BrAgriSpeech: A Corpus of Brazilian-Portuguese Agricultural Reported Speech

Brett Drury (0000-0003-1468-0089)[1], Samuel Morais Drury[2]
brett.drury@gmail.com, and samuel.morais.drury@gmail.com

[1] LIAAD-INESC-TEC, Porto, Portugal
[2] Colégio Puríssimo, Rio Claro, SP, Brasil

Abstract. Agriculture is one of Brazil's largest industries. In Brazil, the price of crops such as sugarcane is driven not only by the production levels but also by speculation and rumour. Also, some crop derivatives such as ethanol have their prices regulated by the government. Reported comments from influential speakers such as government ministers and agricultural-business leaders can impact the prices and in some cases the level of production of food products. Currently, there are no corpora in Brazilian-Portuguese that contains agricultural-related speech, the speakers and their employer. BrAgriSpeech is a corpus that uses linguistic rules and pre-trained models to extract reported speech, the speaker and where available the speaker's employer as well as a discourse connector that connects the speaker with the quote. The resource has 6982 quotes which are in JSONL format. A sample of 50 quotes was manually evaluated and had an accuracy of 0.77 for quote identification, 0.82 for the identification of the speaker and 0.87 for the identification of the discourse connector. The resource is publicly available to encourage further research in the area.

1 Introduction

Agriculture is one of Brazil's largest industries, consequently, the fluctuation of price and quantity of agricultural produce is of national importance. Agricultural prices can fluctuate because of the quantity of produce as well as speculation and rumour. Economic and political actors can influence prices through comment and reported speech in the news media [7]. Brazil regulates the prices [5] of crop derivatives such as ethanol, also it controls the imports of competing agricultural goods [3]. Therefore it is arguable that comments and reported speech from influential economic and political actors can affect the price as well as the quantity of food produced in Brazil. There is currently no corpus of agriculturally related speech in Brazilian-Portuguese. BrAgriSpeech has therefore been developed to bridge this gap in the research literature. It has 6982 examples of reported speech from economic and political actors as well as business leaders. The speech examples have a quote, speaker, discourse connector and where

[3] https://bit.ly/2PrDUtO

available the speaker's employer. The corpus has been made publicly available so that researchers can use the corpus for further investigation on the effect of speech on agricultural production as well as the patterns of speech and use of language in the public commentary about agricultural policy and business.

The remainder of the paper will cover the following 1. Related Work, 2. Extraction Methodology, 3. Evaluation, 4. BrAgriSpeech's Linguistic Characteristics, 5. Resource Overview and 6. Conclusion.

2 Related Work

The related work in this area is sparse because this domain is a niche area of research. The main corpus located in the area of Brazilian agricultural news is BrAgriNews [2]. BrAgriNews is a corpus of Brazilian-Portuguese agricultural news and is the source material for BrAgriSpeech. BrAgriNews spans the period 1996 to 2016 and has 96784 documents. It makes a simple attempt to mark quotations in the news by identifying quote delimiters such as quote marks (") which delimit quotes. It does not attempt to locate the speaker or quotes that are not delimited by quote markers. It does however have entities and sentiment tags for words within the quotes.

There were no directly comparable resources to BrAgriSpeech found in the literature review, however, [8] produced a vocabulary and corpus of what the authors called semi-popularization articles in the agricultural domain. The genre of publication is described by the authors as scientific articles aimed at the lay reader. The corpus however was quite small as consisted of only seven hundred documents and was in English.

The Fame speech corpus [4] is a corpus of audio speech recorded from the radio rather than written speech, however it does contain recorded material about agriculture. The final corpus located in the literature search is the Minho Quotation Resource [1] which is a business-related speech corpus that has direct speech from English Language news stories. It has a small number of quotes that are related to agriculture.

There are a small number of papers that extract quotes or reported speech from news articles. Several approaches use rules, for example, [11] used rules to extract quotes from European-Portuguese texts, and [1] used Open Calais [4] which in turn used rules to extract quotes. And [12] used rules to extract quotations from Indonesian online news. The final approach found in the literature review was [10] who used rules and a dependency parser to identify quotes. The rule-based approach seems to gain high precision at the cost of the recall.

There have been some attempts to use machine learning techniques to extract quotes from text. The unique approach that was found in the literature review was [9] who used a Conditional Random Field to predict if a token is part of a quote or not, and they compared it to a Maxent Classifier. They used features such as verbs and labels of words that are in the current span as the current token.

[4] Open Calais

3 Extraction Methodology

The BrAgriSpeech corpus was extracted using a ruled based approach because it is relatively simple to implement, and has high precision. The rules detected two types of speech: 1. quotes delimited by quote marks ("), and 2. quotes which are not delimited by quote marks. Each quote has quote attribution to a speaker and where possible the speaker's employer.

The rules that extracted the quotes are patterns that identify a sequence of particular linguistic features. The linguistic features are detected at the sentence level, and consequently, the source texts from BrAgriNews were split into sentences using a modified sentence splitter. The modified sentence splitter joined candidate sentences that had one quote mark with the previous sentence. This is because if the candidate quote had more than one sentence then the sentence splitter would split into more than one sentence.

The rules parse the sentence with a Part of Speech of Tagger (POS) and a Named Entity Tagger. Also, a further rule is applied which looks for pairs of quote delimiters such as ″. For there to be a candidate quote one of the following sequence criteria must be met:

- Speaker, Verb, and Quote Marks (opening and closing)
- Quote Marks (opening and closing), Verb and Speaker

The named entity detector identifies the speaker by identifying the person class of a candidate phrase within a sentence. If there is no matching phrase then a pronoun would be identified, for example, (ele/ela) (he / she), by the POS tagger. The POS tagger identifies the verb, between the speaker and the quote, this verb will be referred to as the discourse connector. A further modification was made to the rules where the employer of the speaker is identified by looking for a connector between the speaker and a company entity, which also is detected with the entity tagger. The connectors are *do / da* which is the Portuguese equivalent of the word *of*. For example, " Gilvan Sampaio do Instituto Nacional de Pesquisas Espaciais" where the speaker Gilvan Sampaio is connected to the organisation Instituto Nacional de Pesquisas Espaciais by the connector *do*. A representative example of this rule-based approach is ""Essa estimativa leva em conta a falta de chuvas durante o desenvolvimento fisiológico das plantas" , destacou o presidente da consultoria, Plínio Nastari", which in English is: "This estimate takes into account the lack of rain during the physiological development of plants highlighted the president of consulting, Plínio Nastari.". A workbook with the code for this section of the corpus can be found here.

The second rule set uses similar rules but does not look for quote marks, but takes the discourse connectors from the previous step, and uses this set of verbs to connect to the speaker. For example, "A Cutrale , maior indústria de laranja do Pais , ja processou pouco mais de 30 % das frutas desta safra, diz o diretor corporativo Carlos Viacava" which is in English is "Cutrale, the largest orange industry in the country, has already processed just over 30 % of the fruits of this harvest, says the corporate director". There are no speech delimiters, however

there is a discourse connector (diz) and a candidate speaker (Carlos Viacava). It is possible using the aforementioned sequences to infer the speech portion of the sentence.

The resource is a combination of the output of both sets of rules.

4 Evaluation

An evaluation of the information extraction methodology was made where a random sample of fifty quotes was selected and three domain experts evaluated if the: quote, verb and speaker were correct. The employer was not evaluated because there were insufficient examples. The margin of error from this sample is 0.13, therefore there is a 95 per cent chance that the real evaluation value is within 13 per cent of the evaluation value that is presented here.

The evaluation metric used a discrete score where the results are either correct or not. The evaluation metric was accuracy which is simply $accuracy = \frac{correct_i nstances}{total_i nstances}$. The average accuracy of the evaluation is: Quote, 0.77 ($\pm$0.02), Verb, 0.87 ($\pm$0.02) and Speaker, 0.82 ($\pm$0.08).

In addition to average accuracy and standard deviation, a Fleiss Kappa Coefficient [3] was computed for Quote, Verb and Speaker, and the results were 0.85, 0.76 and 0.68, which indicates that there was strong agreement between the annotators. Although there was stronger disagreement between the annotators for speaker identification than for quote identification. In short, the results demonstrate that the quote was accurately found, however, the discourse connector (verb) and speaker identification were more accurately found than the remainder of the quote.

5 BrAgriSpeech's Linguistic Characteristics

To provide some illustration of the nature of the corpus some basic linguistic analysis was made. The first one was the breakdown of the percentage of Part of Speech Tags (POS) that make up the corpus. The results are in Figure 1. The figures in Figure 1 don't add up to one because of rounding errors, and POS tags that had a value of less than three per cent are excluded so that Figure 1 is not crowded.

The most frequent POS tag is a Noun, which is due to the speaker referring to "objects" such as the ethanol and markets. The verbs and adjectives often refer to actions of objects such as markets rising or falling, or state of objects such as poor harvest. The numeric tag has a quite high value because the speakers often quantified the verb, for example, "the sugar futures market moved down by five per cent today".

Nouns describe some of the topics that are being discussed in the resource, therefore a simple frequency analysis of Nouns is shown in Figure 2.

The figure demonstrates that the most common nouns are related to markets (mercado), government (governo), money (dinheiro) and president (presidente).

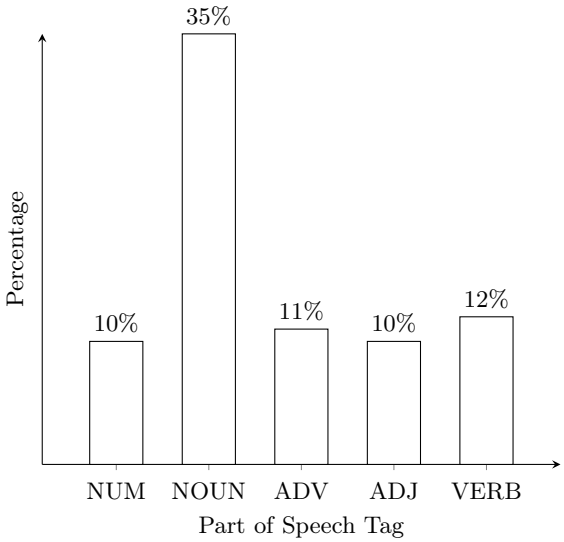

Fig. 1. Percentage of POS Tags in BrAgriSpeech

Fig. 2. Most Frequent Nouns in BrAgriSpeech

It is possible to assume that the majority of the reported speech in the source corpus, BriAgriNews, is mainly related to market news and conditions. Also, the most frequent nouns demonstrate the role of government in agriculture in Brazil. This is not a surprise because the Brazilian government does intervene in the agricultural market, such as paying subsidies to Brazilian farmers [5].

[5] `http://www.oecd.org/brazil/brazil-agriculturalpolicymonitoringandevaluation.htm`

The main product and crop that is shown in the most frequent noun analysis is etanol (ethanol) and Cana-de-açúcar (sugarcane), and this frequency reflects the dependency of Brazil on ethanol which is produced from sugarcane as it is used as a gasoline substitute by low-income consumers.

A frequency analysis of the speakers was made, and it was found that the most frequent speakers in the corpus are the following:

- Fernando Henrique Cardoso
- Michel Temer
- Dilma Rousseff
- Vlamir Brandalizze
- Luiz Inacio Lula

Three of the most frequent speakers: Michel Temer, Dilma Rousseff and Luiz Inacio Lula are former presidents of Brazil. The remainder, Vlamir Brandalizze, is a director of Brandalizze Consulting, and Fernando Henrique Cardoso is an academic. The presence of three ex-presidents in the resource indicates the importance of the source material and agriculture in general to the Brazilian economy.

6 Resource Overview

The corpus is supplied in JSONL format[6] in a text file which is located here. Each line of the text file is a valid JSON statement and is a dictionary, which is a set of key-value pairs, where the key is a unique identifier. An overview of the resource is presented in Figure 3. The Figure shows the keys of the dictionary and its position in the resource hierarchy.

The top level of the resource is "Quote ID", and "BrAgriNews File Name". The Quote ID is a unique numeric identifier for a quote. Quote ID starts at zero and also points to another dictionary. The BrAgriNews File Name key holds the value of the name of the source file in the BrAgriNews[2] corpus. The resource is organised in this manner because a single BriAgriNews file may have more than one quote.

The second level dictionary has the following keys:

- Sentence
- Quote
- Verb
- Company
- Speaker

The sentence key has a value of a string which is the sentence from which the associated quote is taken from. The sentence key always has a value. The keys: Quote, Verb, Company and Speaker have values which are one of the following: 1.

[6] `https://jsonlines.org`

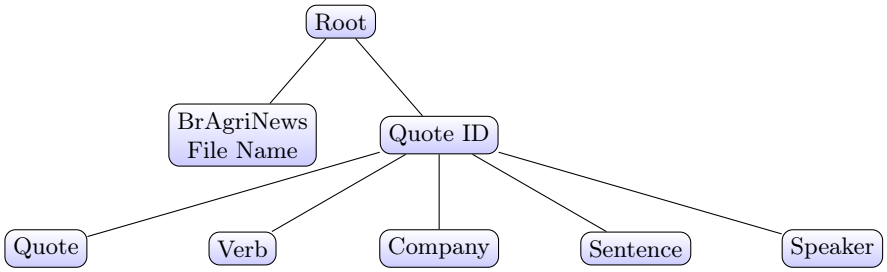

Fig. 3. Overview of BrAgriSpeech Corpus

Null / None when there is no information or 2. A Dictionary. The sub-dictionary has the following keys: start position, end position and value. The value key has a value of a string that reflects the parent key, therefore the value key of the quote parent key will have the quote and the value key of the speaker parent key will have the speaker's name. The start position and end position keys hold the position information in the sentence for the value key pair.

An example of an entry in the resource is shown below. The example shows that the File Name: CIENCIA_2003_6476.txt, has one quote which has the ID "0". The speaker is Marcelo Lopes de Oliveira, and the discourse connector (verb) is "diz" (said). The quote is "Mas isso não é razão para pânico" (but this is no reason to panic).

– {"0": {"Quote": {"End Position": 33, "Value": "Mas isso não é razão para pânico", "Start Position": 1}, "Verb": {"End Position": 39, "Value": "diz", "Start Position": 36}, "Speaker": {"End Position": 65, "Value": "Marcelo Lopes de Oliveira", "Start Position": 40}, "Sentence": " Mas isso não é razão para pânico , diz Marcelo Lopes de Oliveira e Souza , do INPE (Instituto Nacional de Pesquisas Espaciais) . ", "Company": {}}, "BRAGRINEWS_FILE_NAME": "CIENCIA_2003_6476.txt"}

The entry is valid JSON, and although it was generated using Python, it should be readable using other languages.

7 Conclusion

BrAgriSpeech is a resource that has captured speech information from the BrAgriNews corpus. It is focused on reported speech in the Brazilian agricultural domain and has captured the agricultural speech record between the 1990s to 2016. It captures speech around the market and natural events such as price rises and droughts. Speech in this resource is directly reported in the media, and therefore it can be assumed that candid or truthful speech will not be present. The speech that is recorded is planned and uses various techniques to either manipulate audiences, downplay economically damaging events and overplay beneficial

information. It is a resource that captures how public individuals communicate in the mass media to various audiences.

The resource has been freely available to stimulate research in the area of public discourse in Brazilian-Portuguese in an economically significant area.

Future work will be aimed at building an Ontology that contains an end-point for further information about the speaker, company and concepts such as ethanol and sugarcane. These entities will have a unique id so that these entities can be identified within the text. The Ontology will point to further resources such as DBPedia [6] so that further information can be gathered from these resources.

Bibliography

[1] Brett Drury and José João Almeida. The minho quotation resource. In *LREC*, pages 2280–2285, 2012.

[2] Brett Drury, Robson Fernandes, and Alneu de Andrade Lopes. Bragrinews: Um corpus temporal-causal (português-brasileiro) para a agricultura. *Linguamática*, 9(1):41–54, 2017.

[3] JL. FLEISS. The measurement of interrater agreement. *Statistical Methods for Rates and Proportions*, 1981.

[4] H Heuvel, E Yilmaz, DA van Leeuwen, H Velde, and J Dijkstra. *FAME! Speech Corpus*. PhD thesis, Radboud University, 2016.

[5] Mario de Queiroz Monteiro Jales and Cinthia Cabral da Costa. Measurement of ethanol subsidies and associated economic distortions: an analysis of brazilian and us policies. *Economia Aplicada*, 18(3):455–481, 2014.

[6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[7] Anastasios Maligkris. Political speeches and stock market outcomes. In *30th Australasian Finance and Banking Conference*, 2017.

[8] Verónica L Muñoz. The vocabulary of agriculture semi-popularization articles in english: A corpus-based study. *English for Specific Purposes*, 39:26–44, 2015.

[9] Silvia Pareti, Tim O'keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, 2013.

[10] Andrew Salway, Paul Meurer, Knut Hofland, and Øystein Reigem. Quote extraction and attribution from norwegian newspapers. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 293–297, 2017.

[11] Luis Sarmento and Sérgio Nunes. Automatic extraction of quotes and topics from news feeds. In *DSIE'09-4th Doctoral Symposium on Informatics Engineering*, 2009.

[12] Yusuf Syaifudin and Arif Nurwidyantoro. Quotations identification from indonesian online news using rule-based method. In *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 187–194. IEEE, 2016.