning methods and random selection with BioCreative dataset

In terms of deficiency, the AL(proposed) method gains a 36.3% improvement compared to the baseline, as shown in Table 5.2. Note that the previous method AL(Density) shows higher deficiency than the baseline.

Table 5.2: Comparison of deficiencies of active learning methods and random selection

Method	Deficiency
Random Selection	1
AL(Entropy)	0.972
AL(Gibbs)	0.954
AL(Density)	1.194
AL_Cluster+Density	0.968
AL(proposed)	0.637

5.3.3.1 Vector representation model for documents

In this part, we compare the effect of the vector representation in the proposed clustering based active learning method. Specifically,

we compare the learning curves of the active learning method using different data representation models.

The compared document representation models are as follows:

- Vector representation of candidate NE: As proposed in Equations (5.1), (5.2) and (5.3), we use the vector representation of the recognized NEs in the document, designated as AL(VectorData_NE).
- Vector representation without prediction of candidate NE: Not considering the NEs recognized in the document, we only use the vector representation of all the words in the document (i.e. only using Equation (5.3)). This is designated as AL(VectorData_w/o NE).
- Bag-of-word model for documents: We follow the traditional bag-of-word model, which is to first create a feature set using the TF-IDF approach, and represent the documents using the feature set as vectors, which is designated as AL(Bag-of-Word).
- Random selection: This is the same as in the previous experiment, and is designated as Random Selection.

Considering the time efficiency, we carried out this experiment using the GVM clustering approach, and we always selected the first document in the clustering result, rather than selecting the documents close to the centroid or to the boundaries of clusters. The learning curves are plotted in Figure 5.3, and the deficiencies are summarized in Table 5.3.

As shown in Figure 5.3, the learning curves using vector representation model are showing better performance than that of Bag-of-Word model. In particular, by using only the predicted NE rather than averaging over the whole sentence, the performance of the learning curve is further improved, which suggests that the process of prediction of NE helps to filter out redundant information in the sentence. Considering that in this experiment only the first document in the clustering result is selected, the learning curve is not showing steady improvement and affects from sudden drops.

In terms of deficiency, the AL(VectorData_NE) method gains a 18.6% improvement compared to the baseline, as shown in Table 5.3, while all the other methods show better deficiencies than the baseline.

Figure 5.3: Comparison of the performance of active learning methods with different representation models for documents, as well as the baseline of random selection. The vector representation (VectorData) is used either with or without prediction of candidate NE (NE). The performance of active learning using classical bag-of-word model is also presented.

Table 5.3: Comparison of deficiencies of active learning methods with different data representation models for documents

Method	Deficiency
Random Selection	1
AL(Bag-of-Word)	0.929
AL(VectorData_w/o NE)	0.923
AL(VectorData_NE)	0.814

5.3.3.2 Document selection strategy from clustering result

We compare the different document selection strategies from clustering result of active learning method. Concretely, we compare the following document selections strategies:

- First document in clustering result: We consider the speed efficiency and select the first document in the clustering result, skipping the calculation load for finding centroid. This is designated as AL(First in Cluster).
- Document closest to centroid: We follow the traditional approach that select the document closest to the cluster centroid,

i.e., the documents that satisfy the Equation (5.4), which is designated as AL(ClusterCentroid).

- Proposed document selection strategy: We select the documents that meet the requirement in Equation (5.4) or Equation (5.5), and is designated as AL(ClusterCentroid+Edge).
- Document Scoring strategy: We select the document that has the top score in a classifier-independent scoring measurement, which is introduced in [125].
- Random selection: This is the same as in the previous experiment, and is designated as Random Selection.

We carried out this experiment using the GVM clustering approach, and we used the vector representation with the predicted NE, as introduced in Equation (5.1), Equation (5.2) and Equation (5.3). The learning curves are plotted in Figure 5.4, and the deficiencies are summarized in Table 5.4.

Figure 5.4: Comparison of the performance of active learning methods with different document selection strategies, as well as the baseline of random selection.

As shown in Figure 5.4, the learning curves of the methods that select document closest to cluster centroid are better than that of random selection, while the proposed method performs even better. This becomes evident when about 30% to 70% of the documents are selected for manual annotation, where the learning curve of selecting only

the documents close to centroid does not improve significantly. The strategy of selecting the first document in clustering results becomes less stable and drops when 60% of the documents are selected. In contrast, the selection strategy based on a scoring system performs worse than the random selection, which suggests that the scoring system may not be compatible with the clustering approach.

Table 5.4: Comparison of deficiencies of active learning methods with different document selection strategies

Method	Deficiency
Random Selection	1
AL(First in Cluster)	0.814
AL(DocumentScore)	1.241
AL(ClusterCentroid)	0.886
AL(ClusterCentroid+Edge)	0.787

In terms of deficiency, the AL(ClusterCentroid+Edge) method gains a 21.3% improvement compared to the baseline, as shown in Table 5.4, while the rest of the selection methods achieve better deficiencies than the baseline. The only exception is the scoring method, which shows worse deficiency than the random selection.

5.3.3.3 Clustering method for active learning

We compare how the different clustering methods affect the learning curves of active learning method. Particularly, we compared the clustering method of HAC, DBSCAN, SNN and GVM.

In this experiment, we use the vector representation of NE, introduced in Equation (5.1), Equation (5.2) and Equation (5.3), and we select documents using the strategy introduced in Equation (5.4) and Equation (5.5). We plot the learning curves in Figure 5.5 and compare the deficiencies in Table 5.5.

In Figure 5.5, the learning curve of HAC method is performing worse than that of random selection, while the rest method shows better performance. Particularly, the SNN approach shows the best performance for this experiment. One possible explanation for the robust performance of SNN is that such a method may better handle data of heterogeneous densities [122].

Figure 5.5: Comparison of the performance of active learning methods with different clustering methods, as well as the baseline of random selection. For the clustering methods, the HAC, GVM, DVSCAN and SNN are considered.

Table 5.5: Comparison of deficiencies of active learning methods with different clustering methods

Method	Deficiency
Random Selection	1
AL(HAC)	1.194
AL(GVM)	0.787
AL(DBSCAN)	0.861
AL(SNN)	0.637

As shown in Table 5.5, the SNN method shows a 36.3% improvement compared to the baseline, while only the HAC method shows worse deficiency than that of random selection.

5.4 Summary

In this chapter, we propose a novel clustering based active learning method to select the documents that are ‘informative’ and ‘representative’. Comparing to the previous clustering based active learning method, we show that the clustering of named entities surpass the benchmark of entropy, Gibbs error, and density based active learning methods, as well as the random selection. In addition, in terms

of the deficiency gain, the proposed method can achieve a deficiency gain of 36.3% over the random selection. We also show that the performance of proposed active learning method is improved by using the vector representation of named entities in the documents, as well as adopting the SNN clustering approach.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Biomedical text mining may target various kinds of valuable information embedded in the literature, but a critical obstacle to the extension of the mining targets is the cost of manual construction of labeled data, which are required for state-of-the-art supervised learning systems. Active learning is to choose the most informative documents for the supervised learning in order to reduce the amount of required manual annotations.

In this work, we proposed a novel informativity score based document scoring system, following the uncertainty based active learning framework. Our method measured the collective informativity for biological events and relations expressed in documents. We found that with the proposed method using the predicate-argument structure and the ontology hierarchy, the underlying event extraction system reached the plateau of saturated performance with significantly fewer documents, and this reduced the amount of needed documents for manual annotation.

Built upon the informativity score based system, we then followed the committee-based active learning framework by incorporating the information from external event and relation extraction system. We applied the committee-based active learning to the event and relation task, and further improved the performance of the active learning method by integrating information from named entities classifier into the committee of event and relation extraction system. Such integration approach was proved to be effective.

As in the previous uncertainty based, as well as the committee-based, active learning method, the core idea is to select documents that are most ‘informative’. However, there is one limitation of failing to consider whether the selected documents are ‘representative’ or not, where non-representative samples such as outliers may cause overfitting and undermine the learning process of underlying systems. We therefore deploy a clustering based active learning method, which takes into account both the ‘informativeness’ and ‘representativeness’ of the selected documents. This strategy is successful in reducing the needed annotation work for the NER task, as the clustering based active learning method achieved a deficiency gain of 36.3% over the random selection.

Based the results of these experiments, we showed that by using the active learning method, the learning efficiency of the perspective supervised learning system could be greatly improved, which also reduces the needed manual corpus annotation work.

6.2 Future Work

In the future, our work can be extended in several directions.

6.2.1 Ensemble clustering based active learning

We have proposed the clustering based active learning and apply it into the NER task. In fact, the selection of adopted clustering methods are open to question. As each clustering method provide one separation of the documents, it is natural to consider the ensemble clustering based method, in which multiple clustering methods can be adopted in the categorization of the documents.

6.2.2 Real world evaluation for active learning

In this thesis, we have evaluated the active learning method using the dataset from the several shared tasks, which is pre-annotated disregarding the active learning loop. We simulated the active learning loop without considering the practical time needed for documents

annotation. However, to apply the active learning method to real world problem, those document annotation steps can not be ignored. Therefore, it is critical to include those steps into the evaluation in order to thoroughly understand the strength and limitation of the active learning methods.

Publication

1. Xu Han, Jung-jae Kim, "Active learning for ontological event extraction incorporating named entity recognition and unknown word handling", *Journal of Biomedical Semantics*, 2016 Apr. doi: 10.1186/s13326-016-0059-z
2. Xu Han, Chee-Keong Kwoh and Jung-jae Kim, "Clustering Based Active Learning for Biomedical Named Entity Recognition", *IEEE WCCI* 2016.
3. Xu Han, Jung-jae Kim, "Active learning for ontological event extraction", *SMBM*, pp. 45-51, 2014.
4. Jin-Dong Kim, Jung-jae Kim, Xu Han and Dietrich Rebholz-Schuhmann, "Extending the evaluation of Genia Event task toward knowledge base construction and comparison to Gene Regulation Ontology task", *BMC Bioinformatics*
5. Jung-jae Kim, Xu Han, Vivian Lee, Dietrich Rebholz-Schuhmann, "GRO Task: Populating the Gene Regulation Ontology with events and relations", *Proc Workshop BioNLP Shared Task 2013, ACL 2013*, pp. 50-57, 2013.
6. Jung-Jae Kim, Xu Han and Watson Wei Khong Chua, "Annotation of biomedical text with Gene Regulation Ontology: Towards Semantic Web for biomedical literature", *Proceedings of LBM 2011*, pp. 63-70.

References

- [1] L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in biomedical text," *Bioinformatics*, vol. 18, no. 8, pp. 1124–1132, 2002.
- [2] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, "Text mining and ontologies in biomedicine: Making sense of raw text," *Briefings in Bioinformatics*, vol. 6, no. 3, pp. 239–251, 2005.
- [3] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.
- [4] R. Stevens, C. A. Goble, and S. Bechhofer, "Ontology-based knowledge representation for bioinformatics," *Briefings in Bioinformatics*, vol. 1, no. 4, pp. 398–414, 2000.
- [5] M. Bada and L. Hunter, "Desiderata for ontologies to be used in semantic annotation of biomedical documents," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 94–101, 2011.
- [6] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a Robust Part-of-Speech Tagger for Biomedical Text," in *Advances in Informatics*, P. Bozanis and E. Houstis, Eds. Springer Berlin / Heidelberg, 2005, vol. 3746, pp. 382–392.
- [7] P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction," *BMC Bioinformatics*, vol. 10, no. 1, p. 349, 2009.
- [8] S. Jie, F. Xin, and S. Wen, "Active Learning for Semi-supervised Classification Based on Information Entropy," *2009 International Forum on Information Technology and Applications*, pp. 591–595, 2009.

- [9] C. Sun, L. Lin, X. Wang, and Y. Guan, "Using maximum entropy model to extract protein-protein interaction information from biomedical literature," in *Proceedings of the intelligent computing 3rd international conference on Advanced intelligent computing theories and applications*, ser. ICIC'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 730–737.
- [10] J. Kazama and J. Tsujii, "Maximum entropy models with inequality constraints: A case study on text categorization," *Machine Learning*, pp. 159–194, 2005.
- [11] A. Carlson, S. A. Hong, K. Killourhy, and S. Wang, "Active Learning for Information Extraction via Bootstrapping," 2010.
- [12] M. Uschold, M. King, S. Moralee, and Y. Zorgios, "The Enterprise Ontology," *Knowl. Eng. Rev.*, vol. 13, no. 1, pp. 31–89, 1998.
- [13] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Semantic similarity measures as tools for exploring the gene ontology," *Pac Symp Biocomput*, pp. 601–612, 2003.
- [14] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.
- [15] G. Nenadic, I. Spasic, and S. Ananiadou, "Mining term similarities from corpora," *Terminology*, vol. 10, no. 1, pp. 55–80, 2004.
- [16] I. Spasic, S. Ananiadou, and J. Tsujii, "MaSTerClass: a case-based reasoning system for the classification of biomedical terms," *Bioinformatics*, vol. 21, no. 11, pp. 2748–2758.
- [17] L. Yue and W. C. Reisdorf, "Pathway and ontology analysis: emerging approaches connecting transcriptome data and clinical endpoints," *Curr. Mol. Med.*, vol. 5, pp. 11–21, 2005.
- [18] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L.

- Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D258—61, 2004.
- [19] E. Beisswanger, V. Lee, J.-J. Kim, D. Rebholz-Schuhmann, A. Splendiani, O. Dameron, S. Schulz, and U. Hahn, "Gene Regulation Ontology (GRO): design principles and use cases," *Stud Health Technol Inform*, vol. 136, pp. 9–14, 2008.
- [20] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Foster, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li,

- Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarri, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Siden-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirska, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, WoodageT, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin, and J. C. Venter, "The genome sequence of *Drosophila melanogaster*," *Science*, vol. 287, no. 5461, pp. 2185–2195, 2000.
- [21] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology," *BMC Bioinformatics*, vol. 6, no. Suppl 1, p. S1, 2005.
- [22] E. B. Camon, D. G. Barrell, E. C. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, and R. Apweiler, "An evaluation of GO annotation retrieval for BioCreAtIvE and GOA," *BMC Bioinformatics*, vol. 6 Suppl 1, p. S17, 2005.
- [23] M. Liu, A. Liberzon, S. W. Kong, W. R. Lai, P. J. Park, I. S. Kohane, and S. Kasif, "Network-based analysis of affected biological processes in type 2 diabetes models," *PLoS Genet.*, vol. 3, no. 6, p. e96, 2007.
- [24] H. K. Dressman, G. G. Muramoto, N. J. Chao, S. Meadows, D. Marshall, G. S. Ginsburg, J. R. Nevins, and J. P. Chute, "Gene expression signatures that predict radiation exposure in mice and humans," *PLoS Med*, vol. 4, no. 4, p. e106, 2007.
- [25] S. Y. Rhee, V. Wood, K. Dolinski, S. Draghici, and S. Yon Rhee, "Use and misuse of the gene ontology annotations," *Nat Rev Genet*, vol. 9, no. 7, pp. 509–515, 2008.

-
- [26] B. de Bruijn and J. D. Martin, "Getting to the (c)ore of knowledge: mining biomedical literature," *I. J. Medical Informatics*, vol. 67, no. 1-3, pp. 7–18, 2002.
- [27] D. Hanisch, J. Fluck, H. T. Mevissen, and R. Zimmer, "Playing biology's name game: identifying protein names in scientific text," *Pac Symp Biocomput*, pp. 403–414, 2003.
- [28] N. Wacholder, "Spotting and Discovering Terms Through Natural Language Processing," *Information Retrieval*, vol. 6, no. 2, pp. 277–281, 2003.
- [29] A. M. Cohen, W. R. Hersh, C. Dubay, and K. Spackman, "Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts," *BMC Bioinformatics*, vol. 6, no. 1, p. 103, 2005.
- [30] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 142–147.
- [31] O. Bender, F. J. Och, and H. Ney, "Maximum entropy models for named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 148–151.
- [32] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 168–171.
- [33] X. Carreras, L. Màrquez, and L. Padró, "A simple named entity extractor using adaboost," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 152–155.

- [34] F. De Meulder and W. Daelemans, "Memory-based named entity recognition using unannotated data," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 208–211.
- [35] J. Mayfield, P. McNamee, and C. Piatko, "Named entity recognition using hundreds of thousands of features," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 184–187.
- [36] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ser. CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 188–191.
- [37] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363–370.
- [38] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, "Accomplishments and challenges in literature data mining for biology," *Bioinformatics*, vol. 18, no. 12, pp. 1553–1561, 2002.
- [39] L. Smith, L. Tanabe, R. Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. Struble, R. Povinelli, A. Vlachos, W. Baumgartner, L. Hunter, B. Carpenter, R. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Mana-Lopez, J. Mata, and W. J. Wilbur, "Overview of BioCreative II gene mention recognition," *Genome Biology*, vol. 9, no. Suppl 2, p. S2, 2008.
- [40] A. S. Yeh, L. Hirschman, and A. A. Morgan, "Evaluation of text data mining for database curation: lessons learned from

-
- the KDD Challenge Cup," *Bioinformatics*, vol. 19, no. suppl 1, pp. i331—i339, 2003.
- [41] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia, "Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge." *Genome biology*, vol. 9 Suppl 2, no. Suppl 2, p. S1, 2008.
 - [42] H. Yu, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and W. J. Wilbur, "Automatic extraction of gene and protein synonyms from MEDLINE and journal articles." *Proc AMIA Symp*, pp. 919–923, 2002.
 - [43] H. Yu and E. Agichtein, "Extracting synonymous gene and protein terms from biological literature," *Bioinformatics*, vol. 19, no. suppl 1, pp. i340—i349, 2003.
 - [44] H. Liu and C. Friedman, "Mining Terminological Knowledge in Large Biomedical Corpora," in *Pacific Symposium on Biocomputing'03*, 2003, pp. 415–426.
 - [45] D. Ahn, "The stages of event extraction," in *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, ser. ARTE '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 1–8.
 - [46] Z. Chen and H. Ji, "Language specific issue and feature exploration in chinese event extraction," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 209–212.
 - [47] S. Patwardhan and E. Riloff, "A unified model of phrasal and sentential evidence for information extraction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 151–160.

- [48] S. Liao and R. Grishman, "Acquiring Topic Features to improve Event Extraction: in Pre-selected and Balanced Collections." in *RANLP*, 2011, pp. 9–16.
- [49] —, "Using document level cross-event inference to improve event extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 789–797.
- [50] J.-D. Kim, Y. Wang, T. Takagi, and A. Yonezawa, "Overview of Genia Event Task in BioNLP Shared Task 2011," in *Proceedings of BioNLP Shared Task 2011 Workshop*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 7–15.
- [51] G. D. Bader, M. P. Cary, and C. Sander, "Pathguide: a pathway resource list," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D504—D506, 2006.
- [52] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology." *Nucleic acids research*, vol. 32, no. Database issue, pp. D262—6, 2004.
- [53] J.-J. Kim and D. Rebholz-Schuhmann, "Improving the extraction of complex regulatory events from scientific text by using ontology-based inference," *Journal of Biomedical Semantics*, vol. 2, no. Suppl 5, p. S3, 2011.
- [54] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski, "Extracting Complex Biological Events with Rich Graph-Based Feature Sets," *Computational Intelligence*, vol. 27, no. 4, pp. 541–557, 2011.
- [55] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski, "Extracting Contextualized Complex Biological Events with Rich Graph-Based Feature Sets," *Computational Intelligence, Special issue on Extracting Bio-molecular Events from Literature*, 2011.

-
- [56] J. Björne, F. Ginter, and T. Salakoski, “University of Turku in the BioNLP’11 Shared Task,” *BMC Bioinformatics*, vol. 13, no. Suppl 11, p. S4, 2012.
 - [57] J. Björne, A. Airola, T. Pahikkala, and T. Salakoski, “Drug-Drug Interaction Extraction from Biomedical Texts with SVM and RLS Classifiers,” *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*, vol. 761, pp. 35–42, 2011.
 - [58] J. Björne and T. Salakoski, “TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task,” in *Proceedings of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 16–25.
 - [59] E. Buyko, E. Faessler, J. Wermter, and U. Hahn, “Event Extraction from Trimmed Dependency Graphs,” *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, ACL*, pp. 19–27, 2009.
 - [60] H. Kilicoglu and S. Bergler, “Syntactic Dependency Based Heuristics for Biological Event Extraction,” *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, ACL*, pp. 119–127, 2009.
 - [61] R. Sebastian, W. C. Hong, T. Toshihisa, and T. Jun’ichi, “A Markov Logic Approach to Bio-Molecular Event Extraction,” *Proceedings of the Workshop on BioNLP*, pp. 41–49, 2009.
 - [62] S. Van Landeghem, Y. Saeys, B. De Baets, Y. Van de Peer, S. Landeghem, B. Baets, and Y. Peer, “Analyzing text in search of bio-molecular events: a high-precision machine learning framework,” in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. Association for Computational Linguistics, 2009, pp. 128–136.
 - [63] B. Settles, “Active Learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.
 - [64] B. Settles, M. Craven, L. Friedland, and L. Friedl, “Active learning with real annotation costs,” in *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008, pp. 1–10.

- [65] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, no. 4, pp. 319–342, 1988.
- [66] D. Cohn, L. Atlas, and R. Ladner, "Improving Generalization with Active Learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.
- [67] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," Providence, Rhode Island, pp. 591–596, 1997.
- [68] A. McCallum and K. Nigam, "Employing EM and Pool-Based Active Learning for Text Classification," pp. 350–358, 1998.
- [69] S. Argamon-Engelson and I. Dagan, "Committee-Based Sample Selection for Probabilistic Classifiers," pp. 335–360, 2011.
- [70] H. Yu, "SVM selective sampling for ranking with application to data retrieval," Chicago, Illinois, USA, pp. 354–363, 2005.
- [71] A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka, "Selective sampling for example-based word sense disambiguation," *Computational linguistics*, vol. 24, no. 4, pp. 573–597, 1998.
- [72] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," Dublin, Ireland, pp. 3–12, 1994.
- [73] V. Vapnik, *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., 1982.
- [74] S. Tong and D. Koller, "Support Vector machine Active Learning with Applications to Text Classification," *Journal of Machine Learning Research*, pp. 45–66, 2001.
- [75] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Large-scale text categorization by batch mode active learning," Edinburgh, Scotland, pp. 633–642, 2006.
- [76] C. A. Thompson, M. E. Califf, and R. J. Mooney, "Active Learning for Natural Language Parsing and Information Extraction," pp. 406–414, 1999.

-
- [77] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," Ottawa, Canada, pp. 107–118, 2001.
 - [78] R. Yan, J. Yang, and A. Hauptmann, "Automatically Labeling Video Data Using Multi-class Active Learning," p. 516, 2003.
 - [79] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.
 - [80] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," New Brunswick, pp. 148–156, 1994.
 - [81] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," Pittsburgh, Pennsylvania, pp. 746–751, 2005.
 - [82] C. E. SHANNON, "A MATHEMATICAL THEORY OF COMMUNICATION," *BELL SYSTEM TECHNICAL JOURNAL*, vol. 27, no. 3, pp. 379–423, 1948.
 - [83] M. Lindenbaum, S. Markovitch, and D. Rusakov, "Selective Sampling for Nearest Neighbor Classifiers," *Mach. Learn.*, vol. 54, no. 2, pp. 125–152, 2004.
 - [84] V. V. Fedorov, *Theory of optimal experiments*. New York: Academic Press, 1972.
 - [85] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," Pittsburgh, Pennsylvania, United States, pp. 287–294, 1992.
 - [86] N. Abe and H. Mamitsuka, "Query Learning Strategies Using Boosting and Bagging," pp. 1–9, 1998.
 - [87] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
 - [88] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

- [89] P. Melville and R. J. R. J. Mooney, "Diverse ensembles for active learning," ... *international conference on Machine learning*, p. 74, 2004.
- [90] R. Burbidge, J. Rowland, and R. King, "Active Learning for Regression Based on Query by Committee Intelligent Data Engineering and Automated Learning - IDEAL 2007," H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds. Springer Berlin / Heidelberg, 2007, vol. 4881, pp. 209–218.
- [91] B. Cui, H. Lin, and Z. Yang, "Uncertainty sampling-based active learning for protein-protein interaction extraction from biomedical literature," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 344 – 10 350, 2009.
- [92] H.-T. Zhang, M.-L. Huang, and X.-Y. Zhu, "A Unified Active Learning Framework for Biomedical Relation Extraction," *Journal of Computer Science and Technology*, vol. 27, no. 6, pp. 1302–1313, 2012.
- [93] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of bionlp'09 shared task on event extraction," in *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 1–9.
- [94] J.-D. Kim, Y. Wang, T. Takagi, and A. Yonezawa, "Overview of genia event task in bionlp shared task 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*. Association for Computational Linguistics, 2011, pp. 7–15.
- [95] K. Sagae, Y. Miyao, and J. Tsujii, "Hpsg parsing with shallow dependency constraints," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 624–631.
- [96] G. W. Corder and D. I. Foreman, *Nonparametric Statistics for Non-Statisticians: a step-by-step approach*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2009.
- [97] J.-J. Kim, X. Han, V. Lee, and D. Rebholz-Schuhmann, "Gro task: Populating the gene regulation ontology with events

- and relations,” in *Proceedings of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 50–57.
- [98] D. Campos, S. Matos, and J. L. Oliveira, “Gimli: open source and high-performance biomedical name recognition,” *BMC Bioinformatics*, vol. 14, no. 1, p. 54, 2013.
- [99] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2013, pp. 3111–3119.
- [100] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the margin: A new explanation for the effectiveness of voting methods,” *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.
- [101] A. Vlachos, “Active learning with support vector machines,” Master’s thesis, School of Informatics University of Edinburgh, 2004.
- [102] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.
- [103] N. V. Cuong, W. S. Lee, N. Ye, K. M. A. Chai, and H. L. Chieu, “Active learning for probabilistic hypotheses using the maximum gibbs error criterion,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2013, pp. 1457–1465.
- [104] S. Pyysalo, T. Ohta, and S. Ananiadou, “Overview of the cancer genetics (cg) task of bionlp shared task 2013,” in *Proceedings of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 58–66.
- [105] J.-D. Kim, Y. Wang, and Y. Yasunori, “The genia event extraction shared task, 2013 edition - overview,” in *Proceedings*

- of the BioNLP Shared Task 2013 Workshop*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 8–15.
- [106] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, “Active learning with sampling by uncertainty and density for word sense disambiguation and text classification,” in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, August 2008, pp. 1137–1144.
- [107] Y. Baram, R. El-Yaniv, and K. Luz, “Online Choice of Active Learning Algorithms,” *The Journal of Machine Learning Research*, vol. 5, pp. 255–291, 2004.
- [108] J.-D. Kim, J.-j. Kim, X. Han, and D. Rebholz-Schuhmann, “Extending the evaluation of genia event task toward knowledge base construction and comparison to gene regulation ontology task,” *BMC Bioinformatics*, vol. 16, no. 10, pp. 1–13, 2015.
- [109] S. Pyysalo, T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, and S. Ananiadou, “Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013,” *BMC Bioinformatics*, vol. 16, no. 10, pp. 1–19, 2015.
- [110] B. Settles and M. Craven, “An analysis of active learning strategies for sequence labeling tasks,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 1070–1079.
- [111] J. Kang, K. Ryu, and H.-C. Kwon, “Using cluster-based sampling to select initial training set for active learning in text classification,” in *Advances in Knowledge Discovery and Data Mining*, 2004, pp. 384–388.
- [112] L. Qian and G. Zhou, “Clustering-based stratified seed sampling for semi-supervised relation classification,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, October 2010, pp. 346–355.
- [113] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge,

2008, vol. 1, no. 1.

- [114] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 2014, pp. 1532–1543.
- [115] M. R. Anderberg, *Cluster analysis for applications*, 1973.
- [116] R. Xu, D. Wunsch *et al.*, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.
- [117] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*. Springer, 2006, pp. 25–71.
- [118] E. Schubert, A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek, "A framework for clustering uncertain data," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1976–1979, 2015.
- [119] G. Hamerly, "Making k-means even faster." in *SDM*. SIAM, 2010, pp. 130–140.
- [120] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.
- [121] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [122] L. Ertöz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data." in *SDM*. SIAM, 2003, pp. 47–58.
- [123] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia, "Evaluation of text-mining systems for biology: overview of the second biocreative community challenge," *Genome Biology*, vol. 9, no. Suppl 2, p. S1, 2008.
- [124] A. I. Schein and L. H. Ungar, "Active Learning for Logistic Regression: An Evaluation," *Machine Learning*, vol. 68, no. 3, pp. 235–265, Oct. 2007.

- [125] X. Han and J.-j. Kim, "Active learning for ontological event extraction," in *6th International Symposium on Semantic Mining in Biomedicine (SMBM)*, 2014, pp. 45–51.