Extração de Relações utilizando Features Diferenciadas para Português*

Relation Extraction using Different Features in Portuguese

Erick Nilsen Pereira de Souza
Universidade Federal da Bahia
ericknilsen@gmail.com

Daniela Barreiro Claro
Universidade Federal da Bahia
dclaro@ufba.br

Resumo

A Extração de Relações (ER) é uma tarefa da Extração da Informação responsável pela descoberta de relacionamentos semânticos entre conceitos em textos não estruturados. Quando a extração não é limitada por um conjunto predefinido de relações, a ER é dita Aberta, cujo principal desafio consiste em reduzir a proporção de extrações inválidas no universo de relações identificadas. As soluções atuais, baseadas em aprendizado sobre um conjunto de features linguísticas específicas, embora consigam eliminar grande parte das extrações inválidas, possuem como desvantagem a alta dependência do idioma. Tal dependência decorre da dificuldade inerente à determinação do conjunto de features mais representativo para o problema, considerando as peculiaridades de cada língua. Neste sentido, o presente trabalho propõe avaliar as dificuldades da classificação baseada em features na extração de relações semânticas abertas em Português, com o objetivo de embasar novas soluções capazes de reduzir a dependência do idioma nesta tarefa. Os resultados obtidos indicam que nem todas as *features* representativas em Inglês podem ser mapeadas diretamente para a Língua Portuguesa com méritos de classificação satisfatórios. Dentre os algoritmos de classificação avaliados, o J48 apresentou os melhores resultados com uma medida-F de 84,1%, seguido pelo SVM (83,9%), Perceptron (82,0%) e Naive Bayes (79,9%).

Palavras chave

Extração de Relações Abertas, Seleção de Características

Abstract

Relation Extraction (RE) is a task of Information Extraction (IE) responsible for the discovery of semantic relationships between concepts in unstructured text. When the extraction is not limited to a predefined set of relations, the task is called Open Relation Extraction, whose main challenge is to reduce the proportion of invalid extractions in the universe of relationships identified. Current methods based on a set of specific machine learning features eliminate much of the invalid extractions. However, these solutions have the disadvantage of being highly language-dependent. This dependence arises from the difficulty in finding the most representative set of features to the Open RE problem, considering the peculiarities of each language. In this context, the present work proposes to assess the difficulties of classification based on features in open relation extraction in Portuguese, aiming to base new solutions that can reduce language dependence in this task. The results indicate that many representative features in English can not be mapped directly to the Portuguese language with satisfactory merits of classification. Among the classification algorithms evaluated, J48 showed the best results with a F-measure value of 84.1%, followed by SVM (83.9%), Perceptron (82.0%) and Naive Bayes (79,9%).

Keywords

Open Relation Extraction, Feature Selection

1 Introdução

Embora a quantidade de documentos não estruturados publicados na Web cresça a cada ano, a velocidade com que o ser humano consegue interpretar informações permanece constante. Por conta disso, técnicas de Extração da Informação (EI) vêm sendo desenvolvidas com o intuito de identificar conteúdo relevante em grandes quantidades de documentos (Brin, 1998; Feldman e Sanger, 2007; Lutz e Heuser, 2013). Métodos de reconhecimento de conceitos e seus relacionamentos são considerados cruciais em diversas aplicações de processamento linguístico, tais como na construção automática de ontologias e léxicos computacionais (Chaves, 2008),

*Agradecimentos à FAPESB pelo apoio parcial neste projeto.

em sistemas de respostas a perguntas (Hirschman e Gaizauskas, 2001) e na computação forense (Anyanwu, Maduko e Sheth, 2005). Porém, as principais soluções para extração de relações entre conceitos são limitadas por um conjunto predefinido de relações possíveis, o que reduz a aplicabilidade dos métodos a domínios e idiomas específicos.

Um exemplo de aplicação de EI onde a limitação de domínio e idioma constitui um fator proibitivo é no Reconhecimento de Entidades Mencionadas (REM) aplicado à computação forense. Autores em (Dalben e Claro, 2011) afirmam que a identificação de nomes de pessoas e organizações em mídias apreendidas pode reduzir em mais de 90% a quantidade de arquivos analisados manualmente por peritos criminais. Em aplicações deste tipo, é comum que a coleção de documentos contenha vocábulos de domínios e idiomas distintos, pois uma mesma investigação pode envolver organizações com atuações diferentes (como uma clínica médica e um órgão público) em mais de um país. Pelo mesmo motivo, o requisito de independência do domínio se mantém na tarefa de Extração de Relações (ER) entre as entidades identificadas nesses documentos.

Estudos recentes têm sido desenvolvidos com o intuito de contornar as limitações dos métodos tradicionais de ER (Souza e Claro, 2014). Nesse contexto, a Extração de Relações Abertas, derivada da *Open Information Extraction (Open IE)* (Banko e Etzioni, 2008), consiste na tarefa de extrair relações semânticas com vocabulário não-limitado a partir de *corpora* em larga escala. Entretanto, a ambiguidade inerente à linguagem natural tem ocasionado grande proporção de relações inválidas, exemplificadas nas sentenças da Tabela 1.

Uma relação é dita inválida quando é incoerente e/ou incompleta. Intuitivamente, uma extração incoerente ocorre quando a semântica do relacionamento entre as entidades, mesmo sendo completa, não condiz com a interpretação correta da sentença. A primeira linha da Tabela 1 mostra um exemplo de extração incoerente, já que a entidade *Defesa do Criciúma* rebate um objeto que está oculto na frase (a bola), e não a entidade *Maurinho*. Já na segunda linha, *vai emoldurar com* não denota uma relação com sentido completo entre as entidades *PT* e *Luiz Inácio Lula da Silva*.

Sentença	**Extração Inválida**
"Depois de a defesa do Criciúma rebater, Maurinho chutou e marcou."	*(Defesa do Criciúma, rebater, Maurinho)*
"A estrela símbolo do PT vai emoldurar com destaque o cenário dos programas do candidato Luiz Inácio Lula da Silva."	*(PT, vai emoldurar com, Luiz Inácio Lula da Silva)*

Tabela 1: Exemplos de extrações inválidas.

A distinção automática entre relações válidas e inválidas pode ser modelada como um problema de classificação. Trabalhos em (Banko e Etzioni, 2008) e (Fader, Soderland e Etzion, 2011) aplicam algoritmos de aprendizado de máquina sobre *features* extraídas das sentenças para elevar a precisão de classificação das relações. A principal desvantagem dessas abordagens é a dificuldade na seleção de *features* adequadas à tarefa. Além disso, o aprendizado baseado em *features* necessita de bases de treinamento relativamente grandes para gerar resultados satisfatórios. Recursos deste tipo são escassos ou inexistentes na maioria dos idiomas.

Neste trabalho é realizada uma análise do esforço necessário à identificação das *features* mais representativas para a classificação de relações semânticas abertas em textos redigidos em Português, já que a capacidade preditiva de um atributo pode sofrer grande variação em função da mudança de idioma, dificuldade que deve ser considerada em novas soluções capazes de reduzir tal dependência nesta tarefa. O presente artigo está organizado como segue. Na Seção 2 é descrita uma classificação dos métodos, além dos principais conceitos referentes à Extração de Relações (ER). A Seção 3 apresenta as características das abordagens mais recentes de ER Abertas. Na Seção 4 são descritos os experimentos realizados e analisados os resultados obtidos. A Seção 5 conclui este artigo e apresenta alguns trabalhos futuros.

2 Relações Semânticas

A Extração de Relações (ER) consiste na tarefa de descobrir relacionamentos semânticos entre conceitos em documentos não estruturados (Feldman e Sanger, 2007). Embora não exista uma categorização clara dos métodos de ER, é possível agrupá-los a partir dos principais trabalhos apresentados na literatura. Nesta seção são descritos dois tipos de classificação na tarefa de ER: i) Por técnica aplicada; ii) Por tipo de relação extraída.

2.1 Classificação por técnica aplicada

A classificação mais genérica dos métodos de ER distingue as abordagens baseadas em padrões textuais das que utilizam aprendizado de máquina (Taba e Caseli, 2012). A seguir é feita uma breve descrição das principais características de cada tipo de método.

Os métodos de padrões textuais extraem relações utilizando regras formadas por expressões regulares. Um exemplo deste tipo de regra, que pode ser encontrado em (Hearst, 1992), é dado por:

$$NP_1\{,\}\ especially\ \{NP_2, NP_3...\}\ \{or|and\}\ NP_n \quad (1)$$

Com este padrão é possível identificar relações de hiponímia do tipo *is-a* entre as frases nominais NP_i e NP_1, com $i \in \{2, 3, ..., n\}$. Tomando como exemplo a frase "*most countries, especially France, England and Spain*" ("a maioria dos países, especialmente Franca, Inglaterra e Espanha"), a aplicação da regra permite extrair as seguintes relações: *is-a(France, country)*, *is-a(England, country)* e *is-a(Spain, country)*.

É possível elencar uma série de deficiências e limitações nos métodos baseados em padrões textuais. Primeiro, a especificidade das regras resulta em alta precisão, mas baixa cobertura (Freitas e Quental, 2007; Snow, Jurafsky e Ng, 2005). Segundo, devido à grande diversidade das variações linguísticas, certos padrões podem ser associados a diversos tipos de relações, tornando inviável o mapeamento de todas as possibilidades (Girju et al., 2010). Por exemplo, o padrão "tais como" é comumente reduzido à palavra denotativa "como" em textos escritos em Português, que pode pertencer às seguintes classes morfológicas: conjunção, pronome relativo, substantivo, advérbio interrogativo, advérbio de modo, interjeição e preposição. Entretanto, o único sentido da palavra "como" que deve ser reconhecido pelo referido padrão é o equivalente a "por exemplo" (pronome relativo). Por conta disso, a criação de uma base de regras minimamente representativa para esse tipo de método consiste em uma tarefa altamente dispendiosa. Trabalhos recentes vêm apresentando resultados mais efetivos em termos de precisão e cobertura, através de técnicas de aprendizado de máquina.

As abordagens baseadas em aprendizado de máquina selecionam atributos (*features*[1]) a partir de um conjunto de treinamento, a fim de determinar se existe uma relação entre as entidades de uma nova instância (Kambhatla, 2004). Mais precisamente, dada uma sentença $S = w_1, w_2, ..., e_1, ..., w_j, ..., e_2, ...w_n$, onde e_1 e e_2 são entidades existentes entre as palavras $w_1, w_2, ..., w_n$, uma função de mapeamento f é definida por:

$$f_R(\Theta(S)) = \begin{cases} +1, & \text{se existe R entre } e_1 \text{ e } e_2, \\ -1, & \text{caso contrário} \end{cases} \quad (2)$$

Onde $\Theta(S)$ constitui o conjunto de *features* extraídas de S e R representa a relação semântica. Assim, a Equação 2 decide se existe uma relação semântica R entre as entidades e_1 e e_2.

Além das soluções baseadas em *features*, existem trabalhos que utilizam uma generalização da similaridade de subsequências de strings (*string-kernels* (Zelenko, Aone e Richardella, 2003)) para a realização de treinamentos. Considerando duas strings x e y, a similaridade $K(x, y)$ em *string-kernels* é calculada em função do número de subsequências que são comuns a ambas. Ou seja, quanto maior a quantidade de subsequências comuns entre x e y, maior a similaridade entre elas.

Partindo deste princípio, sendo A e B exemplos de sentenças com relação positiva e negativa entre duas entidades, respectivamente, no conjunto de treinamento, a função de similaridade que indica a classe de uma instância de teste T é calculada com base na seguinte equação:

$$f_R(K) = \begin{cases} +1, & \text{se } K({S_A}^+, S_T) > K({S_B}^-, S_T), \\ -1, & \text{caso contrário} \end{cases} \quad (3)$$

Onde ${S_A}^+$, ${S_B}^-$ e S_T representam os respectivos conjuntos constituídos pelos termos que cercam as entidades nas sentenças A, B e T. Como exemplo, considerando a sentença *"O campus da UFBA está situado em Ondina"*, as palavras *campus* e *situado* indicam uma relação do tipo *localidade* entre as entidades *UFBA* e *Ondina*, cujas similaridades com os termos que cercam entidades em outras sentenças podem ser utilizadas para extrair delas o mesmo tipo de relação.

A Figura 1(a) mostra a classificação dos métodos de ER considerando o tipo de método.

2.2 Classificação por tipo de relação extraída

A semântica das relações extraídas varia bastante nos trabalhos de ER. Entretanto, é possível identificar dois tipos de métodos: os que extraem

[1] As features representam propriedades léxicas, sintáticas ou semânticas dos termos de uma sentença.

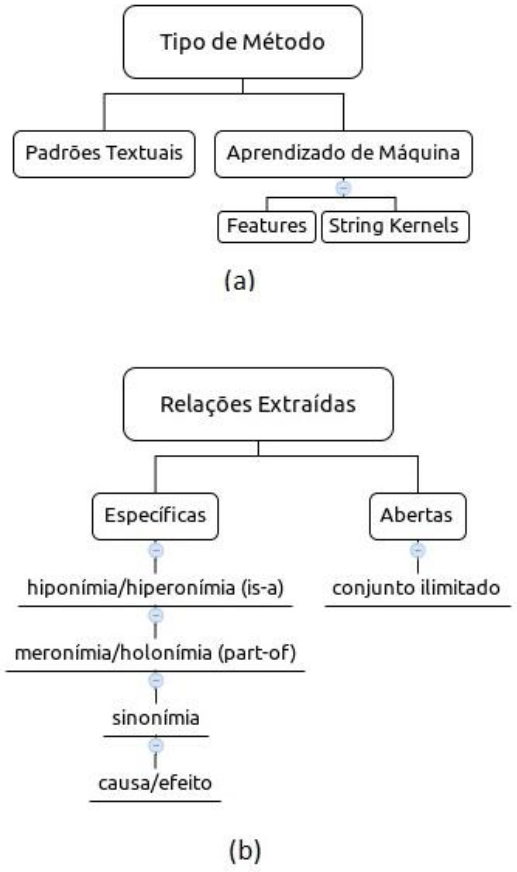

Figura 1: Classificação dos métodos de ER: (a) por tipo de método e (b) por tipo de relação.

relações específicas e os que extraem relações abertas. Um esquema que ilustra esta classificação é mostrado na Figura 1(b).

Na extração de relações específicas, um domínio finito de relações semânticas é definido para a tarefa de ER, conforme exemplos da Tabela 2.

A principal desvantagem dessa abordagem é a limitação da extração a um universo específico. Dessa forma, muitas relações semânticas importantes presentes no texto não são identificadas por não pertencerem ao domínio definido e nem ao conjunto predeterminado de relações.

Relação	location-of(algo/alguém, local)
Exemplo	Um aluno pode ser encontrado na escola
Extração	location-of(aluno, escola)
Relação	isa(subclasse, superclasse)
Exemplo	Salvador é uma cidade
Extração	is-a(Salvador, cidade)
Relação	part-of(todo,parte)
Exemplo	Roda é parte de um carro
Extração	part-of(roda, carro)

Tabela 2: Exemplos de relações específicas.

A descoberta de relações sem restrição de domínio representa um requisito essencial em diversas aplicações de EI. Por conta disso, estudos têm sido conduzidos no sentido de identificar relações de vocabulário não-limitado, caracterizando a Extração de Relações Abertas (do inglês, *Open Relation Extraction*) (Banko e Etzioni, 2008; Nakashole e Mitchell, 2014), tarefa abordada neste trabalho.

Como a categorização apresentada não é mutualmente exclusiva, os métodos de ER se enquadram em ambos os tipos de classificação, sendo possível identificar certas associações entre eles. Por exemplo, todas as abordagens de padrões textuais necessariamente extraem relações específicas (Hearst, 1992; Freitas e Quental, 2007; Girju et al., 2010). Por outro lado, existem abordagens de aprendizado de máquina utilizadas tanto na extração de relações específicas (Kambhatla, 2004; Zelenko, Aone e Richardella, 2003), quanto na extração de relações abertas (Banko e Etzioni, 2008; Fader, Soderland e Etzion, 2011). Nos métodos de extração de relações abertas investigados, as extrações são identificadas através de padrões morfológicos e classificadas utilizando aprendizado supervisionado.

Em relação à ER na Língua Portuguesa, percebe-se que a maioria dos trabalhos utiliza técnicas rudimentares baseadas em padrões textuais, sendo que as abordagem de aprendizado de máquina ainda são pouco exploradas. Isto se deve, possivelmente, à falta de recursos linguísticos em Português, dificultando a construção de bases de treinamento de forma automática ou semi-automática para a tarefa, que necessita de *features* representativas obtidas a partir de conhecimento especializado na língua. Dentre o universo de relações específicas extraídas em Português, as mais frequentes são as relações de hiponímia, meronímia e localidade (Oliveira, Santos e Gomes, 2010; Cardoso, 2008; Chaves, 2008; Bruckschen et al., 2008). Por outro lado, não foi identificada nenhuma pesquisa voltada para a Extração de Relações Abertas neste idioma.

Na próxima seção são descritas as principais características da ER Abertas.

3 Extração de Relações Abertas

Os métodos precursores de ER Abertas obtém extratos na forma *(e_1, frase relacional, e_2)* em três etapas (Fader, Soderland e Etzion, 2011):

1. **Etiquetação:** As sentenças são etiquetadas automaticamente através de heurísticas ou a partir de supervisão distante (treinamento semi-supervisionado);
2. **Aprendizado:** Um extrator de frases relacionais é treinado utilizando um modelo de etiquetação sequencial (e.g. CRF);
3. **Extração:** Um conjunto de argumentos (e_1, e_2) é identificado na sentença de teste. Em seguida, o extrator treinado na etapa 2 é utilizado para etiquetar as palavras contidas entre os argumentos e compor a frase relacional (caso ela exista), extraindo a relação no formato *(e_1, frase relacional, e_2)*.

Uma das desvantagens dessas abordagens reside no fato de que a etiquetação precisa ser realizada em uma quantidade muito grande de sentenças (na ordem de centenas de milhares) para que a etapa de aprendizado seja efetiva. Isto implica em alto custo de construção dos conjuntos de treinamento, além da demanda de recursos linguísticos sofisticados para viabilizar a etiquetação automática, dificilmente encontrados na maioria dos idiomas. Além disso, o método de extração por etiquetação sequencial é pouco eficaz em sentenças maiores, pois há um aumento da incerteza na associação de cada etiqueta a uma palavra à medida que a sequência cresce.

Abordagens mais recentes têm sido desenvolvidas para contornar algumas dessas limitações, por meio de modificações na metodologia e, consequentemente, nas estratégias adotadas nas etapas de extração (Fader, Soderland e Etzion, 2011; Banko et al., 2007; Banko e Etzioni, 2008). Assim, é realizada primeiramente a etapa de extração, seguida pelo aprendizado necessário à posterior classificação das relações, conforme descrito abaixo:

1. **Extração:** Inicialmente, um extrator baseado em padrões linguísticos (e.g. padrões verbais) seleciona uma sequência de palavras que representa a relação semântica entre e_1 e e_2, identificando frases relacionais que casam com esses padrões. Em seguida, se um conjunto de argumentos (e_1, e_2) for identificado na sentença de teste, então é gerada a relação na forma *(e_1, frase relacional, e_2)*;
2. **Aprendizado:** Um classificador de extrações é treinado por meio de um conjunto de *features* linguísticas;
3. **Classificação:** O classificador treinado na etapa 2 é utilizado para distinguir as relações válidas das inválidas geradas na etapa 1.

Essa nova abordagem substitui o aprendizado na etapa de extração pelo processamento de regras baseadas em padrões morfológicos. Em seguida, um classificador é utilizado na remoção das relações inválidas do conjunto que contém todas as relações extraídas. Esta metodologia permite uma redução significativa na cardinalidade do conjunto de treinamento, já que a complexidade do aprendizado para classificação das relações é inferior à do aprendizado para a identificação das relações. Por outro lado, a construção de conjuntos de treinamento a partir de *features* linguísticas eleva o custo de classificação, pois a identificação de *features* representativas requer uma análise mais aprofundada das características da língua no contexto do problema. Neste trabalho, é realizada uma análise do esforço necessário à identificação das *features* mais representativas para a classificação de relações semânticas abertas em textos redigidos em Português, a partir dos experimentos descritos na próxima seção.

4 Experimentos e Resultados

Os experimentos foram realizados utilizando o corpus CETENFolha[2] (Corpus de Extratos de Textos Eletrônicos NILC/Folha de S. Paulo), que contém cerca de 24 milhões de palavras em Português, extraídas de textos do jornal Folha de São Paulo. Foram selecionadas aleatoriamente 500 sentenças do corpus envolvendo diferentes temas, tais como política, economia, esportes e ciência. As classes morfológicas das palavras contidas nas sentenças selecionadas foram obtidas automaticamente pelo etiquetador morfossintático do Cogroo[3], um corretor gramatical acoplável a um editor de texto de código aberto.

Após a etiquetação morfológica, foram extraídas 582 relações do tipo (fn_1, rel, fn_2), onde fn_1 e fn_2 representam as frases nominais contendo entidades mencionadas, encontradas antes e depois da relação, e rel denota a frase relacional da extração. As frases relacionais foram obtidas a partir do procedimento para identificação de padrões morfológicos na etapa de extração descrita na Seção 3 e adaptado para a Língua Portuguesa, sendo as entidades inicialmente identificadas aquelas classificadas como nome próprio pelo Cogroo. Por fim, cada extração foi manualmente classificada como válida ou inválida para compor o conjunto de treinamento. Exemplos de relações válidas e inválidas são mostradas na Tabela 3.

[2] http://www.linguateca.pt/cetenfolha/
[3] http://cogroo.sourceforge.net/

Tipo	Exemplo
Válida	X *matou* Y.
Inválida	X o *matou* enquanto Y assistia.
Válida	X, após o trabalho, *veio buscar* Y.
Inválida	X *veio buscar* os documentos antes de Y retornar.
Válida	X correu, mas *negou ter roubado* Y.
Inválida	X *negou ter roubado*, mas Y confessou o crime.
Válida	No dia seguinte, X *negociou com* Y sobre a venda da empresa.
Inválida	X apresentou o seu sócio, que *negociou com* Y.
Válida	X ainda *deve contar com* Y.
Inválida	X *deve contar com* um novo jogador na partida contra Y.

Tabela 3: Exemplos de relações válidas e inválidas.

Para viabilizar a avaliação de classificação, foram selecionadas 12 *features* de treinamento, definidas originalmente em Inglês por (Fader, Soderland e Etzion, 2011) e adaptadas para a Língua Portuguesa neste trabalho (Tabela 4). Os valores das *features* foram extraídos automaticamente de todas as sentenças selecionadas do corpus e aplicados a quatro classificadores utilizando a ferramenta de mineração de dados WEKA[4].

A efetividade ou mérito das *features* é estimada pelo algoritmo *Correlation-based Feature Selection (CFS)* (Hall, 1999), que utiliza uma heurística baseada em correlação para avaliar a capacidade de cada atributo em predizer a classe de uma instância de teste, dado um conjunto de treinamento. A hipótese que embasa este algoritmo afirma que bons subconjuntos de atributos devem possuir alta correlação com a classe de predição e baixa correlação entre si, já que atributos que possuem alta correlação entre si são considerados redundantes e não contribuem para elevar a capacidade preditiva do subconjunto.

Formalmente, seja S um subconjunto contendo k atributos, o mérito de S é calculado pela Equação 4:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (4)$$

Onde $\overline{r_{cf}}$ representa a correlação média entre cada atributo de S e o atributo de classe, e $\overline{r_{ff}}$ denota a correlação média entre todas as combinações de atributos em S. A correlação entre os atributos pode ser estimada por diversas heurísticas, como o coeficiente de incerteza simétrica (baseado nos conceitos de entropia e ganho de informação) (Kononenko e Bratko, 1991) e o algoritmo *Relief* (Kononenko, 1994) (que utiliza uma abordagem baseada em instâncias para associar pesos às iterações entre os atributos).

[4]http://www.cd.waikato.ac.nz/ml/weka

F_1	tamanho(sentença) - tamanho($fn_1 + rel + fn_2$) < 30 caracteres?
F_2	A última preposição em *rel* é "de"?
F_3	A última preposição em *rel* é "com"?
F_4	A última preposição em *rel* é "por"?
F_5	A última preposição em *rel* é "pela"?
F_6	A última preposição em *rel* é "pelo"?
F_7	A última preposição em *rel* é "para"?
F_8	A última preposição em *rel* é "em"?
F_9	A string $fn_1 + rel$ está contida na sentença?
F_{10}	A string $rel + fn_2$ está contida na sentença?
F_{11}	A string $fn_1 + rel + fn_2$ está contida na sentença?
F_{12}	Há menos de 30 palavras na sentença?

Tabela 4: Features utilizadas para a base de treinamento em Língua Portuguesa.

4.1 Resultados

A Figura 2 mostra o mérito das *features* descritas na Tabela 4, considerando todo o conjunto de dados (582 extrações obtidas de 500 sentenças). É possível notar que as *features* F_9, F_{10} e F_{11} são as que possuem as maiores capacidades de predição. Por outro lado, a *feature* F_1 pode ser eliminada do conjunto de atributos sem prejuízo à qualidade de classificação, já que possui mérito nulo.

Os resultados mostrados na Figura 2 foram obtidos a partir da execução do algoritmo CFS implementado no Weka, usando a estratégia de busca *BestFirst* com parâmetros $D = 1$ (*forward*

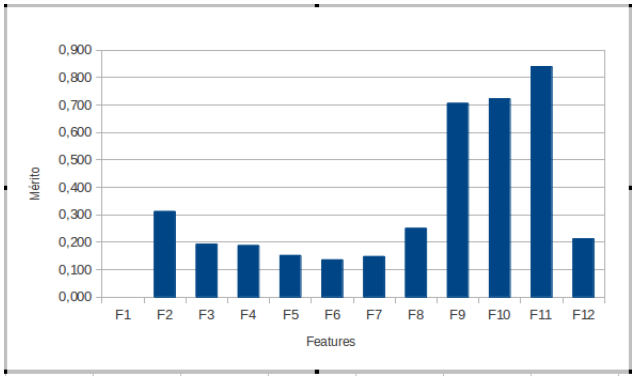

Figura 2: Representatividade das *features* no conjunto de dados.

Sub-conjunto	Feature avaliada	Elementos do melhor subconjunto
CF_1	-	$F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}$
CF_2	F_{11}	$F_9, F_{10}, F_{11}, F_{12}$
CF_3	F_{10}	$F_2, F_6, F_8, F_{10}, F_{11}$
CF_4	F_9	F_4, F_9, F_{11}

Tabela 5: Conjuntos de *features*.

search) e $N = 5$ (número de nós do critério de parada), com seleção de atributos usando todo o conjunto de treinamento. Diante desses resultados, foram selecionados quatro subconjuntos de *features* (Tabela 5) para avaliação. O grupo CF_1 é composto por todas as *features* que possuem mérito não nulo e os grupos CF_2, CF_3 e CF_4 correspondem aos subconjuntos obtidos a partir das melhores *features* avaliadas pelo algoritmo CFS.

É possível notar que nem sempre as *features* que possuem os maiores méritos formam o melhor subconjunto, já que pode haver alta correlação entre elas, redundância que não contribui para elevar a capacidade preditiva do subconjunto como um todo. Dessa maneira, as *features* F_2, F_9, F_{10} e F_{11} não formam um subconjunto com alta capacidade preditiva, devido à alta correlação entre F_2 e F_9.

Na Figura 3 são mostrados os valores médios da medida-F e da área sob a curva ROC (AUC) de quatro algoritmos de classificação avaliados (J48, SVM, Perceptron e Naive Bayes) em cada conjunto de *features*, utilizando o método de validação cruzada com 10 *folds*. Os resultados mostram valores aproximadamente iguais para os três grupos CF_1, CF_2 e CF_3, sendo a maior diferença equivalente a 0,7% para a medida-F e 1,4% para a AUC entre os grupos CF_1 e CF_3, indicando que a dimensionalidade dos atributos pode ser reduzida de 11 para 4 (CF_2) ou 5 (CF_3) *features*, com perdas mínimas na qualidade de classificação. Por outro lado, o grupo CF_4 apresentou

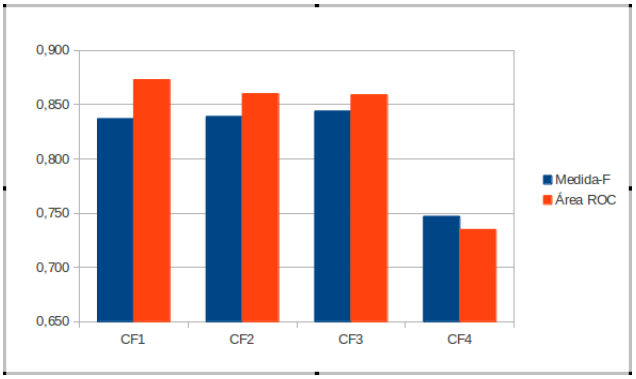

Figura 3: Avaliação dos conjuntos de *features*.

Método	Precisão	Cobertura	Medida-F
J48	0,848 ± 0,014	0,841 ± 0,018	0,841 ± 0,018
Lib SVM	0,848 ± 0,019	0,840 ± 0,018	0,839 ± 0,018
Perceptron	0,823 ± 0,038	0,820 ± 0,041	0,820 ± 0,040
Naive Bayes	0,800 ± 0,037	0,799 ± 0,039	0,799 ± 0,039

Tabela 6: Resultados médios obtidos por validação cruzada com 10 *folds*.

valores médios 9,5% inferiores para a medida-F e 13,8% para a AUC, sendo portanto o menos representativo dentre os conjuntos avaliados.

A Tabela 6 mostra os valores detalhados de precisão, cobertura e medida-F nos métodos de classificação testados em ordem decrescente de desempenho, a partir do conjunto de *features* CF_1. Os valores médios e desvios padrões correspondentes são obtidos pelo processamento de 10 conjuntos de sentenças com tamanhos distintos, que variam de 57 a 582 extrações.

Adicionalmente, as curvas no gráfico da Figura 4 ilustram as variações de precisão, cobertura e medida-F com o aumento do conjunto de treinamento e teste em cada algoritmo. É possível perceber que o algoritmo J48 obteve os melhores resultados na classificação de relações abertas em Português, tendo uma medida-F média 4,2% superior ao classificador bayesiano, que apresentou os piores resultados dentre os métodos testados. Além disso, nota-se um crescimento na medida-F dos algoritmos em função da cardinalidade do conjunto de treinamento.

Os resultados mostram as dificuldades encontradas na identificação de *features* linguísticas representativas para a tarefa de extração de relações abertas, já que o mérito de um atributo pode sofrer grande variação em função da mudança de idioma. Por exemplo, a *feature* F_1 apresentou mérito nulo para a Língua Portuguesa,

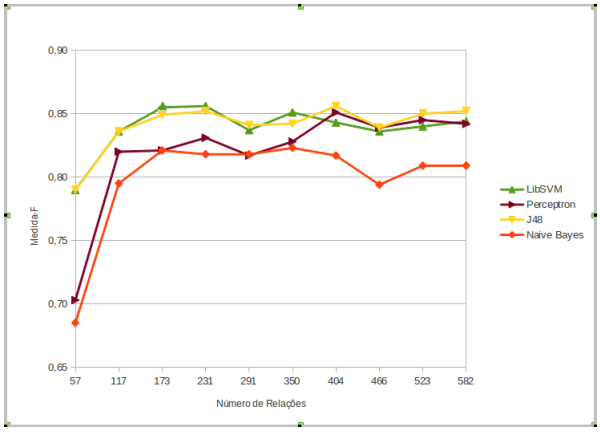

Figura 4: Avaliação da medida-F em função da quantidade de relações classificadas.

mas representa um dos atributos mais significativos para o mesmo problema na Língua Inglesa, como pode ser observado nos resultados obtidos em (Fader, Soderland e Etzion, 2011).

Como a análise do mérito das *features* não foi apresentada nos trabalhos voltados para a Língua Inglesa pesquisados, não foi possível realizar um estudo comparativo direto com os resultados obtidos no presente trabalho. Entretando, é possível comparar indiretamente o desempenho das *features* em Inglês e Português no problema de classificação de relações abertas tratado. Em (Fader, Soderland e Etzion, 2011), a utilização de um classificador de regressão logística treinado com um conjunto de *features* em Língua Inglesa apresentou uma medida-F cerca de 8% superior ao algoritmo de classificação com o melhor desempenho avaliado em Língua Portuguesa neste trabalho.

Essas observações permitem afirmar que o mapeamento direto de um conjunto de *features* de um idioma para outro não implica na seleção dos melhores atributos na classificação de relações abertas. Consequentemente, é necessária uma análise mais profunda das peculiaridades de cada idioma para a escolha de um conjunto representativo de *features*.

5 Conclusões e Trabalhos Futuros

A distinção automática entre relações válidas e inválidas representa um problema recorrente em sistemas de extração de relações em texto não estruturado. Quando as frases relacionais identificadas possuem vocabulário não limitado, a importância da tarefa de classificação na qualidade das extrações se torna mais evidente, já que a ambiguidade inerente à linguagem natural tem ocasionado grande proporção de relações inválidas nos métodos mais recentes que tratam desta tarefa.

Grande parte das soluções atuais extraem relações abertas exclusivamente a partir de textos redigidos em Inglês, idioma que possui os recursos linguísticos mais sofisticados, como etiquetadores morfossintáticos, extratores de entidades mencionadas, frases nominais e correferências, além de léxicos computacionais de alta granularidade e grandes bases de treinamento. Os principais trabalhos do estado da arte eliminam as relações inválidas por meio de classificadores treinados a partir de *features* linguísticas, altamente dependentes do idioma-alvo. Tal dependência decorre da dificuldade inerente à determinação do conjunto de *features* mais representativo para o problema, considerando as peculiaridades de cada língua. Em particular, o presente trabalho avalia esta dificuldade em textos redigidos em Português, que permite identificar duas limitações principais: 1) O mapeamento de um conjunto de *features* de um idioma para outro não implica na seleção dos melhores atributos de treinamento, dadas as especificidades de cada idioma, o que implica em novas análises para cada língua; 2) A qualidade de classificação das relações abertas baseada em *features* depende de conjuntos de treinamento extensos, que possuem alto custo de construção, conforme resultados descritos em trabalhos predecessores voltados para a Língua Inglesa (Wu e Weld, 2010; Banko e Etzioni, 2008). Neste trabalho, o crescimento da medida-F na classificação de relações abertas com o aumento da cardinalidade do conjunto de treinamento indica que esta característica também é válida para corpora redigidos em Língua Portuguesa.

Como trabalhos futuros, pretende-se investigar abordagens capazes de reduzir a dependência do idioma na tarefa de extração de relações abertas, por meio da eliminação da necessidade de construção de conjuntos extensos de treinamento baseados em *features* linguísticas específicas na etapa de classificação das relações.

Referências

Anyanwu, K., A. Maduko, e A. Sheth. 2005. Semrank: Ranking complex relationship search results on the semantic web. *Proc. of the 14th International World Wide Web Conference, ACM Press, 117-127.*

Banko, M. e O Etzioni. 2008. The tradeoffs between open and traditional relation extraction. *In Proceedings of ACL-08: HLT, pages 28-36, Columbus, Ohio, June. Association for Computational Linguistics.*

Banko, M., M. J. J. Cafarella, S. Soderland, M. Broadhead, e O. Etzioni. 2007. Open information extraction from the web. *In the Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 2670-2676, January.*

Brin, S. 1998. Extracting patterns and relations from the world wide web. *In International Workshop on The World Wide Web and Databases, p.172-183, March 27-28, 1998.*

Bruckschen, M., J. Souza, R. Vieira, e S. Rigo. 2008. Sistema serelep para o reconhecimento de relações entre entidades mencionadas. *In Cristina Mota; Diana Santos (ed.), Desafios na avaliação conjunta do reconhecimento de*

entidades mencionadas: O Segundo HAREM. Linguateca, cap. 14, p. 247-260.

Cardoso, N. 2008. Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto, 2008. *In:: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. s.l.:Linguateca, pp. 195-211.*

Chaves, S. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. *In Cristina Mota; Diana Santos (ed.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, cap. 13, p. 231-245.*

Dalben, O. J. e D. B. Claro. 2011. Uma análise do reconhecimento textual de nomes de pessoas e organizações na computação forense. *Proceeding of the Sixth International Conference on Forensic Computer Science - ICoFCS 2011, pp. 7-15.*

Fader, A., S. Soderland, e O. Etzion. 2011. Identifying relations for open information extraction. *In Proceedings of Conference on Empirical Methods in Natural Language Processing.*

Feldman, R. e J. Sanger. 2007. *The text mining handbook: advanced approaches analyzing advanced unstructured data.* New York: Cambridge University Press.

Freitas, C. e V. Quental. 2007. Subsídios para a elaboração automática de taxonomias. *Anais do XXVII Congresso da SBC. Rio de Janeiro, Rio de Janeiro: [s.n.], 2007. (V Workshop em Tecnologia da Informacao e da Linguagem Humana TIL), p. 1585-1594.*

Girju, R., B. Beamer, A. Rozovskaya, A. Fister, e S. Bhat. 2010. A knowledge-rich approach to identifying semantic relations between nominals. *Information Processing and Management,v. 46, n. 5, p. 589-610.*

Hall, M. 1999. *Correlation-based Feature Selection for Machine Learning.* Tese de doutoramento, University of Waikato, Hamilton, NewZealand.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational linguistics - Volume 2. Nantes, France, p. 539-545.*

Hirschman, L. e R. Gaizauskas. 2001. Natural language question answering: the view from here. *Natural Language Engineering 7 (4): 275-300.*

Kambhatla, N. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL.*

Kononenko, I. 1994. Estimating attributes: Analysis and extensions of relief. *In Proceedings of the European Conference on Machine Learning.*

Kononenko, I. e I. Bratko. 1991. Information-based evaluation criterion for classifiers performance. *Machine Learning, 6:67-80.*

Lutz, J. e C. Heuser. 2013. Descoberta de ruído em páginas da web oculta através de uma abordagem de aprendizagem supervisionada. *Simpósio Brasileiro de Banco de Dados (SBBD'13), Recife, PE, Brazil.*

Nakashole, N. e T. Mitchell. 2014. Language-aware truth assessment of fact candidates. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014.*

Oliveira, H., D. Santos, e P. Gomes. 2010. Extração de relações semânticas entre palavras a partir de um dicionário: o papel e sua avaliação. *Linguamática, v. 2, n. 1, p. 77-94.*

Snow, R., D. Jurafsky, e A. Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. *In Advances in Neural Information Processing Systems 17, pages 1297-1304. MIT Press.*

Souza, E. e D. Claro. 2014. Detecção multilíngue de serviços web duplicados baseada na similaridade textual. *Simpósio Brasileiro de Sistemas de Informação (SBSI'14), Maio 27-30, Londrina/PR, Brazil.*

Taba, L. S. e H. Caseli. 2012. Automatic hyponymy identification from brazilian portuguese texts. *In Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR).*

Wu, F. e D. S. Weld. 2010. Open information extraction using wikipedia. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 118-127, Morristown.*

Zelenko, D., C. Aone, e A. Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research 3 1083-1106.*