14º Congresso Nacional BAD

COMUNICAÇÃO

ARCANO - No caminho para adicionar semântica às descrições arquivísticas

Luís Filipe Cunha[a], *José Carlos Ramalho*[b]

[a]*Universidade do Minho, Portugal, lfc@di.uminho.pt*
[b]*Universidade do Minho, Portugal, jcr@di.uminho.pt*

Resumo

Nos últimos anos, o uso de técnicas de *Machine Learning* no Processamento de Linguagem Natural tem sido recorrente no processamento de grandes quantidades de documentos não estruturados. No entanto, nem sempre é fácil encontrar recursos de língua portuguesa, nomeadamente *datasets* anotados, que permitam fazer uso destas tecnologias. Neste trabalho é apresentada uma ferramenta de anotação inteligente ARCANO, que faz uso de algoritmos de *Machine Learning* para assistir o processo de anotação de documentos portugueses. Esta ferramenta foi especializada e treinada no domínio das descrições arquivísticas.

Palavras-chave: Descrições Arquivísticas, Machine Learning, Anotação Inteligente.

Introdução

De momento existe uma quantidade colossal de dados nos arquivos portugueses. Por vezes a exploração destes dados pode tornar-se complexa devido à sua dimensão e estrutura. Uma forma de explorar estes dados consiste no reconhecimento de entidades mencionadas (*Named Entity Recognition* - NER) tais como nomes de pessoas, datas, locais, profissões e organizações, permitindo criar índices sobre esses documentos, de modo a facilitar a pesquisa e a navegação nos mesmos.

Um método de se reconhecer entidades nestes documentos consiste na utilização de mecanismos de Processamento de Linguagem Natural (NLP), fazendo uso de algoritmos de *Machine Learning* (ML) que aprendem a identificar e a classificar estas entidades. Nos últimos anos tem-se observado um grande avanço no estado da arte de várias tarefas de NLP, nomeadamente NER, devido à introdução de novos algoritmos de ML, tais como (Vaswani et al., 2017), (Devlin et al., 2019) e (Radford et al., 2018) que levaram à criação de modelos mais robustos, com maior conhecimento da língua.

Apesar disso, para se gerar um modelo de NER capaz de extrair entidades num determinado domínio, é necessário que este tenha acesso a exemplos de anotação de entidades do domínio correspondente, durante o seu treino. *Datasets* anotados em língua portuguesa são muitas vezes escassos, principalmente se o domínio em que se pretender atuar tiver um elevado grau de especificidade na sua linguagem, a título de exemplo, o domínio arquivístico.

Efetivamente, atualmente já existem alguns recursos que podem ser utilizados para o treino de modelos de NER em língua portuguesa, tais como (Freitas et al., 2010) e (Cunha & Ramalho, 2022), no entanto, estes corpora foram criados a partir de textos atuais de domínio jornalístico, literário, político, etc., pelo

https://doi.org/10.48798/congressobad.2905

que os modelos treinados a partir destes acabam por não ser especializados no domínio pretendido. Para se tentar obter melhores resultados, por vezes é necessário gerar os nossos próprios recursos linguísticos de treino. No caso do reconhecimento de entidades, esta tarefa consiste em selecionar um conjunto de *datasets* do domínio em que se está a trabalhar, e anotar manualmente as entidades mencionadas pretendidas. No entanto, a anotação de corpora, para além de complexa, é uma tarefa extremamente demorada e tediosa obrigando muitas vezes a que o anotador seja alguém com conhecimento do domínio.

Com o objetivo de se tentar agilizar o processo de criação de recursos linguísticos portugueses criamos o anotador inteligente ARCANO especializado para o domínio arquivístico. Este anotador faz uso de algoritmos de ML para ajudar o seu utilizador a anotar documentos mais rapidamente, sem que este necessite de ter competências avançadas de informática.

Metodologia

A metodologia usada para o desenvolvimento desta ferramenta de anotação de corpora baseou-se em aprendizagem iterativa de modelos de ML. Durante este processo, o modelo é trainado múltiplas vezes em diferentes subconjuntos do corpus original de modo a melhorar a sua performance a cada iteração do seu treino. Este tipo de mecanismos é normalmente utilizado em ambientes onde não existe uma grande quantidade de dados anotados disponíveis.

O ARCANO permite que um arquivista com conhecimento do domínio linguístico pretendido, consiga partilhar esse conhecimento com um modelo de ML especializando-o na anotação de entidades mencionadas num domínio específico. Sempre que o modelo de ML recebe texto com entidades corretamente anotadas pelo seu utilizador, essa informação é usada para retreinar o seu modelo interno, fazendo com que este aprenda quais as entidades que este deve anotar automaticamente. Desta forma, quantos mais dados anotados forem enviados para o modelo, melhores serão os resultados obtidos por ele.

O ARCANO funciona da seguinte forma:

1. O anotador recebe como input um documento de texto não estruturado. Este documento é dividido em *N* conjuntos de frases (cada conjunto contém 100 frases por omissão);
2. O primeiro conjunto de frases é processado pelo modelo NER genérico o qual retorna o texto anotado com as entidades que conseguiu extrair;
3. O utilizador é responsável por analisar texto processado, corrigindo as anotações efetuadas pelo modelo;
4. As frases corretamente anotadas são usadas para retreinar o modelo de modo que este aprenda a anotar entidades no domínio em questão.

Depois de treinado o modelo, este irá processar o segundo conjunto de frases não estruturadas, no entanto, espera-se que os resultados sejam mais precisos, visto que o modelo aprendeu com as anotações do primeiro conjunto. Quanto maior for a quantidade de texto anotado corretamente, maior será o corpus de treino do modelo, melhorando cada vez mais o seu desempenho.

É esperado que, a partir de um determinado número de anotações manuais, o modelo ganhe um grau de autonomia suficientemente elevado para que a tarefa do utilizador mude da anotação para a correção, agilizando o processo de criação de *datasets* anotados.

Para além disso foi construído com uma interface gráfica focada na sua usabilidade e simplicidade, de modo a poder ser utilizado por utilizadores menos experientes.

Outras Funcionalidades:

1. Permite guardar o estado da ferramenta, de modo a interromper e continuar o processo de anotação mais tarde, voltando sempre ao estado gravado.

2. Permite visualização de estatísticas associadas às entidades anotadas.

3. Oferece opções de parametrização de modo a adaptar-se às preferências do utilizador, por exemplo, alteração de cores e teclas usadas para anotação.

4. Contém um *dataset* pré-carregado permitindo ao utilizador explorar as funcionalidades do ARCANO de imediato;

5. Permite selecionar que conjunto de dados devem ser utilizados no treino do modelo. O utilizador deve escolher as partes do corpus que considere mais representativas do texto a ser anotado.

De modo a facilitar o uso desta ferramenta, para além da sua interface gráfica, o ARCANO conta com uma API que permite que este possa ser utilizado como um serviço, possibilitando a sua integração noutros sistemas.

De momento, o ARCANO encontra-se hospedado nos servidores do Departamento de Informática da Universidade do Minho estando disponível ao público em http://ner.epl.di.uminho.pt/Annotator.

Resultados

Esta ferramenta já foi utilizada para anotar um corpus extraído do repositório *online* do Arquivo Nacional da Torre do Tombo, correspondente a descrições arquivísticas do Arquivo de Oliveira Salazar. Observando a Tabela 1 temos que no total foram anotadas mais de 7000 entidades mencionadas de um corpus com 71 397 *tokens*. Este processo de anotação demorou aproximadamente 8 horas.

Corpus	Pessoa	Local	Data	Profissão	Organização	Total
Arquivo de Oliveira Salazar	2641	1807	279	1414	1258	7399

Tabela 1- Entidades anotadas no corpus Arquivo de Oliveira Salazar

A interface visual do ARCANO foi desenvolvida com o intuito de tornar o processo de anotação de entidades mencionadas simples, intuitivo e rápido. A Figura 1 ilustra a sua interface principal.

Figura 1- Interface do anotador

Nesta figura temos um exemplo de anotação de um excerto de descrições arquivísticas do Arquivo de Oliveira Salazar, referente a uma troca de correspondência de cartas. Este excerto foi anotado automaticamente pelo modelo de ML e depois corrigido pelo utilizador. Para se remover a anotação de entidades de uma determinada palavra basta fazer um clique sobre essa palavra. Para anotar uma palavra como entidade mencionada, basta colocar o rato sobre essa palavra ou selecionar um conjunto de palavras, e de seguida pressionar a tecla correspondente ao tipo de entidade pretendida, por defeito a tecla 1 para anotar Organizações, 2 para Pessoas, 3 para Locais, 4 para Profissões e 5 para Data. Esta configuração de teclas pode ser configurada consoante as preferências do utilizador.

Com a frase corretamente anotada, o utilizador deve decidir se esta deve ser aceite ou rejeitada, determinando se irá ou não ser utilizada para o treino do modelo de ML.

Depois de se anotar um conjunto de 100 frases, o ARCANO permite retreinar o modelo de ML. Na Figura 2 pode-se observar a interface gráfica que permite selecionar os conjuntos de frase que serão usados nesse treino.

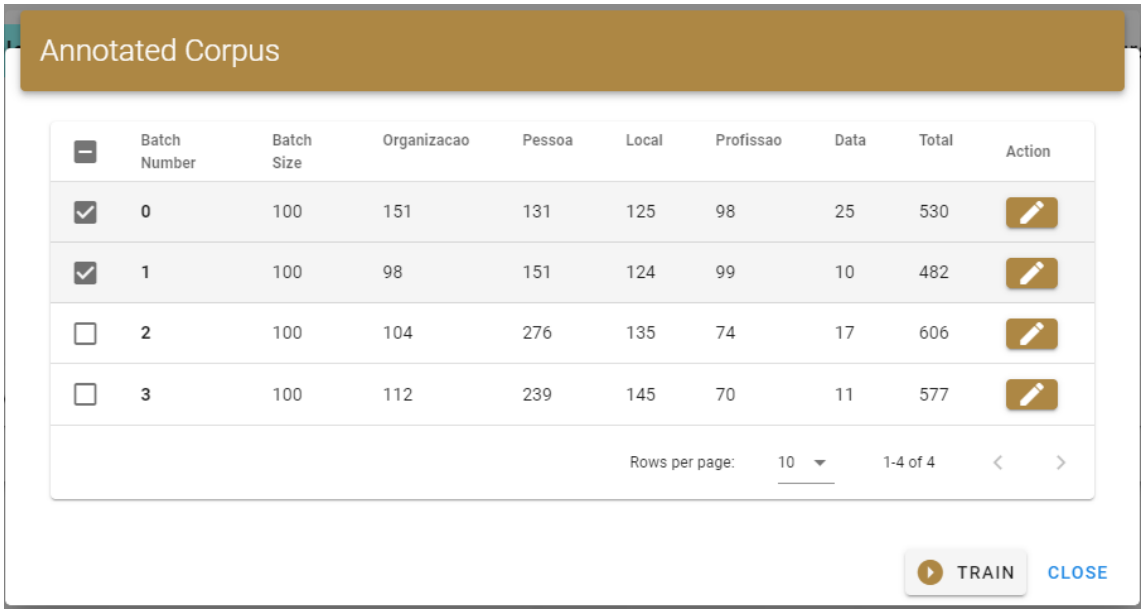

Figura 2 - Interface de treino do modelo ML do ARCANO

Na tabela representada nesta figura, o ARCANO lista os conjuntos de frases já anotados. Nesta vista é possível analisar o número de entidades anotadas de cada tipo de entidade, de cada conjunto de frases. O utilizador deve selecionar os conjuntos de frases que considere mais representativos do contexto que está a anotar, para servirem de exemplo durante a aprendizagem do modelo. A duração do treino do modelo aumenta proporcionalmente em relação à quantidade de dados usados, pelo que à medida que o corpus anotado cresce, limitar os dados usados para treino pode ser uma boa prática, principalmente se o modelo já estiver a reconhecer entidades com elevada eficácia. Por outro lado, pode fazer sentido experimentar diferentes conjuntos de frases selecionadas de modo a explorar combinações que aprimorem a aprendizagem do modelo. Um bom indicador dos conjuntos de frases a selecionar é o número de entidades anotadas de cada tipo. É importante tentar manter o número de entidades de cada tipo balanceado.

Conclusão

Em trabalhos (Cunha & Ramalho, 2022) anteriores surgiu a necessidade de anotar corpora de descrições arquivísticas, um processo lento e tedioso, que se estendeu por vários dias. Com a ferramenta de anotação introduzida neste trabalho, conseguimos criar mecanismos que aceleraram este processo de anotação fazendo uso de algoritmos de ML. Em vez de se anotar todas as entidades manualmente, ensinamos um modelo que irá anotar a maioria das entidades automaticamente, mudando a tarefa do utilizador de anotação para a correção.

Nas experiências realizadas concluímos que o processo de anotação foi notavelmente mais fácil e célere devido à capacidade que o modelo teve de aprender com as entidades anotadas pelo utilizador. Ferramentas deste género são extremamente importantes no contexto linguístico pois assistem o processo de criação de recursos textuais na língua portuguesa.

Como trabalho futuro seria interessante integrar técnicas de *Active Learning* (Lewis et al., 1994) no ARCANO. De momento, a seleção dos dados usados para o treino do modelo é feita de forma sequencial ou pelo utilizador de forma arbitraria. No entanto, para que o modelo aprenda eficientemente, seria interessante definir critérios de seleção destes dados, como por exemplo, selecionar os dados mais

informativos ou que o modelo demonstrasse maior dificuldade a anotar, evitando assim trainar o modelo com informação redundante.

Referências bibliográficas

Cunha, L. F., & Ramalho, J. C. (2022). NER in Archival Finding Aids: Extended. *MDPI*, 1(MAKE), 42–64.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1.

Freitas, C., Mota, C., Santos, D., Oliveira, H. G., & Carvalho, P. (2010). Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, December.

Lewis, D.D., Gale, W.A. (1994). A Sequential Algorithm for Training Text Classifiers. In: Croft, B.W., van Rijsbergen, C.J. (eds*) SIGIR '94*. Springer. https://doi.org/10.1007/978-1-4471-2099-5_1