Understanding Relations using Concepts and Semantics

Jouyon Park
Yonsei University
parkjy0235@yonsei.ac.kr

Hyunsouk Cho
POSTECH
prory@postech.ac.kr

Seung-won Hwang
Yonsei University
seungwonh@yonsei.ac.kr

ABSTRACT

The Financial Entity Identification and Information Integration (FEIII) task aims at the question of understanding relationships among financial entities and their roles using three sentences extracted from each financial contract containing the target word. FEIII task has two challenges - **1) data sparseness:** small training sets (9% of test data) and **2) context sparseness:** limited context (three sentences). Existing statistical approaches, such as Bayes and TF-IDF, cannot evaluate the imporatance of words unobservged in training data, which is vulnerable to the above challenges. We overcome each challenge by considering 1) the concepts of words from knowledge bases (Probase) in addition to the words themselves (**conceptual feature**) and 2) word semantics from distributed representations such as word2vec (**semantic feature**). We empirically evaluate the proposed classification model on the four-class classification (highly relevant, relevant, neutral, and irrelevant), and show that the proposed model increases 18% of F1-score compared to the statistical baselines.

ACM Reference format:
Jouyon Park, Hyunsouk Cho, and Seung-won Hwang. 2017. Understanding Relations using Concepts and Semantics. In *Proceedings of DSMM'17, Chicago, IL, USA, May 14, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3077240.3077250

1 INTRODUCTION

The Financial Entity Identification and Information Integration (FEIII) challenge aims to identify and understand the relationships among financial entities and the roles that they play in financial contracts as represented in documents and databases. The data set consists of 10-K and 10-Q filings, and the task is to identify sentences in the filings that provide evidence for a specific relationship between the filing financial entity and another mentioned financial entity.

The important task is to rank the **triples** (consisting of three sentences, filer name, and role keyword) such that the triples with the context best supports that role assignment and contains financially relevant knowledge are at the top of the ranking, by giving a relevance score between 0 and 1 representing the ranking. To guide for evaluation, experts labeled the training data triples as being Highly relevant, Relevant, Neutral, and Irrelevant. In this paper, we focus on classifying unlabeled working sets into the given four classes.

2 METHOD

Our work adopt a SVM classifier to identify the context (three sentences) using statistic, conceptual, and semantic features. We will describe in the next sections, more specifically, (1) Statistic Features: Bayes, TF-IDF (2) Concept Features: Probase, and (3) Semantic Features: Word2vec. To compute our features, we use a randomly selected training set of 90% of the labeled data. We use three types of features from the training set.

2.1 Statistic Features: Baseline

The first feature type consists of empirical statistical methods: Bayes statistics, and the TF-IDF score.

- Bayes statistics: To classify text from the labeled data, Sahami et al. [2] studies Bayes rule for junk mail. From Bayes rule, we compute $P(c_i|w) = \frac{P(w|c_i)P(c_i)}{P(w)}$, where c_i stands for category ($c_i \in C = \{$ Highly relevant, Relevant, Neutral, Irrelevant$\}$) and w represents a word. Using the probability for each word $P(c|w)$, we compute the category probability of the given sentence as a combination of word probability, *i.e.*, $P(c|s) = P(c|w_1)P(c|w_2), ...P(c|w_n)$, where s represents sentence and $s = (w_1, w_2, w_3, ...w_n)$
- TF-IDF: We leverage TF-IDF score widely adopted for word importance. From each word in the training data, we compute TF-IDF as: $TF(w, c_i) = \frac{Freq(w,c_i)}{\sum_{w_j \in c_i} Freq(w_j, c_i)}$ and $IDF(w) = \log(\frac{|C|}{|c_i \in C:\ w \in c_i|})$.

2.2 Concept Features for data sparseness

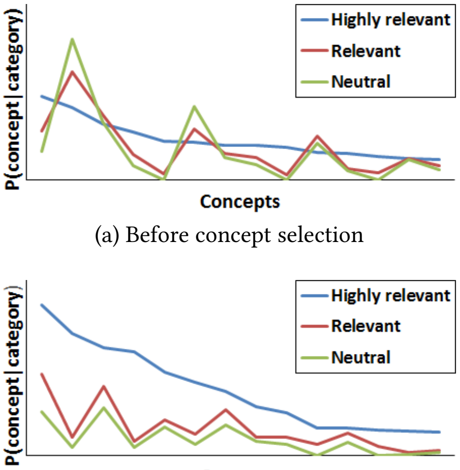
Statistic features (Bayes statistics and TF-IDF) cannot compute the importance words if it does not appear in the training set. However, due to the data and context sparseness mentioned in the abstract, we expect lots of untrained words in the unlabeled sets. In order to give those words some sort of features for classification, we represent a word as a distribution of concepts it belongs to. For this purpose, we adopt a probabilistic knowledge base from Microsoft, namely Probase [3], which contains 2.7 million concepts harnessed automatically from a corpus of 1.68 billion web pages. Table 1 shows the example concept statistics available for an entity in Probase: Each entity corresponds to the list of concept probabilities which are extracted from web pages using 'such as' textual pattern: For example, textual occurrence of 'factor such as abrasion' generates a row in Table 1, and probability $P(o|e)$ suggests the *typicality* of entity 'abrasion' associated with concept 'factor'. In other word, abrasion can be represented as a probability vector, with dimensionality as large as all possible concepts.

Given this concept vector to represent an entity, we can overcome data sparseness, by computing the similarity of words, as the similarity of concept distributions. This enables to compute the word similarity for words unseen in the training data. A sentence

DSMM'17, Chicago, IL, USA
© 2017 ACM. 978-1-4503-5031-0/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3077240.3077250

Table 1: Probase Example

Entity (e)	Concept (o)	Probability $P(o\|e)$
abrasion	factor	0.0348
abrasion	condition	0.0255
⋮	⋮	⋮



(a) Before concept selection

(b) After concept selection

Figure 1: Top-k Frequently Concept Distributions

also can be represented as a concept vector, by using Stanford POS tagger (https://nlp.stanford.edu/software/tagger.shtml) to identify entities, and compute the average concept distribution of all entities in the sentence.

Our hypothesis of using this feature is that, concept distributions should be similar between financial entities with "Relevant" or "Highly Relevant" labels, while those between relevant and irrelevant entities should be dramatically different. Figure 1 (a) weakly confirms this hypothesis, showing similar distributions between Relevant and Highly Relevant, compared to Neutral.

Though a well-designed distribution similarity metric may capture such difference, we aim to make concept feature more robust, by making the difference more drastic. We observed that, the use of abstract concepts, such as 'information', 'datum' and 'service', are frequently observed from all labels, which diminishes the usefulness of this feature. We thus propose to automatically select a more effective subset of concepts O_k, by comparing the top-k concepts observed from the sentences with highly relevant, relevant, and neutral labels, which we denote as c_{HR}, c_R, and c_N respectively.

More specifically, we define O_k as:

$$O_k = arg \max_{o_i \in O_k} \{P(o_i|c_{HR}) - P(o_i|c_R) - P(o_i|c_N)\} \qquad (1)$$

where $|O_k| = k$. We empirically tune k to $k = 14$ and we replace the 14 concepts used in Figure 1 (a) into $O_K = \{$expense, dept, legal, loan, product, credit, transaction, payment, banking, contract, law, finance, secured, obligation$\}$. We can qualitatively observe that these concepts are more relevant to FEIII domain. In addition, Figure 1 (b) shows that using these features makes the distribution difference more dramatic as well.

2.3 Semantic Features for context sparseness

Though concept features enable to compute similarity for unseen words, both statistical and conceptual features treat the three sentences as a *bag of words*. Our research question in this section is whether bag of model is sufficiently effective for modeling the given **triple**, or three sentences? Would having two sentences adjacent or not, in these three sentences, change the meaning? If this is the case, we need an additional feature to capture such *transition*.

As a hypothesis, when denoting three sentences S_1, S_2, and S_3, we define transition vector T_{ij} between two adjacent sentences S_i and S_j. We define T_{ij}, as a vector operation of $S_i - S_j$, where S_i is represented as Word2vec [1]: Word2vec is a continuous Skip-gram model which is an efficient method for learning high-quality distributed vector representations representing a large number of precise syntactic and semantic word relationships. As we empirically report later, these feature add accuracy over statistic and conceptual features.

3 EXPERIMENT

We evaluate the performance of the proposed features with SVM classification. We randomly divide the labeled set by an approximate ratio of 9(train):1(test). We use multi-class C-SVC classification with polynomial kernel type, with gamma value 0.25. The performance is evaluated by the classification accuracy.

Table 2: Divided Validation Performance

	Features		
Statistic	Concept	Semantic	Accuracy (%)
O	X	X	54.59
O	O	X	65.94
O	O	O	**72.71**

Table 2 shows the performance of the proposed features. We increase 11% accuracy by using the concept features compare to baseline performance (only statistic features are used). The best performance is achieved when all three features were used, which is 18% higher than the performance of the baseline.

4 CONCLUSION

In our work, we propose a method using various features for classification. We use statistical, conceptual and semantic features, and used a SVM regression. Our empirical results confirm the effectiveness of each feature, and show the complementary strength of these feature when aggregated for classification.

REFERENCES

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[2] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. 1998. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, Vol. 62. 98–105.

[3] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 481–492.