A Dataset for Open Event Extraction in English

Kiem-Hieu Nguyen[1,*], **Xavier Tannier**[2], **Olivier Ferret**[3], **Romaric Besançon**[3]
1. Hanoi Univ. of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam
2. LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, rue John von Neumann, 91403 Orsay, France
3. CEA, LIST, Vision and Content Engineering Laboratory, F-91191, Gif-sur-Yvette, France

Abstract

This article presents a corpus for development and testing of event schema induction systems in English. Schema induction is the task of learning templates with no supervision from unlabeled texts, and to group together entities corresponding to the same role in a template. Most of the previous work on this subject relies on the MUC-4 corpus. We describe the limits of using this corpus (size, non-representativeness, similarity of roles across templates) and propose a new, partially-annotated corpus in English which remedies some of these shortcomings. We make use of Wikinews to select the data inside the category *Laws & Justice*, and query Google search engine to retrieve different documents on the same events. Only Wikinews documents are manually annotated and can be used for evaluation, while the others can be used for unsupervised learning. We detail the methodology used for building the corpus and evaluate some existing systems on this new data.

1. Introduction

Information Extraction has been defined by the Message Understanding Conference (MUC) evaluations (Grishman and Sundheim, 1996) and its successors, *i.e.* the Automatic Content Extraction (ACE) (Doddington et al., 2004) and Text Analysis Conference (TAC) (Ellis et al., 2014) evaluations, specifically by the task of template filling. The objective of this task is to assign event roles to individual textual mentions. A template defines a specific type of events (*e.g.* earthquakes), associated with semantic roles (or slots) hold by entities (for earthquakes, typically their location, date, magnitude and the damages they caused (Jean-Louis et al., 2011)). This kind of structures is comparable to the schemas of (Schank and Abelson, 1977). *Schema induction* is the task of learning these structures with no supervision from unlabeled texts. We focus here more specifically on *event* schema induction (Chambers and Jurafsky, 2011; Chambers, 2013; Cheung et al., 2013; Nguyen et al., 2015). The idea is to group entities corresponding to the same role into an event template. Figure 1 illustrates this process.

Previous work on event schema induction was evaluated on the MUC-4 corpus (Grishman and Sundheim, 1996). However, this corpus raises two main issues:

- It was annotated with templates describing all events with the same set of slots.
- It doesn't contain redundancy.

The first issue is clearly a limitation due to the fact that all the considered types of events in the MUC-4 corpus are close to each other while the second issue is more a difficulty for applying current machine learning methods.

In this paper, we propose the ASTRE corpus in order to tackle these two issues. We report experimental results on this corpus using state-of-the-art event schema induction methods. The rest of the paper is organized as follows.

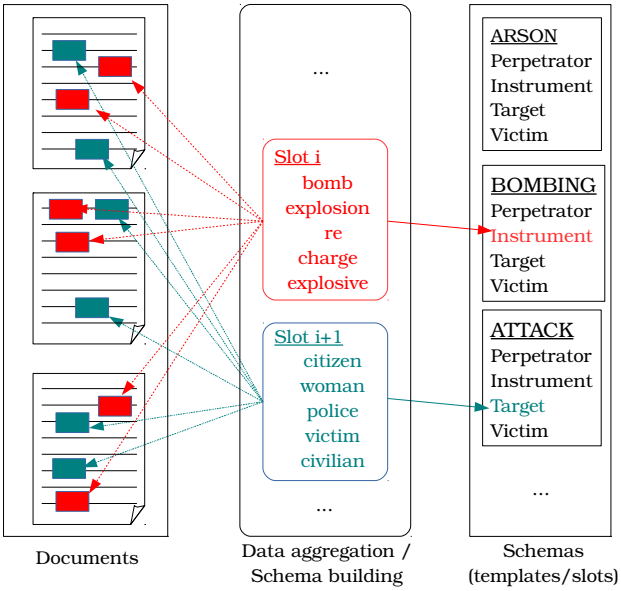

Figure 1: Event induction process (MUC schema example).

Section 2 presents the MUC-4 corpus and its limitations for evaluating schema induction. It also discusses its successors, *i.e.* the ACE and TAC corpora. Section 3 describes the creation of the ASTRE corpus while Section 4 shows the evaluation results of two state-of-the-art systems for open event extraction task on it. Finally, Section 5 concludes the paper.

2. MUC-4 Corpus

A significant part of the work in the field of event schema induction from texts such as (Chambers and Jurafsky, 2011; Chambers, 2013; Cheung et al., 2013; Nguyen et al., 2015) relies on the MUC-4 corpus for its evaluation. This corpus contains 1,700 news articles about terrorist incidents happening in Latin America. The corpus is divided into 1,300 documents for the development set and four test sets, each containing 100 documents.

The evaluation generally focuses on four template types –

* This author was affiliated at LIMSI-CNRS when working on this project.

ARSON, ATTACK, BOMBING, KIDNAPPING – and four slots – Perpetrator, Instrument, Target, and Victim. Perpetrator is merged from Perpetrator_Individual and Perpetrator_Organization. The matching between system answers and references is based on head word matching. A head word is defined as the right-most word of the phrase or as the right-most word of the first 'of' if the phrase contains any. Optional templates and slots are ignored when calculating recall. Template types are ignored in evaluation: this means that a perpetrator of BOMBING in the answers could be compared to a perpetrator of ARSON, ATTACK, BOMBING or KIDNAPPING in the reference.

```
0.  MESSAGE: ID                     TST4-MUC4-0006
1.  MESSAGE: TEMPLATE               1
...
4.  INCIDENT: TYPE                  BOMBING
5.  INCIDENT: STAGE OF EXECUTION    ACCOMPLISHED
6.  INCIDENT: INSTRUMENT ID         "MINE"
7.  INCIDENT: INSTRUMENT TYPE       MINE: "MINE"
8.  PERP: INCIDENT CATEGORY         TERRORIST ACT
9.  PERP: INDIVIDUAL ID             "INSURGENTS"
10. PERP: ORGANIZATION ID           "FMLN"
...
```

Figure 2: MUC-4 annotation example.

The characteristics of the MUC-4 corpus are a limiting factor in the following way:

1. Roles are similar from a template to another. This does not reflect reality and leads to a biased evaluation where only slots are compared and systems do not need to clearly distinguish between templates.

2. The corpus is small and does not contain redundant information. This is due to the initial ambition of the corpus, but raises issues when using it with modern, unsupervised methods. More data and redundant content (*i.e.* several documents relating each event) would open the way to many more different approaches, using for example event clustering and news story aggregation.

Regarding the first issue, the ACE 2005 corpus expanded the set of template types to broader domains, such as LIFE, TRANSACTION and JUSTICE, with specific roles for each of them. Moreover, in ACE, mentions of the same entity in a document are also grouped together. In relation to this last issue, the TAC KBP evaluation includes an *entity linking* task to match different mentions of the same entity across documents through their link to a knowledge base. However, the second issue could not be resolved solely by adding entity linking information. A more in-depth comparison and discussion of the different event annotation schemas can be found in (Aguilar et al., 2014) and (Song et al., 2015).

3. ASTRE Corpus

In order to remedy the shortcomings described in the previous section, we propose a corpus with the following characteristics:

- Redundancy, *i.e.* it contains several documents about the same event.
- Partial annotation in terms of size. The amount of annotated data is sufficient for evaluation purpose while unsupervised training processes for inducing event schemas can benefit from the unannotated data.
- It contains a larger variety of templates.

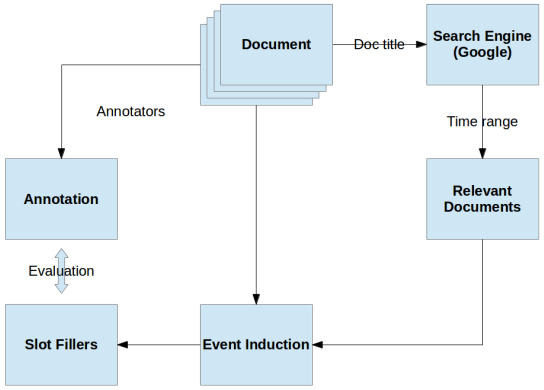

Figure 3: Framework for building and using the ASTRE corpus.

Figure 3 gives an overview of the framework defined for building and using the ASTRE corpus. A collection of documents was first selected from a Wikinews category in English. The Google search engine was then used to retrieve documents from the Web that were similar to these seed documents, with specific time ranges. These retrieved documents were then used for inducing event schemas. At the same time, the Wikinews dataset was manually annotated. The induced schemas were finally evaluated on the annotated dataset.
In this article, we present the corpus, its building procedure as well as the results obtained by two state-of-the-art systems (Chambers, 2013; Nguyen et al., 2015) on this corpus.

3.1. Annotation

3.1.1. Document Acquisition

Wikinews contains news articles from Internet volunteer editors. Its principles are similar to other Wiki sites like Wikipedia and Wiktionary. Each article is composed of a title and a body of several paragraphs, together with rich metadata, among which document creation time (DCT), categorical information, and (external and internal) links to related articles. A Wikinews article is very similar to a newspaper article in its structure and language level.
We chose to annotate articles in the Law & Justice category. First, all the articles in the Law & Justice were collected (except documents in subcategories). Among the 4,000 collected documents, not all of them contain events of the target types. We selected manually 100 documents (*i.e.* 100 different events) containing at least one event of the target types.

3.1.2. Document Annotation

We took a subset of the event types used in TAC KBP 2014 event-argument campaign for annotation, including: LIFE.{Injure, Die}, CONFLICT.Attack, JUSTICE.{Charge-Indict, Arrest-Jail, Release-Parole, Sentence, Convict, Appeal, Acquit, Execute, Extradite}. Moreover, we followed the TAC KBP 2014 guidelines.[1] We also annotated entity coreference chains in documents. However, only the entities appearing at least once as an event argument were annotated with coreference chains.

This annotation scheme is tailored to the type of task we are interested in. Schema induction is a kind of clustering task in which the roles of an induced schema are defined as clusters of entities linked to a type of event. Hence, the annotation of coreference chains is important for determining whether the entity mentions gathered in an induced role actually correspond to the reference role of a type of events. However, contrary to a more classical information extraction task, there is no need for annotating coreference relations between events because we are not interested in extracting the information related to each particular event, which reduces the cost of the annotation task compared to a corpus such as the ACE 2005 corpus.

Figure 4 demonstrates an annotation by our annotators (with the annotation tool Brat (Stenetorp et al., 2012)). Entities (*e.g.* persons and locations) and events (*e.g.* Arrest-Jail and Sentence) are annotated in texts. Directed links from events to entities indicate event-role relationship (as described at the beginning of this Section). In Figure 5, entity coreference chains are traced by the identity field *Note*. For example, 'Ugbogu' mentions in sentences 5 and 7 refer to the same entity ("1-1").[2] 'Masaaki Takahashi' in sentence 5 and 'Takahashi' in sentences 6 and 7 refer the same entity ("1-2").

3.2. Relevant Document Retrieval

In this section, we describe a heuristic technique for retrieving unannotated data from the Web using search engines (*i.e.* Google in this work).

For each annotated document D_Q from Wikinews, we want to select documents from the Web about exactly the same story. For example, if the document is about an attack of a woman on a pet seller using her dead puppy, we want to retrieve several documents relating this exact same story.

The following process was adopted:

- Step 1: the document title, *e.g.* "Woman attacked using her dead puppy", was submitted to the Google search engine as input query. Then, we crawled the documents returned by Google in the descending order of their relevance.
- Step 2: the document creation time of the initial document D_Q (DCT_Q) was extracted. From the results of Step 1, we selected only the documents whose DCTs were in the range $[DCT_Q - \delta, DCT_Q + \delta]$ until we reached K documents for each input query.

The result of this process was a collection of documents about the same story as the query document. In our work we set the time window $\delta = 7$ days and the number of documents per query $K = 20$.

The accuracy of this heuristic is acceptable as long as one do not expect all documents to be relevant to the initial event, but only some of them. The resulting corpus is then a bunch of documents about particular events together with some unrelated documents, which is desirable for making the event extraction task realistic.

3.3. Corpus Building and Statistics

The technique described in Section 3.2 retrieved 1,724 links to relevant articles for our 100 initial annotated documents. 1,347 of them were still accessible. After the application of *boilerpipe* (Kohlschütter et al., 2010) for getting the text content of the retrieved articles and the removal of articles of size less than 1KB, the corpus was made of 1,186 text documents. A deduplication process was then applied as follows: first, the *SpotSigs* tool (Theobald et al., 2008) was used for detecting pairs of possible duplicate documents, *i.e.* pairs of documents whose similarity were equal to 1.0 according to *SpotSigs*' criteria; second, the Markov Clustering algorithm (Dongen, 2000), implemented by the *mcl* tool, was applied for identifying groups of duplicate documents from these pairwise similarities. One document was picked for representing each group, which led to a corpus made of 1,038 documents. A final cleaning of the selected documents was performed for removing boilerplate text not discarded by *boilerpipe*. As this kind of text often consists in recurrent patterns, we sorted all the lines of the corpus and selected the 200 most frequent ones. 61 of them were manually filtered out and the remaining lines were used as reference for cleaning the corpus. Some statistics about this corpus are shown in Table 1.

#docs	#sentences	#words	#tokens
1,038	42.6K	969.5K	1.19M

Table 1: Statistics of the retrieved corpus.

4. Open Event Extraction Evaluation on the ASTRE Corpus

This section describes our first experiments about open event extraction on the ASTRE corpus. Two state-of-the-art systems for this task (as described in (Chambers, 2013) and (Nguyen et al., 2015)) were evaluated on the corpus. Unannotated documents retrieved from the Web were used for model learning. Manually annotated data (as described in Section 3.1.2) were used as development dataset. In addition, we built a separate test dataset. This corpus was defined first by selecting randomly 100 articles from the Wikinews Law & Justice category. Six of them were non-relevant and removed. The remaining 94 articles were manually annotated according to the process described in Section 3.1.2. While they both come from the Wikinews

[1] http://www.nist.gov/tac/2014/KBP/Event/guidelines/TAC_KBP_2014_Event_Argument_Extraction_Assessment_Guidelines_V1.3.pdf

[2] An entity is identified by its first appearance, *e.g.* entity 1-1 means that it is the first coreferable mention in sentence 1.

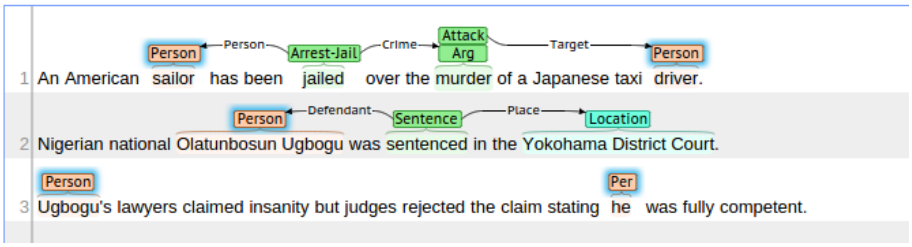

Figure 4: An example of annotations.

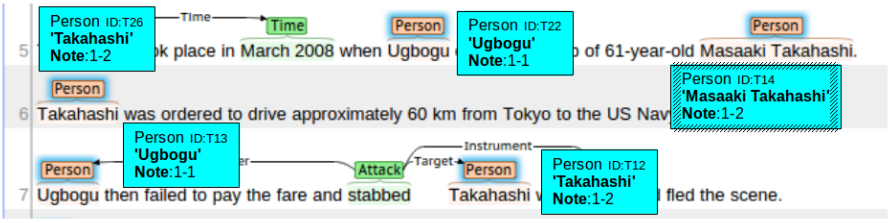

Figure 5: An example of entity coreference annotations.

System	Dev score			Test score		
	P	R	F	P	R	F
(Chambers, 2013)	33	34	34	15	28	19
(Nguyen et al., 2015)	41	30	35	21	26	23

Table 2: Slot filling performance of two state-of-the-art systems on the ASTRE corpus.

Law & Justice category, the *dev* and *test* datasets differs in twofold:

- Articles in the *dev* dataset were manually selected so that all its articles are relevant to the target types of events.
- For taking into account the redundancy issue, the *dev* dataset contains stories similar to stories in the unannotated corpus (as the unannotated corpus was retrieved based on the *dev* dataset). In contrast, the *test* dataset contains totally different stories.

The evaluation of unsupervised slot filling systems includes a slot mapping step to match learnt slots and reference slots (See (Chambers, 2013) and (Nguyen et al., 2015) for more details about the evaluation process). Each reference slot is mapped to the learned slot with the highest F-score for the slot filling task. We used the aforementioned development dataset for this step. In our experiments, we did not use all the slots in the annotated data for evaluation. Slots occurring less than 20 times in the development dataset were omitted.

Table 2 shows the performance of the two systems on manually annotated datasets in terms of Precision (P), Recall (R) and F-score (F). For each one, a model was first learnt from the unannotated corpus described in Section 3.2. We then use the model to evaluate slot filling on two datasets: the development dataset and the test dataset.

The first thing to notice is the significant difference in the results for the test set and the development set, which is a rather classical phenomenon but is particularly strong in the present case. This difference is observed for the two systems but is more important for Chambers (2013). These findings confirm that the task is globally difficult and the fact that the ASTRE corpus is more heterogeneous than the MUC-4 corpus in terms of types of events clearly tends to emphasize this difficulty. This last observation is confirmed by the difference of best F-score values (on respective test datasets) between the two corpora: 19 on the ASTRE corpus compared to 41 on the MUC-4 corpus for Chambers (2013) and 23 compared to 43 for Nguyen et al. (2015)[3]. On the development set, both systems perform rather equally with two different configurations for precision and recall: the results of Chambers (2013) are strictly balanced between precision and recall whereas Nguyen et al. (2015) favor precision over recall. The balance between precision and recall is also different for the two systems on the test set, with a strong imbalance in favor of recall for Chambers (2013) whereas precision and recall are close for Nguyen et al. (2015).

5. Conclusions

We have presented in this paper the ASTRE corpus, a new corpus dedicated to the evaluation of event schema induction. Compared to the MUC-4 corpus, which is classically used for such evaluation but was not developed for it, the ASTRE corpus contains a larger number of types of events with a specific structure for each of them in terms of roles. Moreover, its structure is tailored to the target task, with a large unannotated corpus for inducing event schemas and smaller manually annotated development and test sets built in a controlled way concerning the events they refer to.

[3]The evaluation methods are not strictly similar in both cases but even the comparison with the development dataset of the ASTRE corpus gives far better results for the MUC-4 corpus.

The results of two state-of-the-art event schema induction systems are reported and illustrate the difficulty of the task. As a first extension, we plan to enlarge the ASTRE corpus by considering new types of events (for instance earthquakes) following the same semi-automatic method. More globally, we think that some evolutions of the evaluation methodology could be proposed for taking into account more explicitly that the target task is a clustering task.

6. Acknowledgements

This work has been partially supported by the French National Research Agency (ANR) within the ASRAEL Project, under grant number ANR-15-CE23-0018, by the Foundation for Scientific Cooperation "Campus Paris-Saclay" (FSC) under the project Digiteo ASTRE No. 2013-0774D, and by the French Ministry of Industry under the project REQUEST 018062-25005 FSN-AAP-Big Data n.3. We would like to thank anonymous reviewers for their comments and suggestions.

7. Bibliographical References

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and Ellis, J. (2014). A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Chambers, N. and Jurafsky, D. (2011). Template-Based Information Extraction without the Templates. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*, pages 976–986, Portland, Oregon, USA, June.

Chambers, N. (2013). Event Schema Induction with a Probabilistic Entity-Driven Model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Seattle, Washington, USA, October.

Cheung, K. J. C., Poon, H., and Vanderwende, L. (2013). Probabilistic Frame Induction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 837–846.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *4th Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon, Portugal.

Dongen, S. V. (2000). *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

Ellis, J., Getman, J., and Strassel, S. M. (2014). Overview of linguistic resources for the tac kbp 2014 evaluations: Planning, execution, and results. In *TAC KBP 2014 Workshop*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology.

Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. In *16th International Conference on Computational linguistics (COLING'96)*, pages 466–471, Copenhagen, Denmark.

Jean-Louis, L., Besançon, R., and Ferret, O. (2011). Text Segmentation and Graph-based Method for Template Filling in Information Extraction. In *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 723–731, Chiang Mai, Thailand.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Third ACM international conference on Web search and data mining (WSDM 2010)*, pages 441–450.

Nguyen, K.-H., Tannier, X., Ferret, O., and Besançon, R. (2015). Generative Event Schema Induction with Entity Disambiguation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, Beijing, China, July.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, plans, goals and understanding: an inquiry into human knowledge structures*. The Artificial intelligence series. L. Erlbaum, Hillsdale, N.J.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. (2015). From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado, June. Association for Computational Linguistics.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Theobald, M., Siddharth, J., and Paepcke, A. (2008). Spotsigs: Robust and efficient near duplicate detection in large web collections. In *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, pages 563–570, Singapore, Singapore. ACM.