Desafios da tarefa de Extração de Informação Aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual

Beatriz Paixão Queiroz[1], **Rerisson Cavalcante**[1], **Daniela Barreiro Claro**[2]

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

{beatrizpaixao, rerisson.cavalcante, dclaro}@ufba.br

Abstract. *This work describes the methodology defined for the construction of a corpus, in Portuguese, manually annotated from an automated generation of a corpus for the Open Information Extraction task in Portuguese. Rules were defined for the extraction of triples in order to subsidize the generation of corpus in the creation of models based on machine learning. The results obtained were the generated corpus that has been used in the creation of algorithms for the EIA task.*

Resumo. *Este trabalho descreve a metodologia definida para a construção de um corpus, em português, anotado manualmente a partir de uma geração automatizada de um corpus para a tarefa de Extração de Informação Aberta em Português. Regras foram definidas para a extração de triplas com o objetivo de subsidiar a geração de corpus na criação de modelos baseado em aprendizado de máquinas. Os resultados obtidos foram o corpus gerado que vem sendo utilizado na criação de algoritmos para a tarefa de EIA.*

1. Introdução

A criação de corpora anotados é essencial para o treinamento e validação de recursos voltados para o Processamento de Linguagem Natural (PLN). Este trabalho traz uma abordagem metodológica para a construção de um corpus anotado, desde a geração automatizada de sentenças através da tradução até a geração manual para a tarefa de Extração de Informação Aberta (Open Information Extraction, OpenIE) [Banko et al. 2007] voltadas para o português brasileiro. A Extração de Informação Aberta consiste em gerar informação estruturada de textos não estruturados a qual normalmente é descrita através da tripla (arg0, rel, arg1) [Fader 2011]. Neste contexto, os vínculos semânticos entre entidades (pessoas, organizações, locais, datas etc) são mapeadas como relações.

As novas técnicas de aprendizagem de máquinas exigem grandes conjuntos de dados para que se possam treinar, validar e testar novos os algoritmos, fazendo assim avançar o estado da arte da OpenIE. A abordagem proposta consistiu da premissa de tradução de datasets em inglês para o português com o objetivo de fazer um alinhamento das extrações automatizadas. A partir das extrações inválidas, um novo corpus foi proposto com extrações manuais geradas por um linguista júnior e validadas por um linguista sênior, cujo objetivo é manter a alta qualidade do recurso criado. Esta metodologia foi composta por quatro etapas: (1) tradução do corpus; (2) análise das sentenças traduzidas,

para a extração manual de triplas; (3) anotação sintática manual dos componentes das triplas extraídas; (4) validação do corpus e a proposição de extrações adicionais.

Diversos trabalhos têm discutido metodologias para a criação de corpus grande, cujo objetivo é ser utilizado nas tarefas de PLN, especificamente em OpenIE [Stanovsky and Dagan 2016, Glauber et al. 2018]. Porém a principal limitação é a pequena quantidade de sentenças e de extrações, o que inviabiliza o uso de métodos baseados em aprendizado de máquinas. Assim, o principal objetivo deste trabalho é a criação de um corpus grande para ser utilizado em métodos baseados em redes neurais. As próximas seções detalham o conteúdo deste trabalho.

2. Etapa 1 : Tradução do SQuAD v2

O SQuAD v2 é um conjunto de dados desenvolvido para a tarefa de *Question Answering* (QA) em inglês e tem uma estrutura que permite sua conversão para a tarefa de *Open Information Extraction* (OpenIE). Neste processo de conversão, uma resposta do conjunto de dados original é interpretada como uma relação. Essa relação é então vinculada a um argumento primário (arg0), determinado por uma pergunta W (who, where, what, etc.), e é complementada com um segundo argumento (arg1), conforme adaptado dos autores em [Stanovsky and Dagan 2016].

As 7344 sentenças do SQuAD v2 foram traduzidas para o português por ferramentas automatizadas. Um subconjunto de 360 das sentenças traduzidas serviu de base para a tarefa de extração de construção do corpus manualmente anotado. Nesse processo, foram encontrados alguns erros de tradução nas sentenças iniciais. Com o intuito de validar esta primeira etapa, as sentenças foram revisadas e corrigidas por um linguista senior.

3. Etapa 2: Extração manual de triplas a partir das sentenças

Após a tradução dos dados do SQuAD v2, dois conjuntos de tarefas foram iniciadas, que resultaram em dois corpora de extrações (i) a validação manual das extrações geradas automaticamente pelo algotimo, com a remoção das inválidas; (ii) a geração manual de novas extrações a partir de sentenças para as quais não foram registradas extrações automáticas coerentes ou seja, extrações inválidas.

A segunda dessas duas tarefas resultou no corpus descrito neste trabalho, com o acréscimo das etapas subsequentes: (a) anotação morfossintática manual dos constituintes das novas triplas propostas; (b) revisão das triplas propostas e das anotações. O objetivo principal foi a geração da maior quantidade possível de extrações válidas a partir da mesma sentença.

Para estruturação das regras, também foi utilizada uma das restrições sintáticas mencionadas pelos autores em [Fader 2011], na qual toda relação com múltiplas palavras pode iniciar com um verbo, terminar com uma preposição e consistir em uma sequência contígua de palavras.

3.1. Regras para extrações válidas

Regra 1: O argumento 0 (arg0) deve ser um sintagma nominal à esquerda do verbo, excluídos SN formados apenas por pronomes.
- Sentença (1): "*Clésinger fez a máscara da mortuária de Chopin*".
- Extração: (**Clésinger**; fez; a máscara da mortuária de Chopin).

Regra 2: O argumento 1 deve estar à direita do verbo e pode ser (i) um sintagma nominal, como no exemplo (1) acima; (ii) uma sequência formada por um sintagma nominal e outro(s) sintagma(s) complemento(s) ou adjunto(s), como no exemplo (2), respeitada a contiguidade; (iii) ou uma sentença, como no exemplo (3).
- Sentença (2): "*Peter Stuyvesant entregou Nova Amsterdã aos ingleses*".
- Extração: (Peter Stuyvesant; entregou; **Nova Amsterdã aos ingleses**)
- Sentença (3): "*O sismólogo japonês Yuji Yagi disse que o terremoto ocorreu em duas etapas*."
- Extração: (O sismólogo japonês Yuji Yagi; disse; **que o terremoto ocorreu em duas etapas**).

Regra 3: A relação deve conter pelo menos um verbo. Se houver uma preposição introduzindo o argumento 1, esta deve ser incluída na relação.
- Sentença: "*Léon Escudier escreveu sobre um recital de Chopin*".
- Extraction: (Léon Escudier; **escreveu sobre**; um recital de Chopin).

Regra 4: Se houver vários sintagmas complementos e adjuntos à direita do verbo, várias extrações foram feitas, com diferentes combinações dos elementos, respeitada a contiguidade. - Sentença (4): "*Chopin visitou Berlim com um amigo*".
- Extração 1: (Chopin; visitou; **Berlin**)
- Extração 2: (Chopin; visitou; **Berlin com um amigo**)
- Extração 3: (Chopin; visitou Berlin com; **um amigo**)
- Sentença (5): "*Peter Stuyvesant entregou Nova Amsterdã aos ingleses sem derramamento de sangue*".
- Extração 1: (Peter Stuyvesant; entregou; **Nova Amsterdã aos ingleses**)
- Extração 2: (Peter Stuyvesant; entregou; **Nova Amsterdã aos ingleses sem derramamento de sangue**)
- Extração 3: (Peter Stuyvesant; entregou Nova Amsterdã a; **os ingleses**)
- Extração 4: (Peter Stuyvesant; entregou Nova Amsterdã a; **os ingleses sem derramamento de sangue**)

Regra 5: A negação, outros advérbios e pronomes átonos imediatamente pré-verbais devem ser incluídos na relação.
- Sentença (6): "*A escola Theravada não inclui as escrituras Mahayava em seu cânon*".
- Extração: (a escola Theravada; **não inclui**; as escrituras Mahayava em seu cânon)
- Sentença (7): "*John F. Shea se formou em 1908*".
- Extração: (John F. Shea; **se formou em**; 1908)

Regra 6: A voz passiva permite diferentes extrações, com o particípio incluído na relação e no argumento 1.
- Sentença (8): "*A Batalha de Long Island foi travada em agosto de 1776*"
- Extração: (A Batalha de Long Island"; **foi travada em**; agosto de 1776)
- Sentença (9): "*Todas as rodovias em Wenchuan foram danificadas*".
- Extração: (Todas as rodovias em Wenchuan; foram; **danificadas**).

3.2. Regras para extrações inválidas

Regra 1: Extrações cujos componentes fazem parte de orações diferentes são inválidas.
- Sentença (10): "*Portugal explorou o Oceano Atlântico, explorou a costa africana, colonizou áreas selecionadas da África, descobriu uma rota oriental*

para a Índia (...)".
- Extração inválida: (Portugal; colonizou; área selecionadas da África).
- Extração inválida (Portugal; descobriu; uma rota oriental para a Índia)
- Extração válida: (Portugal; explorou o Oceano Atlântico).

Regra 2: A extração é inválida se não há contiguidade entre os elementos internos que foram os argumentos da tripla.
- Sentença (11): "*Sassou venceu a seguinte eleição presidencial em julho de 2009*".
- Extração inválida: (Sassou; **venceu em**; julho de 2009). - Extração válida (1): (Sassou; venceu; a seguinte eleição presidencial).
- Extração válida (2): (Sassou; venceu; a seguinte eleição presidencial em julho de 2009).
- Extração válida (3): (Sassou; venceu a seguinte eleição presidencial em; julho de 2009).

Regra 3: A extração é inválida se o argumento 0 não estiver à esquerda da relação e/ou se o argumento 2 não estiver à direita da relação
- Sentença (12): "*Existem, na Catalunha, Ilhas Baleares e Valência, padrões regionais*".
- Extração inválida: (**padrões regionais**; existem em; a Catalunha, Ilhas Baleares e Valência).

Regra 4: A extração é inválida se algum dos argumentos for apenas um pronome.
Sentença (13): "*Ela lhe deu um relógio de bolso*.
- Extração inválida: (**ela**; lhe deu; um relógio de bolso).

4. Etapa 3: Anotação morfossintática manual dos componentes das triplas extraídas

Após a identificação manual das extrações de triplas, baseada no conjunto de regras definido, o processo de anotação morfossintática foi realizado utilizando etiquetas (tags) de Partes do Discurso (POS) [Alencar et al. 2018].

5. Etapa 4: Validação automática das extrações manuais

As extrações manuais foram feitas por uma linguista júnior e revistas por um linguista sênior, que também sugeriu novas extrações. Ao final do processo, foram obtidas 663 triplas a partir das 360 sentenças selecionadas dentre as 7344 sentenças traduzidas do SQuAD v2. Após isso, as extrações passaram por um algoritmo que busca correspondencias nas respectivas sentenças, sendo que cada elemento da tripla deve ter seus tokens encontrados juntos e seguindo a ordem (arg0, rel, arg1). Das 663 triplas, 427 foram consideradas validas pelo programa.

6. Conclusão e Trabalhos Futuros

O presente artigo apresentou uma metodologia com um conjunto de regras para auxiliar na tarefa de extração de informação aberta automatizada. Como trabalho futuro, este corpus gerado será validado através de um algoritmo de extração de triplas.

Agradecimentos

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

References

Alencar, L. F., Cuconato, B., and Rademaker, A. (2018). Morphobr: An open source large-coverage full-form lexicon for morphological analysis of portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.

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.

Fader, Anthony, e. a. (2011). *Identifying Relations for Open Information Extraction*. Association for Computational Linguistics.

Glauber, R., de Oliveira, L. S., Sena, C. F. L., Claro, D. B., and Souza, M. (2018). Challenges of an annotation task for open information extraction in portuguese. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, pages 66–76, Cham. Springer International Publishing.

Stanovsky, G. and Dagan, I. (2016). Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.