The Minho Quotation Resource

Brett Drury and J.J. Almeida

LIAAD-INESC, University of Minho,
Rua de Ceuta, 118, 6, Gualtar Campus,
4050-190 Porto,Portugal, Braga, Portugal
Brett.Drury@gmail.com jj@di.uminho.pt

Abstract

Direct quotations from business leaders can provide a rich sample of language which is in common use in the world of commerce. This language used by business leaders often uses: metaphors, euphemisms, slang, obscenities and invented words. In addition the business lexicon is dynamic because new words or terms will gain popularity with businessmen whilst obsolete words will exit their common vocabulary. In addition to being a rich source of language direct quotations from business leaders can have "real world" consequences. For example, Gerald Ratner nearly bankrupted his company with an infamous candid comment at an Institute of Directors meeting in 1993. Currently, there is no "direct quotations from business leaders" resource freely available to the research community. The "Minho Quotation Resource" (MQR) captures the business lexicon with in excess of 500,000 quotations from individuals from the business world. The quotations were captured from October 2009 and April 2011. In this article we describe the MQR corpus and its main characteristics as well as few example experiments. The resource is available in a searchable Lucene index and will be available for download in May 2012.

Keywords: Quotations, Corpus, Business

1. Introduction

The "Minho Quotation Resource" is a detailed linguistic resource which provides the researcher with a comprehensive sample of direct quotations from the business world during the period of 2009 - 2011. The business lexicon can reflect the role of an "economic actor" or the real world pressures he is subject to(Drury, Dias, and Torgo, 2011). For example, analysts may use a more objective tone whereas a CEO may use richer rhetorical language (Drury, Dias, and Torgo, 2011). In addition quotations can have real-world consequences. For example: a series of "colourful" quotations from a speech given by Gerald Ratner at the Institute of Directors in 1991 nearly bankrupted his company. The incident has entered the British lexicon where "doing a Ratner" is a euphemism for committing a serious error or "gaff".

At the current time there is no linguistic resource which incorporates the language found in direct quotations from the business world. The "Minho Quotation Resource" is an attempt to provide the research community with a comprehensive resource which captured over 500,000 quotations over a two year period. The resource is a Lucene index which has the following information: 1.Quote Text, 2. Speaker Name, 3. Job Role, 4. Speaker's Affiliation, 5. Dependency Tree, 6. POS Tagged Quote, 7. Business Sectors in Quote and 8. Companies in Quote Text. The remainder of the paper will discuss: data acquisition, an overview of the resource, linguistic characteristics of quotations and experiments which try to compute the real world consequences of a quote.

1.1. Data Acquisition

The quotations were captured from freely available news sources from the Internet. A web crawler was constructed to crawl a pre-defined list of Really Simple Syndication (RSS) feeds. The RSS feeds provided links to a news story as well as a publication date. The news story text was extracted with an unsupervised text extraction strategy. The news story text was sent to the Open Calais Web Service. Open Calais is a web service which adds "meta data" to textual information. Meta data provided by Open Calais can consist of a number of separate items: for example, classification or named entity extraction. The meta data also includes quote detection and quote assignment. Quote assignment is the process of assigning a quote to a person. The meta data is described in Resource Description Format (RDF). The meta data was processed to extract the quote, the person who made the quote and on occasion the company / organization the person is related to and his job title.

2. Overview of Quotation Resource

The "Minho Quotation Resource" is essentially a Lucene index (Gospodnetic and Hatcher, 2009). Lucene is a open source "full text search" library available from Apache. Lucene is document based where a single document consists of a collection of fields. The "Minho Quotation Resource" uses one document to represent a single quotation, and the fields to represent a different aspect of the quote. The fields are indexed and consequently an individual may search for a given term in a specific field. The layout of a Lucene document in the Minho Quotation Resource can be found in Table 2.. The following text will describe each field in detail.

The **ID Field** is an identification number for the document and does not contain any textual information. The **Quotemaker Field** is the name of the person making the quote. There was no normalization of the person's name and consequently a person may have several entries, for example Bill Gates may be listed William Gates or B. Gates. The format of the persons name will depend upon the source it was extracted from. The **Job Title Field** contains the job title held by the person making the quote. The job title has not been normalized and consequently a quote

maker may have one or more variants of the same job title. The **Quote Company Field** is the name of the company or organization which employs the quote make. This field has not been normalized and therefore a quote maker may have one or more variants of the company name assigned to him. The **Quote Field** contains the actual quote. The **Date Field** contains the published date of the news story from which the quote was extracted. This information was extracted from the original field. The date is expressed in milliseconds and conforms to the milliseconds specification for the Date/Time class in Java. The **POS Field** is the quote with part of speech (POS) tags assigned to each word in the quote. The **Company Field** is the name of the companies mentioned in the quote. The name of the companies in the quote were identified with the ANNIE Gazetteer (Cunningham and Tablan, 2002) which is part of the larger GATE application. The ANNIE Gazetteer is a series of lists which hold a term and a label. The term is a named entity, for example, a name of a company or a person and the label describes the named entity. The **Sector Field** contains sectors which are directly mentioned in the quote. A sector is a term which describes a group of related companies, for example, "the technology industry". The industry sectors were identified by the strategy described by (Drury and Almeida, 2010). The **Tree Field** is a dependency tree representation of the quote text. The dependency tree information conforms to the Stanford Dependency format (Marneffe, MacCartney, and Manning, 2006).

Field	Content
ID	Document Identifier
quotemaker	Name of Person Making the Quote
jobrole	Job title of person making the quote
quotecompany	Name of employer
quote	Quote text
date	Date the quote was made
pos	Pos tagged quote text
company	Names of companies in the quote text
sector	Business sectors in the quote text
tree	Parse tree of quote text

Table 1: Document Fields and Content

2.1. Miscellaneous Resources

In addition to the Lucene index the "Minho Quotation Resource" has some miscellaneous resources to assist the researcher. A possible use of this resource was is to classify quotes into sentiment categories, consequently the resource offers precomputed sentiment bigrams. The sentiment bigrams were generated with the strategy described by (Liu, 2009). The bigrams were pairs of words with a sentiment score which was either positive or negative. A sample of bigrams is demonstrated in Table 2.

The next miscellaneous resource is words or phrases highlighted by journalists who wrote the original article the quotes was extracted from. Journalists occasionally add annotations to words or phrases in a quote which are of interest. The annotations were either single quotes (') or double quotes (""). For example:'Porsche has only settled "substantially below 1 per cent" and hopes to remain under 5 per cent.'. In the example quotation the journalist has highlighted the phrase "substantially below 1 per cent". A sentiment orientation can be computed for the phrase by locating an anchor words which have pre-known sentiment orientation, for example the word 'good' has a positive connotation. The sentiment orientation was computed by comparing the frequency of the term co-occurring with a positive anchor word with the frequency of the term co-occurring with a negative anchor word. The technique is described in full by (Liu, 2009). A sample of terms located by this method is displayed in Table 3.

Category	Bigrams
Positive	strong relationships, welcome news, positive response, global reach, profitable business
Negative	wrong time, cold water, painful recession, unfair competition, sad day

Table 2: Sample Bigrams and Category

Category	Bigrams
Positive	pivotal role, advanced negotiations, very very good, thorough and comprehensive, much better chance
Negative	mixed bag, shockingly bad, put the brakes, unclear, jobless recovery

Table 3: Sample "Special Terms" and Category

The last miscellaneous resource is "sentiment discourse connectors". A discourse connector can join a person with the quote, for example "Mr Smith said", the discourse connector is "said". A discourse connector may have a predetermined sentiment orientation which may determined the sentiment orientation of the following quote. For example "Mr Smith warned", the connector "warned" would determine the following quote was negative. The sentiment orientation of the connectors was computed with the aforementioned anchor technique (Liu, 2009). A sample of sentiment connectors is displayed in Table 4.

Category	Bigrams
Positive	understood, ensures, believes, believed, found
Negative	worried, warned, warns, denied, admitted

Table 4: Sample "Discourse Connectors" and Category

2.2. Tools

The authors have assumed that researchers will integrate the Lucene index into their own applications. To assist users a simple command line application is supplied with the resource which allows the searching and exporting of information from the Lucene index. The user sets the name

of the field and the search term in a properties file. The application reads the property file and exports the matching documents to a XML file.
The web site supporting this resource has links to tools which may assist the researcher using this resource. The web site links to the Tregex GUI tool from the Stanford Parser Tool collection (Levy and Andrew, 2006). The Tregex tools allows the manipulation of the parse trees found in the Tree Field in the Lucene index. The second tool is the Luke Lucene Index Toolbox (http://code.google.com/p/luke/) which allows the manipulation of Lucene indexes.

3. Linguistic Characteristics

The resource represents direct quotations from the business world which is a rich source of language. The language used in the quotations are not governed by strict rules of language or grammar, but represents the everyday use of language in the business world. The language used in this resource can range thoughtfully constructed quotes designed to achieve a certain goal to the crude and ill thought through abusive comments. The section will describe a number of linguistic phenomena discovered in the quote collection. The section will discuss the use of metaphor, stock phrases and informal language.

3.1. Metaphor

A metaphor is a linguistic device which uses a tangible item such as story or image to represent a less tangible item. A metaphor can be used as a rhetorical device to manipulate a target audience, for example shareholders. The associated work to this resource revealed that certain groups represented in this corpus use rhetorical language (Drury, Dias, and Torgo, 2011).
The metaphors identified in this corpus can often be culturally dependent. Quotes from US based commentators often rely upon US specific sports(basketball, baseball, ..) whereas British or Commonwealth commentators will refer to global sports (football, cricket, rugby). A sample of metaphors found in this resource can be found in Table 3.1.. The snapshot shows the typical use of metaphors in business communication. This may allow the person making the quote to communicate abstract ideas to a wider audience.

Quote	Metaphor Domain	Metaphor
They added tremendous value to the project. We threw **curveball** after curveball at them and they **hit it out of the park** each time.	Sports	Curveball, hit it out of the park (Baseball)
Our view is that (the put option) is a **home run** for the company and its shareholders	Sports	Home Run (Baseball)
The economy has gone from being in a freefall and is now **on the road** to recovery	Travel	On the road
We're clearly **digging out of a bigger hole**,	Construction	Digging out of a bigger hole
The biggest issue with compensation at financial firms is that it was like paying people before the **roulette wheel stopped spinning**;	Gambling	roulette wheel stopped spinning
trading futures is emotionally identical to **playing poker**.	Gambling	playing poker

Table 5: Metaphors Domains

3.2. Stock Phrases

The quotes in the resource revealed a number of phrases or terms were repeated used to describe regular events in the business calendar. For example, announcing earnings results or management transition. The terms were used to excess which rendered these phrases meaningless. For example: companies are always "excited" or "pleased" by the addition of a new member of staff and the new member of staff is always "excited" to be working at the company. The may have implications for event detection where stock language could be used in a linguistic pattern to detect a regular business event such as a product launch or earnings announcement. In addition stock language could inhibit sentiment analysis because stock language tends to be composed of words which have positive "connotations" such as pleased or excited. A traditional sentiment approach of accumulating polarity values of words may label these quotes as "positive" instead of neutral as quotes which use stock language are essentially meaningless (Drury, Dias, and Torgo, 2011). A small sample of stock language is displayed in Table 3.2..

Quote	Stock Phrase
I am excited to be continuing my work with Zuora	I am excited
I am excited by the wealth of opportunities in front of us	I am excited
We are pleased to announce the well-deserved promotions of Conan and Kean	We are pleased

Table 6: Examples of Stock Language

3.3. Informal Language

The quotations in this resource were reported verbatim with no interpretation from the journalist reporting the quote. Although the people making the quote often thought through or rehearsed the quote there were many times people made a quote spontaneously or "off message". These

quotes could contain informal language or even obscenities if the person making the quote was "under stress". A definition of informal language was given by Ferguson (Ferguson, 1959) who described informal language as: "spontaneous speech in situations that may be described as natural or real-life" which uses "a low dialect or language in preference to a high one".
The identification of informal language can be assisted by journalists adding notation to a quote. As described in the miscellaneous section of this paper journalists can often put quotes (' or ") around unique words and phrases. For example: " It was: the sales went "phwarr" ". The uniqueness of the word "phwarr" is highlighted by the double quotes. Frequency analysis may assist for the identification of informal language because informal language has no language standards and consequently the words used may be unique and hence infrequent.
The use of informal language may assist in the communication of a message. An example of this was discovered in the Enron scandal were two traders were recorded speaking and they communicated the same message using informal language and then the same message using more formal language. The informal version of the quote was: "He just f—s California. He steals money from California to the tune of about a million". The second version of the quote was: " he arbitrages the California market to the tune of a million bucks or two a day " (Roberts, 2007). It is arguable that the first quote communicates more information to the "lay person" than the second which is couched in industry jargon. The use of informal language and in particular obscenities can be used to detect if a CEO is lying (Larcker and Zakolyukina, 2010). Samples of informal language contained in the resource is shown in Table 3.3..

Quote	Informal Word	Freq.
F**k my victims; I carried them for 20 years	f**k	(12)
Don't screw it up, buddy.	screw	(78)
I was gobsmacked how many people signed up straight away.	gobsmacked	(10)
My vice is speaking to f**kers like you.	f**kers	(1)
Zug is slightly the arse end of nowhere[1]	arse end	(1)

Table 7: Informal Language

4. Experiments

The paper thus far has described the linguistic richness and the lexicon of the quotes in the resource, however the resource also contains information which can be used to infer real world events, for example stock prices, and therefore quotations can be an indicator of the future financial performance of a company. An example of this phenomena is demonstrated by a series of comments made by Juniper Networks (JNPR) CEO Kevin Johson in a presentation made on the 1st June 2011. The quotations are documented below:

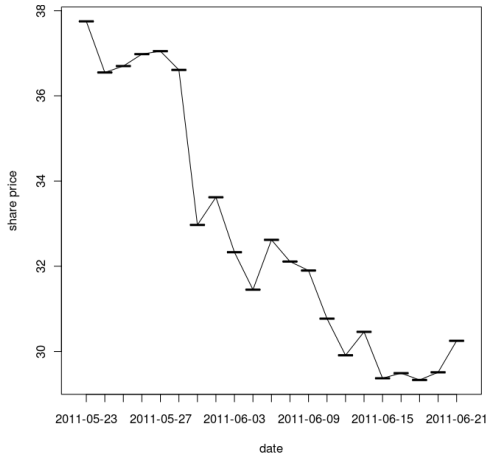

Figure 1: Juniper Networks Share Price

> *look at our Q2, and Q2 is traditionally back-end loaded and I think the linearity of this Q2 would reinforce a backend loaded quarter and were staying focused on that. But, I think theres reason to be somewhat cautious in the near-term but yet maintaining optimism about the long-term growth models that weve outlined and the demand for networking long-term.*
>
> *Were not immune to the focus the government has on budgets and spending*

The share price for Juipter Networks is described in Figure 1. The chart shows the share price before the comments by the CEO and afterwards. There is a sharp dive in the price of the share when the comments made by Kevin Johson from 36.50 to 32.97. A month later the shares were trading at 31.97. On the 27th July the company announced low profits and the share price dropped to by 18%[2].
The Juipter Networks example provides some evidence that quotes from business leaders may be an indicator a specific economic event, for example earnings announcements. The experiments used quotes from the resource to predict the value of a market index. The following text will describe the experimental setup and results.

4.1. Experimental Setup

The experiments sought to predict the direction of the NASDAQ. The NASDAQ is a market index which represents the market value of a group of companies. A value of an index is measured in points. An increase in points represents a gain in value of the group of companies the index represents whereas a drop value is reflected by a loss of points.
The experiments split the resource into two: one part for training and the other for testing. The training data was selected by choosing 300 random days from the training corpus and using the quotes from these days to induce models

[2]http://goo.gl/yRf4w Market Watch consulted September 2011

from a learner. The test data was selected by choosing 100 test days from the test portion of the resource. This was repeated 5 times in process known as Monte Carlo Simulation which attempts to reduce any bias from any individual selection (Glasserman, 2004).

The classification models were induced from quotes using the strategy described by (Drury, Dias, and Torgo, 2011). The models were induced by separating the quotes by the "role" of the person making the quote. The roles were determined by their employment title. There were two roles: "biased" and "independent". Biased model is induced from quotes made by people with a direct connection to a company, for example CEO,CFO or CIO. The independent model is induced from quotes by people whose job roles oblige them to be independent, for example Analyst or Researcher. Two separate strategies were used to induce each role based model. A dictionary approach was used for the independent model where as a semi supervised clustering approach was used for the biased model. A full description of the approach is provided by (Drury, Dias, and Torgo, 2011).

4.2. Evaluation

There were two evaluation measures: trading accuracy and trading profit. The trading accuracy is a number expressed as a percentage which represents the frequency a trading action was correct. A correct decision by a model was decision which correctly predicted the direction of the market. The trading profit was the profit realised by trading actions initiation by the model.

The trading evaluation used quotations which were published when the market was closed. The trading period was from the previous days closing price to the next days opening price. The model predicted the market value was either: higher than the day before or lower than the day before. A successful trade was determined to be either: 1. model predicted "sell" and the opening price was less than the previous days closing price or 2. model predicted "buy" and the opening price was higher than the previous days closing price.

The experiments considered two configurations: classifier confidence and decision boundary. The decision boundary is the difference between the number of "positive quotations" and "negative quotations" before a trading decision is triggered. A decision boundary of 0 infers that a simple majority would be enough to trigger a decision whereas the upper decision boundary of 90 infers that a majority of at least 90 quotes is required before a trading decision is triggered.

4.3. Results

The strategy produced a mean average trading accuracy of 54% and a trading profit of 86.97 points. The optimal configurations returned a trading accuracy of 59%. and a trading profit of 233.81 points. This is higher than expected by chance and comparable to existing systems (M.A.Mittermayer and G.F.Knolmaye, 2006).

The relationship between the decision boundary, classifier confidence and trading accuracy/profit are displayed in Figures 2 and 3. There seems to be no relationship between classifier confidence and trading accuracy/profit, but there seems to be a relationship between decision boundary and trading accuracy/profit. The lower the boundary the higher the trading profit/accuracy. This may due to the habit of journalists providing a counterbalance to the majority view in a news story by quoting a person with an opposite view.

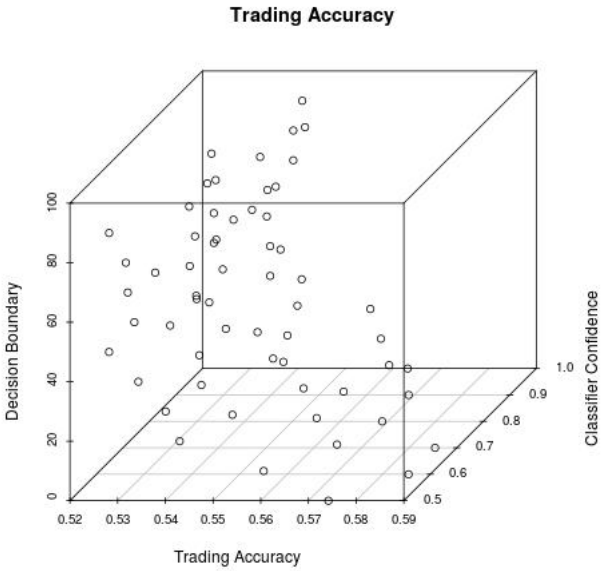

Figure 2: Trading Accuracy by Classifier Confidence / Decision Boundary

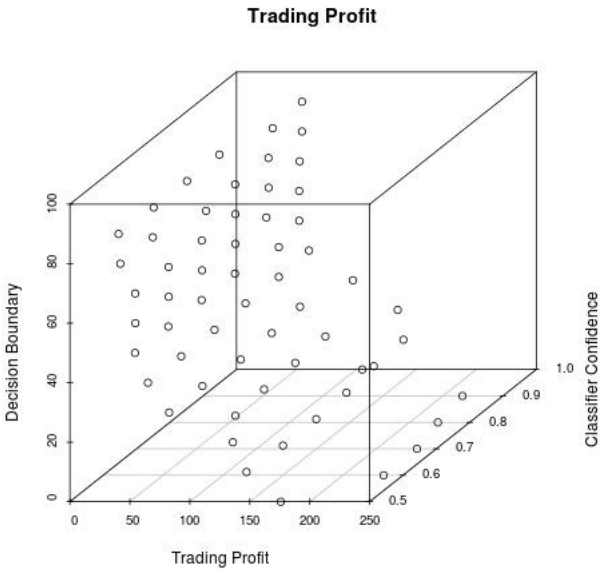

Figure 3: Trading Profit by Classifier Confidence / Decision Boundary

The results indicate that quotations can be used successfully in a trading strategy and the authors are using this work in combination with other strategies they have developed (Drury, Torgo, and Almeida, 2012),(Drury and

Almeida, 2011) to create a more sophisticated trading system.

5. Conclusion

The Minho Quotation Resource is a comprehensive linguistic resource which contains a significant number of quotations which covered the period of 2008 to 2011. The resource represents a rich lexicon derived from the business world. The resource is designed to assist the researcher with routine linguistic tasks such as discourse or sentiment analysis. In addition this resource may be used in "real world" applications such as trading which was highlighted in the experiments section. The resource will be released under the Creative Commons license and is located at http://goo.gl/U6quN.

6. References

Cunningham, Maynard, Bontcheva and Tablan. 2002. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Drury, Brett and Jose Joao Almeida. 2010. Identification, extraction and population of collective named entities from business news. In *Entity2010 – Workshop on Resources and Evaluation for Entity Resolution and Entity Management*.

Drury, Brett and José João Almeida. 2011. Identification of fine grained feature based event and sentiment phrases from business news stories. In *WIMS*.

Drury, Brett, Gael Dias, and Luis Torgo. 2011. A contextual classification strategy for polarity analysis of direct quotations from financial news. In *RANLP 2011 Conference Proceedings*.

Drury, Brett, Luís Torgo, and José João Almeida. 2012. Classifying news stories with a constrained learning strategy to estimate the direction of a market index. *IJCSA*, 9(1):1–22.

Ferguson, Charles. 1959. Diglossia. *Word : Journal of the Linguistic Circle of New York*.

Glasserman, Paul. 2004. *Monte Carlo methods in financial engineering*. Applications of mathematics. Springer.

Gospodnetic, Otis and McCandless Hatcher. 2009. *Lucene in Action*. Manning Publications.

Larcker, David and Anastasia Zakolyukina. 2010. Detecting deceptive discussions in conference calls.

Levy, Roger and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*.

Liu, Bing. 2009. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer.

M.A.Mittermayer and G.F.Knolmaye. 2006. Text mining systems for market response to news: A survey. Technical report, University of Bern.

Marneffe, Marie, Bill MacCartney, and Christopher Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*.

Roberts, Joel. 2007. Enron traders caught on tape.