Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

RePort - Um Sistema de Extração de Informações Aberta para Língua Portuguesa

Victor Pereira, Vládia Pinheiro

Programa de Pós-Graduação em Informática Aplicada
Universidade de Fortaleza
Av. Washington Soares, 1321, Fortaleza, Ceará, Brasil

vsantospro@yahoo.com,br, vladiacelia@unifor.br

Abstract. *An emerging field of research in Natural Language Processing (NLP) proposes Open Information Extraction systems (Open IE). Open IEs follow a domain-independent extraction paradigm that uses generic patterns to extract all relationships between entities. In this work, we present RePort, a method of Open IE for Portuguese, based on the ReVerb, an approach for English. Adaptations of syntactic and lexical rules for Portuguese were performed, using linguistic knowledge and a lexicon of verbal relations extracted from a corpus. The evaluation methodology consisted of two experiments where human evaluators indicated 81% accuracy for relations extracted by RePort, and the second experiment showed 77% similarity between the verbal relations extracted by RePort and its correlated extracted by ReVerb (from texts translated into English).*

Resumo. *Um campo emergente de pesquisa em Processamento de Linguagem Natural (PLN) propõe Sistemas de Extração de Informações Abertos (Open Information Extraction System – Open IE) que segue um paradigma de extração independente de domínio que utiliza padrões genéricos para extrair todas as relações entre entidades. Neste trabalho apresentamos RePort, um método de Open IE para língua portuguesa, baseado na abordagem ReVerb para o inglês. Foram realizadas adaptações das regras sintáticas e lexicais para o português, usando conhecimento linguístico e um léxico de relações verbais extraído de um corpus. A metodologia de avaliação consistiu de dois experimentos, onde avaliadores humanos indicaram 81% de acurácia para as relações extraídas pelo RePort, e o segundo experimento mostrou 77% de similaridade entre as relações verbais extraídas pelo RePort e suas relações correlatas, extraídas pelo ReVerb (dos textos traduzidos em inglês).*

1. Introdução

Precípuos projetos em Inteligência Artificial, como o CYC [Lenat 1995] e o NELL (*Never-Ending Language Learning*) [Mitchell et al. 2015] tem como objetivos a aquisição e representação do conhecimento humano em largas bases de conhecimento. Livros, documentos textuais e a própria Web são importantes fontes para aquisição deste conhecimento e, cada vez mais, torna-se importante a investigação e desenvolvimento de métodos e ferramentas computacionais para extração, integração e representação a partir de conteúdo em linguagem natural [Xavier et al. 2015]. Neste sentido, um campo emergente de pesquisa em Processamento de Linguagem Natural

(PLN) propõe Sistemas de Extração de Informações Abertos (em inglês – *Open Information Extraction System – Open IE*). *Open IE* segue um paradigma de extração independente de domínio que utiliza padrões genéricos para extrair todas as relações entre entidades. Wu e Weld (2010) definem um *Open IE* como uma função que, de um documento *d*, retorna um conjunto de triplas na forma (*arg1*, ***rel***, *arg2*), onde *arg1* e *arg2* são sintagmas nominais e ***rel*** é o fragmento de texto que indica a relação semântica implícita entre os argumentos (sintagmas nominais). Importante ressaltar a diferença entre os tradicionais sistemas de Extração de Informação (*Information Extraction* – IE) e sistemas *Open IE*. Sistemas IE objetivam identificar relações estruturadas, de tipos previamente definidos, a partir de fontes não estruturadas como textos. Tais sistemas são normalmente dependentes de domínio e sua adaptação para novo domínio requer um custo de especificação e implementação de novos padrões, ou mesmo um novo processo de anotação de corpora [Eichler et al. 2008]. *Open IE* vem justamente suplantar estas dificuldades e tem como principal característica não necessitar de definição *a priori* dos tipos de relações semânticas a serem extraídas.

TextRunner [Banko et al. 2007] foi o primeiro sistema *Open IE* e utiliza aprendizagem de máquina para mapear padrões de extração. Um dos mais proeminentes sistemas - *ReVerb* [Fader et al. 2011], utiliza heurísticas sintáticas e lexicais para aprendizagem de relações associadas a uma função de confiança. Mais recentemente, a segunda geração do *ReVerb* [Etzioni et al. 2011], incorporou regras de mapeamento para melhoria da extração dos argumentos. *ReVerb* é fruto de mais de 10 anos de estudo em *machine reading* e *web search* dentro do projeto *KnowItAll* da Universidade de Washington, cujo objetivo é a aquisição de grandes quantidades de informação a partir da web. Poucas abordagens *Open IE* têm sido desenvolvidas para outros idiomas além do inglês. Em geral, abordagens que funcionam bem para uma língua envolvem heurísticas e regras específicas para a língua e, quando aplicadas a outros idiomas, se não forem bem adaptadas, não obterão os mesmos resultados. DepOE [Gamallo et al. 2012] se propõe a extrair relações em outras línguas através da aplicação de regras baseadas em *parser* de dependência. No entanto, o custo computacional e a imprecisão são consideradas desvantagens desta abordagem para extrações para Web [Etzioni et al. 2011]. Pinheiro *et al*. (2013) propõem um processo de aquisição de relações semânticas a partir de textos livres da Wikipédia em português, no entanto, sua abordagem é restrita a textos com *hyperlinks*. Neste trabalho, propomos um método de *Open IE* para o Português - *RePort*. O método proposto é baseado em *ReVerb* e foram realizadas adaptações das regras sintáticas e lexicais para o Português, usando conhecimento linguístico e um léxico extraído de um *corpus*. A metodologia de avaliação consistiu de dois experimentos e os resultados demonstraram a boa performance e a viabilidade do método proposto.

2. Estado da Arte

Relações em geral são conexões entre conceitos, entidades, eventos e aquelas expressas por atributos [Xavier et al. 2015]. Por exemplo, da sentença "Joe comprou uma bonita casa", podem ser apreendidas as relações (Joe, comprou, uma bonita casa), (Joe, comprou, uma casa) e (bonita, é propriedade de, casa). Khoo e Na (2006) extrapolaram o conceito de relações semânticas para "associações significativas entre dois ou mais conceitos, entidades, ou conjunto de entidades". Segundo Murphy (2003), não existe nenhuma forma objetiva de decidir o número e quais são os tipos de relações, o que torna o conjunto de relações semânticas um conjunto aberto. Este paradigma norteia as

pesquisas em modelos e sistemas para *Open IE*, segundo o qual é esperado que tais sistemas extraiam todos os tipos de relações n-árias de um texto livre.

Sistemas *Open IE* se baseiam em três paradigmas principais: (i) *machine learning*, que automaticamente aprendem padrões de extração a partir de um corpus de treinamento; (ii) heurísticas, que possuem regras para a seleção e identificação de padrões em textos; e (iii) híbridas, que buscam combinar as duas outras estratégias [Xavier et al. 2015]. *TextRunner* [Banko et al. 2007] foi o primeiro sistema *Open IE* que segue a abordagem de machine learning. Este sistema foi avaliado por revisores humanos que consideraram 80% das relações como corretas. WOE [Wu and Weld 2010] evoluiu o *TextRunner* com heurísticas para novos atributos do conjunto de treinamento e experimentos apontaram uma melhoria de até 34%. *ReVerb* [Fader et al. 2011] foi o primeiro sistema *Open IE* baseado em heurísticas simples que identificam relações verbais e argumentos, recebendo como entrada sentenças em inglês com anotação morfológica e sintática, e, em seguida, aplica uma função de confiança para melhoria da extração dos argumentos. Os autores reportaram melhoria de 50% e 38% da AUC (*area under Precision-Recall curve*) em relação a *TextRunner* e WOE, respectivamente. DepOE [Gamallo et al. 2012] se propõe a extrair relações em outras línguas através da aplicação de regras baseadas em *parser* de dependência. *DepOE* apresentou acurácia de 68%, enquanto que *ReVerb* alcançou 52%. Trabalhos prévios mostraram que caminhos de dependência realmente melhoram a cobertura de sistemas de *Open IE* por capturarem relações não-contíguas [Wu and Weld, 2010]. No entanto, o custo computacional e a imprecisão em extrações para Web são as principais desvantagens desta abordagem. OLLIE [Schmitz et al. 2012] segue a estratégia híbrida que inicia com um treinamento para aprendizado de *templates* a partir das extrações de ReVerb e aplica-os em um *corpus* obtendo novas triplas. OLLIE também usa informações contextuais como atribuição e modificadores clausais. Experimentos indicaram que OLLIE obtém 1.9 maior AUC do que *ReVerb*. LSOE [Xavier et al. 2015] extrai relações usando padrões inspirados na estrutura *Qualia* de Pustejovsky (1995).

São raros *Open IE* para a língua portuguesa. O trabalho de [Collovini et al. 2014] adquire relações a partir da identificação de entidades nomeadas e o DepOE [de Abreu et al. 2013], *Open IE* multilíngue, sem experimentos publicados para o Português. Há outros sistemas, porém baseados em regras para extração de informações fechada, ou seja, para relações pré-definidas [Freitas et al. 2008], não consistindo de abordagens para *Open IE*. [Pinheiro et al. 2013] propõem um processo de aquisição de relações semânticas a partir de textos livres da Wikipédia em português, mas, no entanto, depende da estrutura de links da Wikipédia para a definição dos argumentos das relações, ficando restrita a textos com *hyperlinks*.

3. *RePort* - Extração de Informações Aberta para Língua Portuguesa

Neste trabalho, trilhamos o caminho de desenvolver um modelo de *Open IE* para o Português –*RePort* - baseado na metodologia do sistema *ReVerb*. O motivo da escolha do *ReVerb* foram a maturidade, robustez e sua arquitetura aberta, que possibilitou a reprodução, experimentação e comparação dos resultados para língua portuguesa.

A Figura 1 apresenta a arquitetura funcional de *RePort*. Um texto livre em português é recebido como entrada do processo que realiza a análise morfológica (*tokenizer* e *POS Tagger*) e a análise de sintagmas nominais (*NP chunker*). Em seguida, *RePort* aplica um conjunto de restrições sintáticas e lexicais para identificar sintagmas

relacionais (relações verbais), objetivando a extração da relação ***rel***, e regras para identificação dos argumentos *arg1* e *arg2*. Por fim, um conjunto de relações da forma (*arg1*, ***rel***, *arg2*) são extraídas. Adicionalmente, uma função de confiança pode ser aplicada para avaliar a qualidade das relações extraídas. Nas subseções seguintes, são detalhadas cada etapa do processo de extração de informações aberta de *RePort*.

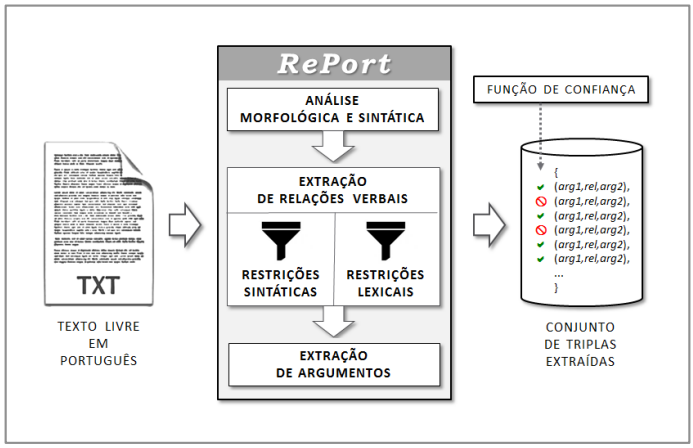

Figura 1. Arquitetura Básica do Modelo de Extração de Informações Aberta – *RePort*.

3.1. Análise Morfológica e Sintática

O método proposto em *RePort* inicia com a etapa de análise morfológica e sintática do texto de entrada em português, aplicando, na ordem a seguir, os seguintes processadores de linguagem natural: (1) Detector de Sentença (*sentence detector*), que realiza a identificação e separação das sentenças do texto; (2) Separador de *tokens* (*tokenizer*), que realiza a identificação palavras ou expressões multi-palavras do texto; (3) Etiquetador morfossintático (*PoS tagger*), que realiza a classificação dos *tokens* e seus modificadores. A anotação de *part-of-speech* é necessária para aplicação das restrições sintáticas – p.ex. em verbos, substantivos, etc; (4) *NP Chunker* – identifica os constituintes sintagmas nominais (NP – *Noum Phrase*) de cada sentença. Associado ao *PoS tagger*, os sintagmas nominais identificados são utilizados na extração dos argumentos.

3.2. Extração de Relações Verbais

O texto anotado na etapa anterior com etiquetas morfológicas e sintáticas segue para esta etapa onde são aplicadas restrições sintáticas e lexicais, com o objetivo de identificar as frases relacionais ***rel***, e regras para extração dos argumentos *arg1* e *arg2*.

(1) Restrições Sintáticas

As restrições sintáticas filtram frases relacionais pouco informativas e incoerentes. Inicialmente, *RePort* identifica, nas sentenças, frases relacionais ***rel*** que combinam com a expressão regular (1), selecionando somente relações verbais.

$$\text{REL} \rightarrow \text{V} \mid \text{V P} \mid \text{V W* P} \qquad \textbf{(1)}$$

A relação verbal ***rel*** pode ser um único verbo ou verbo precedido por verbo de suporte (V) (por exemplo, “publicou” e “precisou publicar”), pode vir seguido por

preposição (V P) (por exemplo, "nasceu em"), ou vir seguido por uma ou mais palavras e finalizado por uma preposição (V W* P) (por exemplo, "abriu inquérito para"). As classes de palavras (W) permitidas são substantivos, adjetivos, advérbios, pronomes ou artigos.

Em seguida, regras heurísticas são aplicadas para refinamento das relações verbais: (1) caso ocorram mais de uma combinação para uma mesma instância (ocorrência) do verbo V, somente a mais longa é selecionada. Por exemplo, no texto "*A Câmara dos Deputados publicou a lei no Diário Oficial da União*", tem-se as relações verbais "publicou" e "publicou a lei em", combinadas pelas expressões regulares (V) e (V W* P), para a mesma ocorrência do verbo "publicar". Nesta situação, apenas a segunda é selecionada; (2) caso sejam encontradas uma ou mais relações adjacentes, todas são agrupadas em uma única relação, como por exemplo ***rel*** = "precisou abrir inquérito para", formada pela relações verbais "precisou" e "abrir inquérito para", as quais aparecem contíguas no texto de entrada.

Este processo de refinamento permite ao modelo considerar relações complexas contendo vários verbos e priorizar relações verbais compostas por combinações de verbo-substantivo, evitando assim relações verbais pouco informativas.

(2) Restrição Lexical

As restrições sintáticas aplicadas no passo anterior podem identificar relações verbais específicas que podem ser pouco representativas na língua. Em outras palavras, quanto mais longa uma relação verbal, mais específica será, e, provavelmente, ocorrerá menos vezes em um *corpus*. No texto: *"A esmagadora maioria dos eleitores quer o PT participando do Governo Fernando Henrique Cardoso"* a relação verbal ***rel*** = "*quer o PT participando de*" pode ser específica e com menor probabilidade de reuso.

Para dirimir este problema, desenvolvemos uma restrição baseada no léxico de relações verbais representativas do português, extraído do *corpus* CetenFolha [Linguateca 2005]. Para a geração deste léxico, são extraídas todas as relações verbais usando as restrições sintáticas (conf. Passo (1)), e contabilizado, para cada relação verbal, um índice de especificidade *k*, o qual expressa o número de instâncias que a relação verbal possui no *corpus* em questão. A intuição é que, quanto maior o valor de *k*, mais inespecífica e representativa a relação verbal será.

A partir do *corpus* CetenFolha, foram extraídas 1.552.791 relações (*arg1*, ***rel***, *arg2*) distintas, as quais foram agrupadas pela relação verbal ***rel***, resultando em 441.230 relações verbais diferentes. A cada grupo foi associado um valor *k* de instâncias, ou seja, um número *k* de pares de argumentos distintos. Por exemplo, ***rel*** = "abrir inscrição para" aparece 56 vezes no *corpus* com argumentos distintos, neste caso *k*=56, enquanto ***rel*** = "implementar reforma para aproveitar tendência de" aparece apenas 1 vez, e, portanto, *k*=1. No léxico gerado, tem-se que 6.159 relações verbais possuem *k*>=20.

Com base no léxico de relações verbais do Português, *RePort* aplica a seguinte restrição lexical: (1) selecionar as relações verbais com nível de especificidade maior ou igual a *k* (definido por parâmetro).

Nos experimentos do *ReVerb* os melhores resultados foram obtidos com parâmetro *k*=20. Para este valor de *k*, no léxico em inglês são consideradas 941.232 relações verbais, enquanto que para o léxico em português, o parâmetro acima selecionaria apenas 6.159 relações verbais, o que tornaria a restrição lexical do *RePort*

muito mais rígida e, consequentemente, muitas triplas seriam descartadas. Em nossas avaliações experimentais, alcançamos os melhores resultados com k=2, importando em um léxico com 84.341 relações verbais passíveis de extração pelo *RePort*. Importante ressaltar que o corpus em inglês utilizado no *ReVerb* possui 384 vezes mais sentenças que o CetenFolha (500 milhões de sentenças contra 1,3 milhão de sentenças).

3.3. Extração de Argumentos

Para cada relação verbal selecionada na etapa anterior, *RePort* aplica as seguintes regras para identificar os argumentos *arg1* e *arg2* da relação:

(1) Atribuir à *arg1* o sintagma nominal mais próximo à esquerda da relação verbal ***rel***, na mesma sentença, desde que o mesmo não seja um pronome relativo, nem um pronome reflexivo, nem a palavra "quem";
(2) Verificar se *arg1*, na mesma sentença, é precedido por outro sintagma nominal e intercalado pela preposição "de". Em caso afirmativo, acrescenta-se este último a *arg1*. Repete-se esta regra até encontrar o último sintagma nominal que satisfaça a condição;
(3) Verificar se *arg1* é um nome próprio, e se, na mesma sentença, está precedido por outro nome próprio e intercalado por conjunção coordenada "e" ou vírgula. Em caso afirmativo, acrescenta-se este último a *arg1*. Repete-se esta regra até encontrar o último nome próprio que satisfaça a condição;
(4) Atribuir à *arg2* o sintagma nominal mais próximo à direita da relação verbal ***rel***, na mesma sentença;
(5) Verificar se *arg2*, na mesma sentença, é sucedido outro sintagma nominal e intercalado pela preposição "de". Em caso afirmativo, acrescenta-se este último a *arg2*. Repete-se esta regra até encontrar o último sintagma nominal que satisfaça a condição.

As regras (2) e (5) são especialmente necessárias, desde que, na língua portuguesa, utiliza-se prioritariamente a preposição "de" para adjetivação de substantivos [Lima 1972] [Junior 2002]. Por exemplo, sem as referidas regras, a relação extraída da sentença "*Filhos de Gandhi é campeã do carnaval de 2013*" é "(Gandhi, ser campeão de, o carnaval)" enquanto, pelas regras (2) e (5), a relação corretamente extraída é "(Filhos_de_Gandhi, ser campeão de, o Carnaval_de_2013)".

4. Avaliação Experimental

Os experimentos realizados visaram avaliar a qualidade das relações extraídas pelo método *RePort*, a partir de textos livres em português, através de uma avaliação manual e da comparação com as triplas extraídas pelo *ReVerb*, a partir dos textos correlatos em inglês.

4.1. Configuração e Metodologia de Avaliação

Nos experimentos realizados foram utilizados o *sentence detector* do OpenNLP[1], e o *tokenizer*, *POS tagger* e *NP chunker* de [Kinoshita et al. 2006] e [Colen 2013], processadores que possuem alta acurácia para língua portuguesa. A metodologia de avaliação seguiu os passos detalhados abaixo:

[1] http://opennlp.sourceforge.net

(1) Extração das relações de um dos artigos do *corpus* multilíngue REVISTA PESQUISA FAPESP PARALLEL CORPORA (NILC) [Aziz and Specia 2011], usando o *ReVerb 1.3* para os textos em inglês, e o *RePort* para os textos correlatos em português. Foram extraídas 93 relações em português e 94 em inglês (com parâmetros k=2 e k=20, respectivamente).

(2) Execução dos seguintes cenários de teste:

CENÁRIO 1 – Sete avaliadores adultos e mestrandos em Ciências da Computação, de posse do texto original em português e da relação semântica extraída pelo *RePort*, qualificavam-na, seguindo a escala Likert – CONCORDO, CONCORDO_PARCIALMENTE, NAO_SEI_DIZER, DISCORDO_PARCIALMENTE, DISCORDO. Pelo menos dois avaliadores distintos qualificaram cada relação, com um terceiro avaliador para resolver casos de divergência. Neste cenário, foram calculadas duas métricas:

$\text{Acurácia}_{\text{Restrita}}$ = (Qtd_CONCORDO) / Qtd_Extracoes

$\text{Acurácia}_{\text{Relaxada}}$ = (Qtd_CONCORDO + Qtd_CONCORDO_PARC) / Qtd_Extracoes

CENÁRIO 2 – Um avaliador humano proficiente em inglês e português comparou as relações extraídas pelo *ReVerb* e pelo *RePort* com o objetivo de verificar a similaridade entre elas. A análise de similaridade considerou os seguintes casos: (1) as relações verbais são iguais ou parcialmente iguais; (2) adicionalmente *arg1* e/ou *arg2* são iguais. Neste cenário, foi calculada a seguinte métrica:

$\text{Similaridade}_{\text{Port-Ing}}$ = (Qtd_SIMILARES) / Qtd_Extracoes

A Tabela 1 apresenta, como exemplo, as relações extraídas para o texto "*O Movitae foi criado em 2003, quando o Congresso Nacional iniciava os debates sobre clonagem terapêutica, técnica que ficou fora da Lei de Biossegurança.*" (em inglês, *"Movitae was created in 2003, when the National Congress was starting the debates on therapeutic cloning, a technique that was left out of the Law on Biosafety"*).

Tabela 1. Exemplo de extrações / avaliações de triplas extraídas pelo *RePort* e ReVerb.

EXTRAÇÕES *RePort* (texto em português)		
arg1	***rel***	*arg2*
O Movitae	foi criado em	2003
o Congresso Nacional	iniciava os debates sobre	clonagem terapêutica
Técnica	ficou fora de	a Lei de Biossegurança

EXTRAÇÕEs *ReVerb* (texto em inglês)		
arg1	***rel***	*arg2*
Movitae	*was created in*	*2003*
the National Congress	*was starting the debates on*	*therapeutic cloning*
a technique	*was left out of*	*a Lei de Biossegurança*

4.2. Análise dos Resultados

No CENARIO 1, foram avaliadas 91 das relações extraídas pelo *RePort* (2 relações foram marcadas como NAO_SEI_DIZER), das quais os avaliadores concordaram plenamente com 57, concordaram parcialmente em 17 casos, discordaram parcialmente em 4 casos e, em 13, discordaram totalmente. Assim, *RePort* obteve $\text{Acurácia}_{\text{Restrita}}$ = 62,6% (considerando apenas as relações avaliadas como completamente corretas por todos os avaliadores) e $\text{Acurácia}_{\text{Relaxada}}$ = 81,3% de acurácia relaxada (considerando as relações em que os avaliadores concordaram com o sistema, mesmo que parcialmente).

No CENARIO 2, foram analisadas as 93 relações verbais extraídas pelo *RePort* e suas correlatas extraídas pelo *ReVerb*. O avaliador humano considerou que 61 relações verbais ***rel*** eram totalmente correspondentes e 10 eram parcialmente. Das 61 relações verbais coincidentes, foram analisadas a similaridade entre os argumentos das respectivas relações: 37 possuíam ambos os argumentos iguais e 21 apresentaram ou *arg1* ou *arg2* iguais. Portanto, considerando apenas as relações verbais, tem-se 66% de similaridade total e 76% de similaridade parcial, e considerando também os argumentos, tem-se 40% similaridade total e 62% de similaridade parcial (quando um dos argumentos das relações coincidem).

Diferença entre os léxicos de relações verbais em inglês e português foi a principal causa das 10 (dez) relações verbais com similaridade parcial (por exemplo, "*will enjoy the results of*" se encontra no léxico em inglês e não foi encontrada no corpus do CetenFolha). Tem-se, ainda, que 22 relações extraídas pelo *RePort* não foram extraídas pelo *ReVerb*, e 23 extrações feitas pelo *ReVerb* não foram extraídas pelo *RePort*. Identificamos que os principais motivos foram: (i) verbos em português expressos como substantivos em inglês; (ii) verbos em inglês expressos como substantivos em português; (iii) uso de sujeito elíptico no português; (iv) restrição de verbos precedidos pela partícula "to". Por exemplo, no texto "(...), suspeitavam estar estimulando a prática do aborto.", a elipse do sujeito em português fez com que a relação não fosse extraída. Noutro exemplo, do texto "*The farmers from Rio Grande do Sul planted RR seeds from Argentina*.", *ReVerb* extrai as relações incorretas - (*the farmers from Rio Grande", do, Sul*) (*the farmers from Rio Grande, planted, RR Seeds*), por entender "do" como verbo. Importante salientar que *RePort* extraiu corretamente a relação (os produtores do Rio Grande do Sul, plantaram, sementes RR argentinas). Importante ressaltar que as 22 novas relações extraídas por *RePort* obtiveram 75% de acurácia.

5. Conclusão

Neste trabalho apresentamos *RePort*, um método de Extração de Informação Aberta para língua portuguesa, baseado na abordagem *ReVerb* para o inglês, e consistindo de regras para seleção da relação verbal e para extração dos argumentos. Foram realizados dois experimentos, onde uma avaliação manual indicou 81% de acurácia para as relações extraídas pelo *RePort*, e o segundo experimento mostrou 76% de similaridade entre as relações verbais extraídas pelo *RePort* e suas correlatas extraídas pelo *ReVerb* (dos textos traduzidos em inglês). Destaca-se que o índice de similaridade decresce para 62%, quando a avaliação considerou também os argumentos das relações. A análise dos casos de erro indiciaram os seguintes trabalhos futuros: (i) novas regras para extração de argumentos, por exemplo, que considere conhecimento linguístico como outras preposições, sujeito oculto, etc; (ii) extensão do léxico de relações verbais a partir de outros *corpora*; (iii) comparação com outros modelos de *Open IE*, como o *DepOE*; (iv) implementação da função de confiança para o português; (v) avaliação a partir de um padrão ouro. Importante reivindicar a relevância do presente trabalho para evolução da área de *Open IE* para o Português. As triplas extraídas pelo *ReVerb* com alto índice de confiança são utilizadas para aprendizado de máquina em outros sistemas, como OLLIE e OpenIE 4 (http://knowitall.github.io/openie), os quais apresentam resultados significativamente melhores para a língua inglesa. Portanto, o caminho trilhado neste trabalho foi necessário para que possamos avançar mais rapidamente nas pesquisas em extração aberta de relações semânticas em língua portuguesa.

Referências

Aziz, Wilker, and Lucia Specia (2011), 'Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation', in *STIL 2011* (Cuiabá, MT, 2011).

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni (2007), 'Open Information Extraction for the Web', in *IJCAI*, 2007, VII, pp. 2670–76.

Colen, W. (2013). Aprimorando o corrector grammatical cogroo. Master's thesis, IME/USP - Inst. de Matemática e Estatística da Universidade de São Paulo.

Collovini, Sandra, et al. (2014) Extraction of Relation Descriptors for Portuguese Using Conditional Random Fields. *Advances in Artificial Intelligence--IBERAMIA 2014*. Springer International Publishing, 2014. 108-119.

Eichler, Kathrin., Hemsen, Holmer., Neumann, Günter. (2008). Unsupervised relation extraction from web documents. In: Proceedings of the International Conference on Language Resources and Evaluation.

Etzioni, Oren, Anthony Fader, Janara Christensen et al. (2011), 'Open Information Extraction: The Second Generation.', in *IJCAI*, 2011, XI, 3–10.

de Abreu, Sandra C., Bonamigo, Tiago Luis., Vieira, Renata. (2013) A review on Relation Extraction with an eye on Portuguese. Journal of the Brazilian Computer Society, Springer, v 19, Issue 4, pp. 553-571. [doi 10.1007/s13173-013-0116-8].

Fader, Anthony, Stephen Soderland, and Oren Etzioni (2011), 'Identifying Relations for Open Information Extraction', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2011), pp. 1535–45.

Freitas, C., Santos, D., Oliveira, H.G., Carvalho, P., Mota, C. (2008) Relações semânticas do ReRelEM: além das entidades no Segundo HAREM, Chapter 4, pp. 75–94. Linguateca.

Gamallo, Pablo, Marcos Garcia, and Santiago Fernández-Lanza (2012), 'Dependency-Based Open Information Extraction', in *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP* (Association for Computational Linguistics, 2012), pp. 10–18.

Junior, E. D. (2002). Preposições no Português Brasileiro: Um Estudo Frequencial. Tese de Doutorado. Universidade Federal do Paraná.

Khoo, C., Na, J.C. (2006) Semantic relations in information science. Annual Review of Information Science and Technology 40, pp.157–228.

Kinoshita, Jorge, L. N. Salvador, and C. E. D. Menezes (2006), 'CoGrOO: A Brazilian-Portuguese Grammar Checker Based on the CETENFOLHA Corpus', in *The Fifth International Conference on Language Resources and Evaluation, LREC*, 2006

Lima, R. (1972) Gramática normativa da língua portuguesa. Rio de Janeiro: José Olympio Editora.

Lenat D (1995) CYC: a large-scale investment in knowledge infrastructure. Communications of the ACM 38(11): pp.33–38.

Linguateca (2005) "CETENFolha", http://www.linguateca.pt/CETENFolha.

Mitchell, T., W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, J. Welling. (2015) 'Never-Ending Learning'. In Proceedings of the Conference on Artificial Intelligence (AAAI), 2015.

Murphy, M.L. (2003) Semantic Relations and the Lexicon. Cambridge University Press, Cambridge, UK.

Pinheiro, V., Furtado, V., Pequeno, T., Franco, W. (2013) A Semi-Automated Method for Acquisition of Commonsense and Inferentialist Knowledge. Journal of the Brazilian Computer Society, Springer, v.19, pp. 75-87 [doi:10.1007/s13173-012-0082-6].

Pustejovsky, J. (1995) The Generative Lexicon. The MIT Press, Cambridge, USA.

Schmitz, Michael, Robert Bart, Stephen Soderland, Oren Etzioni 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* (Association for Computational Linguistics, 2012), pp. 523–34.

Wu, Fei, and Daniel S. Weld (2010), 'Open Information Extraction Using Wikipedia', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2010), pp. 118–27

Xavier, Clarissa Castellã, Vera Lúcia Strube de Lima, and Marlo Souza (2015), 'Open Information Extraction Based on Lexical Semantics. *Journal of the Brazilian Computer Society 2015, 21:4 doi:10.1186/s13173-015-0023-2.*