Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

EventDNA: a dataset for Dutch news event extraction as a basis for news diversification

Camiel Colruyt · Orphée De Clercq · Thierry Desot · Véronique Hoste

Received: date / Accepted: date

Abstract News organizations increasingly tailor their news offering to the reader through personalized recommendation algorithms. However, automated recommendation algorithms reflect a commercial logic based on calculated relevance to the user, rather than aiming at a well-informed citizenry. In this paper, we introduce the EventDNA corpus, a dataset of 1,773 Dutch-language news articles annotated with information on entities, news events and IPTC Media Topic codes, with the ultimate goal to outline a recommendation algorithm that uses news event diversity rather than previous reading behaviour as a key driver for personalized news recommendation. We describe the EventDNA annotation guidelines, which are inspired by the well-known ERE framework and conclude that it is not practical to apply a fixed event typology such as used in ERE to an unrestricted data context. The corpus and related source code is made available at https://github.com/NewsDNA-LT3/.github.

Keywords News recommendation · event annotation

1 Introduction

News recommendation algorithms are increasingly popular tools for online news organizations to tailor their offering to consumers (Thurman et al, 2019; Thurman and Schifferes, 2012; Liemans, 2019). The main news personalization paradigms define good recommendations in terms of similarity to user's previous reading behaviour. In content-based filtering, articles are recommended based on their proximity to other articles the user has read in terms of content, such as topics, keywords etc. (Liu et al, 2010; Adnan et al, 2014). In

Camiel Colruyt, Orphée De Clercq, Thierry Desot, Véronique Hoste*
LT^3, Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
E-mail: firstname.lastname@ugent.be

collaborative filtering, articles are recommended based on what other users with similar interests have read before (Sarwar et al, 2001).

However, defining personalization in terms of similarity to the user's previous reading behaviour presents the risk that users are exposed to an increasingly less diverse news offering, similar to the idea of filter bubbles (Pariser, 2011; Borgesius et al, 2016). This contrasts with the normative concept of journalism as a marketplace of ideas, where users are confronted with a spread of ideas reflecting those present in society (Joris et al, 2020). The NewsDNA project at Ghent University aims to tackle this issue by adopting an interdisciplinary approach to outline a news recommendation algorithm driven by diversity, such that the recommended content reflects the diversity of topics and events that occur in unfiltered news streams[1].

To diversify the content offered by news recommendation algorithms, often referred to as topic diversity, a mechanism is necessary to measure diversity in a collection of news articles and rate the similarity between items. We propose to use news events as the unit of analysis for this mechanism. A granular semantic analysis of news articles based on events can enhance the precision of news recommenders and can be used to increase the diversity of the proposed material. Identifying and linking events across incoming articles could allow us to collect them into buckets by the events they discuss. This allows for a general analysis of the diversity of events over a collection, and could later on allow us to examine e.g. sentiment and viewpoint diversity within the bucket of articles pertaining to a single event, ultimately leading to a more balanced public opinion. For instance once a (cross-document) event extraction system has found all articles referring to the waiving of coronavirus vaccines, a viewpoint diverse system ideally presents the reader with different articles in which the pros and cons of waiving the patents on Covid-19 vaccines are discussed.

For the current publication, we focus on the first step of modeling content diversity. While extracting named entities or topics can give a rough idea of the theme of a news article, identifying the specific events they mention represents a deep level of semantic understanding of the text. Specifically, *event mention extraction* refers to the task of identifying which spans in a text refer to real-world events and extracting certain features of that mention. Mentions are typically linked to entities that play a role in the event as arguments.

While event extraction has received much attention in recent years (Vossen et al, 2016; Peng et al, 2015; Aguilar et al, 2014; Doddington et al, 2004; Song et al, 2015; Pustejovsky et al, 2003a), it remains a challenging field with many unsolved problems. Events are conceptually difficult to delineate, especially in a task-oriented setup which considers the relevance of news events. They may be worded in idiosyncratic ways, making consistent annotation difficult (Vossen, 2018). Poor event mention recall – human annotators not recognising the same events – is an acknowledged issue (Mitamura et al, 2015b; Inel and Aroyo, 2019). In supervised learning contexts, many programs (Doddington

[1] https://www.ugent.be/mict/en/research/newsdna

et al, 2004; Song et al, 2015) define fixed semantic categories of events, such that event mentions are found and sorted into a layered taxonomy of semantic categories such as `Conflict-Attack` or `Transaction-TransferOwnership` (Aguilar et al, 2014). Events that fall outside the typology are considered irrelevant. This closed-domain setting limits more advanced applications for event extraction (Araki and Mitamura, 2018). It also forms a problem in transferring models to unrestricted data contexts, since existing training corpora used in these settings tend to be unnaturally skewed to contain more events from the typology (Grishman, 2010).

In an event extraction task on unrestricted data, we try to extract all relevant news items from incoming data, regardless of semantic type. Designing a taxonomy for this context is a difficult balancing act. A small taxonomy will exclude relevant kinds of events if they have not been foreseen, while a large taxonomy can become difficult to apply and easily lead to data sparsity problems (where classes that rarely occur in the data are difficult to predict). Additionally, the type distribution can change over time. For instance, after a certain terrorist attack, incoming news may be saturated with events of a certain type. A system trained on data from one period may be disadvantaged when dealing with news from a different period. Overall, the scope of annotated events depends more on their relevance as news items rather than their adherence to a taxonomic standard.

This work introduces efforts to design an event extraction system for incoming Dutch news text, the design of which is motivated by the NewsDNA use case. We describe the creation of the EventDNA corpus: a novel Dutch-language corpus of news events, entities, IPTC Media Topic codes [2] and coreference links, consisting of 1,773 news documents. We introduce the idea that event mentions can be described by longer clauses, in an effort to address difficulties with annotation in previous work, where mentions are restricted to one or two tokens. As the use of longer clauses brings additional challenges to match annotation spans over different annotators, we present our solution which takes into account syntactic head information for the calculation of inter-annotator agreement. Given our goal of extracting relevant news items in unrestricted news data, we focus on the annotation of news-relevant events. As part of the annotation effort, we also investigate whether it is possible or practical to further annotate event mentions according to a fixed typology. The Rich ERE (see Section 2) was chosen as a taxonomy for this annotation. We adapted this taxonomy by, among others, allowing for events that are relevant but fall outside the typology to receive an `Unknown` type label. We are the first to apply such a typology to Dutch news text. As the `Unknown` type makes up a large proportion of events in the resulting annotated corpus, we conclude that event type is not a meaningful event attribute in this task. Annotators were also asked to provide Media Topic codes for all documents,using the taxonomy provided by the International Press Telecommunications Council (IPTC). We

[2] IPTC Topics are a standardized taxonomy of news topics, comprising 17 top-level topics (e.g. crime, law and justice, politics or education) that are divided in increasingly granular subtopics (e.g. law enforcement, election or higher education).

find that correlating event mentions and IPTC Topics may provide a more useful bridge to link events across articles. We furthermore perform pilot event span extraction experiments using Conditional Random Fields, framing events as *IOB*-sequences over sentences, achieving a promising F1-score of 0.67.

We present the state-of-the-art for event annotation schemas and event detection in Section 2 and we outline the annotation guidelines and the annotation process in Section 3. Section 4 describes the inter-annotator study we performed to evaluate the guidelines and the consistency of the annotated corpus. In Section 5, we share insights on the composition of the annotations in the experimental corpus. Section 6 describes the pilot event detection experiments performed on the corpus.

2 Related work

Many conceptualizations of events in text have been designed and deployed for different research purposes. While programs such as FrameNet (Ruppenhofer et al, 2016) operate on a very granular level of lexical semantics, other systems are more task-oriented. In the latter, events must answer the more immediate needs of information retrieval or slot-filling tasks, and are typically restricted in scope. In this context, *event mention extraction* refers to the task of identifying which spans in a text refer to real-world events and extracting certain features of that mention.

2.1 Event annotation schemas

The schema used in this work branches off the ACE (Automatic Context Extraction) lineage of annotation systems. The Message Understanding Conferences (MUC) were the first foray into event extraction, casting it as a slot-filling task from 1987 to 1997 (Grishman and Sundheim, 1996; MUC, 2001). The ACE program succeeded MUC in 1999 with more or less the same goals, and inherited a strong connection between entities and events: Entity Detection and Tracking is "the core annotation task, providing the foundation for all remaining tasks" (Doddington et al, 2004). Entity extraction remains a major component of event extraction systems: entities serve important roles as arguments to events, and their presence has been used as predictors for events (Yang and Mitchell, 2016). Event extraction was only introduced in ACE in 2004; the last ACE evaluation took place in 2008 (Aguilar et al, 2014). Starting in 2014, the ACE schema served as a design basis for the family of schemas used in DARPA's Deep Exploration and Filtering of Text (DEFT) program.

A number of approaches to events were designed in the context of DEFT (Bies et al, 2016). First among them, the Entities, Relations, and Events (ERE) schema (Song et al, 2015; Aguilar et al, 2014) builds on the concepts introduced by the ACE evaluation. Its starting point, Light ERE, is a consolidated version of ACE which streamlines its predecessor's more complex aspects. Rich

ERE, introduced in 2015, expands the number of semantic types that define the scope of events, and introduces important considerations towards annotating event co-reference. ERE is meant to be extensible to tasks with different areas of focus. Event-Event Relations (EER) annotation focuses on the relations between ACE/ERE-style events (Bies et al, 2016). The ACE and ERE programs both employ strict taxonomies of event types that are considered relevant to annotate *a priori*, and the datasets used to train these systems are implicitly skewed towards these event types (Grishman, 2010). While the prototype ACE taxonomy was designed with hard news topics in mind, this raises the question whether it is practical to apply a taxonomy to a context where relevant events are not theoretically restricted by a typology. This is exactly what we explore in this research in Section 5.

The output of entity and event extraction systems can be used to populate knowledge bases. After ACE's final evaluation, the Knowledge Base Population (KBP) track of the Text Analysis Conference (TAC) was started to facilitate this kind of downstream work (Aguilar et al, 2014). Originally focusing on entity extraction, evaluations related to events were introduced in 2014. Two are specific to event arguments: Event Argument Extraction and Event Argument Linking (Bies et al, 2016). The ACE, ERE, TAC and KBP workshops and competitions stimulated the creation of data sets labeled with entities and events, e.g. the ACE 2005 corpus (Walker et al, 2006) but initially concentrated on short and fixed event types using *single-word* event spans. Single-word event triggers are usually (main) verbs, nouns, adjectives and adverbs. *Multi-word* event triggers (Mitamura et al, 2015a, 2016, 2015b), which are more challenging to predict, can be *continuous* when the event span consists of consecutive tokens and even complete sentences, or *discontinuous*, meaning that the words belonging to the multi-word event trigger not always lie next to each other. For the Dutch sentence *"XYZ moet extra personeelsleden vinden wegens uitval van werknemers."* [*(The company) XYZ has to find extra staff due to employee absence.*], Table 1 represents an example of a *single-word*, *multi-word discontinuous* and *continuous* annotation schema, where the square brackets denote the annotation boundaries. The infinitive *vinden* [*to find*] is selected as head verb as this contains the most prominent event information and not the auxiliary modal verb *"moet"* [*has to*]. The arguments of the head verb are the subject *"(The company) XYZ"* and object *"extra personeelsleden"* [*extra staff*]. Annotating multi-word triggers allows for a more precise annotation of ambiguous cases, and forms a break in tradition with the majority of other schemas. EventDNA, which is the corpus presented in this paper, applies this concept of an event as a multi-word span to the Dutch language by allowing entire clauses to be annotated as event triggers, in order to eliminate the ambiguity of choosing short triggers.

Event schemas are also found outside the ACE tradition. TimeML (Pustejovsky et al, 2003a) and its successor, ISO-TimeML (Pustejovsky et al, 2010) focus on events and their temporal relations. This ISO-TimeML standard was the basis for the time annotations in the Richer Event Description (RED) annotation project (O'Gorman et al, 2016), in which the focus was not so

Event annotation schemas
Single-word
`XYZ moet extra personeelsleden [vinden] wegens uitval van werknemers.`
[*(The company) XYZ has to find extra staff due to employee absence.*]
Multi-word discontinuous
`[XYZ] moet [extra personeelsleden] [vinden] wegens uitval van werknemers.`
Multi-word continuous (target annotation)
`[XYZ moet extra personeelsleden vinden] wegens uitval van werknemers.`

Table 1 Examples of single-word, multi-word discontinuous and continuous event annotation schemas

much on the relationships between events and the entities participating in them, but rather on the marking of entities, events or times in a document, and the temporal, causal, and coreference relations between them, leading to a timeline of events within a document. The ISO-TimeML guidelines also inspired the MEANTIME schema that was used for annotation in the NewsReader project (Vossen et al, 2016). NewsReader aims to construct knowledge bases from real-time news streams in different languages. In this context, the MEANTIME guidelines and corpus were translated from English into Italian, Spanish and Dutch (Minard et al, 2016), forming, as far as we know, the only Dutch-language event extraction dataset before the one presented here. It consists of 120 Wikinews articles annotated for events, entities, numerical and time expressions, and coreference. MEANTIME events differ from our events in two important aspects. First of all, MEANTIME accepts events triggers in the form of "verbs, nouns, pronouns, adjectives, or prepositional constructions" (Minard et al, 2016) whereas in this work, events can take the form of phrases, clauses or complete sentences. Secondly, MEANTIME events do not carry a semantic category, whereas in EventDNA we investigate the applicability of the ERE typology to the annotated news events.

While English is the traditional focus of event research, annotation has been applied to many languages. NewsReader is multilingual by design and provides parallel data in English, Italian, Spanish and Dutch. ACE has been applied to Chinese and Arabic (Doddington et al, 2004) and ERE has been applied to Spanish and Chinese (Song et al, 2015). TimeML, for its part, has been adapted for use in Hindi (Goud et al, 2019), French (Bittar et al, 2011), Italian (Caselli et al, 2011), Persian (Yaghoobzadeh et al, 2012), Korean (Im et al, 2009) and Basque (Altuna et al, 2018).

Our focus is different from the previously described approaches. In this work, we aim to design an event extraction system for incoming news text, the design of which is motivated by a content-based news recommender use case. This use case also guides our definition of what we consider as a news event, which partly deviates from previous definitions. TimeML, for example, considers events "a cover term for situations that happen or occur. Events can be punctual or last for a period of time" (Pustejovsky et al, 2003b). The ACE and ERE style of event annotation, e.g. Linguistic Data Consortium (2016),

also starts from a very broad definition, by considering events as something that happens or as specific occurrences involving participants, but then narrows down the event annotation to a particular set of types and subtypes. This adherence to a taxonomic standard was not the guiding principle for our annotations, but given our goal of extracting relevant news items from incoming data, we focused on the annotation of news-relevant events. Furthermore, we also wanted to assess to what extent the ERE standard was applicable to these annotated news events.

2.2 Event detection

As previously mentioned, event mention detection systems target event spans with a short length, often using the ACE 2005 corpus that has been annotated with single-word event spans. Initially, knowledge-based event detection methods were based on rule-sets (Valenzuela-Escárcega et al, 2015) or ontologies (Frasincar et al, 2009; Schouten et al, 2010; Arendarenko and Kakkonen, 2012). These also include extracting candidate event words with certain part-of-speech tags (Mihalcea and Tarau, 2004), which can also satisfy predefined syntactic patterns (Nguyen and Phan, 2009). Statistical methods detect event spans by means of n-grams (Witten et al, 2005; Grineva et al, 2009), term frequency inverse document frequency (TF-IDF), word frequency and word co-occurrence (Kaur and Gupta, 2010).

Event detection was recast as a binary classification problem using supervised machine learning approaches (Hasan and Ng, 2014) in order to decide whether an input word is part of an event or not. To that end, maximum entropy (Yih et al, 2006), support vector machines (SVM) (Lopez and Romary, 2010) and Conditional Random Fields (CRF) (Zhang, 2008) were applied. As these feature engineering approaches emerged, a larger scope than one-word event spans was targeted and hand-designed sets of lexical, semantic or syntactic features were extracted and fed into classifiers. More recently, deep neural networks superseded feature engineering-based methods, although the latter ones are still not definitely outperformed. Jacobs et al (2018) and Nugent et al (2017), for example, used lexical, syntactic features, word2vec (Mikolov et al, 2013), GLoVe (Pennington et al, 2014) and fastText (Bojanowski et al, 2017) word embeddings in a SVM classifier and reported better performances for the SVM classifier compared to a Recurrent Neural Network (RNN). Jacobs et al (2018), for example, compared an SVM, integrating rich lexical and syntactic features with a Long Short Term Memory (LSTM) RNN for a discontinuous multi-word economical event span classification task and reported the best performances for the SVM classifier (0.73 versus 0.70 F-score). Nugent et al (2017) also report better performances for an SVM classifier as compared to an attention based Recurrent Neural Network (RNN) approach and a Support Vector Machine (SVM) on a dataset for natural disaster and critical event detection (0.77 vs. 0.76 F-score). In contrast, Nguyen and Grishman (Nguyen and Grishman, 2015) demonstrated that Convolutional Neural Net-

works (CNN) significantly outperformed feature-based methods on the ACE 2005 task.

Other than most of the above-mentioned event detection approaches, in this work we evaluate the quality of our multi-word event span annotations, using a CRF model and by going beyond short event span detection. The CRF model is well capable of exploiting context information and still performs well using smaller sized data sets (Simonnet et al, 2017), other than deep learning models which require vast amounts of data. Section 6 describes how we train a CRF event detection model on the multi-word event mentions of our annotated EventNDA corpus, using a rich set of lexical and syntactic features.

3 Corpus description

In this paper, we present an annotated corpus of Dutch news events, the EventDNA corpus. This section describes the data collection (Section 3.1) and gives more insight in the fine-grained annotation guidelines that were developed to enrich the documents with annotations of entities, events and IPTC-topics (Section 3.2).

3.1 Data collection

The EventDNA corpus consists of 1,773 annotated news documents, where each document consists of the title and lead paragraph of a news article. Newspaper articles generally follow an inverted-pyramid structure, where the most important information is given first. Inspired by the work from van Dijk (1988) on the schematic structure of news texts, we selected the headline and lead paragraph as the text unit under consideration for our annotations as they uncover the most important information about the story and should attract the reader's attention to read the whole piece. We stripped off the subsequent paragraphs of each article, as these tend to elaborate on the lead paragraph with information that is less relevant.

In total, the corpus counts 6,937 sentences. For comparison, the ACE 2005 corpus, which is widely used for English-language event detection, counts 16,375 sentences over 599 documents (Li et al, 2013). MEANTIME – to our knowledge the only other Dutch-language event extraction corpus – consists of 120 news articles and 1,797 sentences (Minard et al, 2016).

Articles were selected for annotation from a large collection of archived news articles provided by Mediahuis, a media company that publishes several major newspapers in Belgium and the Netherlands. This collection comprises news articles from a number of Flemish online newspapers published between 2017 and 2018. Articles were randomly sampled from this collection and manually filtered. Only "hard news" items were retained, with a preference for international news items. We define hard news as serious news, urgent in nature, pertaining for example to economics, politics, war and crime. This can

be contrasted with news stories meant to entertain the reader, such as news on celebrities, sports or popular science. We consider hard news to be more relevant to the broader aims of the NewsDNA project, as recommended news items should reflect important topics in the public debate (Shoemaker, 2006; Patterson, 2000).

3.2 Annotation guidelines

This section summarizes the major features of the EventDNA annotation schema, which cover entities (Section 3.2.1), events (Section 3.2.2) and IPTC Media Topic codes (Section 3.2.3)[3]. The full annotation guidelines have been published as an open-access technical report (Colruyt et al, 2019a).

3.2.1 Entities

All news articles contain mentions of entities. What we call entities are objects in the real or fictional world that a string of text can refer to, like people, companies, places, etc. For each instance, the *entity type*, *mention level type* and attribute *individuality* are annotated. Furthermore, *coreference* links are marked between the annotated entity mentions referring to the same entity.

Mentions can point at three possible *entity types*: persons (abbreviated `PER`), organizations or companies (`ORG`) and places or locations (`LOC`). Additionally, we assign the label miscellaneous (`MISC`) as a fallback for entities that play an important role in news events, but cannot be clearly tagged as `PER`, `LOC` or `ORG`. An example would be the "Titanic" in the headline "RMS Titanic sunk off the coast of Newfoundland." Entity mentions that do not refer to specific, well-defined entities are not tagged. For instance, the companies mentioned in "Some companies weathered the crisis better than others" are ignored.

Entity mentions can be proper names (`NAM`, such as "Joe Biden"), nominal constructions (`NOM`, such as "the president of the United States") or pronouns (`PRON`); this is referred to as their *mention level type*. If the mention is nominal, the semantic head of the construction is also marked (HD). Entity mentions can also overlap with one another, as shown in example [1], in which three location entities are annotated, of which two named entity location mentions are embedded in a nominal location entity mention, which has "kasteel" as semantic head.

The *individuality* attribute indicates whether the mention refers to an individual (`IND`) entity or a group of entities (`GRP`), as exemplified in [2]. Metonymy, in which one expression is used to refer to the referent of a related one, is handled by noting the underlying semantic entity rather than the surface representation. As such, in example [3] "the White House" is considered an organisation entity if it used in that sense.

[3] https://iptc.org/standards/media-topics/

Coreference between entity mentions within the same document is also annotated, following the annotation guidelines for identity-of-reference annotation from Schuurman et al (2010). The conditions for indicating coreference are that the entities intuitively refer to the same entity and that they carry the same entity type. In example [4], the entity Mariano Rajoy is referred to by name, by a descriptive noun phrase, and twice with a pronoun.

[1] Het slachtoffer werd gevonden in [een 19e-eeuws [kasteel]$_{HD}$ bij de [Zweedse]$_{LOC|NAM}$ hoofdstad [Stockholm]$_{LOC|NAM}$]$_{LOC|NOM}$.
English: The victim was found in [a 19th century [castle]$_{HD}$ near the [Swedish]$_{LOC|NAM}$ capital [Stockholm]$_{LOC|NAM}$]$_{LOC|NOM}$.

[2] [Alle drie bedrijven]$_{GRP}$ zeggen de beslissing aan te vechten.
English: [All three companies]$_{GRP}$ say they will fight the decision.

[3] [Het Witte Huis]$_{ORG|NAM}$ legt nieuwe sancties op.
English: [The White House]$_{ORG|NAM}$ imposes new sanctions.

[4] [**Mariano Rajoy**]$_{PER|NAM}$, vorige week nog [**premier van Spanje**]$_{PER|NOM}$, liet gisteren al weten dat [**hij**]$_{PER|PRON}$ opstapt aan het hoofd van [**zijn**]$_{PER|PRON}$ Partido Popular.
*English: [**Mariano Rajoy**]$_{PER|NAM}$, who last week still was [**prime minister of Spain**]$_{PER|NOM}$, announced yesterday that [**he**]$_{PER|PRON}$ resigns at the top of [**his**]$_{PER|PRON}$ Partido Popular.*

3.2.2 Events

As mentioned in Section 2, event schemas usually recognize a single token as the trigger for an event. However, during preliminary testing, we often found it difficult to recognize single-token triggers, because events can be lexicalized in highly ambiguous ways. Other projects have noted the same difficulty and allowed for longer annotations (e.g. Mitamura et al (2015b)). EventDNA therefore generalizes the idea of event triggers further: event mention spans are taken to be multi-word spans or phrases, rather than single tokens. Although one of the annotation guidelines for EventDNA states that relevant semantic information has priority over syntactic information, meaning that the event annotation was not applied on top of a syntactically parsed corpus, the annotations evidently often coincide with grammatical clauses. Event annotation often targets verbal clauses (as part of a sentence or even a complete sentence in itself), as in "a terrorist attack took place in New York". But also nominal phrases (e.g. "the terrorist attack in New York") can be marked as an event mention.

Within EventDNA, event mentions are anchored to a certain textual span and additionally carry the following attributes: *prominence*, *type and subtype*, *negation*, *modality* and *tense*. Besides this, *arguments* to the events are annotated and *coreference* between events inside the document is also marked.

Prominence is an original attribute compared to previous schemas, although this idea of prominence is not new. In the investigation of speech and spoken dialogue, for example, information structure aspects such as saliency, givenness or focus and their acoustic correlates (e.g. pitch accent) are widely used to label prominence (Calhoun et al, 2010; Sridhar et al, 2008). While the

focus there is mainly on words and nominal constituents at the sentence level, our prominence annotation concerns the annotation of events in running news text. The prominence attribute can take two values, `main` or `background`. Main events are those events that cause the reporter to write the article: they are the new information the piece reports on. Background events are those that give background or context to the main event. While being inspired by frameworks as proposed by van Dijk (1988), this fairly shallow prominence annotation was decided because of feasibility concerns, and because we hypothesized it to be sufficient for the content-based news recommendation application we have in mind. In their annotation efforts, Baiamonte et al (2016) concluded that 72% of the main or foreground events were located in the opening sections of the articles.

All selected events were further annotated with type and subtype information. The Rich ERE guidelines were chosen as a basis for this annotation (Song et al, 2015). The ERE standard operates with a fixed typology of 9 types and 38 subtypes. The EventDNA typology inherits this typology, but a few adaptations were made. A full list of event *types and subtypes* found in ERE and their counterparts in NewsDNA can be found in Appendix A. Regarding the adaptations, `Business.DeclareBankruptcy` was merged with `Business.EndOrg`, and `Personell.Nominate` was merged with `Personell.StartPosition` since we consider that one type subsumes the other. Besides these mergers, initial annotation testing rounds revealed the need for a small number of new (sub)types found in hard news texts. The types `Conflict` and `Justice` were each enriched with generic subtypes (`Conflict.Conflict` and `Justice.Justice`) to account for events that could not be brought under an existing subtype. `Journalism.Publication` was created to cover events such as "Turkish media report two incidents that occurred at the polls yesterday". `Journalism.Investigation` to cover journalistic investigations, as in "the Russian press dove into Toporovski's past". The type `Politics` with one subtype `Vote` to cover events of all types of popular elections. Finally, the generic type `Unknown` (with subtype `Unknown`) to cover any other event considered relevant for annotation. This results in a typology with 12 types and 41 subtypes. Especially the addition of an `Unknown` type represents a break with the ACE/ERE tradition.

In an unrestricted data context like the one NewsDNA describes, we thus hypothesized that relevance should be the key factor for deciding which events should be annotated. Throughout the annotation process, the guiding questions the annotators used to determine whether an event should be annotated were the following:

- Is this candidate event *relevant* as a news item? If so, is it an event that caused the author to write the article on or is it an event giving background or context to the main event?
- Can the *candidate* event be classified into one of the adapted ERE typology types and subtypes?

In the example given in Figure 1, the main event is that Kitty Van Nieuwenhuyse's murderer will not be released (a grammatically negated event). An

event mention later in the same article refers to the murder itself. We consider the murder event to be a background event as it provides necessary context.

The attribute *Negation* can take two values: `negative` when the mention is grammatically negative, as in "Kitty Van Nieuwenhuyse's murderer will not be released", or `positive`.

Modality refers to the level of assertion of the event: an event is `asserted` if the writer refers to it as having really occurred or as certainly occurring in the future. It is set to `other` for events that are believed to have happened, that may happen, that have been proposed, or that exist under any other kind of hypothetical condition. Please note that this is not tied to negation, in that negation is an explicit grammatical feature seen in the event mention, while modality refers to the event's real or hypothetical nature. In the sentence "Kitty Van Nieuwenhuyse's murderer will not be released", the event (the non-release of the murderer) comprises a negation, but is definitely asserted. This would be different if the sentence would be "Commenters speculate that Kitty Van Nieuwenhuyse's murderer will not be released".

Tense indicates the event mention's grammatical tense and can thus take the value of `past`, `present` or `future`. In the case of nominal events it takes the value `unspecified`.

This brings us to the *arguments* which are annotated to the events and for which we follow the standard set by Rich ERE[4]. Most arguments correspond to entities that fulfill roles in the event. For instance, the event type `Conflict.Attack` takes as arguments an *Attacker*, a *Target* and a *Place*, which may be filled in by `PER`, `LOC` or `ORG` entities. Other arguments do not correspond to entities, such as the *Time* the attack takes place.

3.2.3 IPTC Media topics

Annotators were also asked to provide Media Topic codes for all documents, using the taxonomy as issued by the International Press Telecommunications Council (IPTC). The International Press Telecommunications Council is a body composed of more than 60 news organizations and companies which aims to develop technical standards to coordinate news activity across the world.[5] The IPTC Media Topic codes framework is a taxonomy of language-agnostic topics that can be used to classify news stories.[6] It defines 17 top-level codes that branch into subordinate topics up to five levels down. For instance, the top-level IPTC topic `Economy, business and finance` is divided into `Business information, Economic sector, Economy` and `Market and exchange` on the second level. `Business information` subdivides into `Business finance, Human resources, Strategy and marketing` and `Corporate social` on the third level, and so on. The deeper into the tree, the more granular and

[4] We refer to Appendix A of the annotation guidelines (Colruyt et al, 2019a) for a complete overview

[5] `https://iptc.org/about-iptc/`

[6] `https://iptc.org/standards/media-topics/`

specific the topic definitions become. A total of 1,226 topics can be defined this way.

Annotators were provided with the taxonomy and asked to mark each document with all relevant IPTC topics, choosing the most specific ones applicable. These specific tags can be traced back up the tree to the desired level of granularity.

Figure 1 gives an example of a fully annotated news headline. The first box shows the annotated **entities**. There are two nested entities in this example: "Moordenaar Kitty Van Nieuwenhuyse / Kitty Van Nieuwenhuyse's murderer" is marked as `PER` (person), `INDIV` (individual) and `NOM` (a nominalized entity mention). It is linked to a head, "moordenaar / murderer". The second entity mention is "Kitty Van Nieuwenhuyse", which is marked as `NAM` (a named entity). The second box represents the **event annotations**, the entire sentence is marked as one event mention trigger of the type `Justice.ReleaseParole`. Regarding the other attributes, the event has been labelled as a main and asserted event which is written in the present tense and comprises a negation. This events carries one argument, "Moordenaar Kitty van Nieuwenhuyse" in the *Person* role, which corresponds to the previously described entity. Finally, the third box illustrates annotion of **IPTC Media Topics**. This headline has been annotated with the IPTC topics `Homicide, Police, Criminal law` and `Prison`, which are all subtopics of the `Crime, law and justice` top-level label.

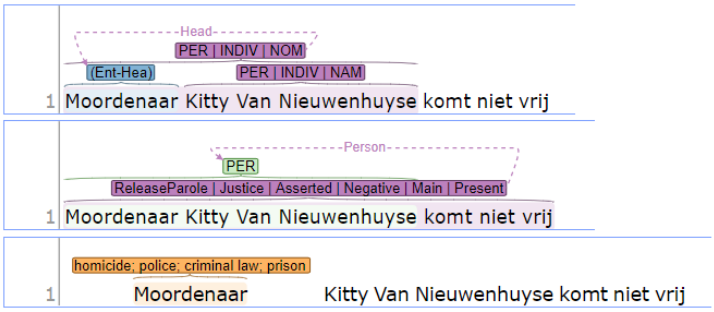

Fig. 1 Fully annotated example of a headline from the EventDNA corpus. *English*: *Kitty Van Nieuwenhuyse's murderer will not be released.*

4 Annotation process and agreement study

Annotation was performed by four annotators using the WebAnno tool (Yimam et al, 2014). These annotators were all native Dutch speakers and university students with a background in linguistics, level Master or Postgraduate

Course. In order to be hired, they first had to successfully pass a test and after selection they were thoroughly trained by an expert supervisor.

In a first phase they all annotated, independently from one another, the same set of 38 documents. These annotations were used to conduct an inter-annotator agreement study, the results of which are presented below. Following this, each of the four workers annotated a portion of the corpus individually and at their own speed. Two were hired full-time, and two part-time. All annotators worked individually, but were urged to call on the help of an expert supervisor to discuss and resolve difficult cases.

4.1 Event mention recognition

This study uses F1-scores to evaluate event mention recognition, i.e. in how far do different annotators identify the same event mentions in a sentence. To do so, we designed and applied a fuzzy matching mechanism to accurately match event annotations from one annotator to another. Only mention spans (as tokens) are compared; other features of the event descriptions are not taken into account.

F1-scores over event annotations were computed by taking each time one annotator as the gold standard and scoring the annotations of the other for precision and recall. Because this is a recall task rather than a categorical classification task, we consider that Cohen's Kappa and related measures, which are popular in agreement studies, are not applicable. Specifically, event mention identification is often phrased as a categorization over tokens, where trigger tokens are labeled as positive instances and other tokens as negative instances. Cohen's Kappa is a natural fit in this scenario. However, in this study, we consider the annotations over a sentence as a "bag of annotations", and then map the annotations of one worker to those of another. The more matches found, the better the consistency between annotators.

Furthermore, EventDNA annotation generalizes event mentions to multi-word spans, i.e. groups of words (rather than single tokens) that can take the form of a nominal constituent or a verbal clause, if they comprise a subject and verb. In this setting, it is possible that annotators recognize the same events mentions, but mark slightly different token spans, either through error or a different interpretation of the mention scope. Table 2 shows an example of two overlapping mentions that refer to the same event while not having the exact same span, due to one annotator omitting the descriptive phrase "in België ondergedoken" "gone into hiding in Belgium"). Achieving a good mapping between two buckets of annotations is therefore not straightforward, and requires a fuzzy matching mechanism.

The matching mechanism used in this work is described in Colruyt et al (2019b). We refine token-based matching methods by using the syntactic heads of each annotation. Intuitively we only consider annotations to match if the "semantic core" of their constructions agree. Given a pair of annotations like [There were several violations — There were several violations severe enough

In België ondergedoken internationaal gezochte hacker opgepakt aan Poolse grens [Internationally hunted hacker who had gone into hiding in Belgium caught on the Polish border]	
Annotator A	Annotator B
`In België ondergedoken internationaal gezochte hacker opgepakt aan Poolse grens`	`internationaal gezochte hacker opgepakt aan Poolse grens`

Table 2 An example of matching overlapping event mentions.

to talk about a breach of confidence], we would roughly identify "violations" as the core element of the event described. If two annotations share the same semantic core, we consider them to match. Conversely, in the pair [There were several violations severe enough to talk about a breach of confidence — a breach of confidence], the semantic cores are different and the annotations do not match. The "breach of confidence" is not the focus of the first event mention. We correlated this idea of semantic cores with the syntactic heads of the mention. These heads are extracted using the state-of-the-art Alpino dependency parser for Dutch (van Noord, 2006). We report the results obtained using two head-selection methods. The first considers as heads all nodes and descendants of nodes that have "HD" (head) as a dependency label. The second constrains heads by discarding heads that are part of adverbial phrases, adjective phrases, modifiers and prepositional phrases.

Given the head sets of each annotation, the matching function proceeds as follows. Given annotations a_1 and a_2, the sets of their tokens is compared. If they overlap perfectly, a_1 and a_2 are assumed to match a priori. If they do not overlap at all, this is counted as a negative match. The fuzzy matching mechanism comes into play when the match is only partial. The similarity between head sets is calculated using the Dice coefficient; if the Dice score exceeds a theshold of 0.8, a positive match is found. The threshold of 0.8 was found to be optimal after testing a series of threshold values against the judgment of a human evaluator. If for any reason no syntactic heads are found in an annotation, the function falls back to comparing the sets of tokens of a_1 and a_2. It similarly reports a positive match if the Dice score between token sets exceeds 0.8. To come back to the example in table 2, the two head sets extracted from the mentions are shown; the Dice score of these two sets is 0.91, and this is judged to be a positive match.

To perform the IAA study, the annotations of each pair of annotators X and Y are collected. Each pair of annotations is considered a potential match and run through the matching function. Results are tallied in a confusion matrix; one of the annotators, X, is arbitrarily considered the gold standard. A positive match is counted as a true positive outcome. An annotation that X found but Y did not is a false negative (since there should have been a positive match), and annotations in Y but not in X are false positives. No true negatives are counted. Recall, precision and F1-score are computed from this matrix.

	All heads			Restricted heads		
Pair	Precision	Recall	F1	Precision	Recall	F1
A-B	0.62	0.84	0.72	0.64	0.87	0.74
A-C	0.80	0.75	0.78	0.84	0.78	0.81
A-D	0.74	0.72	0.73	0.78	0.75	0.76
B-C	0.89	0.63	0.73	0.89	0.63	0.74
B-D	0.81	0.57	0.67	0.85	0.59	0.70
C-D	0.74	0.71	0.72	0.73	0.75	0.74
Avg	0.77	0.70	0.73	0.79	0.73	0.75

Table 3 Precision, Recall and F1 scores over annotator pairs.

Table 3 shows the results of this analysis. Given the difficulty of the task, F1 scores are good overall. There is noticeably higher disagreement between annotators *B* and *D*. During the rest of the annotation process, care was taken to reconcile difficult cases. To this purpose, all annotators were urged to reach out to an expert supervisor whenever they were in doubt about a certain annotation in order to reach a consensus. On the basis of these IAA results, the annotations by annotator *A* were selected and incorporated into the full corpus.

4.2 IPTC code annotation agreement

To evaluate the agreement of the IPTC code annotation, we also gathered the precision, recall and F1 scores over each annotator pair. In this setting we resolved the annotated IPTC codes to their top-level equivalents, i.e. at the most general level of granularity, which implies 17 different labels. Table 4 shows the results of this evaluation. The F1 score over pairs averages out to 0.74. While encouraging, we believe this result also reflects the ambiguous nature of topic annotations.

Pair	Precision	Recall	F1
A-B	0.67	0.82	0.74
A-C	0.74	0.78	0.76
A-D	0.77	0.75	0.76
B-C	0.79	0.69	0.74
B-D	0.77	0.61	0.68
C-D	0.81	0.74	0.77
Avg	0.76	0.73	0.74

Table 4 IAA precision, recall and F1 scores for IPTC topic annotation over annotator pairs.

5 Corpus statistics

In this section we offer more insight into the composition of the annotated EventDNA corpus when it comes to entities (Section 5.1), events (Section 5.2) and IPTC-topics (Section 5.3). We each time present a detailed overview of the annotated instances and their attributes, but start this section by presenting some general corpus statistics in Table 5.

Item	Count
Documents	1,773
Sentences	6,937
Tokens	106,106
Entities	17,491
Events	7,409
IPTC Topics	4,327

Table 5 General statistics of the EventDNA corpus.

Regarding the size of the corpus, the EventDNA corpus is smaller than the popular English ACE 2005 corpus, especially when it comes to the number of sentences. However, when we compare our corpus to the only other Dutch-language event extraction corpus, i.e. MEANTIME, which consists of 120 news articles and 1,797 sentences, we can state that EventDNA presents a substantial new resource to the field of Dutch event extraction.

Regarding the annotations, in total the 1,773 documents have been enriched with no less than 17,491 entity annotations, 7,409 event annotations and 4,327 IPTC Topics.

5.1 Entity annotations

Attribute	Value	Count	% of total
Type	Person (`PER`)	6,986	39.94
	Location (`LOC`)	5,549	31.72
	Organisation (`ORG`)	4,450	25.44
	Miscellaneous (`MISC`)	506	2.89
Mention Level Type	Proper name (`NAM`)	10,095	57.72
	Nominal construction (`NOM`)	5,741	32.82
	Pronoun (`PRON`)	1,655	9.46
Individuality	Individual (`IND`)	14,830	84.79
	Group (`GRP`)	2,661	15.21

Table 6 Distribution of attributes over entity annotations.

As can be derived from Table 5 our corpus comprises a large number of entities. In fact every document in the corpus contains 9.87 entities on average.

Let us now dive deeper into the distribution of the different entity attributes as summarized in Table 6. Regarding the *entity types* we can state that the entity mentions are spread fairly evenly across `PER` (40% of entities), `LOC` (32%) and `ORG` (25%) types, with a minority of `MISC` entities (3%). As explained in Section 3.2.1 `MISC` entities are entities which do not fall under another type, but nevertheless play an important role in the news story. These entities were manually examined and found to consist mostly of mentions referring to abstract entities of a political nature (e.g. the "Trump's new travel ban"), to vehicles that are the agents of events (e.g. "a Ryanair Boeing 737-800"), and to products and objects (e.g. "Bose headphones", "fentanyl").

Looking at the *mention level type* the great majority of mentions refer to named entities; to be exact, 57.72% of the entities are named entities (`NAM`), 32.82% nominal constructions (`NOM`) and 9.46% pronouns (`PRON`). A possible explanation for this skew towards named entities is that the documents in EventDNA consist of only a news article's headline and lead and that named entities frequently occur at the beginning of a text, whereas subsequent entities are more often nominal or pronominal mentions referring to the same entity. A surface level examination revealed that in 56% of the EventDNA document the first entity to occur was indeed a named entity. In order to get more insight into this we also had a closer look at the coreference annotations. In total, there are 6,111 coreference chains of two or more entities in the corpus. Of these, 3,255 (53%) start with a named entity as the first mention in the sentence: 1,255 chains (21%) begin with a named entity and are followed by a nominal or pronominal entity at any point in the chain, while 2,000 chains (33%) are composed entirely of named entities. 1,634 chains (27%) chains consist of only nominal or pronominal mentions, with the rest consisting of a mix of all three mention level types.

Regarding the *individuality* attribute we clearly observe that most entities mentioned throughout the corpus refer to an individual (`IND`).

5.2 Event annotations

In total 7,409 events have been annotated, this boils down to an average of 4.18 events per document and one of 1.3 events per sentence.

Let us again consider the distribution of the annotated event attributes, starting with the *prominence* as presented in Table 7.

Attribute	Value	Count	% of total
Prominence	`Main`	4,249	57.36
	`Background`	3,158	42.64

Table 7 Distribution of the attribute *Prominence* over all event annotations

We observe that the `Main` events outnumber the `Background` events. During the annotation process, it was noticeable that events in the title are usually

main events. Indeed, as shown in Figure 2, 75% of events that are mentioned in the title (labeled as sentence_1) of the article are `Main` events, with this proportion dropping in the following sentences. It is notable that, with 4,249 `Main` events, there are on average 2.4 `Main` events per document. However, many documents are written such that the title gives a first mention of the `Main` event, and a second mention of the same event is given in the following paragraph.

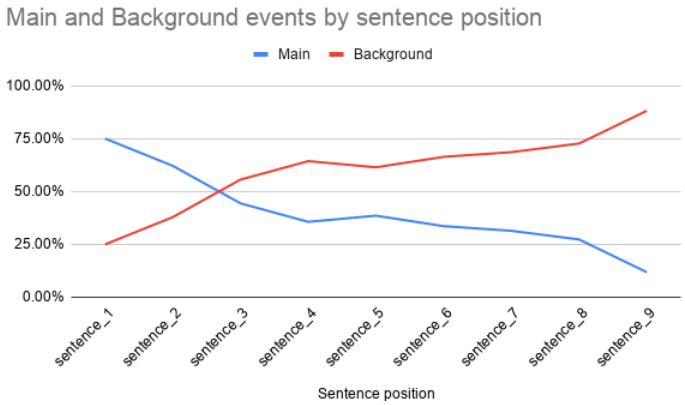

Fig. 2 Percentage of `Main` and `Background` events occurring at each position in the article.

This brings us to a discussion of the attribute *types* as presented in Table 8. What immediately draws attention is that 35% of the annotated events were assigned the type `Unknown`, followed by 17% of `Contact` events.

Type	Count	% of total
`Unknown`	2,559	34.55
`Contact`	1,321	17.83
`Conflict`	821	11.08
`Justice`	798	10.77
`Life`	514	6.94
`Transaction`	358	4.83
`Movement`	352	4.75
`Personnel`	293	3.96
`Politics`	206	2.78
`Journalism`	88	1.19
`Business`	70	0.95
`Manufacture`	27	0.36

Table 8 Distribution of the attribute *types* over the event annotations

Table 9 presents some examples of these `Unknown` events (*UNK*). These events can roughly be broken down into a number of categories. Natural or

Type of UNK event	Examples
Crises and disasters	- *Opnieuw zware aardbeving in Mexico (again a heavy earthquake in Mexico)* - *De bosbranden in de Amerikaanse staat Californië (the forest fires in the American state of California)* - *De almaar ernstiger wordende crisis rond het gebruik van opioïden (pijnstillers) en heroïne in de VS (the increasingly serious crisis around the use of opioids and heroin in the US)*
Political or economic events	- *Doorbraak in Brexit-onderhandelingen (breakthrough in Brexit negotiations)* - *King Mohammed VI wil dienstplicht in Marokko weer invoeren (king Mohammed VI wants to reintroduce military service in Morocco)* - *Het aandeel Bpost verloor gisteren nog maar eens 5,39 procent (shares in Bpost fell by 5.39 percent again yesterday)* - *Merkel begint aan regeringsvorming (Merkel begins government formation)* - *Trump schendt zelf nucleair akkoord (Trump violates nuclear agreement himself)*
Organized sports events	- *Het WK voetbal in Rusland (the football world cup in Russia)*
Other	- *De Noord-Koreaanse rakettest van vorige week (North Korea's roket launch test last week)* - *Chinese Huawei stoot iPhone-maker van tweede plaats op wereldwijde smartphonemarkt (Chinese Huawei takes iPhone-manufacturer's second place in global smartphone market)* - *Turkije moet opnieuw focussen op strijd tegen IS (Turkey must focus again on battle against IS)* - *België wil tegen 2020 Land van de Fair Trade worden (Belgium wants to become Country of Fair Trade by 2020* - *Mexicaanse kartels azen op New York als verdeelcentrum (Mexican cartels aim for New York as distribution center)*

Table 9 Examples of UNK events.

manmade crises and disasters are particularly represented. Unlike sports events (e.g. a goal being made during a soccer match), political and economic events are a rather complex category. Events of this type frequently do not refer to concretely delineated events or use fuzzy wording, as in "breakthrough in Brexit negotiations". Interpreting these `Unknown` events is often dependent on recent developments, and they are abstracted to a degree that they belong more to a topic or sphere of interest (such as "political activity") rather than to any discrete, bounded event type.

`Contact` events are another category found to be disorientating. The generic subtype `Contact` is used to classify `Contact` events that can not be brought cleanly under `Meet`, `Broadcast` or `Correspondence`. There are criteria for inclusion in these subtypes, such as that `Meet` events must involve participants meeting face-to-face. In the dataset, the circumstances of such events are often left unclear. In such cases, the event mention is treated as a generic `Contact.Contact` event. In other cases, the lexicalization of the event is generic in itself, such that it is not possible to recognize a concrete event type other than that some sort of communication is taking place (such as in "May

tries to calm down British citizens"). Some examples of `Contact-Contact` events are shown in Table 10.

Examples
- Bart De Wever reageert op uitspraken Michel (Bart De Wever reacts to Michel's statements)
- Paus Franciscus vraagt vergiffenis voor kindermisbruik door Katholieke Kerk (Pope Francis asks forgiveness for child abuse by Catholic Church)
- [...], zegt Erhan Yilmaz ([...], says Erhan Yilmaz)
- Theo Francken (N-VA) bepleit totale stop op illegale migratie (Theo Francken (N-VA) argues for total halt of illegal immigration)
- de ISIS-propaganda over het idyllische leven in het kalifaat (ISIS propaganda about idyllic life in the caliphate)
- ernstige onderhandelingen over Brexit (serious Brexit negotiations)
- Kim Jong-un onthult zijn grote doel (Kim Jong-un reveals his great goal)
- Charles Michel kritisch voor eredoctoraat Ken Loach (Charles Michel critical about honorary doctorate for Ken Loach)
- Volgens het Russisch ministerie van Defensie is dat ander land Groot-Brittannië (according to the Russian Ministry of Defence, that other country is Great Britain)
- May probeert Britten gerust te stellen (May tries to calm down British citizens)

Table 10 Examples of Contact events.

Our treatment of `Unknown` and `Contact` events reveals a source of friction in the annotation process. When a fixed typology is used, it is natural to mark all event mentions that fall under each type, regardless of its scale or lexicalization. EventDNA annotation used a number of generic types, `Unknown` and `Contact.Contact` being the most prominent, that allow for the annotation of events that do not neatly fall under a type. In these cases, relevance and lexicalization become important considerations. An event must be relevant enough to include, such that it is reasonable to assume the event would be brought up in different articles; and it must be lexicalized or described clearly enough to be understood as a real and concrete event. There are therefore conflicting criteria for the inclusion of event mentions.

One of the research goals of this annotation was to gauge the applicability of pre-existing event typologies to a unrestricted news data context. The coverage and number of the `Unknown` type and the other newly-introduced generic types skew the type distribution of the data in such a way that is not a semantically meaningful or computationally useful feature of the event.

The distribution of the remaining attributes over all events is given in Table 11.

Events are overwhelmingly `Asserted` (91%) and `Positive` (96%), as can be expected from news text which mainly reports on facts. Most events that are expressed as a verbal construction are in the `Present` (32%) or `Past` (30%) tense, with a minority in the `Future` tense (6%). 32% of events have `Unspecified` as tense, which indirectly indicates they are lexicalized as noun phrases. Linking this latter attribute back to *Prominence*, we interestingly found that the distribution in tense differs for `Main` and `Background` events.

Attribute	Value	Count	% of total
Assertion	`Asserted`	6,724	90.78
	`Other`	683	9.22
Negation	`Positive`	7,096	95.80
	`Negative`	311	4.20
Tense	`Present`	2,377	32.09
	`Unspecified`	2,319	31.31
	`Past`	2,243	30.28
	`Future`	468	6.32

Table 11 Distribution of the attribute values for *Assertion*, *Negation* and *Tense* over the event annotations

49% of `Background` events are of `Unspecified` tense, against 18% of `Main` events. On the other hand, 43% of `Main` events are in the `Present` against 17% of `Background` events. During the annotation process, it was noticeable that events in the `Present` usually occur in the titles of articles, as they often use a particular telegraphic writing style.

5.3 IPTC Media Topic annotations

Each document in the corpus is annotated with any number of IPTC Media Topics. As mentioned in section 3.2.3, the IPTC Media Topics taxonomy defines 17 top-level topics that branch into more granular categories. The labels in the document are as specific as possible and can be traced back to more general labels. Additionally, each label is marked as *Certain* or *Uncertain* if there is any possible doubt about its applicability. The corpus contains 4,327 topic annotations in total, which makes 2.44 on average per document.

We derived the top-level IPTC topic from each *Certain* topic annotation in each document (discarding duplicates) and, for each topic, counted the number of documents in which it occurs. Figure 3 shows the results. As mentioned earlier, for the selection of the articles in the corpus, we chose to only select articles which we intuitively considered as hard news. Evidently, the counts mirror our selection of hard news articles from the historical dataset: `Politics` is strongly represented as a topic, followed by `Crime, Economy, Society` and `Conflict` in a smoothly curving trend. However, as the annotators were asked to annotate all relevant IPTC topics for a given article, this sometimes also resulted in the annotation of soft topics next to the hard topics. For instance a given article could be annotated with the soft topic `weather`, but also with topics such as `disaster, accident and emergency_incident`.

We have argued that within EventDNA the attribute event type cannot be considered as meaningful. In addition to the mismatch between the strict typology we employed and the natural diversity of event types in unrestricted data, our analysis of `Contact` events reveals some confusion between the lexical clarity and the concrete character of the described events. However, a semantic classification of events is a desirable feature of an event extraction system as a basis to assess the diversity in events across collections of articles. In the

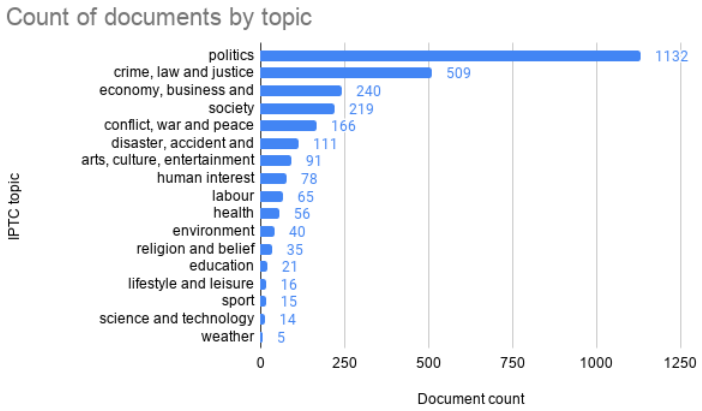

Fig. 3 Number of docs carrying each top-level IPTC topic.

absence of a concrete categorization of events, correlations can be established with IPTC topics. The IPTC topics represent a semantic categorization on the level of articles; we can assume this categorization transfer to its main event(s). In this way, we propose that the widely-applied IPTC topic typology can provide a useful bridge to describe the semantic content of events.

A hypothetical approach to event typology via IPTC topics might be attractive in another way. As we have seen, the ERE-style typology we applied suffered particularly in the case of `Contact` events. The existence of this event type is rooted in the assumption that events are categorized based on their concrete nature as "acts". Considering this, a declaration of war is a clear act of communication and therefore a `Contact` event. It can be argued that if the goal is to identify events in order to increase the diversity in news offering, this type is not very meaningful. A hypothetical event typology via IPTC topics can avoid this, as a declaration of war would first and foremost be recognized as belonging to the broader topic of `Armed conflicts`.

6 Experiments

Pilot experiments were conducted on the annotated corpus, outlined in the previous section, with the aim to identify event mention spans and to evaluate the quality of the annotations. Since our description of event triggers as clauses presumes that the trigger includes all information necessary to understanding the content of the event (e.g. its arguments), we consider *detecting* and *delimiting* the event spans to be the scope of the experiments outlined in this work. In the next sections the architecture of our approach is outlined, followed by a description of the extracted features.

6.1 Architecture

The event detection task is framed as a sequence labeling problem using Conditional Random Fields (CRF), a class of probabilistic classifiers that are well suited to label sequential data (Lafferty et al, 2001). Given an input sequence X, where each item is represented as a bundle of features, the CRF predicts a sequence of target labels Y. The target sequence is given in *IOB* format: tokens which *begin* an event mention are labelled as B, tokens *inside* the mention are I, and tokens *outside* the mention are O.

Based on our observations in Section 5.2 that a sentence on average contains 1.3 events, we chose to consider one event per sentence for the experiments. When multiple events occur in a sentence, the longest event was picked and the rest discarded from the dataset. The `CRFsuite` package[7] was used to implement the CRF learner. The `sklearn-crfsuite` wrapper[8] provided bindings for `CRFsuite` in Python. The CRF model was trained for ten iterations.

To evaluate CRF performance, we performed 10-fold cross-validation on the 6,937-sentence corpus, with a 90%-10% training-test split. As the main target of our experiments is an evaluation of the quality of the EventDNA data annotations, we wanted to evaluate the complete data set. Hence we chose 10-fold cross-validation, with each fold given a chance to be the held-out test set, instead of splitting the data into one training and held-out test set.

Inspired by previous research on Named Entity Recognition (Van de Kauter et al, 2013) using a sequential labeling approach and event extraction (Jacobs et al, 2018) and in order to create a first baseline, a range of lexical and context features were extracted for the pilot experiments:

- **Basic features:**
  - The original form of the token
  - Its lemma or dictionary form
  - Its position in the sentence as an integer index
  - The last 2 characters of the token
  - The last 3 characters of the token
  - A binary feature indicating whether it is the first token in the sentence
  - A binary feature indicating whether it is the last token in the sentence
- **Word shape information** as binary features:
  - Is the token capitalized?
  - Does the token only consist of lower-case characters?
  - Does the token only consist of upper-case characters?
  - Does the token contain any upper-case characters?
  - Does the token consist only of alphabetic characters?
  - Does the token consist only of digits?
- **Syntactic information** extracted using the LeTs toolkit (Van de Kauter et al, 2013):
  - Reduced part of speech information (e.g. N for a noun)

[7] `http://www.chokkan.org/software/crfsuite/`

[8] `https://sklearn-crfsuite.readthedocs.io/en/latest/index.html`

Label	Precision	Recall	F1
B	0.67	0.58	0.62
I	0.72	0.74	0.72
O	0.65	0.64	0.64
Avg.	0.68	0.65	0.67

Table 12 Precision, recall and F1-scores over *IOB* labels (at the word level), macro-averaged over all folds.

Label	Precision	Recall	F1
Gold event	0.78	0.53	0.63
No event	0.58	0.82	0.68
Avg.	0.68	0.67	0.66

Table 13 Precision, recall and F1-scores over extracted event mentions.

  - Full part of speech information (e.g. *N(soort,ev,basis,zijd,stan)*)
  - Reduced chunk information (one of *I, O, B*)
  - Full chunk information, which includes syntactic category (e.g. *B-NP* meaning *token begins a noun phrase*)
  - Full named entity information (e.g. *B-PER* meaning *begins a PER entity)*
  - Named entity type (e.g. *PER*; the label *O* is used for tokens that are not part of a named entity)
- **Context information:** the same sets of features for the previous and next token. Tokens at the beginning or end of the sentence are given a *BOS* or *EOS* feature instead.

6.2 Evaluation and results

Evaluation took place in two settings: once in terms of *IOB* labeling at the word level and once in terms of event mention recall over sentences: correctly predicted event sequences evidently also imply a complete match between predicted and gold *IOB* labels. In the first setting, precision, recall and F1-score where measured for each label (*I, O, B*) compared to the gold sequence. Table 12 shows the results of this evaluation. Scoring was done using methods provided by `scikit-learn`, a popular machine learning toolkit for Python.[9] In this dataset, the *I* label represents the majority of labels with about 54K instances. There are about 5K instances of the *B* label and 46K of the *O* label. The results follow this slight skew. *I* labels were predicted with 0.72 F1-score and *O* labels with 0.64 F1-score. As our goal is to extract those sequences in text as a basis for further event classification, our main interest lies in the B and primarily I scores.

Although the scores at the word level provided in Table 12 give insight in whether the system was able to extract potential event sequences to a certain

[9] `https://scikit-learn.org/stable/`

extent, they do not show whether the system correctly detects event mentions. Therefore, in a second setting, event mentions were also extracted out of the raw sentence context, such that each gold (annotated) and predicted event sequence is represented as a "bag of events" (where an event is encoded as a set of indices over the source sentence). The predicted event mentions were then compared to the gold event mentions . On top of that they were matched using the same syntactic event matching function described in the inter-annotator agreement study (Section 4). A perfect syntactic match and match between predicted and annotated IOB labels indicates a correctly recognized event. This evaluation answers the question *whether the system detects the presence and absence of an event.* If the match is positive – i.e. if the system correctly predicts the gold mention – this is counted as a *true positive*. If the system predicts an event that does not match the gold standard, this is counted as a *false negative*. False negatives also occur when there is a gold event but no mention was predicted. If there is no gold event but the system predicted one, this is counted as a *false positive*. Finally, if neither the gold standard nor the prediction contain events, this is a *true negative*. The recall, precision and F1 score results of this evaluation are shown in Table 13. Event mentions (*Gold Event*) were correctly predicted with a fairly low recall of 0.53, but a high precision of 0.78. Instances without gold event (*No Event*) were predicted with 0.82 recall, but suffer from low precision (0.58). Similar to event *IOB* label prediction, performances show a tendency towards a prediction of the most frequent (54K instances) *I* (*inside event*) *IOB* label of which a predicted *Gold Event* is mainly constituted.

Future work in this evaluation goes in four directions. First, in order to decrease the bias towards *Gold Events* predictions, we will add the raw sentences of the original EventDNA corpus to the training data, that have not been assigned an event type class label during the annotation process and consist completely of *O* (*outside event*) *IOB* labels. Second, for the detection of event spans, training parameters will be tuned to focus on recall more than on precision. Third, a syntactic feature set that is richer than the features generated with the LeTs toolkit (Van de Kauter et al, 2013) will be generated. To that end, we will use Alpino, a Dutch dependency tagger (van Noord, 2006) in order to determine for a given span whether it is an event or not.

Semantic and discourse features will be incorporated to determine whether a given syntactic sequence can be considered as a news event and also to distinguish between main and background events in an event classification task. Finally, work will be done to predict event features such as arguments and co-reference links.

7 Conclusions and future work

In this paper, we described our efforts to design an event corpus and baseline event extraction system for incoming Dutch news text, the design of which was motivated by the NewsDNA use case. The NewsDNA project aims to design

a novel news recommendation algorithm that prioritizes the diversity of news offering over its similarity to the user's previous reading behavior. Identifying the news events that are brought up in incoming news articles and linking them across articles could allow us to perform fine-grained analyses of the diversity within a collection of articles, and retrieve articles for recommendation so as to maximize the diversity of topics the user is exposed to.

In this context, we present the EventDNA corpus, a dataset of 1,773 news articles stripped to their title and lead paragraph, annotated with entities, events, coreference links and IPTC Media Topic codes. The annotated entities are marked as persons, locations, organizations or as belonging to a miscellaneous category. Having the NewsDNA use case in mind, events are marked as being the main news event of an article or a background news event. Further, we investigated whether it is feasible to apply a strict typology of events to unrestricted news data where all relevant events must be captured. To this end, we applied a typology adapted from the ERE framework (Aguilar et al, 2014) and introduced a special `Unknown` type to mark relevant events outside the typology. based on the observation that over a third of all annotated events were marked as `Unknown`, we concluded that the applied typology is not sufficient to cover all events relevant to our task. We therefore suggest that the IPTC Media Topics in the corpus can form a bridge to this kind of classification. IPTC Topics are a standardized taxonomy of news topics, comprising 17 top-level topics (e.g. `Crime, law and justice, Politics` or `Education`) that are divided in increasingly granular sub-topics (e.g. `Law enforcement, Election` or `Higher education`). Each article in the corpus is annotated with any number of relevant topics up to the most specific level of granularity possible. In this way, a semantic categorization is performed on the level of articles. If we assume the main events of articles reflect the article-level IPTC topics, these topic annotations could effectively transfer to events.

We performed pilot experiments on the corpus. A Conditional Random Field learner was trained to identify event mentions as *IOB*-formatted sequences over sentences. No event attributes past the span were predicted, and only lexical and syntactic features were used. The CRF achieved a 0.67 F1-score in assigning *IOB* labels to tokens. A second evaluation was performed on the level of events, such that gold and predicted spans were extracted from the sentence context and matched against each other using an original event-matching function. The system predicted the correct event with an F1-score of 0.56.

In future work, we wish to further develop the experiments using the EventDNA corpus and test the overall effectiveness of incorporating event information into a news recommendation algorithm.

References

Adnan MNM, Chowdury MR, Taz I, Ahmed T, Rahman RM (2014) Content based news recommendation system based on fuzzy logic. In: 2014 Inter-

national Conference on Informatics, Electronics Vision (ICIEV), pp 1–6, DOI 10.1109/ICIEV.2014.6850800, iSSN: null

Aguilar J, Beller C, McNamee P, Van Durme B, Strassel S, Song Z, Ellis J (2014) A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In: Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Association for Computational Linguistics, Baltimore, Maryland, USA, pp 45–53, DOI 10.3115/v1/W14-2907, URL https://aclanthology.org/W14-2907

Altuna B, Aranzabe MJ, Díaz de Ilarraza A (2018) Adapting TimeML to Basque: Event Annotation. In: Gelbukh A (ed) Computational Linguistics and Intelligent Text Processing, Springer International Publishing, Lecture Notes in Computer Science, pp 565–577

Araki J, Mitamura T (2018) Open-Domain Event Detection using Distant Supervision. In: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 878–891, URL http://www.aclweb.org/anthology/C18-1075

Arendarenko E, Kakkonen T (2012) Ontology-based information and event extraction for business intelligence. In: International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Springer, pp 89–102

Baiamonte D, Caselli T, Prodanof I (2016) Annotating content zones in news articles. In: Basile P, Corazza A, Cutugno F, Montemagni S, Nissim M, Patti V, Semeraro G, Sprugnoli R (eds) Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016, CEUR-WS.org, CEUR Workshop Proceedings, vol 1749, URL http://ceur-ws.org/Vol-1749/paper6.pdf

Bies A, Song Z, Getman J, Ellis J, Mott J, Strassel S, Palmer M, Mitamura T, Freedman M, Ji H, O'Gorman T (2016) A comparison of event representations in DEFT. In: Proceedings of the Fourth Workshop on Events, Association for Computational Linguistics, San Diego, California, pp 27–36, DOI 10.18653/v1/W16-1004, URL https://aclanthology.org/W16-1004

Bittar A, Amsili P, Denis P, Danlos L (2011) French TimeBank: An ISO-TimeML Annotated Reference Corpus. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, pp 130–134, URL https://www.aclweb.org/anthology/P11-2023

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5:135–146

Borgesius FJZ, Trilling D, Möller J, Bodó B, Vreese CHd, Helberger N (2016) Should we worry about filter bubbles? Internet Policy Review URL https://policyreview.info/articles/analysis/should-we-worry-about-filter-bubbles

Calhoun S, Carletta J, Brenier JM, Mayo N, Jurafsky D, Steedman M, Beaver D (2010) The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Lang Resour Eval 44(4):387–419, DOI 10.1007/s10579-010-9120-1, URL https://doi.org/10.1007/s10579-010-9120-1

Caselli T, Bartalesi Lenzi V, Sprugnoli R, Pianta E, Prodanof I (2011) Annotating Events, Temporal Expressions and Relations in Italian: the It-Timeml Experience for the Ita-TimeBank. In: Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, Portland, Oregon, USA, pp 143–151, URL https://www.aclweb.org/anthology/W11-0418

Colruyt C, De Clercq O, Hoste V (2019a) EventDNA: Annotation Guidelines for Entities and Events in Dutch News Texts (v1.0). Tech. rep., Ghent University

Colruyt C, De Clercq O, Hoste V (2019b) Leveraging syntactic parsing to improve event annotation matching. In: Aggregating and analysing crowdsourced annotations for NLP : Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, Association for Computational Linguistics (ACL), pp 15–23

van Dijk TA (1988) News as discourse. Lawrence Erlbaum Associates Inc.

Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R (2004) The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In: Proceedings of LREC

Frasincar F, Borsje J, Levering L (2009) A semantic web-based approach for building personalized news services. International Journal of E-Business Research (IJEBR) 5(3):35–53

Goud JS, Goel P, Debnath A, Prabhu S, Shrivastava M (2019) A Semantico-Syntactic Approach to Event-Mention Detection and Extraction In Hindi. In: Workshop on Interoperable Semantic Annotation (ISA-15), p 63

Grineva M, Grinev M, Lizorkin D (2009) Extracting key terms from noisy and multitheme documents. In: Proceedings of the 18th international conference on World wide web, pp 661–670

Grishman R (2010) The Impact of Task and Corpus on Event Extraction Systems. Lrec pp 2928–2931

Grishman R, Sundheim B (1996) Message Understanding Conference- 6: A Brief History. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, URL https://www.aclweb.org/anthology/C96-1079

Hasan KS, Ng V (2014) Automatic keyphrase extraction: A survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1262–1273

Im S, You H, Jang H, Nam S, Shin H (2009) KTimeML: Specification of Temporal and Event Expressions in Korean Text. In: Proceedings of the 7th Workshop on Asian Language Resources (ALR7), Association for Computational Linguistics, Suntec, Singapore, pp 115–122, URL https://www.aclweb.org/anthology/W09-3417

Inel O, Aroyo L (2019) Validation methodology for expert-annotated datasets: Event annotation case study. In: 2nd Conference on Language, Data and Knowledge, LDK 2019, Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, p 12, DOI 10.4230/OASIcs.LDK.2019.12, URL https://research.vu.nl/en/publications/validation-methodology-for-expert-annotated-datasets-event-annota

Jacobs G, Lefever E, Hoste V (2018) Economic event detection in company-specific news text. In: Proceedings of the First Workshop on Economics and Natural Language Processing, pp 1–10

Joris G, Colruyt C, Vermeulen J, Vercoutere S, De Grove F, Van Damme K, De Clercq O, Van Hee C, De Marez L, Hoste V, Lievens E, De Pessemier T, Martens L (2020) News diversity and recommendation systems : setting the interdisciplinary scene. In: Friedewald M, Önen M, Lievens E, Krenn S, Fricker S (eds) Privacy and Identity Management. Data for Better Living : AI and Privacy, Springer, vol 576, pp 90–105

Kaur J, Gupta V (2010) Effective approaches for extraction of keywords. International Journal of Computer Science Issues (IJCSI) 7(6):144

Van de Kauter M, Coorman G, Lefever E, Desmet B, Macken L, Hoste V (2013) LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. Computational Linguistics in the Netherlands Journal 3:103–120

Lafferty J, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning 8(June):282–289, DOI 10.1038/nprot.2006.61, URL http://repository.upenn.edu/cis_papers/159/%5Cnhttp://dl.acm.org/citation.cfm?id=655813

Li Q, Ji H, Huang L (2013) Joint Event Extraction via Structured Prediction with Global Features. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) pp 73–82, DOI 10.1021/bi00231a020, URL http://www.aclweb.org/anthology/P13-1008

Liemans R (2019) Gepersonaliseerd nieuws: matchmaker voor online media of journalistiek-ethisch mijnenveld? URL https://www.vn.nl/gepersonaliseerd-nieuws-matchmaker-of-mijnenveld/

Linguistic Data Consortium (2016) Rich ERE Annotation Guidelines Overview V4.2

Liu J, Dolan P, Pedersen ER (2010) Personalized news recommendation based on click behavior. In: Proceedings of the 15th international conference on Intelligent user interfaces, Association for Computing Machinery, Hong Kong, China, IUI '10, pp 31–40, DOI 10.1145/1719970.1719976, URL https://doi.org/10.1145/1719970.1719976

Lopez P, Romary L (2010) Humb: Automatic key term extraction from scientific articles in grobid. In: SemEval 2010 Workshop, pp 4–p

Mihalcea R, Tarau P (2004) Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 404–411

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781

Minard AL, Speranza M, Urizar R, van Erp M, Schoen A, van Son C (2016) MEANTIME, the NewsReader Multilingual Event and Time Corpus. In: Proceedings of the 10th language resources and evaluation conference (LREC 2016), European Language Resources Association (ELRA), Portorož, Slovenia, p 6

Mitamura T, Liu Z, Hovy E (2015a) Overview of TAC KBP 2015 Event Nugget Track. Kbp Tac 2015 pp 1–31

Mitamura T, Yamakawa Y, Holm S, Song Z, Bies A, Kulick S, Strassel S (2015b) Event Nugget Annotation: Processes and Issues. In: Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Association for Computational Linguistics, Denver, Colorado, pp 66–76, DOI 10.3115/v1/W15-0809, URL http://aclweb.org/anthology/W15-0809

Mitamura T, Liu Z, Hovy E (2016) Overview of TAC-KBP 2016 Event Nugget Track. Tac Kbp 2016

MUC (2001) MUC 7 Proceedings. URL https://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html

Nguyen CQ, Phan TT (2009) An ontology-based approach for key phrase extraction. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp 181–184

Nguyen TH, Grishman R (2015) Event detection and domain adaptation with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp 365–371

van Noord G (2006) At last parsing is now operational. In: Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées, ATALA, Leuven, Belgique, pp 20–42, URL https://aclanthology.org/2006.jeptalnrecital-invite.2

Nugent T, Petroni F, Raman N, Carstens L, Leidner JL (2017) A comparison of classification models for natural disaster and critical event detection from news. In: 2017 IEEE International Conference on Big Data (Big Data), IEEE, pp 3750–3759

O'Gorman T, Wright-Bettner K, Palmer M (2016) Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016) pp 47–56, DOI 10.18653/v1/W16-5706, URL http://aclweb.org/anthology/W16-5706

Pariser E (2011) The filter bubble: How the new personalized web is changing what we read and how we think. Penguin

Patterson T (2000) Doing well and doing good. SSRN Electronic Journal DOI 10.2139/ssrn.257395

Peng H, Chang KW, Roth D (2015) A Joint Framework for Coreference Resolution and Mention Head Detection. In: Proceedings of the Nine-

teenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Beijing, China, pp 12–21, DOI 10.18653/v1/K15-1002, URL http://aclweb.org/anthology/K15-1002

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Pustejovsky J, Castano J, Ingria R, Saurı R, Gaizauskas R, Setzer A, Katz G (2003a) TimeML: Robust Specification of Event and Temporal Expressions in Text. New Directions in Question Answering 3:28–34

Pustejovsky J, Castaño JM, Ingria R, Saurí R, Gaizauskas RJ, Setzer A, Katz G, Radev DR (2003b) Timeml: Robust specification of event and temporal expressions in text. In: New Directions in Question Answering

Pustejovsky J, Lee K, Bunt H, Romary L (2010) ISO-TimeML: An International Standard for Semantic Annotation. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Languages Resources Association (ELRA), Valletta, Malta, URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/55_Paper.pdf

Ruppenhofer J, Ellsworth M, Schwarzer-Petruck M, Johnson CR, Scheffczyk J (2016) Framenet ii: Extended theory and practice. Tech. rep., International Computer Science Institute

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, Association for Computing Machinery, Hong Kong, Hong Kong, WWW '01, pp 285–295, DOI 10.1145/371920.372071, URL https://doi.org/10.1145/371920.372071

Schouten K, Ruijgrok P, Borsje J, Frasincar F, Levering L, Hogenboom F (2010) A semantic web-based approach for personalizing news. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp 854–861

Schuurman I, Hoste V, Monachesi P (2010) Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/162_Paper.pdf

Shoemaker PJ (2006) News and newsworthiness: A commentary. Communications 31(1):105–111, DOI 10.1515/commun.2006.007

Simonnet E, Ghannay S, Camelin N, Estève Y, De Mori R (2017) Asr error management for improving spoken language understanding. arXiv preprint arXiv:170509515

Song Z, Bies A, Strassel S, Riese T, Mott J, Ellis J, Wright J, Kulick S, Ryant N, Ma X (2015) From Light to Rich ERE: Annotation of Entities, Relations, and Events. In: Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015, ACL, Denver, Colorado, pp 89–98

Sridhar KVR, Nenkova A, Narayanan S, Jurafsky D (2008) Detecting prominence in conversational speech: Pitch accent, givenness and focus. Proceedings of the 4th International Conference on Speech Prosody, SP 2008

Thurman N, Schifferes S (2012) The Future of Personalization at News Websites. Journalism Studies 13(5-6):775–790, DOI 10.1080/1461670X.2012.664341, URL https://doi.org/10.1080/1461670X.2012.664341

Thurman N, Moeller J, Helberger N, Trilling D (2019) My Friends, Editors, Algorithms, and I. Digital Journalism 7(4):447–469, DOI 10.1080/21670811.2018.1493936, URL https://doi.org/10.1080/21670811.2018.1493936

Valenzuela-Escárcega MA, Hahn-Powell G, Surdeanu M, Hicks T (2015) A domain-independent rule-based framework for event extraction. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations, pp 127–132

Vossen P (2018) NewsReader at SemEval-2018 task 5: Counting events by reasoning over event-centric-knowledge-graphs. In: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, pp 660–666, DOI 10.18653/v1/S18-1108, URL https://aclanthology.org/S18-1108

Vossen P, Agerri R, Aldabe I, Cybulska A, van Erp M, Fokkens A, Laparra E, Minard AL, Palmero Aprosio A, Rigau G, Rospocher M, Segers R (2016) NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. Knowledge-Based Systems DOI 10.1016/j.knosys.2016.07.013

Walker C, Strassel S, Medero J, Maeda K (2006) Ace 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia 57:45

Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG (2005) Kea: Practical automated keyphrase extraction. In: Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, IGI global, pp 129–152

Yaghoobzadeh Y, Ghassem-sani G, Mirroshandel SA, Eshaghzadeh M (2012) ISO-TimeML event extraction in Persian text. In: Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India, pp 2931–2944, URL https://aclanthology.org/C12-1179

Yang B, Mitchell TM (2016) Joint extraction of events and entities within a document context. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, pp 289–299, DOI 10.18653/v1/N16-1033, URL https://aclanthology.org/N16-1033

Yih Wt, Goodman J, Carvalho VR (2006) Finding advertising keywords on web pages. In: Proceedings of the 15th international conference on World Wide Web, pp 213–222

Yimam SM, Biemann C, Eckart de Castilho R, Gurevych I (2014) Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland, pp 91–96, DOI 10.3115/v1/P14-5016, URL https://www.aclweb.org/anthology/P14-5016

Zhang C (2008) Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems 4(3):1169–1180

A ERE and eventdna types

ERE (v3.0)		EventDNA (v1.0)	
Types	Subtypes	Types	Subtypes
Life	BeBorn Marry Divorce Injure Die	Life	BeBorn Marry Divorce Injure Die
Movement	TransportPerson TransportArtifact	Movement	TransportPerson TransportArtifact
Transaction	TransferOwnership TransferMoney Transaction	Transaction	TransferOwnership TransferMoney Transaction
Business	StartOrg MergeOrg DeclareBankruptcy EndOrg	Business	StartOrg MergeOrg EndOrg
Conflict	Attack Demonstrate	Conflict	Attack Demonstrate Conflict
Contact	Meet Correspondence Broadcast Contact	Contact	Meet Correspondence Broadcast Contact
Personell	StartPosition EndPosition Nominate Elect	Personell	StartPosition EndPosition Elect
Justice	ArrestJail ReleaseParole TrialHearing ChargeIndict Sue Convict Sentence Fine Execute Extradite Acquit Appeal Pardon	Justice	ArrestJail ReleaseParole TrialHearing ChargeIndict Sue Convict Sentence Fine Execute Extradite Acquit Appeal Pardon Justice
Manufacture	Artifact	Manufacture	Artifact
		Journalism	Publication Investigation
		Politics	Vote
		Unknown	Unknown
9	38	12	41

Table 14 Changes in event types and subtypes between Rich ERE and EventDNA.