Cross-document Extraction of Economic Events

Abstract

This document is an appendix of *Cross-document Extraction of Economic Events*. It is meant to be an integral part of the paper, however, due to space constraints it will only be included in the final version (which allows for purchase of up to 2 additional pages).

3 Representing Economic Events

We represent financial transactions using a purpose-built ontology. Our starting point is the REA (Resource, Event, Agent) model (McCarthy 1982) that is often used as a foundational model for describing business-related concepts; it is briefly introduced in §3.1. To be able to capture more fine-grained semantic distinctions about financial transactions, we extend REA with a hierarchy of economic event types in §3.2.

3.1 REA

This section should be shortened to bare minimum.

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. For instance, when buying a car, the agent receives an automobile; the agent has to pay for the car and that decreases her wealth. This reciprocal phenomena, called *duality*, is a fundamental element in REA.

3.2 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., profit-gross → earn → get); subsequent processes can then choose the granularity with which they want the information

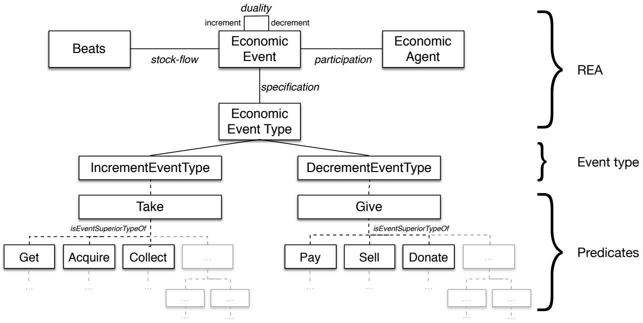

Figure 1: Overview of the Ontology of Economic Events (OEE). Note that the bottom part shows only an excerpt from the instances.

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 (OEE), an extension to REA; see Fig. 1 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 (Miller 1995).

The main class of our economic events ontology is called *EventType*. Following Hruby (Hruby 2006), we differentiate between two major economic event types: events increasing and decreasing the value of agent's resources. These subclasses 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.

Table 1: List of features. Type can be binary (B), categorical (C), or numerical (N).

Feature	Type	Description
sentence_length	N	Length of the sentence
article_length	N	Length of the article
sentence_order	N	Sentence's position within in the article
predicate_tense	C	Tense of the predicate
is_noun_predicate	B	Predicate is expressed by a verb or a noun
s_has_dbpedia_uri	B	Subject has a DBpedia URI
s_has_crunchbase_uri	B	Subject has a CrunchBase URI
s_has_freebase_uri	B	Subject has a Freebase URI
o_has_dbpedia_uri	B	Object has a DBpedia URI
o_has_crunchbase_uri	B	Object has a CrunchBase URI
o_has_freebase_uri	B	Object has a Freebase URI
correct_fin_argument	B	Fin. value is within the correct semantic arg.
correct_temp_argument	B	Temp. value is within the correct semantic arg.
has_event_date	B	Temp. expression was found within the sentence
nytc_descriptor_business	B	Article is classified under "Business" according to the NYTC taxonomy
pred_frequency	N	Relative freq. of the predicate in the corpus
values_ratio	N	Relative freq. of the given monetary value in R_e
dates_count	N	Number of quintuples with the same date in R_e

(c) Find sister 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. 1).[1] In §8.1 (now part of appendix) 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.

5 Creating Structured Representations of Economic Events

5.1 Selecting a Single Quintuple

Supervised learning approach We cast the selection of the best quintuple as a regression task and use a machine learning approach.

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 1 for a detailed list.

8 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 (§8.1), measure the importance of individual features (§8.2), and take a closer look at some successes and failures (§8.3).

Table 2: Feature importance based on Gini importance.

Feature	Gini	Feature	Gini
dates_count	0.186	o_has_dbpedia_uri	0.024
article_length	0.145	nytc_descriptor_business	0.023
sentence_length	0.137	has_event_date	0.021
sentence_order	0.129	correct_temp_argument	0.019
values_ratio	0.088	o_has_freebase_uri	0.016
correct_fin_argument	0.064	is_noun_predicate	0.010
pred_frequency	0.051	s_has_dbpedia_uri	0.007
predicate_tense	0.043	s_has_crunchbase_uri	0.000
o_has_crunchbase_uri	0.041	s_has_freebase_uri	0.000

8.1 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 2 shows the predicates ordered by number of occurrences up to the first three levels of OEE. The most frequent predicate, *pay*, is mentioned in over 66K sentences.

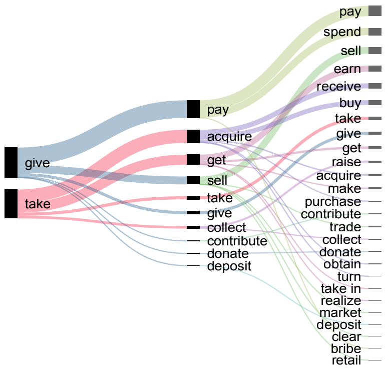

Figure 2: Predicate frequency in the NYTC.

8.2 Features

Table 2 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.

8.3 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

[1] 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.

these events, using relaxed evaluation, are as follows: the earliest baseline fails in 4 cases, the latest baseline fails in 5 cases, while the supervised learning approach was incorrect only in a single case. Table 3 shows a specific example, where only the supervised learning method returned the correct quintuple.

Table 3: Example of a transaction with multiple quintuples: Oracle acquired PeopleSoft.

Predicate	Mon. value	Year	Published	Method	Correct
acquire	$7.3 bn	2003	2003-11-25	BL, earliest	N
acquisition	$7.7 bn	2004	2004-10-26	-	N
acquisition	$7.7 bn	2004	2004-10-26	-	N
acquire	$1.3 bn	2004	2005-12-23	-	N
acquire	$7.038 bn	2004	2005-12-23	-	N
acquire	$10.3 bn	2004	2007-03-01	**SL**	**Y**
acquisition	$10.3 bn	2005	2005-06-30	-	N
purchase	$20 bn	2007	2007-03-21	BL, latest	N

References

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

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

Miller, G. A. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41.