Engenharia de *features* linguísticas para classificação de triplas relacionais

Elian Conceição Luz[1,2], **Camilla Rastely da Silva**[2], **Daniela Barreiro Claro**[1]

[1] FORMAS Research Group, Departamento de Ciência da Computação
Instituto de Computação – Universidade Federal da Bahia,
40170-110 – Salvador, BA, Brazil

[2]Instituto de Letras - Universidade Federal da Bahia
40170-110 – Salvador, BA, Brazil

{elianc,camillars,dclaro}@ufba.br

Abstract. *In this study, linguistic features were listed for classification of triples, based on a parallel corpus in Galician, Brazilian Portuguese (PT-BR) and European Spanish (EE). In the experiments, at the syntactic level, a relevant performance of features that offer greater difficulty to extract valid triples was observed, such as the co-related null object/subject and the verb-subject inversion, as well as relational triples that form ungrammatical sub-sentences. At the morphological level, it was observed that the grammatical class of the initial word of each sub-sentence, especially when they are prepositions, were relevant for the classification of the triples.*

Resumo. *Neste estudo, elencaram-se features morfossintáticas para classificação de triplas, com base em um corpus paralelo em Galego, Português do Brasil e Espanhol Europeu. Nos experimentos, a nível sintático, observou-se desempenho relevante de features que oferecem maior dificuldade para extrair triplas válidas, como as co-relacionadas a objeto/sujeito nulo e a inversão verbo-sujeito, bem como triplas relacionais que formam sub-sentenças agramaticais. A nível morfológico, observou-se que a classe gramatical do vocábulo inicial de cada sub-sentença, sobretudo quando são preposições, foram relevantes para a classificação das triplas.*

1. Introdução

O exponencial crescimento do volume de documentos digitais escritos em linguagem natural impulsiona a demanda por modelos automáticos capazes de extrair informação de dados não-estruturados. Nessa perspectiva, a Extração de Informação Aberta (EIA) possibilita a estruturação da linguagem natural em triplas relacionais ao processar as sentenças de um texto, obtendo uma estrutura composta por três elementos (arg1, relação, arg2) [Barbosa 2018]. Por exemplo, a partir da sentença 'João ama Maria', pode-se extrair a tripla válida ('João', 'ama', 'Maria').

A motivação de compor tarefas de PLN em perspectiva multilíngua é ampliar o alcance dos métodos de EIA para fora do alcance da língua inglesa. Em relação ao Português (PT-BR) e ao Espanhol Europeu (EE), por exemplo, mesmo sendo línguas de

ampla circulação e com números expressivos de falantes, ambas apresentam recursos escassos para automação de tarefas do PLN, o que é ainda mais evidente em relação à língua galega. Nessa perspetiva, este estudo favorece o desenvolvimento de métodos capazes de extrair informação de textos redigidos nessas três línguas. Outrossim, a descrição de características de língua propicia maior independência de domínio, contribuindo para a melhoria da Extração de Informação Aberta (EIA) e para a capacitação linguística dessas línguas [Calvet 2005][Gauger 1989].

Tradicionalmente, a EIA inclui métodos de extração de triplas que, em sua maioria, foram desenvolvidos com base em corpus de língua inglesa, considerando suas especificidades linguísticas, como a maior predominância do preenchimento do sujeito e a ordem sujeito-verbo, características que não são predominantes em línguas românicas[Barbosa 2018]. Muitos estudos propuseram a tradução das *features* para a língua-alvo, o que não permitiu obter resultados satisfatórios. A dificuldade aumenta quando se trata de abordagens multilínguas. Assim, o principal propósito deste trabalho foi determinar as *features* linguísticas por meio das triplas extraídas de corpora em línguas românicas: Galego, Português (PT-BR) e Espanhol Europeu (EE). A metodologia adotada analisou as características linguísticas genéricas por meio de uma revisão de estudos da Linguística Formal, em destaque para os estudos de base gerativista [Chomsky and Lasnik 2008] [Chomsky 2014].

Este artigo está estruturado como segue: a seção 2 apresenta a problemática do estudo e hipóteses para o direcionamento dos experimentos; a seção 3 descreve as *features* elencadas para o desenvolvimento dos experimentos, construídas com base na descrição linguística das línguas-alvo; na seção 4, destacam-se a constituição do corpus, os experimentos e a avaliação das *features*; e , por fim, a seção 5 consolida os principais resultados e finaliza o estudo realizado.

2. *Features* linguísticas nas três línguas românicas

Tradicionalmente, a Extração de Informação em documentos busca explorar padrões que expressam relações predefinidas, como geral-específico, parte-todo, localidade e pertencimento. Em sintonia com o ambiente multifacetado da *Web*, a Extração de Informação Aberta (EIA) não limita o tipo de relação, possibilitando a análise de texto em diferentes domínios. Uma possibilidade é a extração de triplas relacionais por meio de métodos de Processamento de Linguagem Natural (PLN), cuja estrutura composta por três elementos (*arg1*, *relação*, *arg2*) é capaz de representar conhecimento de forma mais flexível, a exemplo de atributos, eventos, fatos e entidades [Barbosa 2018].

No entanto, em sua maioria, os métodos de EIA ainda estão focalizados em domínios linguísticos específicos, enquanto cada vez mais a Web concentra textos de diversas línguas e domínios, evidenciando as especificidades dos métodos que dificilmente são replicáveis a contextos mais amplos. Ao considerar as línguas selecionadas para estudo, observam-se aproximações e divergências com possíveis implicações para extração e validação de triplas relacionais em perspectiva multilíngua. Para tal, faz-se necessária descrição linguística dos domínios de língua para a assertiva execução de tarefas de extração e validação de triplas relacionais. Como um exercício da dificuldade em elencar convergências entre esses domínios, pode-se tomar como exemplo as sentenças e suas respectivas extrações apresentadas logo abaixo:

Galego

Bastas querer atopar os teus funcionarios, que [*sujeito nulo*] aparecen no mesmo instante.
([sujeito nulo], bastas querer atopar, os teus funcionarios)

Português Brasileiro

Basta você querer encontrar seus funcionários, que eles aparecem no mesmo instante.[Fanjul 2014]
(você, basta querer encontrar, seus funcionários)

Espanhol Europeu

Basta que quieras encontrar a tus empleados, que [*sujeito nulo*] aparecen en el mismo instante. [Fanjul 2014]
([sujeito nulo], basta que quieras encontrar a, tus empleados).

Em estudos anteriores, demonstrou-se a importância das preposições para análise da estrutura das sentenças. Observou-se que sentenças maiores apresentam uma maior dificuldade para as tarefas de extração, pois elas, geralmente, formam períodos mais complexos. Pode-se, também, considerar que a presença de determinada classe gramatical é mais frequente em cada elemento da tripla, como são os nomes no argumento 1 e no argumento 2, por outro lado, os verbos flexionados são necessários na relacional entre os argumentos.

Por fim, as classes gramaticais das palavras em posição final e inicial indicam características importantes, como pode ser observado na sétima *feature* do Reverb [Fader et al. 2011]. Por exemplo, o conjunto de *features* proposto no modelo indica a transitividade verbal: se é direta ou indireta. Outrossim, pode ser indicativa de inversão da ordem frasal, bem como deslocamento com formação de tópico, com o deslocamento do complemento adverbial de lugar. Neste caso, é possível identificar o deslocamento pela posição da preposição 'em' no início da sentença.

3. *Engenharia de Features*

A partir da revisão bibliográfica de *features* propostas de trabalhos relacionados, selecionaram-se algumas características possivelmente relevantes para classificação de triplas relacionais no Galego, Português Brasileiro e Espanhol Europeu. Fez-se necessário conhecimento de domínio para análise assertiva das características de língua e correta aplicação das técnicas de análise de dados. O objetivo foi extrair o máximo de informações relevantes para análise, o que foi consolidado nas *features* elencadas na Tabela 1.

Ao observar as *features*, destaca-se, por exemplo, que as preposições ajudam a identificar a função sintática dos itens de uma sentença.

Exemplo 1

(1) *Na segunda, João foi à escola.*

A presença de tópico está relacionada a outros fenômenos sintáticos, como o sujeito nulo [de Castilho et al. 1973] [Chomsky and Lasnik 2008].

Exemplo 2

Table 1. *Features* propostas neste estudo

N	*Feature*
1	O Arg1 não é vazio e está contido na sentença
2	O Arg2 não é vazio e está contido na sentença
3	A Rel está contida na sentença
4	Classe gramatical da palavra no início da sentença
5	Classe gramatical da palavra no início do Arg1
6	Classe gramatical da palavra no início do Arg2
7	Classe gramatical da palavra no início do Rel
8	As preposições em início de sentença
9	As preposições em início de Arg1
10	As preposições em início de Arg2

(2) Onde está João?

(3) No domingo, [*sujeito nulo*] vai para escola.

No corpus estudado, é possível observar esse mesmo fenômeno em uma das sentenças nas três línguas estudadas. A partir dessa amostra, é possível observar um exemplo de como o deslocamento se relaciona com a ocorrência do sujeito nulo nessas línguas.

Excerto do corpus 1

(4) **GL**

Sentença:Na política interna [*sujeito nulo*] mostrouse sempre a favor dun goberno de conciliación nacional e foi contrario á unificación de Moldávia coa Romanía.

Tripla válida: ([*sujeito nulo*], foi contrario á, unificación de Moldávia coa Romanía)

(5) **EE**

Sentença: En política interior [*sujeito nulo*] se mostró siempre a favor de un gobierno de conciliación nacional y fue contrario a la unificación de Moldavia con Romanía.

Tripla válida: ([*sujeito nulo*], fue contrario a, la unificación de Moldavia con Romanía)

(6) **PT-BR**

Sentença: Na politica interna, [*sujeito nulo*] mostrou-se sempre a favor de um governo de conciliação nacional e foi contrário à unificação da Moldávia com a România.

Tripla válida: ([*sujeito nulo*], foi contrário à, unificação da Moldávia com a România)

Nesse ponto, observa-se que há uma relação entre o deslocamento, que pode ser identificado com as *features* 4 e 8. Posto que a presença de um determinante ou nome em início de sentença indica que não houve deslocamento, enquanto a presença da preposição indica que há esse deslocamento, o que está correlacionado com a aparição do sujeito

nulo, que se manifesta no esvaziamento do Arg 1, que pode ser identificado na *feature* 1 [de Castilho et al. 1973] [Chomsky and Lasnik 2008].

O fato dos argumentos (*Arg1* e *Arg2*) e a relação (*Rel*) estarem ou não na sentença apresenta indício do processamento utilizado para a extração de triplas, principalmente por meio dos métodos baseados em regras. Enquanto os casos em que o Arg2 estavam vazios apresentam-se como inválidos em sua totalidade, o fato de a relação (*Rel*) ter sido alterada na tripla indica um processo mais complexo de extração.

Excerto do corpus 2

(7) **GL**

Sentença: Fonte:cronista cumpre do diego video diego el 10 Maradona prepara a lista para enfrentar a España. Coa clasificación ao Mundial, o entrenador Diego Maradona dará a coñecer este venres a lista de convocados para o partido ante España, o 14 de novembre en Madrid.

Tripla inválida: (el 10 Maradona, prepara para, enfrentar a España Coa)

Tripla válida: (el 10 Maradona, prepara, a lista)

(8) **EE**

Sentença: Fuente: cronista cumple del diego video diego el 10 Maradona prepara la lista para enfrentar a España. Con la clasificación al Mundial, el entrenador Diego Maradona dará a conocer este viernes la lista de convocados para el partido ante España, el 14 de noviembre en Madrid.

Tripla inválida: (el 10 Maradona, prepara para, enfrentar a España Con)

Tripla válida: (el entrenador Diego Maradona, dará, la lista)

(9) **PT-BR**

Sentença: Fonte: repórter encontra vídeo de Diego, o 10 Maradona, Diego prepara a lista para enfrentar a Espanha. Com a classificação para a Copa do Mundo, o técnico Diego Maradona anunciará nesta sexta-feira a lista de convocações para a partida contra a Espanha, no dia 14 de novembro, em Madri.

Tripla inválida: (o 10 Maradona, prepara para, enfrentar a Espanha Com)

Tripla válida: (o técnico Diego Maradona, anunciará, a lista)

Por fim, observamos que para cada elemento da tripla, esperam-se determinadas classes gramaticais em posição inicial, sobretudo, no *Arg 1* e na *Rel*, nas quais, respectivamente, são mais frequentes nomes ou determinantes e verbos flexionados.

Excerto do corpus 3:

GL

(10) **Sentença**: Cable audio/video estándar para Xbox 360: Conecta instantáneamente aos xogadores ao mundo de Xbox 360, introduciéndolles en gráficos e xogos de última xeneración utilizando conexión é de definición estándar.

Tripla inválida: (xogos de última xeneración, utilizando, conexión)

EE

(11) **Sentença**: Cable audio/video estándar para Xbox 360: Conecta instantáneamente a los jugadores al mundo de Xbox 360, introduciéndoles en gráficos y juegos de última generación utilizando conexión es de definición estándar.

Tripla inválida: (juegos de última generación, utilizando, conexión)

PT-BR

(12) **Sentença**: Cabo de áudio / vídeo padrão do Xbox 360: Conecta instantaneamente os jogadores ao mundo do Xbox 360, apresentando-os a gráficos e jogos de ponta usando conexão de definição padrão.

Tripla inválida: (jogos de ponta, usando, conexão).

Destaca-se que por outro lado, no *Arg 2*, há uma maior variedade, posto que além dos nomes e determinantes, também são frequentes preposições e outras classes gramaticais. No entanto, notou-se que a presença, por exemplo, de verbos na posição inicial são mais frequentes em triplas inválidas, como pode ser visto no excerto 03: ('o 10 Maradona', 'para enfrentar a Espanha', 'prepara a lista de convocações'). Um ponto de dificuldade dessa extração é justamente a ordem que cada elemento da tripla se apresenta na sentença, mas novamente, a posição inicial da classe gramatical ou da preposição que inicia a *Rel* e o *Arg 2* são pertinentes.

4. Experimentos e resultados

Essa seção descreve o corpus paralelo criado e os experimentos para validar o conjunto de *features* definida para as três línguas.

4.1. Constituição do corpus paralelo em Galego, Português Brasileiro e Espanhol Europeu

Um ponto sensível do desenvolvimento do método proposto é a constituição de *corpora* em Galego, Português (PT-BR) e Espanhol Europeu (EE). A criação de um corpus paralelo permitiu analisar as *features* genéricas. Para tanto, selecionou-se uma base em Espanhol que já fora submetida a experimentos por outros pesquisadores [Barbosa 2018] [Gamallo and Garcia 2011]. O corpus foi criado com 371 triplas em Espanhol Europeu (EE), das quais 271 inválidas e 100 válidas acompanhadas das sentenças das quais foram extraídas. O processo de tradução para o Português (PT-BR) e para o Galego foi realizado por especialistas em Português (PT-BR) e em Galego.

4.2. Experimentos

Antes de realizar os experimentos com os algoritmos de classificação, em pré-processamento, realizaram-se duas tarefas de PLN, a segmentação de sentença em palavras (tokenizer) e classificação morfológica (POS taggers). Logo após esses procedimentos, realizou-se a transformação das variáveis categóricas em binárias, na qual, por exemplo, a *feature* "classe gramatical da palavra no início da sentença" foi substituída por variáveis binárias que indicam a presença (1) ou não (0) de uma determinada classe gramatical na posição inicial das sentenças.

Três experimentos foram realizados com os seguintes modelos de classificação: Regressão Logística, *Lazy Learning* e Árvore de decisão a fim de comparar o desempenho

de cada um na classificação das triplas em válidas ou inválidas. Para o *Lazy Learning*, antes de sua execução, executou-se um algoritmo de busca para localizar a quantidade ideal de vizinhos próximos com base na melhor precisão encontrada.

A seleção de *features* ocorreu de forma iterativa. Ao observar *features* que tratavam de um mesmo aspecto linguístico ou que não colaboram para a melhoria do modelo da classificação, realizaram-se ajustes, mantendo as que apresentavam um melhor desempenho na avaliação, sendo o resultado final apresentado na Tabela 1.

4.3. Avaliação

Com o intuito de avaliar as *features* e averiguar o desempenho das *features*, o corpus foi dividindo em holdlout de 70% para treino e %30 para teste. Além disso, para evitar overfitting, o modelo k-flod 10 foi utilizado. Ao tomar como base estudos realizados por outros pesquisadores, adotou-se a precisão como a medida mais adequada para quantificar os resultados das *features* como relevantes para a EIA. Em ordem de prioridade, o estudo considerou precisão, f1, *recall* e acurácia.

Table 2. *Corpus* paralelo - Galego

Modelo	Precisão	F1	Revocação	Acurácia
Regressão Logística	**0.8296549**	**0.8625751**	0.9029703	0.7650192
Árvore de Decisão	0.7658444	0.8478815	**0.9089476**	**0.7658444**
37 vizinhos próximos	0.7877068	0.8267316	0.8788258	0.7386713

Table 3. *Corpus* paralelo - Português do Brasil

Modelo	Precisão	F1	Revocação	Acurácia
Regressão Logística	0.7727757	**0.8334183**	0.913047	**0.728790**
Árvore de decisão	**0.7754385**	0.8297937	0.9042388	0.7270572
17 vizinhos próximos	0.7445605	0.8222542	**0.9287656**	0.7169777

Table 4. *Corpus* paralelo - Espanhol

Modelo	Precisão	F1	Revocação	Acurácia
Árvore de decisão	0.773257	**0.852036**	**0.9421285**	0.7185568
Regressão Logística	**0.8142896**	0.8193759	0.8368115	0.7210765
57 vizinhos próximos	0.781657	0.8235130	0.8989786	**0.7472554**

A base em Espanhol Europeu foi, também, submetida a teste em experimento realizado por [Barbosa 2018], contudo com um pré-processamento distinto. Nesse experimento, os resultados obtidos pelo modelo Regressão Logística pelas *features* propostas pelo CMULTI foram precisão 0.717, f1 0.828, revocação 0.982 e acurácia 0.714 e pelo ReVerb foram precisão 0.709, f1 0.818, revocação 0.968 e acurácia 0.698. Assim, com excesso da revocação, as demais medidas de avaliação apontam para um melhor desempenho na classificação.

5. Conclusão

Neste trabalho, apresentou-se uma pesquisa dedicada a levantar o maior número de *features* para classificação de triplas relacionais de *corpora* em Galego, Português Brasileiro e Espanhol Europeu. Nesse ponto, destaca-se, também, que aspectos linguísticos comuns ao Galego, Português e Espanhol podem impor maior dificuldade aos modelos de extração arquitetados para realidade da língua inglesa, pois a estrutura Sujeito- Verbo-Objeto (SVO) é menos frequentes nessas línguas.

Assim, as *features* selecionadas nesse estudo possibilitaram a análise das triplas relacionais em todos os conjuntos testados. Entre as *features*, destacaram as relacionadas à presença das preposições e demais classes gramaticas em posição inicial. Dessa forma, os experimentos realizados relacionam estudos da Linguística Formal à Engenharia de *Features*, colaborando para Extração de Informação Aberta em perspectiva multilíngua. Em novos experimentos, os estudo podem ser enriquecidos com a análise de dependência, adoção de novos modelos de classificação e um estudo aprofundado sobre as línguas abordadas.

References

Barbosa, G. C. G. (2018). Utilizando *Features* multi-idioma para classificação de triplas relacionais em português, inglês e espanhol. Masters thesis, Universidade Federal da Bahia, Salvador.

Calvet, L. (2005). *As Politicas Linguísticas*. Parabola, São Paulo.

Chomsky, N. (2014). *The minimalist program*. MIT press, Cambridge.

Chomsky, N. and Lasnik, H. (2008). The theory of principles and parameters. In *Syntax*, pages 506–569. De Gruyter Mouton.

de Castilho, A. T., Kato, M. A., and do Nascimento, M., editors (1973). *Gramática do Português Culto falado no Brasil: a construção da sentença*. Fondo de Cultura Económica, Cidade do México.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 8(4):2011. p. 1535–1545.

Fanjul, A. P. (2014). Conhecendo assimetrias: a ocorrência de pronomes pessoais. In *Syntax*, pages 29–50. Parabola editorial.

Gamallo, P. and Garcia, M. (2011). Multilingual open information extraction. *Portuguese Conference on Artificial Intelligence.*, 8(4):p. 1535–1545.

Gauger, H.-M. (1989). *Introductión a la lingüistica románica*. Gredos, Madrid.