Extracting Fine-Grained Economic Events from Business News

Gilles Jacobs **Véronique Hoste**
Language and Translation Technology Team
Ghent University, 9000 Gent, Belgium
firstname.lastname@ugent.be

Abstract

Based on a recently developed fine-grained event extraction dataset for the economic domain, we present in a pilot study for supervised economic event extraction. We investigate how a state-of-the-art model for event extraction performs on the trigger and argument identification and classification. While F_1-scores of above 50% are obtained on the task of trigger identification, we observe a large gap in performance compared to results on the benchmark ACE05 dataset. We show that single-token triggers do not provide sufficient discriminative information for a fine-grained event detection setup in a closed domain such as economics, since many classes have a large degree of lexico-semantic and contextual overlap.

1 Introduction

We present a pilot study on a novel dataset annotated with economic and financial events in English company-specific news. Event processing automatically obtains the "what, who, where and when" of real-world events described in text. Event extraction consists of identifying event triggers, i.e. the tokens that express an event of a predetermined type, and identifying participant arguments, i.e. the tokens that express prototypical participant roles.

Event extraction is typically an upstream step in pipelines for financial applications: it has been used for news summarization of single (Lee et al., 2003; Marujo et al., 2017) or multiple documents (Liu et al., 2007; Glavaš and Šnajder, 2014), forecasting and market analysis (Nassirtoussi et al., 2014; Bholat et al., 2015; Nardo et al., 2016; Zhang et al., 2018; Chen et al., 2019), risk analysis (Hogenboom et al., 2015; Wei et al., 2019), policy assessment (Tobback et al., 2018; Karami et al., 2018), and marketing (Rambocas and Pacheco, 2018). This work aims to enable these information extraction tasks in the financial domain by making available a dataset and fine-grained event extraction model for company-specific news by classifying economic event triggers and participant arguments in text.

Our SENTiVENT dataset of company-specific events was conceived to be compatible with fine-grained event representations of the ACE benchmark corpora as to enable direct application of advances in the field. In our pilot study, we investigate the portability of an existing state-of-the-art model for event extraction, named DYGIE++ (Wadden et al., 2019b), to the task of financial event extraction.

2 Related Research

Our work on economic event extraction is accommodated within the rich history of automatic event detection. The ACE (Automatic Content Extraction) annotation scheme and programme was highly influential in event processing. Periodically, "Event Detection and Recognition" evaluation competitions were organized where event extraction corpora were released (Consortium, 2005; Walker et al., 2006) to enable automatic inference of entities mentioned in text, the relations among entities, and the events in which these entities participate (Doddington et al., 2004). Some years later, the ERE (Entities, Relations, Events) standard was conceived as a continuation of ACE with the goal of improving annotation consistency and quality. Our annotation scheme is inspired by the Rich ERE Event annotation schemes

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 235–245
Barcelona, Spain (Online), December 12, 2020.

(Linguistic Data Consortium, 2016; Linguistic Data Consortium, 2015a; Linguistic Data Consortium, 2015b) as to provide compatibility with on-going research in event extraction, where the ACE and ERE datasets remain dominant benchmarks.

Many approaches to the detection of economic events are *knowledge-based and pattern-based* (Arendarenko and Kakkonen, 2012; Hogenboom et al., 2013; Du et al., 2016; Chen et al., 2019). These use rule-sets or ontology knowledge-bases which are largely or fully created by hand and do not rely fully on manually annotated supervised datasets for machine learning. The Stock Sonar project (Feldman et al., 2011a) notably uses domain experts to formulate event rules for rule-based stock sentiment analysis. Their approach has been successful in formulating trading strategies (Ben Ami and Feldman, 2017) and in assessing the impact of events on the stock market (Boudoukh et al., 2016). Along the same line, Hogenboom et al. (2013) rely on a hand-crafted financial event ontology for pattern-based event detection in the economic domain and incorporates lexicons, gazetteers, PoS-tagging and morphological analysis.

Several *semi- or distantly supervision* approaches exist in which seed sets are manually labeled or rule-sets are used to generate or enhance training data (Qian et al., 2019; Rönnqvist and Sarlin, 2017; Ein-Dor et al., 2019). For Chinese, Yang et al. (2018) and Chen et al. (2019) rely on a knowledge-base of rules for extracting ACE-like events in stock market prediction. Han et al. (2018) describe a hybrid approach to ACE-like event extraction by labeling triggers, argument types, and event types for Chinese news articles with 8 economic event types using an automatically expanded trigger dictionary and used a pattern-matching approach for argument extraction.

Few strictly *supervised approaches* exist due to the lack of human-annotated ground-truth data. Malik et al. (2011) annotated Dividend and Profit figure slots and detected these types with high precision by combining supervised learning with rule post-processing. The English and Dutch SentiFM business news corpus (Van De Kauter et al., 2015) contains token-span annotations of 10 event types with only one type of relation. Sentence-level event detection experiments have been conducted (Lefever and Hoste, 2016; Jacobs et al., 2018), but supervised fine-grained extraction of triggers and roles as presented here. We are not aware of any other published fine-grained ACE-like event extraction approaches for the economic domain.

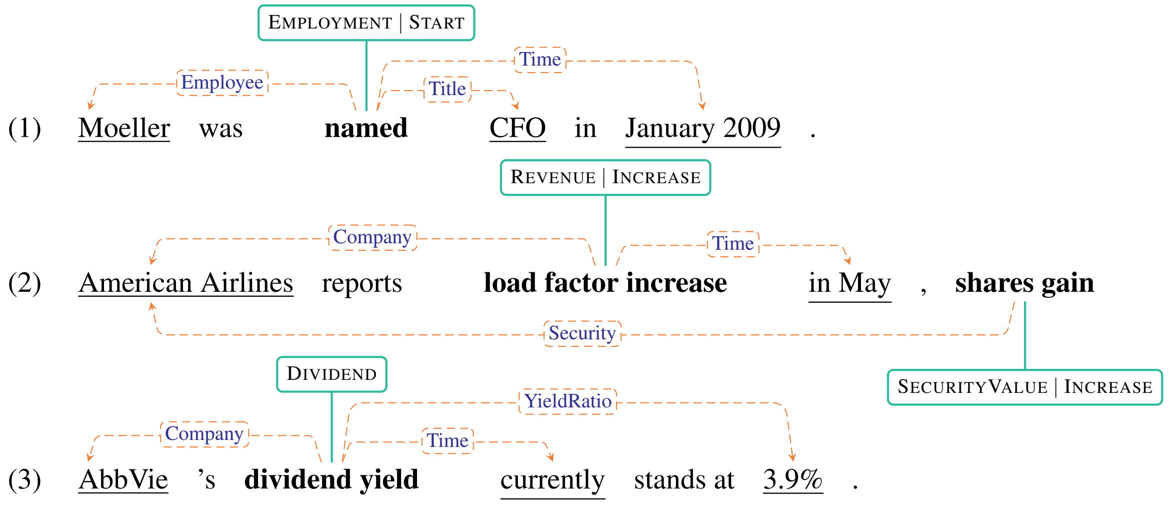
EMPLOYMENT | START
Employee
Time
Title
(1) Moeller was **named** CFO in January 2009 .
REVENUE | INCREASE
Company
Time
(2) American Airlines reports **load factor increase** in May , **shares gain**
Security
SECURITYVALUE | INCREASE
DIVIDEND
Company
YieldRatio
Time
(3) AbbVie 's **dividend yield** currently stands at 3.9% .

Figure 1: Examples of annotated economic event schemata with argument roles. **Boldface** indicates event trigger spans. Underlining indicates argument spans.

3 Dataset

We define economic events as '*textually reported real-world occurrences, actions, relations, and situations involving companies and firms*'. Fine-grained events in our dataset are operationalized as *event triggers*, i.e. the minimal span of text (a single word or a phrase) that most succinctly expresses the occurrence of an event type, and *event arguments*, i.e., participating entities involved in the event by filling a certain prototypical semantic role. Figure 1 shows examples of annotated event triggers with their TYPE | SUBTYPE label linked to argument with their roles.

We annotated 288 documents with structured events containing event types and subtypes, arguments and event attributes (i.e., negation, modality & event co-reference). The news article pertain to 30 S&P500 companies spread over different industries and time between June 2016 to May 2017 to avoid topical specialization. These companies were selected as a starting point for scraping because market data (e.g., key financials, spot prices, dividend, earnings and P/E ratio) for these are readily available, enabling market research. For a full description of the data collection, annotation process, dataset properties, and agreement study, we refer to Jacobs and Hoste (2020). The event typology contains 18 event main types with 42 subtypes for which 48 possible argument roles exist. Not every main event type has subtypes and the arguments "Time", "Place", and "Capital" can belong to any event. The typology was iteratively developed on a sample corpus in which news events were evaluated by financial domain experts. Starting with the set of event types in previous literature (Feldman et al., 2011b; Boudoukh et al., 2019; Hogenboom et al., 2013; Van De Kauter et al., 2015; Du et al., 2016), types were removed and added based on frequency and cohesion of categories. Relevant argument roles were also added in this process of iterative refinement. For a full description of the types and subtypes, we refer to the annotation guidelines (Jacobs, 2020b). Here is a list of main Types → unique Arguments:

- CSR/BRAND → Company
- DEAL → Goal, Partner
- DIVIDEND → Amount, Company, HistoricalYieldRatio, YieldRatio
- EMPLOYMENT → Amount, Employee, Employer, Replacer, Replacing, Title
- EXPENSE → Amount, Company, HistoricalAmount
- FACILITY → Company, Facility
- FINANCIALREPORT → Reportee, Result
- FINANCING → Amount, Financee, Financer
- INVESTMENT → CapitalInvested, Investee, Investor
- LEGAL → Allegation, Adjudicator, Complainant, Defendant, Sentence
- MACROECONOMICS → AffectedCompany, Sector
- MERGER/ACQUISITION → Acquirer, Cost, Target
- PRODUCT/SERVICE → Producer, Product/Service, Trialer
- PROFIT/LOSS → Amount, HistoricalAmount, Profiteer
- RATING → Analyst, HistoricalRating, Security, TargetPrice
- REVENUE → Amount, Company, DecreaseAmount, HistoricalAmount, IncreaseAmount
- SALESVOLUME → Amount, Buyer, GoodsService, HistoricalAmount, Seller
- SECURITYVALUE → DecreaseAmount, HistoricalPrice, IncreaseAmount, Price, Security

The types PROFIT/LOSS, REVENUE, EXPENSE, and SALESVOLUME represent major metrics in financial reporting on the income statement. Of all financial metrics, we found these types to be commonly and generally reported in business news. Other metrics such as debt ratios or cash flow statements as well as highly industry-specific indicators were collected under the FINANCIALREPORT type. Discussion of asset performance are captured by RATING and SECURITYVALUE. MACROECONOMICS is a broad category pertaining to events that do not involve decisions or interactions of specific companies. While this category is not company-specific, it was devised to capture discussions of sector trends, economy-wide phenomena, and governmental policy. In the event extraction pilot study, we only experiment with classifying main event types and not the subtypes.

Table 1 shows the counts of annotation units in our SENTiVENT corpus compared to the benchmark ACE2005-English-Events corpus (henceforth ACE05) (Walker et al., 2006) for the training, development and test set splits used in the pilot experiments.

An important difference in event trigger annotation with ACE05 is the tagging of discontinuous, multi-word triggers (e.g., "upgraded ... to buy", "cut back ... expenses", "EPS decline"). In ACE05, triggers are always single-word and any multi-word idiomatic expressions are joined into one token (e.g. "kick

	SENTiVENT				ACE05			
	Train	Dev	Test	Total	Train	Dev	Test	Total
Documents	228	30	30	288	529	28	40	597
Events	4,603	618	982	6,203	4,202	450	403	5,055
Sentences	5,475	681	727	6,883	17,172	923	832	18,927
Arguments	10,052	1,327	2,296	13,675	4,859	605	576	6040
Entities	12,732*	1,739*	1,441*	15.912*	29,006	2,451	3,017	34,474

	SENTiVENT	ACE05
Domain	Business News	Geo-Political
Entity cat.	-*	7
Trigger cat.	18	33
Argument cat.	48	22

Table 1: Properties for our SENTiVENT dataset and ACE05 in totals, as well as for the training, development and the test evaluation sets. * indicates that SENTiVENT contains no manually annotated entities.

the bucket" becomes "kick_the_bucket"). For SENTiVENT, 42% of triggers are multi-word and 13% are both discontinuous and multi-word.

Comparing annotation counts in ACE05 to our annotations, SENTiVENT has less documents but a higher density of event annotations. This reflects a difference in annotation approach in which events are less constrained by syntactic rules than in ACE05. Unlike ACE, we did not annotate entity tags and do not restrict event arguments definition by entity type. Not restricting events by syntactic rules or entity type results in a conceptually simplified annotation process. It also decreases to degree to which lexico-semantically present events are not tagged because of restrictive rules. Furthermore, the generality of geo-political event types in ACE05 compared to the domain-specific economic events in business news entails that more relevant events will be present. Due to the absence of entity annotations in our dataset, we also have significantly more arguments as in ACE05 argument roles are restricted by specific entity labels. Our argument role annotation is any span within the event scope that expresses the argument role and is thus semantically motivated.

4 Experimental Setup

Task definition Event extraction concerns identifying event triggers and their event types, event arguments and event roles. The input documents are represented as a sequence of tokens D from which the model constructs the set of all possible within-sentence spans $S = \{s_1, ..., s_T\}$ in the document (limited by a threshold-length). Each token d_i is assigned an event type label t_i. Then, for each trigger d_i, event arguments are assigned by predicting and argument role a_ij for all spans s_j in the same sentence as d_i.

Evaluation For ACE05, we follow the evaluation splits (cf. Table 1) and method commonly used in previous research (Nguyen et al., 2016; Sha et al., 2018; Zhang et al., 2019). All reported precision (P), recall (R), and F_1-scores (F_1) are micro-averaged. Our experiments involve the following four subtasks:

Trigger identification (Trig-ID) is the subtask of identifying if a token position matches a ground-truth reference trigger. *Trigger classification (Trig-C)* determines the type of the identified trigger. A trigger is correct when it is correctly identified and its type label matches the reference.

Argument identification (Arg-ID) is the subtask of identifying if a text span belongs to a certain event type. An argument is correctly identified when the span matches and the event type is correct. *Argument classification (Arg-C)* determines the *role* (e.g. "Employee") of an identified argument. An argument is correctly classified if it is correctly identified and its role label matches the ground-truth reference.

Data pre-processing All documents were sentence split and tokenized before annotation. As our event triggers are multi-token spans similar to Rich-ERE triggers, we transformed these multi-token spans into single-token triggers to allow us to use the DYGIE++ implementation in its basic configuration. For argument spans this pre-processing step is not needed as arguments are annotated as continuous, multi-token spans in both ACE05 and SENTiVENT. We selected single tokens by dependency parsing the triggers and selecting verbal and nominal root token within the trigger span using Spacy NLP library (Honnibal and Montani, 2017). When filtering multiple tokens to a single trigger, we use syntactic priority criteria based on PoS in which nominal tokens are prefered over verbal tokens which are prefered over other types. Search for these parts-of-speech goes one level down into the parse tree from the original trigger's root token, defaulting to the original root token if no nominal of verbal child is found. Examples of this pre-processing applied to original multi-token triggers:

• Profit/Loss	"improve margins"	→ "margins"
• Facility	"operating on-site gas stations"	→ "stations"
• SalesVolume	"attracting the greatest volume"	→ "volume"

This priority chain was empirically determined on a subsample of 15% of all triggers by comparing other selection criteria.

The DYGIE++ model also allows the use of named entity labels as features in training the trigger and argument subtasks. Unlike ACE05, our dataset does not include entity annotations (argument roles are unconstrained by entity types). To test if silver-standard entity labels improve identification and classification, we used entity label predictions of a trained entity model on the ACE05 dataset (Entity F_1-score: 90.7%).

Event extraction model For the event extraction pipeline we relied on the state-of-the-art DYGIE++ information extraction framework (Wadden et al., 2019b; Wadden et al., 2019a). As a current state-of-the-art model on ACE event extraction and other information extraction tasks, DYGIE++ is a good candidate for establishing a baseline on our new dataset.

DYGIE++ relies on graph propagation of contextualized embeddings of text spans in the local and global context to jointly model event triggers and arguments. The model enumerates candidate text spans and encodes them using pretrained contextual language models (such as BERT (Devlin et al., 2019)) and task-specific message updates passed over a text span graph. BERT embeddings are encoded at the subword-level by the WordPiece algorithm (Kudo, 2018) and to obtain token-level representations WordPieces are pooled. Tokens are encoded with frozen BERT embeddings using a sliding window, feeding each sentence to BERT together with a size-L window of surrounding sentences. Spans of text are enumerated and constructed by concatenating the tokens together with a learned span width embedding. With event propagation enabled, a graph is generated based on the model's best guess at the relations present among spans in the document. The event graph has two types of nodes: trigger nodes and argument nodes. Trigger nodes pass messages to their likely arguments updating those argument representations and arguments pass messages to their probable triggers. The re-contextualized representations are input into a two-layer feedforward neural net (FFNN) for each subtask. For a trigger token g_i, the final prediction is computed by FFNN$_{\text{trig}}(g_i)$ and for argument role prediction the relevant pair of embeddings is concatenated and FFNN$_{\text{arg}}([g_i, g_j])$ computed.

Model variations We experiment with two main variations in the model architecture: **BERT+LSTM** feeds pretrained BERT embeddings to a bi-directional LSTM layer, while the LSTM parameters are trained jointly with a task specific layer. **BERT Finetune** uses supervised fine-tuning of the BERT model on the end-task. We relied on the BERT implementations provided by the AllenNLP framework (Gardner et al., 2018): BERT$_{\text{Base}}$ with 12 layers (transformer blocks), 12 attention heads, and 110 million parameters. and BERT$_{\text{Large}}$ with 24 layers, 16 attention heads and, 340 million parameters. In initial testing, we found BERT$_{\text{Large}}$ improved performance in the LSTM variant over BERT$_{\text{Base}}$. For finetuning, we relied on the smaller BERT$_{\text{Base}}$ due to memory restrictions.

We examine the impact of **disabling the NER** subtask in the DYGIE++ event extraction pipeline. The SENTiVENT dataset does not contain manually annotated ground-truth entity labels. The labels used in experiments with joint training on the NER subtask are predictions from a pretrained ACE05 model.

We also **enable event propagation** where graph updates between triggers and arguments are computed. The base approach (BERT+LSTM) with event propagation bypasses trigger-argument span updates and directly feeds the embedded span representations from the Bi-LSTM layer to the feedforward scoring layer.

Additionally, we experimented with replacing the pretrained BERT general language model with the in-domain **FinBERT** model (Araci, 2019) as previous state-of-the-art systems have shown in-domain pretraining to increase performance on text classification and information extraction tasks (Howard and Ruder, 2018; Wadden et al., 2019b). FinBERT further pretrains BERT on the TRC2-financial corpus of news articles published by Reuters between 2008 and 2010 resulting in a model that outperforms regular BERT on financial sentiment analysis tasks.[1]

5 Results and Discussion

	SENTiVENT			ACE05		
Task	P	R	F_1	P	R	F_1
Trig-ID	48.23	58.10	52.71	70.22	83.33	76.22
Trig-C	38.86	46.81	42.46	67.04	79.56	72.76
Arg-ID	40.46	38.56	39.49	60.73	63.35	62.01
Arg-C	38.95	37.12	38.01	55.17	57.55	56.33

Table 2: Micro-avg. precision (P), recall (R), and F_1-score of the best system (BERT-Large + LSTM) on SENTiVENT and ACE05 for comparison.

Table 2 shows the results for the best configuration of the SENTiVENT model. The best model configuration was selected by the highest mean F_1-score on the Trig-C and Arg-C subtasks. The scores for the best model variant with similar configuration trained on ACE05 is given as comparison. The best model variation on the different SENTiVENT event subtasks is BERT + LSTM with the NER subtask enabled (Table 3). The same architecture also was best for ACE05.[2]

For our task, using the in-domain FinBERT did not improve performance over BERT on any task. In line with (Wadden et al., 2019b), fine-tuning decreases performance for both BERT and FinBERT configurations. This is likely due to the sensitivity of both the optimization hyper-parameters and BERT finetuning as the trigger detector begins overfitting before the argument detector is finished training. Event propagation also does not improve scores, likely due to the asymmetric relationship between triggers and arguments.

Disabling the NER subtask in which predicted entity labels are used as features lowers performance, showing that the silver-standard entity labels do provide useful features in training.

Variation	Trig-ID	Trig-C	Arg-ID	Arg-C
FinBERT+LSTM	50.97	41.78	37.04	35.44
FinBERT Finetune	49.96	42.34	30.24	29.19
BERT Finetune	50.37	42.22	30.75	29.41
BERT+LSTM	**52.71**	**42.46**	**39.49**	**38.01**
+EventProp	49.54	41.13	32.85	31.61
−NER	49.92	40.03	34.56	33.64

Table 3: Micro-avg. F_1-scores for model variations on the SENTiVENT-Event subtasks.

Interestingly, there is a large gap in performance on the two datasets while similar model settings were used and while a outwardly similar text genre (news text) was under investigation. We believe there

[1] The model weights were obtained from `https://github.com/ProsusAI/finBERT`.

[2] The ACE05 scores differ from those reported in Wadden et al. (2019b) as we did not use ensembles. We did use all the same data pre-processing, splits, and single-model settings provided by the authors.

might be several reasons for this performance gap. First of all, the drop in performance can to a certain extent be explained by the DYGIE++ requirement for single-token triggers, which is the unit of investigation in the ACE05 dataset. Many trigger types in our SENTiVENT corpus are differentiated lexically by the combination of a nominal and verbal phrase. In order to obtain single token triggers, a dependency parsing approach was taken in which the syntactic head noun or verb was preferred, introducing ambiguity. A multi-token trigger such as "conquer consumer spending" for SalesVolume would then lead to a more general and ambiguous trigger "spending" which is common in event type Expense). We hypothesize that single-token triggers delineate general-domain ACE05 geo-political event types (Life.Be-Born, Life.Marry, Movement.Transport, Transaction.Transfer-Money) better than our company-specific events.

Finally, due to the higher event-sentence density of our data (cf. Table 1), there is less global context to learn from, as well as a higher likelihood of overlap in the lexical context (local and global) of triggers and arguments.

Error analysis Table 4 shows the frequency of errors made by the BERT+LSTM system. For event triggers, missing identified triggers (51.7%) constitutes the largest error and 13.6% involves misclassified cases where a trigger was identified but the event type was mistaken (e.g., a Revenue event is assigned Profit/Loss). Missing event arguments is the largest error type (70.8%) for arguments with misclassification playing a minor role (1.3%). Spurious triggers or arguments account for 34.7% and 27.9%, respectively and they occur when token spans are assigned a label where none is present.

Error type	Missing	Spurious	Misclassified
Trigger	51.7%	34.7%	13.6%
Argument	70.8%	27.9%	1.3%

Table 4: Frequency of error types of the best system on the test set.

Trigger Error Type	Specialized/Creative Language	Lexical Sparsity	Plausible Spurious	Ambiguous Trigger	Single-Token Pre-processing
Freq.	12.4%	9.1%	16.4%	20.1%	34.1%

Table 5: Frequency of trigger errors in manual assessment.

We also manually reviewed half of the evaluation test set documents and annotated the errors in more detail (Table 5). *Lexical sparsity errors* where triggers are missing or misclassified because they are rare or unseen in training is less of a problem (associated with 9% of errors). *Highly specialized contexts or creative language use* introduces contextual lexical sparsity. This occurs more frequently with company-specific news than general news because industry, product, or company-specific terminology is used; e.g., "*Besides its **track-tested** suspension and race-ready seats, the Edge ST's looks take a dark turn...*" → true: Product/Service, pred.: Missing. *Ambiguous triggers* are also a common source of errors, e.g., "growth" is often used for various types related to financial metrics such as Profit/Loss, Revenue, Expense, etc.: e.g., "*Expect strongest **growth** from services and Asia.*" → true: FinancialReport, pred.: SalesVolume. "*Apple **made** $64.1 billion before taxes in fiscal 2017.*" → true: Profit/Loss, pred: Revenue. Another example is "buy", which is a trigger often used for Merger/Acquisition but often misclassified as a 'buy/hold/sell' Rating. Better capturing the long context and event co-reference should resolve ambiguous trigger mentions as often the preceding context specifies the metric. A large amount of errors (34%) were due to the *pre-processing* step of converting our ground-truth multi-token triggers into single-tokens. This discards discriminative information regarding type and causes many spurious and missing errors. e.g., "*... organic sales **growth** projections for this year ...*" where "sales growth" is the original annotated SalesVolume trigger reduced to "growth" introducing class ambiguity which has evidently not been resolved by local or global context. Spurious triggers that *plausibly* express an event but are absent in the ground-truth are also fairly common (16%). This is not an artifact of low-quality annotations but occurs with non-salient mentions of events which are generic or unspecific in nature and for which a more concrete example is annotated in the direct vicinity. Our conceptualization of economic events includes specificity and relevancy in annotation, which is not well-captured by the model, e.g., in "*Below is an analysis of Apple's App Store revenue...*" "revenue" is a spurious REVENUE prediction that is not in the ground-truth because it is a generic mention followed by a series of fully-realized events.

6 Data availability and replication

This work's source code, preprocessed replication data in DYGIE-format, and the winning trained model are publicly available at `https://osf.io/j63h9` (Jacobs, 2020a). The original dataset will be freely downloadable at the end of the SENTiVENT project through this repository. Up until then, the fully annotated corpus is available on request for academic research purposes.

7 Conclusion and Future Work

Event extraction is a required step in many data-driven financial tasks in which factual information is needed to capture changes in the real-world. Various general domain fine-grained event extraction corpora are freely available but no economically focused corpus exists. We presented a pilot study on a novel dataset enabling supervised fine-grained extraction of economic events as trigger and argument classification. Using a state-of-the-art information extraction pipeline based on span-based graph propagation of pretrained contextual embeddings, we observed a large drop in performance on our dataset compared to the benchmark ACE05 dataset. After error analysis, we found this is largely caused by missing predictions. For event triggers, many errors are due to ambiguity introduced by the requirement of the DYGIE++ model implementation for single-token triggers. We hypothesize that single-token triggers for the SENTiVENT corpus do not provide sufficient discriminative information for detecting classes which have a large degree of semantic and contextual overlap.

Hence, in future work we will focus on event extraction methods that model arbitrary length triggers. Currently on-going, we are adding "investor sentiment" annotations on top of events as well as separate sentiment expression annotations with their targets. These sentiment annotations will allow us to jointly process the "common-sense" sentiment of events. We will investigate how extracted event schemata can be used upstream from aspect-based sentiment analysis.

Acknowledgements

This work was supported by the Research Foundation Flanders (FWO) under a Ph.D fellowship grant for the SENTiVENT project. We would like to thank anonymous reviewers for their helpful suggestions, as well as David Wadden for publishing and maintaining their model source code and answering questions about the model.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Ernest Arendarenko and Tuomo Kakkonen. 2012. Ontology-Based Information and Event Extraction for Business Intelligence. In *Artificial Intelligence: Methodology, Systems, and Applications*, volume 7557 of *Lecture Notes in Computer Science*, pages 89–102. Springer.

Zvi Ben Ami and Ronen Feldman. 2017. Event-based trading: Building superior trading strategies with state-of-the-art information extraction tools. SSRN Working Paper 2907600.

David Bholat, Stephen Hansen, Pedro Santos, and Cheryl Schonhardt-Bailey. 2015. Text mining for central banks. *Available at SSRN 2624811*.

Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew P Richardson. 2016. Information, trading, and volatility: Evidence from firm-specific news. SSRN Working Paper 2193667.

Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew Richardson. 2019. Information, trading, and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, 32(3):992–1033.

Deli Chen, Yanyan Zou, Keiko Harimoto, Ruihan Bao, Xuancheng Ren, and Xu Sun. 2019. Incorporating fine-grained events in stock movement prediction. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 31–40, Hong Kong, November. Association for Computational Linguistics.

Linguistic Data Consortium. 2005. Ace (automatic content extraction) english annotation guidelines for events version 5.4.3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

George R Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program–tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Mian Du, Lidia Pivovarova, and Roman Yangarber. 2016. PULS: natural language processing for business intelligence. In *Proceedings of the 2016 Workshop on Human Language Technology*, pages 1–8.

Liat Ein-Dor, Ariel Gera, Orith Toledo-Ronen, Alon Halfon, Benjamin Sznajder, Lena Dankin, Yonatan Bilu, Yoav Katz, and Noam Slonim. 2019. Financial Event Extraction Using Wikipedia-Based Weak Supervision. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 10–15, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. 2011a. The stock sonar sentiment analysis of stocks based on a hybrid approach. In *Twenty-Third IAAI Conference*.

Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. 2011b. The Stock Sonar — Sentiment Analysis of Stocks Based on a Hybrid Approach. *Iaai*, pages 1642–1647.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July. Association for Computational Linguistics.

Goran Glavaš and Jan Šnajder. 2014. Event graphs for information retrieval and multi-document summarization. *Expert Systems with Applications*, 41(15):6904–6916.

Songqiao Han, Xiaoling Hao, and Hailiang Huang. 2018. An event-extraction approach for business analysis from online Chinese news. *Electronic Commerce Research and Applications*, 28:244–260.

Alexander Hogenboom, Frederik Hogenboom, Flavius Frasincar, Kim Schouten, and Otto Van Der Meer. 2013. Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications*, 64(1):27–52.

Frederik Hogenboom, Michael de Winter, Flavius Frasincar, and Uzay Kaymak. 2015. A news event-driven approach for the historical value at risk method. *Expert Systems with Applications*, 42(10):4667–4675.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1). Dependency parsing model used: en_core_web_lg v2.3.1.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Gilles Jacobs and Véronique Hoste. 2020. SENTiVENT: Enabling supervised information extraction of company-specific events in economic and financial news. *Manuscript submitted for publication.*

Gilles Jacobs, Els Lefever, and Véronique Hoste. 2018. Economic event detection in company-specific news text. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 1–10.

Gilles Jacobs. 2020a. Replication data for extracting fine-grained economic events from business news, Oct.

Gilles Jacobs. 2020b. SENTiVENT Event Annotation Guidelines v1.1. Technical report, LT3, Ghent University, jun.

Amir Karami, London S Bennett, and Xiaoyun He. 2018. Mining public opinion about economic issues: Twitter and the us presidential election. *International Journal of Strategic Decision Sciences (IJSDS)*, 9(1):18–28.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Chang Shing Lee, Yea Juan Chen, and Zhi W Jian. 2003. Ontology-based fuzzy event extraction agent for Chinese e-news summarization. *Expert Systems with Applications*, 25(3):431–447.

Els Lefever and Véronique Hoste. 2016. A classification-based approach to economic event detection in Dutch news text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 330–335, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Linguistic Data Consortium. 2015a. DEFT Rich ERE Annotation Guidelines: Argument Filler V2.3. Technical report, Linguistic Data Consortium, 11.

Linguistic Data Consortium. 2015b. DEFT Rich ERE Annotation Guidelines: Events V3.0. Technical report, Linguistic Data Consortium, 11.

Linguistic Data Consortium. 2016. Rich ERE Annotation Guidelines Overview V4.2. Technical report, Linguistic Data Consortium. Accessed: 2018-09-05.

Maofu Liu, Wenjie Li, Mingli Wu, and Jun Hu. 2007. Event-based extractive summarization using event semantic relevance from external linguistic resource. *Proceedings - ALPIT 2007 6th International Conference on Advanced Language Processing and Web Information Technology*, pages 117–122.

Hassan H. Malik, Vikas S. Bhardwaj, and Huascar Fiorletta. 2011. Accurate information extraction for quantitative financial events. *International Conference on Information and Knowledge Management, Proceedings*, pages 2497–2500.

Luís Marujo, Ricardo Ribeiro, Anatole Gershman, David Martins de Matos, João P. Neto, and Jaime Carbonell. 2017. Event-based summarization using a centrality-as-relevance model. *Knowledge and Information Systems*, 50(3):945–968.

Michela Nardo, Marco Petracco-Giudici, and Minás Naltsidis. 2016. Walking down wall street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys*, 30(2):356–369.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California, June. Association for Computational Linguistics.

Yu Qian, Xiongwen Deng, Qiongwei Ye, Baojun Ma, and Hua Yuan. 2019. On detecting business event from the headlines and leads of massive online news articles. *Information Processing and Management*, 56(6):102086.

Meena Rambocas and Barney G Pacheco. 2018. Online sentiment analysis in marketing research: a review. *Journal of Research in Interactive Marketing*.

Samuel Rönnqvist and Peter Sarlin. 2017. Bank distress in the news: Describing events through deep learning. *Neurocomputing*, 264:57–70.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *AAAI Conference on Artificial Intelligence*.

Ellen Tobback, Hans Naudts, Walter Daelemans, Enric Junqué de Fortuny, and David Martens. 2018. Belgian economic policy uncertainty index: Improvement through text mining. *International Journal of Forecasting*, 34(2):355 – 365.

Marjan Van De Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42(11):4999–5010.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019a. Dygie++: Span-based system for named entity, relation, and event extraction. `https://github.com/dwadden/dygiepp`.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019b. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China, November. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus.

Lu Wei, Guowen Li, Jianping Li, and Xiaoqian Zhu. 2019. Bank risk aggregation with forward-looking textual risk disclosures. *The North American Journal of Economics and Finance*, 50:101016.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55.

Xi Zhang, Siyu Qu, Jieyun Huang, Binxing Fang, and Philip Yu. 2018. Stock market prediction via multi-source multiple instance learning. *IEEE Access*, 6:50720–50728.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.