TransAlign: tradução e alinhamento de corpora para a língua portuguesa

Alan Rios Melo[1], **Daniela Barreiro Claro**[1]

[1]FORMAS Research Group
Instituto de Computação, Universidade Federal da Bahia
Salvador - Bahia - Brazil

{alan.rios,dclaro}@ufba.br

Abstract. *In this paper, we introduce TransAlign, an innovative framework to enhance Open Information Extraction (OpenIE) in underrepresented languages, such as Portuguese, by leveraging data from resource-rich languages. Utilizing specific grammatical rules and high-quality translation models, we adapted LSOIE, a large-scale dataset, for Portuguese. This approach generated 21.161 high-quality triples for OpenIE in Portuguese. The resulting dataset enabled the training of a new model that improved F1 scores by 50% over existing systems for Portuguese*

Resumo. *Neste artigo, apresentamos o TransAlign, uma estrutura inovadora para ampliar a Extração Aberta de Informações (OpenIE) em idiomas sub-representados, como o português, usando dados de idiomas ricos em recursos. Utilizando regras gramaticais específicas e modelos de tradução de alta qualidade, adaptamos o LSOIE, um conjunto de dados de grande escala, para o português. Essa abordagem gerou 21.161 triplas de alta qualidade para OpenIE em português. O conjunto de dados resultante possibilitou o treinamento de um novo modelo que melhorou em 50% os escores F1 dos sistemas existentes para o português*

1. Introdução

A Extração Aberta de Informações (OpenIE) é um processo essencial no processamento de linguagem natural (PLN) que extrai informações estruturadas de texto não estruturado [Banko et al. 2007, Etzioni et al. 2008]. Apesar dos avanços nos sistemas OpenIE para inglês [Angeli et al. 2015, Stanovsky et al. 2018], ainda há desafios para idiomas sub-representados [Akbik et al. 2019b]. Este artigo apresenta o TransAlign, um framework que utiliza traduções e regras de alinhamento específicas do idioma para criar conjuntos de dados OpenIE em idiomas sub-representados. Utilizamos o português como exemplo, gerando um novo conjunto de dados a partir do LSOIE [Solawetz and Larson 2019], resultando em melhorias significativas no desempenho do OpenIE em português. Nosso trabalho destaca o potencial de grandes conjuntos de dados e ferramentas de tradução para a pesquisa OpenIE em idiomas sub-representados. Este artigo está organizado como segue: a seção 2 descreve a validação dos dados, a seção 3 a construção do dataset e a seção 4 o modelo proposto, tendo a seção 5 os experimentos e resultados.

2. Verificador de correspondencia

Com o objetivo de assegurar o alinhamento correto de todos os dados utilizados no treinamento do modelo, foi desenvolvido um algoritmo de correspondência. Esse algoritmo busca identificar as correspondências dos elementos das triplas ('ARG0', 'REL', 'ARG1'), garantindo que cada elemento seja encontrado em conjunto com todos os tokens pertencentes à mesma sentença, sem qualquer interrupção de palavras externas às triplas. O algoritmo de correspondência possui duas abordagens: busca sequencial e busca não sequencial. Na busca sequencial, são selecionadas as sentenças que possuem exatamente a ordem (ARG0, REL, ARG1) dos elementos da tripla encontrados. Já na busca não sequencial, não há essa limitação, considerando todas as possíveis ordens de elementos como válidas. No método apresentado neste artigo, a abordagem sequencial foi descrita, devido à menor complexidade no treinamento dos modelos.

3. Construção do Dataset

A construção do conjunto de dados foi motivada pela escassez de dados em grande quantidade para a tarefa de OpenIE na língua portuguesa do Brasil. Tentativas anteriores de obtenção desses dados, utilizando diferentes técnicas, resultaram em dados inconsistentes e de qualidade mediana. Diante da necessidade de construir um conjunto de dados amplo e de melhor qualidade, a abordagem proposta traduziu e alinhou os conjuntos de dados do inglês para o português brasileiro.

3.1. TransAlign

Na primeira tentativa, o 'PTOIE', originado da tradução do 'SQuAD v2'[Rajpurkar et al. 2016], apresentou ruídos, levando a 7.344 extrações de média qualidade e 2.472 correspondidas de 130.000 instâncias iniciais.

O TransAlign, um conjunto de dados traduzido do inglês para o português, minimiza o ruído de conversão, embora a tradução possa introduzir algumas incompatibilidades. Utilizando a API do Google Tradutor e o GPT 3.5 para a tradução, foram obtidas 7.000 e 22.124 extrações de válidas, respectivamente, de um total de 49.566. Um desafio significativo é a preservação das características de anotação das sentenças durante a tradução, o que requer um processo de alinhamento cuidadoso. Este processo é composto por três etapas principais:

Busca de Combinações de Relações: O algoritmo busca todas as combinações possíveis de relações na sentença traduzida, que podem ocorrer entre o primeiro e ultimo token.

Correspondência de Triplas e Frases: O algoritmo procura correspondências entre a tripla gerada e a frase inteira. Se uma correspondência é encontrada, a 'pos tag' é anotada com base na analise feita na sentença, utlizando o modelo treinado pelo spacy [Honnibal and Montani 2017].

Verificação de 'Pos Tags': O algoritmo percorre a 'postag' anotada da combinação de relação, a etapa é uma adaptação das regras apresentadas no reverb para a lingua pt-br [Fader et al. 2011], verificando se a relação começa com um verbo e se o restante das tags da relação pertence a um advérbio, adjetivo, verbo, pronome ou substantivo. A estrutura da relação determina o último token necessário para a relação ser considerada válida.

Table 1. Estatisticas de Conversão TransAlign

	Original	TransAlign
LSOIE Train [Solawetz and Larson 2019]	49.566	15.006
LSOIE Test[Solawetz and Larson 2019]	10.783	3.282
LSOIE Dev[Solawetz and Larson 2019]	9.459	2.873
CARB[Bhardwaj et al. 2019]	3.497	715
S2 Train[Kolluru et al. 2022]	166.032	77.805
S2 Valid[Kolluru et al. 2022]	1.872	923
all	231.750	100.604

Se houver mais de uma combinação possível de relação, a maior delas é escolhida, pois a escolha da relação com maior quantidade de tokens melhora e complementa o contexto da extração. Após a seleção da relação, a extração é realinhada, com todos os tokens antes do primeiro token da relação considerados como argumento 0, e todos os tokens após o último token da relação considerados como argumento 1. Gerando assim, um conjunto de dados anotados OpenIE para o português do Brasil

- *Inglês:*

Sentença: *"English longbow was also used against the English by their Welsh neighbours."*

Tripla: *(English longbow – was also used – against the English by their Welsh neighbours)*

- *Português:*

Sentença: *"O arco longo inglês também foi usado contra os ingleses por seus vizinhos galeses."*

Tripla: *(O arco longo inglês – foi usado contra – os ingleses por seus vizinhos galeses)*

4. Modelo

O modelo utilizado para o treinamento foi desenvolvido com base na framework Flair-NLP [Akbik et al. 2019a], adotando uma metodologia de Rotulagem de Sequência. A arquitetura do modelo inicialmente incorpora um Word Embedding [Akbik et al. 2018], seguido por embeddings direcionais, tanto para frente quanto para trás [Akbik et al. 2018]. Esses três conjuntos de embeddings são então concatenados e direcionados para uma camada linear, que unifica e representa todos os codificadores. Após essa reprojeção, adicionamos duas camadas de Redes Neurais Recorrentes (RNN) para processamento subsequente. Na extremidade da arquitetura, posicionamos um classificador que conta com uma camada CRF (Campo Aleatório Condicional).

4.1. Fine-Tunning

Após a fase de treinamento, o modelo passa por um processo de ajuste fino, ou fine-tuning. Esse processo emprega o dataset Pud-200[Cabral et al. 2022], uma seleção de extrações "silver" anotadas manualmente. Essa etapa de ajuste ocorre durante 20 épocas. Sendo este processo importante por se tratar de dados nobres que ajudam na precisão das extrações.

4.2. Datasets utilizados

O modelo TransAlign foi treinado com datasets traduzidos e alinhados, conforme apresentado na Tabela 1. O modelo PTOIE, por outro lado, foi treinado usando a técnica de conversão no conjunto de dados SQuAD v2. Todos os modelos foram validados usando o conjunto de dados PUD-100 [Cabral et al. 2022], que inclui 100 sentenças com extrações 'gold', que foram anotadas e validadas manualmente por três anotadores distintos. Apenas as extrações que receberam dois ou mais votos positivos foram selecionadas.

5. Experimentos e Resultados

Os modelos experimentais são comparados usando métodos estatísticos e qualitativos. As métricas estatísticas incluem precisão, recall e f1-score, gerados a partir da validação com o dataset Pud-100 na ferramenta CaRB [Bhardwaj et al. 2019]. A análise qualitativa envolve a revisão manual das triplas extraídas pelos modelos, proporcionando uma visão realista da qualidade e aplicabilidade prática dos modelos treinados.

O conjunto de dados exibiu um desempenho robusto no treinamento do modelo, resultando em triplas coerentes e contextualmente relevantes. O modelo alcançou métricas de benchmark, com f1-score de 0.3228 e 0.3766, precisão de 0.4137 e 0.4827, e cobertura de 0.2647 e 0.3088, no *identical match* e *lexical match*, respectivamente. Ao avaliar a capacidade do modelo de extrair fatos de um texto gerado pelo GPT-4 sobre furtos fictícios em uma cidade, o modelo demonstrou habilidade em realizar inferências precisas e extrair informações relevantes, o texto possui 249 palavras.

Algumas extrações:
"(Adrian, estava fora de, a cidade)"
"(Sua casa, estava, vazia)"
"(Clara uma senhora idosa, cuidava de, a casa)"
"(Clara, voltou de, sua caminhada noturna)"
"(A porta de a frente, estava, entreaberta)"
"(Clara, chamou, a polícia)"
"(Os policiais, chegaram em, 10 minutos)"
"(Os indícios, sugeriram que, os criminosos eram profissionais)"
"(O roubo, pareceu ter ocorrido, entre as 20h e as 20h30)"

6. Conclusão e Trabalhos Futuros

Este estudo produziu resultados significativos para a língua portuguesa na tarefa de EIA, demonstrando a viabilidade de gerar dados de alta qualidade e enriquecer idiomas subrepresentados. A expectativa futura é superar desafios relacionados ao aumento da complexidade dos dados e à cobertura de diferentes estruturas, o que parece ser uma meta alcançável com base nos resultados atuais.

Agradecimentos

O presente trabalho conta com o apoio da CAPES-Brasil - Código de Financiamento 001 e da FAPESB - Projeto TIC.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019a). Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Akbik, A., Bergmann, T., and Vollgraf, R. (2019b). Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728. Association for Computational Linguistics.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Angeli, G., Premkumar, M. J. 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 and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354. Association for Computational Linguistics.

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical intelligence*, pages 2670–2676. University of Washington.

Bhardwaj, S., Aggarwal, S., and Mausam, M. (2019). CaRB: A crowdsourced benchmark for open IE. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.

Cabral, B., Souza, M., and Claro, D. B. (2022). Portnoie: A neural framework for open information extraction for the portuguese language. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 243–255, Cham. Springer International Publishing.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. pages 1535–1545.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kolluru, K., Mohammed, M., Mittal, S., Chakrabarti, S., and ., M. (2022). Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Solawetz, J. and Larson, S. (2019). LSOIE: A large-scale dataset for supervised open information extraction. *arXiv preprint arXiv:2101.11177*.

Stanovsky, G., Michael, J., Zettlemoyer, L., and Dagan, I. (2018). Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895. Association for Computational Linguistics.