Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment

Julio Cesar Salinas Alvarado **Karin Verspoor** **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
Australia
jsalinas@student.unimelb.edu.au
karin.verspoor@unimelb.edu.au
tb@ldwin.net

Abstract

Risk assessment is a crucial activity for financial institutions because it helps them to determine the amount of capital they should hold to assure their stability. Flawed risk assessment models could return erroneous results that trigger a misuse of capital by banks and in the worst case, their collapse. Robust models need large amounts of data to return accurate predictions, the source of which is text-based financial documents. Currently, bank staff extract the relevant data by hand, but the task is expensive and time-consuming. This paper explores a machine learning approach for information extraction of credit risk attributes from financial documents, modelling the task as a named-entity recognition problem. Generally, statistical approaches require labelled data for learn the models, however the annotation task is expensive and tedious. We propose a solution for domain adaption for NER based on out-of-domain data, coupled with a small amount of in-domain data. We also developed a financial NER dataset from publicly-available financial documents.

1 Introduction

In the years 2007–2008, the GFC (Global Financial Crisis) affected a vast number of countries around the world, causing losses of around USD$33 trillion and the collapse of big-name banks (Clarke, 2010). Experts identified that one of the main causes of the GFC was the use of poor financial models in risk assessment (Clarke, 2010; news.com.au, 2010; Debelle, 2009).

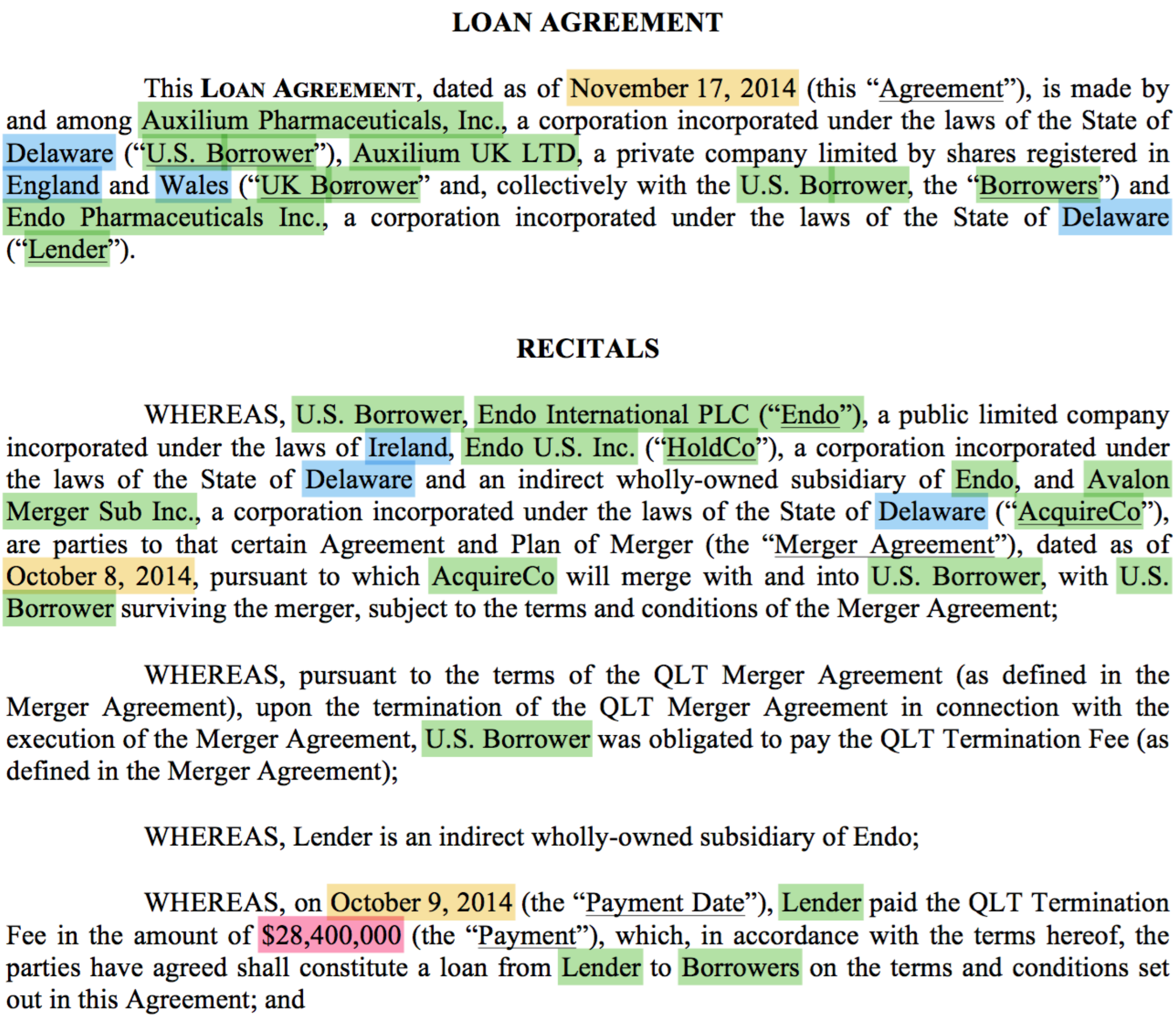
Risk assessment helps banks to estimate the amount of capital they should keep at hand to promote their stability and at the same time to protect their clients. Poor risk assessment models tend to overestimate the capital required, leading banks to make inefficient use of their capital, or underestimate the capital required, which could lead to banks collapsing in a financial crisis.

Financial documents such as contracts and loan agreements provide the information required to perform the risk assessment. These texts hold relevant details that feed into the assessment process, including: the purpose of the agreement, amount of loan, and value of collateral. Figure 1 provides a publicly available example of a loan agreement, as would be used in risk assessment.

Currently, bank staff manually extract the information from such financial documents, but the task is expensive and time-consuming for three main reasons: (1) all documents are in unstructured, textual form; (2) the volume of "live" documents is large, numbering in the millions of documents for a large bank; and (3) banks are continuously adding new information to the risk models, meaning that they potentially need to extract new fields from old documents they have previously analyzed.

Natural language processing (NLP) potentially offers the means to semi-automatically extract information required for risk assessment, in the form of named entity recognition (NER) over fields of interest in the financial documents. However, while we want to use supervised NER models, we also want to obviate the need for large-scale annotation of financial documents. The primary focus of this paper is how to build supervised NER models to extract information from financial agreements based on pre-existing out-of-domain data with partially-matching labelled data, and small amounts of in-domain data.

There are few public datasets in the financial domain, due to the privacy and commercial value of the data. In the interest of furthering research on information extraction in the financial domain, we



LOAN AGREEMENT

This LOAN AGREEMENT, dated as of November 17, 2014 (this "Agreement"), is made by and among Auxilium Pharmaceuticals, Inc., a corporation incorporated under the laws of the State of Delaware ("U.S. Borrower"), Auxilium UK LTD, a private company limited by shares registered in England and Wales ("UK Borrower" and, collectively with the U.S. Borrower, the "Borrowers") and Endo Pharmaceuticals Inc., a corporation incorporated under the laws of the State of Delaware ("Lender").

RECITALS

WHEREAS, U.S. Borrower, Endo International PLC ("Endo"), a public limited company incorporated under the laws of Ireland, Endo U.S. Inc. ("HoldCo"), a corporation incorporated under the laws of the State of Delaware and an indirect wholly-owned subsidiary of Endo, and Avalon Merger Sub Inc., a corporation incorporated under the laws of the State of Delaware ("AcquireCo"), are parties to that certain Agreement and Plan of Merger (the "Merger Agreement"), dated as of October 8, 2014, pursuant to which AcquireCo will merge with and into U.S. Borrower, with U.S. Borrower surviving the merger, subject to the terms and conditions of the Merger Agreement;

WHEREAS, pursuant to the terms of the QLT Merger Agreement (as defined in the Merger Agreement), upon the termination of the QLT Merger Agreement in connection with the execution of the Merger Agreement, U.S. Borrower was obligated to pay the QLT Termination Fee (as defined in the Merger Agreement);

WHEREAS, Lender is an indirect wholly-owned subsidiary of Endo;

WHEREAS, on October 9, 2014 (the "Payment Date"), Lender paid the QLT Termination Fee in the amount of $28,400,000 (the "Payment"), which, in accordance with the terms hereof, the parties have agreed shall constitute a loan from Lender to Borrowers on the terms and conditions set out in this Agreement; and

Figure 1: Example of a loan agreement. Relevant information that is used by risk assessment models is highlighted. The example is taken from a loan agreement that has been disclosed as part of an SEC hearing, available at `http://www.sec.gov/Archives/edgar/data/1593034/000119312514414745/d817818dex101.htm`

construct an annotated dataset of public-domain financial agreements, and use this as the basis of our experiments.

This paper describes an approach for domain adaption that includes a small amount of target domain data into the source domain data. The results obtained encourage the use of this approach in cases where the amount of target data is minimal.

2 Related Work

Most prior approaches to information extraction in the financial domain make use of rule-based methods. Farmakiotou et al. (2000) extract entities from financial news using grammar rules and gazetteers. This rule-based approach obtained 95% accuracy overall, at a precision and recall of 78.75%. Neither the number of documents in the corpus nor the number of annotated samples used in the work is mentioned, but the number of words in the corpus is 30,000 words for training and 140,000 for testing. The approach involved the creation of rules by hand; this is a time-consuming task, and the overall recall is low compared to other extraction methods.

Another rule-based approach was proposed by Sheikh and Conlon (2012) for extracting information from financial data (combined quarterly reports from companies and financial news) with the aim of assisting in investment decision-making.

The rules were based on features including exact word match, part-of-speech tags, orthographic features, and domain-specific features. After creating a set of rules from annotated examples, they tried to generalize the rules using a greedy search algorithm and also the Tabu Search algorithm. They obtained the best performance of 91.1% precision and 83.6% recall using the Tabu Search algorithm.

The approach of Farmakiotou et al. (2000) is similar to our approach in that they tried to address an NER problem with financial data. However, their data came from financial news rather than the financial agreements, as targeted in our work. The focus of Sheikh and Conlon (2012) is closer to that in this paper, in that they make use of both financial news and corporate quarterly reports. However, their extraction task does not consider financial contracts, which is the key characteristic of our problem setting.

Somewhat further afield — but related in the sense that financial agreements stipulate the legal terms of a financial arrangement — is work on information extraction in the legal domain. Moens et al. (1999) used information extraction to obtain relevant details from Belgian criminal records with the aim of generating abstracts from them. The approach takes advantage of discourse analysis to find the structure of the text and linguistic forms, and then creates text grammars. Finally, the approach uses a parser to process the document content. Although the authors do not present results, they argue that when applied to a test set of 1,000 criminal cases, they were able to identify the required information.

In order to reduce the need for annotation, we explore domain adaptation of an information extraction system using out-of-domain data and a small amount of in-domain data. Domain adaptation for named entity recognition techniques has been explored widely in recent years. For instance, Jiang and Zhai (2006) approached the problem by generalizing features across the source and target domain to way avoid overfitting. Mohit and Hwa (2005) proposed a semi-supervised method combining a naive Bayes classifier with the EM algorithm, applied to features extracted from a parser, and showed that the method is robust over novel data. Blitzer et al. (2006) induced a correspondence between features from a source and target domain based on structural correspondence learning over unlabelled target domain data. **?**) showed that a graph transformer NER model trained over word embeddings is more robust cross-domain than a model based on simple lexical features.

Our approach is based on large amounts of labelled data from a source domain and small amounts of labelled data from the target domain (i.e. financial agreements), drawing inspiration from previous research that has shown that using a modest amount of labelled in-domain data to perform transfer learning can substantially improve classifier accuracy (**?**).

3 Background

Named entity recognition (NER) is the task of identifying and classifying token-level instances of named entities (NEs), in the form of proper names and acronyms of persons, places or organizations, as well as dates and numeric expressions in text (Cunningham, 2005; Abramowicz and Piskorski, 2003; Sarawagi, 2008). In the financial domain, example NE types are LENDER, BORROWER, AMOUNT, and DATE.

We build our supervised NER models using conditional random fields (CRFs), a popular approach to sequence classification (Lafferty et al., 2001; Blunsom, 2007). CRFs model the conditional probability $p(s|o)$ of labels (states) s given the observations o as in Equation 1, where t is the index of words in observation sequence o, each k is a feature, w_k is the weight associated with the feature k, and $Z_w(o)$ is a normalization constant.

$$p(s|o) = \frac{\exp(\sum_t \sum_k w_k f_k(s_{t-1}, s_t, o, t))}{Z_w(o)} \quad (1)$$

4 Methods

4.1 Data

In order to evaluate NER over financial agreements, we annotated a dataset of financial agreements made public through U.S. Security and Exchange Commission (SEC) filings. Eight documents (totalling 54,256 words) were randomly selected for manual annotation, based on the four NE types provided in the CoNLL-2003 dataset: LOCATION (LOC), ORGANISATION (ORG), PERSON (PER), and MISCELLANEOUS (MISC). The annotation was carried out using the Brat annotation tool (Stenetorp et al., 2012). All documents were pre-tokenised and part-of-speech (POS) tagged using NLTK (Bird et al., 2009). As part of the annotation, we automatically tagged all instances of the tokens *lender* and *borrower* as being of entity type PER. We have made this dataset available in CoNLL format for research purposes at: `http://people.eng.unimelb.edu.au/tbaldwin/resources/finance-sec/`.

For the training set, we use the CoNLL-2003 English data, which is based on Reuters newswire data and includes part-of-speech and chunk tags (Tjong Kim Sang and De Meulder, 2003).

The eight financial agreements were partitioned into two subsets of five and three documents, which we name "FIN5" and "FIN3", respectively. The former is used as training data, while is used exclusively for testing.

Table 1 summarizes the corpora.

4.2 Features

For all experiments, we used the CRF++ toolkit (Kudo, 2013), with the following feature set (optimized over the CoNLL-2003 development set):

- Word features: the word itself; whether the word starts with an upper case letter; whether the word has any upper case letters other than the first letter; whether the word contains digits or punctuation symbols; whether the word has hyphens; whether the word is all lower or upper case.
- Word shape features: a transformation of the word, changing upper case letters to *X*, lower case letters to *x*, digits to *0* and symbols to *#*.
- Penn part-of-speech (POS) tag.
- Stem and lemma.
- Suffixes and Prefixes of length 1 and 2.

4.3 Experimental Setup and Results

We first trained and tested directly on the CoNLL-2003 data, resulting in a model with a precision of 0.833, recall of 0.824 and F1-score of 0.829 (**Experiment1**), competitive with the start-of-the-art for the task.

The next step was to experiment with the financial data. For that, first we applied the CoNLL-2003 model directly to FIN3. Then, in order to improve the results for the domain adaption, we trained a new model using the CONLL +FIN5 data set, and test this model against the FIN3 dataset.

A summary of the experimental results over the financial data sets is presented in Table 2.

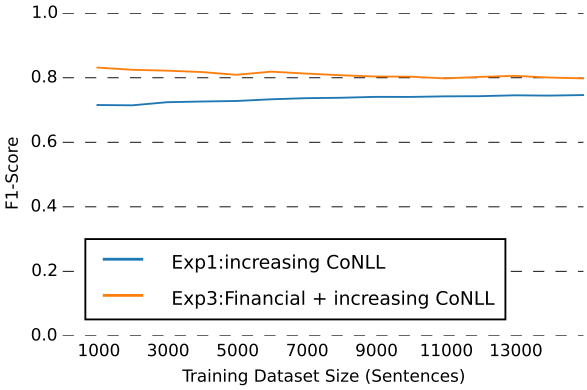


Figure 2: Learning curves of the performance of Experiment5 (over FIN3), starting with the FIN5 point data and incrementally adding CONLL data. Also, it shows Experiment1, Experiment 2 and Experiment 3 just with increasing CoNLL data.

5 Discussion

Table 2 summarizes the results of directly applying the model obtained by training only over out-of-domain data to the two financial data sets. The difference in the domain composition of the CONLL data (news) and the financial documents can be observed in these results. With out-of-domain test data, a precision of 0.247 and a recall of 0.132 (**Experiment2**) was observed, while testing with in-domain data achieved a precision of 0.833 and recall of 0.824 (**Experiment1**).

As a solution to the difference in the nature of the sources in the context of limited annotated in-domain data, we experimented with simple domain adaptation, by including into the source domain (CONLL) data a small amount of the target domain data — i.e. including data from FIN5— generating a new training data set (CONLL +FIN5). When trained over this combined data set, the results increased substantially, obtaining a precision of 0.828, recall of 0.770 and F-score of 0.798 (**Experiment3**).

As additional analysis, in Figure 2, we plot learning curves based on F-score obtained for Experiment2 and Experiment3 as we increase the training set (in terms of the number of sentences). We can see that the F-score increases slightly with increasing amounts of pure CONLL data (Experiment2), but that in the case of the mixed training data (Experiment3), the results actually drop as we add more CONLL data.

Figure 3 shows the learning curves for Experi-

Name	Description
CoNLL	CoNLL-2003 training data
CoNLL$_{\text{test}}$	CoNLL-2003 test data
CoNLL +Fin5	CoNLL-2003 training data + five financial agreements
Fin5	Five financial agreements
Fin3	Three financial agreements

Table 1: Description of the data sets used.

Name	Training Data	Test Data	P	R	F1
Experiment1	CoNLL	CoNLL$_{\text{test}}$	0.833	0.824	0.829
Experiment2	CoNLL	Fin3	0.247	0.132	0.172
Experiment3	CoNLL +Fin5	Fin3	0.828	0.770	0.798
Experiment4	Fin5	Fin3	0.944	0.736	0.827

Table 2: Results of testing over the financial data sets.

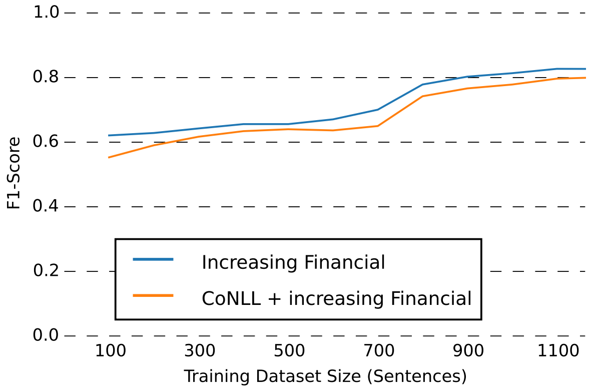


Figure 3: Learning curves showing the F-score as more financial training data is added for Experiment3 and Experiment 4.

ment3 and Experiment4, as we add more financial data. Here, in the case of Experiment3, we start out with all of the CoNLL data, and incrementally add Fin5. We can see that the more financial data we add, the more the F-score improves, with a remarkably constant absolute difference in F-score between the two experiments for the same amount of in-domain data. That is, even for as little as 100 training sentences, the CoNLL data degrades the overall F-score.

Confusion matrices for the results of the predictions of **Experiment3** are shown in Table 3.

Analysis of the errors in the confusion matrix reveals that the entity type MISC has perfect recall over the financial dataset. Following MISC, PER is the entity type with the next best recall, at over 0.9. However, generally the model tends to suffer from a high rate of false positives for the entities LOC and ORG, affecting the precision of those classes and the overall performance of the model.

One interesting example of error in the output of the model is when the tokens refer to an address. One example is the case of *40 Williams Street*, where the correct label is LOC but the model predicts the first token (*40*) to be NANE and the other two tokens to be an instance of PER (i.e. *Williams Street* is predicted to be a person).

In the model generated with just the CoNLL data, one notable pattern is consistent false positives on tokens with initial capital letters; for example, the model predicts both *Credit Extensions* and *Repayment Period* to be instances of ORG, though in the gold standard they don't belong to any entity type. This error was reduced drastically through the addition of the in-domain financial data in training, improving the overall performance of the model.

Ultimately, the purely in-domain training stratagem in Experiment4 outperforms the mixed data setup (Experiment3), indicating that domain context is critical for the task. Having said that, the results of our study inform the broader question of out-of-domain applicability of NER models. Furthermore, they point to the value of even a small amount of in-domain training data (**?**).

6 Conclusions

Risk assessment is a crucial task for financial institutions such as banks because it helps to estimate the amount of capital they should hold to promote their stability and protect their clients. Manual extraction of relevant information from text-

		Predicted					
		LOC	MISC	ORG	PER	O	Recall
Actual	LOC	**20**	0	3	2	14	0.513
	MISC	0	**7**	0	0	0	1.000
	ORG	0	0	**16**	0	40	0.286
	PER	0	0	0	**202**	14	0.935
	NANE	12	2	24	8	–	
	Precision	0.625	0.778	0.372	0.953		

Table 3: Confusion matrix for the predictions over FIN3 using the model from Experiment3, including the precision and recall for each class ("NANE" = Not a Named Entity).

based financial documents is expensive and time-consuming.

We explored a machine learning approach that modelled the extraction task as a named entity recognition task. We used a publicly available non-financial dataset as well as a small number of annotated publicly available financial documents. We used a conditional random field (CRF) to label entities. The training process was based on data from CoNLL-2003 which had annotations for the entity types PER (person), MISC (miscellaneous), ORG (organization) and LOC (location). We then assembled a collection of publicly-available loan agreements, and manually annotated them, to serve as training and test data. Our experimental results showed that, for this task and our proposed approach, small amounts of in-domain training data are superior to large amounts of out-of-domain training data, and furthermore that supplementing the in-domain training data with out-of-domain data is actually detrimental to overall performance.

In future work, we intend to test this approach using different datasets with an expanded set of entity types specific to credit risk assessment, such as values and dates. An additional step would be carry out extrinsic evaluation of the output of the model in an actual credit risk assessment scenario. As part of this, we could attempt to identify additional features for risk assessment, beyond what is required by the financial authorities.

References

Witold Abramowicz and Jakub Piskorski. 2003. Information Extraction from Free-Text Business. *Effective databases for text & document management*, page 12.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Sebastopol, USA.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip Blunsom. 2007. *Structured classification for multilingual natural language processing*. Ph.D. thesis, University of Melbourne Melbourne, Australia.

Thomas Clarke. 2010. Recurring crises in anglo-american corporate governance. *Contributions to political Economy*, 29(1):9–32.

Hamish Cunningham. 2005. Information extraction, automatic. *Encyclopedia of language and linguistics,*, pages 665–677.

Guy Debelle. 2009. Some effects of the global financial crisis on australian financial markets. `http://www.rba.gov.au/speeches/2009/sp-ag-310309.html`.

Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pages 75–78. Citeseer.

Jing Jiang and ChengXiang Zhai. 2006. Exploiting domain structure for named entity recognition. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 74–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

Taku Kudo. 2013. Crf++: Yet another crf toolkit. *Software available at http://wing. comp. nus. edu. sg/ forecite /services /parscit -100401 /crfpp/ CRF++-0.51 / doc/*. Accessed: 2015-05-26.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, USA.

Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. 1999. Information extraction from legal texts: the potential of discourse analysis. *International Journal of Human-Computer Studies*, 51(6):1155–1171.

Behrang Mohit and Rebecca Hwa. 2005. Syntax-based semi-supervised named entity tagging. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

news.com.au. 2010. Poor risk assessment 'led to global financial crisis'. `http://www.news.com.au/finance/business/poor-risk-assessment-led-to-global-financial-crisis/story-e6frfkur-1225912209942`.

Sunita Sarawagi. 2008. Information Extraction. *Found. Trends databases*, 1(3):261–377, March.

Mahmudul Sheikh and Sumali Conlon. 2012. A rule-based system to extract financial information. *Journal of Computer Information Systems*, 52(4):10–19.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.