An Overview of the Active Gene Annotation Corpus and the BioNLP OST 2019 AGAC Track Tasks

Yuxing Wang, Kaiyin Zhou, Mina Gachloo and **Jingbo Xia***
Hubei Key Lab of Agricultural Bioinformatics, College of Informatics,
Huazhong Agricultural University, 430070, Wuhan, China
Mailto: xiajingbo.math@gmail.com

Abstract

The active gene annotation corpus (AGAC) was developed to support knowledge discovery for drug repurposing. The AGAC track of the BioNLP Open Shared Tasks 2019 was organized, to facilitate cross-disciplinary collaboration across BioNLP and Pharmacoinformatics communities, for drug repurposing. The AGAC track consists of three subtasks: 1) named entity recognition, 2) thematic relation extraction, and 3) loss of function (LOF) / gain of function (GOF) topic classification. The AGAC track was participated by five teams, of which the performance is compared and analyzed. The results revealed a substantial room for improvement in the design of the task, which we analyzed in terms of "*imbalanced data*", "*selective annotation*" and "*latent topic annotation*".

Keywords: corpus annotation, shared task, gene mutation, drug repurposing

1 Introduction

Biomedical natural language processing (BioNLP) has long been recognized as effective method to accelerate drug-related knowledge discovery (Vazquez et al., 2011; Gachloo et al., 2019). Particularly, PubMed is regarded as a main source for knowledge discovery as it stored a vast amount of reports on scientific discovery, and the size keeps constantly growing (Hunter and Cohen, 2006; Cohen et al., 2016). Various corpora used texts from PubMed. Examples include GENIA (Kim et al., 2003), CRAFT (Cohen et al., 2017), and BioCreative task corpora (Li et al., 2016), to name just a few.

The growing interest in developing corpus annotation also has led to the development of public annotation platform in the BioNLP community. An example of recent progress is PubAnnotation (Kim and Wang, 2012; Kim et al., 2019), which offers a versatile platform for corpus construction, annotation, sharing the data, and offering them as open shared tasks (https://2019.bionlp-ost.org/tasks).

In the context of drug-related knowledge discovery, various corpora were developed. Examples include annotated corpora for adverse drug reactions (ADR) (Roberts et al., 2017; Demner-Fushman et al., 2018; Karimi et al., 2015; Ginn et al., 2014; Gurulingappa et al., 2012), and those for drug-drug interactions (DDI) (Herrero-Zazo et al., 2013). However, as far as the authors know, there has been no work of corpus annotation (except AGAC-related ones) for drug repurposing. Drug repurposing (AKA drug repositioning) is to find new indications of approved drugs, which is now recognized as an important mean for investigating novel drug efficiency in the pharmaceutical industry.

This paper presents the Active Gene Annotation Corpus (AGAC) corpus and a shared task (the AGAC track of BioNLP Open Shared Tasks 2019) based on it. The design of AGAC is highly motivated by the LOF-agonist/GOF-antagonist hypothesis proposed by Wang and Zhang (Wang and Zhang, 2013), which states:

> For a given disease caused by driven gene with Loss of function (LOF) or Gain of function (GOF), an targeted antagonist/agonist is a candidate drug.

The hypothesis was well supported by experiments, which encouraged large scale automatic knowledge curation.

Actually, the hypothesis represented the ideas of tracking the phenotypic information of gene and it shared the similar motivation of phenome-wide association studies (PheWAS) (Rastegar-Mojarad et al., 2015). In PheWAS, the international classification of diseases (ICD) codes was assigned as

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 62–71
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

the form of the phenotype to candidate single nucleotide polymorphisms (SNPs) so as to investigate the relevance of phenotypes and gene mutation.

AGAC is a corpus annotated by human experts, with an aim at capturing function changes of mutated genes in a pathogenic context. The design of the corpus and the guidelines were published in 2017 (Wang et al., 2018), and a case study of using such an annotated corpus for drug repurposing was successfully performed in 2019, unveiling potential associations of variations with a wide spectrum of human diseases (Zhou et al., 2019). Since then, the whole annotation work took 20 months, with involvement of four annotators.

Using the corpus the AGAC track of BioNLP Open Shared Tasks 2019 was organized, which was participated by 5 teams. In this paper, both the AGAC corpus and AGAC track are introduced, and the performance of the participants are presented. The full information of the AGAC track is available at the website, `https://sites.google.com/view/bionlp-ost19-agac-track`.

2 The AGAC corpus and shared task

2.1 Corpus preparation

We collected abstracts by Mesh terms "Mutation/physiopathology" and "Genetic Disease". AGAC is annotated for eleven types of named entities, which categorized into bio-concepts, regulation types, and other entities, and for two types of thematic relations between them. All the types of named entities and thematic relations are defined in the AGAC ontology (see Figure 1).

While the full description of the named entity types can be found in the AGAC guideline book (Wang et al., 2018), briefly speaking, it is designed to include the entities which are relevant to genetic variations and forthcoming phenotype changes at molecular and cellular levels, with a focus on tracing the biological semantics of LOF and GOF mutations.

Since AGAC aims to annotate mutations and the subsequent bio-processes caused by the mutations, the two thematic role types, `themeOf` and `causeOf`, of which the original use are introduced by the GENIA event annotation (Kim et al., 2008), are adopted to represent relations between AGAC entities. Note that here the use of the `themeOf` and `causeOf` relations are a little bit different from their use in linguistic analysis, in the sense that they are *not* confined to be used only around verbs. In AGAC, the thematic relations may be used to connect two named entities, both in noun forms. Below is the semantics of the two thematic relations:

- `ThemeOf`: a theme of an event (or a regulatory named entities) is the object which undergoes a change of its state due to the event.

- `CauseOf`: a cause of an event (or a regulatory named entities) is the object which leads the event to happen.

In order to help understanding of the semantics of the AGAC entities, they are mapped to corresponding MeSH terms (Lipscomb, 2000) whenever possible (see Figure 1).

In addition to the annotations for named entities and relations, each abstract in AGAC is annotated with a statement of a LOF/GOF-classified gene-disease association. The statement is expressed by a triple: a gene, the type of function change (GOF or LOF), and a disease. For example, if an abstract reports an association between a mutation of SHP-2, which causes a GOF type of function change, and leukemia, the abstract is annotated with the triple, *SHP-2; GOF; leukemia*. Note that it is the most straightforward form of knowledge piece to apply the LOF-agonist/GOF-antagonist hypothesis to discovery of candidate chemicals for diseases, which is the primary application scenario of AGAC.

2.2 Statistics and characteristics of AGAC corpus

AGAC corpus is annotated by four annotators: a main annotator and three fellow annotators. To evaluate the quality of the annotations, inter-annotator agreement (IAA) was measured in an asymmetric way: the performance of the main annotator was assumed as the "oracle", to which the performance of each fellow annotator was compared. The IAAs of the three annotators were 0.68, 0.78 and 0.70, respectively, in F-score.

To serve as the training and test data sets of the AGAC shared task, the corpus was randomly divided into halves: 250 abstracts for each of the training and the test data sets. The basic statistics of the abstracts, sentences, and annotations are shown in Table 1.

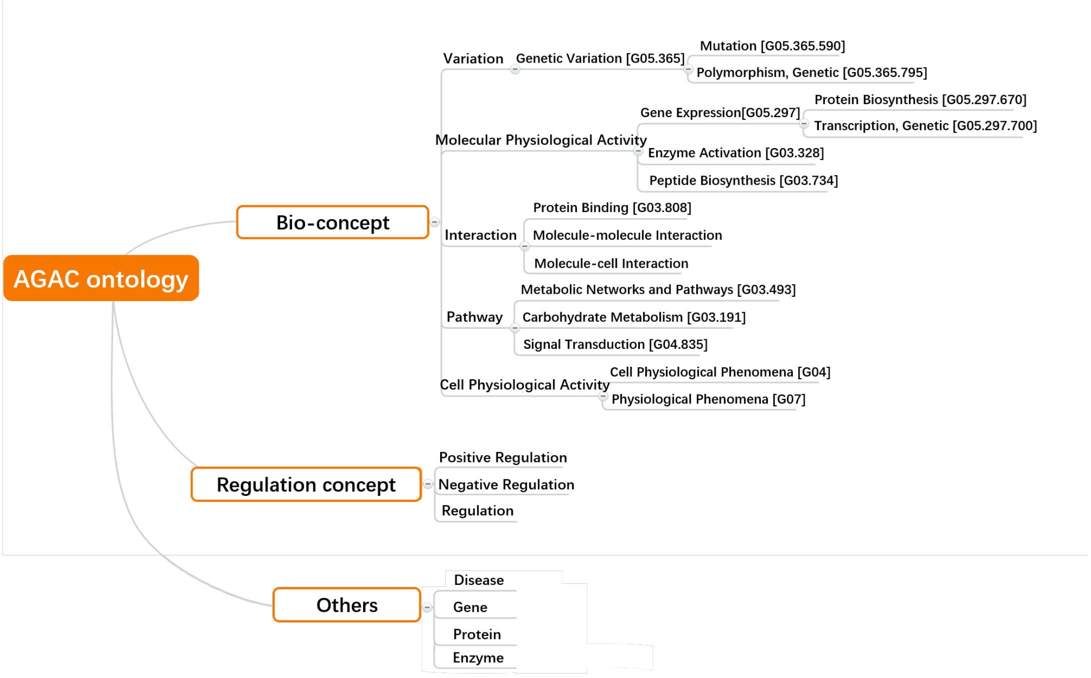

AGAC ontology
Bio-concept
Variation
Genetic Variation [G05.365]
Mutation [G05.365.590]
Polymorphism, Genetic [G05.365.795]
Molecular Physiological Activity
Gene Expression[G05.297]
Protein Biosynthesis [G05.297.670]
Transcription, Genetic [G05.297.700]
Enzyme Activation [G03.328]
Peptide Biosynthesis [G03.734]
Interaction
Protein Binding [G03.808]
Molecule-molecule Interaction
Molecule-cell Interaction
Pathway
Metabolic Networks and Pathways [G03.493]
Carbohydrate Metabolism [G03.191]
Signal Transduction [G04.835]
Cell Physiological Activity
Cell Physiological Phenomena [G04]
Physiological Phenomena [G07]
Regulation concept
Positive Regulation
Negative Regulation
Regulation
Others
Disease
Gene
Protein
Enzyme

Figure 1: AGAC ontology.

Table 1: Statistics of annotations in total, training and test sets

	Total	**Training set**	**Test set**
# of Abstracts	**500**	**250**	**250**
# of Sentences	**5,080**	**2,534**	**2,546**
# of Named entities	**5,741**	**3,317**	**2,424**
.Bio-concept Named Entities	**2,274**	**1,428**	**846**
`Var (Variation)`	1,304	735	569
`MPA (Molecular Physiological Activity)`	618	418	200
`Interaction`	35	28	7
`Pathway`	38	24	14
`CPA (Cell Physiological Activity)`	279	223	56
.Regulatory Named Entities	**1,514**	**905**	**609**
`Regulation`	613	215	398
`Positive Regulation`	406	323	83
`Negative Regulation`	495	367	128
.Other Entities	**1,953**	**984**	**969**
`Disease`	751	336	415
`Gene`	1,004	529	475
`Protein`	150	90	60
`Enzyme`	48	29	19
# of Thematic roles	**4,677**	**2,729**	**1,948**
`ThemeOf`	2,986	1,698	1,288
`ThemeOf` (Intra/inter sentential)	(2910/76)	(1657/41)	(1253/35)
`CauseOf`	1,691	1,031	660
`CauseOf` (Intra/inter sentential)	(1581/110)	(961/70)	(620/40)

The AGAC corpus is characterized in three terms: *imbalanced data*, *selective annotation*, and *latent topic annotation*.

i) **Imbalanced Data:** The statistics in Table 1 clearly shows that the entity distribution is imbalanced over the entity types, e.g. 1,304 `Var` vs. 35 `Interaction` annotations, and across the training and test data sets, e.g., 481 vs. 200 `MPA` annotations in the training and test data sets, respectively. In the mean time, the distribution of several named entities shows imbalance between training set and test set. For instance, there are 418 `MPA` in training set, while the amount is 200 in test set. Similarly, the amount ratio of `Interaction` and `Pathway` is 28:7 and 24:14. As in the thematic roles, the amount of `CauseOf` in training set is mostly doubled than that in test set.

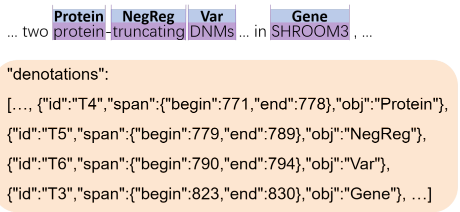
ii) **Selective Annotation:** According to the AGAC guidelines (Wang et al., 2018), annotations are made only to the sentences which carry sufficient information to mine a gene-disease association with LOF/GOF specification, i.e., a sentence is annotated only if it contains specific gene, mutation, disease mentions. In other words, the named entities appearing in a sentence are *not* annotated if the sentence misses any of the required entities. Later, it has turned out to be a tricky feature, which makes the NER task based on the corpus a much more complicated one compared to typical NER tasks (See Section 5).

iii) **Latent Topic Annotation:** The annotation of each abstract with a LOF/GOF-classified gene-disease association may be regarded as a kind of latent topic annotation, in the sense that the LOF/GOF context of a gene-disease association may not be directly visible from the text. This feature makes the AGAC annotation unique: the annotation is really geared toward knowledge discovery for drug repurposing based on the LOF-agonist/GOF-antagonist hypothesis. Note that the agonist or antagonist information of a chemical is available in various databases like *Drugbank* (Wishart et al., 2017) or *Therapeutic Target Database (TTD)* (Li et al., 2017), which means, if mining of LOF/GOF-classified gene-disease association is possible in a large scale, mining of drug candidates for diseases also will be possible in a large scale.

2.3 Task Definition of AGAC Track

AGAC track consists of three tasks: Task 1: named entity recognition, Task 2: thematic relation extraction, and Task 3: mutation-disease knowledge discovery. While participants were allowed to choose the tasks they would participate, due to the dependency between the tasks, it was expected that participating all the three tasks might maximize the chance of high performance: Task 2 requires the result of Task 1, and Task 3 may be benefited from the result of Task 1 and 3. Below is the details of the three tasks:

Task 1. NER: To recognize named entities appearing in given texts, and to assign them their entity class, based on the AGAC ontology. Figure 2 shows an example, where four spans, *“protein”*, *“Truncating”*, *“DNMs”*, and *“SHROOM3”* are annotated as `Protein`, `Negative Regulation`, `Variation`, and `Gene`, respectively. The participants are required to produce the result in the PubAnnotation JSON format. Note that while compound nouns are common, there is no discontinuous or overlapping spans annotated as named entities, in AGAC.

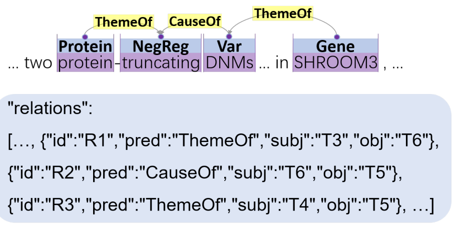


```
"denotations":
[..., {"id":"T4","span":{"begin":771,"end":778},"obj":"Protein"},
{"id":"T5","span":{"begin":779,"end":789},"obj":"NegReg"},
{"id":"T6","span":{"begin":790,"end":794},"obj":"Var"},
{"id":"T3","span":{"begin":823,"end":830},"obj":"Gene"}, ...]
```

Figure 2: Annotation example for Task 1.

Task 2. Thematic relation identification: To identify the thematic relation, `ThemeOf`, `CauseOf`, between named entities. Figure 3 shows an example, where two `ThemeOf` relations, `Protein` → `Negative regulation` and `Gene` → `Variation`, and one `CauseOf` relation, `Negative regulation` → `Variation`, are annotated. Note that the relation annotations are added on top of the NER annotations. Note also that relations may be intra- or inter-sentential, and in AGAC, 3.98% of the relations are inter-sentential.

Task 3. Mutation-disease knowledge discovery: To extract the triples of a gene, a function change,



"relations":
[..., {"id":"R1","pred":"ThemeOf","subj":"T3","obj":"T6"},
{"id":"R2","pred":"CauseOf","subj":"T6","obj":"T5"},
{"id":"R3","pred":"ThemeOf","subj":"T4","obj":"T5"}, ...]

Figure 3: Annotation example for Task 2

and a disease. A function change is classified into four classes: Loss of Function(LOF), Gain of Function(GOF), Regulation(REG), and Complex(COM). Figure 4 shows an example, where the PubMed abstract, 25805808, is annotated with the triple, *SHROOM3; LOF; Neural tube defects*. Participants are requried to produce a text file where a quadraple (a PubMed Id, plus a triple) takes one line. Note that while this task is independent from Task 1 and 2, syntactically, it may be benefited from the results of the two tasks, semantically.

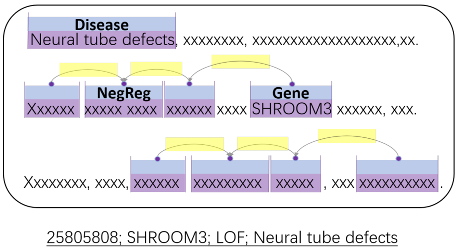

Disease
Neural tube defects, xxxxxxxx, xxxxxxxxxxxxxxxxxxx,xx.
NegReg
Gene
Xxxxxx xxxxx xxxx xxxxxx xxxx SHROOM3 xxxxxx, xxx.
Xxxxxxx, xxxx, xxxxxx xxxxxxxxx xxxxx , xxx xxxxxxxxxx.
25805808; SHROOM3; LOF; Neural tube defects

Figure 4: Annotation example for Task 3

For better understanding, let us pick a sentence, *"Mutations in SHP-2 phosphates that cause hyperactivation of its catalytic activity have been identified in human leukemia, particularly juvenile myelomonocytic leukemia."* From a biological view, hyperactivation of catalytic activity is clearly a description of Gain-of-Function. Henceforth, this sentence carries clear semantic information that, a gene *"SHP-2"* after mutation plays a GOF function related to the disease *"juvenile myelomonocytic leukemia"*. Therefore, the Task 3 requires the triple from this sentence, i.e., `SHP-2;GOF;juvenile myelomonocytic leukemia`.

In another sentence, *"Lynch syndrome (LS) caused by mutations in DNA mismatch repair genes MLH1."*, it describes the association between disease *"Lynch syndrome"* and gene *"MLH1"*, but the phrase *"caused by"* means no loss or gain, hence the triple from this sentence should be `MLH1;REG;Lynch syndrome`.

In a COM example, *"Here, we describe a fourth case of a human with a de novo KCNJ6 (GIRK2) mutation, who presented with clinical findings of severe hyperkinetic movement disorder and developmental delay. Heterologous expression of the mutant GIRK2 channel alone produced an aberrant basal inward current that lacked G protein activation, lost K+ selectivity and gained Ca2+ permeability."* , the description *"lost K+ selectivity and gained Ca2+ permeability"* shows both LOF and GOF, therefore the function change can not be labeled as LOF or GOF but COM, `GIRK2;COM;hyperkinetic movement disorder`.

2.4 Sample data for task 1, 2, and 3

Figure 5 shows a sample text of AGAC corpus, the format of which is JSON. The bold term "target" is the address of the annotated text. "sourcedb" is where the text original from, all the text in AGAC corpus are from PubMed. "sourceid" is pmid of the text. "text" contains the raw abstract.

1) "denotations" for Task 1:

   "denotations" contains the named entity annotations corresponding to Task 1. Each named entity annotation has an "id"; a "span": its position in the abstract; an "obj": the named entity it belongs to.

2) "relations" for Task 2:

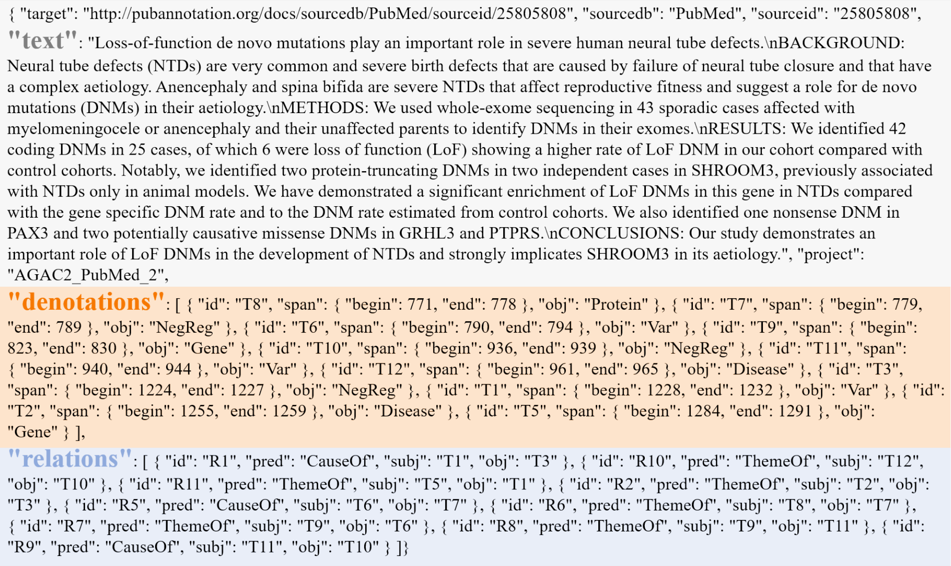
   "relations" contains the thematic roles between the named entities, which corresponds to Task 2. Each relation contains an "id"; a "pred": the thematic roles; "subj" and "obj": the named entity "id" that the relation associates, and the direction of the relation is from "subj" to "obj".

   Note that Task 2 requires the result of Task 1.

3) Triples for Task 3:

   25805808;SHROOM3;LOF;Neural tube defects Triples showed above is the result of Task 3, which is required to be extracted from the sample text. So, for the result template during evaluation, the standard format of triples is: `pmid;gene;function change;disease`.

The visualization of part of this sample text is shown in Figure 5, which is presented by the annotation platform PubAnnotation.



{ "target": "http://pubannotation.org/docs/sourcedb/PubMed/sourceid/25805808", "sourcedb": "PubMed", "sourceid": "25805808", "text": "Loss-of-function de novo mutations play an important role in severe human neural tube defects.\nBACKGROUND: Neural tube defects (NTDs) are very common and severe birth defects that are caused by failure of neural tube closure and that have a complex aetiology. Anencephaly and spina bifida are severe NTDs that affect reproductive fitness and suggest a role for de novo mutations (DNMs) in their aetiology.\nMETHODS: We used whole-exome sequencing in 43 sporadic cases affected with myelomeningocele or anencephaly and their unaffected parents to identify DNMs in their exomes.\nRESULTS: We identified 42 coding DNMs in 25 cases, of which 6 were loss of function (LoF) showing a higher rate of LoF DNM in our cohort compared with control cohorts. Notably, we identified two protein-truncating DNMs in two independent cases in SHROOM3, previously associated with NTDs only in animal models. We have demonstrated a significant enrichment of LoF DNMs in this gene in NTDs compared with the gene specific DNM rate and to the DNM rate estimated from control cohorts. We also identified one nonsense DNM in PAX3 and two potentially causative missense DNMs in GRHL3 and PTPRS.\nCONCLUSIONS: Our study demonstrates an important role of LoF DNMs in the development of NTDs and strongly implicates SHROOM3 in its aetiology.", "project": "AGAC2_PubMed_2",
"denotations": [{ "id": "T8", "span": { "begin": 771, "end": 778 }, "obj": "Protein" }, { "id": "T7", "span": { "begin": 779, "end": 789 }, "obj": "NegReg" }, { "id": "T6", "span": { "begin": 790, "end": 794 }, "obj": "Var" }, { "id": "T9", "span": { "begin": 823, "end": 830 }, "obj": "Gene" }, { "id": "T10", "span": { "begin": 936, "end": 939 }, "obj": "NegReg" }, { "id": "T11", "span": { "begin": 940, "end": 944 }, "obj": "Var" }, { "id": "T12", "span": { "begin": 961, "end": 965 }, "obj": "Disease" }, { "id": "T3", "span": { "begin": 1224, "end": 1227 }, "obj": "NegReg" }, { "id": "T1", "span": { "begin": 1228, "end": 1232 }, "obj": "Var" }, { "id": "T2", "span": { "begin": 1255, "end": 1259 }, "obj": "Disease" }, { "id": "T5", "span": { "begin": 1284, "end": 1291 }, "obj": "Gene" }],
"relations": [{ "id": "R1", "pred": "CauseOf", "subj": "T1", "obj": "T3" }, { "id": "R10", "pred": "ThemeOf", "subj": "T12", "obj": "T10" }, { "id": "R11", "pred": "ThemeOf", "subj": "T5", "obj": "T1" }, { "id": "R2", "pred": "ThemeOf", "subj": "T2", "obj": "T3" }, { "id": "R5", "pred": "CauseOf", "subj": "T6", "obj": "T7" }, { "id": "R6", "pred": "ThemeOf", "subj": "T8", "obj": "T7" }, { "id": "R7", "pred": "ThemeOf", "subj": "T9", "obj": "T6" }, { "id": "R8", "pred": "ThemeOf", "subj": "T9", "obj": "T11" }, { "id": "R9", "pred": "CauseOf", "subj": "T11", "obj": "T10" }]}

Figure 5: Sample data for Task 1, 2 and 3

3 Evaluation methods

The performance of the participants was evaluated in standard precision, recall, and F-score. For Task 1 and 2, the *PubAnnotation Evaluator*[1] tool was used, with a parameter setting for strict span matching (*soft_match_characters* = 0 & *soft_match_words* = 0). For task 2, for a predicted relation to be counted as a true positive, the two entities participating in the relation have to be correctly predicted, together with the type of the relation. Note that the evaluation criteria applied to Task 1 and 2 are very strict.

For Task 3, a custom evaluation tool was provided by the organizers Unlike Task 1 and 2, for Task 3, a relaxed matching criteria was applied: a "Function-Classified Gene-Disease Assciation" (FCGDA) statement is counted as correct one if the function classification (LOF or GOF) is correctly recognized. The motivation of using the relaxed matching criteria was that it was fairly a new type of task, making a highly challenging one, and and that prediction of the LOF/GOF context was of the primary interest.

4 Results and observations

Overall, five teams participated in the tasks of the AGAC track: three teams in both Task 1 and 2, one team only in Task 1, and one team (through a late submission) only in Task 3. The results of Task 1, 2, and 3 are presented in Table 2, 3, 4, respectively.

Looking into the methods used by the participants, it is observed that, although the number of participants is not so high, various methods are well mixed: a probabilistic sequence labeling model, e.g., *CRF* (Lafferty et al., 2001)), a kernel-based linear classification model, e.g., *SVM*, modern neural network models, e.g., *CNN* (Lawrence et al., 1997) and *Bi-LSTM* (Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012),We collected abstracts by Mesh terms "Mutation/physiopathology" and "Genetic Disease".

and also a joint learning. It is also observed that use of BERT (Devlin et al., 2018), a pre-trained language representation model, was popular.

4.1 Task 1

In Task 1, DX-HITSZ used "JFB-NER" model which was a joint learning model with parameters fine tuned bioBert. Zheng-UMASS used a hierarchical multi-task learning model for both Named entity recognition and Relation Extraction. In this model 12 entities were decomposed into three subtasks: (1) Var, MPA,CPA,Enzyme for part one (2) Gene, Pathway, Protein, Disease for part two (3) PosReg, Interaction, NegReg, Reg for part three. Besides, they used Bert embedding, customized embedding, and Char level embedding to represent inputs sentences. Then, the bi-LSTM encoders were used as encoders for each of the subtasks. YaXXX-SiXXX/LMX used Bi-LSTM CRF with linguistic features and ensemble 3 best models on 3 data splits. Finally, DJDL-HZAU used traditional CRF method and combined with some

[1] https://github.com/pubannotation/pubannotation_evaluator

Table 2: Participants Performance of Task 1

	Participants	Precision	Recall	F-score	Main NLP techniques
1st	DX-HITSZ	0.63	0.56	0.60	Bert, joint learning
*	Baseline	0.50	0.51	0.50	Bert, joint learning
2nd	Zheng-UMASS	0.36	0.59	0.45	Bert, CNN, Bi-LSTM
3rd	YaXXX-SiXXX/LMX	0.55	0.28	0.37	CRF, Bi-LSTM
4th	DJDL-HZAU	0.16	0.25	0.20	CRF

*: Baseline.

Table 3: Participants Performance of Task 2

	Participants	Precision	Recall	F-score	Main NLP techniques
1st	Zheng-UMASS	0.40	0.31	0.35	Bert, CNN, Bi-LSTM
2nd	DX-HITSZ	0.61	0.16	0.25	Bert, joint learning
3rd	YaXXX-SiXXX/LMX	0.05	0.02	0.03	SVM

Table 4: Participants Performance of Task 3

	Participants	Precision	Recall	F-score	Main NLP techniques
*	Baseline	0.72	0.59	0.65	Bert, joint learning
L	Ashok-BenevolentAI	0.26	0.20	0.23	Bert

*: Baseline

L: Late submission.

linguistic features.

4.2 Task 2

In Task 2, Zheng-UMASS used a hierarchical multi-task learning model for both Named entity recognition and Relation Extraction. In relation extraction part the model shared the same encoding layers with Named entity recognition part. DX-HITSZ used a simple fine tuned bioBert, refer as ”SB-RE”. The F-score they obtained is 0.35 and 0.25, respectively. Furthermore, YaXXX-SiXXX/LMX converted the task 2 into a classification model and used the traditional support vector machine to obtain a F-score of 0.03.

4.3 Task 3

In Task 3, Ashok-BenevolentAI used BERT as well to extract “gene function change disease triples. They encoded the pair of mentions and their textual context as two consecutive sequences and then used a single linear layer to classify their relation into five classes. It is noted that none of the results in Task 1 and Task 2 were jointly learned in this model.

As the task organizer, AGAC team provided baseline method for Task 1 and 3. We used BERT to learn semantic structure of the sentences, and use joint learning for output sequence labeling in Task 1 and triple recognition in Task 3.

4.4 Summary

To sum up, the best performance for Task 1 was 0.6 in F-score, which was obtained by DX-HITSZ. It outperformed the reference method provided by the organizers by 0.10 in F-score. For task 2, the base performance was 0.35, which was acheived by Zheng-UMASS. The best performance for Task 1 and 2 are quite low compared to other NER and RE tasks. We attribute the reason to the strict evaluation criteria and the selective annotation characteristics of the AGAC corpus, the latter of which is discussed in Section 5. For Task 3, while the reference performance provided by the organizers achieved a moderate performance, 0.65 in F-score, the only participant achieved a much lower performance, 0.26. We attribute the reason to the fact that the team did not use the results of Task 1 and 2 which we expected critical to perform

Task 3.

5 Discussion and Conclusion

In this section, the "*selective annotation*" and "*latent topic annotation*" features of AGAC are reviewed and future research directions are discussed.

5.1 *Selective annotation* makes NER challenging

As suggested in the previous discussion, state-of-art methods in NLP community, like BERT and joint learning, are frequently tested in AGAC track. Comprehensive investigation of the performance results show the effectiveness and disadvantages of these method.

Unlike normal sequence labelling task, AGAC track requires the artificial intelligence method to perform NER only when the sentence exactly fit the GOF/LOF topic. Here, "*selective annotation*" attribute refers that only the core named entities or phrase within a sentence which carries clear function change semantics is annotated. Actually, the design with this attribute stem from real scenario of the drug knowledge discovery where curators need to trace and extract exact relevant function change information of a mutated gene among texts. Unfortunately, this attribute also make AGAC track a fairly challenging task to fulfill.

The performances comparison in AGAC track shows that the modern NLP strategies like BERT propel the traditional sequence labeling task to the full strength. Both the team won the first position and the baseline method use BERT and joint learning model. As a conclusion, sophisticated language representative model is an effective way to handle sequence labeling in AGAC research. In addition, LOF/GOF recognition without using results of Task 1 and 2 failed to outperform the baseline method which make good use of the named entities in AGAC. It hints that joint learning model is a proper integrated tasks solution for NER, thematic role recognition and LOF/GOF triplet recognition.

In all, the "*Selective annotation*" attribution make AGAC track more challenging than traditional sequence labeling task. Just mocking the human annotator who make annotation with sufficient LOF or GOF semantics consideration, a successful model should discern the full semantics when correctly performing the labeling. Hopefully, the performance of the AGAC track will be enhanced by a design of a more intellectual learning model, which is capable of capturing both the sequence labeling and the triple information, and therefore making tactical adjustment.

5.2 The potential of *latent topic annotation*

The purpose of AGAC track for drug repurposing requires comprehensive cooperation among BioNLP and Bioinformatics communities, even in general, NLP and Biology communities. Though none of the participants attempts to solve Task 3 due to the domain gap of computer science and life science, a cross disciplinary cooperation is still promising, especially in the era of Multi-Omics data (Groen et al., 2016).

"*Latent topic annotation*" attribute refers to comprehensive integration of drug related knowledge and deep cooperation in a cross-disciplinary manner. As mentioned in the introduction, the biological idea of the AGAC design is consistent with the mainstream phenotype mining strategy as PheWAS (Rastegar-Mojarad et al., 2015). In addition, the literature review as well suggests that BioNLP and computational method shed light to drug-related knowledge discovery (Gachloo et al., 2019). In our early attempt of AGAC application (Zhou et al., 2019), a PubMed-wide GOF and LOF recognition is successfully achieved by using AGAC as training data. Specifically, AGAC corpus offers abundant semantic information in the function change recognition, and helps to evaluate the GOF/LOF topic of a Pubmed abstract.

All of the above facts hint that well formed knowledge structure in AGAC is capable of ensuring nice application of function change investigation, and good commanding of the domain knowledge is the key point to propel the research of drug repurposing. Henceforth, it is promising to develop deep cooperation among BioNLP and Bioinformatics communities based on the outcome of AGAC track competition.

6 Data Availability

The AGAC corpus is developed and made available in the PubAnnotation platform, which is technically supported by Database Center for Life Science (DBCLS), Japan. Link to retrieve the data: `http://pubannotation.org/projects/`

AGAC_test/annotations.tgz.

7 Acknowledgement

This work is funded by the Fundamental Research Funds for the Central Universities of China (Project No. 2662018PY096) and Hubei Province Funds for Natural Science (Project No.2019CFB552) .

The authors would also like to express their gratitude for Jin-Dong Kim and Kevin Bretonnel Cohen for many instructive discussion in the corpus design and BioNLP open shared task setting; Yuxing Ren and Shanghui Nie for efforts in corpus annotation; Hong-yu Zhang and Xuan Qin for discussion on the *"LOF/atagonist-GOF/antagonist"* hypothesis; Yao Xinzhi for his participation in guideline design; Sheng Zhang, Qi luo and Panzhong Lu for the discussion and coding of AGAC applications.

References

Kevin Bretonnel Cohen, Karin Verspoor, Karën Fort, Christopher Funk, Michael Bada, Martha Palmer, and Lawrence E Hunter. 2017. The Colorado richly annotated full text (CRAFT) corpus: multi-model annotation in the biomedical domain. In *Handbook of Linguistic Annotation*, pages 1379–1394. Springer.

Kevin Bretonnel Cohen, Jingbo Xia, Christophe Roeder, and Lawrence E Hunter. 2016. Reproducibility in natural language processing: a case study of two R libraries for mining PubMed/MEDLINE. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2016, page 6. NIH Public Access.

Dina Demner-Fushman, Sonya E Shooshan, Laritza Rodriguez, Alan R Aronson, Francois Lang, Willie Rogers, Kirk Roberts, and Joseph Tonning. 2018. A dataset of 200 structured product labels annotated for adverse drug reactions. *Scientific data*, 5:180001.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mina Gachloo, Yuxing Wang, and Jingbo Xia. 2019. A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition. *Genomics & Informatics*, 17(2).

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen OConnor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. 2014. Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*, pages 1–8. Citeseer.

Nathalie Groen, Murat Guvendiren, Herschel Rabitz, William J Welsh, Joachim Kohn, and Jan De Boer. 2016. Stepping into the omics era: opportunities and challenges for biomaterials science and engineering. *Acta biomaterialia*, 34:133–142.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lawrence Hunter and Kevin Bretonnel Cohen. 2006. Biomedical language processing: what's beyond pubmed? *Molecular cell*, 21(5):589–594.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpus-a semantically annotated corpus for bio-text mining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9.

Jin-Dong Kim and Yue Wang. 2012. PubAnnotation: a persistent and sharable corpus and annotation repository. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, Toyofumi Fujiwara, Shujiro Okuda, Tiffany J Callahan, and Kevin Bretonnel Cohen. 2019. Open agile text mining for bioinformatics: the pubannotation ecosystem. *Bioinformatics*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. 1997. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Ying Hong Li, Chun Yan Yu, Xiao Xu Li, Peng Zhang, Jing Tang, Qingxia Yang, Tingting Fu, Xiaoyu Zhang, Xuejiao Cui, Gao Tu, et al. 2017. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic acids research*, 46(D1):D1121–D1127.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Majid Rastegar-Mojarad, Zhan Ye, Jill M Kolesar, Scott J Hebbring, and Simon M Lin. 2015. Opportunities for drug repositioning from phenome-wide association studies. *Nature biotechnology*, 33(4):342.

Kirk Roberts, Dina Demner-Fushman, and Joseph M Tonning. 2017. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *TAC*.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Miguel Vazquez, Martin Krallinger, Florian Leitner, and Alfonso Valencia. 2011. Text mining for drugs and chemical compounds: methods, tools and applications. *Molecular Informatics*, 30(6-7):506–519.

Yuxing Wang, Xinzhi Yao, Kaiyin Zhou, Xuan Qin, Jin-Dng Kim, Kevin B Cohen, and Jingbo Xia. 2018. Guideline design of an active gene annotation corpus for the purpose of drug repurposing. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI 2018), Oct, 2018, Beijing.(2018, accepted)*.

Zhong-Yi Wang and Hong-Yu Zhang. 2013. Rational drug repositioning by medical genetics. *Nature biotechnology*, 31(12):1080.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2017. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Kaiyin Zhou, Yuxing Wang, Sheng Zhang, Mina Gachloo, Jin-Dong Kim, Qi Luo, Kevin Bretonnel Cohen, and Jingbo Xia. 2019. Gof/lof knowledge inference with tensor decomposition in support of high order link discovery for gene, mutation and disease. *Math Biosci Eng*, 16(16):1376–1391.