Unsupervised Event Clustering and Aggregation from Newswire and Web Articles

Swen Ribeiro, Olivier Ferret, Xavier Tannier

Unsupervised Event Clustering and Aggregation from Newswire and Web Articles

Swen Ribeiro
LIMSI, CNRS
Univ. Paris-Sud
Université Paris-Saclay
swen.ribeiro@limsi.fr

Olivier Ferret
CEA, LIST,
Gif-sur-Yvette,
F-91191 France.
olivier.ferret@cea.fr

Xavier Tannier
LIMSI, CNRS
Univ. Paris-Sud
Université Paris-Saclay
xavier.tannier@limsi.fr

Abstract

In this paper, we present an unsupervised pipeline approach for clustering news articles based on identified event instances in their content. We leverage press agency newswire and monolingual word alignment techniques to build meaningful and linguistically varied clusters of articles from the Web in the perspective of a broader event type detection task. We validate our approach on a manually annotated corpus of Web articles.

1 Introduction

In the context of news production, an event is the characterization of a significant enough change in a space-time context to be reported as newsworthy content. This definition fits with definitions proposed in other contexts such as the ACE 2005 and TAC KBP Event evaluations or work such as (Cybulska and Vossen, 2014; Mitamura et al., 2015), which generally view each event as "something that happens at a particular place and time", implying changes in the state of the world and involving participants. In accordance with ontologies about events such as the Simple Event Model (SEM) ontology (van Hage et al., 2011), events can be categorized into different *types*, for example "elections" or "earthquakes", gathering multiple real-life *instances*, for example the "2017 UK General Election" or the "2012 French Presidential Election". These *instances* are reported by journalists through varying textual *mentions*. Event extraction is a challenging task that has received increasing interest in the past years through many formulations such as event identification or event detection. It is also an important subtask of larger NLP applications such as document summarization and event schema induction. Several approaches have been used to tackle the different aspects of this task, particularly in an unsupervised fashion, from linguistic pipelines (Filatova et al., 2006; Huang et al., 2016) to topic modeling approaches (Chambers and Jurafsky, 2011; Cheung et al., 2013) and more recently neural networks (Nguyen et al., 2016). While the definition and granularity of an event varies with the task and objectives at hand, most event identification systems exploit *mentions* to produce *type*-level representations.

We propose to address the unsupervised event extraction task through two subtasks: first, unsupervised event instance extraction and second, event type extraction. This paper will focus on our efforts regarding the first step, *e.g.* unsupervised event instance extraction. In this perspective, we present a method based on clustering algorithms leveraging news data from different sources. We believe that this first step might act as a bridge between the surface forms that are mentions and the more abstract concept of *instances* and *types* of events. Moreover, the context of this work is the ASRAEL project, which aims at providing operational tools for journalists, and this instance/type segmentation seems relevant in the perspective of further event-driven processing developments.

Our clustering approach considers three dimensions: time, space and content. A content alignment system is adapted from Sultan et al. (2014) and a time and space-aware similarity function is proposed in order to aggregate articles about the same event.

We work with a large collection of English news

Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 62–67
Copenhagen, Denmark, September 7, 2017.

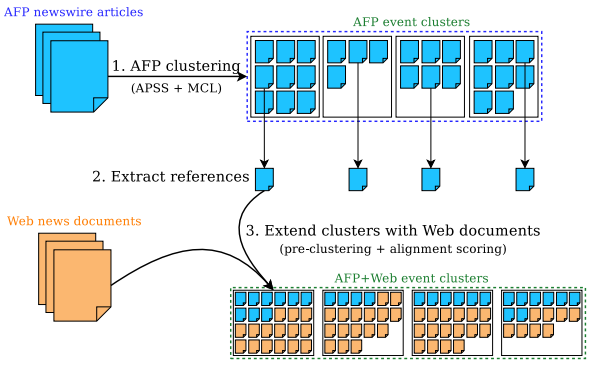

Figure 1: Overview of the system.

and Web articles, where each article describes an event: the main topic of the article is a specific event, and other older events are mentioned in order to put it into perspective. Thus, we consider an event associated with an article.

Our system's objective is to build clusters of articles describing the same exact real-life event, *e.g* the same event *instance*. We adopt two definitions of the relation "same event" (strict and loose) and evaluate through these two definitions.

2 Two-step Clustering

Our approach is structured as a pipeline including a two-step clustering with an additional filtering step at the end. The first step leverages an homogeneous corpus of news articles for building focused and "clean" clusters corresponding to event instances. The second step exploits these focused clusters for clustering documents coming from the Web that are more noisy but also more likely to bring new information about the considered events. Figure 1 illustrates this pipeline.

2.1 Corpora

The first clustering step (represented in blue on Figure 1) is performed on a corpus from Agence France-Presse (AFP) news agency. Each news article comes with several metadata providing additional information about its time-space context of creation, such as its UTC time-stamp, and its content, through International Press Telecommunications Council (IPTC) NewsCodes. NewsCodes are a standard subject taxonomy created and maintained by the IPTC, with a focus on text.

From the 1,400+ existing NewsCodes, we selected 72 that can be viewed as event types[1], covering as many event types as possible without overlapping with one another, and retrieved all news articles tagged with at least one of these NewsCodes. This resulted in a corpus of about 52,000 documents for the year 2015.

The second clustering step (in orange on Figure 1) takes as input news articles crawled from a list of Web news feeds in English. We used a corpus of 1.3 million Web news articles published in 2015, from about 20 different Web news sites (3,700 documents/day in average) including the RSS feeds of the New-York Times, the BBC or the Wall Street Journal.

In both corpora, we process only the title and first paragraph (usually one or two sentences) of the documents, under the assumption that they follow the journalistic rule of the 5Ws: the lead of an article must provide information about *what*, *when*, *where*, *who* and *why*.

2.2 Approach

2.2.1 Press Agency Clustering

The first clustering step computes the similarity matrix of the AFP news by the means of the All Pairs Similarity Search (APSS) algorithm (Bayardo et al., 2007) and applies to it the Markov Clustering (MCL) algorithm (van Dongen, 2000). News are represented by a bag-of-word representation including the lemmatized form of their nouns, adjectives and verbs.

The similarity function between two documents d_1 and d_2 is the following:

$$sim(d_1, d_2) = \frac{cos(d_1, d_2)}{e^{\delta/24}}$$

where $cos(d_1, d_2)$ is the cosine similarity and δ is the difference between the documents creation times (in hours). This time decay ensures that two similar but different events, occurring at different moments, will not be grouped together. Only similarities above a threshold τ have been considered[2].

This first step yields small and instance-focused clusters of press agency news articles only. While they can be considered high quality content, they are quite homogeneous and lack variety in their wording, and could not be used for broader tasks such as event type-level detection. An example of output for this step is provided in Figure 2.

[1] A user-friendly tree visualization of all the NewsCodes is available at http://show.newscodes.org/index.html?newscodes=subj.

[2] A grid search led to $\tau = 0.5$.

Hundreds dead in Nepal quake, avalanche triggered on Everest. A massive 7.8 magnitude earthquake killed hundreds of people Saturday as it ripped through large parts of Nepal, toppling office blocks and towers in Kathmandu and triggering an avalanche that hit Everest base camp.

Nepal quake kills 1,200, sparks deadly Everest avalanche. A massive earthquake killed more than 1,200 people Saturday as it tore through large parts of Nepal, toppling office blocks and towers in Kathmandu and triggering a deadly avalanche at Everest base camp.

Hundreds dead in Nepal quake, deadly avalanche on Everest. A massive 7.8 magnitude earthquake killed more than 900 people Saturday as it ripped through large parts of Nepal, toppling office blocks and towers in Kathmandu and triggering a deadly avalanche that hit Everest base camp.

Figure 2: 3 of 5 AFP news articles clustered together. While they indeed cover the same event instance, there are few wording variations between them, limiting their interest for broader event detection and assimilated tasks.

2.2.2 Web Article Extension

In this step, we aim to alleviate the lack of variability of our AFP news article clusters by leveraging their high focus to aggregate Web documents about the same event instances.

To do so, we identify the first article published in each AFP cluster (using the time-stamp) and retrieve all Web articles in the next 24 hours. This is based on the assumption that press agencies are a primary source of trustworthy information for most news feeds, so it would be rare to find mentions of an event instance before an article was released, especially in an international context. We call this article the "reference".

We first perform a first "coarse-grain" agglomeration by performing low-threshold cosine similarity-based clustering between the AFP reference and all Web articles for the given 24-hour timespan. This results in smaller subsets of data to feed the next module in the pipeline.

We then use the monolingual word alignment system described in Sultan et al. (2014). This system performs a word-to-word alignment between two sentences by applying a series of alignment modules focusing each on a specific type of linguistic units. The alignment process starts with n-grams of words (with n $\geqslant$ 2) including at least one content word. Then, named entities are considered, followed by content words and finally, stopwords. While alignment of n-grams of words and named-entities is based only on string matching (exact match for n-grams, partial for named entities as the system uses Stanford NER to resolve acronyms and matching partial mentions), the system also relies on contextual evidence for other linguistic units, e.g: syntactic dependencies and textual neighborhood. Textual neighborhood is defined as a window of the next and previous 3 content words surrounding each word being considered for an alignment. The system then computes a similarity score between each candidate pair available based on this evidence, and selects the highest scored pair for a given word as the chosen alignment. We adapted the system to better fit our needs by extending the stopword list, first aligning unigram exact matches and using the absence of matching content words or named entities as an early stopping condition of the alignment process.

For each AFP cluster, we perform alignment between the reference (earliest article) and each Web article from the subset. This allows us to build a word alignment matrix where each column contains the words in a document and each line shows how each word of the reference has aligned across all documents.

We then compute a score for each document, taking into account how many words in a document have been aligned with the reference, and how many times a reference word has found an alignment across all documents.

Figure 3 illustrates how this score is computed. We first build the binary alignment matrix B where columns represent documents and rows represent term alignments. If a term i (out of M aligned terms) from document j (out of N documents) has been aligned with a term from the reference, then $B_{i,j} = 1$, otherwise $B_{i,j} = 0$. We then compute a weight for each alignment, leading to a vector $Align$ such as for each term i:

$$Align_i = \sum_{j=0}^{N} B_{i,j}$$

The absolute alignment score of each document j is then:

$$s_j = \sum_{i=0}^{M} W_{i,j}$$

where $W = B \times Align$. Finally, we normalize these by the scores that the reference itself would have obtained.

Once we have scored the documents of a cluster, we sort them and find the greatest gap between

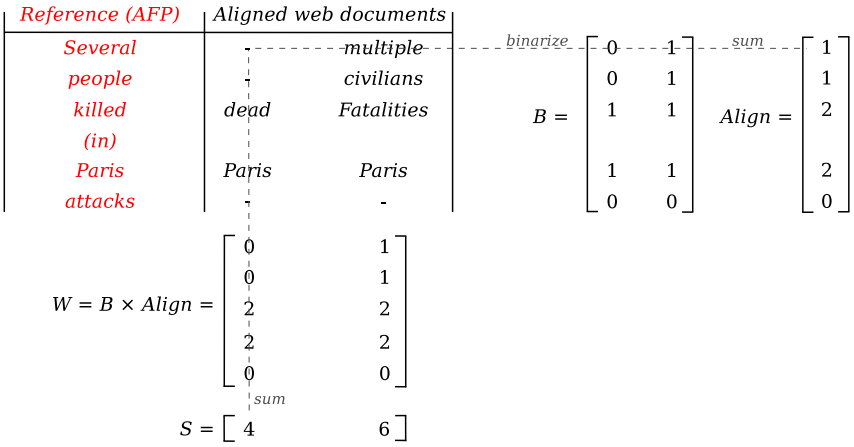

Figure 3: Document scoring.

two consecutive scores (scree test). Only the best-ranked documents before this elbow value are kept as event instance driven document clusters.

3 Evaluation and Results

In our evaluation, we focus on assessing the quality of the clusters produced at the end of the alignment filtering step. We performed our experiments on the AFP and Web data for the whole year 2015. Considering that the AFP corpus sometimes develops more "France-" and "Europe-centric" content while our Web corpus is more "Anglo-Saxon-centered", we need to ensure that we evaluate on event instances that are covered in both corpora, which is the case in the resulting outputs of the coarse-grain agglomeration phase, by construction. We therefore selected 12 of these "pre-clusters" of event instances, based on the notable events of the year 2015 as per Wikipedia[3]. This selection is described in Table 1. The Web articles in these intermediary outputs are sorted by descending order of their cosine similarity to the AFP reference. This ordering will serve as a baseline to evaluate the capacity of the alignment module to produce more relevant clusters, the documents processed at both steps being the same.

We ran AFP clustering and "coarse-grain" agglomeration, identified the resulting intermediary outputs that corresponded to our 12 selected event instances (content and time-stamp wise). We then ran the alignment phase, picked the 50 best-ranked Web articles in each cluster obtained from the selected outputs and tagged them manually with a relevance attribute as follows:

- 0: The document is not related to the reference event considered;
- 1: The document has a *loose* relation to the reference event;
- 2: The document has a *strict* relation to the reference event.

[3] https://en.wikipedia.org/wiki/2015

France seizes passports of would-be jihadists. *February 23rd*	**Protesters clash with police in St Louis, Mo., USA.** *August 20th*
Cyclone Pam hit Vanuatu archipelago. *March 15th*	**Facebook vows to combat racist content on German platform.** *September 14th*
UK General Election campaign start. *March 30th*	**Wildfires rampage across northern California.** *September 14th*
Magnitude 7.9 earthquake hits Nepal. *April 25th*	**Paris Attacks.** *November 13th*
Pakistan police kill head of anti-Shiite group. *July 7th*	**Swedish police arrest man for plotting terror attack.** *November 20th*
ISIS Truck bombing in Baghdad market. *August 13th*	**Typhoon Melor causes heavy flooding in Philippines.** *December 16th*

Table 1: The 12 events of our gold standard.

We define *strict* and *loose* relation as follows: a *strict* relation means that the document is focused on the event and differ from the reference news article only by its wording or additional/missing information; a *loose* relation designates a document that is not focused on the event, but provides a news that is so specific to this event that its mention is core to the overall information provided. Examples of strict and loose relations are provided in Figure 4.

This distinction was introduced when facing two particular types of documents: death toll updates and responsibility claims for terrorist attacks. In both cases, the causal events (attack or natural disaster) are first released as they are in-

Magnitude 7.5 earthquake hits Nepal: USGS. A powerful 7.5 magnitude earthquake struck Nepal on Saturday, the United States Geological Survey said, with strong tremors felt across the Himalayan nation and parts of India.
101 dead as 7.8 quake hits Nepal, causing big damage. A powerful earthquake struck Nepal Saturday, killing at least 71 people as the violently shaking earth, collapsed houses, leveled centuries-old temples and triggered avalanches in the Himalayas.
Nepal quake toll reaches 688: government. KATHMANDU (Reuters) - The death toll from a powerful earthquake that struck Nepal on Saturday has risen to 688, a senior home ministry official told Reuters, with 181 people killed in the capital Kathmandu.

Figure 4: Examples of strict and loose relations. The first text is from the reference news article, the second one is assessed as "strict" relation, the third one as a "loose" relation.

formation of their own. Afterwards, death tolls and claims become stand-alone newsworthy content and are updated independently, yet remaining tightly connected to their causal event.

We use the same metrics as described in Glavaš and Šnajder (2013): mean R-precision (*R-prec.*) and mean average precision (*MAP*) are computed over the complete ordering of all the documents in the cluster with:

$$R\text{-}prec = \frac{r}{R}$$

where r = number of relevant retrieved documents and R = total number of relevant documents to retrieve. Average Precision (AP) is given by:

$$AP = \frac{\sum_{k=1}^{n} (P(k) * rel(k))}{R}$$

where k = rank of the document, $P(k)$ is the precision at cut-off k and $rel(k) = 1$ if document k is relevant, 0 otherwise. We also compute precision, recall and F-score after applying the elbow splitting to evaluate it separately.

Our results are detailed in Table 2 by distinguishing for each reference (strict or loose) the figures with (*align*) and without (*no align*) the use of our final alignment algorithm. From that perspective, Table 2 clearly shows the interest of this last step, with a significant increase of both MAP and R-precision when the final alignment algorithm is applied. This increase is particularly noticeable for R-precision, which emphasizes the ability of this last step to rerank the Web documents in a relevant way. Unsurprisingly, the strict reference is globally more difficult than the loose one, especially for precision: as *loose* documents are close

	Strict		**Loose**	
	no align	**align**	**no align**	**align**
MAP	58.6	62.2	63.7	66.9
R-prec.	50.2	60	56.5	63.5
Precision	–	70.7	–	77.1
Recall	–	80.3	–	76.3
F-score	–	75.2	–	77.7

Table 2: Performance of our event instance clustering system. Average values for the 12 events.

to *strict* documents, the overall system tends to select more false positives with the *strict* reference. Logically, the *loose* reference makes recall decrease, but very slightly.

From a qualitative perspective, we observed several phenomena. Sometimes, the journalistic coverage of an event extends greatly from the time-space context of the mentioned instance, which tends to have a negative impact on precision. For example, in our corpus, the 13 November terrorist attacks of Paris have caused many official reactions worldwide as well as actions taken through social media that have been covered on their own, all in a very short period of time. Moreover, the event itself might be complex in nature: while the event "Paris Attacks" can be restricted to the city of Paris on one particular night (unified time-space context), it is in fact composite, consisting in multiple attacks of different natures (shootings and bombings). For our system, this results in clusters of abnormal sizes (700+ documents clustered in this case, against an usual maximum of 100+). In such cases, the number of annotated documents in the gold standard can be too low, which is an obstacle to the correct evaluation of the output. These abnormal clusters also have another characteristic: being composed of significantly more documents, the distribution of their alignment scores tends to be smoother, making the scree-test less reliable.

4 Conclusion and Perspectives

In this paper, we introduced an unsupervised pipeline aiming at producing event instance driven clusters of news articles. To do so, we leverage homogeneous high-quality news agency articles to identify event instances and find linguistic variations in their expression from Web news articles. Our experimental results validate our approach as a groundwork for future extensions in the broader task of grouping events according to their type and inducing a shared representation of each type of

event by identifying and generalizing the participants of events.

5 Acknowledgment

This work has been partially funded by French National Research Agency (ANR) under project ASRAEL (ANR-15-CE23-0018). We would like to thank the French News Agency (AFP) for providing us with the corpus.

References

Roberto J. Bayardo, Yiming Ma, and Ramakrishnan Srikant. 2007. Scaling up all pairs similarity search. In *16th International World Wide Web Conference (WWW'07)*. pages 131–140.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Portland, Oregon, USA, pages 976–986.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwenden. 2013. Probabilistic frame induction. In *Proceedings of NAACL-HLT 2013*. Atlanta, Georgia, USA, pages 837–846.

Agata Cybulska and Piek Vossen. 2014. Using a Sledgehammer to Crack a Nut? Lexical Diversity and Event Coreference Resolution. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland.

Elena Filatova, Vasileios Hatzivassiloglou, and Kathleen McKeown. 2006. Automatic creation of domain templates. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. pages 207–214.

Goran Glavaš and Jan Šnajder. 2013. Recognizing identical events with graph kernels. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia, Bulgaria, pages 797–803.

Lifu Huang, Taylor Cassidy, Feng Xiaocheng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin, Germany, pages 258–268.

Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event Nugget Annotation: Processes and Issues. In *3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. Denver, Colorado, pages 66–76.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL-HLT 2016*. San Diego, California, USA, pages 300–309.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the Association for Computational Linguistics (TACL)* 2:219–230.

Stijn van Dongen. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web* 9(2):128–136.