11

Exploring Machine Learning Techniques for Linking Event Templates

Jakub Piskorski, Fredi Šarić, Vanni Zavarella, and Martin Atkinson

Abstract. Traditional event detection systems typically extract structured information on events by matching predefined event templates through slot filling. Automatically linking of related event templates extracted from different documents over a longer period of time is of paramount importance for analysts to facilitate situational monitoring and manage the information overload and other long-term data aggregation tasks. This chapter reports on exploring the usability of various machine learning techniques, textual, and metadata features to train classifiers for automatically linking related event templates from online news. In particular, we focus on linking security-related events, including natural and man-made disasters, social and political unrest, military actions and crimes. With the best models trained on moderate-size corpus (ca. 22,000 event pairs) that use solely textual features, one could achieve an F1 score of 93.6%. This figure is further improved to 96.7% by inclusion of event metadata features, mainly thanks to the strong discriminatory power of automatically extracted geographical information related to events.

11.1 Introduction

With the rapid proliferation of large digital archives of textual information on what happens in the world, a need has risen recently to apply effective techniques that go beyond the classification and retrieval of text documents in response to profiled queries. Systems already exist that automatically distill structured information on events from free texts; e.g., with the goal of monitoring disease outbreaks (Yangarber et al., 2008), crisis situations (King and Lowe, 2003), and other security-related events from online news.

Classical event extraction engines typically extract knowledge by locally matching predefined event templates in text documents, by filling template slots with detected entities. However, when not coupled with modules for event

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

coreference detection, these systems tend to suffer of the event duplication problem, consisting of extracting several mentions referring to the same occurring event. That makes their output misleading for both real-time situation monitoring and long-term data aggregation and analysis.

Though event coreference is a semantically well-defined relationship (Mitamura et al., 2015), additional, more fuzzy types of links typically occur between events and potentially contribute to the information overload of the user of an event extraction engine. Capturing such kinds of relationships may be crucial in order to further compress the information and increase the validity of event data.

Imagine a scenario where, given a large set of news reports about a major terrorist attack event, an event extraction engine returns a number of event templates like the ones shown in Figure 11.1. As can be noticed from title and text of the source articles, templates a and b describe the same main fact (the attack itself), c provides updates on some police operations following it, d tells about some public reactions to the event, and e is about an official claiming of the attack by one terrorist organization. Recognizing a and b as duplicate reporting of the same event would help mitigate the information redundancy in the system. At the same time, though c, d, and e should be regarded as semantically distinct events from a, extracting them as independent templates would result in a loss of information. That loss could prevent a user from obtaining a complete picture of the ongoing situation. On the contrary, we envision a user-centered process, where an end-user fed with an initial event template is allowed to explore additional event templates, by calling an on-the-fly computation of related events.

In this context, we explore the possibility of merging a number of distinct event–event relationships (Caselli and Vossen, 2017) into a more general, user-centered definition of event linking and experiment on training statistical classifiers for automatically detecting those links based on textual and non-textual contents of event templates.

The motivation behind our work is fourfold. Firstly, we are interested in elaboration of techniques for linking event information in existing event data sets, such as the one presented in Atkinson et al. (2017), in order to improve their usability by the analysts. Therefore, we have exploited this corpus to carry out the presented work. Secondly, because the event extraction engine underlying the corpus presented in Atkinson et al. (2017) is multilingual, we focus on exploring linguistically lightweight event similarity metrics. Thirdly, we are interested in exploring how inclusion of automatically extracted event metadata (e.g., location) impacts the performance of the trained event linking models. Finally, our aim is to provide a publicly available data set resembling

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

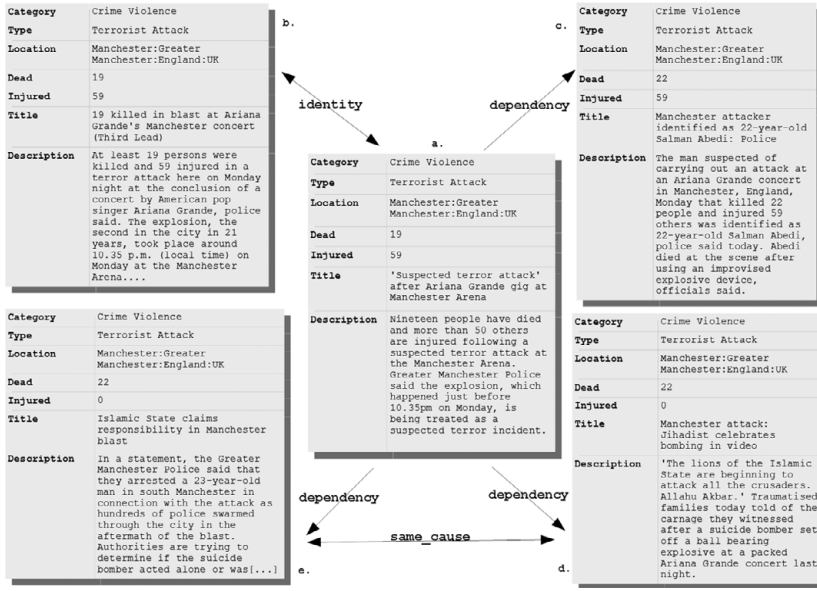

Figure 11.1 Event templates extracted from news reports following the 2017 Manchester terrorist attack, and the different relations linking them to the initial event report in a.

a real-world scenario where end-users are primarily interested in having access to all relevant event information rather than being provided with fine-grained labeling of event relations (e.g., temporal and causal).

Event linking has been modeled as the task of matching monolingual clusters of news articles, describing the same event, across languages. For example, Rupnik et al. (2016) used a number of techniques, including canonical correlation analysis, exploiting comparable corpora such as Wikipedia.

Work similar to ours was performed in the context of the event co-reference resolution task, which consists of clustering of event mentions that refer to the same event (Bejan and Harabagiu, 2010; Cybulska and Vossen, 2014). We diverge from both task formulations in that our underlying representation of events is richer than local event mentions, including metadata and text slots from clusters of articles. Guo et al. (2013) proposed a task of linking tweets with news articles to enable other natural language processing (NLP) tools to better understand Twitter feeds. Related work to event linking was also reported in Nothman et al. (2012), Krause et al. (2016), and Vossen et al. (2016).

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

The article is structured as follows. Section 11.2 gives an overview of the event linking task. The event similarity metrics explored are introduced in Section 11.3. Subsequently, the experimental setup and evaluation results are presented in Section 11.4. Finally, we end with conclusions in Section 11.5.

11.2 Task Description

The event linking task is defined as follows: given an event e and a set of events $E = \{e_1, \ldots, e_n\}$, compute $E^* = \{E^R, E^U\}$ a partition of E into two disjoint subsets of **related** (E^R) and **unrelated** (E^U) events to e. Each event e is associated with an event template $Temp(e)$ consisting of attribute–value pairs describing e, some of which are mandatory – e.g., TYPE, CATEGORY, and LOCATION of the event – while others are optional and event-specific, e.g., PERPETRATOR, WEAPONS_USED. An event template contains three string-valued mandatory slots, namely, TITLE, DESCRIPTION, and SNIPPET, which contain, respectively, the title and the first two sentences from the body of a news article on the event[1] and some text snippet that triggered the extraction of the event.[2]

Figure 11.1 shows a simplified version of a target event template (a) and a number of additional templates (b through e), all belonging to the subset of events related to a.

The semantics of the **related** relationship in our context is defined in a rather broad manner. An event $e' \in E$ is considered to be related to e if the corresponding event templates $Temp(e)$ and $Temp(e')$ refer to (a) the same event (identity); (b) reporting about different aspects of the same ongoing situation/focal event (cooccurrence); (c) two events, where one event occurrence is temporally following and is induced by that of the other event (dependency) with an explicit mention of the prior event; e.g., a trial following a man-made disaster; and (d) two distinct events that were triggered by the same event (same cause).

Due to the application scenario sketched in Section 11.1, the event linking task is modeled here as a classification task applied over a set of events E that does not coincide with the whole search space of events gathered over time but is rather a subset thereof retrieved as some function of the target event e (e.g., events within the same time window as e). This differentiates our approach from clustering methods that attempt to build a partition of the entire event search space based on some relatedness criteria.

[1] The centroid article of the cluster of articles from which the event template was extracted
[2] Please refer to Atkinson et al. (2017) for more details.

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

11.3 Event Similarity Metrics

In our experiments, we explore two types of event similarity metrics. Text-based event similarity metrics compute similarity based on the content of the textual slots in the event templates (e.g., TITLE, DESCRIPTION, etc; see Section 11.2), whereas meta-data-based event similarity metrics compare information contained in the slots that were automatically computed, such as event type and location. The specific metrics of both types are described in Sections 11.3.1 and 11.3.2.

11.3.1 Text-Based Metrics

To determine semantic similarity of text-based event slots, we exploit a wide range of similarity measures, including string similarity metrics, measures that exploit knowledge bases (e.g., WORDNET, BABELNET), and corpus-based similarity metrics, among others. Because our working definition of event linking emphasizes that events should be recognizable even in different languages, we did not explore measures not easily portable across languages; e.g., ones relying on syntactic parsing (Šarić et al., 2012). The remainder of this section briefly introduces the text-based metrics.

First, we use two string distance metrics, namely, **Levenshtein Distance (LT)**, an edit distance metric given by the minimum number of character-level operations needed to transform one string into another (Levenshtein, 1966), and **Longest Common Substrings (LCS)**, which recursively finds and removes the longest common sub-string in the two texts compared (Navarro, 2001). Next, we use **Word Ngram Overlap (WNO)**, which measures the fraction of common word ngrams in both texts, and **Weighted Word Overlap (WWO)**, which measures the overlap of words between the two texts, where words bearing more content are assigned higher weight (Šarić et al., 2012). The formal definitions of LCS, WNO, and WWO are presented in Figure 11.2.

The second pool of event similarity measures exploits various knowledge bases and includes (a) **Named-Entity Overlap (NEO)**, which computes similarity of the named entities found in both texts; (b) **Hypernym Overlap (HO)**, which computes an overlap of the set of hypernyms associated with named entities and concepts found in the texts being compared; and (c) **WordNet Similarity Word Overlap (WSWO)**, a metric that exploits semantic similarity of word pairs computed using WORDNET.[3] The respective formal definitions of NEO, HO and WSWO are provided in Figure 11.3.

[3] We deployed the WS4j library for this purpose: https://github.com/Sciss/ws4j

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

$$LCS(T_1, T_2) = \begin{cases} 0 \text{ if } |lcs(T_1, T_2)| < 3, \\ \frac{|lcs(T_1,T_2)|}{max(|T_1|,|T_2|)} + LCS(T_{1-lcs(T_1,T_2)}, T_{2-lcs(T_1,T_2)}), \end{cases}$$

where $lcs(T_1, T_2)$ denotes the first longest common substring in T_1 and T_2, and T_{i-p} denotes a text obtained by removing from T_i the first occurrence of p in T_i,

$$WNO(T_1, T_2) = \frac{2 \cdot |Ngrams(T_1) \cap Ngrams(T_2)|}{|Ngrams(T_1)| + |Ngrams(T_2)|},$$

where $Ngrams(T_i)$ denotes the set of consecutive ngrams in T_i,

$$WWO(T_1, T_2) = \frac{2 \cdot WoCov(T_1, T_2) \cdot WoCov(T_2, T_1)}{WoCov(T_1, T_2) + WoCov(T_2, T_1)},$$

where $WoCov(T_1, T_2)$ denotes *Weighted Word Coverage* of T_2 in T_1 and is defined as

$$WoCovC(T_1, T_2) = \frac{\sum_{w \in Words(T_1) \cap Words(T_2)} InfoContent(w)}{\sum_{x \in Words(T_2)} InfoContent(x)},$$

where $InfoContent(w) = \ln \sum_{x \in C} frequency(x)/frequency(w)$ (with C and $frequency(x)$ being the set of words in the corpus and the frequency of x in C, respectively[4]) and $Words(T_i)$ denotes the set of words occurring in T_i.

Figure 11.2 Longest Common Substrings (LCS), Word Ngram Overlap (WNO), and Weighted Word Overlap (WW) similarity metrics.

As regards recognizing names in order to compute NEO, a combination of three lexico-semantic resources is used in the respective order on the unconsumed part of the text: (a) JRC Variant Names database (ca. 4 million entries; Ehrmann et al., 2017), (b) a collection of multiword named entities from BabelNet (Navigli and Ponzetto, 2012; ca. 6.8 million entries) that have been semi-automatically derived using the method described in Chesney et al. (2017), and (c) toponyms (only populated places) from the GeoNames[5] gazetteer (ca. 1.4 million entries). Additionally, heuristics are used to join adjacent named entities (NEs). The aforementioned lexical resources cover a wide range of languages and the metric as such can be directly used on texts in other noninflected languages.

To retrieve the hypernyms in the context of computing HO measure, we use version 3.6 of BabelNet (Navigli and Ponzetto, 2012).[6]

[4] The event corpus introduced in Atkinson et al. (2017) was used for this purpose.
[5] www.geonames.org/
[6] We used BabelNet API method which returns all hypernyms for a given synset (depth one).

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

$$NEO(T_1, T_2) = \frac{2 \cdot NeCo(T_1, T_2) \cdot NeCo(T_2, T_1)}{NeCo(T_1, T_2) + NeCo(T_2, T_1)},$$

where *Named-Entity Coverage (NeCo)* of T_1 in T_2 is defined as follows:

$$NeCo(T_1, T_2) = \frac{1}{|N(T_1)|} \cdot \sum_{n \in N(T_1)} max_{m \in N(T_2)} sim(n, m),$$

where $N(T_i)$ is the set of named entities in T_i and $sim(n, m)$ denotes the similarity score of n and m.[7]

$$HO(T, S) = \frac{2 \cdot HypCov(T, S) \cdot HypCov(S, T)}{HypCov(T, S) + HypCov(S, T)},$$

where $HypCov(T, S)$ denotes *Hypernym Coverage* of T in S, defined as

$$HypCov(T, S) = \frac{1}{|T^*|} \cdot \sum_{t \in T^*} \max_{s \in S^*} hypSim(t, s),$$

where T^* and S^* denote the set of potentially overlapping text fragments in T and S, respectively, which can be associated either with a named entity or a concept in a knowledge base. Furthermore, $hypSim(t, s)$ denotes the hypernym similarity between t and s and is computed as follows:

$$hypSim(t, s) = \begin{cases} 1, & t = s \\ x, & \alpha + \beta \cdot \frac{|hyp(t) \cap hyp(s)|}{|hyp(t) \cup hyp(s)|} \\ 0, & hyp(t) \cap hyp(s) = \emptyset \end{cases},$$

where $hyp(s)$ denotes the set of hypernyms for s and α and β have been set to 0.2 and 0.5, respectively, based on empirical observations.

$$WSWO(T, S) = \frac{2 \cdot WnCov(T, S) \cdot WnCov(S, T)}{WnCov(T, S) + WnCov(S, T)},$$

where *WordNet Coverage (WnCov)* of T_1 in T_2 is defined as follows:

$$WnCov(T_1, T_2) = \frac{1}{|Words(T_1)|} \cdot \sum_{w_1 \in Words(T_1)} \max_{w_2 \in Words(T_2)} sim(w_1, w_2),$$

where $sim(w_1, w_2)$ denotes WordNet-based semantic similarity measure between w_1 and w_2.

Figure 11.3 Named Entity Overlap (NEO), Hypernym Overlap (HO), and WordNet Similarity Word Overlap (WSWO) metrics.

Regarding computing WSWO we exploited various WordNet-based semantic similarity measures between pair of words, including, inter alia: **Path**,[8] **WP** presented in Wu and Palmer (1994), **Lesk** (Banerjee and Pedersen,

[7] To compute $sim(n, m)$, we used a weighted version of the LCS metric called *Weighted Longest Common Substrings* introduced in Piskorski et al. (2009).

[8] Counting the length of the path in "is-a" Verb and Noun hierarchy.

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

$$ANO(T_1, T_2) = \frac{2 \cdot |Num(T_1) \cap Num(T_2)|}{|Num(T_1)| + |Num(T_2)|},$$

where $Num(T_i)$ denote the set of numerical expressions found in T_i,

$$RNO(T_1, T_2) = \frac{2 \cdot |NumClos(T_1) \cap NumClos(T_2)|}{|NumClos(T_1)| + |NumClos(T_2)|},$$

where $NumClos(T_1, T_2)$ denotes *Numerical Closeness* between T_1 and T_2 and is defined as follows:

$$NumClos(T_1, T_2) = \frac{1}{|Num(T_1)|} \cdot \sum_{t \in T_1} \max_{s \in T_2} closeness(t, s),$$

where $closeness(t, s)$ is defined as follows:

$$closeness(t, s) = \begin{cases} 1 - \log_2(1 + \frac{|t-s|}{\max(t,s)}), & type(t) = type(s) \\ 0, & type(t) \neq type(s). \end{cases}$$

Figure 11.4 Absolute (ANO) and Relative Numerical Overlap (RNO) similarity metrics.

2002), and **HirstStOnge**, **LeacockChodorow**, **Resnik**, **JiangConrath**, and **Lin** (Budanitsky and Hirst, 2001).

Because an overlap of numerical information contained in texts might constitute an indication of relatedness thereof, we also compute metrics that compute an overlap of the set of numerical expressions found in the texts being compared. Though reported features for computing "similarity" of sets of numerical expressions do not differentiate between the specific types of such expressions (Socher et al., 2011), we do exploit numerical expression–type information. To be more precise, all recognized numerical expressions are classified into one of the following categories: currency (e.g., *200mln$*), percentage, measurement (*one million kilograms*), age (e.g., *20-year-old*), number (e.g., *20 thousand*), whereas numbers being part of temporal references (e.g., *1 May 2017*) are discarded. We computed two numerical overlap measures, namely, **Absolute Numerical Overlap (ANO)** and **Relative Numerical Overlap (RNO)**, whose definitions are presented in Figure 11.4.

Finally, we also use **Cosine of Text Vectors (CTV)**, defined as $CTV(T_1, T_2) = Cosine(Doc2Vec(T_1), Doc2Vec(T_2))$, where $Doc2Vec(T_i) = \frac{1}{|T_i|} \sum_{w \in T_i} embedding(w)$ (Le and Mikolov, 2014) is computed using GloVe (Pennington et al., 2014) word embeddings.

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

Text Preprocessing

Before applying most of the metrics, we deploy preprocessing of the text, which mainly boils down to (a) lowercasing it, (b) normalizing whitespaces, (c) removing constructs such as URLs, etc. For WNO, WSWO, and WWO, some initial/final token characters are stripped (e.g., brackets), and to compute WSWO, WWO, and CVT we remove stop words using a list of ca. 250 English word forms. In the case of NEO and HO the texts are not downcased because this might deteriorate NE recognition performance, which relies on orthographic features. To compute ANO and RNO, no preprocessing is carried out because nonalphanumeric characters often constitute part of numerical expressions.

11.3.2 Metadata-Based Metrics

For metadata information we define four metrics that exploit event location, category, and type information. Because the reported quality of extraction of event-type specific slots (e.g., number of injured, perpetrators, etc.) is not very high, we decided not to exploit such information in the experiments.

Location Administrative Similarity (LAS) computes the administrative distance between locations. It is a modification of the WUP metric presented in Wu and Palmer (1994) and aims to reflect how close two locations are with respect to an administrative hierarchy of geographical references. Let T_{GEO} denote the four-level (Country, Region, Province, and Populated Place) administrative hierarchy in the GeoNames gazetteer,[9] let $lcs(x, y)$ denote the lowest common subsumer for nodes x and y in T_{GEO}, and let $Loc(e)$ denote the node in T_{GEO} that corresponds to the location of the event e. LSA is then defined as follows:

$$LSA(e_1, e_2) = \frac{2 \cdot \omega(lcs(Loc(e_1), Loc(e_2)))}{\omega(Loc(e_1)) + \omega(Loc(e_1))}, \tag{11.1}$$

where $\omega(v) = \sum_{i=0}^{depth(v)} \delta/2^i$ is a weighted depth of a node v in T_{GEO}, with δ empirically set to 10. The intuition behind LSA is to apply a higher weight to path segments closer to the root of T_{GEO}; e.g., distance paths at the Country level are penalized more than paths at the level of Province.

[9] www.geonames.org

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

Location Geographical Similarity (LGS) computes geographical distance between two event locations:

$$LSG(e_1, e_2) = (\ln(dist(coord(e_1), coord(e_2)) + e))^{-1}, \quad (11.2)$$

where $coord(e)$ denotes the coordinates of the location of the event e as found in the GeoNames gazetteer, and $dist(p_1, p_2)$ denotes the physical distance in kilometers between points p_1 and p_2.

Event Category Similarity (ECS) and **Event Type Similarity (ETS)** are two metrics that exploit the event category and type information. Let $cat(e)$ and $type(e)$ denote event category and type, respectively. The metrics are then defined as follows:

$$ECS(e_1, e_2) = Prob(REL(e_1, e_2)|(cat(e_1), cat(e_2)) \quad (11.3)$$

$$ETS(e_1, e_2) = Prob(REL(e_1, e_2)|(type(e_1), type(e_2)), \quad (11.4)$$

where $REL(e_1, e_2)$ denotes events e_1 and e_2 being related. The respective probabilities for category and type pairs were computed using the Gold data set (see Section 11.4.1). In case a certain combination of types (categories) was not observed the respective probability was set to zero, whereas in case of type/category equality the respective probability was set to 1.

11.4 Experiments

11.4.1 Data Set

We built a Gold corpus consisting of event template pairs taken from the event data set described in Atkinson et al. (2017) and labeled as either related or unrelated. First, we attempted to create balanced groups of event templates, where initial groups were built by extracting events (not less than five) around keys consisting of a category, location (country), and a time slot (e.g., time window of ± 2 days) in 2017. All such initial groups G were subsequently amended with a set of max. $|G|/6$ most similar events from the same time window and another set of max. $|G|/6$ most similar events from 2017 but outside of the original time window. The events were selected through computing cosine similarity with the centroid template in G.[10] Finally, G was amended by adding $|G|/3$ of randomly selected events (disjoint from the

[10] Vector representations of event templates and thus centroid templates of groups are derived by computing Doc2Vec on joined Description, Title and Snippet textual slots and converting each word with GloVe word embeddings (Pennington et al., 2014)

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

TITLE: *Militants attack police party in Srinagar*
DESCRIPTION: *Two cops were injured tonight when militants attacked a police party in the Hyderpora area of the city here, police said. Unidentified militants fired upon a night police party near the branch in Hyderpora tonight, resulting in injuries to two policemen, a police official said.*

TITLE: *Civilian gunned down by militants in J-K's Pulwama, 3rd death this week*
DESCRIPTION: *This was the third civilian killed in firing incidents this week. Earlier, one civilian was killed in Srinagar's Rangreth area as security personnel allegedly opened fire to disperse stone-pelters, while another died during an encounter in Arwani village in Bijbehara area.*

Figure 11.5 An example of two events perpetrated by the same group as part of the same armed conflict.

previous groups) from the same time window, regardless of location, category, and similarity.

This original sampling method had some limitations, namely: (a) a significantly unbalanced distribution of the event types, (b) a relatively low number of unique event templates, and (c) a limited range of unique locations of the events. In order to alleviate the aforementioned shortcomings, we repeated the event grouping mechanism described above after selecting groups for a wider range of locations, including event templates that do not have any location information extracted.[11] Moreover, we only annotated a smaller fraction of potential event pair combinations for each event group, to increase the number of unique events. Finally, an additional set of event template pairs with relatively high lexical similarity was sampled and a fraction thereof tagged as unrelated was added to the new corpus in order to increase the fraction of nonobvious event pairs.

All event pairs in each group were computed and subsequently labeled by four annotators, by taking into account only the textual and metadata information available in the templates. The average pairwise κ score for inter-annotator agreement on a sample of around 13,400 event pairs was over 0.85. Questionable cases were typically due to event granularity issues. For example, the two events in Figure 11.5 were arguably perpetrated by the same armed group as part of a same armed conflict on the same day and in the larger area. Whether the two killing incidents should be considered as different consequences of the same larger event, and thus only related, or be considered as distinct events is an open question. We used pairs with at least two non-conflicting judgments to build the GOLD data set.

Detailed statistics of the corpus are provided in Table 11.1.

[11] This resulted in principle due to a failure of event extraction system to detect locations, i.e., the event location was not covered in the underlying linguistic resource or the location information was considered by the system as other type of named entity.

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

Table 11.1. GOLD *data set statistics*

#REL	#UN	#EVT	#CRI	#CIV	#MM	#NAT	#MIL
13051	9587	4214	38.2%	16.2%	16.9%	11.1%	17.6%

The first two columns (REL and UN) provide the number of related (unrelated) event pairs, the third (EVT) provides the total number of unique events, and the others provide the percentage of events falling into crisis-violence (CRI), civic-political action (CIV), man-made disasters (MM), natural disasters (NAT), and military actions (MIL).

11.4.2 Discriminative Power of the Metrics

In order to have a preliminary insight into the discriminative power of the various event similarity metrics, we exploit an objective measure of *absolute Distance* (*absDistance*). Let for some event similarity metric histogram h, $\{u_h\}$, and $\{r_h\}$ denote the sequences of heights of the bars for unrelated and related event pairs, respectively, for all considered bins $i \in I$. *absDistance* is then defined as follows:

$$absDistance(h) = \sum_{i \in I} |u_i^h - r_i^h|/200. \tag{11.5}$$

This metric computes the fraction of the area under the histogram curves being compared that corresponds to the symmetric difference between them, where the area under each histogram has 100 units. The higher values of $absDistance$ indicate better discriminative power of a metric being considered.

We have considered five different modes as regards computation of the features corresponding to the text-based event similarity metrics, namely: (a) only text from the event DESCRIPTION and SNIPPET slots (see Section 11.2) from the event template is used (D); (b) only event TITLE slot is used (T); (c) in addition to (a), the TITLE slot is exploited as well (D+T); (d) similarity score for the TITLE, DESCRIPTION/SNIPPET slot is computed separately and an average thereof is returned (AVG(D,T)); and (e) similarity score for the TITLE, and DESCRIPTION/SNIPPET slot is computed separately and the maximum thereof is returned (MAX(D,T)).

Figure 11.6 provides a comparison of the discriminative power computed using $absDistance$ on GOLD data set for all event similarity metrics and four aforementioned modes in which text-based metrics are calculated. One can observe high potential of some of the metadata metrics, namely, LSG (close to 90% of the area under the curve [AUC]) and ETS (close to 40% of the AUC), whereas NEO and WWO (both of which can be computed efficiently) lead

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

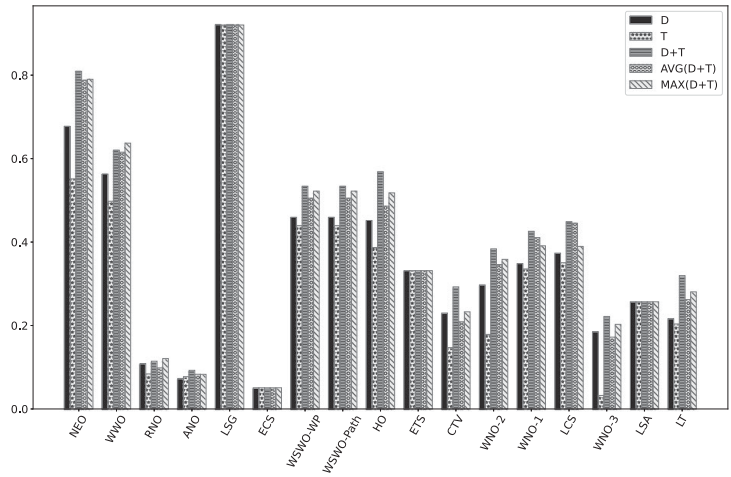

Figure 11.6 Discriminative power of the event similarity metrics. The different versions of WNO depending on the n-gram type are denoted with WNO-i.

the ranking of text-based metrics, followed by metrics exploiting WordNet, BabelNet, which also have relatively high discriminatory power (in the range of 45% – 80% of the AUC). In particular, HO discriminative power is very similar to the WordNet-based distance metrics, which is due to the fact that BabelNet encompasses WordNet resources. Interestingly, the surface-level LCS metric exhibits much higher discriminative power vis-a-vis CTV. Numerical overlap features seems to be least attractive in this comparison, most likely due to the fact that a large fraction of event template pairs tagged as related do not refer to same events but rather to different events linked through the same cause or being in some other type of dependency and thus more likely reporting on different numerical values. Nevertheless, we hypothesize that exploitation of numerical overlap metrics might come in handy in case of natural and man-made disaster events, which unfortunately constitute only a small fraction of all events in our corpora.

11.4.3 Experimental Setup

Experiments were carried out using five different machine learning (ML) models, namely: support vector machine (SVM), stochastic gradient descent classifier (regularized linear model learned), decision tree, random forest, and AdaBoost classifier. All models were implemented using scikit learn

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

library (Pedregosa et al., 2011). Each model was trained using a full set of event similarity metrics as features[12] and on a subset of features obtained using feature selection with *SelectFromModel* base estimator being random forest. All models consistently exhibited better performance when using all features vis-a-vis a subset of features obtained through feature selection. All models were trained and evaluated on the same training–test split (80:20) in order to ensure the comparability of the models. In order to tune the hyper-parameters of the models, we performed grid search over hyperparameter space for each model and evaluated it using fivefold cross-validation using only the training data. The performance metrics of best performing models were further evaluated on held-out test set and used for model comparison.

Noteworthy, in case of missing features – i.e., whenever an event metric could not be computed (e.g., due to missing elements such as named entities or numerical expressions to be compared) – we set the respective values to the mean in the corresponding feature distribution assuming that lack of elements to compare should be scored higher than zero overlap (e.g., different named entities in both texts).

Finally, we carried out the evaluation in two setups, one with text-based features only and a second one with both textual and metadata features.

11.4.4 Results

The performance of the models on the GOLD data set is shown in Table 11.2. The observed results indicate that the task is well modeled by the different classification paradigms, with random forest being in general the top scoring model across all settings. We trained additional models using the random forest paradigm using subsets of the text-based features set by excluding in each run a single feature in order to explore how the exclusion of each feature impacts the performance. The resulting significance order of the features matched to a larger extent the discriminative power ranking depicted in Figure 11.6; i.e., NEO and WWO topped the rankings and ANO and RNO ranked lowest.

As expected (see Section 11.4.2), adding metadata features (in particular, given the discriminatory power of LSG) on top of the text-based features significantly boosts the performance, raising the upper bound from 93.6% to as much as 96.7%. Nevertheless, this is a remarkable finding considering that the metadata features (i.e., the slots LOCATION, TYPE, and CATEGORY)

[12] For the WSWO metric family we finally considered only WSWO-Path and WSWO-WP variants based on some empirical observations that revealed that the other variant returns very similar scores.

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

Table 11.2. *Performance on the* GOLD *data set (F1 scores)*

	Text and Metadata features				
ML Paradigm	D	T	D + T	AVG(D,T)	MAX(D,T)
SVM	95.29%	94.76%	96.02%	95.87%	95.55%
SDG	95.25%	94.73%	95.85%	96.01%	94.73%
RANDOM FOREST	**96.34%**	**96.24%**	**96.62%**	**96.70%**	**96.48%**
DECISION TREE	95.12%	94.28%	95.34%	95.82%	95.93%
ADABOOST	95.21%	94.92%	95.47%	95.64%	95.17%
	Text-based features				
ML Paradigm	D	T	D + T	AVG(D,T)	MAX(D,T)
SVM	84.55%	75.20%	92.29%	89.27%	90.23%
SDG	84.44%	76.61%	91.77%	89.76%	90.31%
RANDOM FOREST	**87.54%**	**82.16%**	**93.60%**	**91.59%**	**91.53%**
DECISION TREE	83.66%	77.50%	91.94%	90.25%	90.59%
ADABOOST	85.02%	78.38%	92.49%	90.69%	90.80%

are automatically generated by an event extraction engine and their extraction is more error prone vis-a-vis the computation of similarity metrics on the textual slots. One needs to emphasize in this context that the surprisingly high discriminative power of the LSG metric that contributed to the overall performance might have been due to the way in which the evaluation corpora were built (see Section 11.4.1); i.e., the discriminative power of the LSG metric might be lower when applied on an event corpus with a higher number of distinct (unrelated) events of the same type that happened in the same location.

Moreover, D + T mode seems to be the best choice overall as regards the various modes for computing text-based features and is statistically different with $p < 0.05$ compared to other modes on the GOLD data set with only textual features. Exploiting only title information (T mode) when using text-based features resulted in a respectable F1 score of 82.2%.

A rudimentary error analysis on the output of the GOLD data set–trained random forest classifier with metadata features and D+T option revealed that most of the false negatives consisted of event pairs referring to different, related aspects of the same target event, like in the article titles in Figure 11.7 (top). This was expected because the text pairs have little lexical overlapping and (more) background knowledge (e.g., access to full news articles) is required in

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

TITLE: *"Concert bomber targeted children"*
DESCRIPTION: *British Prime Minister Theresa May said police know the identity of the bomber, who died in the blast late Monday, and believed he acted alone. [. . .]*

TITLE: *Miley Cyrus "more cautious" after terror attack at Ariana Grande's gig*
DESCRIPTION: *Miley Cyrus says the terror attack at Ariana Grande's concert has made her "more cautious". A bomb was detonated after Ariana's gig. [. . .]*

TITLE: *Ex-Qaeda affiliate leaders among 25 dead in Syria strike*
DESCRIPTION: *An air strike in Syria on Tuesday killed at least 25 members of former Al-Qaeda affiliate Fateh al-Sham Front including senior figures, a monitor said. Unidentified aircraft hit one of the group's most important bases in Syria, in the northwestern province of Idlib. [. . .]*

TITLE: *Syrian air strikes kill at least six civilians*
DESCRIPTION: *ALEPPO - Syrian government air strikes killed at least six civilians, including four children, in Aleppo province on Thursday, despite a fragile two-week-old truce, a monitor said. In neighbouring Idlib province, at least 22 jihadists were killed in air strikes over the past 24 hours. [. . .]*

Figure 11.7 A sample false negative (top) and false positive (bottom) event pair.

order to draw a relatedness link. On the other hand, the models struggled to set apart individual incidents (see Figure 11.7, bottom) belonging to a larger event context, which typically share lexical profile, LOCATION, and TYPE slots. Among all false classifications 60% were false negatives and 40% were false positives.

11.5 Conclusions

This chapter reported on experiments of testing ML methods using a wide range of textual and metadata features to train classifiers for linking related event templates that have been automatically extracted from online news. Though exploiting solely textual features resulted in a 93.6% F1 score evaluated on a 22,000 event-pair corpus, adding metadata features allowed to improve it up to 96.7% on the same corpora, mainly thanks to exploitation of an event similarity metric that computes geographical distance between events with high discriminatory power.

Future research envisaged encompasses (a) adaptation and evaluation of the approach on event data in other languages; (b) consideration of additional lightweight features (e.g., exploitation of country/region size assuming that events occurring in larger countries are less likely to be related, utilization of the structure of the URLs to the related sources that might hint at reporting over time on some bigger events/stories over certain period of time); (c) based on the work carried out, elaboration of additional event similarity metrics to train models for cross-lingual event linking (Rupnik et al., 2016; Al-Badrashiny

https://doi.org/10.1017/9781108854221.014 Published online by Cambridge University Press

et al., 2017); and (d) introducing an additional subclassification of the related class. As a matter of fact, we carried out an initial attempt to subclassify a sample of 150 event pairs (e_1, e_2) labeled as related into one of the four subclasses: IDENTITY (reporting on the same event), SAME_CAUSE (e_1 and e_2 were triggered by the same event; e.g., arrests/investigations and visit of a political leader, both following a terrorist attack), e_1 UPDATES_OR_DEPENDS_ON e_2 and the symmetric case (terrorist attack followed by an introduction of an emergency situation). However, the bilateral κ scores between the three annotators involved ranged from 0.45 to 0.63, which indicates the complexity of the task.

The event linking data set GOLD used for carrying out the reported study, which includes both labeled event pairs and the full set of the event templates from which the event pairs were selected, is available at: http://piskorski.waw.pl/resources/eventLinking/GoldNew.zip.

References

Al-Badrashiny, Mohamed, Bolton, Jason, Chaganty, Arun Tejasvi, et al. (eds.). 2017. *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13–14, 2017*. NIST.

Atkinson, Martin, Piskorski, Jakub, Tanev, Hristo, and Zavarella, Vanni. 2017. On the Creation of a Security-Related Event Corpus. Pages 59–65 of: Caselli, Tommaso, Miller, Ben, van Erp, Marieke, et al. (eds.), *Proceedings of the Events and Stories in the News Workshop 2017*. Vancouver: Association for Computational Linguistics.

Banerjee, Satanjeev, and Pedersen, Ted. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. Pages 136–145 of: Gelbukh, Alexander F. (ed.), *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*. CICLing '02. Berlin: Springer.

Bejan, Cosmin Adrian, and Harabagiu, Sanda. 2010. Unsupervised Event Coreference Resolution with Rich Linguistic Features. Pages 1412-1422 of: Hajič, Jan, Carberry, Sandra, Clark, Stephen, and Nivre, Joakim (eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Budanitsky, Alexander, and Hirst, Graeme. 2001. Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures. In: *Proceedings of the Workshop on WORDNET and Other Lexical Resources, NAACL 2001*.

Caselli, Tommaso, and Vossen, Piek. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. Pages 77–86 of: Caselli, Tommaso, Miller, Ben, van Erp, Marieke, et al. (eds.), *Proceedings of the Events and Stories in the News Workshop*. Vancouver: Association for Computational Linguistics.

Chesney, Sophie, Jacquet, Guillaume, Steinberger, Ralf, and Piskorski, Jakub. 2017. Multi-word Entity Classification in a Highly Multilingual Environment. Pages 11–20 of: Markantonatou, Stella, Ramisch, Carlos, Savary, Agata, and Vincze, Veronika (eds.), *MWE@EACL*. Association for Computational Linguistics.

Cybulska, A., and Vossen, P. 2014. Using a Sledgehammer to Crack a Nut? Lexical Diversity and Event Coreference Resolution. In: *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*.

Ehrmann, Maud, Jacquet, Guillaume, and Steinberger, Ralf. 2017. JRC-Names: Multilingual entity name variants and titles as Linked Data. *Semantic Web*, **8**(2), 283–295.

Guo, Weiwei, Li, Hao, Ji, Heng, and Diab, Mona. 2013. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. Pages 239–249 of: Schuetze, Hinrich, Fung, Pascale, and Poesio, Massimo (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Volume 1. Long Papers*. Sofia, Bulgaria: Association for Computational Linguistics.

King, Gary, and Lowe, Will. 2003. An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders. *International Organization*, **57**, 617–642.

Krause, Sebastian, Xu, Feiyu, Uszkoreit, Hans, and Weissenborn, Dirk. 2016. Event Linking with Sentential Features from Convolutional Neural Networks. Pages 239–249 of: Riezler, Stefan, and Goldberg, Yoav (eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin: Association for Computational Linguistics.

Le, Quoc, and Mikolov, Tomas. 2014. Distributed Representations of Sentences and Documents. Pages II–1188–II–1196 of: Xing, Eric P., and Jebara, Tony (eds.), *Proceedings of the 31st International Conference on International Conference on Machine Learning. Volume 32*. ICML'14. JMLR.org.

Levenshtein, V.I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, **10**, 707.

Mitamura, Teruko, Liu, Zhengzhong, and Hovy, Eduard H. 2015. Overview of TAC KBP 2015 Event Nugget Track. In: *TAC 2015*.

Mohamed Al-Badrashiny, Jason Bolton, Arun Tejasvi Chaganty Kevin Clark Craig Harman Lifu Huang Matthew Lamm Jinhao Lei Di Lu Xiaoman Pan Ashwin Paranjape Ellie Pavlick Haoruo Peng Peng Qi Pushpendre Rastogi Abigail See Kai Sun Max Thomas Chen-Tse Tsai Hao Wu Boliang Zhang Chris Callison-Burch Claire Cardie Heng Ji Christopher D. Manning Smaranda Muresan Owen Rambow Dan Roth Mark Sammons, and Durme, Benjamin Van (eds). 2017. *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*. NIST.

Navarro, Gonzalo. 2001. A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, **33**(1), 31–88.

Navigli, Roberto, and Ponzetto, Simone Paolo. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, **193**, 217–250.

Nothman, Joel, Honnibal, Matthew, Hachey, Ben, and Curran, James R. 2012. Event Linking: Grounding Event Reference in a News Archive. Pages 228–232 of: Li,

Haizhou, Lin, Chin-Yew, Osborne, Miles, Lee, Gary Geunbae, and Park, Jong C. (eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Volume 2. Short Papers*. Jeju Island, Korea: Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. 2014. Glove: Global Vectors for Word Representation. Pages 1532–1543 of: Moschitti, Alessandro, Pang, Bo, and Daelemans, Walter (eds.), *Proceedings of EMNLP 2014*, Vol. 14.

Piskorski, Jakub, Wieloch, Karol, and Sydow, Marcin. 2009. On Knowledge-Poor Methods for Person Name Matching and Lemmatization for Highly Inflectional Languages. *Information Retrieval*, **12**(3), 275–299.

Rupnik, Jan, Muhič, Andrej, Leban, Gregor, et al. 2016. News Across Languages – Cross-lingual Document Similarity and Event Tracking. *Journal of Artificial Intelligence Research*, **55**(1), 283–316.

Šarić, Frane, Glavaš, Goran, Karan, Mladen, Šnajder, Jan, and Bašić, Bojana Dalbelo. 2012. TakeLab: Systems for Measuring Semantic Text Similarity. Pages 441–448 of: Agirre, Eneko, Bos, Johan, Diab, Mona, et al. (eds.), *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics: Volume 1. Proceedings of the Main Conference and the Shared Task, and Volume 2. Proceedings of the Sixth International Workshop on Semantic Evaluation*. SemEval '12. Stroudsburg, PA: Association for Computational Linguistics.

Socher, Richard, Huang, Eric H., Pennington, Jeffrey, Ng, Andrew Y., and Manning, Christopher D. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. Pages 801–809 of: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. Red Hook, NY: Curran Associates.

Vossen, Piek, Agerri, Rodrigo, Aldabe, Itziar, et al. 2016. NewsReader: Using knowledge resources in a Cross-lingual Reading Machine to Generate More Knowledge from Massive Streams of News. *Knowledge-Based Systems*, **110**, 60–85.

Wu, Zhibiao, and Palmer, Martha. 1994. Verbs Semantics and Lexical Selection. Pages 133–138 of: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. ACL '94. Stroudsburg, PA: Association for Computational Linguistics.

Yangarber, Roman, Von Etter, Peter, and Steinberger, Ralf. 2008. Content Collection and Analysis in the Domain of Epidemiology. In: *Proceedings of DrMED 2008: International Workshop on Describing Medical Web Resources at MIE 2008: the 21st International Congress of the European Federation for Medical Informatics 2008*.