Towards Building a Knowledge Base of Monetary Transactions from a News Collection

Jan R. Benetka
Norwegian University of Science and Technology
benetka@idi.ntnu.no

Krisztian Balog
University of Stavanger
krisztian.balog@uis.no

Kjetil Nørvåg
Norwegian University of Science and Technology
kjetil.norvag@idi.ntnu.no

***Abstract*—We address the problem of extracting structured representations of economic events from a large corpus of news articles, using a combination of natural language processing and machine learning techniques. The developed techniques allow for semi-automatic population of a financial knowledge base, which, in turn, may be used to support a range of data mining and exploration tasks. The key challenge we face in this domain is that the same event is often reported multiple times, with varying correctness of details. We address this challenge by first collecting all information pertinent to a given event from the entire corpus, then considering all possible representations of the event, and finally, using a supervised learning method, to rank these representations by the associated confidence scores. A main innovative element of our approach is that it jointly extracts and stores all attributes of the event as a single representation (quintuple). Using a purpose-built test set we demonstrate that our supervised learning approach can achieve 25% improvement in F1-score over baseline methods that consider the earliest, the latest or the most frequent reporting of the event.**

I. Introduction

Financial columns are an essential part of every major news portal. Even people who are not employed in the business sector tend to enjoy catchy headlines about acquisitions of start-ups by the big market players. Information about such economic events is partially captured in a (semi-)structured form, for example, in Wikipedia list pages[1] and in domain-specific knowledge bases, like CrunchBase. These resources, however, are typically limited to a particular genre of business entities and to a handful of transaction types (e.g., CrunchBase focuses on startups and considers only investments, funding, and acquisitions as financial relations). Furthermore, the above resources are constructed manually and thus require a continuous editorial work in order to remain up-to-date. News collections contain millions of articles and it is not humanly possible to explore and extract all information about economic events manually. Realistically, only the most prominent transactions with high publicity are likely to be extracted and organized. Yet, it is the whole space of transactions that provides a complete picture about economic entities.

Having a comprehensive knowledge base (KB), which organizes information about monetary transactions in a structured and semantically meaningful way would, therefore, be of great value. This knowledge base could be utilized, among others, for mining and exploring financial entities and trends. Our interest is in developing an automated approach that is able to assist in populating a financial knowledge base with previously unseen events, and with updating existing events, if new facts surface. Imagine, for instance, that a person responsible for updating a KB has a tool similar to the one depicted in Fig 1. By selecting a company (A) and a particular financial relation (B), this tool would mine a predefined set of sources (e.g., The New York Times corpus) and discover companies which are in the given relation with the query company (C). For each of the returned companies, it would identify the attributes of the financial event (D). One challenge KB population faces is due to the very nature of news: they tend to report on the same event multiple times, with slight differences. As we will show later in this paper, simply trusting either the earliest or the latest reporting of the event does not yield the best results. The proposed tool would extract all possible interpretations of the event, as they appear in the text sources and would assign a confidence score to each of these interpretations. This information, together with an interactive preview of the original text (E), would help the editor in deciding which event to include in the KB and which to ignore. Importantly, the editor would update the entire event record with a single click, as opposed to manipulating individual attribute values.

We address the aforementioned challenges by first extracting information from news articles using a natural language processing pipeline. The proposed pipeline is rather typical in terms of its architecture and components, but is tailored specifically to the financial domain. It comprises monetary value recognition, economic event recognition, named entity recognition, date extraction, and semantic role labeling steps.

The main conceptual and technical novelty of the paper lies in that all attributes of an event are extracted jointly and a single structured representation is created from them. We start with grouping all sentences together from the entire corpus that discuss a given event. From these, we generate all possible structured representations of the event, i.e., quintuples comprising subject, predicate, object, monetary value, and date. All elements of the quintuple are extracted collectively, from a single sentence (which serves as provenance). To select a single structured representation (quintuple), we employ a supervised learning approach with a set of innovative features to rank the possible quintuples, and then the one with the

[1] https://en.wikipedia.org/wiki/List_of_mergers_and_acquisitions_by_Alphabet

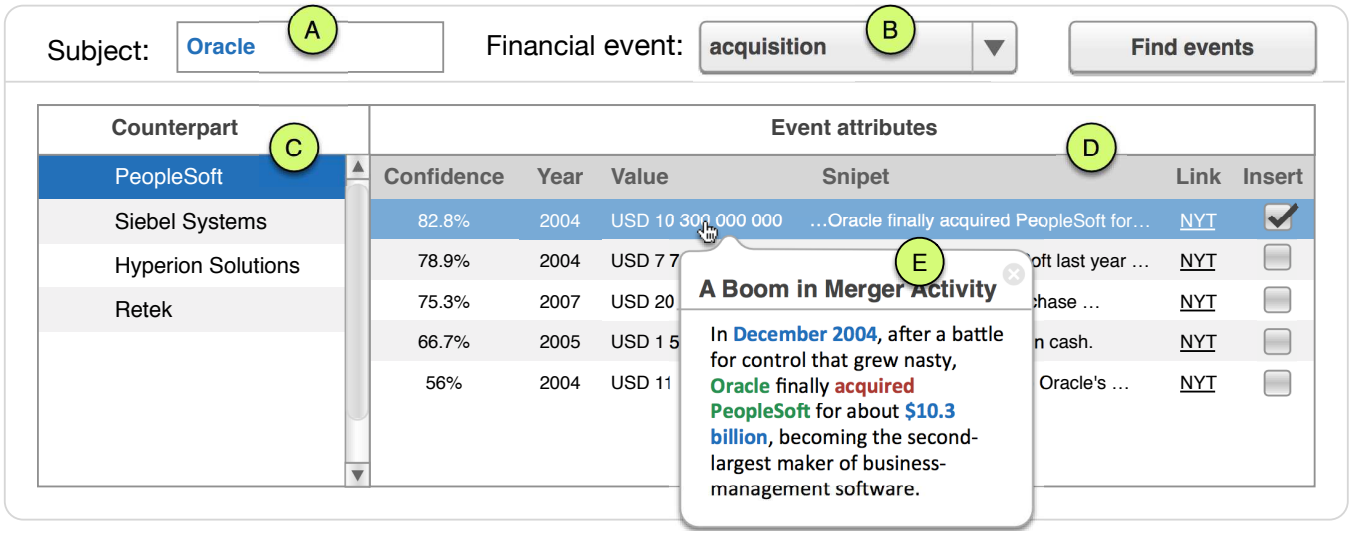

Fig. 1. An interactive system for populating a financial knowledge base from a news corpus. For a selected subject entity (A) and financial relation (B), the system finds object entities (C) and lists the extracted events with attributes and an associated confidence score (D); the origins of the extracted data can be checked (E).

highest confidence score is chosen. Importantly, our approach can also identify when none of the candidate quintuples would serve as an accurate representation of the event, using a confidence threshold. We demonstrate the effectiveness of our method using a purpose-built test collection.

In summary, this work makes the following contributions: (1) we present a natural language processing pipeline tailored to financial information extraction, (2) we develop a supervised learning approach and a rich set of features for ranking representations of an economic event and selecting the best one, (3) we provide a test dataset and evaluation methodology, and (4) we perform an experimental evaluation and offer insights on our methods. All resources developed in this paper are made publicly available at https://github.com/benetka/kbmt.

II. Related Work

The present work lies in the intersection of information extraction, news stream monitoring, financial text mining, and knowledge base population.

Event extraction is a specialized branch of information extraction [26] that has attracted a lot of attention in recent years. Automated extraction techniques play a crucial role in aiding humans in knowledge-intensive activities in various domains, including global crisis monitoring [30], and algorithmic trading [22]. The main approaches and implementations of event extraction from text are well summarized in [13]. Our work focuses on the extraction of financial transactions. This is not an entirely unexplored research area. Hogenboom et al. [12] present a semantic-based pipeline for the detection of economic events (SPEED). They utilize a traditional language processing pipeline combined with an ontology of economic concepts extracted from Yahoo! Finance. SPEED is focused solely on the information extraction part; unlike our approach, the authors do not deal with the ambiguity introduced by multiple and possibly conflicting mentions related to one event. Vossen et al. [31] describe the NewsReader project and the design of a system aiming at representing events in news streams in a knowledge graph. NewsReader aligns extracted information on a timeline in a story-telling fashion, which is convenient for visual browsing of the data. The paper offers overall statistics of extracted data, but no evaluation is performed. Given the chronological nature of news, the temporal dimension is a common perspective for event exploration [6]. Strötgen and Gertz [28] combine time and location for deriving events, however, compared to our work, they do not consider multiple reporting of the same event.

Ontologies are a means to formally model knowledge in the form of (hierarchical) classes of concepts and relations between them. Domain-specific ontologies can provide fine-grained conceptualization for a specific field of knowledge, e.g., biology [2], music [25], or law [24]. Concerning the financial domain, several ontologies have been proposed. The Resource Event Agent (REA) ontology, based on the model developed by [17], represents economic events in an organization from an accounting perspective. This model was further analyzed from the ontological perspective using Sowa's conceptual terminology [27] and is widely used since, either in its core form or in extended versions. The Timely Ontologies for Business Relations (TOB) framework [33] focuses on business relations and extends the well-known YAGO ontology [29] with a means to represent underspecified time intervals. This feature allows for temporal relation inference. A pattern-based approach for financial relation extraction from Wikipedia infoboxes is also presented and evaluated in this work. Finally, there are

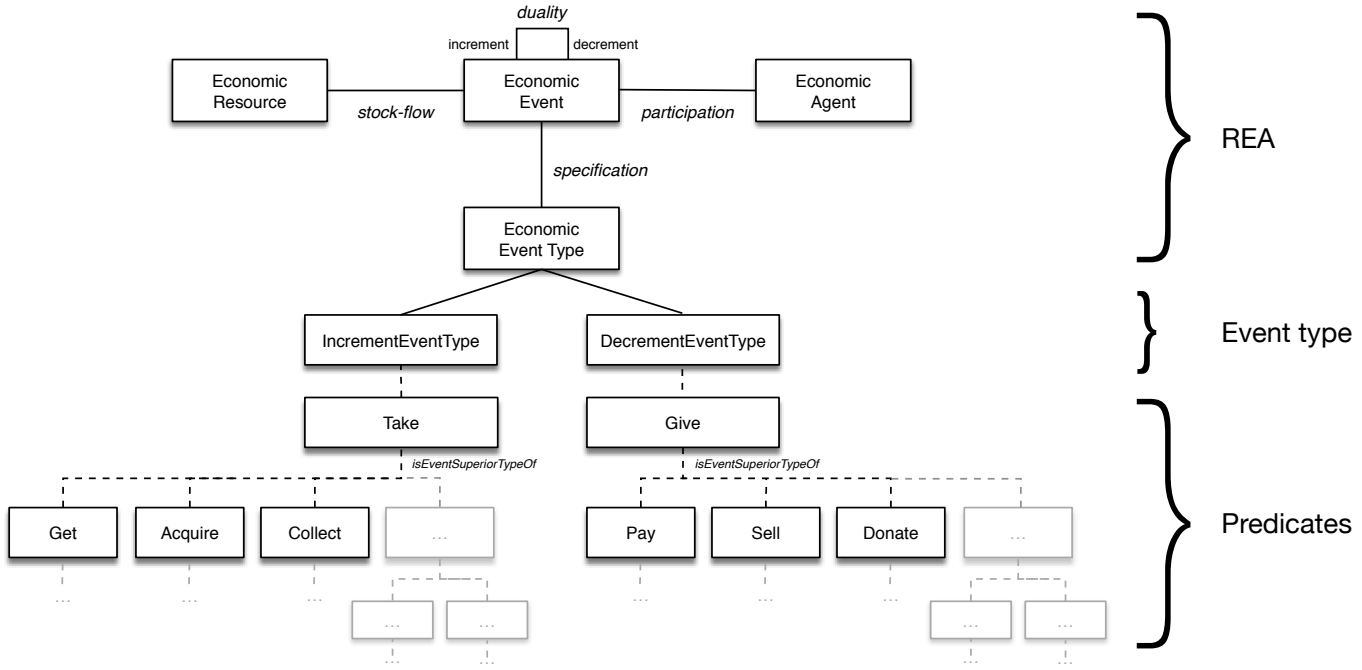

Fig. 2. Overview of the Ontology of Economic Events (OEE). Note that the bottom part shows only an excerpt from the instances.

ongoing efforts towards standardization of financial reporting. XBRL (eXtensible Business Reporting Language) is a markup language, with provided taxonomies, that is nowadays widely used by some of the world's largest economies [21].

Knowledge bases, containing rich semantic knowledge about entities, their properties, and relationships, have become great assets for many applications, including semantic search [18] and business intelligence [32]. Knowledge base construction and maintenance have been of increasing interest in both academia and industry, see, e.g., [8, 11, 15, 4, 3]. General-purpose knowledge bases, such as DBpedia [16], Freebase, or YAGO [29], cover thousands of business entities; however, they contain limited information regarding financial transactions. CrunchBase[2] is a publicly accessible knowledge base containing comprehensive information about startup companies. At the time of writing, it contains about 650K profiles of people and companies. Information about financial transactions, which are also part of the data set, are restricted to investments, funding, and acquisitions. Neither of the aforementioned knowledge bases contain provenance information; our approach supplements each extracted record with provenance data.

III. Representing Economic Events

Before we proceed with the description of our extraction pipeline, we present the ontology we developed for conceptual organization of economic events.

Our starting point is the REA (Resource, Event, Agent) model [17] that is often used as a foundational model for describing business-related concepts; it is briefly introduced in Sect. III-A. To be able to capture more fine-grained semantic distinctions about financial transactions, we extend REA with a hierarchy of economic event types in Sect. III-B.

A. REA

REA has emerged from a framework for accounting systems to one of the standard models in the business domain. The main concepts of this model are *resources* (e.g., services or money), *events* (e.g., transactions), and *agents* (e.g., companies or people). Economic events are processes, where economic resources are changing their owners. It is assumed that there are always two events in a business activity. One which increases the value of the agent's resources and another, which, in turn, decreases value of another resource belonging to the agent.

B. Ontology of Economic Events

In the scope of this project, we deal with a broad spectrum of economic events (i.e., *predicates*) with fine semantic distinctions (e.g., profit-gross). At the same time, we aim to organize economic events in a hierarchical manner (e.g., get $\rightarrow$ earn $\rightarrow$ profit-gross); subsequent processes can then choose the granularity with which they want the information to be processed. Currently, there is no ontology available that would allow for such detailed representation of financial activities. To fill this gap, we propose the Ontology of Economic Events

[2]https://www.crunchbase.com/

(OEE), an extension to REA; see Fig. 2 for a graphical overview. OEE is created using a semi-supervised method that starts with a set of seed verbs and then expands them using the WordNet lexical ontology [20].

The main class of our economic events ontology is called *EventType*. Following Hruby [14], we differentiate between two major economic event types: events increasing and decreasing the value of agent's resources. These sub-classes are called *IncrementEventType* and *DecrementEventType*, respectively. We populate these two classes with predicates that represent specific economic events, organized in a hierarchical fashion, using the following procedure.

1) Select a set of *seed verbs* that are frequently used in a finance-related context. We construct this set by extracting verbs (automatically) from all sentences in our corpus that contain a monetary value; then, we select (manually) the most common verbs as predicates that describe a financial transaction (e.g., buy, sell, invest).
2) For each seed verb:
   a) Create an instance of the verb in the ontology.
   b) Find hypernyms (more general words) of the verb in WordNet; these are added as predicates with a parent-child relation to the verb.
   c) Find adjacent terms (word sharing the same hypernym) of the verb in WordNet; these are also added as predicates and linked to the same parent hypernym by a parent-child relation.
3) Manually revise the placement of verbs.

This process has led to a hierarchy of 50 most common business-related verbs, organized into 5 levels (see the bottom layer on Fig. 2).[3]

Example Consider the following financial statement: *Apple acquires Beats for $3.2 billion.* This information is represented in OEE as three triples:

(Agent) `Apple`	`participates`	(Event) `EventID_1`
(Event) `EventID_1`	`isClassified`	(IncrementEventType) `acquire`
(Event) `EventID_1`	`inflow`	(Resource) `Beats`

In Sect. VIII-A we evaluate the coverage of our ontology using a large news corpus and present further analysis on the usage of predicates in this collection. OEE is made publicly available in OWL format.

IV. Extracting Economic Events

Our goal is to extract structured information about economic events from unstructured text (in our case, a large news archive). An economic event, as understood in this work, is an unambiguous quintuple:

$$(< subject >, < predicate >, < object >, monetary\_value, date),$$

where subjects and objects are unique entity identifiers, predicates come from a purpose-built ontology of monetary transactions, and monetary values and dates are normalized literal values. Our extraction process consists of several steps, organized in a pipeline architecture, as shown in Fig. 3. We deal with the first two steps, semantic annotations (Step 1) and event identification (Step 2), in this section. Steps 3 and 4 are presented in Sect. V.

A. Semantic Annotations

The first step of our pipeline is responsible for the semantic annotation of text using natural language processing techniques: recognizing financial events, entities, monetary values, and dates. We operate on the sentence level; sentences serve as provenance information for the extracted information. Another pragmatic reason for using sentences is that they can be presented as short summaries on the user interface, as it is shown in Fig. 1 (E). We generate annotations in a sequential order; sentences missing the required piece of information (i.e., monetary value, financial event, or entities) are excluded from subsequent processing steps. Some of the components in the pipeline have multiple possible configurations; these are evaluated in Sect. VII-A.

(a) **Monetary Value Recognition**
Each sentence is tested on the presence of monetary value. We define monetary value as a tuple consisting of a numerical value and a currency identifier (e.g., '€1000' or 'two billion US dollars'). A grammar capable of recognition of both verbal and nominal forms of numbers, extended with a list of currency names and symbols, is used for annotation. Beyond recognition, value and currency normalization are also performed in this step using an extended Numbers Tagger[4] in GATE [10].

(b) **Event Recognition**
To identify financial transactions in text, we use predicates from a purpose-built ontology of monetary transactions that we constructed in a semi-supervised manner (see Sect. III) which starts with a set of seed verbs and then expands them using the WordNet lexical ontology [20]. These predicates are used to create a gazetteer for a predicate tagger that, by default, labels verbs. For each of the predicates, the corresponding semantic frame set from PropBank [23] is extracted. The frame set contains specifications of arguments, referred to as role sets, for possible meanings of the predicate. The specific meaning of each predicate (i.e., role set) is determined later in our annotation pipeline, in the Semantic Role Labeling step. Further, we extend our annotator with the possibility of recognizing noun predicates as well (e.g., 'acquisition of').

[3]We wish to point out that predicates from all levels of the hierarchy may be used, not only the leaf nodes. Obviously, more specific predicates should be preferred over less specific ones.

[4]https://gate.ac.uk/sale/tao/splitch23.html#sec:misc-creole:numbers:numbers

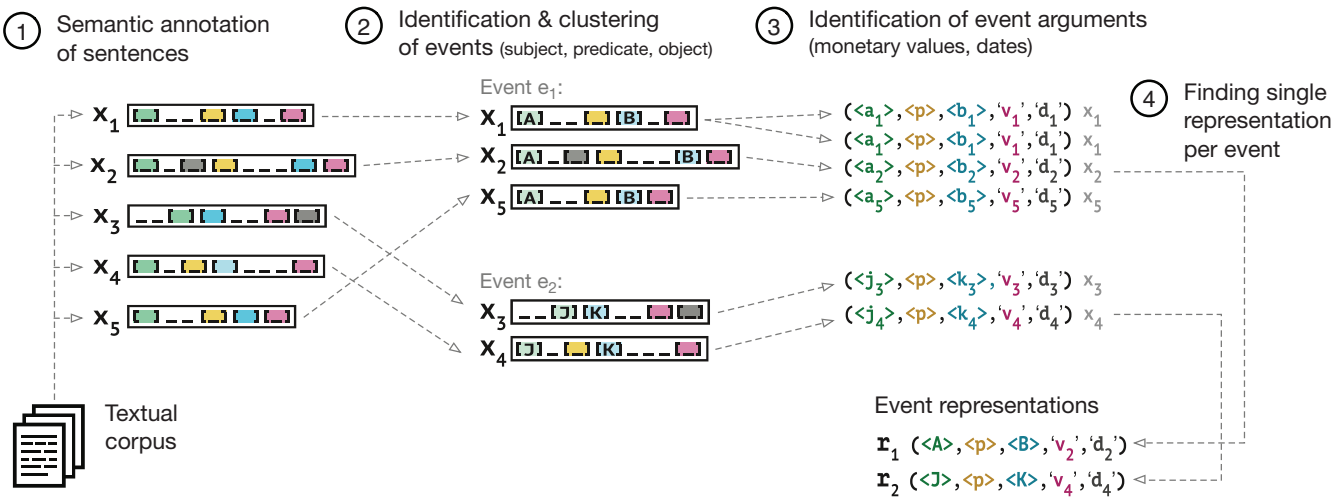

Fig. 3. Economic event extraction pipeline.

We do so by leveraging the NomBank dataset [19]; each noun in NomBank, provided it originates from a verb, contains an identifier of its source (verb counterpart) in PropBank.

(c) **Entity Recognition**
After having a monetary value and an economic event identified in the sentence, the next step is to recognize the participants of the financial transaction. Since we are only interested in economic subjects, such as companies and organizations, we use a repository of entities assembled from multiple knowledge bases, specifically, DBpedia, Freebase, and CrunchBase; we refer to Sect. VI-B for details. We resolve each entity mention to its most common sense, i.e., the entity that is most commonly referred to by that mention. Despite being a naive way of resolving ambiguity, this technique works well in practice [15]. We consider two settings: one where we require entities to have descriptions, i.e., their knowledge base entry is more complete, and another where such requirement is not imposed.

(d) **Date Extraction**
Each economic event is timestamped with at least one date. Absolute (e.g., 6/7/2015) and relative (e.g., 'two days ago') temporal expressions within the sentence are annotated and normalized by the Stanford Temporal Tagger (SUTime) [9]. Note that the sentence might contain multiple explicit temporal expressions, all of which are recorded. We also consider the article's publication date (this is always available).

(e) **Semantic Role Labeling**
Sentences containing all the necessary ingredients mentioned earlier are good candidates for being able to extract a structured representation of the event from them. To verify that all the components (i.e., money, relation verb/noun, and entities) are mutually related, we employ semantic role labeling (SRL), specifically, the system by Björkelund et al. [5]. SRL is capable of recognizing predicates and their semantic arguments. Given, for example, the verb *sell*, it will identify a subject (who is selling), an object (to whom) and possibly several other arguments (price, date, manner, etc.) as well. This final step decides whether the identified components fit a correct semantic pattern. It is a configuration setting in our pipeline whether the correct semantic roles for monetary value and date are enforced.

B. Event Identification

Events, and especially those that involve popular entities, are often reported multiple times in the news media; see Table I for an illustrative example. Information covered by individual sentences may be redundant, may be conflicting, or may develop over time. The second step of our pipeline is responsible for identifying events and clustering sentences that discuss the same economic event.

We use subject, predicate, object triples to uniquely identify events: $e = (s, p, o)$.[5] Subject and object are unique identifiers from the entity repository. Our entity repository is constructed in such way that the mapping from surface forms of the recognized entities to unique identifiers is unambiguous (see Sect. VI-B). We consider predicates equivalent as long as they have a common ancestor on the second level of the hierarchy; we use the same conditions in our experimental evaluation (cf. Sect. VI-D).

V. Creating Structured Representations of Economic Events

To this point, we have recognized economic events in sentences, along with their participants (object and subject) and attributes (monetary value and date). Further, we have

[5]We note that this identifier allows for a single economic event, with the given predicate, between the two companies; this is currently not an issue in our dataset. It could be easily generalized by including the date as well in the event identifier.

TABLE I
ECONOMIC EVENT (GOOGLE ACQUIRES YOUTUBE) EXPRESSED BY MULTIPLE SENTENCES.

Publ. date	Sentence
2006-10-11	Before Google agreed to buy YouTube for $1.65 billion in stock, it paid $1 billion for 5% of AOL...
2007-02-08	Google bought YouTube in October for $1.65 billion.
2007-04-05	YouTube was purchased by Google in November for $1.6 billion.

grouped sentences together that correspond to the same event. What is left for us to do is to create, for each event, a quintuple representing that event. This might appear a straightforward exercise at first sight; however, there might be multiple sentences describing the same event (see Table I for an example). Matters are further complicated by the fact that even in a single sentence there might be multiple financial values or dates, leading to multiple possible interpretations. We approach this problem in two phases. First, we form one or more quintuples representing the event (Sect. V-A). Then, in case there are multiple quintuples, we select one that constitutes the best representation of the event (Sect. V-B). These phases correspond to Steps 3 and 4 in Fig. 3, respectively.

A. Generating Candidate Quintuples

A single sentence might contain multiple financial values and dates. In such cases, a quintuple is generated for each possible combination of attributes. Formally, let e be an event and S_e the set of annotated sentences describing this event. Each sentence $x \in S_e$ has the following information extracted: subject (s), predicate (p), object (o), publication date (d_x), explicit date mentions (D_x), and monetary values (V_x). Note that s, p, and o are the same across all sentences in S_e, because of how sentence grouping works (cf. Sect. IV-B). Further note that D_x might be an empty set, while V_x always has at least one element. Let then R_e denote the set of possible structured representations for event e:

$$R_e = \big\{(s, p, o, v, d) | x \in S_e, v \in V_x, d \in D_x \cup \{d_x\}\big\}$$

For sentences without an explicit date mention, the publication date is used ($d = d_x$); for sentences with one or more dates extracted from the content ($|D_x| \geq 1$), the article's publication date is ignored ($d \in D_x$).

B. Selecting a Single Quintuple

At the end of the processing pipeline, each event e may be represented by a single quintuple. For events with multiple possible representations (i.e., where $|R_e| > 1$) we need a mechanism to select the quintuple $r \in R_e$ that best describes the given economic transaction. We present three baseline methods and a supervised learning approach.

(a) **First reporting of the event**
Our first baseline selects the first reporting of event e:

$$r^* = \underset{d_x \in S_e}{\arg\min} \big\{(s, p, o, v, d) | v \in V_x, d \in D_x \cup \{d_x\}\big\} \quad (1)$$

In case there are multiple financial values and dates present in the sentence with the earliest publication date, they are chosen arbitrarily.

(b) **Last reporting of the event**
One might argue that the most recent report is likely to be the most accurate one. Our second baseline method implements this intuition by considering the last reporting of the event. This goes analogously to the previous case, except that we write $\max$ instead of $\min$ in Eq. (1).

(c) **Most frequent reporting of the event**
Intuitively, an information that is repeated multiple times has a strong potential to hold true. The third baseline selects the most frequent reporting of the event. In the case of a tie, the earliest reporting is selected from the pool of the most frequent reportings.

(d) **Supervised learning approach**
We cast the selection of the best quintuple as a regression task and use a machine learning approach. Specifically, we use the Random Forests algorithm [7], given its robustness and good empirical performance across a wide range of application domains. Our training data comprises a set of instances, $\mathcal{L} = \{(\mathbf{r}_i, y_i)\}$, where $\mathbf{r}_i$ is a feature vector and $y_i = \{0, 1\}$ is a ground truth label corresponding to the quintuple r_i. The learned model is then used to make a prediction $\hat{y} = \varphi(\mathbf{r})$ on an unseen instance $\mathbf{r}$. We select the quintuple with the highest estimated score:

$$r^* = \underset{r \in R_e}{\arg\max}\, \varphi(\mathbf{r})$$

Further, we introduce a confidence threshold γ, and return the quintuple r^* iff $\varphi(\mathbf{r}) \geq \gamma$. This is an important feature of our approach, as it allows events to be ignored if there is a lack of support. Moreover, this parameter can be used to control the performance trade-off between precision and recall to suit specific applications. The value of γ is determined empirically and is set to 0.3. Our feature vector contains a total of 18 features, developed specifically for this task; it includes (i) simple descriptive statistics (sentence and article length), (ii) linguistic features (predicate tense, noun/verb predicate), (iii) semantic features, related to automatic as well as explicit semantic annotations (entity identification, semantic roles, temporal value, article category), and (iv) cross-document features considering global predicate frequency and attributes across all sentences describing the event (dates and values). See Table V for a detailed list.

VI. EXPERIMENTAL SETUP

The problem we address in this paper is the automatic extraction of economic events from unstructured text. This is a restricted and specialized information extraction task for which no standard evaluation resources exist to date. Next, we describe the text corpus that serves as our input data (Sect. VI-A), the entity repository (Sect. VI-B), the test collection (Sect. VI-C), and our evaluation methodology (Sect. VI-D).

TABLE II
ENTRY FOR THE COMPANY SKYPE FROM OUR ENTITY REPOSITORY.

ID:	Skype
Surface forms:	{Skype, Skype Technologies, Skype Limited}
URIs:	{<dbpedia:Skype_Technologies>, <crunchbase:org/skype-technologies> <crunchbase:org/skype>, <freebase:m/026wfg>, <freebase:m/06whf7>}

A. Text Corpus

We use the New York Times Annotated Corpus (NYTC)[6] as our input text collection. This data set contains over 1.8M news articles spanning over 20 years, beginning in 1987. Apart from its volume, a great benefit of the corpus lies in the annotations, both automatically generated and manually assigned, accompanying a subset of the articles. In the scope of this work, we leverage the following annotations as features in our supervised learning step (see Sect. V-B): publication date, (online) descriptors, and word count. We parsed all documents of the NYTC which yielded 2.1M sentences containing a monetary value. Out of these, 383K sentences describe an economic event.

B. Entity Repository

We employ an entity repository that is constructed from three sources: DBpedia, Freebase, and CrunchBase. From DBpedia and Freebase, we only include entities that are of type organization. CrunchBase contains only companies (over 160K), so we consider all of them. Some organization names can be expressed in many ways which makes their identification a non-trivial task (e.g. 'The Times', 'The New York Times', 'NYT'). The above-mentioned knowledge bases, however, only hold the official organization names and their unique identifiers. Therefore, on top of the known surface forms, we generate additional name variants using a set of heuristics, similar to those described in [1]. Finally, we group URIs as well as surface forms together that refer to the same entity and assign a unique identifier to each entity. Our entity repository contains 989K unique entities, 1.35M unique surface forms, and same-as links to 1.24M DBpedia, Freebase, and CrunchBase URIs in total. An example entry is shown in Table II.

C. Test Collection

We created a test collection by capitalizing on what is already available in CrunchBase. We hand-picked 30 target companies from CrunchBase that are known to have participated in financial transactions during the period covered by the NYTC. Importantly, the gold standard we need to compare to is not CrunchBase, but what could potentially be extracted from the text corpus (by a human). Therefore, CrunchBase transactions are checked for their presence in the NYTC; transactions absent from CrunchBase, but covered by the NYTC, are added to the ground truth. For each target company, we extracted all monetary sentences from the NYTC mentioning the given company and manually grouped sentences by events. Then, sentences were individually inspected, and a single supporting sentence was selected for each event; the data regarded as ground truth is extracted from this sentence. In sum, our test data set contains information about investments and acquisitions for 30 companies, 132 events in total.

[6]https://catalog.ldc.upenn.edu/LDC2008T19

D. Evaluation Methodology

We evaluate event extraction as a binary classification task, using standard measures: precision (P), recall (R), and the F1-measure (F1). The first part of the evaluation (Sect. VII-A) is focused on the correct identification of economic events. In order to label an instance as correct, both participating entities as well as the relation type need to correspond with the ground truth. When comparing predicates, we considered them equivalent if they had a common ancestor on the second level of the hierarchy. The second stage of the evaluation (Sect. VII-B) examines the extraction of the attributes of economic events. We consider two settings: (1) *strict*, where the financial value and event date have to match exactly, and (2) *relaxed*, where a certain tolerance is allowed, specifically, only the year part of date is considered and 10% difference in monetary values is allowed.

VII. EVALUATING EVENT EXTRACTION

Evaluation is divided into two main steps: (1) *event identification*, which considers the extent to which subject-predicate-object triples are successfully identified (Sect. VII-A), and (2) *event extraction*, which focuses on the end-to-end task of creating structured representations of economic events, including their attributes (Sect. VII-B).

A. Economic Event Identification

We compare different configurations of our NLP pipeline (in Sect. IV). Specifically, we have control over the following options: (1) whether noun predicates are also included for event recognition (Y) or only verbs are used (N); (2) whether semantic roles are enforced for monetary value and date (Y) or not (N); (3) whether entities are required to have descriptions (Y) or not (N).

Table III presents the results. Due to space constraints we do not include all possible combinations, but on having all options 'off' or 'on' (rows 1 vs. 5), and all but one option 'on' (rows 2–4). We observe that both the addition of noun predicates (NP) and the relaxed treatment of semantic roles (SRL) increase the number of extracted quintuples and events while reducing precision. Accepting only entities with description (ED) has the exact opposite effect. The combination of all three methods ensures results with the highest possible recall, without sacrificing precision too much. We need high recall in downstream processing (where unwanted quintuples can still be filtered out), therefore we use the setting with all options on (row 5) in the remainder of the section.

TABLE III
ECONOMIC EVENT EXTRACTION RESULTS. HIGHEST SCORES ARE IN BOLDFACE.

NP	SRL	ED	#events	#quintuples	P	R	F1
N	N	N	170	268	0.26	0.39	0.31
Y	N	N	185	312	0.24	0.40	0.30
N	Y	N	**316**	**496**	0.16	**0.44**	0.23
N	N	Y	117	194	**0.37**	0.38	**0.38**
Y	Y	Y	217	377	0.23	**0.44**	0.31

TABLE IV
ATTRIBUTE EXTRACTION RESULTS FOR THE BASELINES (BL) VS. OUR SUPERVISED LEARNING APPROACH. HIGHEST SCORES ARE IN BOLDFACE.

Method	Events only			Attr. strict			Attr. relaxed		
	P	R	F1	P	R	F1	P	R	F1
BL/earliest	0.23	**0.44**	0.31	0.18	**0.34**	0.23	0.22	**0.42**	0.29
BL/frequent	0.23	**0.44**	0.31	0.16	0.31	0.21	0.21	0.41	0.28
BL/latest	0.23	**0.44**	0.31	0.16	0.31	0.21	0.21	0.40	0.27
Our approach	**0.51**	0.31	**0.39**	**0.34**	0.20	**0.25**	**0.49**	0.29	**0.36**

B. Economic Event Extraction

The second step of the evaluation focuses on the extraction of event attributes, i.e., financial value and date. We shall point out that the extraction mechanism is the same for all methods, but they differ in how a single structured description for the event (quintuple) is selected (cf. Sect. V-B). The baseline methods always choose the earliest/latest information. The supervised method attempts to learn how to select the best quintuple from annotated data; it can also choose not to return any quintuple for a given event. We use leave-one-out cross-validation, i.e., use all but one company for training and test on the remaining one; this is repeated for all companies in the test set.

Table IV reports the results for extracting events only (columns 2–4) and extracting attributes as well, using both strict and relaxed evaluation (columns 5–10). Focusing on the event extraction part first, we can observe the effectiveness of the filtering mechanism of the supervised learning approach; it doubles precision and improves F1-score by 26% (the baselines correspond to the last row in Table III). Next, when event attributes are also considered, we find again that the supervised learning approach achieves better results in terms of F1-score than any of the baselines. The improvements over the earlier baseline (the better of the two) are 10% in strict mode and 25% in relaxed mode.

VIII. ANALYSIS

This section provides further analysis of the data and of the results. Specifically, we check the coverage of our ontology and the frequency of predicates, measure the importance of individual features, and take a closer look at some successes and failures.

A. Ontology

The type of each economic event is defined by a predicate in the OEE ontology. In order to evaluate the coverage of the ontology, we created a list of the most frequent verbs from the 2.1M sentences of the NYTC with monetary value, manually inspected the top 200 verbs and deemed 81 of them as finance-related. 84% of these finance-related verbs is covered by our ontology. Further, we measured the frequency of the various predicates in the NYTC. Figure 4 shows the predicates ordered by number of occurrences up to the first three levels of OEE. The most frequent second-level predicate, *pay*, is mentioned in over 66K sentences. The average amount of sentences per transaction type from our ontology is $2,339$.

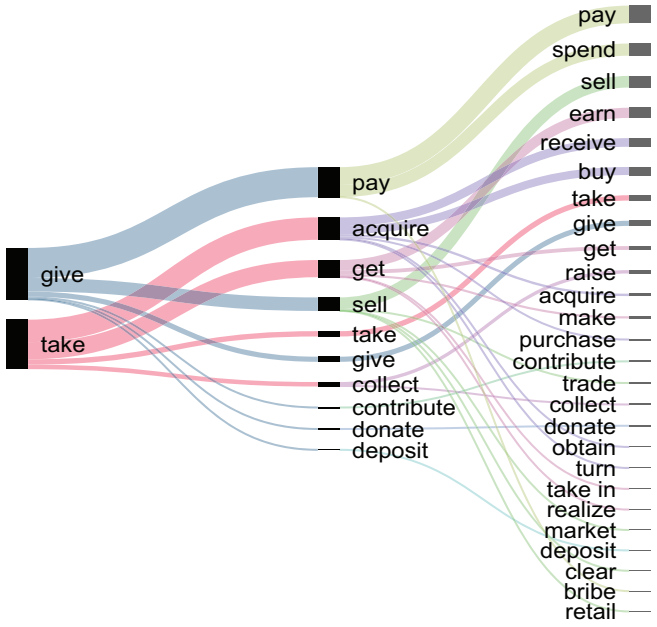

Fig. 4. Predicate frequency in the NYTC.

B. Features

Table V lists our features ordered by their Gini importance. We find that features that consider information from all quintuples for the given event are especially useful (dates_count and values_ratio), and so are global predicate statistics (pred_frequency). The most important linguistic feature is whether monetary values stand in the correct semantic argument (correct_fin_argument); semantic roles seem far less crucial for dates (correct_temp_argument). Article and sentence length are among the strongest features.

C. Successes and Failures

We now take a closer look at cases where our supervised learning approach can really make a difference: events for which multiple structured representations (quintuples) are generated. Our data set contains 24 such events; the number of quintuples for these range from 2 to 17. The results for these events, using relaxed evaluation, are as follows: the earliest baseline fails in 4 cases, the latest baseline fails in 5 cases,

TABLE V
LIST OF OUR FEATURES ALONG WITH THEIR GINI IMPORTANCE.

Feature	Description	Type	Gini
dates_count	Quintuple # with the same date in R_e	numerical	0.1859
article_length	Length of the article	numerical	0.1447
sentence_length	Length of the sentence	numerical	0.1365
sentence_order	Sentence's position within in the article	numerical	0.1291
values_ratio	Relative freq. of the given monetary value in R_e	numerical	0.0883
correct_fin_arg	Fin. value is within the correct semantic arg.	binary	0.0636
pred_frequency	Relative freq. of the predicate in the corpus	numerical	0.0507
predicate_tense	Tense of the predicate	categorical	0.0425
object_has_cb_uri	Object has a CrunchBase URI	binary	0.0405
object_has_dbp_uri	Object has a DBpedia URI	binary	0.0237
nytc_desc_bus	Article is classified under "Business" according to the NYTC taxonomy	binary	0.0232
has_event_date	Temp. expression was found within the sentence	binary	0.0207
correct_temp_arg	Temp. value is within the correct semantic arg.	binary	0.0186
object_has_fb_uri	Object has a Freebase URI	binary	0.0156
is_noun_predicate	Predicate is expressed by a verb or a noun	binary	0.0096
subject_has_dbp_uri	Subject has a DBpedia URI	binary	0.0067
subject_has_cb_uri	Subject has a CrunchBase URI	binary	0.0000
subject_has_fb_uri	Subject has a Freebase URI	binary	0.0000

TABLE VI
EXAMPLE OF A TRANSACTION WITH MULTIPLE QUINTUPLES: ORACLE ACQUIRED PEOPLESOFT.

Subject	Predicate	Object	Monetary value	Year	Published	Returned by method	Correct
Oracle	acquire	PeopleSoft	$7.3 billion	2003	2003-11-25	baseline, earliest	N
Oracle	acquisition	PeopleSoft	$7.7 billion	2004	2004-10-26	-	N
Oracle	acquisition	PeopleSoft	$7.7 billion	2004	2004-10-26	-	N
Oracle	acquire	PeopleSoft	$1.3 billion	2004	2005-12-23	-	N
Oracle	acquire	PeopleSoft	$7.038 billion	2004	2005-12-23	-	N
Oracle	acquire	PeopleSoft	$10.3 billion	2004	2007-03-01	**supervised learning**	**Y**
Oracle	acquisition	PeopleSoft	$10.3 billion	2005	2005-06-30	-	N
Oracle	purchase	PeopleSoft	$20 billion	2007	2007-03-21	baseline, latest	N

while the supervised learning approach was incorrect only in a single case. Table VI shows a specific example, where the same event is reported multiple times. The supervised learning method was able to identify the correct quintuple.

IX. CONCLUSIONS

In this paper, we have addressed the task of extracting economic events from a large news corpus. We have presented a natural language processing pipeline for the semantic annotation of text. To create a single structured representation for each economic event, we have employed a supervised learning approach and have developed a set of innovative features. Using a purpose-built test collection, we have demonstrated that our approach is superior to two intuitive baselines, i.e., earliest and latest published information, and can achieve 25% improvement in F1-score.

Our work represents an important step towards building domain-specific knowledge bases in an automated manner. Even if the system may not have reached the necessary level of performance yet for fully automated operation, it could aid human editors in their tasks by displaying verifiable structured records in a ranked order. The next direction for future work is to consider multiple textual sources, not just a single newspaper.

REFERENCES

[1] R. Ananthanarayanan, V. Chenthamarakshan, P. M. Deshpande, and R. Krishnapuram. Rule based synonyms for entity extraction from noisy text. In *Proc. of AND*, 2008.

[2] M. Ashburner. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[3] K. Balog and H. Ramampiaro. Cumulative citation recommendation: Classification vs. ranking. In *Proc. of SIGIR*, 2013.

[4] K. Balog, H. Ramampiaro, N. Takhirov, and K. Nørvåg. Multi-step classification approaches to cumulative citation recommendation. In *Proc. of OAIR*, 2013.

[5] A. Björkelund, L. Hafdell, and P. Nugues. Multilingual semantic role labeling. In *Proc. of CoNLL*, 2009.

[6] T. Bögel and M. Gertz. Time will tell: Temporal linking of news stories. In *Proc. of JCDL*, 2015.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1): 5–32, Oct. 2001.

[8] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Proc. of AAAI*, 2010.

[9] A. X. Chang and C. D. Manning. SUTime: a library for recognizing and normalizing time expressions. In *Proc.*

of LREC, 2012.

[10] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE*. 2011.

[11] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[12] A. Hogenboom, F. Hogenboom, F. Frasincar, K. Schouten, and O. van der Meer. Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications*, 64(1):27–52, 2013.

[13] F. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong. An overview of event extraction from text. In *Proc. of DeRiVE*, 2011.

[14] P. Hruby. *Model-Driven Design Using Business Patterns*. Springer, 2006.

[15] H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In *Proc. of ACL*, pages 1148–1158, 2011.

[16] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

[17] W. E. McCarthy. The REA accounting model: A generalized framework for accounting systems in a shared data environment. *Accounting Review*, 57(3):554, 1982.

[18] E. Meij, K. Balog, and D. Odijk. Entity linking and retrieval for semantic search. In *Proc. of WSDM*, 2014.

[19] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The NomBank project: An interim report. In *Proc. of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*, 2004.

[20] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[21] N. Müller-Wickop, M. Schultz, and M. Nüttgens. XBRL: impacts, issues and future research directions. In *Proc. of FinanceCom*, 2012.

[22] V. Nuij, V. Milea, F. Hogenboom, F. Frasincar, and U. Kaymak. An automated framework for incorporating news into stock trading strategies. *IEEE TKDE*, 26(4): 823–835, 2014.

[23] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Comp. Linguistics*, 31(1):71–106, 2005.

[24] R. Poli and J. Seibt. *Theory and Applications of Ontology: Philosophical Perspectives*. Springer, 2010.

[25] Y. Raimond, S. A. Abdallah, M. B. Sandler, and F. Giasson. The music ontology. In *Proc. of ISMIR*, 2007.

[26] S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008.

[27] J. F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole, 1999.

[28] J. Strötgen and M. Gertz. Event-centric search and exploration in document collections. In *Proc. of JCDL*, 2012.

[29] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a large ontology from Wikipedia and WordNet. *Web Semantics*, 6(3):203–217, 2008.

[30] H. Tanev, J. Piskorski, and M. Atkinson. Real-time news event extraction for global crisis monitoring. In *Proc. of NLDB*, 2008.

[31] P. Vossen, G. Rigau, L. Serafini, P. Stouten, F. Irving, and W. V. Hage. NewsReader: recording history from daily news streams. In *Proc. of LREC*, 2014.

[32] G. Weikum, S. Bedathur, and R. Schenkel. Temporal knowledge for timely intelligence. In *Proc. of BIRTE*, 2011.

[33] Q. Zhang, F. M. Suchanek, L. Yue, and G. Weikum. TOB: timely ontologies for business relations. In *Proc. of WebDB*, 2008.