BrAgriNews: Um Corpus Temporal-Causal (Português-Brasileiro) para a Agricultura

BrAgriNews: A Temporal-Causal Brazilian-Portuguese Corpus for Agriculture

Brett Drury
Faculty of I.T.,National University of Ireland Galway,Ireland
brett.drury@gmail.com

Robson Fernandes
ICMC, University of Sao Paulo, Sao Carlos,Brazil
robs.fernandes@outlook.com

Alneu de Andrade Lopes
ICMC, University of Sao Paulo, Sao Carlos,Brazil
alneu@icmc.usp.br

Resumo

Recentemente tem havido um aumento no interesse, tanto no meio acadêmico quanto na indústria, em aplicações de aprendizagem de máquina e técnicas de inteligência artificial relacionadas com problemas agrícolas. Mineração de texto e técnicas relacionadas com o processamento da língua natural, raramente foram usadas para resolver problemas agrícolas, e muito menos para a língua portuguesa. É possível que um dos fatores que influenciam a escassez no uso técnicas de mineração de texto, para analisar textos em português e resolver problemas agrícolas, pode ser devido à falta de um corpus anotado livremente disponível. Para colmatar a falta de um corpus agrícola em língua portuguesa, estamos liberando um recurso em português-brasileiro voltado para agricultura, descrito neste artigo. O corpus abrange um período parcialmente contínuo de tempo entre 1996 e 2016, consistindo de notícias em português-brasileiro que foram anotadas com o seguinte tipo de informação: causal, sentimento, entidades nomeadas que incluem expressões temporais. O corpus tem recursos adicionais como: treebank, listas de termos frequentes (sem stop-words): unigramas, bigramas e trigramas, bem como palavras ou frases que foram identificados por jornalistas como de domínio específico. Espera-se que a liberação do corpus estimule a adoção da mineração de texto na agricultura na comunidade de pesquisa lusófona.

Keywords

Mineração de Texto Agricultura Relações causais

Abstract

There has been a recent sharp increase in interest in academia and industry in applying machine learning and artificial intelligence to agricultural problems. Text mining and related natural language processing techniques, have been rarely used to tackle agricultural problems, and at the time of writing there was a single project in the Portuguese language. It is possible that the failure of researchers to use text mining techniques to analyze Portuguese texts to resolve agricultural problems may be due to a lack of freely available corpora. To correct the lack of a Portuguese language agriculture centric corpus we are releasing a Brazilian-Portuguese agricultural language resource, which is described by this paper. The corpus is partially non-contiguous and spans a time period from 1996 to 2016. It consists of news stories that have been scraped from Brazilian News sites that have been annotated with the following information types: causal, sentiment, named entities that include temporal expressions. The corpus has additional resources such as a: treebank, lists of frequent: unigrams, bigrams and trigrams, as well words or phrases that have been identified by journalists as either: "important"or domain specific. It is hoped that the release of this corpus will stimulate the adoption of text mining in agriculture in the Lusophonic research community.

Keywords

Text Mining Agriculture Causal Relations

1 Introdução

Este artigo descreve um corpus em português-brasileiro, em que se pretende ser útil para incentivar a pesquisa em mineração de texto para a agricultura.

O *BrAgriNews* é um corpus parcialmente não contíguo que abrange o período de 1997 a 2016. O corpus anota as seguintes informações: sentimento, informações temporais, causais e entidades nomeadas em notícias agrícolas. O corpus contém: Um "treebank" e documentos com parte de etiquetas de fala, bem como: modelos de tópicos e representações vetoriais de termos. Também fornece recursos léxicos, tais como:

1. Palavras frequentes;
2. Bigramas frequentes;

DOI: 10.21814/lm.9.1.245

3. Trigramas frequentes;
4. Palavra/frases que são considerados "importantes" pelos jornalistas com a adição de delimitadores, como aspas.

O restante do artigo está organizado da seguinte forma: Seção 2: Trabalhos Relacionados; Seção 3: Aquisição do Corpus e Visão Geral; Seção 4: Metodologia de Anotação; Seção 5: Recursos Léxicos; Seção 6: Treebank; Seção 7: Recursos de Relações entre Palavras; Seção 8: Informações de Nível de Documento; Seção 9: Licenciamento; Seção 10: Trabalhos Futuros; Seção 11: Conclusão.

2 Trabalhos Relacionados

Este corpus contém uma variedade de fenômenos da linguagem, incluindo causalidade, expressões temporais, bem como sentimento. O trabalho relacionado, portanto, concentra-se nas seguintes áreas:

1. Causalidade na linguagem.
2. Representação temporal no texto.
3. Sentimento na linguagem.

Causalidade

Há uma série de definições de causalidade. Uma definição bem conhecida foi preferida pelo filósofo escocês David Hume que afirmou que a causalidade tem três propriedades específicas: "(i) contiguidade no tempo e no lugar; (ii) prioridade no tempo, e (iii) constante conjunção entre a causa e o efeito" (Khoo et al., 2002). A causalidade na linguagem é expressa como "relações causais." As relações causais são relações dependentes entre eventos, fatos ou objetos (Vendler, 1967; Altenberg, 1984), onde um evento, fato ou objeto é a causa de outro evento, fato ou objeto (Altenberg, 1984).

As relações causais no texto como explicado anteriormente são relações dependentes entre eventos, fatos ou objetos. Os objetos de causa (eventos, fatos ou objetos) são ligados através de uma ligação causal aos objetos de evento (eventos, fatos ou objetos). Uma ligação causal é uma palavra ou frase que contém propriedade causal. Ligações causais são tipicamente verbos causais (Shams-Eddien, 2002), nos quais a causa ou objetos de evento podem ser expressos como frases nominais. As relações causais podem, portanto, ser expressas como simples padrões de extração, como:

$$NP\ CV\ NP,$$

no qual NP = *Frases Nominais* e CV = *Verbo Causal* (Shams-Eddien, 2002). O fluxo de causalidade neste padrão é da esquerda para a direita, onde o lado esquerdo (LHS) NP é o objeto de causa e o lado direito (RHS) é o objeto de efeito. Em português esta ordem pode ser alterada por uma preposição, por exemplo a expressão "por causa de" inverterá a ordem de causalidade em uma relação causal. A maior parte da pesquisa sobre a causalidade na língua foi realizada em inglês, por exemplo por Khoo et al. (2002); Altenberg (1984); Thomson (1987); Shams-Eddien (2002), sendo que poucos foram os estudos conduzidos em Português (Drury & de Andrade Lopes, 2015).

Representação e Extração do Tempo

Uma característica dos corpora disponíveis são as anotações temporais. Uma suposição deste artigo é que a representação temporal no texto é uma maneira de descrever expressões multi-palavras que representam:

1. Duração;
2. Expressão do tempo.

Por exemplo: "21 de maio de 2001" é uma expressão do tempo e "12/04/75 – 12/05/76", é uma duração de tempo. O tempo pode cobrir: segundos, minutos, horas, dias, décadas, anos e assim por diante.

Expressões de tempo podem ser feitas em linguagem natural em uma série de maneiras diferentes, consequentemente houve um padrão desenvolvido que tenta ter uma maneira uniforme de expressar informação temporal e de evento. Este padrão é o *TimeML* (Pustejovsky et al., 2003a)[1]. O *TimeML* é um dialeto XML, que permite a expressão padrão de:

1. Marcação de tempo de eventos;
2. Ordem de eventos com relação a um outro;
3. Raciocínio com expressões temporais contextualmente sub-especificadas;
4. Raciocínio sobre a persistência de eventos.

Além da padronização das expressões temporais, o consórcio *TimeML* lançou uma série de ferramentas que podem ser usadas para anotar ou extrair expressões de tempo no texto. O site documenta a Ferramenta de anotação (TANGO) e o *Tarsqi Toolkit*.

[1] http://www.timeml.org

O *Tarsqi Toolkit* contém um conjunto de ferramentas que podem ser usadas para extrair expressões de tempo, bem como garantir a sua consistência.

A literatura de pesquisa contém uma série de estratégias para extrair expressões temporais. Essas estratégias podem ser agrupadas em duas categorias: 1. aprendizagem de máquina (Bethard, 2013; Kolya et al., 2013; Llorens et al., 2010; UzZaman & Allen, 2010) e 2. híbrida de aprendizagem de máquina e linguística (Laokulrat et al., 2013; Jung & Stent, 2013).

Uma abordagem comum de aprendizado de máquina na literatura de pesquisa é a aprendizagem supervisionada com campos aleatórios condicionais (conditional random fiels — CRF) (Kolya et al., 2013; Llorens et al., 2010; UzZaman & Allen, 2010). As abordagens híbridas usam características linguísticas de dados rotulados para gerar modelos em uma estratégia de aprendizagem supervisionada. As duas principais características linguísticas utilizadas nas técnicas híbridas são as estruturas de dependência (Laokulrat et al., 2013) e informação semântica (Jung & Stent, 2013).

Existem vários corpora que podem ser usados para avaliar estratégias de extração temporal. Os dois principais corpora para o Inglês são: TimeBank (Pustejovsky et al., 2003b) e o AQUAINT Corpus[2]. Esses corpora são relativamente pequenos, com 183 e 73 notícias, respectivamente. Existem corpora em línguas não-inglesas, tais como para o Francês (Bittar, 2010), Italiano (Caselli et al., 2011), Romeno (Forascu & Tufiș, 2012), Espanhol[3] e Catalão.[4] Para o Português temos o HAREM (Carvalho et al., 2008), com 129 notícias.

Análise de Sentimentos

A análise do sentimento, de acordo com Liu e Zhang, é o estudo computacional das opiniões, avaliações, atitudes e emoções das pessoas em relação a entidades, indivíduos, questões, eventos, tópicos e seus atributos (Liu & Zhang, 2012). O campo é vasto, consequentemente esta pesquisa será limitada à análise de sentimentos da língua portuguesa.

Existêm vários métodos para a análise do sentimento, porém a abordagem dominante para o português descoberta nessa revisão é a baseada em dicionário. A análise do sentimento baseada em dicionário utiliza-se de recursos lexicais que possuem palavras ou frases com uma orientação de sentimento pré-definida. Existem três dicionários principais: dois multilíngue: *SentiLex* (Silva et al., 2012), *Opinion Lexicon* (Souza et al., 2011) e *LIWC* (Balage Filho et al., 2013), que é parte de uma aplicação de software. Avaliou-se os três dicionários e os principais pontos constatados foram que o *Sentilex* foi superior para a classificação de sentimento de documentos e *LIWC* produziu os melhores resultados para a classificação de opinião de sentenças. A análise do sentimento baseado no dicionário para o português foi aplicada a uma série de áreas que incluíram hotéis (Chaves et al., 2012), finanças (Alvim et al., 2010), crítica de cinema (Freitas & Vieira, 2013) e política (Silva et al., 2009).

As estratégias supervisionadas de classificação do sentimento de aprendizado de máquina exigem dados de treinamento. Um possível impedimento para o uso dessas técnicas é a falta de corpos anotados na língua portuguesa. Esta revisão da literatura descobriu um pequeno número de recursos que continham relativamente poucos recursos: Petronews (1500 documentos) (Alvim et al., 2010), ReLi (2056 documentos) (Freitas et al., 2012) e o conjunto de dados de Drury & de Andrade Lopes (2014) (500 documentos).

3 Aquisição do Corpus e Visão Geral

O corpus, como já comentado, contém notícias relacionadas à agricultura escritas em português-brasileiro. O corpus foi construído a partir de recursos inéditos pré-existentes e com notícias coletadas na Internet. As notícias foram coletadas com um "scraper" de sites respeitáveis, como:

1. Revista Canavieiros (Sugarcane Magazine).
2. Jornal Cana (Sugarcane Newspaper).

O "scraper" rodava às 8 horas da manhã, antes do início da bolsa de São Paulo. Esta decisão foi tomada para garantir que todas as experiências de negociação que foram feitas com modelos derivados deste corpus seriam "justas". O "scraper" correu de 2014 a 2016. O corpus final contém 96.784 documentos.

Características da Linguagem

Coleções de documentos ou corpus têm características específicas de linguagem que são determinadas pelo assunto e estilo do autor. Uma

[2] https://tac.nist.gov//data/data_desc.html#AQUAINT

[3] Disponível em https://catalog.ldc.upenn.edu/docs/LDC2012T12/

[4] Disponível em https://catalog.ldc.upenn.edu/docs/LDC2012T10/

maneira de comparar a linguagem é comparar a frequência de:

1. Advérbios com adjetivos.
2. Substantivos com verbos.

Os rácios foram 0,52 e 2,24, respectivamente. Uma comparação com outros textos pode ser encontrada nas Figuras 1 e 2[5].

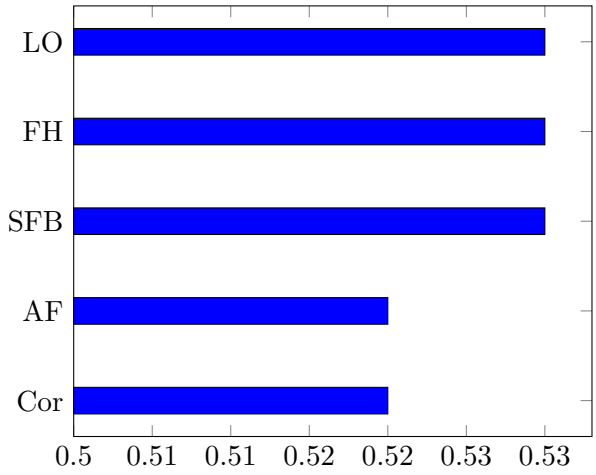

Figura 1: Relação entre advérbios e adjetivos, onde Cor = Corpus, AF = O Triunfo dos Porcos (Animal Farm), SFB = (Escândalo do Padre Brown) Scandal Of Father Brown, FH = História de Fanny Hill (Fanny Hill) e LO = Romance Lady Oracle (Lady Oracle).

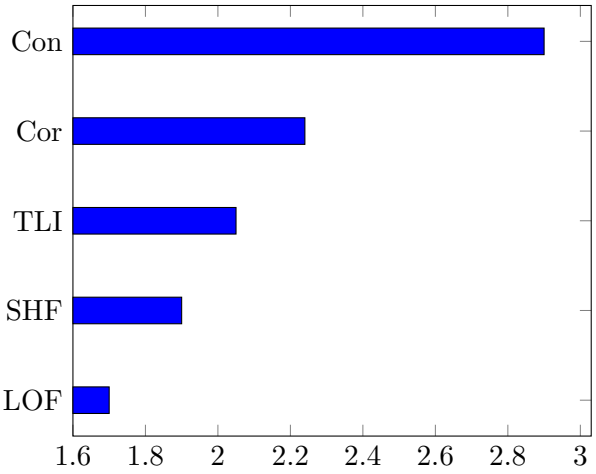

Figura 2: Relação entre Substantivo e Verbo, onde LOF = Vida de Johnson (Life Of Johnson), SHF = Forma das Coisas Por Vir (Shape Of Things To Come), TLI = O Instinto da Linguagem (The Language Instinct), Cor = Corpus (Corpus) e Con = Constituição (Constitution).

[5]Uma lista completa de rácios para textos alternativos para: substantivo/verbo e adjetivos/advérbios podem ser encontrados em: 1. https://goo.gl/1OZpNH e 2. https://goo.gl/6hzYPd.

Uma técnica de análise de linguagem complementar é listar as palavras mais frequentes no corpus. As palavras frequentes no corpus são boas indicadoras do assunto porque a frequência da palavra segue uma distribuição zipf, como demonstrado na Figura 3. A análise de palavras frequentes removeu *stop-words* (um, isto, o, etc), uma vez que elas não têm um significado específico de domínio, pois ocorrem com frequência relativa similar na maioria dos corpora ou coleções de textos. As palavras mais comuns neste corpus foram: Brasil; Milhões; Governo; Presidente; Mercado; Produção; Nacional; Acordo; Estado e Safra.

Uma representação visual da frequência de palavras na coleção de corpus é representada no diagrama de Nuvem de Palavra na Figura 4.

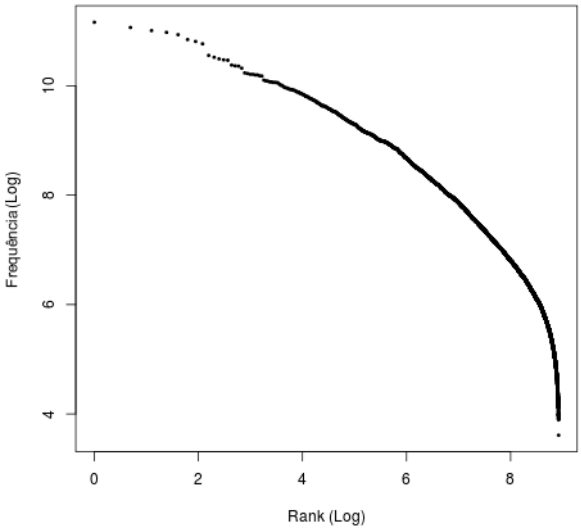

Figura 3: Relação entre a frequência das palavras e o seu rank.

Figura 4: Nuvem de Palavras de termos frequentes no corpus *BrAgriNews*.

A análise final considerou o tamanho do documento (número de palavras), frequência média das palavras e número total de palavras. Os valores foram: 1.127,14 palavras por documento,

frequência média de 1.617,82 palavras ±3504.12 e 12.305.150 de palavras no corpus *BrAgriNews*.

As técnicas de análise simples acima referidas forneceram uma visão geral das características linguísticas do corpus. A razão entre frequencias de substantivos e verbos indicam um corpus em liguagem objetiva, no qual a relação entre adjetivos e advérbios é similar a da literatura clássica. A contagem de frequência indica que os assuntos dominantes são: Estado; Comércio; e Agricultura. E que o comprimento médio do documento é relativamente pequeno.

Visão Geral do Corpus

O *BrAgriNews* está disponível em `https://goo.gl/lc0PzS`, e é distribuído como um arquivo compactado. A organização de pastas de nível superior é apresentado na Figura 5.

A pasta de nível superior contém: notícias, previsões meteorológicas e um *treebank*. As pastas *Weather Forecasts* e *Trees* contêm previsões meteorológicas e representação de árvore de dependência de sentenças aleatórias, respectivamente. A pasta *News Stories* tem um segundo nível de pastas que é demonstrado na Figura 5. O conteúdo das pastas será descrito posteriormente neste artigo.

Resumo da Etiqueta

A principal contribuição deste corpus é a anotação de notícias. A anotação delimita informações que podem ser úteis para categorização supervisionada ou técnicas de extração de relação. As notícias anotadas são armazenadas na pasta *Annotated Texts*. As anotações assumem a forma de marcações do tipo XML (etiquetas) que delimitam: uma única palavra ou uma sequência de palavras. As etiquetas anotam:

1. Sentimento.
2. Relações causais.
3. Porções de causa e efeito de relações causais.
4. Expressões de tempo.
5. Expressões de moeda.

Um resumo das etiquetas é descrito na Tabela 1.

Etiqueta	Explicação
Positive	Uma palavra que foi determinada como tendo uma orientação positiva
Negative	Uma palavra que foi determinada como tendo uma orientação negativa
Entity	Uma palavra ou n-grama que foi determinado como uma entidade nomeada
CRelation	Delimitação de uma relação causal
Effect	A parte de um efeito de uma relação causal
Cause	A parte de uma causa de uma relação causal
DOW	Dia da Semana
TOD	Hora do Dia
Season	Estação
Week	Expressão semanal
Date	Expressão diária
Currency	Expressão monetária
Quote	Discurso direto

Tabela 1: Resumo da Etiqueta.

4 Metodologia de Anotação

Esta seção discute as estratégias que foram usadas para anotar os documentos neste corpus. Haviam duas escolhas metodológicas possíveis para anotar este corpus:

1. Anotação manual.
2. Anotação automatizada.

A anotação manual é laboriosa e lenta, consequentemente seria impraticável usar esta técnica para este corpus e a anotação automatizada foi selecionada.

O resumo de etiqueta descrito na Tabela 1 revela que são 6 áreas de anotações principais:

1. Entidades nomeadas.
2. Anotação de sentimento.
3. Expressões de tempo.
4. Relações causais.
5. Discurso direto.
6. Parte da fala.

Entidades Nomeadas

As entidades nomeadas são palavras únicas ou expressões multi-palavras, que podem ser classificadas em uma categoria pré-existente, tais como: pessoa, empresa, organizações, e assim por diante. O suporte de entidade nomeada para o português-brasileiro é limitado e no momento da construção do corpus não havia nenhum classificador/extrator de entidade nomeada livremente disponível. Consequentemente, uma técnica baseada em regras foi desenvolvida para identificar candidatos de entidades nomeadas.

A técnica usou o seguinte procedimento:

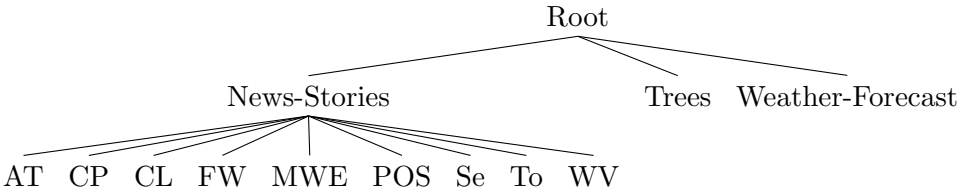

Figura 5: Organização das pastas, onde AT = Textos Anotados, CP = Frases de Causa, CL = Clusters, FW = Palavras Frequentes, MWE = Expressão Multi-Palavras, POS = Parte da Fala, Se = Sentimento, To = Topic and WV = Vetor de Palavras.

1. Identificar palavras maiúsculas que não iniciam sentenças.
2. Juntar os candidatos da Regra 1 se for separado por uma palavra de ligação.
3. Repetir a Regra 2 com entidades unidas geradas a partir dessa mesma regra.

O processo de união descrito nas Regras 2 e 3 pode ser ilustrada com o seguinte exemplo: uma entidade denominada candidata gerada por esta técnica é *Procuradoria-Geral da República*, que contém duas entidades candidatas denominadas *Procuradoria-Geral* e *República*, que é acompanhado por uma palavra de ligação *da*.

Uma pequena avaliação manual feita por um único especialista em domínios, onde 10 documentos foram escolhidos aleatoriamente, constatou que a técnica tinha uma precisão de 73.25%. A avaliação identificou manualmente as entidades em um documento, e verificou que a técnica as identificou corretamente. Correspondências parciais, bem como a falha ao identificar as entidades foram marcadas como incorretas.

Anotações de Sentimento

A anotação de sentimento foi alcançada usando um dicionário pré-compilado de sentimento: *Sentilex*. O dicionário contém palavras que têm uma orientação pré-determinada do sentimento. A estratégia divide as palavras em um documento e verifica a palavra contra a entrada no *Sentilex*. A estratégia aplica-se a uma das duas etiquetas: positiva ou negativa, as palavras com orientação de sentimento neutro são ignoradas. Por exemplo, <negative> ruim </negative>, ruim tem uma conotação de sentimento negativo e consequentemente é encapsulado com uma etiqueta negativa.

O dicionário *Sentilex* foi avaliado por Balage Filho et al. (2013), verificou-se que o Sentilex tem uma precisão de 44.17% no nível da sentença e 53.35% no nível do documento. *Sentilex* é um dos melhores dicionários de sentimentos para a língua portuguesa.

Expressões temporais

As expressões temporais para este corpus foram extraídas usando uma abordagem padrão baseada em regras. A expressão diária foi extraída com expressões regulares que identificaram sequências de números com separadores comuns. As expressões típicas captadas por esta abordagem foram “12/04/2016” e “12/04/16”.

As demais categorias de expressões de tempo foram capturadas usando listas codificadas de palavras. A lista de palavras foi compilada por um especialista em domínio.

As técnicas de anotação de expressão temporal baseadas na expressão regular, relataram exatidão muito alta, por exemplo, Strötgen & Gertz (2010) relataram que sua técnica de expressão regular registrou uma precisão de 85.00%.

Relações Causais

As anotações de causalidade seguem a noção de que as relações causais entre os eventos, e que a relação causal contém duas partes: (i) Evento de causa, e (ii) Evento de efeito.

Consequentemente, as anotações causais têm três anotações: (i) Toda a relação causal; (ii) Evento de causa; e (iii) Evento de efeito.

A estratégia de anotação causal foi uma estratégia de aprendizagem supervisionada descrita por Drury & de Andrade Lopes (2015). A estratégia utilizou uma visão local e global da causalidade no corpus. Dois separadores são criados a partir dessas duas visões. Os dois classificadores rotulam as relações causais no corpus e, quando os dois classificadores concordam com uma relação causal, uma anotação causal é feita. Exemplos das relações causais são demonstrados na Tabela 2.

Esta técnica foi avaliada por Drury & de Andrade Lopes (2015), verificou-se que tem uma precisão de 67.00% na anotação do nível da frase e 81.00% na classificação da relação causal.

Expressão Causal (Português)
preços gasolina alta aumentando demanda biocombustivel
politicas diminuído industria biocombustivel
consumo problemas logisticos causa destaca
surgiram oportunidades curto prazo exportacoes brasileiras biocombustivel

Tabela 2: Relações Causais relacionado com Biocombustíveis

Discurso Direto

Discurso direto para este artigo, é o discurso que foi citado diretamente no texto. Por exemplo, "Eu não estou em seu comitê de estratégia" Watson respondeu (https://goo.gl/VLeH18). O discurso é delimitado por marcas de fala, e seguido por uma entidade nomeada e um verbo.

A estratégia para anotar a fala direta foi outra técnica baseada em regras que identificou delimitadores de fala que foram as aspas, e as marcas de fala.

As palavras entre esses delimitadores foram assumidas como sendo de fala direta se a frase extraída tivesse uma contagem de palavras mínima de 6.

Uma pequena avaliação manual de 10 documentos que continham uma etiqueta de citação, realizadas por um único especialista de domínio, descobriu que as seqüências de texto que foram marcadas com aspas estavam corretas 86.66% do tempo. Uma citação correta foi assumida para ter um orador, como uma pessoa ou outra entidade, como uma empresa ou organização, bem como um elemento de fala. Marcação indevida ou obviamente incorreta foi marcada como um erro pelo anotador.

Marcação de Parte da Fala

A Marcação do papel morfo-sintático (part-of-speech tagging) aplica uma categoria de palavra como substantivo, adjetivo, advérbio, etc. a uma palavra. Para as marcações foi usado o *nlpnet* (Fonseca & Rosa, 2013) que é um rotulador baseado em rede neural. O rotulador foi treinado no corpus *mac-morpho* e tem: "97.33% a precisão de um token, 93.66% exatidão do token fora do vocabulário".

Um exemplo das anotações tipicamente encontradas no corpus pode ser encontrado na Tabela 3. A anotação é uma citação direta por "um consultor". A citação é encapsulada pela etiqueta "quote". A citação contém uma série de palavras sentimentais, em particular: "sofra" e "stress".

Essas palavras têm conotação negativa e, consequentemente, são encapsuladas por uma etiqueta "negativa". A citação contém uma relação causal: "o stress durante a pré-polinização pode resultar em produtividades menores.". Esta relação causal contém um evento de causa: "o stress durante a pré-polinização" e um evento de efeito: "produtividades menores". A citação também contém informações sobre o tempo: "Maio" e informação da entidade tal como: "Kansas" e "Iowa".

Exemplo Anotado

A agregação das anotações pode fornecer uma descrição detalhada dos dados. Um exemplo de anotações agregadas pode ser encontrado na Tabela 3.

Exemplo anotado
<Quote> "Minha preocupação é de que algum milho <Negative> sofra </Negative> com o <Negative> stress </Negative> hídrico durante a polinização, quando a planta está definindo o tamanho da orelha. Uma vez que este tamanho está definido, ele não pode ficar <Month> maio </Month> , assim sendo, <CRelation> <Cause> o <Negative> stress </Negative> durante a pré-polinização </Cause> pode resultar <Effect> em produtividades menores </Effect> </CRelation> . Eu acredito que isso possa já estar ocorrendo em alguns locais com o leste do <Entity> Kansas, </Entity> norte do <Entity> Missouri, </Entity> sul de <Entity> Iowa </Entity> e oeste de <Entity> Illinois, Indiana, Ohio </Entity> e <Entity> Michigan" </Quote> , </Entity> diz o consultor.

Tabela 3: Exemplo anotado

O exemplo de anotação demonstra claramente o esquema de anotação e como ele é usado dentro do corpus *BrAgriNews*, onde:

1. Etiqueta 'Quote' indica citação.
2. Etiqueta 'Negative' indica palavras com conotação negativa.
3. Etiqueta 'CRelation' indica citações que contém relação causal.
4. Etiqueta 'Month' indica citações que contém informações sobre o tempo.
5. Etiqueta 'Entity' indica informações sobre a entidade.

5 Recursos Léxicos

Há uma série de recursos léxicos que complementam o corpus principal. Os recursos léxicos estão localizados nas pastas *Multi-word Expressions* e *Frequent Words*.

Os recursos léxicos são: Palavras frequentes (não *stop-words*); Bigramas frequentes; e Trigramas frequentes.

Palavras Frequentes

As palavras frequentes, como descrito anteriormente, são palavras frequentes que não são *stop-words*. A técnica para identificar palavras frequentes eliminou qualquer palavra do corpus que estivesse em listas de stop-words comuns[6]. A frequência para o restante das palavras foi calculada. As 7499 palavras mais frequentes são armazenadas em um arquivo de texto e em formato "pickle" em Python (dicionário) e localizado na pasta *Frequent Words*.

Expressões Multi-palavras

Expressões multi-palavras são expressões que contêm 2 palavras ou mais. Existem várias estratégias para calcular expressões multi-palavras (MWE), e para os recursos MWE fornecidos com este corpus foram utilizadas três estratégias: Associação estatística; Co-ocorrência de palavras; Delimitadores de frases.

Associação estatística

É uma estratégia que identifica relações estatísticas entre palavras que aparecem em sequência (pares de palavras). Os pares de palavras que têm uma relação estatística significativa são susceptíveis de ser uma expressão de multipalavras (multi-word expression)(MWE) ou parte de um MWE. A técnica utilizada para calcular as MWEs foi *Pointwise Mutual Information* (PMI). O cálculo do PMI pode ser representado

$$PMI = \log\left(\frac{P(a,b)}{P(a)P(b)}\right)$$

onde "a" é a primeira palavra em uma sequência de duas palavras, "b" é a segunda palavra em uma sequência de duas palavras e "prob" é a probabilidade de uma palavra no corpus. Pares de palavras que tiveram um $PMI > 0$ foram considerados como bigramas. Os trigramas foram computados pelo cálculo da média PMI Para cada relação da sequência de 3 palavras. Esta técnica produziu: 6141 trigramas e 6491 bigramas. O bigrama e o trigrama estão localizados na pasta *Multi-Word Expressions* e estão disponíveis como: Arquivos de texto e formato de "pickle" em Python (Dicionários). Exemplos de MWEs extraídos com este método estão documentados na Tabela 4.

Bigramas
aparelhos celulares, principal adversário, laudo técnico, menor disponibilidade, tão difícil, investimento social, maior processadora, momento oportuno, agências internacionais, jogadas ofensivas, clubes participantes, primeira greve
Trigramas
contra a corrupção, dados foram divulgados, postos de combustíveis, investiga um esquema, abriu as portas, mês passado foi, plantio de mudas, área de educação, reduziu sua estimativa

Tabela 4: Amostra de MWE Extraído com Associação Estatística.

Co-ocorrência de palavras

Co-ocorrência é outra técnica a partir da qual os MWEs podem ser detectados. As palavras podem ser representadas como vetores, onde os valores no vetor são pesos que representam co-ocorrência com outras palavras. Esta representação combinada com skip-gramas pode ser usada para identificar frases (Mikolov et al., 2013) dentro de um fluxo de unigramas.

Este corpus vem com dois modelos que permitem a detecção de bigramas ou trigramas. Os modelos foram gerados a partir de Gensim[7]. Os modelos estão localizados na pasta *Word Vectors* e estão disponíveis como um formato Python "pickle".

Delimitadores de frases

Delimitador de frase é a pontuação que delimita palavras ou frases. Esta técnica identifica pares de marcas de citação ou sinais de pontuação que delimitam palavras, bigramas ou trigramas. Suponha-se que esses delimitadores fossem utilizados por jornalistas para indicar frases específicas de "domínio". Esta técnica identificou 1026 palavras, bigramas ou trigramas.

[6] Tais como https://snowballstem.org/algorithms/portuguese/stop.txt

[7] http://radimrehurek.com/gensim/models/phrases.html

6 Treebank

Uma árvore de dependência é uma forma de representação de dependências léxicas entre palavras e/ou frases. Uma coleção de árvores de dependência é conhecida como *treebank*. São relativamente poucos os *treebanks* portugueses quando comparados com o inglês. A mais conhecida *treebank* portuguesa é "Floresta" (Afonso et al., 2002).

Árvores de dependência têm sido usadas em tarefas comuns de processamento de língua natural (Qiu et al., 2009), tais como extração de relação causal (Khoo et al., 2000), área de pesquisa que a liberação deste corpus se destina a incentivar.

O *treebank* fornecido com este corpus consiste de 27931 sentenças que foram selecionadas aleatoriamente e analisadas com o analisador *LX-Dependency* (Rodrigues et al., 2014) cuja saída está em conformidade com a do analisador de *Stanford (Stanford Parser)*.

Em termos de avaliação do analisador *LX-Dependency*, o mesmo possui o UAS (*Unlabeled Attachment Score*) de 94,42 e a sua LAS (*Label Attachment Score*) é de 91,23 (Silva et al., 2010).

Uma saída típica do analisador é a seguinte:

```
(ROOT (S (NP (N' (N' (N Produção) (A
global)) (PP (P de) (NP (N açúcar)))))
(VP (V deve) (VP (V crescer) (PP (P
para) (NP (N' (N 165,1) (N' (N milhões)
(PP (P de) (NP (N toneladas)))))))))))
```

As dependências representadas por esta saída são apresentados na Figura 6.

7 Recursos de Relações entre Palavras

Este corpus contém modelos que podem ajudar na detecção de relações entre palavras ou frases. Os recursos liberados são métodos estatísticos, que são Vetores de palavras e Modelagem de tópicos; Estes modelos foram gerados com a biblioteca Gensim Python. Os recursos estão localizados nas pastas *Word Vector* e *Topic Resources*, respectivamente.

Vetores de Palavras

A representação de vetor de palavra é uma representação que trata palavras como vetores. Os vetores representam a co-ocorrência de uma determinada palavra com outras palavras no vocabulário. A frequência de co-ocorrência é representada como um peso. Os vetores são sistemas de coordenadas, portanto as semelhanças entre os vetores de palavras podem ser representadas como um ângulo. Isso permite o uso de medidas de similaridade como a similaridade de Cosseno para calcular a semelhança semântica entre as palavras.

O corpus tem um modelo de vetor de palavras que foi treinado a partir da informação no corpus. Para ilustrar a capacidade do modelo de vetor de palavras para identificar palavras relacionadas, um simples experimento foi conduzido para calcular o vizinho mais próximo com uma pequena seleção de palavras. As pontuações de similaridade foram computadas usando as chamadas de função Gensim[8]. A faixa de pontuação possível foi $0.0 \leq s \leq 1.00$, onde 1 com o maior índice de similaridade e 0 o menor. Os resultados são apresentados na Tabela 5. Os resultados mostram claramente que os pares de palavras com alta pontuação tinham similaridade, no entanto, os pares de palavras com as pontuações mais baixas não tinham relações óbvias. Os recursos de vetores de palavras estão localizados: */Data/News Stories/Topic Resources/Word Vectors/*.

Palavra	Palavras mais próximas	Palavras mais distantes
Etanol	Biocombustível (0.85), Álcool hidratado (0.84), Combustível (0.81), Álcool (0.87), Álcool anidro (0.81)	Vice-liderança(-0.47), Limão (-0.55), Sábado (-0.48), Rocher (-0.48)
Milho	Trigo (0.83), Soja (0.88), Grão de bico (0.84), Algodão (0.84)	Jogador Real (-0.46), Atenção (-0.45), Bonito (-0.44), Frutal, MG (-0.46)
Gasolina	Diesel (0.79), Combustível (0.81), Álcool (0.80)	Eroles (-0.46), PM (-0.42), Exultos (-0.42), Titã (-0.42)
Chuva	Tempestades (0.75), Sopros (0.78), Nuvens (0.74), Chuva (0.73), Isolado (0.74)	Discrepante (-0.39), Estradas (-0.39), T.M. (-0.36)

Tabela 5: Palavras com vizinhos mais próximos e mais distantes.

Os experimentos foram repetidos para verbos causais. Os verbos causais são verbos que descrevem uma relação causal entre eventos de causa e efeito. Os resultados para a experimento do verbo causal são demonstrados na Tabela 6. Os resultados mostram claramente que os vizinhos mais próximos têm propriedades causais. Isso tem implicações para a extração de relação causal, já que no momento da escrita não havia uma estratégia de extração de relação causal publicada que usasse vetores de palavras.

[8] https://radimrehurek.com/gensim/models/word2vec.html

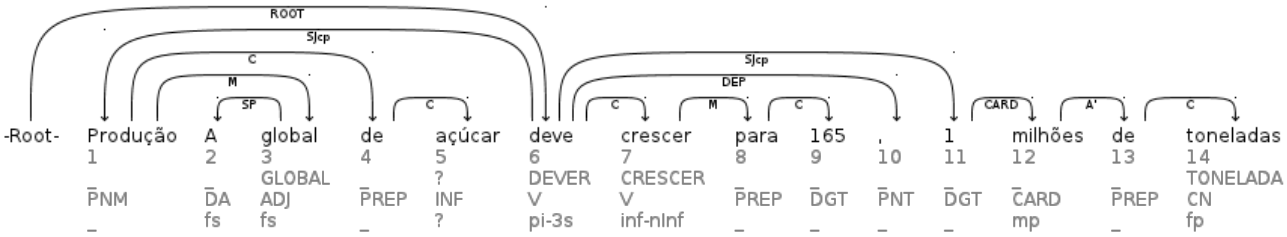

Figura 6: Dependências Léxicas

Verbo	Palavras mais próximas
causar	provocar (0.83), causam (0.67), sofrer (0.68), afetar (0.65), provoca (0.63)
afetar	prejudicar (0.85), comprometer (0.74), favorecer (0.73)
provocar	causar (0.82), gerar (0.71), sofrer (0.74)
causam	provoca (0.70), provocam (0.83)
provocam	causam (0.84), provoca (0.73)

Tabela 6: Verbos causais e seus vizinhos mais próximos.

Os vetores de palavras também podem ser usados para identificar frases semelhantes a uma frase de origem. A biblioteca Gensim fornece uma função de similaridade para n-gramas, que foi utilizada nos experimentos de bigramas e trigramas conduzidos neste artigo.

Os experimentos de bigramas usaram os seguintes bigramas de fonte selecionados aleatoriamente: Aparelhos celulares; Maior processadora; Dilma Rousseff; Receita bruta. A partir da qual foram calculados os bigramas mais próximos. Os resultados estão documentados na Tabela 7.

Bigramas	Mais próximos
aparelhos celulares	telefones móveis, canais eletrônicos, aparelhos eletrônicos, paredes celulares, equipamentos eletrônicos, caixas eletrônicos
maior processadora	maior importadora , maior produtora, maior produção, maior trading , maior exportadora, maior comercializadora, produção maior
Dilma Rousseff	Michel Temer, possível impeachment, eventual afastamento, recém-eleito presidente
receita bruta	captação líquida, dívida líquida, renda líquida, margem líquida

Tabela 7: Bigramas frequentes e seus vizinhos mais próximos.

O experimento do trigrama selecionou aleatoriamente trigramas e calculou seus vizinhos mais próximos. A técnica utilizada foi idêntica à utilizada para o experimento com bigrama. Os trigramas para esta experimento foram: Ministério da Cultura; Moagem de cana; Cultivares de soja. Os resultados estão descritos na Tabela 8.

Trigramas	Mais próximos
Ministério da Cultura	secretário-executivo do Ministério , Secretaria da Educação, ministro da Educação, Secretaria da Fazenda
moagem de cana	toneladas de cana-de-açúcar volume de moagem, safra de cana-de-açúcar, oferta de cana-de-açúcar, produção de cana-de-açúcar, capacidade de moagem
cultivares de soja	plantio de milho lavouras de milho lavouras de café

Tabela 8: Trigramas frequentes e seus vizinhos mais próximos.

Os experimentos com múltiplas palavras mostram que, embora os n-gramas mais próximos fossem compostos de sinônimos semelhantes semânticos, embora houvesse alguns erros óbvios. Exemplos de erros:

1. Aparelhos celulares e paredes celulares.
2. Aparelhos celulares e caixas eletrônicos.

Apesar dos erros, é claro que os experimentos retornam informações semânticas semelhantes nos n-gramas.

Modelagem de Tópicos

A modelagem de tópicos é um método não-supervisionado para agrupar palavras que co-ocorrem no mesmo tópico. A modelagem de tópicos pode ser usada para calcular semelhanças entre: frases e documentos.

Este corpus contém um número de modelos pré-treinados, bem como a distribuição de

tópicos pré-computados para cada documento no corpus. Os modelos pré-treinados têm uma série de variações de hiper parâmetros. As duas principais variáveis são: técnica de amostragem estatística *Latent Dirichlet Allocation* (LDA) ou *Latent Semantic Indexing* (LSI) (Blei et al., 2003) e 2. número de tópicos. Existem 5 modelos que usam LDA. Os modelos usam uma variedade de tópicos na faixa $500 \leq s \leq 2500$. O número de tópicos é incrementado em 500 para cada incremento do modelo. O modelo LSI tem um número de tópicos de 2000, o número de tópicos foi determinado pelo trabalho realizado por (Drury et al., 2015).

8 Informações de Nível de Documento

Informações de nível de documento no contexto deste artigo são aquelas que descrevem informações contidas em um documento individual. Existem 4 tipos de informações do documento: Distribuição do tópico; Orientação do sentimento; Número do grupo; e Frases de causa.

Os recursos estão localizados respectivamente nas pastas *Topic Resources*, *Sentiment*, *Clusters* e *Cause Phrases*.

Distribuição do Tópico

As informações do documento de distribuição de tópicos estão contidas em um arquivo de texto. Cada linha dentro do arquivo de texto representa um único documento. Cada linha contém o nome do documento e uma coleção de números de tópicos com uma probabilidade. O separador entre o número do tópico e sua probabilidade é um espaço, e o separador entre o número de tópicos e os pares de probabilidade é uma tabulação. A distribuição de probabilidade foi calculada com LDA e 2000 tópicos. Estes valores foram derivados do trabalho realizado por Drury et al. (2015).

Orientação do Sentimento

A orientação do sentimento para um documento foi alcançada contando o número de palavras com uma orientação sentimental. As palavras com uma orientação do sentimento neste caso são palavras com uma orientação positiva ou negativa do sentimento. As palavras com uma orientação neutra são ignoradas porque dominariam o documento. O cálculo pode ser representado:

$$S = freq(W_p) - freq(W_n),$$

onde $freq$ é a frequência de palavras com uma determinada orientação de sentimento,W_p são palavras com uma orientação positiva, W_n são palavras com orientação negativa e S é a orientação do sentimento. Documentos com uma pontuação de: 1. $S < 0$ recebem uma orientação negativa, 2. $S > 0$ recebem uma orientação positiva e 3. $S = 0$ recebem uma orientação neutra. O recurso é um arquivo de dicionário "pickled". O arquivo contém: a localização relativa de um documento, nome do arquivo e orientação de sentimento. Os valores das chaves são o local do arquivo e os valores são a orientação do sentimento.

Agrupamento

Documentos relacionados podem ser detectados por um processo de agrupamento. O processo de agrupamento para este corpus foi conseguido usando K-means, e a distribuição tópica acima mencionada. K foi ajustado para 200 usando Davies Bouldin Index (DBI) para calcular a "qualidade" de várias configurações de agrupamento. A medida de distância que foi usada para computar os agrupamentos foi a distribuição de tópicos de cada documento.

Os *clusters* e seus documentos componentes são fornecidos em um formato de dicionário "pickled". A chave é um número de cluster nominal e o valor são os documentos. Para ilustrar a semelhança de documentos que fazem parte do mesmo cluster são apresentados na Tabela 9.Os documentos contêm o mesmo tema da predição de colheita. O uso de tópicos em vez de semelhança de palavras produziu clusters que contêm o mesmo tema, ao invés da mesma palavra.

Documento 1	**Documento 2**
As usinas e destilarias do Centro-Sul do Brasil dão início nesta sexta, dia 1º de abril, a mais uma safra de cana-de-açúcar, com perspectivas favoráveis. A principal região produtora do país irá processar em 2016/2017 619,37 milhões de toneladas de cana (+2,3%).	A Organização Internacional do Café (OIC), em sua primeira estimava para a produção mundial no ano-safra 2015/2016, prevê colheita de 143,4 milhões de sacas de 60 kg, indicando um aumento modesto de 1,4% em relação ao ano-safra de 2014/2015 (141,4 milhões).....

Tabela 9: Fragmentos de texto dos documentos no mesmo grupo (*cluster*).

Relações Causais

Os documentos anotados fornecem uma relação de causa anotada, mas para extrair todas as relações de causa pode ser uma tarefa onerosa. O

corpus fornece uma lista de relações de causa pré-extraídas. A relação de causa é um arquivo delimitado por tabulação que representa a relação de causa como um triplo:

1. Evento de causa.
2. Ligação causal.
3. Evento de efeito.

Cada triplo tem um nome de documento que é o documento onde reside a relação causal. As palavras de parada *(stop-words)* foram removidas das relações causais. Uma amostra de relações causais pode ser encontrada na Tabela 10.

Relações Causais
governo aumente etanol anidro gasolina
clima seco produzidas milhoes toneladas acucar
taxa declinio diminuído levantando expectativas setor
chuvas últimos causa máquinas conseguem entrar lavoura

Tabela 10: Amostra de Relações Causais

9 Licenciamento

Este corpus é lançado sob a *Creative Commons License (4.0)* (https://wiki.creativecommons.org/wiki/Text).
É intenção dos autores que este corpus seja utilizado em sua amplitude, consequentemente esta licença foi escolhida porque permite o uso comercial e de redistribuição.

Este corpus se qualifica para a liberação de acordo com a legislação de uso justo[9] porque: é transformador, e nenhum ganho monetário será exigido para sua liberação.

10 Trabalhos Futuros

Pretende-se em trabalhos futuros considerar a avaliação de outras ferramentas que realizam detecção de entidades nomeadas, assim como outras formas de detecção de expressão multi-palavras, considerando o uso de opções como: OpenNLP, FreeLing, PALAVRAS e etc. Aplicar anotações baseadas em XML em relações causais que apresentam estruturas fracas. Além disso, vamos considerar alternativas abertas ao LX-Dependency, como por exemplo o UDPortugueseBR[10]

11 Conclusão

Este artigo descreve um corpus português-brasileiro que contém notícias relacionadas a agricultura. Essas notícias têm anotações causais e sentimentais relacionadas a informações temporais, bem como anotações de entidades nomeadas. O corpus contém recursos de linguagem, tais como: árvores de dependência, modelos de tópicos e modelos de vetor de palavras, bem como meta-informações, como distribuição de tópicos. Além disso, contém informações sobre o nível do documento, como distribuição de tópicos e informações sobre o sentimento.

Este recurso que acreditamos ser único e substancial, foi liberado para incentivar pesquisas de mineração de texto no campo da agricultura, bem como pesquisas em áreas relacionadas, como relação de causalidade e extração de conhecimento.

Agradecimentos

Esta pesquisa teve apoio financeiro das agências brasileiras: FAPESP (processos 15/14228-9 e 11/20451-1) e CNPq (processo 302645/2015-2). Somos gratos aos árbitros pelos comentários e sugestões no desenvolvimento deste trabalho.

Referências

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002. Floresta sintá (c) tica: A treebank for portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*, 1698–1703.

Altenberg, Bengt. 1984. Causal linking in spoken and written English. *Studia Linguistica* 38(1). 20–69.

Alvim, Leandro, Paula Vilela, Eduardo Motta & Ruy Luiz Milidiú. 2010. Sentiment of financial news: a natural language processing approach. Em *1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology*, edição online.

Balage Filho, Pedro P., Thiago A. S. Pardo & Sandra M. Alusio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. Em *9th Brazilian Symposium*

[9] https://www.copyright.gov/fair-use/more-info.html

[10] https://github.com/UniversalDependencies/UD_Portuguese-BR

in Information and Human Language Technology (STIL), 215–219.

Bethard, Steven. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 10–14.

Bittar, André. 2010. *Building a TimeBank for French: a reference corpus annotated according to the ISO-TimeML standard*: Paris 7. Tese de Doutoramento.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3. 993–1022.

Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas & Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 11–31. Linguateca.

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta & Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. Em *5th Linguistic Annotation Workshop*, 143–151.

Chaves, Marcírio Silveira, Larissa A. de Freitas, Marlo Souza & Renata Vieira. 2012. Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. Em *International Conference on Application of Natural Language to Information Systems*, 296–301.

Drury, Brett & Alneu de Andrade Lopes. 2014. A comparison of the effect of feature selection and balancing strategies upon the sentiment classification of Portuguese news stories. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 413–417.

Drury, Brett & Alneu de Andrade Lopes. 2015. The identification of indicators of sentiment using a multi-view self-training algorithm. *Oslo Studies in Language* 7.

Drury, Brett, Jorge Carlos Valverde-Rebaza & Alneu de Andrade Lopes. 2015. Causation generalization through the identification of equivalent nodes in causal sparse graphs constructed from text using node similarity strategies. Em *International Symposium on Information Management and Big Data*, 58–65.

Fonseca, Erick R. & João Luís G. Rosa. 2013. A two-step convolutional neural network approach for semantic role labeling. Em *International Joint Conference on Neural Networks*, 2955–2961.

Forascu, Corina & Dan Tufiș. 2012. Romanian TimeBank: An annotated parallel corpus for temporal information. Em *Eight International Conference on Language Resources and Evaluation (LREC)*, 3762–3766.

Freitas, Cláudia, Eduardo Motta, R. Milidiú & Juliana César. 2012. Vampiro que brilha... rá! desafios na anotaçao de opinião em um corpus de resenhas de livros. Em *XI Encontro de Linguística de Corpus*, s/p.

Freitas, Larissa A. & Renata Vieira. 2013. Ontology based feature level opinion mining for Portuguese reviews. Em *22nd International Conference on World Wide Web (WWW)*, 367–370.

Jung, Hyuckchul & Amanda Stent. 2013. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 20–24.

Khoo, Christopher, Syin Chan & Yun Niu. 2002. The many facets of the cause-effect relation. Em Rebecca Green, Carol A. Bean & SungHyon Myaeng (eds.), *The Semantics of Relationships*, vol. 3 Information Science and Knowledge Management, 51–70. Springer.

Khoo, Christopher S. G., Syin Chan & Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. Em *38th Annual Meeting on Association for Computational Linguistics*, 336–343.

Kolya, Anup Kumar, Amitava Kundu, Rajdeep Gupta, Asif Ekbal & Sivaji Bandyopadhyay. 2013. JU_CSE: A CRF based approach to annotation of temporal expression, event and temporal relations. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 64–72.

Laokulrat, Natsuda, Makoto Miwa, Yoshimasa Tsuruoka & Takashi Chikayama. 2013. Uttime: Temporal relation classification using deep syntactic features. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 88–92.

Liu, Bing & Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. Em Charu C. Aggarwal (ed.), *Mining text data*, 415–463. Springer.

Llorens, Hector, Estela Saquete & Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. Em *5th International Workshop on Semantic Evaluation (SemEval)*, 284–291.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in neural information processing systems*, 3111–3119.

Pustejovsky, James, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer & Graham Katz. 2003a. TimeML: robust specification of event and temporal expressions in text. Em Mark T. Maybury (ed.), *New directions in question answering*, 28–34. AAAI Press.

Pustejovsky, James, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro & Marcia Lazo. 2003b. The TIMEBANK corpus. Em *Corpus linguistics*, 647–656.

Qiu, Guang, Bing Liu, Jiajun Bu & Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. Em *International Joing Conference on Artificial Intelligence (IJCAI)*, vol. 9, 1199–1204.

Rodrigues, João, Francisco Costa, João Silva & António Branco. 2014. Automatic syllabification of portuguese. *Encontro Anual da Associação Portuguesa de Linguística* 715–720.

Shams-Eddien, Katrin. 2002. *Beth Levin's English verbs classes and alternations*. Free University of Berlin.

Silva, Joao, António Branco, Sérgio Castro & Ruben Reis. 2010. Out-of-the-box robust parsing of Portuguese. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 75–85.

Silva, Mário J., Paula Carvalho & Luís Sarmento. 2012. Building a sentiment lexicon for social judgment mining. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 218–228.

Silva, Mário J., Paula Carvalho, Luís Sarmento, Pedro Magalhães & Eugénio Oliveira. 2009. The design of OPTIMISM, an opinion mining system for Portuguese politics. Em *New trends in artificial intelligence: Proceedings of EPIA*, 12–15.

Souza, Marlo, Renata Vieira, Débora Busetti, Rove Chishman & Isa Mara Alves. 2011. Construction of a Portuguese opinion lexicon from multiple resources. Em *8th Brazilian Symposium in Information and Human Language Technology*, 59–66.

Strötgen, Jannik & Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. Em *5th International Workshop on Semantic Evaluation*, 321–324.

Thomson, Judith Jarvis. 1987. Verbs of action. *Synthese* 72(1). 103–122.

UzZaman, Naushad & James F. Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. Em *5th International Workshop on Semantic Evaluation (SemEval)*, 276–283.

Vendler, Zeno. 1967. Causal relations. *The Journal of Philosophy* 64(21). 704–713.