Hybrid Feature Factored System for Scoring Extracted Passage Relevance in Regulatory Filings

D. Proux
Xerox Research Centre Europe
6 Chemin de Maupertuis, 38240 Meylan
France
denys.proux@xrce.xerox.com

C. Roux
Xerox Research Centre Europe
6 Chemin de Maupertuis, 38240 Meylan
France
claude.roux@xrce.xerox.com

A. Sándor
Xerox Research Centre Europe
6 Chemin de Maupertuis, 38240 Meylan
France
agnes.sandor@xrce.xerox.com

J. Perez
Xerox Research Centre Europe
6 Chemin de Maupertuis, 38240 Meylan
France
julien.perez@xrce.xerox.com

ABSTRACT

We report in this paper our contribution to the FEIII 2017 challenge addressing relevance ranking of passages extracted from 10-K and 10-Q regulatory filings. We leveraged our previous work on document structure and content analysis for regulatory filings to train hybrid text analytics and decision making models. We designed and trained several layers of classifiers fed with linguistic and semantic features to improve relevance prediction. We discuss in this paper our experiments and results on the competition data set.

1 INTRODUCTION

Information is the lifeblood of financial markets, and the amount of data available to decision-makers is increasing exponentially, with 90% of global information created in the last decade [1]. Efficient and effective financial market operations rely on information, much of which involves unstructured texts (e.g., annual reports, press releases, web pages, research reports, regulatory guidelines, financial and social media, internal documents, blogs, etc.). Textual content can range from a few words (e.g., a newspaper headline or Tweet) to detailed and complex documents (e.g., annual reports). Failure to analyse and take into account information disclosed in such channels leads to ignored risks, uninformed decisions and missed opportunities. The ability to locate and process large quantities of text based information is therefore a growing issue in financial markets [2]. This is an opportunity for text analytics (TA) to develop methods that support decision-makers in understanding market dynamics, predicting outcomes and trends, formulating strategies, curbing fraud and managing risk. However, despite the fact that TA is gaining prominence in all industries, the financial sector still faces many challenges when moving beyond simple document retrieval and classification tasks. Various attempts are nevertheless trying to address this domain mostly driven by the expectation to be able to predict stock price variations [3]. These attempts mostly focus on the analysis of short text such as financial news articles and Tweets [4, 5]. Addressing longer and more complex documents, such as financial reports (e.g. 10-Ks, N-CSRs), raise additional challenges [6].

These challenges come from the multiplicity of communication channels, the high complexity (and sometimes subjectivity) of the information disclosed and the frequent verbosity (not to say opacity) of the style of regulatory filings. Dealing with these challenges often requires a combination of multidisciplinary expertise addressing linguistics, computer science, regulation, sociology and financial markets.

In an industrial project internal to our company, in a multidisciplinary team, we developed a text analytics platform designed to process N-CSR and N-Q fillings (related to mutual funds reporting for the Security and Exchange Commission) to spot and extract specific information such as lists of mentioned key peoples (e.g. corporate officers, members of the board, trustees), names of service providers, asset valuations data and investment figures. For this competition, we integrated some of the linguistic analysis components developed at that time with a combination of logistic regression based classifiers to train a system designed to rank relevant information within the Financial Entity Identification and Information Integration (FEIII) data set. The system design, the challenges we faced and the results we get are described in this paper.

DSMM'17, May 14, 2017, Chicago, IL, USA

ACM 978-1-4503-5031-0/17/05...$15.00
http://dx.doi.org/10.1145/3077240.3077251

2 ABOUT REGULATORY FILLINGS

Regulation organizations such as the Security Exchange Commission (SEC) request registered mutual funds or public companies to communicate on a quarterly or yearly basis their key financial and operational information through specific forms (e.g. 10-K, 10-Q, N-CSR, N-Q). For this purpose, the SEC maintains the Electronic Data Gathering and Reporting ("EDGAR") system into which regulated entities submit about 15 million documents each year. 10-k documents, for public companies, are well defined forms specifying the nature of what should be reported. However, the way this information is reported or its layout is up to each company. These documents contain various pieces of information structured in free text paragraphs, tables, lists of items, graphs, footnotes, etc. Information in such documents could be called semi-structured. In terms of document content processing, this raises some challenges, since the way information needs to be processed in a paragraph is different from the way it needs to be processed in tables, or in lists. Furthermore, each specific section (e.g. a set of paragraphs) carries its own semantics. Assumptions discussed in "risk factor" sections, or expectations in "management's discussion" sections should be considered differently from facts discussed in the "Financial Statement" part.

Similarly, information disclosed in tables must be processed taking into account the semantics carried out by the structure. For example, searching for the current position of some key people mentioned in such a document requires the ability to make a difference with other positions occupied in other companies (which is often required by regulation). Footnotes also bring additional highlights to the way some facts or figures should be understood (e.g., modifications in the way inventory is performed, moving from a LIFO to a FIFO method). Therefore, the meaning and usefulness of information extracted from such documents is really strongly dependent on the context and structure where it is discussed.

3 SYSTEM DESCRIPTION

3.1 Architecture

As previously discussed, 10-K filings are complex documents to process. The way information should be understood strongly depends on the part where this information is discussed and the existence of additional contextual materials to disambiguate disclosed elements. However, for this competition only sets of extracted texts were available without their context. A deeper analysis on the training data set revealed that these texts originate from various places such as paragraphs, footnotes sections and sometimes tables or lists of items. This implies that not all of this content is syntactically well-formed.

Therefore we decided to use a hybrid method to benefit from available syntactic and semantic features when possible but still using a broader statistical approach based on other categories of features. We combined a linguistic parser, a semantic analysis component and several layers of different logistic regression based classifiers. These components are detailed below.

3.2 Linguistic Features

As for linguistic features, we wanted to extract elements such as part of speech tags and syntactic dependencies to evaluate if they benefit to the decision making process. To do so we used the Xerox Incremental Parser (XIP), a robust syntactic and semantic parser [7] as one component of our system. XIP has previously been adapted for many domains of application, including Aspect Based Sentiment Analysis, and for several competitions such as SemEval [8, 9]. This parser provides a full processing chain including tokenization, morpho-syntactic analysis, POS tagging, chunking and finally the extraction of dependency relations such as subject, object and modifiers. XIP also allows semantic analysis like for Named Entity Recognition: it detects person names, organizations, locations, dates and measures. The output yields a rich set of linguistic features associated with every layer of the analysis. This information is very valuable when trying to identify specific relations among Named Entities or among facts. But the quality of the analysis is of course strongly dependent on the quality and nature of the input: results are best when the input consists of syntactically well-formed sentences.

However in the context of this competition the corpus is composed of a mixture of well-formed sentences and fragments of text possibly coming from tables and lists. Furthermore, even well-formed sentences are generally long and complex which makes long distance syntactic dependencies difficult to disambiguate. Therefore a decision making process based only on such elements would have a negative impact on recall. This is why we decided to use these features only as an additional input to a set of classifiers. Furthermore we did not kept all features generated by XIP. In the model adaptation process we filtered some to promote those with the highest impact on classification scores for the training dataset as discussed in section 4.2.

3.3 Semantic Features

In addition to the syntactic level we wanted to enrich further the diversity of input of our classifiers with additional categories of features. To do so we reused a previous work targeting financial reports related to mutual funds. Indeed we developed in the past a platform that extracts some information from N-CSR documents - which are similar to 10-Ks - and from Accounting and Auditing Enforcement Releases (AAER) disclosed by the SEC. For this, we developed a document structure analysis component and on top of the general XIP parser a set of semantic rules for detecting key people and their roles, as well as service providers, asset valuation, investment holdings information, companies that violate laws, types of litigation and periods of fraud.

To associate a person's or an organization's name with their own roles at the right dates, the semantic rules take into account the context (both syntactic and structural). Moreover, we used lexical resources provided by financial experts, like synonym lists and terminology lists.

Such rules thus allow us to identify in example 1 "Nuveen Fund Advisors, LLC", and not the former company "Nuveen Fund Advisors, Inc." as investment advisor, based on the fact that investment advisor is syntactically related (as a subject

complement) to the first company name in the sentence, and the XIP parser detects this relationship.

> ***Nuveen Fund Advisors, LLC**, formerly known as Nuveen Fund Advisors, Inc., is the registrant's **investment adviser** (also referred to as the Adviser).*

Example 1: text sample from an NCSR report

Besides using the previously integrated lexical resources and semantic rules, we implemented some dedicated semantic rules that address the current FEII challenge task. These rules assign special features to lexical elements and syntactic dependencies that play a role in the relevance categorization. This work is based on explanations provided by expert in their ratings, as well as evidence in the corpus of annotated data. Thus we designed special relevance features targeting:

- named Entities (person, location and organization names),
- dates,
- monetary expressions,
- financial terms (service provider functions, roles in the company, investment-related terms),
- lexical units related to litigation.

We also included in the set of semantic features dependencies that include relevant lexical units and dependencies that are likely to be characteristic of some criteria of relevance: e.g. the dependency that consists of a present perfect verb whose subject is the filing company is likely to indicate "a change from the status quo or current situation", as specified in the description of highly relevant segments, since such structures may refer to a recent action of the company. For instance, in example 2 from the annotated dataset, segments in bold are marked by the rules as indicators of relevance:

> *At **December 31, 2012**, for loans originated between **2004** and **2008**, the unpaid principal balance of loans related to unresolved monoline **repurchase** claims was **$2.4 billion**, substantially all of which we **have reviewed and declined to repurchase** based on an assessment of whether a material breach exists.*

Example 2: text sample from the training dataset

This semantic information is a good added-value to help decision making especially when the size of the training corpus is small. It provides additional information. However due to the nature of texts within the corpus, only taking into account this dimension would have impact negatively on the recall. We therefore used this as an additional input for our classifiers.

3.4 Inference Models

As previously stated, the use of syntactic and semantic features is generally an added-value for precision but it fails to address recall and more specifically when sentences are complex, not always syntactically correct and/or with long distance dependencies. Therefore we decided to use a mixture of input evaluated through various combinations of classifiers. We decided to test several configurations of logistic regression based classifiers and to combine some of them into a layered architecture managed by a meta-classifier trained to take the best decision based on the output of the previous layers.

Different sets of features have been considered and tested to measure their impact on the predication quality: These sets of features are:

- the surface form of the bag of words coming from the dataset,
- the lemmatized form of the bag of words coming from the dataset,
- the name of the mentioned financial entity,
- the normalized role of this entity,
- syntactic features,
- semantic features.

We tested several configurations of classifiers using all or partial categories of features. We made several cross-validation evaluations as discussed in section 4 and then used a filtering to keep those providing the highest impact.

3.4.1 Classifier 4 categories. The first obvious approach was to use a 4 categories classifier trained to categorize texts among: highly relevant, relevant, neutral and irrelevant classes.

We trained 3 different logistic regression classifiers, with a L2 regularization. The first classifier utilizes all information as input, features, dependencies, surface forms and lemmas. The second one utilizes only dependencies and their features, the last one only surface forms and lemmas.

3.4.2 Binary Classifiers. Then, we trained four binary classifiers, for each of the three feature combinations (12 in total). Each classifier is a variant of "one against all".

3.4.3 Meta classifier. We then applied the 15 classifiers back to the original input vectors and used the output to build a new vector into a last classifier. The main idea is to use this classifier as a way to detect, which of all these sub-classifiers benefit the most to the final decision.

4 PRELIMINARY EXPERIMENTS AND FEATURE FACTORING

Prior to any training of our system we decided to refine the training dataset. We decided to keep triples with one expert annotation label only, or triplets with several labels only if there were an agreement among experts. This allowed us to produce a set of 816 triples assumed to contain no disagreement with respect to labelling.

Once our gold standard was defined, we designed for each classifier combinations a set of experiments varying selected

features and evaluating impact using a 20-fold cross-validation. Then we performed a feature filtering taking into account the real impact on results for each of the 4 targeted categories (highly relevant, relevant, neutral and Irrelevant). We examined the data produced by each classifier and ranked each feature according to its positive or negative relevance to the results to identify the most effective threshold. .

4.1 About Data Impact

With respect to the objective of validating the link between a given Mentioned Financial Entity (MFE) and its role with respect to the Reporting Entity (RE), a deeper analysis on annotations and comments provided by experts revealed that some triples were scored as relevant or even highly relevant despite the fact that associated comments indicate a non-valid link. We considered removing such triples from our gold standard, but this drives down the size of our training set to 663 triples, which is quite small especially considering the fact that these roles are distributed among 10 sub categories. This impacts negatively the number of annotated triples per category.

Then we considered moving these "incorrect" labels (with respect to the MFE-role-RE relationship) to the "Irrelevant" category in our training set. But our experiments demonstrated that it introduces significant noise in the prediction, as elements taken into account by experts to rank relevance seems to be orthogonal to the notion of validity for MFE-Role-RE relationship. Therefore we decided to train our system using the 816 triples data set considering that the most important aspect is the information about relevance provided by experts.

4.2 About Feature Impact and Feature Factoring

Once our gold standard was defined we designed for each classifier combination (4 categories classifier only, set of binary classifiers only and combination of all of them) a set of experiments varying selected features and evaluating impact using a 20-fold cross-validation. We also used weight variations per categories of features (BOW + lemma, syntactic features, semantic features and MFE + role) to evaluate the impact on precision.

Our initial intuition was that due to the small size of the dataset, syntactic and semantic features would have a strong impact on result quality, but in fact a configuration using syntactic and semantic features only produces scores 4.5 % lower compared to a configuration using BOW and lemma only. On the opposite direction, adding altogether BOW + lemma, syntactic features, semantic features and MFE-Role relation tends to improve the overall results by 4% compared to the BOW + lemma only configuration.

Also, a deeper analysis of the cross-validation results indicates that syntactic and semantic features introduce more variance and tend to provide more capabilities to the model. Therefore, we decided to keep this "all-categories of features" configuration and then fine tune our model hyper-parameters. To do so we performed feature filtering taking into account the real impact on results for each of the 4 targeted categories (Highly relevant, Relevant, Neutral and Irrelevant). We looked at results produced by each classifier, and ranked each feature according to its positive or negative relevance to results.

We considered to train and tune a dedicated classification configuration for each of the 10 sub categories of role, but samples in the training dataset are not heterogeneously distributed among these sub categories. The 3 main sub categories of role represent 76% of all triples in our training set, and the first one only contains 44.5 %. Therefore we decided not to make any distinction among roles and train only one combination of classifiers.

5 EVALUATION AND RESULTS

5.1 Experiment and ranking metrics

We applied on the data set released for the competition, which was composed of 900 rows, the best iteration of our system. The output of our system provides for each line of data a category (highly relevant, relevant, neural or irrelevant) and a confidence score. We combined this information to compute a global ranking score according to the following mapping rules: highly relevant ranging from 0.76 to 1.0, relevant from 0.51 to 0.75, neutral from 0.26 to 0.5 and irrelevant from 0.0 to 0.25. Our results are detailed in the next section. .

5.2 Results

Table 1 details the distribution of triples generated by our system from the FEIII competition dataset. Our results are organized according to highly relevant (H), relevant (R), neutral (N) and irrelevant (I) categories.

Role	H	R	N	I
Total	315	365	192	28
Affiliate	47	42	35	5
Agent	16	9	15	0
Counterparty	24	73	9	2
Guarantor	8	15	5	0
Insurer	7	39	1	0
Issuer	18	38	25	17
Seller	8	40	1	0
Servicer	5	34	18	0
Trustee	165	54	81	4
Underwriter	17	21	2	0

Table 1: categorization summary per role

Table 2 details the dispatch made by experts on the same dataset. This table makes a difference between sets of triples annotated as highly relevant or relevant and triples where the link between MFE, role and RE has also been validated (V). All details about this competition including evaluation metrics and global result discussions are provided in the FEIII report [10].

Label	
Highly Relevant and validating triple [H+V]	149
Highly Relevant (partial or no validation) [H]	160
Relevant and validating triple [R+V]	215
Relevant (partial or no validation) [R]	154
Neutral [N]	142
Irrelevant [I]	80

Table 2: Triples annotated by experts

Considering the 5 NDCG scoring variants (indicated below) used by competition organizers, we got scores summarized in table 3.

1. **gt1**: H+V and H sentences must be ranked higher than all other sentences (R, N, I). Order of R, N, I are ignored. H+V=4; H=3; R+V,R,N, I = 0.
   (gt1_500 only considers the top 500 triples).
2. **gt2**: H+V and H sentences must be ranked higher than R or R+V; R or R+V must be ranked higher than N; N must be ranked higher than I. H+V=4; H=3, R+V=2; R=2; N=1; I = 0.
3. **gt3**: H+V sentences must be ranked higher than all other sentences; order of H, R+V, R, N and I are ignored. H+V=4; H, R+V, R, N, I = 0.
4. **gt4**: H+V sentences must be ranked higher than all other sentences; R+V must be ranked higher than H, R and other sentences. H must be ranked higher than R and other sentences. R, N and I are ignored. H+V=4; H=3; R+V=3.5; R, N, I = 0.
5. **gt5**: H+V sentences must be ranked higher than all other sentences; R+V must be ranked next. Order of H, R, N and I are ignored. H+V=4; R+V=3.5; H, R, N, I = 0

gt1	gt2	gt3	gt4	gt5
0.921	0.9583	0.7383	0.9395	0.806

Table 3: NDCG scores.

With respect to these variants of NDCG scoring methods, the one that is most relevant to the way our system has been designed and trained is gt2 (and to a greater extend gt1). Indeed, as discussed in section 4.1. we did not had enough samples and clear indications to train our system to make a difference between validated and not validated MFE-role-RE relationships for highly relevant and relevant texts. Therefore our system has been trained to promote highly relevant texts first (whatever the link is validated or not), then relevant ones (whatever validated or not), then Neutral and finally Irrelevant ones. With respect to this objective our system ranked very high, close to the maximum with an NDCG score of 0.9583 for gt2, compared to 0.9593.

6 CONCLUSIONS

In this paper we present our contribution to address the 1st task of the 2017 Financial Entity Identification and Information Integration challenge. We used for this work a hybrid system designed for text analytics combining a robust parser generating rich syntactic and semantic features and a layered combination of logistic regression based classifiers. The semantic component leverage previous work done to address information extraction from mutual fund reports (N-CSR). Semantic features have been used in combination with others (bag of words, lemma, and syntactic dependencies) through various configurations to train a set of 15 LR2 classifiers. Final decision was made by an additional meta-classifier. We report in this paper our results on the competition dataset.

Initially our assumption was that syntactic and semantic features could bring a lot for the decision making process especially when the size of the annotated dataset is small, but experiments demonstrated that it was not the case. This seems to be related to the complexity and nature of extracted text paragraphs. Used in isolation these features produce lower scores than when using standard bag of words. But when added to BOW, lemma, and MFE-role features they increase the model capability which can then be fine-tuned through a feature filtering process to reach higher scores.

With respect to the objective of MFE-role-RE link validation, due to the complexity of the text within extracted paragraphs, this task could have benefited from a broader analysis of full document contents in order to take into account both structure and repetitions to support the disambiguation process.

REFERENCES

[1] Bank of England. 2015. One Bank Research Agenda 2015 Available at: http://www.bankofengland.co.uk/research/Documents/onebank/discussion.pdf
[2] T. Loughram, and B. McDonald. 2014. Textual Analysis in Finance and Accounting: A Survey. Available at: http://ssrn.com/abstract=2504147
[3] H. Lee, M. Surdeanu and B. MacCartney. 2014. On the Importance of Text Analysis for Stock Price Prediction. In Proceedings of LREC, pp. 1170-1175.
[4] R. Schumaker. 2010. An analysis of verbs in financial news articles and their impact on stock price. In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. Association for Computational Linguistics.
[5] W. Zhang and S. Skiena. 2010. Trading Strategies to Exploit Blog and News Sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp 375-378.
[6] I.E. Fisher, M.R. Garnsey, and M.E. Hughes. 2016. "Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research." Intelligent Systems in Accounting, Finance and Management.
[7] S. Ait-Mokhtar, J.P. Chanod and C. Roux. 2002. Robustness beyond shallowness: incremental deep parsing. In Natural Language Engineering, 8(2-3):121-144
[8] C. Brun, D. Nicoleta Popa and C. Roux. 2014. Xrce: Hybrid classification for aspect-based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 838–842, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University
[9] C. Brun, J. Perez and C. Roux. 2016. XRCE at SemEval-2016 Task 5: Feedbacked Ensemble Modeling on Syntactico-Semantic Knowledge for Aspect Based Sentiment Analysis. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT, San Diego, USA, June 16-17, 2016

[10] L. Raschid, D. Burdick, M. Flood, J. Grant, J. Langsam, I. Soboroff and E. Zotkina. 2016. Financial entity identification and information integration (FEIII) challenge 2017: The report of the organizing committee. In Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD), 2017.