Automated Sequence Tagging: Applications in Financial Hybrid Systems

Automated Sequence Tagging: Applications in Financial Hybrid Systems

Peter Hampton, Hui Wang, William Blackburn and Zhiwei Lin

Abstract Information released by a company regarding internal financials, governance, events and their reaction to market conditions are all believed to impact the underlying value of the business. This data is important to numerous stakeholders external to the business including the investment industry. In a world of heterogeneous information coupled with the high processing power of machines, software is now managing trade execution and making decisions on behalf of humans. As unstructured information grows, the need to disambiguate structure for machine processing remains an attractive research goal in the financial profession. In this paper we describe an approach for automatic entity classification for reporting based applications that make sense of unstructured data using a hybrid system. We critically evaluate our approach and recommend a future research directions based on the experimental results.

1 Introduction

Human beings regard the ability to produce and understand natural and formal languages as a central product of their intelligence. As consumers and institutions generate new information increased data is made available to for investment decision support, many of such is text-based. This generation of information has increased recently, so much so that profit-oriented organizations are relying on the high processing power of machines for modeling and analysis of natural language to support risk-based decisions [1][2][3]. In turn, stock market investors are burdened with multifaceted quantities of information that are typically delivered in unpredictable formats, from dispersed sources around the world. As a result, Text Processing has become an integral part of next generation technologies and working with unstruc-

P Hampton, H Wang, W Blackburn, Z Lin
Artificial Intelligence and Applications Research Group, Ulster University, BT37 0QB
e-mail: hampton-p1@email.ulster.ac.uk, {h.wang, wt.blackburn, z.lin}@ulster.ac.uk

tured data. In turn, research in this area is telling us more about how humans understand and interact with language [4][5].

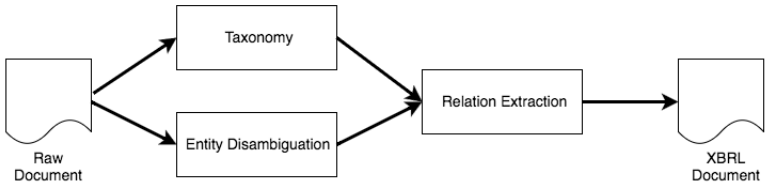

Fig. 1 The overall classification and extraction process. The conversion of a raw document into an XBRL document. The first step takes a raw document and then in parallel identifies the appropriate taxonomy while classifying the entities. Once this is complete, the entities and their relations are extracted to compose the final XBRL document.

Advances in Behavior Based Artificial Intelligence has enabled automatic or semi-automatic collection, extraction, aggregation and categorization of various text based documents [6]. This has undoubtedly had a large impact on how companies store and retrieve information, but further processing and analysis of such information is still an on-going research challenge for academic researchers and industry practitioners alike. Intelligent systems that exhibit understanding of language are becoming more and more important to the decision-making process when buying and selling financial instruments [7][8]. Yet little is published on how to analyse multi-structured documents in this specialist domain.

This paper surveys related works and methods that can be used to extract critical financial information from annual (10-K) accounts. We then build on the previous work using a Hybrid Conditional Random Fields instead of a Maximum Entropy Markov Model (MEMM) [9]. We conclude this paper by discussing the advantages and limitations of such hybrid-graphical models and propose a future research direction for extracting actionable information from financial documents using event disambiguation and relation extraction.

2 Background

In the early days of Artificial Intelligence, academic and industry pundits prophesied and as promised machines just as intelligent as their human counterparts. As of the time of writing, their predictions are yet to fully materialize. One explanation for the delay in machines that exhibit behaviour similar to humans is that we don't yet fully understand human *intelligence*. The sub-field in Artificial Intelligence that tackles the problem of understanding language is known as Natural Language Processing which either take a deep or naive understanding to language. Early practitioners in Artificial Intelligence used rule based methods to interact with and understand language. This restrictive approach led researchers to focus on shallow methods

in statistical machine learning which make heavy use of feature engineering. With increased computing power, unsupervised methods have become an interesting area of research after experiencing their own AI winter [10].

Named Entity Recognition is a type of classification task, which is the process of categorizing a set of n-grams while only using a non-exhaustive list of data features that describe them. Early work in this area found that it is imperative to identify meaningful information units like names, including person, organization and locations, and numeric expressions including time, date, money and percent expressions [11][12][13][14]. To date, practitioners in this area have been using word level features to identify Named Entities which can be seen in Table 1.

One area that is suffering from information overload is the finance and investment industry. New technologies and entire industries are evolving around the idea that is *big data* management and analysis. One area in the research domain that remains relatively untouched is the analysis of financial reports [15]. We analysed a small sample of 6 companies and found that on average between the years of 2011 - 2015 that the average annual report was 97 pages. These documents are information rich, have inconsistent formatting and contain a mix of factual and subjective information from internal company.

Table 1 Natural and Engineered Word-level features

Feature	Type*	Description
Case	Natural	The casing reflects the grammatical function performed by a noun or pronoun in a phrase, clause, or sentence in Engish and some other languages.
Punctuation	Natural	Used in text to separate sentences and their elements. Other punctuation marks are used to clarify meaning.
Digit	Natural	These are numbers in text which can include digit patterns, cardinal and ordinal numbers, roman numbers and words with digits.
Character	Natural	These features tend to be possessive marks or greek letters.
Morphological Feature	Engineered	Morphological features include prefixes, suffexises and stems of words.
Part-of-Speech Tag	Engineered	The category assigned in accordance with a words syntactic function.
Function	Engineered	Typically functions defined by the programmer, such as token / phrase length, case-insensitivity, hand crafted patterns and so on.

* *Natural* types are words that can be identified using morphological functions whereas *Engineered* requires a system to generate the feature in a preprocessing stage.

2.1 Financial Reporting

The aim of this project is to computationally *understand* the language contained within financial reports through Information Extraction as the means. The typical objective of a financial statement is to give an overview of the financial position, financial performance, and cash flows of a business that is useful to a wide range of users, human or machine, in making economic decisions.

Strategies The strategic part of accounts typically holds information such as the Chairmans and Chief Executives professional analysis of the business. This information is typically presented in an unstructured nature.

Governance The governance section of a financial report looks at information such as information about the board of directors, discussions around corporate governance, etc. This information is typically presented in an unstructured nature.

Financials Financial information includes sections as previously discussed such as the cash flow statement, balance sheet, specific types of incomes and outgoings. The information is typically presented in a structured nature, however, the format is unpredictable.

There are various types of information that would be of interest to a decision maker. Some that would be of interest include:

Assets Recorded on the Balance Sheet, the assets section of a report lists anything of value that can be converted into cash. This can include property, stock and materials

Liabilities Recorded on the balance sheet, liabilities include loans, accounts payable, mortgages, deferred revenues and accrued expenses. Liabilities are used to finance various activities and other assets that keep the organization running.

Equity Equity is value of a company less it's liabilities. This can found be subtracting the liabilities from the assets.

Revenues Listed in the Profit and Loss statement, Revenue is the amount of cash that is brought into a company through income

Expenses An expense consists of the economic costs that a business incurs through its operations to earn revenue which is usually listed in the Profit and Loss table.

Gains An increase in the value of an asset or property. A gain arises if the selling or disposition price of the asset is higher than the original purchase or acquisition price.

From the above short descriptions it is reasonable to infer that a financial document such as a 10-K (Annual Statement) or 10-Q (Quarterly Statement) is a long a complex document. Take the following excerpt from the First Derivatives PLC (AIM: FDP.L) annual released in 2016:

Revenue for the year increased by 40.6% to 117.0m (2015: 83.2m), while adjusted EBITDA rose by 50.5% to 23.3m (2015: 15.5m) and adjusted earnings per share increased by 33.2% to 51.7p (2015: 38.8p).

Annotated, this **sentence** would look very different with 3 percentage entities, 6 money entities, 2 year (time based) entities and 2 accounting entities. From this The unstructured information appears to be dense with rich information tightly coupled together in financial reports. Extracting information into predefined ontological templates as previously practiced would be incredibly difficult, if not impossible.

Even with an expert or team of experts, designing a template for the extraction would be a very long and very complicated task. Using the shallow parsing described in most Named Entity Classification experiments would also prove incredibly challenging due to the heterogeneous markup of the document ranging from free-flowing sentences to tabular formatted data, with different table formats [6].

Initial work in this area comes from the work of the EU Musings project which experimented with Ontology based Information Extraction as a means to creating business intelligence from static documents. The areas their research focused on where financial risk management, internationalization and IT operational risk management. The end result is a populated ontology (Knowledge base) that can be queried. Although this is semantically guided, our approach does not take a ontological approach. Trough our use of Ontology Based information Extraction, we found the basis of the ontology to be far too restrictive although it proves to be very successful in narrow applications. Our approach deviates from their approach by leveraging the XBRL taxonomy as an ontology, discarding relations for atomic facts, or terms.

2.2 Semantic Representation

From a computational standpoint (and the end goal of this project), this is a very challenging task regardless of the documents integrity. Such reasons include:

- Entites can have a non-numerical relation to another entity
- Some facts might relate to non-financial facts.
- Entities have fidelity
- There are many namespaces to choose from.

XBRL (eXtensible Business Reporting Language) is a relatively new standard for exchanging business information in an XML like format. XBRL allows the expression of semantic meaning commonly required in financial reporting. It is believed XBRL is a suitable way to communicate and exchange business information between different business systems [16]. These communications are defined by metadata set out in taxonomies, which capture the definition of individual reporting

concepts as well as the relationships between concepts and other semantic meaning. Information being communicated or exchanged is provided within an XBRL instance.One use of XBRL is to define and exchange financial information, such as a financial statement. An short snippet of XBRL is displayed below for easy conceptualization.

```
<!-- revenue for the years 2007 / 08 -->
<us-gaap:OperatingRevenue contextRef="FY08Q1"
unitRef="USD">1376200000</us-gaap:OperatingRevenue>
<us-gaap:OperatingRevenue contextRef="FY07Q1"
unitRef="USD">1081100000</us-gaap:OperatingRevenue>

<!-- Cost of revenue for the years 2007 / 08 -->
<us-gaap:CostOfRevenue contextRef="FY08Q1"
unitRef="USD">2675000000</us-gaap:OperatingRevenue>
<us-gaap:CostOfRevenue contextRef="FY07Q1"
unitRef="USD">1696000000</us-gaap:OperatingRevenue>
```

The goal is eventually to be able to detect the appropriate XBRL taxonomy for a given document, and use the XBRL namespace entities as the filling template for the extracted information of the document. This paper focuses on identifying entities, rather than extracting them. The complete process however is discussed in detail in Section 7.

3 Methodologies

Information Extraction typically involves several stages: Preprocessing, Named Entity Recognition (NER), Relation Extraction and Template Filling [17]. Named Entity Recognition is the process of identifying n-grams of interest that belong to some predefined category of interest and is what we are most interested in at this point of the study.

Information Extraction is an eclectic field that focuses on extracting structure from heterogeneous documents. Typical approaches are to directly lift information from a Natural Language document and fill a template. With the continued evolution of the World Wide Web into areas such as offline-first and the Internet of Things (IoT) is creating more potential application areas for Information Extraction. Although the AI sub-field of Natural Language Processing typically focuses on unstructured data, extracting information from structured and semi-structured sources can still prove a very challenging research goal.

Common approaches include human designed algorithmic and statistically driven methodology in Information Extraction. Both approaches exhibit promising results and strengths that have inspired recent uses described in Section 3.3. [18] identify

information extraction on several different dimensions: deterministic to stochastic, domain-specific to general, rule based to learned, and narrow to broad test situations. In this section we survey the methods adopted by other researchers in this field.

3.1 Rule Based

Rule based approaches have been the most popular approach in industry and are largely considered obsolete in academic research. Chiticariu et al (2013) noted that only 6 publications relied solely on rules across a 10-year period reflecting the popularity for machine learning. This reflects the growing popularity of statistical techniques.

Table 2 List Based Features

List Type	Description
Gazetteer	A generalized list of words such as stopwords, swearwords.
Entity List	Lists of organization names, people names and countries.
Entity Cues	Cues that include things such as name prefixes, post-nominal letters, the location of a typical word.

It is believed that manually constructing such IE systems is a laborious and error prone task. This has led researchers to believe that a more appropriate solution would be to implement machine learning instead of human defined rules.

Another domain that heavily uses gazetteers in Named Entity Recognition tasks includes the BioMathematics and BioInformaitics (BMI). The approach outlined by [19] uses a similar approach to this paper implementing a Conditional Random Fields with a dictionary to extract chemical names, a research topic that very much remains virgin territory.

3.2 Statistical

Machine Learning approaches to Named Entity Recognition are currently the most prominent form of research in academia, whereas rule based approaches are favored in industry. [20] believe that the domain specific knowledge and system expertise needed for this task makes the area a prime target for machine learning. The idea is to let machines self discover and learn from large corpora; annotated and real time text data, say from company websites. However, IBMs industrial survey found that only one third of industry relied exclusively on Machine Learning techniques for extraction products [21]. Here we briefly review the two of the most successful

machine learning methods to date in the area of Named Entity Recognition: Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) [22].

3.2.1 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states and have been used liberally in research with sequential and temporal data which tends to serve as a strong baseline. Hidden Markov Models are famous for playing a big role in the Human Genome Project and various speech recognition studies before Deep Neural Networks became the state of the art [23][24].

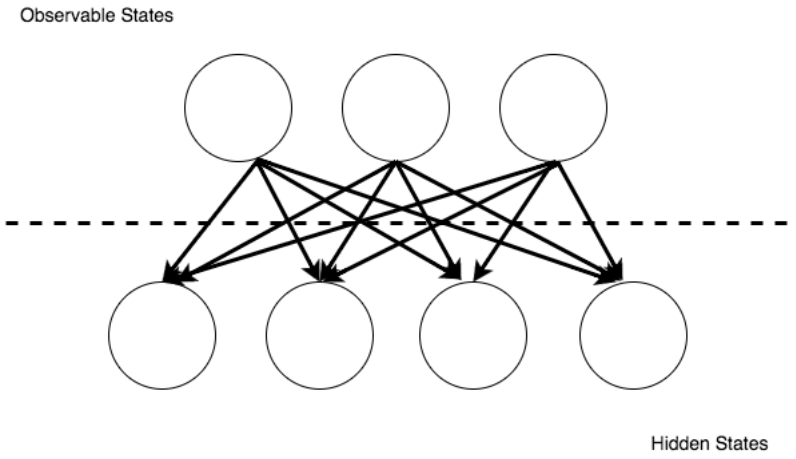

Fig. 2 A simple depiction of a Hidden Markov Model (HMM).

$$z_1..z_n \varepsilon \{1,..,n\} \tag{1}$$

$$x_1,...,x_n \varepsilon \chi \tag{2}$$

$$p(x_1..x_n, z_1..z_n) = p(z_1)p(x_1|z_1)\prod_{k=2}^{n} p(z_k|z_{k-1})p(x_k|z_k) \tag{3}$$

(1) Shows the hidden states whereas (2) shows the depicts the observed data. The joint distribution (3) of all the following factors in the following way can be easily conceptualized in fig 2.

3.2.2 Conditional Random Fields

When examining the results of various Named Entity Recognition competitions such as CoNLL and Biocreative it's clear that are Conditional Random Fields or similar conditional taggers are the state of the art. [25] introduced this sequence modelling framework that has all the advantages of a Hidden Markov Model and manages to overcome the labelling bias problem. The authors describe their creation as a *finite state model with normalized transition probabilities* that is trained by maximum likelihood estimation.

3.3 Future Methods

Although Named Entity Recognition systems do not attempt a holistic understanding of human language, are still time-consuming to build and generally contain highly domain-specific components and expert bias, making the models very difficult to port to new application areas. This is troublesome in this project because different financial jurisdictions use different terminology and therefore different XBRL namespaces. In this section we briefly review methodologies believed to rapidly evolve over the coming years and overcome various limitations in rigid rules and statistical estimators.

3.3.1 Hybrid Systems

Hybrid Systems are dynamic systems that can exhibit both continuous and discrete dynamic behavior a system that can both flow (described by a differential equation) and jump (described by a difference equation or control graph). Hybrid systems are an attractive alternative for Named Entity Recognition in a Financial setting, as it is easily adoptable, scalable, and has transparent debugging features compared to machine learning. A common approach is to complement a black-box, machine learning based approach with rules to handle exceptions, employ a hybrid approach where rules are used to correct classifier mistakes in order to increase the precision of the overall system. These researchers use rules to cover cases that the system cannot yet handle.

3.3.2 Human Guidance

Human guided models (or *Human-in-the-loop* is defined as a model that requires human interaction on some level of the training and re-training process as depicted in Fig 4. Human guidance is considered a future method of working with machine learning models because the processes are typically black boxes, and human influ-

ence can prove to be an important contribution to the overall performance of the models [26].

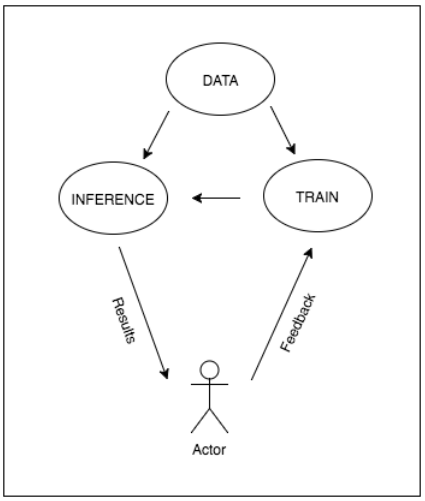

Fig. 3 Human Guidance systems are systems used to give feedback to a classifier, offering corrections to the outputs of the system.

3.3.3 Unsupervised Approaches

Although there is a variety of Neural Network topologies and implementations, few have proved very powerful for the task of language understanding. The purpose of a Neural Network is to take a set of inputs, perform various complex computations to them and then use the output to solve some sort of predefined problem. They are highly structured and made up of layers, typically a input layer, hidden layer and output layer.

Recurrent Neural Networks have proven to be very effective in some text classifications tasks including NER studies [27][28]. Although these new studies with unsupervised methods are gaining traction, at the time of writing they are yet to still beat the performance of heavily engineered CRFs. It has also been noted that organizations struggle to implement various aspects of machine learning algorithms and prefer machine transition models such as a rule base / gazetteer.

4 Architecture

We propose implementing a rule-based, recursive pattern matching module alongside a machine-learning algorithm to improve the overall results for NER. We expect that this will yield significantly better results for entities that are typically expressed numerically in text such as monetary values, percentages, dates and times based on the research.

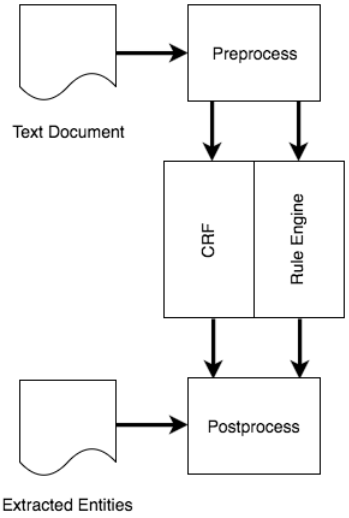

Fig. 4 The text document is sent for preprocessing (OCR extraction, tokenization, Part-of-Speech tagged before being classified by the CRF and Rule base in parallel. Once both models are finished, the document entities are merged and the annotated document is outputted.

5 Experiments

When evaluating Information Extraction tasks, researchers tend to use the Precision, Recall and F-Score metrics and a train / test split. There are 4 states which we can separate the extracted output into; *True Positive, False Positive, False Negative, False Positive*. Precision and Recall is a measure to understand the relevance of extracted text.

- **True Positive**: Selected and Correctly Classified.
- **True Negative**: Not Selected and Not Correct
- **False Positive**: Selected, but incorrect.
- **False Negative**: Correct, but not selected.

The precision is the instances classified that are correct. That is the true positives divided by the sum of the true positives plus the false positives.

$$Precision = \frac{tp}{tp + fp} \tag{4}$$

$$Recall = \frac{tp}{tp + fn} \tag{5}$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6}$$

The recall, or sensitivity in this case, is the fraction of the entities that were classified as relevant.

Table 3 The result of the Hybrid CRF.

	Hyrbid MEMM*	Hybrid CRF
F-Score	77.86%	84.81%

* The baseline used in the initial experiments.

6 Discussion

Although conditional models such as CRFs are unbiased (typically), and high variance. Quite the opposite is true for the rule base which exhibit high amounts of bias and little variance. It can be concluded that the method presented in this paper is good at trading of Bias / Variance where appropriate by combining the two methods. Further, Lafferty (2003) describe in their paper that CRFs generalize easily to analogues of stochastic free-grammars which could prove an interesting research direction.When developing and evaluating the Named Entity Recognition part of our work, we found the following advantages and disadvantages in our hybrid implementation:

Advantages:

1. The rules enhance the performance of the CRF.
2. It is simple to maintain, and retrain if needed.
3. The rules are portable.
4. The rules can be expanded by any novice programmer with limited proficancy in data structures and regular expressions.

Disadvantages:

1. CRFs require tedious and laborious feature engineering to outperform a Hidden Markov Model.

2. a typical training method makes training and model inference extremely computationally intensive.
3. And finally, very high model complexity and variance.
4. Portability is nearly impossible because of the training set and hand-tuned features.

To conclude, researchers focusing on the automatic generation of XBRL documents have struggled to deal with the imprecise and ambiguous nature of natural languages. Many have relied on semi-structured documents which are fairly easy to build a semantic web service to interpret and act on these documents is arguably not a research challenge, rather an engineering pursuit. Others have used an ontological approach, hard wiring expert-guided knowledge to guide the extraction process. Recent works have focused on merging methods together such as the work of Hampton et al (2015), where deterministic rules are used to gauge the predictable nature of certain entities such as dates and money, and introduce stochastic methods to handle imperfections in written text.

7 Further Research

Further research is needed on this topic to progess to a relation extraction stage. Halevy et al (2009) promotes representation that can use unsupervised learning on unlabeled data, and represent the data with a non-parametric model. This may be wise and a strong ground for inspiration for all future work in this area. However, the case for hybrid systems shouldn't be discarded or ignored. Because the XBRL taxonomy is a human designed discrete space, a predefined template influenced by various economic and sociological factors such as tax jurisdiction, turnover, formation type, and the use of hybrid characteristics or reinforcement learning. One possible further direction would be to train a model on large amounts of unlabeled but mapped data from financial accounts as depicted in Fig 5 .

References

1. Li, Nan, and Desheng Dash Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forecast." Decision support systems 48.2 (2010): 354-368.
2. Loughran, Tim, and Bill McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10Ks." The Journal of Finance 66.1 (2011): 35-65.
3. Andriy Bodnaruk, Tim Loughran and Bill McDonald, 2015, Using 10-K Text to Gauge Financial Constraints, Journal of Financial and Quantitative Analysis, 50:4.
4. Fisher, Ingrid E., Margaret R. Garnsey, and Mark E. Hughes. "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 (2016).
5. Hirschberg, Julia, and Christopher D. Manning. "Advances in natural language processing." Science 349.6245 (2015): 261-266.

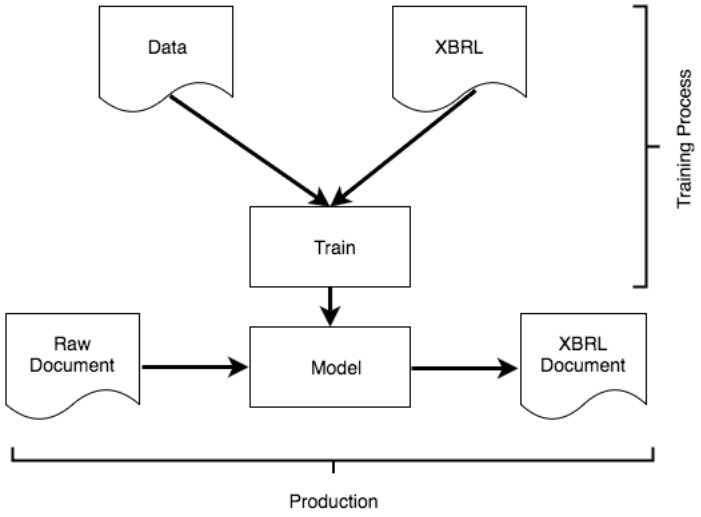

Fig. 5 The overall (simplified) desired pipeline of extracting entities into a predefined document. The training process takes a pair of raw accounts (data points) and XBRL documents (target) and fits their together together to create a model. In the production process, a raw document can be delivered to the model and outputs an XBRL document.

6. Saggion, Horacio, et al. Ontology-based information extraction for business intelligence. Springer Berlin Heidelberg, 2007.
7. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.
8. Kearney, Colm, and Sha Liu. "Textual sentiment in finance: A survey of methods and models." International Review of Financial Analysis 33 (2014): 171-185.
9. Hampton, Peter John, Hui Wang, and William Blackburn. "A Hybrid Ensemble for Classifying and Repurposing Financial Entities." Research and Development in Intelligent Systems XXXII. Springer International Publishing, 2015. 197-202.
10. Burr, Geoffrey W., et al. "Experimental demonstration and tolerancing of a large-scale neural network (165 000 Synapses) using phase-change memory as the synaptic weight element." Electron Devices, IEEE Transactions on 62.11 (2015): 3498-3507.
11. Shaalan, Khaled. "A survey of arabic named entity recognition and classification." Computational Linguistics 40.2 (2014): 469-510.
12. Marrero, Mnica, et al. "Named entity recognition: fallacies, challenges and opportunities." Computer Standards Interfaces 35.5 (2013): 482-489.
13. Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.
14. Reeve, Lawrence, and Hyoil Han. "Survey of semantic annotation platforms." Proceedings of the 2005 ACM symposium on Applied computing. ACM, 2005.
15. Loughran, T., McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research.
16. Wisniewski, Tomasz Piotr, and Liafisu Sina Yekini. "Stock market returns and the content of annual report narratives." Accounting Forum. Vol. 39. No. 4. Elsevier, 2015.
17. Troshani, I., Parker, L. D., Lymer, A. (2015). Institutionalising XBRL for financial reporting: resorting to regulation. Accounting and Business Research, 45(2), 196-228.
18. Burdick, Douglas, et al. "Extracting, linking and integrating data from public sources: A financial case study." Available at SSRN (2015).

19. Rao, Delip, Paul McNamee, and Mark Dredze. ”Entity linking: Finding extracted entities in a knowledge base.” Multi-source, Multilingual Information Extraction and Summarization. Springer Berlin Heidelberg, 2013. 93-115.
20. Russell, Stuart, Peter Norvig, and Artificial Intelligence. ”A modern approach.” Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs 25 (2013): 27.
21. Rocktschel, Tim, Michael Weidlich, and Ulf Leser. ”ChemSpot: a hybrid system for chemical named entity recognition.” Bioinformatics 28.12 (2012): 1633-1640.
22. Califf, Mary Elaine, and Raymond J. Mooney. ”Bottom-up relational learning of pattern matching rules for information extraction.” The Journal of Machine Learning Research 4 (2003): 177-210.
23. Chiticariu, Laura, Yunyao Li, and Frederick R. Reiss. ”Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!.” EMNLP. No. October. 2013.
24. Nadeau, David, and Satoshi Sekine. ”A survey of named entity recognition and classification.” Lingvisticae Investigationes 30.1 (2007): 3-26.
25. Morwal, Sudha, Nusrat Jahan, and Deepti Chopra. ”Named entity recognition using hidden Markov model (HMM).” International Journal on Natural Language Computing (IJNLC) 1.4 (2012): 15-23.
26. Morwal, Sudha, and Deepti Chopra. ”NERHMM: A Tool For Named Entity Recognition based on Hidden Markov Model.” International Journal on Natural Language Computing (IJNLC) 2 (2013): 43-49.
27. Lafferty, John, Andrew McCallum, and Fernando CN Pereira. ”Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” (2001).
28. Clancy, Seamus, Sam Bayer, and Robyn Kozierok. Active Learning with a Human In The Loop. No. MTR120603. MITRE CORP BEDFORD MA, 2012.
29. Lample, Guillaume, et al. ”Neural architectures for named entity recognition.” arXiv preprint arXiv:1603.01360 (2016).
30. Yao, Kaisheng, et al. ”Recurrent neural networks for language understanding.” INTERSPEECH. 2013.