Proposta recebida em Setembro 2024 e aceite para publicação em Dezembro 2024.

Anotação, análise e aprendizagem de Entidades Nomeadas em textos históricos portugueses (séc. XVIII)

Annotating, analysing and learning named entities in Portuguese historical texts (18th century)

Renata Vieira
Universidade de Évora, CIDEHUS

Fernanda Olival
Universidade de Évora, CIDEHUS

Helena Freire Cameron
Instituto Politécnico de Portalegre, CIDEHUS

Fátima Farrica
Universidade de Évora, CIDEHUS

Joaquim Santos
Universidade de Vale do Rio dos Sinos, PPCA

Daniel De Los Reyes
Universidade de Évora, CIDEHUS

Resumo

Este artigo apresenta um estudo baseado em textos portugueses do século XVIII, através da análise de entidades nomeadas, tendo em vista potenciá-las para análise histórica.

Para isso foi elaborado um *corpus* anotado, a partir de uma fonte (*Memórias Paroquiais*) transcrita, revista e normalizada. Posteriormente, realizou-se uma análise da distribuição das entidades nomeadas na fonte em apreço, para refletir sobre os significados da variação das categorias definidas tendo presente os requisitos dos historiadores. Em seguida, o *corpus* anotado foi usado para desenvolver modelos de Reconhecimento de Entidades Nomeadas (REN) que respeitem a referida complexidade da análise histórica. Foram treinados e avaliados diferentes soluções e modelos de linguagem para a tarefa de REN, onde os melhores modelos atingem $F_1 = 0.70$. Dessa forma, este trabalho demonstra a utilidade do reconhecimento de entidades nomeadas nas análises de textos históricos e disponibiliza um modelo com capacidades de estender as anotações para um conjunto maior de textos com as mesmas características.

Palavras chave

reconhecimento de entidades nomeadas; Século XVIII

Abstract

This article presents a study based on 18th-century Portuguese texts, focusing on the analysis of named entities to enhance their value for historical research. For that, an annotated *corpus* was developed using a primary source (*the Parish Memories*), which was transcribed, revised, and standardised.

The distribution of named entities in the source was then analysed to reflect on the variations in the defined categories, which were established according to historians' requirements. The annotated *corpus* was subsequently employed to develop Named Entity Recognition (NER) models that accommodate the complexity of historical analysis. Several solutions and language models for the NER task were trained and evaluated, where the best models achieve $F_1 = 0.70$. Thus, this work demonstrates the usefulness of named entity recognition in the analysis of historical texts and provides a model with the capabilities to extend annotations to a larger set of texts with the same characteristics.

Keywords

named entities recognition; 18th century

1. Introdução

Este artigo é uma extensão do trabalho apresentado em março de 2024, em Santiago de Compostela, na Conferência Internacional do Processamento Computacional do Português, PROPOR 2014, entitulado *Named entity recognition specialised for Portuguese 18th-century History research* (Santos et al., 2024).

Apresenta-se um estudo realizado numa coleção de textos históricos portugueses chamada *Memórias Paroquiais*, produzida entre 1758–1761, e muito usada por investigadores de diversas áreas. Os originais em apreço têm um valor cultural e histórico significativo. Correspondem às respostas a um inquérito feito pela Coroa, contendo 60 perguntas. Os textos foram redigidos à mão pelos párocos de todo o reino de Portugal, e cada um deles respondia com os múltiplos dados da sua paróquia. Estamos pois perante uma diversidade de grafias que foram meticulosamente transcritas. Para este estudo, o texto foi normalizado para o padrão ortográfico contemporâneo, tendo em vista facilitar a análise (Olival et al., 2023a).

DOI: 10.21814/lm.17.1.445
This work is Licensed under a Creative Commons Attribution 4.0 License

Constituiu-se um *corpus* textual que, pela sua natureza, pelo seu valor linguístico e histórico e pela diversidade de autores, é uma representação plural do português escrito em meados do século XVIII. É uma fonte útil não apenas para historiadores e linguistas, mas também para arquitetos, demógrafos e gestores de território, entre outros. O *corpus* é composto por 71 textos que correspondem às paróquias dos concelhos das principais cidades do Alentejo. O corpus total do Alentejo reúne 420 textos. A coleção das *Memórias Paroquiais* é composta por 4 087 textos de todo o país continental. Note-se que a Região do Alentejo, embora a maior em extensão, é constituída por paróquias/freguesias de grande dimensão territorial, pelo que o número de paróquias desta região é bastante menor face às regiões centro e norte de Portugal. O subcorpus de estudo foi anotado manualmente com recurso à plataforma de anotação, INCEpTION. Os resultados obtidos constituíram um *dataset*.

Em estudos anteriores foram desenvolvidos experimentos envolvendo três categorias principais (PESSOA, LOCAL, ORGANIZAÇÃO) (Vieira et al., 2021). Mais tarde, com o alargamento do número de textos em estudo, surgiram novos desafios, que não conseguiam ser respondidos apenas com as tradicionais categorias de anotação. Definiu-se, assim, uma extensão das mesmas (Cameron et al., 2022), subdividindo-as em classes mais especializadas, que possam traduzir uma etiquetagem mais fiel da realidade da época. Neste artigo discute-se uma análise da distribuição das entidades nas novas categorias e sub-categorias definidas.

O dataset foi utilizado também para construir novos modelos de aprendizagem de máquina. Para tal, avaliaram-se configurações previamente estudadas (Santos et al., 2019) e opções alternativas, com modelos de linguagem mais recentemente disponíveis.

Com este passo, pretende-se aplicar os melhores modelos no futuro imediato, com vista a uma maior automatização do processo e, assim, acelerar a anotação de toda a coleção, mantendo o rigor da análise histórica, usando os modelos por meio de um sistema de anotação semiautomatizado, assistido por análise humana.

Este trabalho aborda: i) a definição de um conjunto especial de categorias de entidades para anotação com base nas características de trabalho dos historiadores (ver seção 3.1), ii) a descrição do processo de anotação manual realizado do subconjunto da coleção das *Memórias Paroquiais*, iii) uma análise da distribuição por categorias resultante da anotação, e iv) a avaliação de modelos de aprendizagem de máquina para a tarefa de anotação dessas categorias.

O objetivo é demonstrar a utilidade das entidades nomeadas na análise de uma fonte histórica e disponibilizar um modelo com capacidade de estender a anotação para um volume maior de texto, possibilitando assim uma análise mais detalhada.

2. Trabalhos relacionados

A tarefa de REN (Reconhecimento de Entidades Nomeadas) é uma tarefa usual do Processamento de Linguagem Natural, existindo trabalhos dedicados à língua portuguesa neste âmbito de estudo e aplicação. No entanto, é mais comum encontrar estudos relacionados com o português contemporâneo. Vejam-se, a título de exemplo, estudos recentes sobre REN para o português atual, por Albuquerque et al. (2023) e Silva (2023).

No entanto, o reconhecimento de entidades nomeadas para a investigação do domínio da história está a ganhar algum relevo, com o avanço das humanidades digitais. Em Ehrmann et al. (2023), apresenta-se um estudo sobre reconhecimento e classificação de entidades nomeadas em documentos históricos considerando uma variedade de línguas. Entre outros estudos, veja-se também o trabalho relativo ao Corpus Histórico do Português, BDCamões (Grilo et al., 2020). Este *corpus* foi anotado automaticamente com ferramentas de processamento de linguagem natural e inclui as categorias usuais de EN. Contudo, não há ainda uma avaliação da precisão da anotação realizada de forma automática, o que não nos permite verificar a qualidade dessa anotação.

Registem-se, ainda, outros estudos que consideram o português histórico, apresentados em: Grilo et al. (2020); Aguilar et al. (2017); Zilio et al. (2022). Estes trabalhos, porém, não abrangem uma maior especificação de categorias conforme apresentado aqui neste presente estudo, onde descreve-se um *corpus* do século XVIII, anotado com categorias e subcategorias de elevado valor para o fazer História (ver seção 3.1), e apresenta-se uma análise dos resultados bem como uma avaliação da precisão dos modelos treinados para a sua classificação automática.

3. Anotação das *Memórias Paroquiais*

As *Memórias Paroquiais* são o resultado de um inquérito solicitado no começo de 1758, com dois objetivos principais: 1) obter informações sobre

o estado do território após o grande terramoto de 1755; 2) reunir elementos para criar um Dicionário Geográfico de Portugal, dando continuidade ao trabalho do padre oratoriano Luís Cardoso (Pernes, 1697- Lisboa, 1769).

Atualmente, no site do Arquivo Nacional Português da Torre do Tombo (ANTT), os manuscritos das *Memórias Paroquiais* estão disponíveis online, como cópias digitalizadas de microfilmes. Neste estudo, considerou-se um subconjunto textual, com as respostas dos párocos oriundos de paróquias das principais cidades do Alentejo. Os originais foram transcritos manualmente, dada a complexidade da diversidade das grafias bem como o mau estado de alguns documentos. Uma primeira versão foi disponibilizada no CIDEHUS Digital[1], com a ortografia do século XVIII e com algumas intervenções dos transcritores. Posteriormente, como parte do processo para produzir uma versão anotada com entidades nomeadas, os textos foram normalizados manualmente para a ortografia do século XXI. As alterações efetuadas foram-no apenas ao nível gráfico; no plano vocabular, mantiveram-se palavras que, embora desusadas, estão registadas nos dicionários e são ainda usadas em contextos determinados, como por exemplo: "el-rei", "mui" e "cousa". Conservou-se toda a variação lexical, como a alternância ou/oi e outras. Normalizaram-se as maiúsculas e foram atualizados na grafia os nomes geográficos, sempre que possível. Uma das mais frequentes intervenções foi ao nível da normalização da representação das sibilantes que, à época, eram grafadas de forma não sistemática. Também foi normalizado o registo dos ditongos nasais em posição final da palavra: *am*, *aõ* ¿ *ão* e *oens*, *õens* ¿ *ões*.

3.1. Sub-categorização das entidades nomeadas para estudos em História

O processo de anotação seguido visou traduzir e respeitar a complexidade de eras passadas, expressas nas fontes históricas, pois elas diferem das contemporâneas. Consideraram-se cinco categorias principais: PESSOA, LUGAR, ORGANIZAÇÃO, TEMPO e OBRA DE AUTOR. As quatro primeiras procuram responder a questões históricas fundamentais: quem?, onde?, o quê?, quando?, e a última permite-nos tratar as fontes mencionadas no *corpus*. Contemplar este último tópico era essencial para revelar a arqueologia do próprio texto: em que fontes se escorou o discurso dos párocos? No entanto, é nas especializações das referidas categorias que o processo de anotação arquitetado se diferencia.

As entradas principais foram divididas em várias subcategorias, para melhor captar a especificidade das sociedades outras do passado.

A categoria pessoa (PER) considera referências por nome, categoria social ou ocupação. Além disso, foram definidas subcategorias específicas para menções de santos, divindades, grupos de pessoas (coletivos com identidade) e autores citados pelos párocos. Incluir a categoria social é um dado essencial para a época em estudo. No século XVIII, e ao contrário de hoje, a desigualdade estava inscrita na lei, e pautava não só as formas de tratamento, como os mais ínfimos aspetos do quotidiano. Ao privilégio que marcava, que favorecia uns em detrimento de outros, somava-se a hierarquia. Por isso, captar estas particularidades é essencial em estudos sobre realidades pretéritas. Também frequentemente, títulos e posições ocupacionais eram quase parte do nome e da identidade de uma pessoa. Um exemplo de categoria social anotada no *corpus* é *Conde de Portalegre* ou *fidalgo da Casa Real*.

Relativamente à ocupação, vejam-se exemplos de menções a pessoas através da ocupação:

- Arcebispo de Évora
- Presidente da Mesa da Consciência

A subcategoria grupos de pessoas foi usada ainda para anotar grupos orgânicos, parentelas e membros de uma organização, entre outros, como nos exemplos seguintes:

- Jesuítas [os Jesuítas]
- Sequeiras [a parentela ou família Sequeira]
- Almas [as Almas do Purgatório]
- Mouros [os Mouros]

Em relação aos indicadores de lugar, generalizou-se localização (LOC) para lugar (PLC), por ser mais abrangente. Esta categoria inclui entidades geopolíticas, aquíferos, montanhas, instalações e uma subcategoria extra para outros locais que não conseguem ser categorizados nas sub-etiquetas estabelecidas:

- termo da vila de Monsaraz
- outeiro de São Bento
- chafarizes dos Leões
- Palácio da Inquisição

A categoria ORG inclui todas as tipologias de organizações, como, por exemplo:

- Convento de Santo António
- Inquisição de Évora
- Confraria de São Pedro

As organizações tinham diferentes papéis sociais e a diferença entre uma localização e uma organização podia ser ténue. Assim, utilizou-se sempre a entidade geopolítica, mais abrangente, e não apenas os lugares, como já foi aclarado.

[1] http://www.cidehusdigital.uevora.pt

Etiquetaram-se todas as referências a pontos geográficos, como rios e montanhas, nesta fase apenas como rótulos, mas já essenciais para criar georeferenciação, tendo na mira produzir cartografia histórica.

Para a categoria Tempo, anotaram-se apenas referências específicas a datas, por exemplo, o ano de 1755.

Estas foram algumas das razões que apoiaram a necessidade de repensar as ENs para descrever melhor os elementos da fonte e tornar o processo de anotação mais relevante do ponto de vista da História e de outras áreas do conhecimento. No entanto, esta é uma questão desafiadora. O refinar da categorias frequentemente implica mais complexidade na anotação e nos processos computacionais, o que foi assumido pela equipa de investigação logo desde o início. A categoria TEMPO, por exemplo, pode vir a ser alargada para contemplar outros marcadores temporais, como tempo religioso, tempo dinástico, etc.

3.2. Diretrizes de anotação

Como é habitual neste tipo de estudos, foram previamente definidas diretrizes de anotação antes do processo de anotação manual se iniciar, e estas foram revistas ao longo do mesmo, capitalizando a experiência para produzir melhorias.

Deste modo, a construção das *guidelines* foi uma fase vital, pois havia vários anotadores, e todos deviam ter o mesmo suporte a sustentar a decisão. Para todas as categorias e subcategorias geraram-se exemplos de diferentes textos do *corpus*, esclarecendo situações complexas. Tudo isto exigiu tempo e trabalho de uma equipa multidisciplinar a atuar em simultâneo. Seguiram-se alguns princípios norteadores do trabalho para respeitar ao máximo o contexto. Assim, a delimitação deve abarcar a totalidade da expressão, incluindo informações sequenciais adicionais, como o aposto. Esta pode ser uma boa forma de enfrentar a desambiguação da entidade. Vejam-se dois exemplos:

- Morgado Francisco José Cordovil - onde "Morgado" não faz parte do nome, mas é uma identificação adicional que permite destrinçar entre homónimos, e a homonímia era frequente no século XVIII.
- Dom Frei João de Azevedo bispo - neste caso manteve-se Dom e Frei pois são uma menção do estatuto, e ambos fazem parte do nome.
- Francisco José Cordovil, natural de Évora - aqui incluiu-se a informação adicional "natural de Évora" [nascido em Évora], para facilitar a correta identificação.

Nestas diretrizes gerais, também foi assumido que apenas ENs que incluem nomes próprios deviam ser anotadas. Por exemplo, anotar-se-á a expressão "cabido da Sé de Évora", mas não os usos isolados da palavra "cabido". Igualmente será marcada a denominação da organização "Santa Casa da Misericórdia de Beja", mas não apenas o nome geral "misericórdia". Assim, etiquetar EN não é o mesmo que produzir um índice das palavras presentes no texto.

3.3. Processo de anotação

Após a fase de normalização gráfica dos textos, a anotação manual iniciou-se, primeiramente conduzida como um processo consensual, com quatro anotadores compartilhando sincronicamente o ecrã e decidindo o que anotar. A equipa de anotadores era composta por uma linguista, uma historiadora, uma paleógrafa e uma cientista da computação. Durante esse processo, as diretrizes foram revistas sempre que necessário.

Após essa fase inicial de definição de critérios e construção de uma anotação consensual, uma historiadora e paleógrafa prosseguiu com a tarefa, trazendo dúvidas para a equipa. A discussão sobre casos ambíguos permitiu melhorias significativas, com incorporação de maior especificação nas *guidelines* e, até, pontualmente, de revisão de anotação já efetuada. Este trabalho revelou-se fundamental, pois permitiu não só uma revisão dos dados, garantindo a sua fiabilidade, como também contribuiu para uma construção progressiva dos *guidelines*, suportados em exemplos reais do texto. Importa criar regras, com exemplos, para facilitar o trabalho em equipa e a harmonia entre os vários intervenientes. A ferramenta de anotação usada foi a plataforma INCEpTION.[2]

4. Análise de distribuição

4.1. Descrição do *corpus* anotado

O subconjunto anotado reúne 71 paróquias do Alentejo, correspondendo a 17% do total das paróquias desta região. No entanto, qualitativamente, estas pertencem aos concelhos mais importantes: Beja, Évora, Portalegre, e também Vila Viçosa. Os três primeiros são capitais de distrito, que do século XIX chegaram aos dias de hoje. Até cerca de 1640, Évora era a segunda cidade do Reino em termos políticos. Vila Viçosa, no passado, foi sede do Ducado de Bragança, a maior casa senhorial do país no início do século XVII.

[2] https://inception-project.github.io

Como expresso na Tabela 1, e Figura 1, relativamente às categorias anotadas, a distribuição das frequências é muito desigual. As principais categorias representadas no *corpus* são as entidades geopolíticas, seguidas dos nomes de pessoas e de santos. A categoria social e as montanhas são as que têm menor número de expressões anotadas. Observe-se que, para o processo de aprendizagem, descrito na sequência, elas tiveram que ser separadas para treino, desenvolvimento e teste, considerando aproximadamente uma distribuição de 70, 10 e 20%.

CATEG	# Casos	Tipo
AUTWORK	137	Obra de autor
ORG	393	Organização
PER_AUT	129	Autor
PER_CAT	49	Categoria social
PER_DIV	184	Divindade
PER_NAM	718	Nome de pessoa
PER_OCC	124	Ocupação
PER_PGRP	199	Grupo de pessoas
PER_SAINT	644	Santo
PLC_AQU	228	Aquífero
PLC_FAC	289	Construção
PLC_GPE	1101	Entidade Geo Política
PLC_LOC	447	Lugares em geral
PLC_MOUNT	73	Montanhas
TIM_CRON	316	Tempo cronológico (data)
Total	5031	Todos

Tabela 1: Distribuição das Entidades Nomeadas

A partir dos arquivos de saída gerados no INCEpTION, constituiu-se um dataset com as informações resultantes da anotação, quer por categoria, quer com todas as categorias. Mantiveram-se, igualmente, os dados individualizados por paróquia, agrupados por concelho, uma organização que permitiu analisar as entidades em cada paróquia e por concelho.

Criou-se um *parser*, que se disponibilizou à comunidade científica,[3] para simplificar a extração, análise e organização de entidades nomeadas dos ficheiros de saída do INCEpTION. Após aplicar o analisador, exploram-se os resultados da anotação e categorização em geral e também por concelho. A Tabela 3 mostra a distribuição nos quatro concelhos. Obtiveram-se 5031 NEs anotadas como resultado da anotação manual.

4.2. Distribuição das entidades nomeadas por categoria

Entender a distribuição geral de categorias de entidades nomeadas é essencial para contextualizar o estudo feito e o que se continua a desenvolver. Na Tabela 2 pode observar-se a predominância das categorias PLC e PER (lugares e pessoas) em comparação com as outras, totalizando mais de 80% das entidades nomeadas observadas nesses textos.

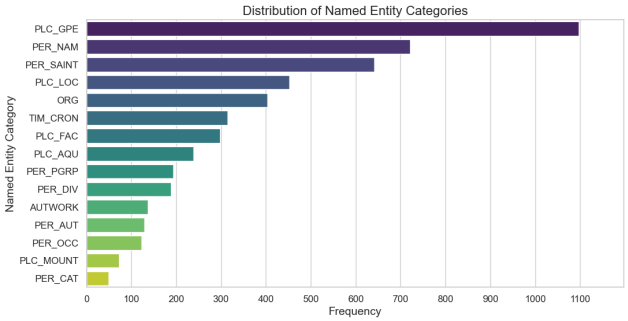

Figura 1: Distribuição das categorias de ENs

Categoria	Distribuição (%)
PLC	43
PER	40
ORG	8
TIM	6
AUTWORK	3

Tabela 2: Distribuição das ENs por categoria

Na Figura 1 pode ver-se que, dentro da categoria PLC, a subcategoria GPE (entidade geopolítica) é a que tem maior número de ocorrências. A segunda subcategoria mais abundante neste corpus é PER_NAM (nome de pessoa). A terceira subcategoria com maior número de expressões anotadas é relativa aos Santos. A categoria social foi a subcategoria que registou menor número de expressões anotadas. Esta menor frequência não significa, todavia, que os indicadores desta natureza estivessem quase ausentes nos textos. Dado que a equipa anotou sempre a maior expressão, muitos marcadores de estatuto social foram incluídos na subcategoria PER_NAM, facto que explica esta menor incidência. A segunda subcategoria com menor número de expressões anotadas é PLC_MOUNT. A região do Alentejo é essencialmente de planície, com algumas serras. A menor quantidade de expressões anotadas nesta categoria traduz, neste caso, a efetiva realidade orográfica local.

A Figura 2 mostra as 15 entidades nomeadas mais referenciadas em todos os textos de todos os concelhos analisados, oferecendo uma perspetiva sobre os elementos mais recorrentes e importantes.

[3] https://github.com/DanielReeyes/inception-entity-parser

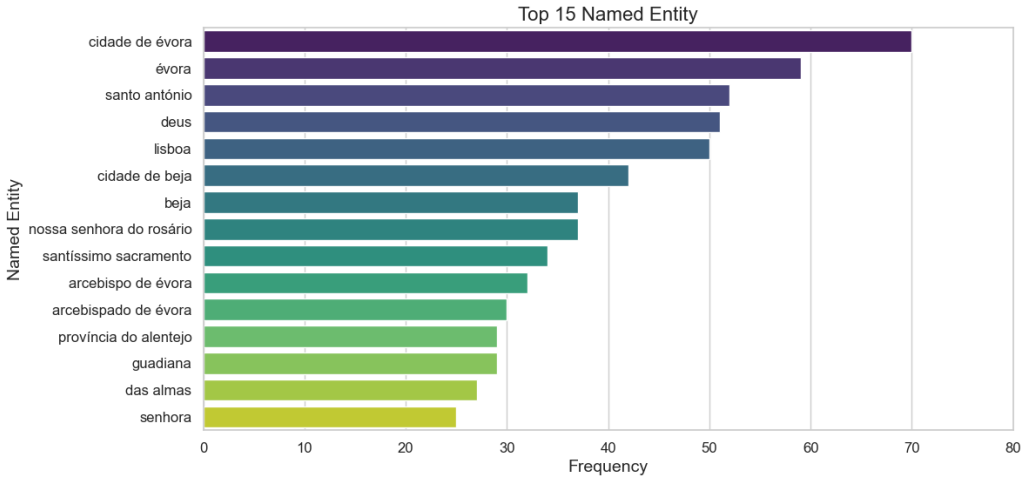

Figura 2: Top 15 ENs

No topo das mais mencionadas está a cidade de Évora, sede do arcebispado do mesmo nome, a que se junta o vocábulo Évora isolado. Especificamente, Évora, Lisboa e Beja foram as entidades nomeadas mais frequentemente mencionadas na categoria GPE, no geral. Elas foram seguidas por entidades nomeadas nas categorias de santos e divindade, como Santo António, Deus e Nossa Senhora do Rosário, uma devoção amplamente difundida em Portugal após a Contra-Reforma. Ainda relativamente à categoria GPE, Lisboa também recebeu menções significativas, embora não pertença ao Alentejo. Esta maior frequência explica-se pelo facto de uma questão colocada no inquérito para todas as paróquias do Reino ter sido: "Quão longe está a paróquia da cidade capital episcopal, e quão longe está de Lisboa, a capital do Reino?". Como resultado, Lisboa foi consistentemente citada em cada uma das paróquias, pois a pergunta induzia o resultado. O mesmo se passou com Évora, em muitos locais, por ser sede arquidiocesana e, como tal, referida amiúde.

A alusão a Beja, com elevando número de ocorrências, foi considerada um pouco surpreendente, já que, naquela época, esta cidade não era capital episcopal, mas o concelho tinha muitas freguesias (29).

É ainda de assinalar o grande peso da expressão "Província do Alentejo", uma vez que aparece entre as designações do top 15, não só no concelho de Évora, mas também em Beja e Portalegre. Tudo isto constituía um indicador de que a identidade por província já estava presente nesta altura na zona sul de Portugal, embora a primeira pergunta do interrogatório também condicionasse o resultado, ao querer saber em que província ficava a terra.

4.3. Distribuição de entidades nomeadas por concelho

Concelhos	Paróquias	ENs
Beja	29	1895
Évora	22	1807
Portalegre	14	855
Vila Viçosa	6	474
Total	71	5031

Tabela 3: Distribuição das ENs por concelho

Agrupando-se as paróquias por concelhos a que pertencem, podem analisar-se as variações temáticas entre concelhos, expressas na Figura 3.

A análise do gráfico revela que todos os concelhos seguem o mesmo padrão de distribuição de subcategorias do contexto global. Todos têm uma prevalência das categorias e subcategorias PLC_GPE. As outras duas subcategorias destacadas na análise geral, PER_NAM e PER_SAINT, também estão presentes quando analisamos os dados por concelho. Mesmo Vila Viçosa, de menor dimensão, segue esta tendência.

A identificação das principais entidades nomeadas em cada concelho é crucial para obter uma visão mais específica das particularidades regionais. A Figura 4, representando as 15 principais ENs por concelho, fornece uma visualização que destaca as entidades mais proeminentes em cada local. Deste modo, é possível observar a importância de Évora em todos os concelhos, bem como a referência a divindades e santos mencionados em todas as regiões. Citem-se: Deus é uma entidade referida em todos os concelhos; Santo António é anotado em Beja, Évora e Portalegre. Também São Pedro neste último concelho. Estes dois santos masculinos representam dois dos três

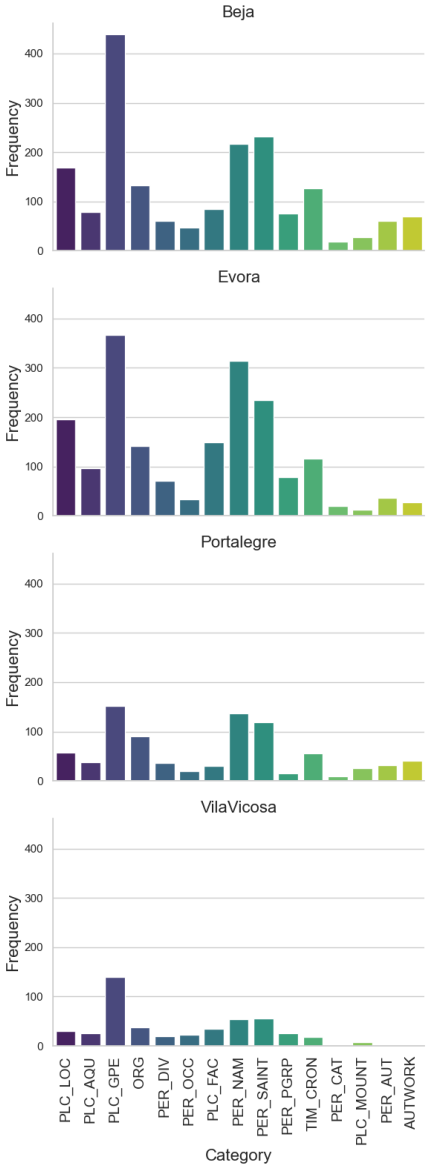

Figura 3: Distribuição de subcategorias por concelho

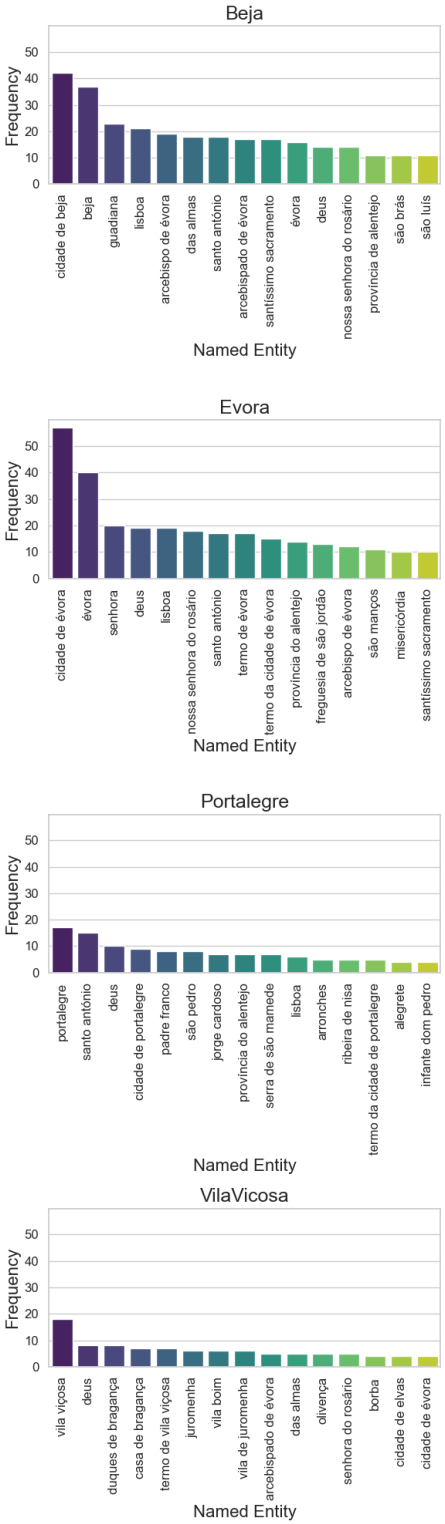

Figura 4: Top 15 ENs por concelho

mais venerados em Portugal. Santo António, que se acredita ter nascido em Lisboa, seria muito invocado desde o século XVII para localizar objetos perdidos; por sua vez, São Pedro era o guardião das chaves do céu, o santo padroeiro da Igreja e do papado, simbolizando a reafirmação da Igreja Católica após a Contra-Reforma (Farmer, 1997). Era em Beja, uma terra de muitos cristãos-novos, que mais elementos do panteão religioso figuravam nas 15 entidades com mais relevo (7, incluindo as Almas, ao passo que em Évora eram 4, e 3 nos outros concelhos).

Para além dos dados apresentados no quadro que se acaba de referir, a Casa de Bragança foi copiosamente mencionada em Vila Viçosa. Esta Casa representou um elemento-chave da identidade local, devido aos seus laços diretos com a realeza e ao seu patrocínio no estabelecimento de conventos, capelas e outras instalações comunitárias. Foi consistentemente referenciada de forma positiva e apenas numa das paróquias houve menção à obrigação da população de pagar impostos quando construísse moinhos e outros engenhos nos aquíferos locais (Olival et al., 2023b).

Em Portalegre aparecem 3 nomes nas 15 entidades mais citadas. São eles: o Jesuíta António Franco (1662–1732), como autor de obras que inventariaram jesuítas insígnes; Jorge Cardoso (1606–1669), na qualidade de autor do *Agiológio Lusitano*; o Infante D. Pedro (1717–1786), por ser prior do Crato e dispor de vários direitos e jurisdições naquele concelho. Na realidade, nem este último era uma entidade plenamente laica. O peso da esfera religiosa era grande nas *Memórias Paroquiais*, tal como acontecia na sociedade portuguesa de então.

5. Construção de modelos para anotação automática

A anotação realizada possibilitou uma análise de apenas uma parte desta fonte histórica. Contudo, o trabalho é alargável a todas as regiões de Portugal, expandindo e alargando a anotação. Por ser uma tarefa custosa, e aproveitando o fato de haver um subconjunto de anotação manual já desenvolvido, pretende-se avançar no processo com o auxílio computacional. Com o subcorpus anotado manualmente e revisto por historiadores e linguistas, passou-se a uma nova fase, de treino e avaliação de modelos de anotação, conforme exposto nas seguintes secções.

5.1. Arquiteturas para aprendizagem

Os experimentos neste trabalho basearam-se na biblioteca Flair(Akbik et al., 2019) de REN desenvolvida em PyTorch[4] para muitos idiomas. Ela oferece modelos de linguagem pré-treinados, modelos de reconhecimento de entidades nomeadas e redes neurais para treino de etiquetação sequencial. Utilizando o modelo Flair, podem construir-se *pipelines* de treino de classificadores de *tokens* e alimentá-los com modelos de linguagem de vários tipos, como *Word Embeddings*, modelos baseados em *Transformer* e o próprio *Flair Embeddings*.

Flair é um framework de fácil uso, de baixo consumo de GPU e suporta inúmeros embeddings, facilitando a proposta do artigo de avaliar mais do que um modelo de linguagem. Note-se, ainda, que o Flair é um forte concorrente com o atual estado-da-arte[5] em REN, distando apenas 0,21% de diferença na avaliação CoNLL.[6]

Em relação ao uso de modelos de linguagem, o artigo seminal do *Flair Embeddings*(Akbik et al., 2018) mostra que a sua combinação para REN é benéfica. No *framework* Flair, a ferramenta Stacking Embeddings permite combinar diferentes tipos de modelos de linguagem: modelos baseados em *transformers*; *flair embeddings*; e *shallow WE*. Dessa forma, cada palavra é representada pela concatenação dos vetores fornecidos por cada modelo de linguagem que foram carregados para o Stacking Embeddings.

Nos experimentos usou-se, ainda, a arquitetura LSTM-CRF, composta, basicamente, por dois componentes: a estrutura neural Long-short Term Memory (LSTM)(Hochreiter & Schmidhuber, 1997) e um classificador Conditional Random Fields (CRF)(Lafferty et al., 2001). Primeiramente, uma camada recebe o *Stacked Embeddings* e depois converte os tokens de entrada em vetores enriquecidos de contexto. Então, os vetores são alimentados à LSTM, que aprende os padrões de anotação e, por fim, o classificador CRF recebe as saídas e retorna a sequência de rótulos.

A estrutura neural Transformer-Linear é composta pelo modelo de linguagem baseado em *transformers*, onde é adicionada uma camada final (linear) que retorna a sequência de rótulos. Essa estratégia é a mesma aplicada no artigo seminal sobre o BERT (Devlin et al., 2019). Essa forma de fazer *fine-tuning* também está disponível no framework Flair e foi integrada ao Flair como Flert(Schweter & Akbik, 2020). Assim, também se usou o *Flair* para treinar o modelo com o *Flert*.

Uma forma menos explorada de se fazer etiquetagem de sequências é usar algoritmos *text-to-text*. Esses algoritmos recebem texto (como entrada) e devolvem texto (como saída). São também conhecidos como algoritmos de sequência-a-sequência (Seq2Seq). O uso de modelos Seq2Seq pode ser vantajoso porque permite formular a tarefa como uma transformação de sequência para sequência, onde a entrada é o texto original e a saída é uma sequência estruturada contendo as entidades identificadas e suas respetivas categorias. Essa abordagem pode capturar dependências de longo alcance e aproveitar melhor o contexto global da frase, o que é especialmente útil em línguas morfologicamente ricas como o Português. Em REN isso significa que uma palavra pode depender de algo que foi dito antes ou depois no texto. Estes modelos Seq2Seq podem lidar bem com este tipo de rotulação devido ao seu mecanismo de atenção. Neste sentido, usou-se o *framework* HappyTransformer[7] para realizar e treinar o nosso modelo e o Seq2Seq para reconhecer entidades nomeadas.

[4] https://pytorch.org/
[5] https://aclanthology.org/2020.emnlp-main.523/
[6] https://conll.org
[7] https://happytransformer.com

Neste trabalho, foram usados os modelos de linguagem: Shallow Word Embeddings e Contextual Embeddings. Usou-se ainda os grandes modelos de linguagem LlaMa2 e mT5. Os modelos são apresentados em seguida com as suas configurações.

5.2. Modelos de Linguagem

O uso de WE na tarefa de REN vem desde o advento destes modelos de linguagem e é amplamente usado com redes neurais recorrentes. Neste trabalho, foram utilizados dois tipos de modelos Word Embeddings pré-treinados: Word2Vec(Mikolov et al., 2013) (Skip-gram) e Glove(Pennington et al., 2014), ambos com 300 dimensões. Os modelos foram obtidos no repositório de embeddings do NILC.[8]

Como modelo de linguagem do tipo *Flair Embeddings*, foram usados os modelos *FlairBBP*[9] treinado por um dos autores, conforme Santos et al. (2019). Os autores treinaram o modelo com cerca de quatro milhões de tokens. Os Flair Embeddings são treinados através de uma BiLSTM, onde o modelo é treinado para prever o próximo caracter de uma sequência de *tokens*. Cada modelo *Flair Embeddings* é composto por dois ficheiros: um modelo *forward* e outro *backward*, numa operação linear que combina os dois modelos e fornece uma representação para cada palavra, em que a representação é sensível ao contexto. Isso faz com que esse tipo de modelo seja um *contextual embedding*, ou seja, as representações mudam de acordo com o contexto. Esse tipo de embedding diferencia-se dos WE, uma vez que são vetores fixos, onde a representação da palavra é sempre a mesma, independente do contexto. Escolheu-se experimentar modelos *Flair Embeddings*, sobretudo por terem versões exclusivas para o português e também por apresentarem uma reduzida complexidade na sua implementação, diminuindo o tempo investido nesta tarefa.

O XLM-RoBERTa(Conneau et al., 2020) (XLM-R) é um LLM multilingue do tipo RoBERTa. Foi pré-treinado num *corpus* de 2,5 TB de dados com cem idiomas. Do total de dados de treino, 49.1 GB foram de dados em português, correspondendo a aproximadamente 8.4 bilhões de tokens. Dado que os dados deste nosso estudo estão em português, e face ao treino já efetuado previamente no modelo, usou-se essa versão multilingue do RoBERTa.

Pode descrever-se o XLM-R caracterizando, primeiramente, o modelo original RoBERTa. Este é baseado em *transformers* e pré-treinado num grande *corpus* de forma não-supervisionada. O RoBERTa herda do BERT a estratégia de treino por máscara, ou seja, o objetivo do modelo durante o treino é prever os *tokens* mascarados de uma frase. Assim, durante a fase de treino, 15% dos *tokens* de entrada foram mascarados para se sujeitarem à predição.

Neste estudo, utilizou-se a versão Large do XLM-R, disponível no repositório HuggingFace.[10] Usou-se esse tipo de modelo por ser um modelo extremamente competitivo com o estado-arte atual no REN em inglês.

BERTimbau(Souza et al., 2020) é um modelo de linguagem pré-treinado estilo BERT treinado para o português. Esse modelo foi treinado no *corpus brWaC*(Filho et al., 2018), somando um total de 2.6B de *tokens* que, depois de pré-processado, totalizou 17.5GB. Usou-se a versão Large do BERTimbau, disponível no HuggingFace.[11]

BERTimbau é um modelo baseado em transformers e também foi treinado com máscara de tokens nas frases de entrada, sendo este também um Masked Language Model (MLM). Escolheu-se esse modelo, uma vez que o atual estado-da-arte (Souza et al., 2019) em REN para o português usa esse modelo.

5.3. Grandes Modelos de Linguagem

Realizaram-se alguns experimentos com base em grandes modelos de linguagem (LLMs). Foram usadas duas versões do LLaMa2 (Touvron et al., 2023) através do HuggingFace: a versão original[12] (disponibilizada pela Meta) e uma versão treinada pela NousResearch.[13] Em ambos os casos, neste estudo usou-se a versão *chat* com 7B de parâmetros. Os modelos pré-treinados do Llama 2 foram treinados em 2 trilhões de *tokens*. Os modelos foram ajustados (*fine-tuned*) em mais de 1 milhão de anotações humanas.

O treino do Llama 2-Chat começa com o pré-treino usando arquitetura *transformer* em fontes de dados *online* disponíveis publicamente. Em seguida, é feito um ajuste fino supervisionado para criar uma versão inicial do LLaMa 2-chat.

[8] http://nilc.icmc.usp.br/nilc/index.php
[9] https://github.com/jneto04/ner-pt
[10] https://huggingface.co/xlm-roberta-large
[11] https://huggingface.co/neuralmind/bert-large-portuguese-cased
[12] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[13] https://huggingface.co/NousResearch/Llama-2-7b-chat-hf

Por fim, inicia-se então a fase de refinamento por um processo interativo utilizando metodologias de Aprendizagem por Reforço com Feedback Humano (RLHF).

O modelo mT5 é multilingue e expande o modelo T5 original, utilizando o *corpus* mC4 (Multilingual Common Crawl)[14], que abrange 101 idiomas, permitindo um treino em larga escala com dados diversificados. O modelo adota o paradigma *text-to-text*, em que tanto a entrada quanto a saída são representadas como sequências textuais, proporcionando uma abordagem unificada para diferentes tarefas, como tradução automática, sumarização e resposta a perguntas. O treino do mT5 utiliza a tarefa de *span corruption*, que consiste em mascarar trechos de texto para que o modelo aprenda a reconstruí-los, promovendo uma aprendizagem eficaz de representações linguísticas.

O uso de LLMs, hoje na ordem do dia, tem muitas vantagens, conforme publicado na literatura quase diariamente. Contudo, esses modelos têm também desvantagens. Uma das principais tem a ver com o poder computacional requerido para o seu uso, seja para inferência ou *fine-tuning*. É nesse sentido que, neste estudo, foram usadas técnicas de redução de parâmetros e redução de precisão dos pesos do modelo, adiante descritas. Essas duas técnicas foram utilizadas apenas nos dois modelos LLaMa avaliados neste trabalho.

A técnica da quantização oriunda do domínio da estatística, pode ser brevemente descrita como o processo de mapear valores contínuos infinitos num conjunto discreto finito. No contexto de LLMs, a redução acontece na precisão dos pesos que, no caso do LLaMa2, são, originalmente, 32 bits. Neste sentido, converteu-se o nosso modelo para uma precisão de 8 bits, usando a biblioteca bitsandbytes (Dettmers et al., 2022).

Depois da quantização, foi feito um ajuste (*fine-tuning*) de forma eficiente usando o PEFT-LoRA(Mangrulkar et al., 2022; Hu et al., 2022), onde se mostrou que congelar pesos do modelo e reduzir a complexidade das matrizes das camadas de *transformer* conduz a uma redução significativa do número de parâmetros e pode ainda apresentar resultados iguais ou superiores ao modelo original. Em outras palavras, PEFT-LoRA reduz o número de parâmetros treináveis durante o *fine-tuning*. Neste estudo considerou-se *rank* $r = 64$ e $\alpha = 16$.

O algoritmo Needleman–Wunsch(Needleman & Wunsch, 1970) é um algoritmo de programação dinâmica que tem por objetivo realizar o alinhamento de duas sequências. Esse algoritmo é frequentemente utilizado para alinhar sequências de proteínas ou nucleotídeos. Neste trabalho, retomou-se esse algoritmo, usando-o para alinhar os textos rotulados pelos modelos LLaMa e mT5 ao texto padrão ouro, possibilitando a extração das métricas de avaliação. Usou-se a versão implementada pela Genalog[15] em Python.

6. Experimentos

6.1. Configuração

Os experimentos realizados podem ser divididos em dois conjuntos: (*i*) Experimentos com empilhamento de *embeddings*, e (*ii*) Experimentos com LLMs.

No que respeita ao conjunto de experimentos (*i*), foi usada o modelo Vanilla LSTM-CRF implementado no Flair. Desse modo, montaram-se dois empilhamentos de *embeddings*: FlairBBP + Word2Vec (Skip-gram), que doravante será chamado FlairBBP+W2V-SKPG e FlairBBP + Glove. Foi feita a combinação desses *embeddings*, uma vez que Santos et al. (2019) mostrou que combinar o FlairBBP com o Word2Vec (skip-gram) resultava num melhor empilhamento de *embeddings* para o reconhecimento de entidades nomeadas no *corpus* HAREM(Santos & Cardoso, 2007). No trabalho original sobre o Flair, os autores empilham um modelo Flair Embeddings com um modelo de linguagem Glove. Tal experimento não foi feito por Santos et al. (2019). Assim, decidiu-se avaliar tal empilhamento de *embeddings*.

Para os experimentos feitos com o XLM-R e o BERTimbau, usou-se o Flert, onde a etiquetação sequencial é composta pelo próprio modelo mais uma camada final linear, que retorna a sequência de rótulos. Chama-se a esses experimentos *Transformer-Linear*, pois ambos os modelos avaliados são baseados em transformers. Usaram-se os hiperparâmetros padrão[16] e executaram-se esses experimentos numa GPU RTX 4090 24GB.

Para o conjunto de experimentos (*ii*), avaliaram-se dois LLMs: LLaMa 2 e mT5. Sobre os experimentos com o LLaMa 2, foi feito um *instruct-tuning*, onde o *prompt* é composto pela instrução, *input* e *response*. A Figura 5 mostra um exemplo de *prompt*.

[14] https://arxiv.org/abs/2010.11934

[15] https://microsoft.github.io/genalog/text_alignment.html

[16] Model_max_length = 512, hidden_size = 256, learning_rate = 5.0e-6, mini_batch_size = 4, mini_batch_chunk_size = 1, max_epochs = 20.

Instrução: "Reconheça as entidades nomeadas e reescreva cada token de entrada seguido de seu rótulo até o final da sentença de entrada.'

Input: "Tem catorze moinhos , na Ribeira de Caia , e Caldeirão , e três pisões ."
Response: "`Tem <|O|> catorze <|O|> moinhos <|O|> , <|O|> na <|O|> Ribeira <|B-PLC_AQU|> de <|I-PLC_AQU|> Caia <|I-PLC_AQU|> , <|O|> e <|O|> Caldeirão <|B-PLC_AQU|> , <|O|> e <|O|> três <|O|> pisões <|O|> . <|O|>`"

Figura 5: Exemplo de instrução

Para gerar os *prompts*, fez-se um *script* que lê o ficheiro original em formato CoNLL, e que, primeiramente fornece a frase sem anotações e, depois, outra com as anotações. Para cada exemplo do *corpus*, adiciona-se a mesma instrução: "Reconheça as entidades nomeadas e reescreva cada token de entrada seguido de seu rótulo até o final da frase de entrada". Ou, em inglês, *Recognize the named entities and rewrite each input token followed by its label to the end of the input sentence*. Foram adicionados três *tokens* especiais ao tokenizador: $< s >$, $< /s >$, $< unk >$, respetivamente *bos*, *eos* e *pad*. Definiu-se o *token* de início de frase para ser o primeiro *token* do *prompt* e o *token* de final de frase para ser o último. Deve ser tido em consideração a definição do final da frase com a introdução de um *token* especial para que o modelo aprenda a parar de gerar texto, evitando alucinações. Ainda no tokenizador, definiu-se um tamanho de entrada de 1024 *tokens*. No momento da predição, definiu-se o máximo de novos *tokens* de 512. Uma vez terminado o *corpus* de instruções, fez-se então o *instruct-tuning*, usando o *pipeline* de treino do HuggingFace para modelos causal. Para tentar reduzir os custos computacionais, usou-se a técnica de quantização que converte o modelo para uma precisão de 8 bit. Usou-se igualmente o PEFT-LoRa, que reduz o número de parâmetros treináveis. Ao usar o PEFT-LoRa, o modelo foi treinado com cerca de 33.5M de parâmetros. Com estas reduções conseguiu-se executar o *fine-tuning* numa GPU Tesla T4 com 16GB.

Com relação ao ensaio feito com o mT5, usou-se um *pipeline* de algoritmos *Text-to-Text* disponibilizado pelo framework HappyTransformer.[17] Apenas o tamanho da entrada e saída foram modificados para 512 *tokens*, sendo mantidos os

[17] https://github.com/EricFillion/happy-transformer

Input:ner: Tem catorze moinhos , na Ribeira de Caia , e Caldeirão , e três pisões ."
Target: "`Tem <|O|> catorze <|O|> moinhos <|O|> , <|O|> na <|O|> Ribeira <|B-PLC_AQU|> de <|I-PLC_AQU|> Caia <|I-PLC_AQU|> , <|O|> e <|O|> Caldeirão <|B-PLC_AQU|> , <|O|> e <|O|> três <|O|> pisões <|O|> . <|O|>`"

Figura 6: Exemplo de treino *text-to-text*

demais hiper-parâmetros. De modo semelhante ao que foi feito aquando da preparação dos dados para o *instruct-tuning* do LLaMa2, realizou-se um *script* que devolve dois tipos de frases a partir dos dados originais em CoNLL. O algoritmo devolve as frases de entrada (que contêm apenas o texto, sem anotações) e as frases-alvo (que contêm os *tokens* seguidos de seus respetivos rótulos). A Figura 6 mostra um exemplo. Assim, o algoritmo *Seq2Seq* recebe a frase sem as entidades nomeadas e é treinado para gerar uma frase com as entidades identificadas e classificadas. Note-se que a frase de entrada recebe um prefixo *ner:*, assim aprende que a tarefa que é solicitada é de reconhecimento de entidades. Executou-se esse ensaio numa GPU RTX 4090 24GB.

6.2. Processo de avaliação

Os modelos treinados pela abordagem *Transformer-Linear* e pela vanilla LSTM-CRF foram diretamente avaliados usando o script de avaliação do reconhecimento de entidades nomeadas do CoNLL-2002(Sang, 2002). As razões que levaram à escolha desse script prenderam-se com o facto de este já ser empregado em trabalhos para REN em português e também é amplamente usado para avaliar modelos de REN em inglês. Deste modo, o script retorna as métricas Precision (PRE), Recall (REC) e F_1 para cada categoria e de todo o *corpus* predito.

Relativamente às métricas utilizadas, pode entender-se Precisão como a percentagem de entidades nomeadas encontradas pelo sistema de aprendizagem que estão corretas. O parâmetro Recall pode ser entendido como a percentagem de entidades nomeadas presentes no *corpus* que foram encontradas pelo sistema.

A avaliação dos modelos mT5 e LLaMa2 requerem um pré-processamento antes de serem avaliados pelo script. O pré-processamento consiste em:

Architecture	Model	PRE	REC	F_1	$\Delta\uparrow$	$\Delta\downarrow$
Transformer-Linear	**XLM-R-Large**	**68.31**	73.38	**70.76**	+0.23	*sota*
	BERTimbau-Large	67.36	**74.00**	70.53	+3.03	−0.23
LSTM-CRF	**FlairBBP + W2V-SKPG**	67.77	67.23	67.50	+1.23	−3.03
	FlairBBP + Glove	66.50	66.04	66.27	+17.24	−1.23
Causal LM	**LLaMa 2 (8bit) + LoRa**	68.01	38.34	49.03	+6.28	−17.24
Text-to-Text	**mT5-Large**	48.55	38.19	42.75	*bl*	−6.28

Tabela 4: Métricas gerais. *bl* = baseline e*sota* = estado da arte.

	XLM-R			BERTimbau		
	PRE	**REC**	F_1	**PRE**	**REC**	F_1
AUTWORK	47.83	55.00	51.16	45.83	52.38	48.89
ORG	53.23	55.93	54.55	48.05	67.27	56.06
PER_AUT	78.95	93.75	85.71	77.78	87.50	82.35
PER_CAT	50.00	75.00	60.00	87.50	87.50	87.50
PER_DIV	69.57	80.00	74.42	76.74	82.50	79.52
PER_NAM	66.23	71.83	68.92	61.04	67.63	64.16
PER_OCC	60.71	62.96	61.82	44.12	60.00	50.85
PER_PGRP	55.17	76.19	64.00	50.00	61.90	55.32
PER_SAINT	75.69	78.99	77.30	77.37	79.10	78.23
PLC_AQU	72.73	76.71	74.67	66.20	67.14	66.67
PLC_FAC	59.52	66.67	62.89	65.33	67.12	66.22
PLC_GPE	78.84	77.87	78.35	77.87	81.55	79.66
PLC_LOC	60.00	72.53	65.67	65.35	74.16	69.47
PLC_MOUNT	75.00	92.31	82.76	56.25	69.23	62.07
TIM_CRON	66.67	65.71	66.19	69.33	77.61	73.24

Tabela 5: Melhores resultados por categoria

- Alinhamento das frases preditas pelos modelo. Esse alinhamento é feito pelo algoritmo Needleman-Wunsch.
- Separação das pontuações que estão juntas com *tokens*. Esse foi um problema comum nas predições do mT5.
- Eventualmente algumas vezes os *labels* podiam conter o símbolo @ por conta da fase de alinhamento. Neste caso, substituiam-se os *labels* da frase alinhada pelos *labels* da frase predita.
- Reescrita das frases no formato CoNLL. Não foram escritas no ficheiro final de avaliação as linhas em que o *token* ou rótulo-chave ou rótulo predito contém o símbolo @.

A partir de Paolini et al. (2021), estruturou-se o *pipeline* utilizado de pré-processamento para a avaliação de REN oriundos de modelos geradores.

Uma vez terminado o pré-processamento, aplicou-se o script de avaliação CoNLL-2002 e obtiveram-se as métricas.

6.3. Resultados

Nesta secção apresentam-se os resultados obtidos. A Tabela 4 mostra as medidas gerais de cada modelo avaliado. A partir da análise destes resultados gerais, estabeleceu-se o melhor modelo e o menos favorável ao reconhecimento de entidades nomeadas para o *corpus* das *Memórias Paroquiais*. Dois modelos tiveram o parâmetro $F_1 > 70\%$ com uma pequena margem de diferença entre eles, como mostram as colunas $\Delta\uparrow$ e $\Delta\downarrow$.

Analisando a métrica Precision (PRE), o modelo *XLM-R-Large* teve a métrica mais alta, pelo que foi o melhor modelo a identificar entidades corretamente. Por outro lado, o modelo *BERTimbau-Large* destacou-se na métrica Recall, tendo obtido a maior percentagem de entidades nomeadas reconhecidas. Já relativamente à métrica F_1, que une as duas métricas, o *XLM-R-Large* foi o melhor modelo.

Com relação ao uso do *Glove*, verificou-se que usar o *W2V-SKP* continua a ser a melhor opção.

Do ponto de vista dos dois modelos geradores (LLaMa2 e mT5), somente foram mostradas as métricas do modelo LLaMa2 da NousResearch, pois teve desempenho superior considerável em relação ao original da Meta. A avaliação mostra que o modelo LLaMa 2 (original) teve $F_1 = 42.71$, uma redução de $6,32$ pontos em relação a F_1 do LLaMa extra-oficial. Tais LLMs tiveram resultados significativamente menores que os outros modelos. Estima-se que se deva a um menor número de exemplos disponíveis no momento para algumas categorias, bem como a complexidade inerente a algumas categorias, que pode levar a ambiguidades (a mesma expressão ser anotada em duas categorias, consoante o contexto, por exemplo), dificultando a performance dada a realidade histórica em que opera. Esta hipótese alinha-se com o trabalho de Paolini et al. (2021), que mostrou resultados competitivos em várias tarefas de rotulação sequencial, mas com uma quantidade muito mais vasta de dados de treino. Assim, pode concluir-se, com base na métrica F_1, que o modelo *XLM-R-Large* foi o melhor modelo.

CATEG	Max		Min	
	Model	F_1	Model	F_1
AUTWORK	XLM-R	51,16	mT5	10,53
ORG	Glove	56,41	LLaMa2	13,11
PER_AUT	XLM-R	85,71	LLaMa2	66,67
PER_CAT	BERTimbau	87,50	Glove	52,17
PER_DIV	BERTimbau	79,52	mT5	33,33
PER_NAM	XLM-R	68,92	LLaMa2	40,35
PER_OCC	W2V-SKPG	71,43	mT5	7,41
PER_PGRP	XLM-R	64,00	LLaMa2	10,00
PER_SAINT	BERTimbau	78,23	mT5	67,98
PLC_AQU	XLM-R	74,67	LLaMa2	51,06
PLC_FAC	XLM-R	62,89	LLaMa2	42,31
PLC_GPE	BERTimbau	79,66	mT5	51,69
PLC_LOC	BERTimbau	69,47	mT5	19,87
PLC_MOUNT	Glove	85,71	BERTimbau	62,07
TIM_CRON	BERTimbau	73,24	mT5	30,00

Tabela 6: Melhores e piores modelos por categoria.

A Tabela 5 apresenta os melhores resultados detalhados para cada categoria do *corpus*.

A Tabela 6 apresenta o modelo que alcançou a máxima medida F_1 para cada categoria. Apresenta-se, igualmente, qual o modelo teve a menor medida $F_1 > 0\%$ para cada categoria. Pode notar-se que os modelos XLM-R e BERTimbau empataram quando nos referimos ao número de máxima F_1 por categoria, seguidos pelos empilhamentos de *embeddings* com os modelos Glove e W2V-SKPG. Essa análise permitiu identificar que o empilhamento de *embeddings* composto pelo Glove teve melhores métricas globais do que o empilhamento que contém o W2V-SKPG, mantendo-se o modelo W2V-SKPG mais estável.

Sobre as mínimas, o mT5 teve a maior quantidade de mínimas acima de zero, seguido pelo LLaMa2. Note-se, também, que o empilhamento que contém o Glove teve o pior resultado acima de zero na categoria PER_CAT, enquanto o FlairBBP+W2V-SKP não teve pior comportamento em nenhuma categoria. Destaca-se também que o BERTimbau teve o pior desempenho na categoria TIM_CRON.

Assim, após os vários ensaios realizados, pode concluir-se que, para estas tarefas, ainda é muito mais vantajoso utilizar um modelo do tipo BERT com uma camada linear.

7. Conclusão

Neste trabalho, apresenta-se um estudo sobre textos do século XVIII, escritos por párocos do Alentejo, Portugal, com base no reconhecimento de entidades nomeadas. Este estudo, motivado pelos objetivos de investigação dos historiadores, levou à definição de novas categorias e subcategorias de EN.

Apresentou-se uma análise das Entidades Nomeadas anotadas quanto à sua distribuição nas paróquias da região do Alentejo. A predominância da ocorrência de expressões com o rótulo da categoria GPE, além de reveladora de todo um território, parece reforçar igualmente a ideia de que o inquérito que foi lançado em 1758 tinha também o fim de retomar o projeto de um Dicionário Geográfico de Portugal, iniciado antes do terremoto de 1755 e interrompido por esta catástrofe (Olival et al., 2023a).

Do ponto de vista da História, o estudo dessas Entidades Nomeadas propicia a criação de conjuntos de dados robustos e confiáveis, que estarão aptos para a realização de comparações entre paróquias ou entre áreas do território português. Permite ainda localizar informações de interesse global que, de outra forma, estariam perdidas na mera escala local. Ficam disponíveis para serem ligadas a outras.

Do ponto de vista linguístico, a normalização gráfica efetuada permitiu gerar uma duplicação de dados textuais, em registo ortográfico original do século XVIII, e os correspondentes normalizados aplicando a norma ortográfica do século XXI, na variante do português europeu. Tal operação irá permitir, noutros estudos, o treino de modelos que possam lidar melhor com a variação gráfica e, assim, aumentar a performance dos modelos, diminuindo o trabalho de normalização manual dos textos. Este processo de regularização gráfica, embora apenas uma fase do *pipeline* do processo, foi também essencial para

um melhor conhecimento do estádio da língua portuguesa no século XVIII. As características gráficas e linguísticas deste *corpus* resultaram das condições da sua constituição textual, plural, escrito a várias mãos, por padres de diversas idades e formações, oriundos de várias regiões.

Do ponto de vista computacional, como não foram encontrados modelos anteriores treinados com as categorias estabelecidas no estudo, foi necessário treinar novos modelos. Neste processo, avaliaram-se várias alternativas e o melhor modelo foi o XLM-R-Large, que pode ser treinado numa única GPU, sem a necessidade de técnicas de redução de parâmetros, e em apenas algumas horas. As várias avaliações produzidas envolveram modelos multilingues e específicos para o português, com apenas uma pequena margem de diferença nas métricas dos dois melhores modelos, que são multilingues e monolingues (para o português), respetivamente. Face aos resultados atuais obtidos, visa-se, agora, usar os modelos num sistema de anotação assistida para acelerar o processo de anotação de toda a coleção das *Memórias Paroquiais*.

Como limitações, apontamos o uso de modelos treinados com *corpus* contemporâneo do Brasil para textos históricos de Portugal, e a etapa de normalização textual ser realizada de forma manual. Em trabalhos futuros, planeia-se, assim, refinar modelos para o português do século XVIII, tentar eliminar a necessidade de normalização manual, e expandir a anotação do *corpus*. Dessa forma pretende-se alargar este estudo para outras coleções, aproveitando o investimento manual e computacional que vem sendo feito. Como se demonstrou, estas etapas intermédias são essenciais. Permitem uma reflexão analítica que pode melhorar as escolhas e acelerar o processo. Por outro lado, ao ser desenvolvido um estudo que parte de um pré-treino validado cientificamente e feito por humanos, permite que todo o processo computacional subsequente, ainda que eventualmente mais complexo, possa ser aproveitado para outros estudos em *corpora* históricos em português.

Agradecimentos

Este trabalho recebeu apoio financeiro da Fundação para a Ciência e a Tecnologia (FCT) no contexto dos projetos CEECIND/01997/2017 e UIDB/00057/2020 - https://doi.org/10.54499/UIDB/00057/2020.

Referências

Aguilar, Gustavo, Suraj Maharjan, Adrian Pastor López Monroy & Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. Em *3rd Workshop on Noisy User-generated Text*, 148–153. 10.18653/v1/W17-4419

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter & Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. Em *North American Chapter of the Association for Computational Linguistics (NAACL)*, 54–59. 10.18653/v1/N19-4010

Akbik, Alan, Duncan Blythe & Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. Em *27th International Conference on Computational Linguistics (COLING)*, 1638–1649.

Albuquerque, Hidelberg O., Ellen Souza, Carlos Gomes, Matheus Henrique de C. Pinto, P.S. Ricardo Filho, Rosimeire Costa, Vinícius Teixeira de M. Lopes, Nádia F. F. da Silva, André C. P. L. F. de Carvalho & Adriano L. I. Oliveira. 2023. Named entity recognition: a survey for the Portuguese language. *Procesamiento del Lenguaje Natural* 70. 171–185. 10.26342/2023-70-14

Cameron, Helena Freire, Fernanda Olival, Renata Vieira & Joaquim Francisco Santos Neto. 2022. Named entity annotation of an 18th century transcribed corpus: problems, challenges. Em *2nd Workshop on Digital Humanities and Natural Language Processing (DHandNLP)*, 18–25.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. Em *58th Annual Meeting of the Association for Computational Linguistics, ACL*, 8440–8451. 10.18653/v1/2020.acl-main.747

Dettmers, Tim, Mike Lewis, Younes Belkada & Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. ArXiv [cs.LG/cs.AI]. 10.48550/arxiv.2208.07339

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for

language understanding. Em *North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. 10.18653/v1/N19-1423

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello & Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys* 56(2). 1–47. 10.1145/3604931

Farmer, David. 1997. *The oxford dictionary of saints*. Oxford University Press 4th edn.

Filho, Jorge A. Wagner, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. Em *11th Language Resources and Evaluation Conference (LREC)*, 4339–4344.

Grilo, Sara, Márcia Bolrinha, João Silva, Rui Vaz & António Branco. 2020. The BDCamões collection of Portuguese literary documents: a research resource for digital humanities and language technology. Em *12th Language Resources and Evaluation Conference (LREC)*, 849–854.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8). 1735–1780. 10.1162/neco.1997.9.8.1735

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang & Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. Em *10th International Conference on Learning Representations (ICLR)*,

Lafferty, John D., Andrew McCallum & Fernando C. N. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Em *18th International Conference on Machine Learning (ICML)*, 282–289.

Mangrulkar, Sourab, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul & Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. GitHub Repository.

Mikolov, Tomás, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Em *1st International Conference on Learning Representations (ICLR)*,

Needleman, Saul B. & Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3). 443–453. 10.1016/0022-2836(70)90057-4

Olival, Fernanda, Helena Freire Cameron, Fátima Farrica & Renata Vieira. 2023a. As memórias paroquiais (1758) do atual concelho de Vila Viçosa. *Callipole: Revista de Cultura* 29. 85–128.

Olival, Fernanda, Helena Freire Cameron & Renata Vieira. 2023b. As memórias paroquiais: Do manuscrito ao digital. Em *Jornada de Humanidades Digitais do CIDEHUS*, 75–92.

Paolini, Giovanni, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang & Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. Em *9th International Conference on Learning Representations, (ICLR)*,

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. GloVe: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162

Sang, Erik F. Tjong Kim. 2002. Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. Em *6th Conference on Natural Language Learning (CoNLL)*, 155–158.

Santos, Diana & Nuno Cardoso. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Santos, Joaquim, Helena Freire Cameron, Fernanda Olival, Fátima Farrica & Renata Vieira. 2024. Named entity recognition specialised for Portuguese 18th-century history research. Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 117–126.

Santos, Joaquim, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini & Renata Vieira. 2019. Assessing the impact of contextual embeddings for portuguese named entity recognition. Em *8th Brazilian Conference on Intelligent Systems (BRACIS)*, 437–442. 10.1109/BRACIS.2019.00083

Schweter, Stefan & Alan Akbik. 2020. FLERT: Document-level features for named entity recognition. ArXiv [cs.CL]. 10.48550/arXiv.2011.06993

Silva, Andressa Vieira. 2023. Uma revisão para o reconhecimento de entidades nomeadas aplicado à língua Portuguesa. *Linguamática* 15(2). 69–85. 10.21814/lm.15.2.396

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. Em *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 403–417. 10.1007/978-3-030-61377-8_28

Souza, Fábio, Rodrigo Frassetto Nogueira & Roberto de Alencar Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. ArXiv [cs.CL/cs.IR/cs.LG]. 10.48550/arXiv.1909.10649

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov & Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. ArXiv [cs.CL/cs.AI]. 10.48550/arxiv.2307.09288

Vieira, Renata, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira & Ivo Santos. 2021. Enriching the 1758 Portuguese parish memories (Alentejo) with named entities. *Journal of Open Humanities Data* 7. 20. 10.5334/johd.43

Zilio, Leonardo, Maria José Finatto & Renata Vieira. 2022. Named entity recognition applied to Portuguese texts from the XVIII century. Em *2nd Workshop on Digital Humanities and Natural Language Processing (DHandNLP)*, 1–10.