The Genia Event Extraction Shared Task, 2013 Edition - Overview

Jin-Dong Kim and Yue Wang and Yamamoto Yasunori
Database Center for Life Science (DBCLS)
Research Organization of Information and Systems (ROIS)
{jdkim|wang|yy}@dbcls.rois.ac.jp

Abstract

The Genia Event Extraction task is organized for the third time, in BioNLP Shared Task 2013. Toward knowledge based construction, the task is modified in a number of points. As the final results, it received 12 submissions, among which 2 were withdrawn from the final report. This paper presents the task setting, data sets, and the final results with discussion for possible future directions.

1 Introduction

Among various resources of life science, literature is regarded as one of the most important types of knowledge base. Nevertheless, lack of explicit structure in natural language texts prevents computer systems from accessing fine-grained information written in literature. BioNLP Shared Task (ST) series (Kim et al., 2009; Kim et al., 2011a) is one of the community-wide efforts to address the problem. Since its initial organization in 2009, BioNLP-ST series has published a number of fine-grained information extraction (IE) tasks motivated for bioinformatics projects. Having solicited wide participation from the community of natural language processing, machine learning, and bioinformatics, it has contributed to the production of rich resources for fine-grained BioIE, e.g., TEES1 (Björne and Salakoski, 2011), SBEP2 (McClosky et al., 2011) and EVEX3 (Van Landeghem et al., 2011).

The Genia Event Extraction (GE) task is a seminal task of BioNLP-ST. It was first organized as the sole task of the initial 2009 edition of BioNLP-ST. The task was originally designed and implemented based on the Genia event corpus (Kim et al., 2008b) which represented domain knowledge around NFκB proteins. There were also some efforts to explore the possibility of literature mining for pathway construction (Kim et al., 2008a; Oda et al., 2008). The GE task was designed to make such an effort a community-driven one by sharing available resources, e.g., benchmark data sets, and evaluation tools, with the community.

In its second edition (Kim et al., 2011b) organized in BioNLP-ST 2011 (Kim et al., 2011a), the data sets were extended to include full text articles. The data sets consisted of two collections. The abstract collection, that had come from the first edition, was used again to measure the progress of the community between 2009 and 2011 editions, and the full text collection, that was newly created, was used to measure the generalization of the technology to full text papers.

In its third edition this year, while succeeding the fundamental characteristics from its previous editions, the GE task tries to evolve with the goal to make it a more “real” task toward knowledge base construction. The first design choice to address the goal is to construct the data sets fully with recent full papers, so that the extracted pieces of information can represent up-to-date knowledge of the domain. The abstract collection, that had been already used twice (in 2009 and 2011), is removed from official evaluation this time. Second, GE task subsumes the coreference task which has long been considered critical for improvement of event extraction performance. It is implemented by providing coreference annotation in integration with event annotation in the data sets.

The paper explains the task setting and data sets, presents the final results of participating systems, and discusses notable observations with conclusions.

1https://github.com/jbjorne/TEES/wiki
2http://nlp.stanford.edu/software/eventparser.shtml
3http://www.evexdb.org/
2 Task setting

This section explains the task setting of the 2013 edition of the GE task with a focus on changes to previous editions. For comprehensive explanation, readers are referred to Kim et al. (2009).

The changes made to the task setting are three-fold, among which two are about event types to be extracted. Table 1 shows the event types and their arguments targeted in the 2013 edition. First, four new event types are added to the target of extraction: the Protein modification type and its three sub-types, Ubiquitination, Acetylation, Deacetylation. Second, The Protein modification types are modified to be directly linked to causal entities, which was only possible through Regulation events in previous editions.

The modifications were made based on analysis on preliminary annotation during preparation of the data sets: in recent papers on $NF_kB$, discussions on protein modification were observed with non-trivial frequency. However, in the end, it turned out that the influence of the above modifications was trivial in terms of the number of annotated instances in the final data sets, as shown in section 3, after filtering out events on non-individual proteins, e.g., protein families, protein complexes.

Third change made to the task setting is addition of coreference and part-of annotations to the data sets. It is to address the observation from 2009 edition that coreference structures and entity relations often hide the syntactic paths between event triggers and their arguments, restricting the performance of event extraction. In 2011, the Protein coreference task and Entity Relation were organized as sub-tasks, to explicitly address the problem, but this time, coreference and part-of annotations are integrated in the GE task, to encourage an integrative use of them for event extraction. Figure 1 shows an example of annotation with coreference and part-of annotations. Note that the event representation in the figure is relation-centric, which is different from the event-centric representation of the default BioNLP-ST format. The two representations are interchangeable, and the GE task provides data sets in both formats, together with an automatic converter between them. Below is the corresponding annotation in the BioNLP-ST format:

T8 Protein 933 938 TRAF1
T9 Protein 940 945 TRAF2
T10 Protein 947 952 TRAF3
T11 Protein 958 963 TRAF6
T12 Protein 1038 1042 CD40
T41 Anaphora 1058 1072 These proteins
T48 Binding 1112 1119 binding
T49 Entity 1127 1143 cytoplasmic tail
T13 Protein 1147 1151 CD40
R1 Coreference Subject:T41 Object:T8
R2 Coreference Subject:T41 Object:T9
R3 Coreference Subject:T41 Object:T10
R4 Coreference Subject:T41 Object:T11
E4 Binding:T48 Theme:T8 Theme2:T13 Site2:T49
E5 Binding:T48 Theme:T9 Theme2:T13 Site2:T49
E6 Binding:T48 Theme:T10 Theme2:T13 Site2:T49
E7 Binding:T48 Theme:T11 Theme2:T13 Site2:T49

In the example, the event trigger, binding, denotes four binding events, in which the four proteins, TRAF1, TRAF2, TRAF3, and TRAF6, bind to the protein, CD40, respectively, through the site, cytoplasmic tail. The links between the four

---

Table 1: Event types and their arguments for Genia Event Extraction task. The type of each filler entity is specified in parenthesis. Arguments that may be filled more than once per event are marked with “+”, and optional arguments are with “?”.

| Event Type                  | Primary Argument          | Secondary Argument       |
|-----------------------------|----------------------------|--------------------------|
| Gene_expression             | Theme(Protein)            |                          |
| Transcription               | Theme(Protein)            | Loc(Protein)             |
| Localization               | Theme(Protein)            |                          |
| Protein_catabolism          | Theme(Protein)            |                          |
| Binding                     | Theme(Protein)+           | Site(Protein)*           |
| Protein_modification        | Theme(Protein), CSite(Protein) | Site(Protein)?       |
| Phosphorylation             | Theme(Protein), CSite(Protein) | Site(Protein)?       |
| Ubiquitination              | Theme(Protein), CSite(Protein) | Site(Protein)?       |
| Acetylation                 | Theme(Protein), CSite(Protein) | Site(Protein)?       |
| Deacetylation               | Theme(Protein), CSite(Protein) | Site(Protein)?       |
| Regulation                  | Theme(Protein), CSite(Protein) | Site(Protein)?       |
| Positive_regulation         | Theme(Protein), CSite(Protein) | Site(Protein)?       |
| Negative_regulation         | Theme(Protein), CSite(Protein) | Site(Protein)?       |
proteins and the event trigger are however very hard to find, without being bridged by the demonstrative noun phrase (NP), These proteins. In the case, if the link between the demonstrative NP, These proteins and its four antecedents, TRAF1, TRAF2, TRAF3, and TRAF6, can be somehow detected, the remaining link, between the demonstrative NP and the trigger, may be detected by their syntactic connection. A key point here is the different characteristics of the two step links: detecting the former is rather semantic or discoursal while the latter may be a more syntactic problem. Then, solving them using different processes would make a sense. To encourage an exploration into the hypothesis, the coreference annotation is provided in the training and development data sets.

Based on the definition of event types, the entire task is divided into three sub-tasks addressing event extraction at different levels of specificity:

**Task 1. Core event extraction** addresses the extraction of typed events together with their primary arguments.

**Task 2. Event enrichment** addresses the extraction of secondary arguments that further specify the events extracted in Task 1.

**Task 3. Negation/Speculation detection** addresses the detection of negations and speculations over the extracted events.

For more detail of the subtasks, readers are referred to Kim et al. (2011b).

| Item                      | Training | Devel | Test  |
|---------------------------|----------|-------|-------|
| Articles                  | 10       | 10    | 14    |
| Words                     | 54938    | 57907 | 75144 |
| Proteins                  | 3571     | 4138  | 4359  |
| Entities                  | 121      | 314   | 327   |
| Events                    | 2817     | 3199  | 3348  |
| Gene_expression           | 729      | 391   | 619   |
| Transcription             | 122      | 98    | 101   |
| Localization              | 44       | 197   | 99    |
| Protein_catabolism        | 23       | 30    | 14    |
| Binding                   | 195      | 376   | 342   |
| Protein_modification      | 8        | 1     | 1     |
| Phosphorylation           | 117      | 197   | 161   |
| Ubiquitination            | 4        | 2     | 30    |
| Acetylation               | 0        | 3     | 0     |
| Deacetylation             | 0        | 5     | 0     |
| Regulation                | 299      | 284   | 299   |
| Positive_regulation       | 780      | 883   | 1144  |
| Negative_regulation       | 496      | 532   | 538   |
| Coreferences              | 178      | 160   | 197   |
| to Protein                | 152      | 123   | 169   |
| to Entity                 | 5        | 6     | 6     |
| to Event                  | 18       | 27    | 13    |
| to Anaphora               | 3        | 4     | 9     |

Table 2: Statistics of annotations in training, development, and test sets

### 3 Data Preparation

As discussed in section 1, for the 2013 edition, the data sets are constructed fully with full text papers. Table 2 shows statistics of three data sets for training, development and test. The data sets consist of 34 full text papers from the Open Access subset of PubMed Central. The papers were retrieved using lexical variants of the term, “NFκB” as primary keyword, and “pathway” and “regulation” as secondary keywords. The retrieved papers were given to the annotators with higher priority...
Table 3: Statistics of annotations in different sections of text: the Abstract column is of the abstraction collection (1210 titles and abstracts), and the following columns are of full paper collection (14 full papers). TIAB = title and abstract, Intro. = introduction and background, R/D/C = results, discussions, and conclusions, Methods = methods, materials, and experimental procedures. Some minor sections, supporting information, supplementary material, and synopsis, are ignored. Density = relative density of annotation (P/W = Protein/Word, E/W = Event/Word, and E/P = Event/Protein).

Figure 2: Event distribution in different sections to newer ones. Note that among 34 papers, 14 were from the full text collection of 2011 edition data sets, and 20 were newly collected this time. The annotation to the all 34 papers were produced by the same annotators who also produced annotations for the previous editions of GE task.

The annotated papers are divided into the training, development, and test data sets; 10, 10, and 14, respectively. Note that the size of the training data set is much smaller than previous editions, in terms of number of words and events, while the size of the development and test data sets are comparable to previous editions. It is the consequence of a design choice of the organizers with the notion that (1) relevant resources are substantially accumulated through last two editions, and that (2) therefore the importance of training data set may be reduced while the importance of development and test data sets needs to be kept. Instead, participants may utilize, for example, the abstract collection of the 2011 edition, of which the annotation was produced by the same annotators with almost same principles. As another example, the data sets of the EPI task (Ohta et al., 2011) also may be utilized for the newly added protein modification events.

Table 3 shows the statistics of annotated event types in different sections of the full papers in the data sets. For the analysis, the sections are classified to five groups as follows:

- The TIAB group includes the titles and abstracts. In the GE-2011 data sets, the corresponding files match the pattern, PMC-*TIAB*.txt.
- The Intro group includes sections for introduction, and background. The corresponding files match the pattern, PMC-*@(-|._)@(I|Back)*.txt.
The R/D/C group includes sections on results, discussions, and conclusions. The files match the pattern, PMC-@(-|_).@(R|D|Conc).*.txt

- The Methods group includes sections on methods, materials, and experimental procedures. The files match the pattern, PMC-@(-|_).@(Met|Mat|MAT|E).*.txt

- The Caption group includes the captions of tables and figures. The corresponding files match the pattern, PMC-aption*.txt.

Figure 2 illustrates the different distribution of annotated event types in the five section groups. It shows that the Methods group has significantly different distribution of annotated events, confirming a similar observation reported in Kim et al. (2011b).

4 Participation

The GE task received final submissions from 12 teams, among which 2 were withdrawn from final report. Table 4 summarizes the teams. Unfortunately, the subtasks 2 and 3 did not meet a large participation.

Table 5 profiles the participating systems. The systems are roughly grouped into SVM-based pipeline (EVEX, TEEs-2.1, and DlutNLp), rule-based pipeline (BioSEM and UZH), mixed pipeline (USheff and HCUMUS), joint pattern matching (NCBI and NICTANLM), and joint SVM (HDS4NLp) systems. In terms of use of external resources, 5 teams (EVEX, TEEs-2.1, NCBI, DlutNLp, and USheff) utilized data sets from 2011 edition, and two teams (HDS4NLp and NICTANLM) utilized independent resources, e.g., UniProt (Bairoch et al., 2005), IntAct (Kerrien et al., 2012), and CRAFT (Verspoor et al., 2012).

5 Results and Discussions

Table 6 shows the final results of subtask 1. Overall EVEX, TEEs-2.1, and BioSEM show the best performance only with marginal difference between them. In detail, the performance of BioSEM is significantly different from EVEX and TEEs-2.1: (1) while BioSEM show the best performance with Binding and Protein_modification events, EVEX and TEEs-2.1 show the best performance with Regulation events which takes the largest portion of annotation in data sets; and (2) while the performance of EVEX and TEEs-2.1 is balanced over recall and precision, BioSEM is biased for precision, which is a typical feature of rule-based systems. It is also notable that BioSEM has achieved a near best performance using only shallow parsing. Although it is not shown in the table, NCBI is the only system which produced Ubiquitination events, which is interpreted as a result of utilizing 2011-EPI data sets (Ohta et al., 2011) for the system development.

Table 7 shows subtask 1 final results only within TIAB sections. It shows that the systems developed utilizing previous resources, e.g., 2011 data sets, and EVEX, perform better for titles and abstracts, which makes sense because those resources are title and abstract-centric.

Tables 8 and 9 show evaluation results within Methods and Captions section groups, respectively. All the systems show their worst performance in the two section groups. Especially the drop of performance with regulation events is huge. Note the two section groups also show significantly different event distribution compared to other section groups (see section 3). It suggests that language expression in the two section groups may be quite different from other sections, and an extensive examination is required to get a reasonable performance in the sections.

Table 10 and 11 show final results of Task 2 (Event enrichment) and 3 (Negation/Speculation detection), respectively, which unfortunately did not meet a large participation.

6 Conclusions

In its third edition, the GE task is fully changed to a full text paper centric task, while the online evaluation service on the abstract-centric data sets.
Table 5: System profiles: SnowBall=SnowBall Stemmer, CNLP=Stanford CoreNLP (tokenization), McCCJ=McClosky-Charniak-Johnson Parser, Stanford=Stanford Parser, S.=Speculation, N.=Negation

| Team    | Simple Event | Binding | Prot-Mod. | Regulation | All     |
|---------|--------------|---------|-----------|------------|---------|
| EVEX    |              |         |           |            | 48.44   |
| TEE2-2.1|              |         |           |            | 58.03   |
| BioSEM  |              |         |           |            | 50.97   |
| NCBI    |              |         |           |            |         |
| DlinFLP |              |         |           |            |         |
| HDSANLP |              |         |           |            |         |
| NICTANLM|              |         |           |            |         |
| UShef   |              |         |           |            |         |
| UZH     |              |         |           |            |         |
| HCMUS   |              |         |           |            |         |

Table 6: Evaluation results (recall / precision / f-score) of Task 1. Some notable figures are emphasized in bold.

is kept maintained. Unfortunately, the coreference annotation, which has been integrated in the event annotation in the data sets, was not exploited by the participants, during the official shared task period. An analysis shows that the performance of systems significantly drops in the Methods and Captions sections, suggesting for an extensive examination in the sections.

As usual, after the official shared task period, the GE task is maintaining an online evaluation that can be freely accessed by anyone but with a time limitation; once in 24 hours per a person. With a few new features that are introduced in 2013 editions but are not fully exploited by the participants, the organizers solicit participants to continuously explore the task using the online evaluation. The organizers are also planning to provide more resources to the participants, based on the understanding that interactive communication between organizers and participants is important for progress of the participating systems and also the task itself.

References

Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O’Donovan, Nicole Redaschi, and Lai-Su L. Yeh. 2005. The universal protein resource (uniprot). Nucleic Acids Research, 33(suppl 1):D154–D159.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In Proceedings of BioNLP Shared Task 2011 Workshop, pages 183–191, Portland, Oregon, USA, June. Association for Computational Linguistics.

Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C. Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeifferenger, Pablo Porras, Arathi Raghu Nath, Bernd Röchert, Sandra Orchard, and Henning Hermjakob. 2012. The intact molecular interaction database in 2012. Nucleic Acids Research, 40(D1):D154–D159.

Jin-Dong Kim, Tomoko Ohta, Kanae Oda, and Jun’ichi Tsujii. 2008a. From text to pathway: corpus annotation for knowledge acquisition from biomedical literature. In Proceedings of the 6th Asia Pacific Bioinformatics Conference, Series on Advances in Bioin-
Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

Overview of BioNLP’09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Overview of the Genia Event task in BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1635, Portland, Oregon, USA, June. Association for Computational Linguistics.

New challenges for text mining: Mapping between text and manually curated pathways.

Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Evex: A pubmed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of BioNLP 2011 Workshop*, pages 28–37, Portland, Oregon, USA, June. Association for Computational Linguistics.

A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(1):207.
Table 7: Evaluation results (recall / precision / f-score) of Task 1 in titles and abstracts. Some notable figures are emphasized in bold.

| Team    | Simple Event | Binding | Prot-Mod. | Regulation | All     |
|---------|--------------|---------|-----------|------------|---------|
| NICTANLM| 70.83 / 90.44 / 79.44 | 48.24 / 53.93 / 50.93 | 74.17 / 82.41 / 78.07 | 28.74 / 25.17 / 35.63 | 42.97 / 46.90 / 51.70 |
| TEES-2.1| 73.51 / 83.26 / 78.08 | 43.72 / 47.80 / 45.67 | 66.67 / 66.12 / 66.39 | 32.79 / 46.79 / 38.56 | 45.29 / 58.05 / 50.88 |
| NCBI    | 74.09 / 83.37 / 78.46 | 43.72 / 47.80 / 45.67 | 66.67 / 65.04 / 65.84 | 33.24 / 44.48 / 38.04 | 47.50 / 56.34 / 50.46 |
| DuEnLP  | 70.06 / 84.49 / 76.60 | 39.20 / 44.32 / 41.60 | 67.50 / 74.31 / 70.74 | 27.78 / 43.23 / 33.83 | 41.01 / 56.70 / 47.60 |
| NICTANLM| 75.24 / 57.14 / 64.95 | 35.68 / 41.76 / 38.48 | 52.50 / 76.83 / 62.38 | 22.33 / 46.83 / 30.24 | 37.73 / 52.30 / 43.84 |
| UShef   | 56.81 / 80.43 / 66.59 | 32.66 / 48.15 / 38.92 | 45.00 / 94.74 / 61.02 | 21.67 / 53.55 / 30.85 | 32.75 / 63.93 / 42.89 |
| HDS4NLP | 76.20 / 84.65 / 80.20 | 41.21 / 38.14 / 39.61 | 75.83 / 75.21 / 75.52 | 16.58 / 30.16 / 21.40 | 36.19 / 51.26 / 42.42 |
| UZH     | 63.53 / 78.25 / 70.13 | 32.19 / 23.82 / 25.63 | 66.67 / 74.07 / 70.18 | 10.61 / 29.39 / 15.59 | 27.36 / 50.89 / 35.58 |
| HCMUS   | 67.18 / 62.84 / 64.94 | 38.19 / 28.15 / 32.41 | 67.50 / 70.13 / 64.54 | 19.45 / 20.11 / 19.78 | 35.09 / 33.95 / 34.51 |

Table 8: Evaluation results (recall / precision / f-score) of Task 1 in Methods section group. Some notable figures are emphasized in bold.

| Team    | Simple Event | Binding | Prot-Mod. | Regulation | All     |
|---------|--------------|---------|-----------|------------|---------|
| TEES-2.1| 76.67 / 67.65 / 71.88 | 53.19 / 46.30 / 49.50 | 60.61 / 67.92 / 67.80 | 22.68 / 39.29 / 28.76 | 43.41 / 53.74 / 48.02 |
| BioSEM  | 60.00 / 76.26 / 67.92 | 68.09 / 58.18 / 62.75 | 69.70 / 82.14 / 75.41 | 23.20 / 34.35 / 27.69 | 42.31 / 54.42 / 47.60 |
| EVEX    | 76.67 / 67.65 / 71.88 | 53.19 / 46.30 / 49.50 | 48.48 / 72.73 / 58.18 | 21.13 / 39.81 / 27.61 | 41.48 / 53.74 / 46.82 |
| DuEnLP  | 70.00 / 67.02 / 68.48 | 55.32 / 48.15 / 51.49 | 57.58 / 79.17 / 66.67 | 18.04 / 46.67 / 26.02 | 39.29 / 57.89 / 48.61 |
| NCBI    | 80.00 / 58.54 / 67.61 | 40.43 / 41.30 / 40.86 | 66.67 / 70.97 / 68.75 | 14.95 / 44.62 / 22.39 | 39.01 / 53.58 / 45.15 |
| HDS4NLP | 78.89 / 78.02 / 78.45 | 48.94 / 29.49 / 36.80 | 66.67 / 68.75 / 67.69 | 06.19 / 14.63 / 08.70 | 35.16 / 45.25 / 39.57 |
| UZH     | 57.78 / 66.42 / 62.65 | 23.40 / 26.19 / 24.72 | 69.70 / 74.19 / 71.88 | 12.89 / 43.10 / 19.84 | 30.49 / 52.62 / 38.88 |
| UShef   | 47.78 / 74.14 / 58.11 | 36.17 / 45.95 / 40.48 | 30.30 / 100.00 / 46.51 | 13.40 / 45.61 / 20.72 | 26.37 / 59.26 / 36.50 |
| NICTANLM| 75.56 / 53.12 / 62.39 | 40.33 / 27.94 / 33.04 | 18.18 / 44.55 / 27.27 | 11.34 / 36.67 / 17.32 | 31.59 / 43.07 / 36.45 |
| HCMUS   | 73.33 / 52.80 / 61.40 | 53.19 / 25.51 / 34.48 | 63.64 / 53.58 / 58.33 | 15.46 / 17.96 / 16.62 | 39.01 / 33.10 / 35.81 |

Table 9: Evaluation results (recall / precision / f-score) of Task 1 in Captions section group. Some notable figures are emphasized in bold.

| Team    | Simple Event Binding Site-Phosphorylation Loc-Localization Total |
|---------|-----------------|-----------------------------|----------------------------|----------|
| TEES-2.1| 31.37 / 56.14 / 40.25 | 37.21 / 82.05 / 51.20 | 36.67 / 78.57 / 50.00 | 22.03 / 61.90 / 32.50 |
| EVEX    | 31.37 / 56.14 / 40.25 | 32.56 / 80.00 / 46.28 | 36.67 / 78.57 / 50.00 | 20.90 / 61.67 / 31.22 |

Table 10: Evaluation results (recall / precision / f-score) of Task 2

| Team    | Negation | Speculation | Total |
|---------|----------|-------------|-------|
| TEES-2.1| 21.68 / 36.84 / 27.30 | 18.46 / 33.96 / 23.92 | 19.53 / 35.59 / 25.22 |
| EVEX    | 20.98 / 38.03 / 27.04 | 18.46 / 32.73 / 23.61 | 19.82 / 34.41 / 25.15 |
| NICTANLM| 15.38 / 32.76 / 20.94 | 14.36 / 34.15 / 20.22 | 14.79 / 33.57 / 20.54 |

Table 11: Evaluation results (recall / precision / f-score) of Task 3