An Update to the Minho Quotation Resource

Brett Drury

LIAAD-INESC-Tec, Campus da FEUP
Rua Dr. Roberto Frias, 4200 - 465 Porto, Portugal
brett.drury@gmail.com

Samuel Morais Drury

Colégio Puríssimo, R. Sete, 881 - Centro
Rio Claro - SP
Brazil

April 15, 2021

Abstract

The Minho Quotation Resource was originally released in 2012. It provided approximately 500,000 quotes from business leaders, analysts and politicians that spanned the period from 2008 to 2012. The original resource had several failings which include a large number of missing job titles and affiliations as well as unnormalised job titles which produced a large variation in spellings and formats of the same employment position. Also, there were numerous duplicate posts. This update has standardised the job title text as well as the imputation of missing job titles and affiliations. Duplicate quotes have been deleted. This update also provides some metaphor and simile extraction as well as an emotion distribution of the quotes. This update has also replaced an antiquated version of Lucene index with a JSONL format as well as a rudimentary interface that can query the data supplied with the resource. It is hoped that this update will encourage the study of business communication in a time of a financial crisis.

1 Introduction

The Minho Quotation Resource is a linguistic resource that contains 500,000 quotes from business and government leaders as well as analysts and economists that covers the period 2008 to 2012 which we released in 2012 [1]. The period covered by this resource is an economically interesting time because it is the start of the recovery from the financial recession of 2007, and contains quotes
related to the economically significant events such as the bailouts of General Motors and Freddie Mae and Mac.

The original resource used Open Calais to extract reported speech contained in business news. Open Calais returned a speaker name, and if available a job title and an organisation as well as a quote. An example of a complete relation is Ben Bernanke, Chairman, US Federal Reserve, “Considering some alternative forms for the government-sponsored enterprises - or for mortgage securitisation generally during this time out seems worthwhile.” In this example, the Speaker is Ben Bernanke who held the post of Chairman at the US Federal Reserve.

In the original resource, there was no preprocessing of the data as well as no information extraction of significant words and phrases from the quote. The original release stored data in a Lucene Index and a CSV file. The resource now suffers from several issues and shortcomings. For example, the Lucene Index supplied with the resource is antiquated because it uses the Lucene format from 2012. In addition to the ageing format, there are some issues with the data itself such as missing job titles and affiliations as well as variations in speakers’ names. And finally, the extractor has conflated affiliation with the job role.

These project notes describe a significant update of The Minho Quotation Resource. This update: imputes missing job titles and affiliations, metaphors and similes are extracted, the storage format is updated and an emotion distribution is estimated for each quote.

These project notes will adhere to the following format: Corpus Characteristics, Phrase and Word Extraction, Resource Structure and Conclusion.

2 Corpus Characteristics

The original release of The Minho Quotation Resource was a selection of 500,000 quotes from business news, and it spans the period 2008 to 2012. It has several issues which are demonstrated by a simple frequency analysis which is shown in Table 1. The frequency analysis counted the number of entries for the most frequent Speakers, Job Roles and Affiliation. The most frequent speaker is Barack Obama with 5979 attributed quotes. There are 131860 individual speakers in the resource. The most frequent Job Role is President with 9536 attributed quotes and the most frequent affiliation is Chief Executive with 5783 attributed quotes.

The frequency analysis demonstrates an error with the original resource. The error is the transposition of the Job Role with Affiliation. The underlying relation extraction has erroneously identified a job role as an affiliation which is shown by the most frequent affiliations. There are also a large number of speakers, job roles and affiliations which may be caused by non-normalised speaker name, job role and affiliation.
2.1 Corpus Preprocessing

It is clear from the previous section that a preprocessing step was required to remove the transposition error, reduce speaker name and job title ambiguity, remove duplicated and non-English sentences, and impute missing job titles and affiliations.

The transposition error was resolved by matching the entries in the affiliation category with the entries in the job title category and if they matched the affiliation term was moved to the job title category. For example, in Table 1 it is clear that Chief Executive is present in both the Affiliation and Job Role category, consequently any Chief Executive entry in the Affiliation category would be moved to the Job Role category.

The speaker name ambiguity was reduced by changing all names to title case. Therefore the entry “barack Obama” would be changed to “Barack Obama”. Manual rules changed entries of people who are known by a single name to their complete name. For example, the entry “Obama” was expanded to “Barack Obama”.

The final step in name normalisation was to expand single names to similar full names who have the same affiliation. These individuals are not known by a single name. For example, Dick Fuld who was the CEO of Lehman Brothers was often referred to as Fuld. If the Quote where he was referred to as Fuld had his affiliation as Lehman Brothers, then his name would be expanded to Dick Fuld. If the quote had his affiliation missing then his name would remain the same. This is the case for the Quote with the ID 341819.

The job title ambiguity was reduced by contracting three-word job titles to an acronym such as “Chief Operating Office” to “COO”. Also, there were manual rules that transformed known job titles that had less than three words to an acronym. For example, the job role of Chief Executive would be transformed to CEO. Compound job titles were simplified to one job title by splitting the job title at the conjunction and selecting the job title on the left-hand side. For example, the job title “President and CEO” would be simplified to “President”.

Missing job titles were imputed by matching a candidate name and affiliation with an entry that has the corresponding name and affiliation as well as a job title. The job title is then transferred to the entry with the candidate name and affiliation. For example, if there was an entry for “Dick Fuld” with the

| Attribute | Most Frequent | Number of Entries |
|-----------|---------------|------------------|
| Speaker   | B.Obama       | G.Brown 131860   |
|           | 5979          | 1783 1498        |
| Job Role  | President     | Chief Executive CEO 14050 |
|           | 9536          | 9064 8131        |
| Affiliation | Chief Executive | President Minister 65004 |
|           | 5783          | 3362 2073        |

Table 1: Count of Attribute Entries
affiliation “Lehman Brothers”, but no job role, then he would be assigned to the job title “CEO” because the majority of Quotes which had the name “Dick Fuld” and affiliation “Lehman Brothers” had a job role of “CEO”.

The same process was followed for missing affiliations, but the matching was by using name and job title rather than name and affiliation. For example, using the “Dick Fuld” example, an entry that had “Dick Fuld” and the job role “CEO” would be assigned the affiliation “Lehman Brothers” because the majority of entries the name “Dick Fuld” and the job role “CEO” have the affiliation “Lehman Brothers”.

Some entries were duplicated because of the repetition of quotes reported in different media sources. The removal duplicate quotes which was achieved by matching speaker name and quotes. If more there was more than one quote with the same speaker name, then one quote would be removed. In the much used Dick Fuld example, if there were more than one entry with the following details: Dick Fuld CEO Lehman Brothers You know what, let them line up; I can handle it., one of them would be deleted.

There were quotes not reported in English. These quotes were identified by detecting the language of the quote with the LangDetect Python Library. Any quotes identified by the library as non-English were removed.

The final step of the preprocessing stage was the unescaping of HTML entities and the removal of HTML from the quotes. Text which is marked up in HTML often escaped and the escaped text is then rendered by the browser. For example, &lt; is the escaped form of <. The unescaping returned the escaped character to its original form. Therefore the quote text is represented uniformly. Any remaining HTML was removed with the Beautiful Soup Python Library as well as manual rules. Any quotes which were empty of this process were removed.

The results of the preprocessing are shown in Table 2, and it is clear from the Table that the amount of different total entries for each attribute have been reduced, and that the most frequent affiliations now match that of the most frequent speakers.

| Attribute   | Most Frequent         | Number of Entries |
|-------------|-----------------------|-------------------|
| Speaker     | B.Obama               | G.Brown           | B.Bernanke       | 130942 |
|             | 6524                  | 1685              | 1256             |       |
| Job Role    | CEO                   | President         | Chairman         | 5358  |
|             | 69175                 | 38155             | 18427            |       |
| Affiliation | White House           | UK Government     | US Federal Reserve | 50256 |
|             | 7996                  | 2841              | 2229             |       |

Table 2: Count of Attribute Entries After Preprocessing

The imputation of missing job titles and affiliations reduced the number of missing entries in the aforementioned columns. This is demonstrated in Table 3.
which shows the number of missing entries before and after the imputation step. The number of missing entries for Job Role and Affiliation has been reduced significantly. The missing entries in the job role category have been reduced by 71.40% (378114 to 108138) and the affiliation category by 82.90% (293485 to 50256).

The preprocessing phase has reduced the amount of ambiguity and the number of missing entries and after preprocessing the number of quotes have been reduced to 458602 entries due to the removal of duplicate, blank and non-English quotes.

| Attribute     | Blank Entries (Unprocessed) | Blank Entries (Processed) |
|---------------|----------------------------|---------------------------|
| Speaker       | 0                          | 0                         |
| Job Role      | 378114                     | 108138                    |
| Affiliation   | 293485                     | 50256                     |

Table 3: Count of Blank Entries

3 Phrase and Word Extraction

Phrases and words from the quote may carry information that may assist the interpretation of the quote. The word and phrase types that were extracted from each quote were: highlighted terms, neologisms, similes and metaphors.

Highlighted terms are words and phrases that have been given an indicator of importance or derision by the journalist by using inverted commas. For example, “A $10bn credit line that it drew down earlier this year is ”needed for VW” and is not invested in options.”, the highlighted term is “needed for VW”, and it can be inferred from the quote that the journalist is derisive of the claim by Frank Gaube that VW needs a $10Bn credit line is needed by VW [2]. These terms were extracted using manual rules.

Neologisms are invented words. Neologisms are arguably an attempt by the speaker to convey a meaning that is not available in the English lexicon. For example, “The **measurability** really sets it apart.; You can track any money you spend online” [7]. In this case, the neologism is the term **measurability** which is a portmanteau of the words “measure” and “ability”, which conveys the ease a process has to be measured. The neologisms were detected by identifying words that are not in English Language dictionaries, and are not a named entity such as company or person name as well the word not being a typo, spelling error or formatting error. The neologism identification step checked the candidate word against US and UK English dictionaries, as well as checking the word against the speaker’s name, job role and affiliation as well as against named entities extracted from the quote by a pre-trained model. If the candidate word fails each one of these checks it is assumed to be a candidate neologism. Spelling and formatting errors are migrated by ensuring that the candidate neologism appears in the corpus a minimum number of times.
3.1 Simile Extraction

A simile is a figure of speech that compares one thing with a different thing using the connectors “like a” or “as a”. The use of similes in business communication is an aid to communication with a wide audience because it can make “an arid topic concrete and immediate” [3]. And therefore the communicator will not lose the attention of the audience. Similes can often be crude such as the example provided by [3] which drew a comparison between sex and labour relations.

The importance of similes in business communication was the motivation to extract it and provide it as part of the resource. The extraction technique was a rule-based approach which used the following patterns to extract similes.

1. \( \text{ADJ}_{0:1} \text{NOUN}_1 \text{like}_1 \text{DET}_{0:1} (\text{ADJ}_{0:1} \text{NOUN}_1)_1 \).

2. \( \text{as}_1 \text{ADJ}_1 \text{as}_1 \text{DET}_{0:1} (\text{ADJ}_{0:1} \text{NOUN}_1)_1 \).

The aforementioned Rules uses the following nomenclature: Noun: Noun, ADJ: Adjective and DET: Determiner. The numeric arguments determine the minimum and maximum times the term can appear in the extracted phrase. For example, \(0:1\) indicates that the minimum amount of times it appears is 0 and the maximum frequency is 1, whereas 1 indicates that the term must appear exactly once. Where is there is no upper limit the term can appear as many times as it exists in the quote.

The first rule identifies similes that compare Nouns or Adjective-Noun pairs with the term “like” with other Nouns or Adjective-Noun Pairs. The rule also allows for the optional use of determiners. Rule one for example returns this simile: currency like the dollar, where the word currency is the Left Hand Side (LHS) Noun, and the is the optimal determiner and the word dollar is the the Right Hand Side (RHS) Noun. The simile transformational technologies like broadband which is also returned by Rule one, demonstrates the use of the optional adjective transformational.

The second rule identifies similes that compare an Adjective with Nouns or Adjective-Noun pairs with the phrase as a. In common with Rule One, Rule Two also allows for the use of optional determiners. An example returned by Rule Two is as good as a coin flip where the adjective good is compared with the noun phrase “coin flip”. This phrase does not have an optional determiner.

3.2 Metaphor Extraction

A metaphor is a word or phrase that describes an action that is applied to an object which is not physically possible, such as leap of faith. The role of metaphor in business and financial language is to illustrate complex or dry language with visual stories that indicate the thought process of the speaker [10]. The resource has two types of metaphors extracted for each quote. They are: “is a” and “of” metaphors.

The “is a” metaphor extraction strategy follows the technique described by [5] where the metaphor is detected by the following pattern: \( NP \text{ is a } NP \), where
NP represents a Noun Phrase. The last step of the technique is to ensure that the nouns in the Noun Phrases are not hyponyms of the other and that they do not share any common hyponyms. An example of a “is a” metaphor returned by this technique is *This tax credit is a lifeline*.

The “of” metaphor was computed by the technique described by [8]. With this technique candidate pairs of Noun and Noun Phrases are harvested with the pattern: *NP of NP* where *NP* represents a Noun Phrase. A pointwise mutual information score is computed by: \( P(x, y) = \log \frac{p(y|x)}{p(x)} \), where \( x \) is the RHS Noun Phrase and \( y \) is the LHS Noun Phrase, and \( p(y|x) \) is the probability of two Noun Phrases co-occurring whereas \( p(x) \) is the probability of the Noun Phrase \( x \) appearing in the whole corpus. Noun Phrase pairs that are above a predetermined score are accepted as constituents of a “of” metaphor such as “leap of faith” and “ghettos of poverty”.

### 3.3 Emotion Detection

Emotion detection is a technique of classifying a word or phrase into an emotion category such as joy or anger. The motivation to add this information is that emotions drive the financial markets and can be an indicator of the underlying prospects of the company [9]. The tool that was used is called Limbic which classifies a quote into one of the four following emotion categories: anger, fear, joy and sadness. The score for each category is a continuous number from zero to one. With a score of one being the highest score possible. The emotion category the quote is classified into is the emotion with the highest score.

The quotes that attracted high scores for each emotion is shown in Table 4. The emotions estimated from the economic actors represented in this corpus indicates Keynes “Animal Spirits” stalked the financial markets between 2008 and 2012.

### 3.4 Entity Detection

A pre-trained entity detection model was used to identify named entities present in the quote. The full list of named entity classes that are detected can be found [here](#) and include Person, Location and Product. Sample output is the following: “Eric”, ”PERSON”, ”Larry”, ”PERSON”, ”Natti”, ”PERSON”, which was generated from the quote: “Eric, Larry, and Natti are some of the more talented equity sales and trading professionals on the Street”.

### 4 Resource Structure

The updated resource is located [here](#). The resource contains the following data files: Raw Data, Sentiment Bigrams, Discourse Connectors and the JSONL file which contains the updated corpus.

The Raw Data file is a tab-delimited file (TSV) that contains a Quote ID, a Speaker Name, an Affiliation and a Quote. The Sentiment Bigrams and Dis-
| Emotion | Score | Quote |
|---------|-------|-------|
| Anger   | 0.92  | The technology industry would be likely to fare better this time than during its last previous big downturn in the aftermath of the tech bubble in 2001, partly owing to advances in how inventory is managed. "We all feel the rage of the storm" |
| Fear    | 0.84  | Some analysts attribute the correlation to continuing foreign capital outflows, while others say it is panic among domestic investors. "It’s all Gulf sentiment now,"
| Joy     | 0.89  | the forecast fall in retention numbers was a prudent step. "Our customers tend to be elderly, insurance-minded people who are happy to pay to avoid the worry of a household emergency." |
| Sadness | 0.90  | While risk appetite may be getting a bit frothy ... the pound may remain resilient versus the euro with the still festering problems for the eurozone largely left unaddressed, especially the terrible costs that the strong euro is having on its weaker members. |

Table 4: Emotive Quotations

course Connectors were created for the first release of this resource [1]. The discourse connectors contain a discourse connector and a sentiment score, whereas the sentiment bigrams are word pair candidates with a sentiment score computed by the method described by [6]. Each file has a single entry per line with an unbounded number where a score of less than zero indicates a negative sentiment and a score above zero indicates a positive score. A representative entry for a discourse entry is “worried” which has a score of -2.49. The bigram “strong reputation” which has a score of 16.10 is a typical entry for the sentiment bigrams.

The JSONL file contains markup that describes a Python dictionary for each quote. It contains the following keys:

- **ID:** Unique identifier for the entry.
- **Speaker Name:** The name of the person who made the quote.
- **Affiliation:** The organisation that the speaker is associated with.
- **Sentences:** The sentences of quote
- **Neologism:** Neologisms present in the quote
- **Quotedphrases:** Word and phrases in the quote that have been highlighted by the journalist
- **Simile:** Similes present in the quote
• Metaphor: Metaphors present in the quote
• Emotion: Emotion distribution of the quote
• Entities: Entities present in the quote

The resource contains one code file, “main.py” which provides a rudimentary interface to the JSONL file. The user passes a command-line argument of a dictionary key with a comma-delimited of words that will be searched for as well as an output location of an image that a Word Cloud will be outputted to.

5 Conclusion

These Project Notes describe an update to the Minho Quotation Resource. This update includes removal of duplicate quotes, normalisation of job titles as well as imputation of job roles and affiliations. In addition to the resources supplied with the first iteration, metaphors and similes, as well as emotion distribution for each quote, have been added.

The resource covers an important period of financial history where the world suffered a depression. The resource contains quotes about some significant events such as the failure of Lehman Brothers and the bailout of General Motors as well as quotes from important economic actors such as Ben Bernanke and Richard Fuld. As the world economy begins to enter an economic crisis this update is a timely release for scholars who wish to study how economic actors communicate in a period of a severe financial crisis.

References

[1] Brett Drury and José João Almeida. The minho quotation resource. In Proceedings of Language Resources and Evaluation, pages 2280–2285, 2012.

[2] Financial Times. Porsche intentions baffle markets. https://www.ft.com/content/db57be26-a5ea-11dd-9d26-000077b07658 Online; accessed 24 June 2020.

[3] Craig Kallendorf and Carol Kallendorf. The figures of speech, ethos, and aristotle: Notes toward a rhetoric of business communication. The Journal of Business Communication (1973), 22(1):35–50, 1985.

[4] John Maynard Keynes. The general theory of employment. The quarterly journal of economics, 51(2):209–223, 1937.

[5] Saisuresh Krishnakumaran and Xiaojin Zhu. Hunting elusive metaphors using lexical resources. In Proceedings of the Workshop on Computational approaches to Figurative Language, pages 13–20, 2007.

[6] Bing Liu. Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media, 2007.
[7] Nick Clark. Online advertising overtakes TV sales for first time ever. [https://bit.ly/2NtM551](https://bit.ly/2NtM551) Online; accessed 24 June 2020.

[8] Ashequl Qadir, Ellen Riloff, and Marilyn A Walker. Automatically inferring implicit properties in similes. In *Proceedings of NAACL-HLT*, pages 1223–1232, 2016.

[9] Nadine Strauß, Rens Vliegenthart, and Piet Verhoeven. Lagging behind? emotions in newspaper articles and stock market prices in the netherlands. *Public Relations Review*, 42(4):548–555, 2016.

[10] Mihaela Vasiloaia, Michelle Gaisoa-Issa, and Nora Vergara-Issa. Metaphors business linguistic culture lives by. *Economy Transdisciplinarity Cognition*, 14(1), 2011.