Utilizando Features Linguísticas Genéricas para Classificação de Triplas Relacionais em Português

George C. G. Barbosa[1], **Daniela Barreiro Claro**[1]

[1]Formalismos e Aplicações Semânticas Research Group (FORMAS)
LaSiD - Departamento de Ciência da Computação
Instituto de Matemática e Estatística – Universidade Federal da Bahia
Av. Adhemar de Barros, s/n, Ondina, Salvador - Bahia - Brasil

gcgbarbosa@gmail.com, dclaro@ufba.br

Resumo. *A quantidade de textos gerados diariamente na web torna cada vez mais difícil a análise e extração de informações desses dados. Retirar informação útil de forma automática de textos é uma tarefa difícil, dada a complexidade e infinidade de formas com que as pessoas podem se expressar utilizando a linguagem natural. A tarefa de Extração de Informação Aberta tem o papel de automatizar o processamento de repositórios tais como a Web. Esta abordagem pode ser classificada em duas etapas: (i) extração e (ii) classificação. A proposta desse trabalho é, na etapa de classificação, utilizar um conjunto de features genéricas que não contém termos presentes em um idioma específico. Experimentos foram realizados em Português do Brasil nos quais as features genéricas obtiveram uma acurácia média de 70% contra 55% das features propostas em [Fader et al. 2011].*

1. Introdução

Mais de 80% das informações da Web são armazenadas em formato de texto nos mais diferentes idiomas [Barion and Lago 2008]. Estima-se que 50% do conteúdo disponível em *websites* está escrito em Inglês[1]. Os principais trabalhos da área de Extração da Informação (IE, do Inglês *Information Extraction*) utilizam metodologias desenvolvidas com base no Inglês [Banko et al. 2007] [Wu and Weld 2010] [Fader et al. 2011] [Schmitz et al. 2012] [Del Corro and Gemulla 2013] [Angeli et al. 2015]. Os dados textuais em diferentes idiomas, que somam a outra metade do conteúdo disponível, têm recebido pouca atenção [Gamallo et al. 2012] e muitos esforços têm sido realizados na tentativa de analisá-los [Fader et al. 2011]. A Extração da Informação (IE) é a tarefa de aquisição de informação a partir de dados não estruturados ou semi-estruturados. É possível classificá-la em aberta ou tradicional. A IE tradicional tem como objetivo a extração de informação em um domínio específico, geralmente um conjunto pré-especificado de expressões [Schmitz et al. 2012]. Já a IE aberta (OIE, do Inglês *Open Information Extraction*) tem como principais objetivos: (i) independência de domínio, (ii) extração não supervisionada e (iii) escalabilidade para grandes bases de dados [Del Corro and Gemulla 2013].

As tarefas de Processamento de Linguagem Natural (NLP, do Inglês *Natural Language Processing*) tais como: *Tokenization*, *Sentence Splitting* e *Part-of-Speech tagging* -

[1]https://w3techs.com/technologies/overview/content_language/all

POS [Manning et al. 2014] são essenciais para a IE em dados textuais, porém, são dependentes do idioma no qual o texto foi escrito.

Os trabalhos recentes em OIE podem ser classificados em quatro tipos. São eles (i) dados de treinamento e análise rasa, (ii) dados de treinamento e análise de dependência, (iii) baseado em regras e análise rasa e (iv) baseado em regras e análise de dependência [Gamallo 2014]. Essa classificação é feita de acordo com a metodologia empregada para a extração das triplas relacionais.

Os métodos de OIE baseados em análise rasa são realizados em duas etapas, sendo a primeira etapa a extração e, posteriormente, a classificação das relações extraídas. A classificação é a tarefa que define se uma extração realizada é válida ou inválida com o objetivo de conferir ao método uma melhor precisão nos resultados. Alguns trabalhos encontrados na literatura utilizam métodos de classificação baseados em *features* dependentes de características linguísticas [Fader et al. 2011] [Xu et al. 2013] [Pereira and Pinheiro 2015]. Entende-se por dependência de idioma a utilização de funções linguísticas que estão presentes no idioma alvo do estudo, mas não fazem parte de outros idiomas. Por exemplo, o Português não apresenta nenhum recurso similar à apóstrofe (*genitive marker* ('s)) do Inglês. Com isso, a utilização dessa função linguística em alguma *feature* tornaria difícil a adaptação do método para o Português.

O Inglês possui ferramentas e recursos linguísticos sofisticados que outros idiomas ainda não possuem. Em geral, as ferramentas construídas para o Inglês não são aplicáveis a outros idiomas. Assim, este trabalho propõe um método de classificação baseado em *features* independentes do idioma. A hipótese é de que *features* genéricas em relação ao idioma podem apresentar resultados superiores a *features* dependentes. As principais contribuições do presente trabalho são: (i) desenvolver um método para classificação de triplas relacionais através de *features* genéricas independentes de idioma e (ii) avaliar este método para outro idioma diferente do Inglês, neste trabalho o Português do Brasil.

Este trabalho está organizado como segue: a Seção 2 traz os trabalhos relacionados. A seção 3 define o problema que este trabalho trata e a Seção 4 descreve a metodologia utilizada. Na Seção 5 os experimentos realizados são apresentados. A Seção 6 apresenta os resultados obtidos para cada experimento e por fim a Seção 7 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Os primeiros trabalhos em OIE faziam uso da metodologia de dados de treinamento e análise rasa (categoria (i)). Tendo como o pioneiro o *TextRunner* [Banko et al. 2007], que usava uma abordagem baseada em etiquetagem morfossintática (POS, do inglês *Part-of-Speech*) e etiquetagem de sintagmas nominais (NP, do inglês *Noun Phrase*). O *TextRunner* utilizava como método de classificação das extrações realizadas o *Naïve Bayes*, tendo como base de treino exemplos gerados a partir do *Penn Tree Bank*. Outros sistemas como o *ReVerb* [Fader et al. 2011] e o WOE^{POS} [Wu and Weld 2010] utilizaram uma abordagem similar ao *TextRunner*, apresentando melhorias como o desenvolvimento de classificadores mais robustos.

Em seguida, foram introduzidos trabalhos na literatura baseados em análise de dependência (categoria (ii)). Os trabalhos mais conhecidos nesta classe são o WOE^{Parse} e

o OLLIE. O WOE^{POS} faz uso de dados etiquetados do Wikipedia como treinamento para a detecção de triplas relacionais [Wu and Weld 2010]. O *OLLIE* é baseado em extrações de alto grau de confiança obtidos pelo *ReVerb* para detecção de padrões derivados da análise de dependência [Schmitz et al. 2012]. Na categoria (iii) os trabalhos mais relevantes são o *ExtrHech* [Zhila and Gelbukh 2013] e o LSOE [Xavier et al. 2013] que são baseados em padrões léxicos e sintáticos extraídos manualmente a partir de etiquetagem morfossintática.

Utilizando regras extraídas manualmente a partir do método de análise de dependência (categoria (iv)) destacam-se o CSD [Gamallo et al. 2012] e o *ClausIE* [Del Corro and Gemulla 2013]. Em [Angeli et al. 2015] uma abordagem similar ao *ClausIE* é utilizada, com a diferença das sentenças serem separadas em núcleos semânticos, de forma que as relações extraídas possuam a menor quantidade de *tokens* possível. Isso resulta no aumento da qualidade das extrações e facilita a utilização das triplas resultantes para outros fins (e.g. construção de ontologias e sistemas de pergunta e resposta).

Os trabalhos recentes baseados em metodologias mais robustas (análise de dependência e anotação de papéis semânticos) não utilizam a tarefa de classificação [Del Corro and Gemulla 2013][Gamallo et al. 2012][Schmitz et al. 2012]. Estes trabalhos extraem um número muito maior de triplas relacionais quando comparados a outros baseados em análise rasa, porém, em termos de acurácia ambos possuem desempenho similar [Del Corro and Gemulla 2013].

O presente trabalho pode ser aplicado às triplas extraídas pelos trabalhos citados nesta seção, aumentando a qualidade em seus resultados através da classificação binária (válida ou inválida), evitando que informações inválidas sejam disponibilizadas no resultado final.

3. Definição do Problema

Sistemas de OIE extraem triplas do tipo *(E1, SR, E2)*, onde *E1* e *E2* são sintagmas nominais reconhecidos no texto, e *SR* é um sintagma relacional que relaciona *E1* e *E2* [Gamallo 2014]. Para ilustrar, tem-se a sentença:

"A cidade de São Paulo detém 15% das indústrias de produtos químicos do país, parte dos 53% do total de empresas desse setor instaladas no Estado."

Considera-se uma extração válida:

(São Paulo, detém, 15% das indústrias de produtos químicos do país)

A seguinte extração é considerada inválida pois traz em *E1* uma *string* contendo uma porcentagem ao invés de um sintagma nominal, o que a torna incoerente:

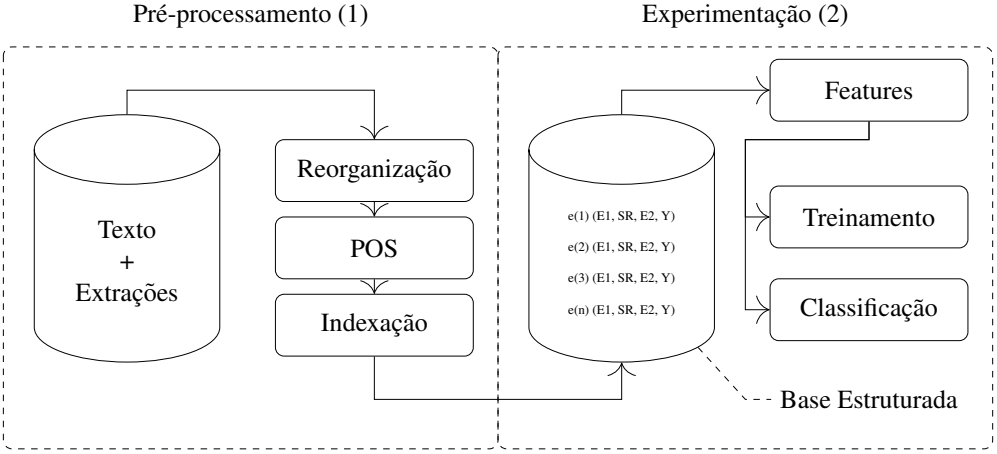
(53%, instaladas no, Estado)

O presente trabalho tem como objetivo classificar em válidas e inválidas através de algoritmos de aprendizado de máquina as relações extraídas a partir de sentenças escritas em linguagem natural. A finalidade do método de classificação desenvolvido é ser utilizado por qualquer sistema que realiza extrações em textos, garantindo a este maior acurácia em seus resultados. O intuito é que o método de classificação baseado em *features* genéricas possa ser aplicado a alguns dos principais idiomas do mundo (Inglês, Espanhol, Francês e Português).

4. Metodologia Proposta

A Figura 1 descreve o fluxo da experimentação das *features* genéricas propostas neste trabalho. Ela está dividida em (1) pré-processamento e (2) experimentação. O Corpus da Folha de São Paulo foi utilizado como fonte de dados [CETENFOLHA 2008]. A primeira etapa realiza a reorganização das relações, a etiquetagem morfossintática e a indexação com o objetivo de preparar o banco para o cálculo das *features*. Em seguida, o banco de dados é utilizado para treinamento e classificação. Na classificação são avaliados os métodos de aprendizado de máquina *Logit*, SVM, *NaiveBayes* e a Árvore de Decisão C5.0. Os resultados obtidos na etapa 2 são avaliados através das métricas de Acurácia, Precisão, Revocação e F1 [Forman 2003].

Figura 1. Fluxo da experimentação das *features* genéricas.



4.1. Conjunto de Dados

A quantidade de recursos disponíveis em NLP para o Inglês é consideravelmente maior do que para Português. Não foram encontrados conjuntos de dados manualmente etiquetados em Português do Brasil para a tarefa de OIE. Assim, para viabilizar o experimento foi utilizado um conjunto de dados com 500 sentenças aleatórias obtidas do [CETENFOLHA 2008], denominado CETENFOLHA-500.

A partir de (CETENFOLHA-500) foram extraídas 904 triplas relacionais utilizando uma ferramenta baseada em uma adaptação do ReVerb para Português do Brasil. A base de dados resultante está organizada como segue: (i) uma linha contendo a sentença original S_1 e N linhas subsequentes contendo as relações extraídas a partir de S_1, sendo este padrão repetido para as sentenças S_2 até S_n. As triplas extraídas foram avaliadas por dois especialistas e uma coluna contendo o resultado da análise foi adicionada a base de dados (1 = válida, 0 = inválida)

4.2. Features Genéricas

As Tabelas 1 e 2 apresentam uma comparação das *features* apresentadas por [Fader et al. 2011] e as *features* genéricas avaliadas neste trabalho para o Português. Como as *features* presentes em [Fader et al. 2011] foram aplicadas apenas ao Inglês, é possível observar que a maioria delas refere-se as características específicas deste idioma. Por exemplo, observa-se as *features* 2-4, que possuem palavras do Inglês e dificilmente

terão correspondentes em outras línguas ('for', 'on', 'of'). Já a *feature* 6 cita palavras do tipo "WH" (e.g. 'What', 'Why', 'Where'), que são marcadores de perguntas comuns no Inglês. Isso dificulta a adaptação de um classificador baseado nessas *features* para outro idioma, por exemplo, o Português.

As *features* genéricas apresentadas em [Barbosa et al. 2016] foram adaptadas do ReVerb para não ter dependência de características do Inglês. Cada *feature* dependente na Tabela 2 foi analisada e, quando possível, uma *feature* considerada não dependente foi proposta em seu lugar, dando origem as *features* da Tabela 1.

	Feature
1	Tamanho de S - Tamanho de E1+SR+E2
2	Número de verbos na SR
3	Tamanho de SR
4	Existe uma pergunta a esquerda da SR em S
5	A sentença tem 10 palavras ou menos
6	Distância entre E1 e SR
7	Existe uma preposição a esquerda de E1
8	Tamanho de E2
9	Distância entre E2 e SR
10	Número de preposições na SR
11	Número de substantivos a direta de E2
12	Tamanho de E1
13	Tamanho de S
14	Número de nomes próprios em E1
15	Número de nomes próprios em E2

Tabela 1. *Features* genéricas propostas em [Barbosa et al. 2016]

S: sentença na qual é feita a extração
$E1$ e $E2$: sintagmas nominais da tripla da relação
SR: sintagma relacional da extração

	Feature
1	Extração cobre todas as palavras da sentença
2	A ultima preposição na relação é 'for'
3	A ultima preposição na relação é 'on'
4	A ultima preposição na relação é 'of'
5	A sentença tem 10 palavras ou menos
6	Existe uma palavra com 'WH' a esquerda da relação na sentença
7	A relação corresponde ao padrão VW*P
8	A ultima preposição na relação é 'to'
9	A ultima preposição na relação é 'in'
10	A sentença tem entre 10 e 20 palavras
11	A sentença começa com E1
12	E1 é um nome próprio
13	E2 é um nome próprio
14	Existe um sintagma nominal a esquerda de E1 na sentença
15	A sentença tem mais de 20 palavras
16	A relação corresponde ao padrão V
17	Existe uma preposição a esquerda de E1 na sentença
18	Existe um sintagma a direita de E2 na sentença
19	Existe uma conjunção coordenativa a esquerda da relação na sentença

Tabela 2. Features utilizadas no ReVerb [Fader et al. 2011].

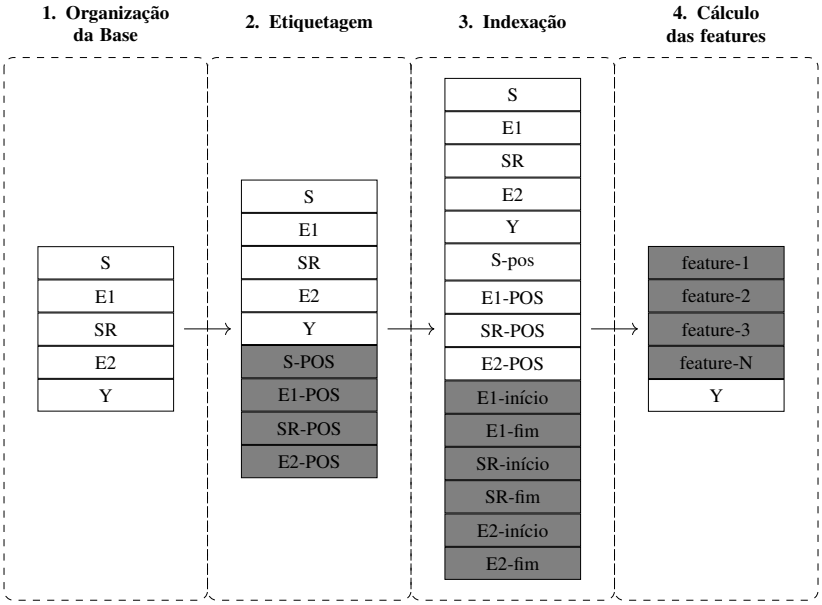
4.3. Pré-processamento

Neste trabalho foi avaliado o desempenho das *features* genéricas propostas por [Barbosa et al. 2016] para o Português. A Figura 2 detalha a etapa de pré-processamento. A primeira etapa consiste na reorganização do conjunto de dados presente na Figura 1, onde S = sentença, E1 e E2 são os sintagmas nominais e Y é a coluna contendo o resultado da análise manual (1 = válido e 2 = inválido). Na etapa 2, o conjunto de dados é etiquetado e na etapa 3, ele é indexado para tornar possível o cálculo de algumas *features*. Na etapa 4 as *features* são calculadas.

Para que algumas *features* fossem calculadas, foi necessário empregar tarefas de NLP (etapa 2 na Figura 2) no conjunto de dados citado na Seção 4.1. Para as *features* 2, 7, 10, 11, 14 e 15 na Tabela 1 que necessitam de etiquetagem morfossintática foi utilizado a ferramenta CoGrOO [Kinoshita et al. 2006]. As células em cinza na etapa 2 indicam colunas com as etiquetas morfossintáticas da sentença e da tripla relacional. Cada palavra dentro da sentença/relação é etiquetada individualmente.

A Tabela 2 apresenta *features* que usam o posicionamento das palavras de E1, SR ou E2 dentro da sentença como entrada para o cálculo (*features* 4, 6, 7, 9, e 11). Por essa razão, faz-se necessário o cálculo dos índices de início e fim de E1, E2 e SR dentro de S

Figura 2. Pré-processamento da base original



antes da etapa do cálculo das *features*. Esses índices são adicionados a base na etapa 3 (Figura 2, destacados em cinza).

O cálculo de cada *feature* consiste em executar a operação sintetizada em sua descrição e armazenar o valor obtido para ser utilizado mais tarde nas etapas de treinamento e teste (Figura 2, destacados em cinza). Apenas os valores de cada *feature* e o Y são necessários a etapa de experimentação.

Por fim, o etiquetador CoGroo faz a separação de algumas palavras durante a etiquetagem (Tabela 3). Isso resulta em um número de etiquetas maior do que o de palavras presentes na sentença. Com isso, a introdução dos índices na etapa de indexação fica prejudicada. Para solucionar este problema, as duas etiquetas das palavras divididas foram re-mapeadas em apenas uma. Este mapeamento está descrito na Tabela 3. As palavras na coluna "Exemplo" foram divididas em duas pelo CoGrOO (eg. "no" = "em" + "o"). A função de mapeamento baseada na Tabela 3 uniu as classes gramaticais das duas palavras de acordo com a coluna "Categorias" e a divisão da palavra foi desfeita.

Tabela 3. Palavras divididas na etiquetagem morfossintática via CoGrOO

Palavras	Exemplo	Palavra_POS 1	Palavra_POS 2	Categoria
no na nos nas	no	em_prp	o_art	prp+art
daí	daí	de_prp	aí_art	
pelo pela pelos pelas	pelo	por_prp	o_art	
ao aos	ao	a_prp	o_art	
do da dos das	do	de_prp	o_art	
num numa	num	em_prp	um_art	
um uma uns umas	num	em_prp	um_num	prp+num
dele dela deles delas	dele	de_prp	ele_pron-pers	prp+pron-pers
neste nesta nestes nestas	neste	em_prp	este_pron-det	prp+pron-det
naquele naquela naqueles naquelas	naquele	em_prp	aquele_pron-det	
daquele daquela daqueles daquelas	daquele	de_prp	aquele_pron-det	
nesse nessa nesses nessas	nesse	em_prp	esse_pron-det	
à	à	a_prp	a_prp	prp+prp

5. Experimentos

Dois experimentos foram realizados com a finalidade de verificar a hipótese de que o conjunto de *features* genérico apresentava acurácia mais alta que o ReVerb. O primeiro experimento objetivou verificar a significância estatística dos resultados obtidos em [Barbosa et al. 2016] para o Inglês. O segundo experimento teve por objetivo verificar o desempenho das *features* genéricas para Português e comparar seu resultado com as do ReVerb. Os testes estatísticos foram feitos utilizando a ferramenta R na versão 3.1.1. Os experimentos de classificação utilizaram o Scikit[2], uma ferramenta que disponibiliza uma implementação de alguns dos algoritmos de Aprendizado de Máquina (ML, do Inglês *Machine Learning*) mais populares nos dias atuais.

Na tentativa de apresentar maior generalização, as *features* foram testadas com vários métodos para observar o comportamento em diferentes abordagens de ML. Foram eles: Árvore de Decisão, SVM, Regressão Logística e *NaiveBayes*. Os algoritmos foram escolhidos baseados na descrição de desempenho encontrada em [Nikam 2015]. O algoritmo *Logit* foi considerado por fazer parte do artigo utilizado como referência [Fader et al. 2011].

A. Teste de significância estatística

As features genéricas utilizadas neste trabalho foram obtidas do trabalho de [Barbosa et al. 2016]. Porém, observou-se que os autores [Barbosa et al. 2016] não realizaram um teste que demonstrasse a significância estatística dos resultados apresentados. Assim, com o intuito de validar as *features* genéricas que foram utilizadas neste trabalho para o Português do Brasil, o teste de Wilcoxon foi aplicado para comparar as medianas dos valores obtidos por cada conjunto de *feature*, avaliando se um dos conjuntos tende a ter valores maiores do que o outro. Para a realização desse teste, o experimento *Cross-fold Validation* utilizando as bases descritas em [Barbosa et al. 2016] foi refeito. No teste, foram comparadas as métricas de acurácia do classificador que apresentou o melhor resultado para acurácia nos dois conjuntos de *features* (Logit).

B. Cross-fold Validation utilizando (CETENFOLHA-500)

O segundo experimento avaliou as *features* genéricas para o Português do Brasil, comparando os conjuntos das *features* das Tabelas 1 e 2 através de validação cruzada (10-fold *cross-validation*) utilizando as extrações feitas no CETENFOLHA-500. Foram calculadas as métricas de Acurácia, Precisão, Revocação e Medida-F (F1) [Forman 2003].

6. Resultados

Nesta seção são apresentados os resultados das avaliações dos conjuntos de *features* genéricos e do ReVerb. A média aritmética das métricas de Acurácia, Precisão, Revocação e F1-score são apresentados para cada um dos Algoritmos testados, bem como o teste de Wilcoxon para as acurácia do classificador Logit.

A. Teste de significância estatística

A Tabela 4 apresenta os resultados obtidos em [Barbosa et al. 2016] para o Inglês. A partir da reexecução desse experimento, foram obtidas as acurácias de cada um dos K-Folds que compõe a média apresentada para cada métrica. Dadas as medidas de acurácia do

[2]http://scikit-learn.org/

classificador Logit, o teste de Wilcoxon evidenciou que, ao nível de 5% de significância, as *features* genéricas propostas em [Barbosa et al. 2016] apresentaram desempenho superior as do ReVerb (p=0.01).

Tabela 4. Comparação entre as *features* utilizando validação cruzada em Inglês [Barbosa et al. 2016].

	Algoritmo	Acurácia	Precisão	Revocação	F1
Features Genéricas	Logit	0.689 ± 0.042	0.707 ± 0.034	0.851 ± 0.060	0.772 ± 0.033
	SVM	0.685 ± 0.025	0.698 ± 0.019	0.864 ± 0.053	0.771 ± 0.022
	C5.0	0.648 ± 0.049	0.727 ± 0.038	0.701 ± 0.062	0.718 ± 0.039
Features ReVerb	Logit	0.653 ± 0.037	0.672 ± 0.027	0.853 ± 0.036	0.752 ± 0.026
	SVM	0.643 ± 0.023	0.639 ± 0.014	0.967 ± 0.017	0.770 ± 0.013
	C5.0	0.607 ± 0.036	0.659 ± 0.025	0.743 ± 0.039	0.698 ± 0.025

B. Cross-fold Validation utilizando (CETENFOLHA-500)

A Tabela 5 apresenta o desempenho do experimento realizado utilizando a base de dados (CETENFOLHA-500). Os resultados obtidos pelas *features* genéricas são superiores aos do ReVerb em todos os classificadores e métricas avaliadas. O teste de Wilcoxon foi realizado e, ao nível de 5% de significância, observa-se que o desempenho das *features* genéricas, em relação a acurácia, é superior ao desempenho das *features* ReVerb (p=0.0001).

Tabela 5. Resultados obtidos utilizando *features* genéricas e validação cruzada para Português.

	Algoritmo	Acurácia	Precisão	Revocação	F1
Features Genéricas	SVM	0.703 ± 0.037	0.712 ± 0.023	$\mathbf{0.932 \pm 0.041}$	$\mathbf{0.807 \pm 0.024}$
	Logit	$\mathbf{0.707 \pm 0.046}$	0.736 ± 0.029	0.879 ± 0.086	0.798 ± 0.041
	C5.0	0.667 ± 0.046	$\mathbf{0.759 \pm 0.028}$	0.720 ± 0.084	0.749 ± 0.038
	NaiveBayes	0.668 ± 0.043	0.742 ± 0.034	0.731 ± 0.104	0.743 ± 0.030
Features ReVerb	SVM	0.548 ± 0.067	0.559 ± 0.103	0.370 ± 0.060	0.440 ± 0.060
	Logit	0.551 ± 0.065	0.551 ± 0.093	0.399 ± 0.086	0.458 ± 0.081
	C5.0	0.522 ± 0.063	0.515 ± 0.088	0.360 ± 0.087	0.415 ± 0.073
	NaiveBayes	0.525 ± 0.062	0.519 ± 0.086	0.357 ± 0.089	0.416 ± 0.072

7. Conclusão e Trabalhos Futuros

Neste trabalho foi avaliado o desempenho de um conjunto de *features* genéricas para classificação de triplas relacionais para o Português do Brasil. Esse conjunto tem a finalidade de ser aplicado a métodos de extração de relações garantindo-lhes uma melhor precisão, reduzindo o número de extrações inválidas. Foram realizados experimentos com o Português do Brasil, no qual as *features* genéricas apresentaram um resultado superior no Inglês, quando comparados a um conjunto de *features* dependentes do idioma. Um teste estatístico demonstrou que as *features* genéricas apresentam resultados superiores as *features* dependentes de características do idioma (p=0.0001).

A falta de grandes conjuntos de dados etiquetados para Português diminui as evidências de experimentos com este idioma. Como trabalho futuro, pretende-se utilizar métodos que não necessitem de conjuntos de dados etiquetados grandes, como, por exemplo aprendizado de máquina semi supervisionado.

Referências

Angeli, G., Premkumar, M. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction for the web. In *IJCAI*, volume 7, pages 2670–2676.

Barbosa, G. C. G., Glauber, R., and Claro, D. B. (2016). Classificação de relações abertas utilizando features independentes do idioma. In *Symposium on Knowledge Discovery, Mining and Learning*, pages 234–241.

Barion, E. C. N. and Lago, D. (2008). Mineração de textos. *Revista de Ciências Exatas e Tecnologia*, 3(3):123–140.

CETENFOLHA (2008). Corpus de extratos de textos eletrônicos nilcs/folha de são paulo. Disponível em: <http://www.linguateca.pt/cetenfolha/>. Acesso em: 2 de Maio de 2016.

Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. ACM.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305.

Gamallo, P. (2014). An Overview of Open Information Extraction (Invited talk). In Pereira, M. J. V., Leal, J. P., and Simões, A., editors, *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 13–16, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-based open information extraction. In *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, pages 10–18. Association for Computational Linguistics.

Kinoshita, J., Salvador, L., and Menezes, C. (2006). Cogroo: a brazilian-portuguese grammar checker based on the cetenfolha corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), Genoa, Italy*, pages 2190–2193.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science & Technology*, 8(1):13–19.

Pereira, V. and Pinheiro, V. (2015). Report-um sistema de extração de informações aberta para língua portuguesa. In *Proceedings of Symposium in Information and Human Language Technology*, pages 191–200. Sociedade Brasileira de Computação.

Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.

Xavier, C. C., de Lima, V. L. S., and Souza, M. (2013). Open information extraction based on lexical-syntactic patterns. In *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*, pages 189–194. IEEE.

Xu, Y., Kim, M.-Y., Quinn, K., Goebel, R., and Barbosa, D. (2013). Open information extraction with tree kernels. In *HLT-NAACL*, pages 868–877.

Zhila, A. and Gelbukh, A. (2013). Comparison of open information extraction for english and spanish. In *19th Annual International Conference Dialog*, pages 714–722.