Comparable Study of Event Extraction in Newswire and Biomedical Domains

Makoto Miwa[†,‡] **Paul Thompson**[†] **Ioannis Korkontzelos**[†] **Sophia Ananiadou**[†]
[†]National Centre for Text Mining and School of Computer Science,
University of Manchester, United Kingdom
[‡]Graduate School of Engineering, Toyota Technological Institute, Japan
{makoto.miwa, paul.thompson, ioannis.korkontzelos, sophia.ananiadou}@manchester.ac.uk

Abstract

Event extraction is a popular research topic in natural language processing. Several event extraction tasks have been defined for both the newswire and biomedical domains. In general, different systems have been developed for the two domains, despite the fact that the tasks in both domains share a number of characteristics. In this paper, we analyse the commonalities and differences between the tasks in the two domains. Based on this analysis, we demonstrate how an event extraction method originally designed for the biomedical domain can be adapted for application to the newswire domain. The performance is state-of-the-art for both domains, with F-scores of 52.7% for the biomedical domain and 52.1% for the newswire domain in terms of their primary evaluation metrics.

1 Introduction

Research into event extraction was initially focussed on the general language domain, largely driven by the Message Understanding Conferences (MUC) series (e.g., Chinchor (1998)) and the Automated Content Extraction (ACE) evaluations[1]. More recently, the focus of research has been widened to the biomedical domain, motivated by the ongoing series of biomedical natural language processing (BioNLP) shared tasks (STs) (e.g., Kim et al. (2013)).

Although the textual characteristics and the types of relevant events to be extracted can vary considerably between domains, the same general features of events normally hold across domains. An event usually consists of a trigger and arguments (see Figures 1 and 2.) A trigger is typically a verb or a nominalised verb that denotes the presence of the event in the text, while the arguments are usually entities. In general, arguments are assigned semantic roles that characterise their contribution towards the event description.

Until now, however, there has been little, if any, effort by researchers working on event extraction in different domains to share ideas and techniques, unlike syntactic tasks (e.g., (Miyao and Tsujii, 2008)) and other information extraction tasks, such as named entity recognition (e.g., (Giuliano et al., 2006)) and relation extraction (e.g., (Qian and Zhou, 2012)). This means that the potential to exploit cross-domain features of events to develop more adaptable event extraction systems is an under-studied area. Consequently, although there is a large number of published studies on event extraction, proposing many different methods, no work has previously been reported that aims to adapt an event extraction method developed for one domain to a new domain.

In response to the above, we have investigated the feasibility of adapting an event extraction method developed for the biomedical domain to the newswire domain. To facilitate this, we firstly carry out a detailed static analysis of the differences that hold between event extraction tasks in the newswire and biomedical domains. Specifically, we consider the ACE 2005 event extraction task (Walker et al., 2006) for the newswire domain and the Genia Event Extraction task (GENIA) in BioNLP ST 2013 (Kim et al., 2013) for the biomedical domain. Based on the results of this analysis, we adapt the biomedical event

[1] itl.nist.gov/iad/mig/tests/ace

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2270–2279, Dublin, Ireland, August 23-29 2014.

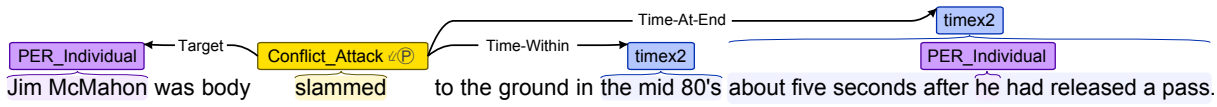

Figure 1: ACE 2005 event example (ID: MARKBACKER_20041220.0919)

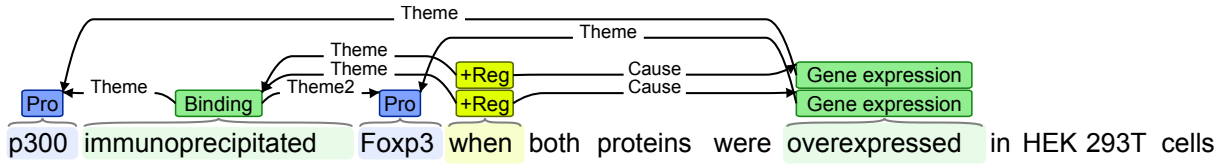

Figure 2: GENIA event example (ID: PMC-1447668-08-Results)

extraction method to the task of extracting events in the newswire domain, according to the specification of the ACE 2005 event extraction task. The original method consists of a classification pipeline that has previously been applied to extract events according to task descriptions that are similar to GENIA. In order to address the differences between this task and the ACE task, we have made a number of changes to the original method, including modifications to the classification labels assigned, the pipeline itself and the features used. We retrained the model of the adapted system on the ACE task, compared the performance, and empirically analysed the differences between the two tasks in terms of entity-related information. We demonstrate that the resulting system achieves state-of-the-art performance for tasks in both domains.

2 Related Work

In this section, we introduce the two domain specific event extraction tasks on which we will focus, i.e., the ACE 2005 event extraction task, which concerns events in the newswire domain, and the GENIA event task from the BioNLP ST 2013, which deals with biomedical event extraction. We also examine state-of-the-art systems that have been developed to address each task.

2.1 Newswire Event Extraction

The extraction of events from news-related texts has been widely researched, largely due to motivation from the various MUC and ACE shared tasks. Whilst MUC focussed on filling a single event template on a single topic by gathering information from different parts of a document, ACE defined a more comprehensive task, involving the recognition of multiple fine-grained and diverse types of entities and associated intra-sentential events within each document.

A common approach to tackling the MUC template filling task has involved the employment of pattern-based methods, e.g., Riloff (1996). In contrast, supervised learning approaches have constituted a more popular means of approaching the ACE tasks[2]. In this paper, we choose to focus on adapting our biomedical-focussed event extraction method to the ACE 2005 task. Our choice is based on the task definition for ACE 2005 having more in common with the BioNLP 2013 GENIA ST definition than the MUC event template task definition.

In terms of the characteristics of state-of-the-art event extraction systems designed according to the ACE 2005 model, pipeline-based approaches have been popular (Grishman et al., 2005; Ahn, 2006). Grishman et al. (2005) proposed a method that sequentially identifies textual spans of arguments, role types, and event triggers. This pipeline approach has been further extended in several subsequent studies. For example, Liao et al. (2010) investigated document-level cross-event consistency using co-occurrence of events and event arguments, while Hong et al. (2011) exploited information gathered from the web to ensure cross-entity consistency.

[2]Note that there are also approaches using few or no training data (e.g., (Ji and Grishman, 2008; Lu and Roth, 2012)) for the ACE 2005 task, but they are not so many and we will focus on the supervised learning approaches in this paper.

Li et al. (2013) recently proposed a joint detection method to detect both triggers and arguments (together with their role types) using a structured perceptron model. The system outperformed the best results reported for the ACE 2005 task in the literature, without the use of any external resources.

2.2 Biomedical Event Extraction

The task of event extraction has received a large amount of attention from BioNLP researchers in recent years. Interest in this task was largely initiated by the BioNLP 2009 ST, and has been sustained through the organisation of further STs in 2011 and 2013. The STs consist of a number of different sub-tasks, the majority of which concern the extraction of events from biomedical papers from the PubMed database. Events generally concern interactions between biomedical entities, such as proteins, cells and chemicals.

Similarly to newswire event extraction systems, pipeline-based methods have constituted a popular approach to extracting events in the biomedical domain (Björne and Salakoski, 2013; Miwa et al., 2012). The pipeline developed by Miwa et al. (2012) consists of a number of modules, which sequentially detect event triggers, event arguments, event structures and hedges (i.e., speculations and negations). The system has been applied to several event extraction tasks, and has achieved the best performance on most of these, in comparison to other systems. It should be noted that the ordering of the components in biomedical event extraction pipelines often differs from pipelines designed for news event extraction, e.g., Grishman et al. (2005), which was described above.

As in newswire event detection, some joint (non pipeline-based) approaches have also been proposed for biomedical event extraction. For example, McClosky et al. (2012) used a stacking model to combine the results of applying two different methods to event extraction. The first method is a joint method, similar to Li et al. (2013), that detects triggers, arguments and their roles. However, in contrast to the structured perceptron employed in Li et al. (2013), McClosky et al. (2012) use a dual-decomposition approach for the detection. The second method is based on dependency parsing and treats event structures as dependency trees.

3 Adaptation of Biomedical Event Extraction to Newswire Event Extraction

In this section, we firstly analyse the differences between the domain-specific ACE 2005 and GENIA event extraction tasks. Based on our findings, we propose an approach to adapting an existing event extraction method, originally developed for biomedical event extraction, to the ACE 2005 task, by resolving the observed differences between the two task definitions.

3.1 Differences in event extraction tasks

Both the ACE 2005 and GENIA tasks concern the task of event extraction, i.e., the identification of relationships between entities. For both tasks, the requirement is to extract events from text that conform to the general event description introduced earlier, i.e., a trigger and its arguments, each of which is assigned a semantic role. Despite this high-level similarity between the tasks, their finer-grained details diverge in a number of ways. Apart from the different textual domain, the tasks adopt varying annotation schemes. The exact kinds of annotations provided at training time are also different, as are the evaluation settings.

Several variants of the official task setting for the ACE 2005 corpus have been defined. This is partly due to the demanding nature of the official task definition, which requires the detection of events from scratch, including the recognition of named entities participating in events, together with the resolution of coreferences. Alternative task settings (such as Ji and Grishman (2008); Liao and Grishman (2010))) generally simplify the official task definition, e.g., by omitting the requirement to perform coreference resolution. A further issue is that the test data sets for the official task setting have not been made publicly available. As a result of the multiple existing variations of the ACE 2005 task definition that have been employed by different research efforts, direct comparison of our results with those obtained by other state-of-the art systems is problematic. The solution we have chosen is to adopt the same ACE 2005 event extraction task specification that has been adopted in recent research, by Hong et al. (2011) and Li et al. (2013). For GENIA, we follow the specification of the original GENIA event extraction task.

	ACE 2005	GENIA
# of entity types	13 (type) / 53 (subtype)	2
Argument	Entity/Nominal/Value/Time	Entity
# of event types	8 (type) / 33 (subtype)	13
# of argument role types	35	7
Max # of arguments for an event	11	4
Nested events	None	Possible
Overlaps of events	None	Possible
Correspondences of arguments	None	Possible
Entity	Available (Given)	Available (Partially given)
Entity attributes	Available (Given)	Not available
Event attributes	Available (Not given)	Available (Not given)
Entity coreference	Available (Given)	Available (Not given)
Event coreference	Available (Not given)	Not available
Evaluation	Trigger/Role	Event

Table 1: Comparison of event definitions and event extraction tasks. *"Available annotations"* are annotations available in the corresponding corpus, while *"Given annotations"* are annotations provided during (training and) prediction. *"Given annotations"* do not need to be predicted during event extraction.

Event annotation examples for ACE 2005 and GENIA are shown in Figures 1 and 2, respectively. Table 1 summarises the following comparison between the two event extraction tasks.

Semantic types There are more event, role and entity types and a greater potential number of arguments in ACE 2005 events than in GENIA events. There is also a hierarchy of event types and entity types in ACE 2005. For example, the *Life* event type has *Be-Born*, *Marry*, *Divorce*, *Injure*, *Die* event subtypes. Some GENIA event types can also be arranged to have a hierarchy but they are limited. Events in ACE 2005 can take non-entity arguments, e.g., *Time*.

Nested events/Overlapping events Event structures are flat in ACE 2005, but they can be nested in GENIA, i.e., an event can take other events as its arguments. Events in GENIA can also be overlapping, in the sense that a particular word or phrase can be a trigger for multiple events. Figure 2 illustrates both nesting and overlapping in GENIA events. These properties of GENIA events are not addressed by methods developed for event extraction according to the ACE 2005 specification, making direct application of these methods to the GENIA task impossible.

Links amongst arguments A specific feature of the GENIA event extraction task, which is completely absent from the ACE 2005 task, is that links amongst arguments sometimes have to be identified. For example, the *Binding* event type in the GENIA task can take the following argument role types: *Theme*, *Theme2*, *Site* and *Site2*. The number *2* is attached to differentiate specific linkages between arguments: *Site* is the location of *Theme*, while *Site2* is the location of *Theme2*.

Entities, events and their attributes Entities in ACE 2005 have rich attributes associated with them. For example, the *Time* entity type has an attribute to store a normalised temporal format (e.g., *2003-03-04* for entities "20030304", "March 4" and "Tuesday") while the *GPE (Geo-Political Entity)* type has attributes such as subtypes (e.g., *Nation*), mention type (proper name, common noun or pronoun), roles (location of a group or person) and style (*literal* or *metonymic*). In contrast, GENIA entities have no attributes[3]. In ACE 2005, all entities are provided (gold) in the training and test data and they do not need to be predicted. In GENIA, some named entities (i.e., *Protein*s) are also provided, but other types of named or non-named entities that can constitute event arguments, such as locations and sites of proteins, are not provided in the test data and thus need to be predicted as part of the extraction process. Events in both corpora also have associated attributes: modality,

[3]Types are not counted as attributes in this paper.

polarity, genericity and tense in ACE 2005 and negation and speculation in GENIA. The GENIA task definition requires event attributes to be predicted, but the ACE 2005 task definition does not.

Coreference Both entity and event coreference are annotated in ACE 2005, but only entity coreference is annotated in GENIA. Events in ACE 2005 can take non-entity mentions, such as pronouns, as their arguments. However, events in GENIA can take only entity mentions as arguments. Thus, instead of non-entity mentions, coreferent entity mentions that are the closest to triggers are annotated as arguments in GENIA. For example, in Figure 2, "*p300*" and "*Foxp3*" are annotated as *Theme*s of *Gene_expression* events instead of "*both proteins*".

Evaluation In ACE 2005, the accuracy of extracted events is evaluated at the level of individual arguments and their roles. Completeness of events is not taken into consideration (Li et al., 2013), presumably because each event can take many arguments. Evaluation is performed by taking into account the 33 event subtypes, rather than the 8 coarser-grained event types. In contrast, evaluation of events according to the GENIA specification considers only the correctness of *complete* events, after nested events have been broken down.

In summary, the ACE 2005 task is in some respects more complex than the GENIA task, because it concerns a greater number event types, whose arguments may constitute a greater range of entity types, and whose semantic roles are drawn from a larger set, some of which are specific to particular event types and entities. In other respects, the task is more straightforward than the GENIA task, because of the simpler nature of the event structures in ACE 2005, i.e., there are no nested or overlapping event structures.

3.2 Adaptation of event extraction method

Since event structures are simpler in ACE 2005 than GENIA, we choose to adapt a biomedical event extraction method to the ACE 2005 task rather than the other way around. The inverse adaptation, starting from a newswire event extraction method, is considered more complex, since we would need to extend the method to capture the more complex event structures required in the GENIA task. It would additionally be inappropriate to employ domain adaptation methods (Daumé III and Marcu, 2006; Pan and Yang, 2010) to allow GENIA-trained models to be applied to the ACE 2005 tasks. This is because such methods require that there is at least a certain degree of overlap between the target information types, which is not the case in this scenario.

We employ the biomedical event extraction pipeline method described in Miwa et al. (2012) as our starting point. Our motivation is that, due to their modular nature, pipeline approaches are often easier to adapt to other task settings than joint approaches, e.g., (McClosky et al., 2012; Li et al., 2013). In addition, the method has previously been shown to achieve state-of-the-art performance in several biomedical event extraction tasks (Miwa et al., 2012).

The pipeline consists of four detectors, i.e., trigger/entity, event role, event structure, and hedge detectors. The trigger/entity detector finds triggers and entities in text. The event role detector determines which triggers/entities constitute arguments of events, links them to the appropriate event trigger and assigns semantic roles to the arguments. The event structure detector merges trigger-argument pairs into all possible complete event structures, and determines which of these structures constitute actual events. The same detector determines links between arguments, such as *Theme2* and *Site2*. The hedge detector finds negation and speculation information associated with events. Each detector solves multi-label multi-class classification problems using lexical and syntactic features obtained from multiple parsers. These features include character n-grams, word n-grams, and shortest paths between triggers and participants within parse structures. More detailed information can be found in Miwa et al. (2012).

We have updated the original method by simplifying the format of the classification labels used by both the event role detector and event structure detector modules. We refer to this method as *BioEE*, which we have applied to the GENIA task. We use only the role types (e.g., *Theme*) as classification labels for instances in the event role detector, instead of the more complex labels used in the original version of the module, which combined event types, roles and semantic entity types of arguments (e.g.,

Binding:Theme-Protein). Similarly, in the event structure detector, we use only two labels ("EVENT" or "NOT-EVENT"), instead of the previously used composite labels, which consisted of the event type, together with the roles and semantic entity types of all arguments of the event (e.g., *Regulation:Cause-Protein:Theme-Protein*.) We employed the simplified labels, since they increase the number of training instances for each label. The use of such labels, compared to the more complex ones, could reduce the potential of carrying out detailed modelling of specific aspects of the task. However, this was found not to be an issue, since the use of the simplified labels improved the performance of the pipeline in detecting events within the GENIA development data set (about 1% improvement in F-score). The simplification of the set of classification labels was also vital to ensure the tractability of the classification problems within the context of the ACE 2005 task. For example, using the same conventions to formulate classification labels as in the original system would result in 345 possible labels (compared to 91 in GENIA) to be predicted by the event role detector (and an even greater number of labels for the event structure detector), based on event-role-semantic type combinations found in the ACE training/development sets.

In order to adapt the system to extract events according to the ACE 2005 specification, we modified BioEE in several ways, making changes to both the pipeline itself and the features employed by the different modules. We refer to this method as *Adapted BioEE*, and we applied this method to the ACE 2005 task. These changes were made in an attempt to address the two major differences between the GENIA and ACE 2005 tasks, i.e., the simpler event structures and the availability of entity attribute and coreference information in ACE.

The pipeline-based modifications consisted of removing certain modules from the original pipeline, such that only two modules remained, i.e., the trigger/entity and event role detectors. The other two modules of the original pipeline, i.e., the event structure and hedge detectors, were designed to deal with problems that do not exist in the ACE 2005 extraction task, and thus their usage would be redundant. Instead of using the event structure detector to piece the different elements of an event, we simply aggregate all the arguments of the same trigger into a single event structure, after the event role detector has been applied.

As mentioned above, the ACE 2005 task definition includes rich information about entities, including attributes and coreference information. Existing systems developed to address this task have exploited this information to generate rich feature sets for classification (Liao and Grishman, 2010; Li et al., 2013). Based on the demonstrated utility of this information within the context of event extraction, we also choose to use it, by adding binary feature that indicate the presence of base forms, entity subtypes, and attributes of the entities and their coreferent entities to features in both detectors above. We choose to use base forms, since surface forms of entities are not used by most biomedical event extraction systems, including BioEE. We also add the features for Brown clusters (Brown et al., 1992) following Li et al. (2013). Further details can be found in Li et al. (2013).

4 Evaluation

4.1 Evaluation settings

To assess the performance of Adapted BioEE on the ACE 2005 task, we followed the evaluation process and settings used in previously reported studies (Hong et al., 2011; Li et al., 2013). ACE 2005 consists of 599 documents. In order to facilitate direct comparison with other systems trained on the same data, we conducted a blind test on the same 40 newswire documents that were used for evaluation in (Ji and Grishman, 2008; Li et al., 2013), and used the remaining documents as training/development sets. We use precision (P), recall (R) and F-score (F) to report the performance of the adapted system in classifying triggers and argument roles. We use the latter F-score as our primary metric for comparing our system with other systems, since this score better reflects the performance of the extraction of event structures.

GENIA consists of 34 full paper articles (Kim et al., 2013). To evaluate the performance of BioEE on the GENIA task, we followed the task setting in BioNLP ST 2013 and used the official evaluation systems provided by the organisers. We also used the same partitioning of data that was employed in the official BioNLP ST 2013 evaluation, with 20 articles being used as the training/development set, and the remaining 14 articles being held back as the test set. For brevity, we show the only the primary P,

	Arg. Role Decomposition			Event Detection		
	P	R	F	P	R	F (%)
BioEE	71.76	47.44	57.12	64.36	44.62	52.71
BioEE (+Entity)	69.47	46.94	56.02	61.81	44.11	51.48
EVEX	64.30	48.51	55.30	58.03	45.44	50.97
TEES-2.1	62.69	49.40	55.26	56.32	46.17	50.74

Table 2: Overall performance of BioEE on the GENIA data set

	Trigger Classification			Arg. Role Classification			Event Detection		
	P	R	F	P	R	F	P	R	F (%)
Adapted BioEE	59.9	72.6	65.7	54.2	50.2	52.1	20.7	21.7	21.2
Adapted BioEE (-Entity)	57.9	71.5	64.0	51.0	48.1	49.5	19.7	19.3	19.5
Li et al. (2013)	73.7	62.3	67.5	64.7	44.4	52.7	-	-	-
Hong et al. (2011)	72.9	64.3	68.3	51.6	45.5	48.4	-	-	-

Table 3: Overall performance of Adapted BioEE on the ACE 2005 data set

R and F scores in the shared task, i.e., the EVENT TOTAL results obtained using the approximate span & recursive evaluation method, as recommended by the organisers. The method individually evaluates each complete *core* event, i.e., event triggers with their *Theme* and/or *Cause* role arguments, with relaxed span matching, after nested events have been broken down as explained in Section 3.1. Note that the scores do not count the non-named entities, hedges, and links between arguments, since only core events are considered in the official evaluation.

We applied both a deep parser, Enju (Miyao and Tsujii, 2008) and a dependency parser, ksdep (Sagae and Tsujii, 2007) to generate features for the ACE 2005 task, and their bio-adapted versions for the GENIA task. We also employed the GENIA sentence splitter (Sætre et al., 2007) for sentence splitting, and the snowball (Porter2) stemmer[4] for stemming. We did not make use of any other external resources, such as dictionaries, since this would hinder direct comparison of the two versions of the system.

4.2 Evaluation on GENIA

The "Event Detection" column in Table 2 shows evaluation results of BioEE on GENIA. The effects on performance by including entity-related features, i.e., entity base forms and Brown clustering, as introduced in Section 3.2, are shown as "BioEE (+Entity)". The inclusion of these features slightly degrades the performance.

For completeness, we also show in Table 2 the best and second best performing systems that took part in the official BioNLP 2013 ST evaluation: EVEX (Hakala et al., 2013) and TEES-2.1 (Björne and Salakoski, 2013). TEES-2.1 consists of a modular pipeline similar to BioEE, but it uses a different set of features. EVEX enhances the output of TEES-2.1, by using information obtained from the results of large-scale event extraction. The comparison shows that BioEE achieves state-of-the-art event extraction performance on the GENIA task.

4.3 Evaluation on ACE 2005

The "Trigger Classification" and "Arg. Role Classification" columns of Table 3 summarise the evaluation results of the Adapted BioEE system (as described in Section 3.2) on the ACE 2005 task.

We analysed the effects of incorporating features based on entity-related information into the extraction process, by repeating the experiments with such features omitted (-Entity). As can be observed in Table 3, the removal of entity-related features led to 3% performance decrease in F-score.

For completeness, Table 3 also illustrates the results of state-of-the-art systems that were specifically developed for ACE 2005: the system based on a joint approach (Li et al., 2013) and the pipeline-based system enhanced with web-gathered information (Hong et al., 2011). The difference between the

[4] snowball.tartarus.org

Adapted BioEE and the best system is small and insignificant and the Adapted BioEE achieved performance that is comparable to or better than these other systems, in terms of the F-scores in argument role classification.

5 Discussion

To further investigate the differences in performance of the BioEE and Adapted BioEE systems on the two tasks, we evaluate the scores achieved for each task using the evaluation criteria originally designed for the other task. Specifically, we apply the ACE 2005 argument role classification criteria to the output of GENIA task, and we apply the complete event-based evaluation, originally used to evaluate the GENIA task, to the events extracted for the ACE 2005 task. The "Arg. Role Decomposition" column of Table 2 depicts the former evaluation, while the "Event Detection" column of Table 3 shows the latter.

Table 2 also shows the performance of the other biomedical event extraction systems introduced above in carrying out argument role classification, since such information was provided as "Decomposition" within the results of the original task evaluation[5]. Although the results shown for "Arg. Role Decomposition" in Table 2 are not directly comparable to those shown for "Arg. Role Classification" in Table 3 (given the different characteristics of GENIA and ACE 2005 tasks), the scores are broadly comparable. This demonstrates that the task of argument role classifications is equally challenging for both tasks.

The "Event Detection" column of Table 3 illustrates event-based evaluation scores on ACE 2005. The event structure detector was added to the pipeline to facilitate comparison of the results of the two different tasks in a similar setting, and performance was evaluated according to the GENIA evaluation criteria. Evaluation scores on ACE 2005 are unexpectedly low compared to those in Table 2. Considering that the performance of argument role classification is similar in both tasks, this low performance is likely to be due to the large number of potential event arguments in ACE 2005. This means that, in comparison to GENIA events, which have a small number of possible argument types, there is a greater chance that some arguments of more complex ACE 2005 events will fail to be detected. According to the GENIA evaluation criteria, even if the majority of arguments has been correctly identified, the complete event structure will still be evaluated as incorrect. This helps to explain why such evaluation criteria may have been deemed inappropriate in the original ACE 2005 evaluations.

Subsequently, we analysed the effects of utilising entity-related features. We show the results obtained by adding entity information (+Entity) in Table 2 and the results obtained by removing entity information (-Entity) in Table 3. The positive or negative effect on performance of adding or removing these features is consistent across all subtask evaluations shown in the two tables, although the exact level of performance improvement or degradation depends on the subtask under evaluation. Overall, the inclusion of the features degraded the performance of BioEE on the GENIA task, but improved the performance of Adapted BioEE on the ACE 2005 task. These differences may be due to the increased richness of entity information in the ACE 2005 corpus, suggesting that enriching entities in the GENIA corpus with attribute information could be a possible way to further improve the performance of the system on this task.

6 Conclusions and Future Work

In this paper, we have described our adaptation of a biomedical event extraction method to the newswire domain. We firstly evaluated the method on a biomedical event extraction task (GENIA), and showed that its performance was superior to other state-of-the-art systems designed for the task. We then adapted the method to a newswire event extraction task (ACE 2005), by addressing the major differences between the tasks. With only a small number of adaptations, the resulting system was also able to achieve state-of-the-art performance on the newswire extraction task. These results show that there is no need to develop separate systems for event extraction tasks in different domains, as long as the types of tasks being addressed exhibit domain-independent features. However, further discussion and evaluation is needed to better understand how different potential methods for adapting such tools from one domain to another can be used and/or combined effectively.

[5] bionlp-st.dbcls.jp/GE/2013/results

As future work, we intend to further investigate the adaptation of alternative methods proposed for use in one domain to another domain. Several interesting approaches have been described, such as the utilisation of contextual information beyond the boundaries of individual sentences in the newswire domain (Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011) and joint approaches in the biomedical domain (McClosky et al., 2012), but their adaptability to other domains has not yet been investigated. We also intend to investigate the possibility of discovering and utilising shared information between the two domains (Goldwasser and Roth, 2013). Encouraging greater levels of communication between researchers working on NLP tasks in different domains will help to stimulate such new directions of research, both for event extraction and for other related information extraction tasks, such as relation extraction and coreference resolution.

Acknowledgements

This work was supported by the Arts and Humanities Research Council (AHRC) [grant number AH/L00982X/1], the Medical Research Council [grant number MR/L01078X/1], the European Community's Seventh Program (FP7/2007-2013) [grant number 318736 (OSSMETER)], and the JSPS Grant-in-Aid for Young Scientists (B) [grant number 25730129].

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia, July. ACL.

Jari Björne and Tapio Salakoski. 2013. Tees 2.1: Automated annotation scheme learning in the bionlp 2013 shared task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 16–25, Sofia, Bulgaria, August. ACL.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Nancy A. Chinchor. 1998. Overview of MUC-7/MET-2. In *Proceedings of the 7th Message Understanding Conference (MUC-7/MET-2)*.

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Simple information extraction (sie): A portable and effective ie system. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, pages 9–16, Trento, Italy, April. Association for Computational Linguistics.

Dan Goldwasser and Dan Roth. 2013. Leveraging domain-independent information in semantic parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 462–466, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. NYU's english ACE 2005 system description. In *Proceedings of ACE 2005 Evaluation Workshop*, Washington, US.

Kai Hakala, Sofie Van Landeghem, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. Evex in st'13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 26–34, Sofia, Bulgaria, August. ACL.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th ACL-HLT*, pages 1127–1136, Portland, Oregon, USA, June. ACL.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio, June. ACL.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria, August. ACL.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st ACL*, pages 73–82, Sofia, Bulgaria, August. ACL.

Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th ACL*, pages 789–797, Uppsala, Sweden, July. ACL.

Wei Lu and Dan Roth. 2012. Automatic event extraction with structured preference modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 835–844, Jeju Island, Korea, July. Association for Computational Linguistics.

David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, and Christopher Manning. 2012. Combining joint models for biomedical event extraction. *BMC Bioinformatics*, 13(Suppl 11):S9.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80, March.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Longhua Qian and Guodong Zhou. 2012. Tree kernel-based protein–protein interaction extraction from biomedical literature. *Journal of biomedical informatics*, 45(3):535–543.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence*, pages 1044–1049.

Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. 2007. AKANE System: Protein-protein interaction pairs in BioCreAtIvE2 Challenge, PPI-IPS subtask. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 209–212, CNIO, Madrid, Spain, April.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1044–1050, Prague, Czech Republic, June. ACL.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium*.