On Training Classifiers for Linking Event Templates

Jakub Piskorski[1], Fredi Šarić[2], Vanni Zavarella[1], Martin Atkinson[1]
[1]Text and Data Mining Unit, Joint Research Centre of the European, Ispra, Italy
{firstname.lastname}@ec.europa.eu
[2]Text Analysis and Knowledge Engineering Lab, University of Zagreb, Croatia
fredi.saric@fer.hr

Abstract

The paper reports on exploring various machine learning techniques and a range of textual and meta-data features to train classifiers for linking related event templates automatically extracted from online news. With the best model using textual features only we achieved 94.7% (92.9%) F1 score on GOLD (SILVER) dataset. These figures were further improved to 98.6% (GOLD) and 97% (SILVER) F1 score by adding meta-data features, mainly thanks to the strong discriminatory power of automatically extracted geographical information related to events.

1 Introduction

With the rapid proliferation of large digital archives of textual information on what happens in the world, a need has raised recently to apply effective techniques that go beyond the classification and retrieval of text documents in response to profiled queries. Systems already exists 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 co-reference 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.

While event co-reference is a semantically well-defined relationship (Mitamura et al., 2015), capturing some additional kinds of relationships, although more fuzzy, that link together events, may be crucial in order to reduce the information overload of the user of an event extraction engine.

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 1. As it can be noticed from Title and Text of the source articles, while 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, while 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 mitigating the information redundancy in the system. At the same time, while 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 preventing a data user to obtain a complete picture of the ongoing situation. On the contrary, we envision an user-centered process, where an analyst is fed with a target event template and is allowed to explore on demand additional event templates, by calling an on-the-fly computation of related events in order to update the information from the original record.

In this context, we explore the possibility to merge 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 content of event templates.

Proceedings of the Workshop on Events and Stories in the News, pages 68–78
Santa Fe, New Mexico, USA, August 20, 2018.

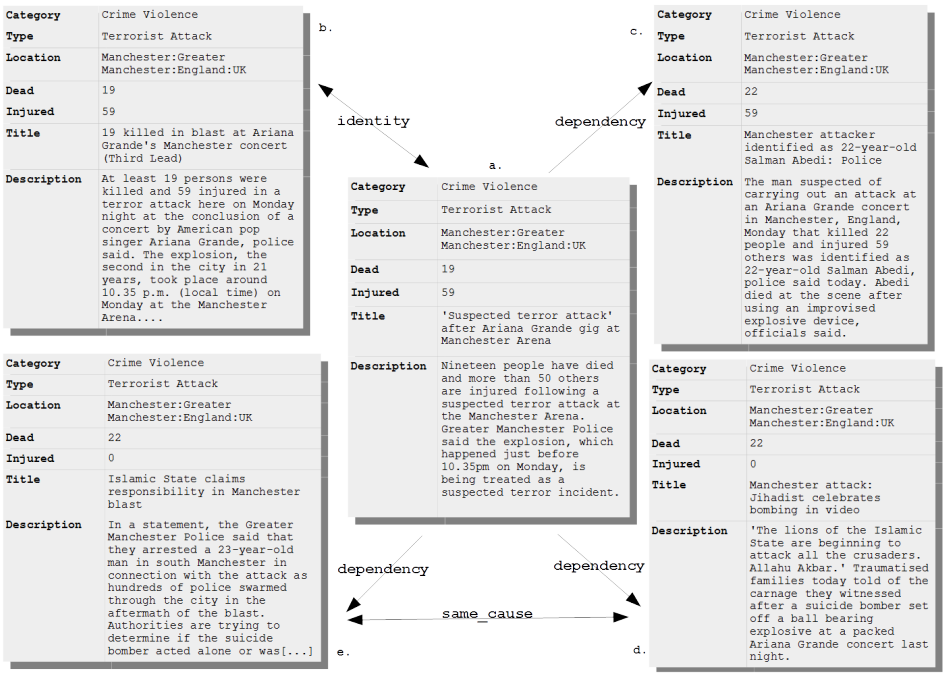

Figure 1: Event templates extracted from news reports following the 2017 Manchester terrorist attack, and the different relations linking them to the initial event report in a.

The motivation behind our work is four fold. Firstly, we are interested in elaboration of techniques for linking event information in existing event datasets, 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, as the event extraction engine underlying the (Atkinson et al., 2017) corpus is multilingual, we focus on exploring linguistically-lightweight event similarity metrics. Thirdly, we are interested in exploring how inclusion of automatically extracted event meta data (e.g., location) impacts the performance of the trained event linking models. Finally, due to scarcity of publicly available resources for carrying out research on event linking our intention was to contribute to the provision of such a resource, focusing particularly on creating a dataset that resembles a real-world scenario of event data for analysts, who 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., 2017) use 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, that consists of clustering of event mentions that refer to the same event (Bejan et al., 2010). We diverge from both task formulations in that our underlying representation of events is richer than local event mentions, including meta-data and text slots from clusters of articles. (Weiwei Guo et al., 2013) proposes a task of linking tweets with news articles to enable other NLP tools to better understand Twitter feeds. Related work to event linking was also reported in (Nothman et al., 2012; Krause et al., 2016; Vossen et al., 2016).

The paper is structured as follows. Section 2 gives an overview of the event linking task. The event similarity metrics explored are introduced in Section 3. Subsequently, the experiments set-up and evaluation results are presented in Section 4. Finally, we end up with conclusions in Section 5.

2 Task description

The Event Linking task is defined as follows: given an event e and a set of events $E = \{e_1, ..., e_n\}$ 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, and optional event-specific ones, e.g., PERPETRATOR, WEAPONS_USED. An event template contains three string-valued mandatory slots, namely, TITLE, DESCRIPTION and SNIPPET which contain in the resp. order: 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. Please refer to (Atkinson et al., 2017) for more details.

Figure 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 (co-occurrence), (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 1, the event linking task is modeled here as a classification task applied over a set of events E which 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.

3 Event Similarity Metrics

3.1 Text-based Metrics

For determining semantic similarity of text-based event slots we exploit a wide range of similarity measures, including, i.a., string similarity metrics, measures that exploit knowledge bases (e.g., WORDNET, BABELNET), and corpus-based similarity metrics. 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 introduces each of the measures used. Let T_1 and T_2 denote two texts being compared.

Levenshtein Distance (LT) is well-known edit distance metric given by the minimum number of character-level operations needed to transform one text into another (Levenshtein, 1965).

Longest Common Substrings (LCS) is a similarity distance metric, which recursively finds and removes the longest common sub-string in the two texts compared (Navarro, 2001). Let $lcs(T_1, T_2)$ denote the first longest common sub-string in T_1 and T_2 and let T_{i-p} denote a text obtained by removing from T_i the first occurrence of p in T_i. The LCS metric is then calculated as follows.

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

Word Ngram Overlap (WNO) is a fraction of common word ngrams in both texts and is defined as:

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

where $Ngrams(T_i)$ denotes the set of consecutive ngrams in T_i. In particular, we computed ngram overlap for unigrams **WNO-1**, bigrams **WNO-2** and trigrams **WNO-3**.

Weighted Word Overlap (WWO) measures the overlap of words between the two texts, where words bearing more content are assigned higher weight (Šarić et al., 2012) using the following formula.

$$InfoContent(w) = \ln \frac{\sum_{x \in C} frequency(x)}{frequency(w)} \quad (3)$$

where C and $frequency(x)$ denote the set of words in the corpus and the frequency of x in C resp. The word frequencies were computed using the entire event corpus introduced in (Atkinson et al., 2017). We then define *Weighted Word Coverage (WWC)* of T_2 in T_1 as follows.

[1]The centroid article of the cluster of articles from which the event template was extracted

$$WWC(T_1, T_2) = \frac{\sum_{w \in Words(T_1) \cap Words(T_2)} InfoContent(w)}{\sum_{x \in Words(T_2)} InfoContent(x)} \quad (4)$$

where $Words(T_i)$ denotes the set of words occurring in T_i. The $WeightedWordOverlap(T_1, T_2)$ is then computed as the harmonic mean of $WWC(T_1, T_2)$ and $WWC(T_2, T_1)$.

Named-Entity Overlap (NEO) is a metric that computes similarity of the named entities found in both texts. Let us first define *Named-Entity Coverage (NEC)* of T_1 in T_2 as follows.

$$NEC(T_1, T_2) = \frac{1}{|Names(T_1)|} \cdot \sum_{n \in Names(T_1)} max_{m \in Names(T_2)} sim(n, m) \quad (5)$$

where $Names(T_i)$ denotes the set of named entities found in T_i and $sim(n, m)$ denotes a similarity score of n and m. The $NamedEntityOverlap(T_1, T_2)$ is then defined as harmonic mean of $NEC(T_1, T_2)$ and $NEC(T_2, T_1)$. In order to compute $sim(n, m)$ we used a weighted version of the LCS metric called *Weighted Longest Common Substrings*[2] introduced in (Piskorski et al., 2009).

For recognising names a combination of 3 lexico-semantic resources have been used in the respective order on the unconsumed part of the text: (a) JRC Variant Names database (ca. 4 mln entries) (Ehrmann et al., 2017), (b) a collection of multi-word named entities from BABELNET (Navigli et al., 2012) (ca. 6.8 mln 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[3] gazetteer (ca. 1.4 mln entries). Additionally, heuristics are used to join adjacent NEs. The aforementioned lexical resources cover a wide range of languages and the metric as such can be directly used on texts in other non-inflected languages.

Hypernym Overlap (HO) is an overlap of the set of hypernyms associated with named entities and concepts found in the texts being compared. Let $T = t_1 \dots t_n$ and $S = s_1 \dots s_n$ denote two texts, where $t_i(s_i)$ denote tokens. Further, let T^* and S^* denote the set of potentially overlapping text fragments (i.e., sequences of tokens) in T and S resp. which can be associated either with a named entity or a concept encoded in a knowledge base. The aforementioned text fragments are computed by identifying at each position in a text the longest sequence of tokens that can be associated with a name or concept. In particular, for computing the sets T^* and S^* we exploit version 3.6 of BABELNET (Navigli et al., 2012)[4]. The hypernym coverage of T in S is then defined as follows:

$$HypCoverage(T, S) = \frac{1}{|T^*|} \cdot \sum_{t \in T^*} \max_{s \in S^*} hypSim(t, s) \quad (6)$$

where $hypSim(t, s)$ denotes the hypernym similarity between t and s and is computed as follows:

$$hypSim(t, s) = \begin{cases} 1, & t = s \\ x, & \alpha + \beta \cdot \frac{|hyp(t) \cap hyp(s)|}{|hyp(t) \cup hyp(s)|} \\ 0, & hyp(t) \cap hyp(s) = \emptyset \end{cases} \quad (7)$$

where $hyp(s)$ denotes the hypernyms for s returned by BabelNet and α and β has been set to 0.2 and 0.5 resp. based on empirical observations. Finally, we define hypernym overlap between T and S as weighted harmonic mean of $HypCoverage(S, T)$ and $HypCoverage(T, S)$.

WordNet Similarity Word Overlap (WSWO) is a metric that exploits semantic similarity of word pairs computed using WORDNET[5]. To be more precise, we compute for each word in a one text a word in the second one with maximum semantic similarity and then normalise the sum of such similarity scores. We first define *WordNet Coverage (WNC)* of T_1 in T_2 as follows.

$$WNC(T_1, T_2) = \frac{1}{|Words(T_1)|} \cdot \sum_{w_1 \in Words(T_1)} \max_{w_2 \in Words(T_2)} sim(w_1, w_2) \quad (8)$$

[2]Common substrings closer to the beginning of the text are scored higher.

[3]http://www.geonames.org/

[4]For computing hypernyms we used BabelNet API method which returns all hypernyms for a given synset (depth one).

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

where $sim(w_1, w_2)$ denotes WordNet-based semantic similarity measure between w_1 and w_2. In particular, we explored the following measures: **Path**[6], **WP** (Wu and Palmer, 1994), **Lesk** (Banerjee and Pedersen, 2002) and **HirstStOnge**, **LeacockChodorow**, **Resnik**, **JiangConrath** and **Lin** (Budanitsky et al., 2001). Finally, $WSWO(T_1, T_2)$ is defined as harmonic mean of $WNC(T_1, T_2)$ and $WNC(T_2, T_1)$.

Numerical Overlap is an overlap of the set of numerical expressions found in the texts being compared. While 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. Let $Num(T_i)$ denote the set of numerical expressions found in T_i. We then define **Absolute Numerical Overlap (ANO)** as follows.

$$AbsoluteNumericalOverlap(T_1, T_2) = \frac{2 \cdot |Num(T_1) \cap Num(T_2)|}{|Num(T_1)| + |Num(T_2)|} \tag{9}$$

We then define closeness of numerical expressions in T_i with numerical expressions in T_2 as follows.

$$NumericalCloseness(T_1, T_2) = \frac{1}{|Num(T_1)|} \cdot \sum_{t \in T_1} \max_{s \in T_2} closeness(t, s) \tag{10}$$

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

$$closeness(t, s) = \begin{cases} 1 - \log_2(1 + \frac{|t-s|}{\max(t,s)}), & type(t) = type(s) \\ 0, & type(t) \neq type(s) \end{cases} \tag{11}$$

Finally, we define **Relative Numerical Overlap (RNO)** between T_1 and T_2 as a weighted harmonic mean of $NumericalCloseness(T_1, T_2)$ and $NumericalCloseness(T_2, T_1)$.

Cosine of Text Vectors (CTV) is computed as $Cosine(Doc2Vec(T_1), Doc2Vec(T_2))$, 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., 2011) word embeddings.

3.1.1 Text Preprocessing

In the case of most of the metrics we deploy pre-processing of the text, which mainly boils down to: (a) lowercasing it, (b) normalising whitespaces, (c) removing constructs such as urls, etc. As regards WNO, $WSWO$ and WWO some initial/final token characters are stripped (e.g., brackets), while for computing $WSWO$, WWO and CVT one removes stop words using a list of ca. 250 English word forms. In the case of NEO and HO the texts are not downcased since this might have had deteriorated NE recognition performance which relies on orthographic features. For computing ANO and RNO no pre-processing is carried since non-alphanumeric characters often constitute part of numerical expressions.

3.2 Meta-data based Metrics

As regards meta-data information we define four metrics that exploit event location, category and type information. Since 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 (LSA) computes the administrative distance between locations. It is a modification of WUP metric presented in (Wu and Palmer, 1994) and it aims to reflect how close two locations are with respect to an administrative hierarchy of geographical references. Let T_{GEO} denote the 4-level (Country, Region, Province and Populated Place) administrative hierarchy in the GeoNames gazetteer[7] and let $lcs(x, y)$ denote the lowest common subsumer for nodes x and y in T_{GEO} and $Loc(e)$ denote the node in T_{GEO} that corresponds to the location of the event e. LSA is then defined as follows:

[6] Counting the length of the path in 'is-a' Verb and Noun hierarchy

[7] http://www.geonames.org

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-Ks Pulwama, 3rd death this week*
DESCRIPTION: *This was the third civilian killed in firing incidents this week. Earlier, one civilian was killed in Srinagars 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 2: An example of two events perpetrated by the same group as part of the same armed conflict.

$$LSA(e_1, e_2) = \frac{2 \cdot \omega(lcs(Loc(e_1), Loc(e_2)))}{\omega(Loc(e_1)) + \omega(Loc(e_1))} \tag{12}$$

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.

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

$$LSG(e_1, e_2) = (\ln(dist(coord(e_1), coord(e_2)) + e))^{-1} \tag{13}$$

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 km. between the 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 resp. The metrics are then defined as follows.

$$EventCategorySimilarity(e_1, e_2) = Prob(RELATED(e_1, e_2)|(cat(e_1), cat(e_2)) \tag{14}$$

$$EventTypeSimilarity(e_1, e_2) = Prob(RELATED(e_1, e_2)|(type(e_1), type(e_2)) \tag{15}$$

The respective probabilities for category and type pairs have been computed using the GOLD dataset (see Section 4.1). In case certain combination of types (categories) was not observed the respective probability was set to zero, whereas in case the type/category equality the resp. probability was set to 1.

4 Experiments

4.1 Dataset

We built 2 corpora consisting of event template pairs taken from the event dataset 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 5) around keys consisting of a category, location (country) and a timeslot (e.g. time window of +- 2 days) in 2017. Each of such initial groups G was 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[8]. Finally, G was amended by adding $|G|/3$ of randomly selected events (disjoint from the previous groups) from the same time window, regardless of location, category and similarity.

For each resulting group, all event pairs were computed, which were then labeled by 4 annotators, who were asked to consider only textual and meta-data information in the templates. The average pairwise κ score for inter-annotator agreement on a sample of around 13.4K event pairs was over 0.85. Questionable cases were typically due to event granularity issues. For example, the two events in Figure 2 were arguably perpetrated by the same armed group as part of a same armed conflict in the same day and larger area. Whether the two killing incidents should be considered as different consequences of the same larger armed conflict event and thus be considered as related, or should they be considered as distinct events sharing a large number of slot values is an open question.

[8]Vector representations of event templates and thus centroid templates of groups are derived by computing Doc2Vec on joined DESCRIPTION, TITLE and SNIPPET textual slot and converting each word with GloVe word embeddings (Pennington et al., 2011)

Corpus	#RELATED	#UNRELATED	#CRI-VIO	#CIV-POL	#MM-DIS	#NAT-DIS	#MIL
GOLD	10705	6074	76.0%	3.65%	12.71%	4.0%	3.65%
SILVER	10606	11060	67.47%	7.0%	16.45%	4.15%	4.89%

Table 1: GOLD/SILVER dataset statistics. The first 2 columns provide number of related (unrelated) event pairs, the others provide % of events falling into: crisis-violence (CRI-VIO), civic-political action (CIV-POL), man-made disasters (MM-DIS), natural disasters (NAT-DIS) and military actions (MIL).

We used pairs with at least 2 non-conflicting judgments to build a GOLD dataset, whereas SILVER dataset was created on top of it through adding event pairs annotated only by one annotator. Detailed statistics are provided in Table 1.

4.2 Discriminative power of the Event Similarity Metrics

In order to have a preliminary insight into the discriminative power of the various event similarity metrics we exploit an objective measure *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 resp. for all considered bins $i \in I$. *absDistance* is then defined as follows.

$$absDistance(h) = \sum_{i \in I} |u_i^h - r_i^h|/200 \tag{16}$$

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 event description with the snippet is used (D), (b) only event title is used (T), (c) in addition to (a) the title is exploited as well (D+T), (d) similarity score for the title and description/snippet is computed separately and an average thereof is returned (AVG(D,T)), and (e) similarity score for the title and description/snippet is computed separately and the maximum of the two is returned (MAX(D,T)).

Figure 3 provides a comparison of the discriminative power computed using $absDistance$ on GOLD dataset 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 meta-data metrics, namely LSG (more than 90% of the AUC) and ETS (more than 30% of the AUC), whereas NEO and WWO (both of which can be computed efficiently) lead 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 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.

4.3 Experiment Setup

Experiments were carried out using five different ML models, namely: SVM, Stochastic gradient descent classifier (regularized linear model learned), Decision Tree, Random Forest and AdaBoost classifier. All models were implemented using (Scikit-learn, 2011). Hyper-parameters of each model were tuned using grid search. Each model was trained using full set of event similarity metrics as features[9] and on a subset

[9]As regards $WSWO$ metric family we finally considered only $WSWO - Path$ and $WSWO - WP$ variants based on some empirical observations which revealed that the other variant produce very similar scores

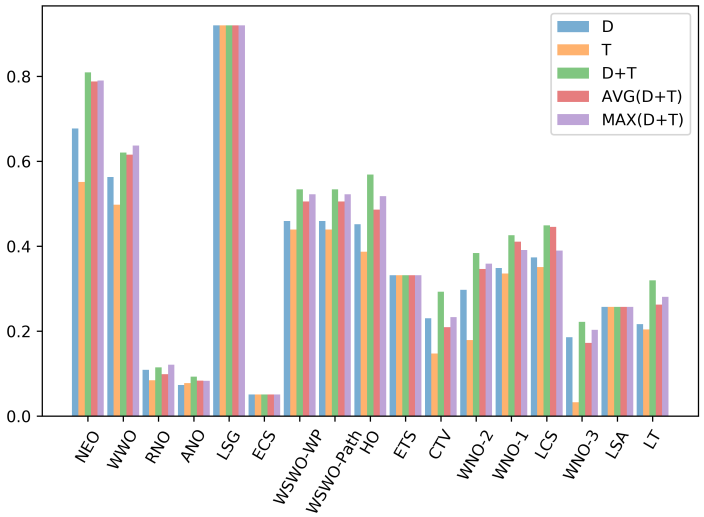

Figure 3: Discriminative power of the event similarity metrics.

of features obtained using feature selection *SelectFromModel* with base estimator being Random Forest. All models consistently exhibited better performance when using all features vis-a-vis subset of features obtained through feature selection. All models were trained on the same train-test split (80:20) and 5-fold cross-validation was performed.

Noteworthy, in case of 'missing' features, i.e., whenever 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 on both datasets described in 4.1 in two set-ups, one with text-based features only, and second one with both textual and meta-data features.

4.4 Results

The performance of the models on GOLD and SILVER datasets is shown in Table 2. The observed results indicate that the task is well modeled by the different classification paradigms, with a Random Forest being in general the top scoring model across all settings. We have also 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 matches to a larger extent the discriminative power ranking depicted in Figure 3, i.e., NEO, WWO topping the rankings, and ANO and RNO ranking lowest.

As expected (see 4.2) adding meta-data 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 94.7% (92.9% - SILVER) to as much as 98.6% (97% - SILVER). Nevertheless, this is a remarkable finding considering that the meta-data features (i.e., the slots LOCATION, TYPE and CATEGORY) are automatically generated by an event extraction engine and their extraction is more error prone vis-a-vis computation of similarity metrics on the textual slots. One needs to emphasize in this context that the surprisingly high discriminative power of LSG metric that contributed to the overall performance might have been potentially due to the way how the evaluation corpora were built (see Section 4.1).

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 GOLD dataset with only textual features. Exploiting only title information (T mode) when using text-based features resulted in achieving a respectable F1 score of 84.4% (GOLD) and 80.7% (SILVER).

A rudimentary error analysis on the output of GOLD dataset-trained Random Forest classifier with meta-data 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 4 (top). This was expected as the text pairs have little lexical overlapping and (more) background knowledge (e.g.

ML Paradigm	Text and Meta-data features D	T	D+T	AVG(D,T)	MAX(D,T)	Text-based features D	T	D+T	AVG(D,T)	MAX(D,T)
SVM	97.39%	97.17%	97.98%	97.66%	97.66%	86.22%	75.20%	93.90%	92.01%	92.24%
SDG	97.42%	97.37%	97.27%	97.83%	97.88%	85.20%	76.11%	93.53%	90.99%	91.80%
RANDOM FOREST	**98.38%**	**98.54%**	**98.57%**	**98.43%**	**98.49%**	**88.69%**	**84.40%**	**94.71%**	**93.62%**	**93.61%**
DECISION TREE	97.92%	97.87%	98.06%	98.18%	98.16%	86.36%	78.93%	94.32%	92.73%	92.99%
ADABOOST	97.85%	97.67%	98.09%	98.04%	98.02%	87.19%	79.68%	93.94%	92.31%	92.55%

ML Paradigm	Text and Meta-data features D	T	D+T	AVG(D,T)	MAX(D,T)	Text-based features D	T	D+T	AVG(D,T)	MAX(D,T)
SVM	95.78%	95.47%	96.62%	96.07%	96.22%	83.40%	73.02%	91.17%	88.71%	88.93%
SDG	96.08%	95.78%	96.68%	96.28%	96.25%	82.99%	73.26%	90.42%	87.61%	88.54%
RANDOM FOREST	**96.76%**	**96.80%**	**97.01%**	**96.96%**	**96.87%**	**85.39%**	**80.74%**	**92.89%**	**91.42%**	**91.60%**
DECISION TREE	96.23%	96.50%	96.50%	96.21%	96.41%	83.70%	76.39%	92.05%	90.66%	90.59%
ADABOOST	95.96%	96.27%	96.27%	96.06%	96.20%	83.63%	75.22%	91.28%	89.98%	89.84%

Table 2: Performance on the GOLD (top) and SILVER (bottom) dataset (F1 scores).

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 at Manchester Arena earlier this week, leaving 22 people dead and over 50 injured and Miley, 24, admitted it has affected [...]*

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 groups most important bases in Syria, in the northwestern province of Idlib, Syrian Observatory for Human Rights director Rami Abdel Rahman told AFP.*

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, the Syrian Observatory for Human Rights said.*

Figure 4: A sample false negative (top) and false positive (bottom) event pair.

access to full news articles) is required in order to draw a relatedness link. On the other hand, the models struggled to set apart individual incidents (see Figure 4 - 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.

5 Conclusions

This paper reported one experiments of testing ML methods using a wide range of textual and meta-data features to train classifiers for linking related event templates that have been automatically extracted from online news. While exploiting solely textual features resulted in achieving 94.7% F1 score, adding meta-data features allowed to improve it up to 98.6%, 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 bigger countries are less likely to be related, utilization of the structure of the urls to the related sources which 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., 2017; Al-Badrashiny et al., 2017), and (d) introducing an additional sub-classification of the 'related' class. As a matter of fact we carried out an initial attempt to sub-classify a sample of 150 event pairs (e_1, e_2) labelled as related into one of the four sub-classes: 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 3 annotators involved ranged from 0.45 to 0.63 which indicates the complexity of the task.

All the resources used in the experiments, i.e., the annotated corpora, files with event similarity metric values in ARFF format and feature histograms can be accessed at: `http://labs.emm4u.eu/eventlinking/event-linking_version_1.0_29.06.2018.zip`.

References

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 Benjamin Van Durme. 2017. TinkerBell: Cross-lingual Cold-Start Knowledge Base Construction, *Proceedings of the 2017 Text Analysis Conference, TAC 2017*, 59–65.

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017. On the Creation of a Security-Related Event Corpus, *Proceedings of the Events and Stories in the News Workshop*, 59–65.

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, 136–145.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1412–1422.

Evgenia Belyaeva, Aljaž Košmerlj, Andrej Muhič, Jan Rupnik, and Flavio Fuart. 2015. Using semantic data to improve cross-lingual linking of article clusters, *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:64–70.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *Proceedings of NAACL 2001 Workshop on Wordnet and Other Lexical Resources*.

Tommaso Caselli and Piek Vossen. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction, *Proceedings of the Events and Stories in the News 2017 Workshop*.

Sophie Chesney, Guillaume Jacquet, Ralf Steinberger, and Jakub Piskorski. 2017. Multi-word Entity Classification in a Highly Multilingual Environment, *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 11–20.

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

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media *Proceedings of ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics*

Gary King and Will Lowe. 2017. An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders, *International Organization*, 57:617-642.

Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn. 2016. Event Linking with Sentential Features from Convolutional Neural Networks, *Proceedings of CoNLL 2016*, 239–249.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents, *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*.

Vladimir Levenshtein. 1965. Binary codes for correcting deletions, insertions, and reversals, *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2015. Overview of TAC KBP 2015 Event Nugget Track, *Text Analysis Conference*.

Gonzalo Navarro. 2001. A guided tour to approximate string matching, *ACM Comput. Surv.*, 33(1):31–88.

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

Joel Nothman, Matthew Honnibal, Ben Hachey, and James R. Curran. 2012. Event Linking: Grounding Event Reference in a News Archive, *Proceedings of ACL 2012*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation, *Proceedings of Empirical Methods in Natural Language Processing 2014*, 1532–1543.

Jakub Piskorski, Karol Wieloch, and Marcin Sydow. 2009. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages, *Information Retrieval*, 12(3):275–299.

Jan Rupnik, Andrej Muhič, Gregor Leban, Blaz Fortuna, and Marko Grobelnik. 2017. News Across Languages: Cross-lingual Document Similarity and Event Tracking, *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 5050–5054.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for Measuring Semantic Text Similarity, *Proceedings of the First 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*, 441–448.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, *Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS 2011)*, 801–809.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 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.

Zhibiao Wu and Marta Palmer. 1994. Verbs Semantics and Lexical Selection, *Proceedings of ACL 1994 - 32nd Annual Meeting on Association for Computational Linguistics*, 133–138.

Roman Yangarber, Peter Von Etter, and Ralf Steinberger. 2008. Content Collection and Analysis in the Domain of Epidemiology, *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*.