Cross-document Extraction of Economic Events

1 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 (§1.1), measure the importance of individual features (§1.2), and take a closer look at some successes and failures (§1.3).

1.1 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 1 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.

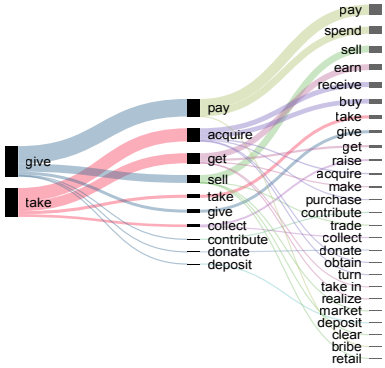

Figure 1: Predicate frequency in the NYTC.

1.2 Features

Table 1 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.

Table 1: 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

1.3 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 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 2 shows a specific example, where only the supervised learning method returned the correct quintuple.

Table 2: 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