Reconhecimento Automático Multilingue de Entidades Mencionadas em Diversos Domínios, para Efeitos de Anonimização de Tradução Automática

Miguel Menezes[1,2], Vera Cabarrão[3], Helena Moniz[1,2], Pedro Mota[3]

[1]Universidade de Lisboa, Faculdade de Letras, Lisboa, Portugal
[2]INESC-ID, Lisboa, Portugal
[3]Unbabel, Lisboa, Portugal

Abstract

The following article describes the research developed at Unbabel, a Portuguese Machine-Translation start-up, that combines Machine Translation (MT) with human post-edition with a focus on customer service content. With the work carried out within a real multilingual AI powered, human-refined, MT industry, we aim to contribute to furthering MT quality and good-practices, by exposing the importance of having continuously-in-development, robust Named Entity Recognition systems for General Data Protection Regulation (GDPR) compliance. We will report three different experiments, resulting from a shared work with Unbabel´s linguists and Unbabel´s Artificial Intelligence (AI) engineering team, matured over a year. The first experiment focused on developing a methodology for the identification and annotation of domain-specific Named Entities (NEs) for the Food-Industry. The devised methodology allows the construction of gold standards for building domain-specific NER systems and can be applied for a myriad of different domains. With the implementation of the designed method, we were able to identify the following domain-specific NEs set: *Restaurant Names*; *Restaurant Chains*; *Dishes*; *Beverage*, *Ingredients.* The second and third experiments explored the possibilities of constructing, in a semi-automatically way, multilingual NER gold standards for different domains and language pairs, using aligners that project Named Entities across a parallel *corpus.* Both experiments made it possible to benchmark four different open-source aligners (SimAlign; Fastalign; AwesomeAlign; Eflomal), allowing to identify the one with better performance and, simultaneously, validate the aforementioned approach. This work should be taken as a statement of multidisciplinary, proving and validating the much-needed articulation between different scientific fields that compose and characterize the area of Natural Language Processing (NLP).

Keywords: Machine-Translation; Named Entities; Annotation; Gold Standards; Aligners.

Palavras-chave: Tradução Automática; Entidades Mencionadas; Anotação; *Gold* Standards; Sistemas de Alinhamento.

1. Introdução

Profissionais do serviço de apoio ao cliente lidam com todo o tipo de problemas decorrentes da interação humana, que vão desde responder a perguntas ou reclamações de clientes, ao processamento de encomendas e devoluções, até à partilha de informações e serviços, entre outros. São, de certa forma, uma linha direta entre um cliente e um prestador de serviços, pelo que devem ser eficientes, rápidos e, acima de tudo, ter um discurso claro e inteligível, enquanto trabalham de forma remota. De forma a potencializar o serviço de apoio ao cliente, hoje em dia, são aplicadas diversas técnicas de Inteligência Artificial (AI), nomeadamente, a de Tradução Automática (TA), como forma de ultrapassar barreiras linguísticas.

Para o efeito, destacamos o papel das Entidades Mencionadas (EM), que compõem uma parte vital das arquiteturas de TA, sendo que promovem um aumento da qualidade nas traduções e asseguram que todos os

requisitos exigidos pela Regulamentação de Proteção de Dados de 2018[1] sejam devidamente cumpridos. Dada a sua importância, este tema será frequentemente mencionado ao longo do presente artigo.

Tendo em conta a relevância das EM para a TA, a falta de ferramentas automáticas de Reconhecimento de Entidades Mencionadas (REM) adequadas a domínios específicos e a escassez de dados de treino/teste multilingues para a construção de modelos de REM multilingues funcionais, foram desenvolvidos três estudos experimentais para ultrapassar as dificuldades acima mencionadas.

Numa primeira fase, foram traçados os caminhos para a construção de sistemas de REM adaptáveis a domínios específicos, permitindo a validação da metodologia concebida para o efeito. Esta experiência prevê que, de futuro, se possam desenvolver e implementar sistemas dedicados a diferentes conteúdos, *i.e.*, de valor mais genérico ou de cariz mais especializado, podendo estes sistemas ser treinados, de igual forma, para corresponderem às convenções específicas de determinados clientes.

Adicionalmente, foram investigados métodos alternativos para gerar semiautomaticamente anotações de EM num *corpus* paralelo, com recurso a sistemas automáticos de alinhamento. O recurso a esta técnica permite obter conjuntos de dados para treino e teste de sistemas de REM de forma rápida e menos dispendiosa, ultrapassando o obstáculo da falta de dados multilingues anotados. De forma a atingirmos os objetivos propostos, quatro sistemas de alinhamento baseados em algoritmos foram requisitados de forma a projetar anotações de EM de um documento-fonte para múltiplos documentos-alvo.

1.1 O que são EM.

O conceito de Entidades Mencionadas aparece na segunda metade do século XX, durante o período da Guerra Fria, como parte de um programa de investigação levado a cabo com o apoio do Departamento de Defesa dos Estados Unidos. Esperava-se, com a criação e implementação de novas técnicas, entre as quais a identificação de EM, conseguir extrair de forma automática informação essencial e relevante de grandes volumes de dados, o que se traduziria num avanço estratégico, especialmente relevante dada a conjuntura política da altura (Nouvel *et al*, 2016).

As possibilidades e aplicações futuras deste tipo de técnicas acabariam por levar, em 1987, à primeira conferência de extração de dados, *Message Understanding Conference* (MUC)[2], com o intuito de incrementar o conhecimento e melhorar os métodos de extração de informação disponíveis, iniciando-se, assim, uma tradição que se mantém até aos dias de hoje. Durante essas conferências, eram levadas a cabo múltiplas campanhas de avaliação, nas quais se comparavam os desempenhos dos diferentes sistemas automáticos propostos com os resultados levados a cabo por profissionais, permitindo uma percepção real da qualidade e *performance* dos sistemas apresentados na conferência. No entanto, e apesar dos esforços exercidos para levar este género de tarefas a bom porto, havia bastantes restrições tecnológicas, uma vez que os recursos disponíveis não eram suficientes para corresponder às necessidades que tais tarefas implicavam. Apesar das limitações, as experiências realizadas durante as conferências MUC permitiram uma tomada de consciência da importância e aplicabilidade futura de sistemas automáticos associados à informação, acabando por fomentar o nascimento da subárea de Processamento de Língua Natural (PLN) denominada como REM.

Hoje em dia, apesar de ser considerada uma subárea com plena implementação, ainda não existe um verdadeiro consenso sobre uma definição concreta de EM, o que se pode atribuir ao facto de estas serem elementos polivalentes e adaptáveis, podendo assumir diferentes formas, *i.e.*, estruturas lexicais de diferentes tipos, cuja relevância e seleção é determinada pelo objetivo da tarefa para a qual são requisitadas. Tendo em conta estes atributos, uma unidade lexical específica pode ser tomada como EM para uma tarefa particular, mas não para outra, como veremos mais à frente. Há, no entanto, características partilhadas por todas as EM que, *grosso modo*, auxiliam na identificação e conceptualização destas estruturas: "unidades textuais que correspondem a categorias semânticas pré-definidas" (Nouvel *et al*, 2016), apresentando-se como

[1] https://gdpr-info.eu
[2] https://dl.acm.org/conference/muc

"identificadores unívocos" (Nouvel *et al*, 2016), frequentemente referindo-se a "sintagmas contendo nomes de pessoas, organizações e localizações" (Sang *et al*, 2003) entre outras, funcionando como objetos linguísticos com informação relevante e essencial. É comum, seguindo as convenções adotadas na área, separar as EM em três grandes grupos: *ENAMEX*, referindo-se a *Nomes*, *Localizações*, *Produtos e Serviços*, entre outros; *TIMEX*, referindo-se a EM que denotam tempo, como por exemplo as *Horas*; e uma terceira categoria referindo-se exclusivamente a expressões numéricas, *NUMEX*, como por exemplo *Percentagens*, *Identificadores Bancários* como NIB ou o *Número de Cartão de Crédito.*

A versatilidade inerente às EM apresenta-se como uma mais-valia, permitindo diversas aplicações com diferentes fins, que passaremos a descrever.

1.2 Quais os Impactos de um Sistema de REM?

Como anteriormente mencionado, as EM são de extrema importância para a indústria da tradução automática. No entanto, a sua importância não termina aqui. Atualmente, implementam-se sistemas de REM para diversos fins. Um bom exemplo é a tarefa de sumarização automática, em que um sistema de REM é implementado de forma a identificar as EM num documento, e às quais estão acoplados os elementos-chave de um texto, ou seja, a informação chave, permitindo explicitar e sumarizar os temas fulcrais do documento de uma forma rápida e automática. Simultaneamente, estes sistemas começaram a ser aplicados ao setor da saúde, de forma a conseguir identificar e caracterizar doenças, medicamentos e sintomas, desta forma auxiliando os profissionais de saúde a chegar a diagnósticos e tratamentos mais facilmente.

Continuando na área da saúde, um sistema de REM pode ainda ser empregado para detetar possíveis surtos infeciosos, através da sua aplicação na análise de conteúdos nas redes sociais, como o Twitter. Neste caso, o sistema é aplicado às mensagens geradas pelos utilizadores, permitindo destacar e extrair as EM relevantes, que, mais uma vez, estão indubitavelmente associadas aos segmentos-chave e à informação pretendida para aquele fim. Outro benefício, já anteriormente mencionado, é o facto de atender à necessidade de respeitar a Lei da Proteção de Dados, evitando que sejam implementadas penalizações por incumprimento da lei.

1.3 Tipologia das Entidades Mencionadas

Para a tarefa de REM, é importante i) definir quais as EM relevantes para o trabalho e ii) como anotá-las. Este processo requer a criação de novos tipos de EM, isto é, "uma formalização descritiva das categorias selecionadas" (Nouvel et al., 2016: 48), que normalmente vem sob a forma de diretrizes de anotação. O projeto multifacetado levado a cabo permitiu desenvolver a atual tipologia genérica de Entidades Mencionadas criada pela Unbabel, seguindo a tríade universal de Categorias de EM: Enamex, Numex e Timex. De seguida, uma descrição mais pormenorizada destas:

Enamex	**Timex**	**Numex**
Name		*Measurements*
Username		*Currencies*
Locations		*Percentages*
Cities	*Dates*	*Passwords*
Countries	*Date-Relatives*	*IP-Addresses*
Addresses	*Hours*	*ID-Numbers*
Organizations	*Duration*	*Tax-Number*
Phone-Number		*Creditcard*
Email		*Bank-ID*
Products and Services		*Ref-Number*

URL		*Device-ID*

Tabela 1: Conjunto de EM usadas pela Unbabel

Além do mais, o trabalho de análise de dados dos vários clientes da mesma área permitiu aos investigadores a criação de novas categorias de EM e correspondentes diretrizes de anotação (ver secção 3.1.1), sendo estas posteriormente usadas para atualização das existentes na Unbabel.

1.4 Aplicações de Sistema de REM em Arquiteturas de Tradução Automática

De uma forma sumarizada, o processo de preparação de um documento para tradução começa após a receção do mesmo, geralmente, através de plataformas de integração. Desta forma, o documento é sujeito a uma fase de análise, sendo a arquitetura de TA ajustada conforme as especificações e características da encomenda. É nesta fase que se vai aplicar, entre outros processos, o sistema de REM.

O processo de REM é composto por 2 passos: o primeiro corresponde à identificação das entidades relevantes. Neste passo, o sistema é aplicado, permitindo a identificação da entidade em contexto, sendo esta posteriormente bloqueada para tradução, no caso particular da empresa. Este passo é motivado pelo facto de as EM serem elementos particularmente suscetíveis a erros de tradução realizados pelos sistemas de TA.

Adicionalmente, algumas EM particulares são sujeitas a um segundo passo, a anonimização. Para estas, é-lhes atribuído um hiperónimo, conhecido como *placeholder*, substituindo a entidade pela tipologia pertencente, como por exemplo, *Date*; *Email; Country*, entre outros. Para os nomes próprios, a anonimização é feita recorrendo a um equivalente semântico, ou seja, nomes fictícios, de forma a manter as características de género ao longo do documento. Veja-se o caso seguinte apresentado por Mota et al. (2022), de uma arquitetura de anonimização recorrendo a um equivalente semântico:

Input:	Hi Zéphyrin
Reconhecimento de EM:	Hi **Zéphyrin**
Tradução da EM:	Hi **[Zéphyrin ➢Zéphyrin]**
Equivalente Semântico:	Hi **[Thomas ➢Thomas]**
Sistema de TA:	Bonjour **Thomas**
Output:	Bonjour **Zéphyrin**

Figura 1: Arquitetura de anonimização recorrendo a um equivalente semântico.

Este segundo passo vai então permitir a uma empresa de tradução: i) assegurar que nenhuns dados sensíveis sejam revelados a terceiros (editores, anotadores etc.) ou utilizados para aprendizagem dos sistemas de tradução automática; ii) assegurar concordância de género ao longo do documento, no caso dos nomes próprios; iii) fomentar a legibilidade do documento; e iv) prevenir erros de tradução automática.

2. Estado-da-Arte

Nos últimos anos, os sistemas de aprendizagem automática têm sido predominantemente usados para alcançar resultados de REM de última geração, conseguindo grandes avanços desde as primeiras iniciativas MUC. Um fluxo constante de trabalhos na área, tanto a nível da indústria como em ambientes mais académicos, tem produzido mudanças significativas que vão desde novas tecnologias de alto desempenho associadas à subtarefa de REM, até novos objetivos e aplicações, como os já anteriormente mencionados para a área da

saúde. Recentemente, a dicotomia domínio genérico *vs.* domínio específico tem sido um tema de particular interesse para muitos estudiosos na área, gerando conhecimento sobre as implicações da utilização de sistemas dedicados a domínios particulares, como, por exemplo, a geração de mapas a partir de fontes históricas ou de obras literárias (Harper et al., 2020), ou, como observável através do trabalho desenvolvido por Cheng et al. (2019) e no trabalho de Tian et al. (2019), construção de *corpora* de EM em língua inglesa, a partir de várias línguas fonte, abrangendo vários domínios para melhorias de desempenhos dos sistemas de REM.

Apesar de todos os avanços tecnológicos atualmente em implementação, uma grande parte do trabalho desenvolvido na área ainda depende de uma forte intervenção humana, sendo o desenvolvimento influenciado pela disponibilidade, ou não, de dados anotados para diversas línguas e domínios distintos. Como resposta, foram propostas soluções semiautomáticas para superar a necessidade de "recursos manuais dispendiosos" (Santos & Guimarães, 2015: 1). De igual forma, surgiram novas estratégias para resolução dos obstáculos acima mencionados, nomeadamente, a utilização de técnicas de projeção de EM em *corpora* paralelos, bitextos, *i.e.*, a projeção de anotações de EM de um documento-fonte para um documento-alvo, através de sistemas de alinhamento, permitindo a criação de dados de anotação multilingues (Eskin et al., 2019). À semelhança, têm sido desenvolvidos com resultados satisfatórios novos modelos, baseados em arquiteturas de *encoder/decoder*, como, por exemplo, *Transformers*, em que é feito um mapeamento de frases a vetores, resultando numa representação da sequência de entrada da língua fonte.

Akbik et al. (2018), por sua vez, para a tarefa partilhada CoNLL03, propõem um novo tipo de *word embeddings*, designadas como *string embeddings* em contexto, de forma a potencializar os modelos linguísticos, que aplicados às tarefas de REM permitem registar valores de F-1[3] associados apenas a modelos de última geração. Os autores propõem que tais *embeddings* "sejam treinados sem qualquer noção explícita de palavra, modelando palavras como sendo exclusivamente uma sequência de caracteres (...) contextualizadas pelo texto circundante, o que significa que uma mesma sequência de caracteres terá diferentes *embeddings*, dependendo do contexto em que é usada". De acordo com o proposto pelos autores, os *embeddings* permitem i) ser pré-treinados em grandes volumes de *corpora* não classificados, ou seja, dados não caracterizados, não identificados ou não etiquetados; ii) captar o significado da palavra em contexto, conseguindo, assim, diferentes *embeddings* para palavras polissémicas; iii) modelar palavras e contextos de forma a lidar com erros de ortografia, tais como prefixos e sufixos.

Wang et al. (2019) propõem a utilização do M-BERT (Multilingual Bidirectional Encoder Representations from Transformers), para transferência multilingue "sem direção multilingue e sem alinhamento de dados" (ibidem). A experiência foi feita recorrendo a três línguas diferentes (espanhol, hindi e russo) e mostrou que o M-BERT "parece generalizar bem entre línguas para uma variedade de tarefas posteriores" (Wu & Dredze, 2019), como REM e etiquetagem morfossintática (*POS tagging)*.

Como acima referido, vários modelos de REM foram propostos ao longo dos anos, alguns com objetivos mais académicos, como no caso do Grupo NLP Stanford[4], que desenvolveu uma arquitetura de processamento de língua natural (Stanza), na qual o sistema REM utilizado é apenas um dos muitos passos disponibilizados. Outros, por sua vez, são mais orientados para a indústria, como no caso da Google[5] e do Spacy[6], que se concentram no cumprimento dos requisitos de memória e velocidade de inferência. Em todos os sistemas, é tida em conta uma grande variedade de EM, que vão desde *Address; Date-Time; Credit Cards*, no caso da Google, ou *Location; Facilities; Law; Language*, entre outros, no caso do Spacy. Além disso, tanto o Stanza como o Spacy permitem a personalização dos seus modelos pré-definidos.

[3] Média harmónica entre precisão e cobertura.
[4] https://nlp.stanford.edu/software/CRF-NER.html
[5] https://cloud.google.com/natural-language/docs/analyzing-entities
[6] https://spacy.io/api/entityrecognizer

3. Metodologia

Esta secção descreve as três experiências realizadas para a subtarefa de REM. Numa primeira fase, foi delineada uma experiência no âmbito do projeto desenvolvido, para o qual se deram os primeiros passos na identificação de EM de domínios específicos, com vista à criação de um sistema de REM dedicado a domínios específicos.

Numa segunda fase, realizámos duas experiências com base em projeções de EM, recorrendo à utilização de sistemas de alinhamento, que alinham de forma automática EM de um documento-fonte com as EM homólogas no documento-alvo. Para esta tarefa foram utilizados dados disponíveis na Unbabel nos domínios do Turismo e do ramo Tecnológico, com o objetivo de criar e avaliar conjuntos de treino e teste multilingues.

3.1 Experiência Piloto de Anotação na Área da Indústria Alimentar

Para esta tarefa, foi selecionado um conjunto de dados em inglês (EN) da área da indústria alimentar, mais concretamente a de entrega ao domicílio de bens perecíveis. A partir da análise do conjunto de dados, novas tipologias de EM do domínio foram propostas, para as quais se procedeu à criação de novas diretrizes de anotação com vista à criação de um conjunto de treino e teste do novo sistema de REM. A metodologia desenvolvida procurou ser replicável e robusta o suficiente de forma a ser aplicável a outros domínios.

Para a experiência, compilou-se um extenso *corpus* de três clientes do domínio para a análise de EM particulares ao domínio e comuns aos três clientes. As tabelas seguintes representam o conjunto total de dados do domínio específico, recolhidos para cada cliente da área e utilizados para identificação das EM. Os tipos de dados têm várias proveniências, nomeadamente, de MT (Tabela 2), Glossários (Tabela 3), e Memórias de Tradução (Tabela 4). Com a análise dos dados de diferentes proveniências, *i.e.,* diferentes clientes, diferentes glossários e memórias da área da indústria alimentar foi possível determinar as categorias de EM de domínio específico prevalentes em todos os conjuntos de dados. Os dados analisados encontram-se distribuídos da seguinte forma:

Clientes	Número de segmentos (frases)
Cliente 1	816
Cliente 2	763
Cliente 3	444

Tabela 2: Conjunto de dados de Traduções Automáticas

Cliente	Entradas de glossários
Cliente 1	870
Cliente 2	25 974
Cliente 3	21 864

Tabela 3: Número de entradas do glossário por cliente

Clientes	Memórias de Tradução
Cliente 1	25 757

Cliente 2	610
Cliente 3	61

Tabela 4: Número total de memórias de tradução por cliente de domínio

Da tarefa de prospecção acima descrita foram identificadas as seguintes categorias de EM: *Restaurant Names, Restaurant Chains, Dish Names, Beverages, Ingredients.*.

3.1.1 Diretrizes de Anotação

As diretrizes de anotação propostas para cada nova categoria alimentar foram concebidas para ajudar e facilitar a experiência de anotação de EM, levando a uma maior concordância entre anotadores, sendo estas definidas da seguinte forma:

- A etiqueta *Restaurant Name* foi concebida para este domínio específico como uma entidade referente a um local onde as refeições são preparadas e servidas aos clientes. Propusemos a etiqueta *Restaurant Name*, em vez de utilizar a etiqueta mais genérica *Location*, uma vez que a etiqueta *Restaurant* fornece uma descrição mais adequada do domínio específico;
- A etiqueta *Restaurant Chains* foi concebida para identificar todas as entidades referentes a lojas de retalho que operam com o mesmo nome e que vendem produtos semelhantes. Não nos esquecemos de assinalar que, em alguns casos, um determinado restaurante, pertencente a uma determinada cadeia alimentar, poderá ter associado ao seu nome o local onde se encontra situado, como elemento diferenciador. Nestes casos, pediu-se ao anotador para fazer uso da categoria *Restaurant Chains*;
- A etiqueta *Dish Name* foi considerada particularmente relevante para o domínio alimentar para o processo de anotação de EM devido à sua frequência. Esta etiqueta foi definida como o nome atribuído a um prato em particular podendo este compreender, no seu nome, conjuntos de ingredientes;
- A etiqueta *Beverages* foi concebida para identificar bebidas, podendo esta categoria ser aplicada a nomes de diferentes marcas de bebidas;

3.2 Projeção de EM para os domínios do Turismo e da Tecnologia

Para as duas experiências, com o objetivo de testar a viabilidade de construir de forma semiautomática conjuntos de treino multilingues, foram utilizados quatro sistemas de alinhamento de palavras de última geração e de acesso livre. Além disso, e antes de enviar os conjuntos de dados para alinhamento, foi possível determinar o nível de confiança e exatidão entre os anotadores designados para a tarefa, *i.e.*, recorrendo a uma análise da concordância inter-anotadores, expressa sob um valor percentual, o que permitiu estabelecer um bom nível de confiança sobre o profissionalismo e competência dos profissionais requisitados para a tarefa em questão. Os valores de concordância inter-anotadores poderão ser observados em 4.2.1.

Para a tarefa de alinhamento do conjunto de dados proveniente da área do Turismo, foram utilizados dados paralelos (bitexto) em inglês (EN), como língua fonte, e em alemão (DE) como língua alvo. Antes da implementação dos sistemas de alinhamento, os conjuntos de dados, (EN-DE), compreendendo 2500 frases cada, foram submetidos a uma anotação manual de EM por dois linguistas, em que um foi responsável pela anotação dos dados em EN, enquanto um segundo foi responsável pela versão em DE. Para a tarefa de anotação em DE, foram utilizados dois conjuntos de dados provenientes do mesmo documento-fonte (EN), um traduzido com recurso exclusivo à tradução automática, ao outro foi adicionada uma tarefa extra de pós-edição humana (PE). Tanto a versão EN como a DE sofreram uma fase de pré-processamento, em que os conjuntos de dados

foram divididos em frases, permitindo que a anotação fosse feita frase a frase, utilizando o Prodigy[7], uma plataforma de anotação. Ambos os anotadores utilizaram as diretrizes de anotação internas de EM da Unbabel para evitar enviesamento de resultados, não houve contacto entre os anotadores durante a fase de anotação, nem estes tiveram acesso ao trabalho um do outro. No entanto, e tendo em vista uma tarefa de anotação equitativa, ambos os anotadores utilizaram as mesmas diretrizes de anotação e foram autorizados a aceder a informação *online*, nomeadamente, a dicionários, mapas e outras fontes de informação relevantes para a tarefa. Os resultados da anotação foram então entregues à equipa de engenharia de PLN, para se proceder ao alinhamento do *corpus* paralelo anotado.

3.2.1 Sistemas Automáticos de Alinhamento

Para compreender o impacto da utilização de uma abordagem de alinhamento na construção de um sistema de REM multilingue, foram testados quatro alinhadores de última geração: FastAlign (https://github.com/dgel/fastalign), o sistema de alinhamento utilizado atualmente pela Unbabel, o eflomal, (github.com/robertostling/eflomal); o SimAlign, (github.com/cisnlp/simalign), e o AwesomeAlign (github.com/neulab/awesome-align). Cada sistema de alinhamento tinha disponível diferentes configurações que, quando combinadas, totalizavam um total de 53 possibilidades de alinhamento diferentes para cada categoria de entidade. As diferentes configurações variam desde:

- Heurística, permitindo diferentes direções de alinhamento: da fonte ao alvo e vice-versa, com o objetivo, de acordo com Mota et al., (2022) de "fornecer alinhamentos de palavras melhores/consistentes".
- Dados de treino que vão desde dados mais genéricos até dados de clientes ou dados mistos (tanto genéricos como dados de clientes)

ou

- Modelos pré-treinados para a compreensão multilingue.

Utilizando os alinhamentos das palavras, as EM identificadas na frase do documento-fonte foram projetadas no documento-alvo com base num algoritmo de min-max. Isto significa que se considerou a extensão da entidade no documento-alvo, variando entre o maior e menor alinhamento de palavras. Os resultados da Tarefa de Projeção de EM foram apresentados para avaliação utilizando um software online, desenvolvido pela equipa de engenharia de Inteligência Artificial (AI) da Unbabel, que mostrava todos os resultados para cada sistema de alinhamento, juntamente com as configurações associadas. Os resultados de cada alinhamento foram disponibilizados num sistema quantitativo de melhor (em posição 0), a pior (em posição 53). Adicionalmente, a interface permitiu uma análise comparativa de dois modelos de alinhamento (ver Figura 2), dando um panorama sobre a qualidade do alinhamento para cada categoria (ver Figura 3 para a categoria *Name*)

[7] https://prodi.gy

Figura 2: Comparação entre dois sistemas de alinhamento, *Model A e Model B*. possibilitando escolher entre diferentes configurações

Figura 3: Melhor pontuação de alinhamentos para a categoria *Name*, considerando as diferentes configurações do modelo (Model; Heuristics; Train Data)

Com acesso à informação exibida pela interface acima mencionada, conseguimos identificar as diferenças nos alinhamentos gerados entre EM do conjunto de dados em EN e o seu homólogo em DE. Além disso, pudemos comparar o conjunto de dados em DE com e sem pós-edição, determinando se tal tarefa interfere positiva ou negativamente nos resultados de projeção de EM. Adicionalmente, pudemos avaliar as configurações dos sistemas de alinhamento com melhor desempenho de entre as 53 possíveis combinações, comparando com o sistema usado pela Unbabel. A tarefa de projeção foi avaliada utilizando uma definição de classificação com as seguintes métricas de desempenho padrão: Precisão (*Precision*), Cobertura (*Recall*) e F-1 (Makhoul et al., 1999), a fim de ter uma perspetiva de desempenho mais fina dos resultados do modelo aplicado:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Figura 3: Fórmula de precisão e cobertura

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figura 4: Fórmula para calcular a medida F1

O valor de precisão é definido como o número de previsões de categorias de EM positivas (verdadeiros positivos) dividido pela soma de verdadeiros positivos e falsos positivos. Esta fórmula é utilizada para compreender a precisão do classificador em relação às categorias. A questão a que o conceito de precisão responde é: de todas as categorias de EM identificadas pelo algoritmo de projeção, quantas foram identificadas com a devida categoria? Valores mais baixos de precisão indicam um maior número de falsos positivos.

O valor de cobertura é definido como o rácio de verdadeiros positivos, corretamente previstos, divididos pela soma dos verdadeiros positivos e falsos negativos. A resposta à pergunta da cobertura é: de todas as EM presentes no conjunto de dados de teste, quantas foram identificadas pelo algoritmo de projeção de EM?

A medida F, também conhecida como F-1, por sua vez, é definida como a média harmónica da precisão e da cobertura, que, de uma forma simplista, podemos definir como uma ideia geral da qualidade.

4. Resultados

O capítulo seguinte centra-se nos resultados obtidos nas três experiências realizadas no âmbito do REM. Em primeiro lugar, serão apresentados os resultados referentes à construção de sistemas de REM dedicados, passando à apresentação dos resultados dos sistemas de alinhamento para a experiência realizada na área do Turismo e da Tecnologia.

4.1 Sistema de REM para a Indústria Alimentar

No conjunto de dados dos clientes de entrega de alimentos ao domicílio, constituído por 9797 frases, os dois anotadores identificaram 122 *Dish Names*; 23 *Beverages*; 18 *Restaurant Names*; 12 *Restaurant Chains* e mais 12 *Ingredients*; um número de EM abaixo do esperado. O conjunto de dados anotados foi então utilizado como conjunto de treino (80%) e teste (20%).

Dois modelos foram treinados de forma a permitir uma análise comparativa. Um foi treinado com o conjunto de dados de treino de domínio específico, o outro foi treinado com o mesmo conjunto e todos os dados anotados com EM disponíveis na Unbabel.

Para a fase de teste, foram utilizados quatro conjuntos de dados de teste diferentes: i) um contendo dados do cliente 1, ii) um segundo com dados do cliente 2, iii) um terceiro com dados do cliente 3, e iv) um quarto conjunto de dados que combinava todos os dados acima referidos.

Devido às baixas ocorrências de EM no conjunto de dados de treino do domínio específico (Mota et al., 2022), o que dificultou a construção de um modelo NER preciso e robusto capaz de identificar um número significativo de EM alimentares relevantes, só foi possível verificar os resultados de qualidade dos modelos de REM para a categoria Dish Name. Os resultados poderão ser observados nas seguintes tabelas:

Conjunto de teste	Cobertura	Precisão	F-1
Cliente 1	50	100	66.67
Cliente 2	0	0	0
Cliente 3	66.6	75	70.59
Combinação dos dados	63.64	77.78	70

Tabela 5: Resultados para categoria *Dish Name* para o sistema REM treinado com o conjunto de treino genérico e de domínio específico, dividido por quatro conjuntos de dados de teste.

Conjunto de teste	Cobertura	Precisão	F-1
Cliente 1	70	77.8	73.68
Cliente 2	0	100	73.68
Cliente 3	70.59	85.71	77.42
Combinação de dados	65.52	82.61	70

Tabela 6: Resultados para categoria *Dish Name* para o sistema REM treinado com o conjunto de treino de domínio específico

Para esta categoria, os resultados mostram que os dois modelos se comportaram de forma diferente, mostrando, para alguns conjuntos de teste, discrepâncias na medida F-1. Comparando os resultados, consegue identificar-se que, para os quatro conjuntos de dados utilizados nos testes, o modelo de REM treinado exclusivamente com o conjunto de dados de domínio específico superou o modelo treinado com conjunto de treino genérico e de domínio, mostrando valores F-1 constantes acima de 70%. Além disso, é possível salientar o facto de que, como se vê na Tabela 5, para o cliente 2, nenhum *Dish Name* foi identificado, resultando num valor de F-1 de 0, contrapondo com o modelo de REM expresso na Tabela 6, para o mesmo cliente, que obteve um valor de F-1 de 73.68%.

Em relação à combinação de dados de EM (genérico e de domínios) usada para teste, os resultados parecem ser mais consensuais, com resultados de F-1 idênticos. O modelo treinado exclusivamente com o conjunto de dados de domínio específico, ainda assim, apresenta valores de Cobertura e Precisão mais elevados, com 65,52 e 82.61%, enquanto o sistema de REM representado pela Tabela 5 (treinado com dados genéricos e de domínio específico) apresenta valores de 63.64% e 77.78%, respectivamente.

Embora os resultados da experiência tenham ficado aquém das expectativas para ambos os modelos, mostrando a necessidade de conjuntos de treino e teste mais adequados para melhoria dos resultados, foi conseguido, ainda assim, determinar uma metodologia adequada para a identificação das EM específicas de domínio. A metodologia desenvolvida no nosso projeto foi, entretanto, já aplicada para a identificação de EM particulares à indústria de jogos (Silva, 2021).

4.2 Tarefas de Projeção

As secções seguintes mostram os resultados de alinhamentos e análises de ambas as nossas experiências (domínios do Turismo e da Tecnologia) no âmbito da retroprojeção de EM. De igual forma, é apresentada a interface *online* de análise utilizada para comparar os sistemas de alinhamento e a partir da qual conseguimos

identificar o sistema com melhor desempenho e comparando-o com o sistema de alinhamento utilizado atualmente pela Unbabel.

4.2.1 Concordância inter-anotadores

Para a análise de concordância inter-anotadores, consideramos uma entidade mencionada compatível dentro de ambos os conjuntos de treino (EN/DE) sempre que ambos os anotadores tenham concordado: i) no intervalo de identificação da entidade e ii) na categoria.

A análise realizada permitiu-nos identificar uma elevada concordância inter-anotadores, entre o conjunto de treino e teste em EN, e ambos os conjuntos de treino e teste em DE: 90% de concordância para o conjunto traduzido automaticamente e 91% para o conjunto de dados traduzido automaticamente e com pós-edição humana. Veja-se a seguinte tabela:

Named Entities Categories	Named Entities	Inter-Annotator Agreement Results	
	EN GS	DE MT GS	DE PE GS
Organizations	183	161	167
Currencies	284	276	278
Percentages	9	9	9
Refnumber	64	52	53
Names	45	43	43
Dates	106	102	102
Address	26	22	23
E-mail	12	12	12
Phone Number	15	15	15
Time	26	21	21
URL	18	17	17
City	56	39	39
Country	3	3	3
Products and Services (PRS)	13	4	4
Credit Card	1	1	1
Password	1	1	1
Username	1	1	1
App. Ref. Number Code (ARN)	1	0	0
Total	865	781	789

Tabela 7: Concordância inter-anotadores para as Entidades Mencionadas

Observando o conjunto de treino e teste de EM, conseguimos contabilizar 865 entidades identificadas pelo anotador um e 781 entidades pelo anotador dois para o conjunto de teste em DE, traduzido automaticamente, e 789 para o conjunto de teste traduzido automaticamente e pós-editado. Comparando o número de EM identificadas entre a versão EN e a versão DE, determinamos que o anotador dois anotou menos 9,72% de EM na versão traduzida com o sistema de tradução automática (MT) e menos 8,72% de EM na versão de tradução automática e pós-editada do que o anotador um.

Foram observados valores altos de concordância para EM específicas, nomeadamente as que identificam dígitos (Numex), tais como:

1. *Percentage*s: 100% de concordância entre EN e ambos os conjuntos de teste em DE.
2. *Currencies*: 97.1% de concordância entre o EN e o conjunto de teste apenas com MT e 97.8% MT e pós-editado;

3. *Phone numbers*: 100% de concordância.

As expressões temporais, (Timex), como *Dates* ou *Time* parecem seguir o mesmo padrão, ascendendo a um valor de concordância de 96,22% no caso de *Dates*, e 80,76%, para a categoria *Time*, tanto no conjunto de teste exclusivo de tradução automática como no pós-editado. Para entidades da categoria Enamex, *Countries* obteve-se 100% de concordância entre os anotadores e para *Names*, um valor de 95%. Aparentemente, parece haver uma consciência intuitiva partilhada em relação à anotação destas entidades, corroborada pelo material lexical circundante, auxiliando o anotador na sua tarefa e justificando os valores de concordância. Para um melhor entendimento, foi selecionado um conjunto de exemplos retirado dos dados e que passamos a descrever:

(1) En: "Dear Manuela Frieda Kalo"
(2) De: "Sehr geehrte(r) Manuela Frieda Kalo".

Saudações como no exemplo acima, *Dear ...*, ou em alemão *Sehr geehrte(r) ...*, insinuam que a seguinte palavra é uma EM, especificamente um nome, sendo relevante tanto para o processo de anotação humana como para o processo de aprendizagem do sistema de TA.

Com base nos valores de concordância de anotação para as categorias acima mencionadas, concluímos que todas estas EM reúnem mais consenso entre os anotadores. Nestes casos, houve poucas dúvidas quanto às etiquetas a escolher. Por outro lado, as EM como *Products and Services* (PRS) apresentam valores de concordância entre anotadores mais baixos, 30%. Muitas das EM anotadas como PRS no conjunto de teste em EN foram etiquetadas como *Organizations* (ORG) tanto no conjunto de teste de TA como no conjunto de teste pós-editado, sendo, assim, consideradas como anotações assimétricas. Além disso, para estas categorias, a mesma EM pode assumir ambas as categorias em frases diferentes, denotando características ambíguas. Nestes casos, a interpretação da frase completa, ou das palavras na vizinhança da entidade, é fulcral para determinar a categoria da mesma. Esta abordagem pode, no entanto, não ser suficiente, como se pode observar pelos seguintes exemplos:

(3) En: "Kindly make sure that one of the accepted cards like[Union pay credit card]$_{\textbf{organization}}$ is saved in your [HolidayConsultee]$_{\textbf{Organziation}}$ account."
(4) De: "Bitte stellen Sie sicher, dass eine der alzeptierten Karten [Union Pay Kredit-, die HolidayConsultee – Karte]$_{\textbf{Products and Services}}$ in Ihrem-Konto gespeichert ist."

Nos casos acima, cada entidade mencionada foi identificada como uma *Organization* (ORG) para o conjunto de dados de teste em EN, enquanto em DE foram etiquetados como *Products and Services*(PRS).

As diferenças na anotação residem no facto de que, nos dados em EN, a entidade foi tomada pelo anotador um como uma entidade que presta um serviço, enquanto nos dados em DE, o anotador dois interpretou a entidade como um serviço em si.

De um modo geral, a concordância entre anotadores foi elevada. Evidenciamos, no entanto, o facto de que para algumas categorias, como PRS, ORG e *Locations* (LOC) a anotação nem sempre foi totalmente consensual, levando a assimetrias e discordância entre os anotadores que, depois de devidamente discutidas, foram incluídas nas diretrizes de anotação para futuras tarefas na empresa.

4.2.2 Tarefa de Alinhamento do Conjunto de Dados da Área de Turismo

O nosso estudo produziu resultados muito promissores, mostrando que a abordagem concebida é robusta quando um sistema de alinhamento bem validado, aliado a configurações específicas, é implementado.

Com base nos resultados F-1 para cada EM, foram elaboradas para cada categoria de EM duas tabelas que permitiram identificar não só qual o sistema com melhor desempenho, mas também deram a informação necessária para uma análise comparativa entre o FastAlign (sistema de alinhamento utilizado pela Unbabel) e os três restantes sistemas. A primeira tabela apresenta os cinco melhores resultados de alinhamento. A segunda

tabela, dedicada exclusivamente ao FastAlign, apresenta os seus cinco melhores alinhamentos. Por razões de clareza e principalmente devido a restrições de espaço, apenas descreveremos e discutiremos os resultados para a categorias de EM *Currency*:

Model	Mode	Heuristic	Train data	category	precision	Recall	F1	time
SimAlign	**Bert**	**inter**	**No-data**	**CRR**	**0.981**	**0.975**	**0.976**	**0.0205**
SimAlign	kiwi	inter	No-data	CRR	0.981	0.974	**0.974**	0.0284
SimAlign	Kiwi	intermax	No-data	CRR	0.976	0.978	**0.974**	0.318
SimAlign	xlmr	mwmf	No data	CRR	0.976	0.977	**0.973**	0.4719
SimAlign	Kiwi	mwmf	No data	CRR	0.976	0.977	**0.973**	0.3695

Tabela 8: Os cinco melhores resultados de alinhamento para a *Currencies*, recorrendo às diferentes configurações

Model	Mode	Heuristic	Train data	category	Precision	Recall	F1	time
FastAlign 17th	**production**	**Grow-diag-final**	**No data**	**CRR**	**0.934**	**0.894**	**0.899**	**0.0007**
FastAlign 18th	production	intersect	No data	CRR	0.973	0.853	**0.889**	0.0007
FastAlign 19th	train	Grow-diag-final	Mixed data	CRR	0.914	0.883	**0.883**	0.0005
FastAlign 20th	Train	Grow-diag-final	generic	CRR	0.906	0.881	**0.878**	0.0005
FastAlign 21th	train	intersect	Mixed-data	CRR	0.975	0.824	**0.866**	0.0005

Tabela 9: Os cinco melhores resultados de alinhamento para FastAlign para a entidade *Currencies*, recorrendo às diferentes configurações

Com base na Tabela 2, podemos concluir que para a categoria *Currencies*, o SimAlign superou os restantes sistemas, produzindo os cinco melhores resultados de alinhamento a nível geral. Por outro lado, o FastAlign só conseguiu atingir a 17.ª posição no *Ranking Model*, resultando numa diferença de 0,076% entre ambos os sistemas.

Com a análise contrastiva, conseguiu aferir-se a qualidade de alinhamento para cada categoria:

Descriptives

	SimAlign	FastAlign	AwesomeAlign	eflomal
N	6	5	3	3

Tabela 10: Número de categorias para as quais cada sistema de alinhamento alcançou os melhores resultados.

Com base nestas análises de resultados, pudemos verificar que o SimAlign provou ser o melhor modelo de alinhamento para seis categorias: *Organizations; Currencies; City; Time; Products and Services* e *Dates*, gerando os alinhamentos mais fiáveis utilizando o modelo pré-treinado: XLM-R e a *heuristic: intersect simmetrization*.

FastAlign foi classificado como segundo melhor alinhador, obtendo os melhores alinhamentos para as seguintes categorias: *Countries; Credit-Card; Address; Percentages; Username.* as restantes categorias (seis) foram divididas entre o eflomal e o AwesomeAlign, o que nos levou a descartá-los.

A análise dos resultados levou à conclusão de que o SimAlign tem um comportamento mais consistente do que os demais, com resultados de F-1 consistentemente muito altos.

4.2.3 Tarefa de Alinhamento do Conjunto de Dados da Área da Tecnologia

De acordo com os dados recolhidos, foi possível concluir que o AwesomeAlign produziu os melhores alinhamentos, seguido imediatamente pelo SimAlign, que apenas para a categoria *Organizations*, utilizando o BERT e o itermax heurístico, não atingiu um valor F-1 de 1 (valor máximo). No entanto, é importante ter em mente que o conjunto de dados utilizado para o alinhamento incluía apenas 360 frases, traduzindo-se numa presença reduzida de EM. A falta de EM suficientes representando uma categoria pode explicar o F-1 obtido por AwesomeAlign e SimAlign, independentemente das configurações utilizadas.

Quanto ao modelo que produziu melhores resultados de alinhamento, todas as configurações do AwesomeAlign produziram alinhamentos com resultados F-1 de 1, à semelhança do SimAlign, indicando que a tarefa era trivial de resolver.

No que diz respeito ao FastAlign, o seu desempenho ainda é inferior ao dos outros alinhadores, sendo para algumas categorias o alinhador que apresentou os piores resultados de alinhamento. Presumivelmente, o fraco desempenho de FastAlign poderá estar relacionado com a sua dificuldade em lidar com palavras que lhe são desconhecidas, que tipicamente são instâncias de Entidades Mencionadas. As abordagens com os modelos pré-treinados são mais robustas quando confrontadas com esta questão, uma vez que operam ao nível das sub-palavras e são expostas a conjuntos de dados muito maiores durante a fase de treino.

5. Conclusões e Trabalho Futuro

Com este trabalho, concentramo-nos em dar uma visão geral sobre a importância central das Entidades Mencionadas numa perspetiva linguística e histórica, destacando a sua relevância dentro de um cenário de tradução automática. Além disso, em conjunto com a equipa de PLN, pudemos testar sistemas de alinhamento diferentes usados para a criação semiautomática de conjuntos de treino e teste multilingues, recorrendo à projeção de EM num *corpus* paralelo. Com os resultados da investigação, foi possível substituir o sistema de alinhamento mais comumente usado pela Unbabel, Fastalign, pelo SimAlign. Ao fazê-lo, garantimos uma integração fiável desta técnica para a criação de conjuntos de treino e teste de EM multilingues. As tarefas de anotação manual, realizadas ao longo das experiências, permitiram-nos realçar o facto de determinadas EM poderem desempenhar papéis ambíguos e serem responsáveis por assimetrias de anotação, necessitando uma discussão de critérios entre anotadores e inclusão de diretrizes de casos ambíguos nas diretrizes de anotação.

Também vemos possibilidades de os sistemas de REM poderem ser usados para alavancar as memórias de tradução em empresas de TA. Acreditamos que a identificação de EM em memórias de tradução, seguida pela substituição por expressões equivalentes, levará a um aumento de correspondências nas memórias de tradução existentes, promovendo resultados de TA mais precisos, diminuindo a necessidade de pós-edição, logo diminuindo custos acessórios.

Por último, uma nota final sobre a contribuição do nosso trabalho para o módulo de anonimização. O trabalho realizado reflete, em última análise, melhorias no módulo de anonimização, crucial para qualquer empresa que cumpra os Princípios da Inteligência Artificial Responsável. Os fundamentos e abordagens desenvolvidos no âmbito do nosso projeto relativamente à identificação e anonimização da Informação Pessoal Identificável já foram implementados pelo Projeto MAIA (Multilingual AI Agent Assistant), permitindo o processamento e partilha de informação de uma forma segura. Como tal, continuaremos o nosso trabalho relativo à tarefa REM, com especial ênfase na etapa de anonimização.

6. Agradecimentos

Este trabalho foi apoiado por fundos nacionais em Portugal através da Fundação para Ciência e a Tecnologia (FCT), com a referência UIDB/50021/2020 e através da FCT e Agência Nacional de Inovação com o Projecto Multilingual AI Assistants (MAIA), número 045909.

7. Referências

Agerri, R., Chung, Y., Aldabe, I., Aranberri, N., Labaka, G., & Rigau, G. (2018, May). Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (pp. 3529-3533).

Akbik, A., Blythe, D., & Vollgraf, R. (2018, August). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638-1649).

Chinchor, N., & Robinson, P. (1997, September). MUC-7 named entity task definition. In *Proceedings of the SeventhConference on Message Understanding* (Vol. 29, pp. 1-21).

Chung, Y. L. (2017). Automatic generation of named entity taggers leveraging parallel corpora. Stanford University.

Data Protection Act, 2018. *Data Protection Act 2018*. [online] GOV. U.K. Disponível em: <https://www.gov.uk/government/collections/data-pro tection-act-2018>

Finkel, J. R., Grenager, T., & Manning, C. D. (2005, June). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)* (pp. 363-370).

Goyal, A., Kumar, M., Gupta, V (2017). Named Entity Recognition: Application, Approaches and Challenges. *International Journal of Advance Research in Science and Engineering. 6 (*10).

Harper, C., & Gorham, R. B. (2020). From Text to Map: Combing Named Entity Recognition and Geographic Information Systems. *Code4Lib Journal*, (49).

Joseph, C. (2019). *What are the Benefits of delivering excellent customer service?*. Chron. Disponível em https://smallbusiness.chron.com/benefits-delivering-e xcellent-customer-service-2086.html. Consultado em 18, janeiro de 2020.

Kenny, D. (2018). The Routledge Handbook of Translation and Philosophy. *Chapter 26: Machine Translation.* London, Routledge.

Mansouri, A., Affendey, L., Mamat, A. (2008). Named Entity Recognition Approaches. In *International Journal of Computer Science and Network Security. 8* (2).

Makhoul, J., Kubala, F., Scwartz, R., and Weischedel, R. (1999). Performance measure for information extraction. In *Proc. of the DARPA Broadcast News Workshop, Herndon, VA.*

Martins, F. T. André, Graça, J., Dimas, P., Moniz, H., Neubig, G. (2020). Project MAIA: Multilingual AI Agent Assistant. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, (pp. 495-496).

Modrzejewski, M., Exel, M., Buschbeck, B., Ha, T. L., Waibel, A. (2020, November). Incorporating external annotation to improve named entity translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 45-51).

Mota, P., Cabarrão, V., Farah, E. (2022). Fast-Paced Improvements to Named Entity Handling for Neural Machine Translation. In *Proceedings of EAMT*.

NER Annotation Guidelines. (2020). Unbabel´s Internal Company Document.

Nouvel, D., Ehrmann, M., Rosset, S. (2016). Named entities for computational linguistics. ISTE.

Qin, Y., Lin, Y., Takanobu, R., Liu, Z., Li, P., Ji, H., & Zhou, J. (2020). Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning. arXiv preprint arXiv:2012.15022.

Sang, E. F., De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.
Santos, C. N. D., Guimarães, V. (2015). Boosting named entity recognition with neural character embeddings. arXiv preprint arXiv:1505.05008.
Stengel-Eskin, E., Su, T. R., Post, M., Van Durme, B. (2019). A discriminative neural model for cross-lingual word alignment. arXiv preprint arXiv:1909.00444.
Silva, R. (2021). *Anotação de Entidades Mencionadas.* Relatório de Estágio da Licenciatura em Tradução, Faculdade de Letras da Universidade de Lisboa.
Tarcar, A. K. Tiwari, A., Dhaimodker, V. N., Rebelo, P., Desai, R., & Rao, D. (2019). Healthcare NER models using language model pretraining. arXiv preprint arXiv:1910.11241.
Tian, T., Dinarelli, M., Tellier, I., & Cardoso, P. D. (2016, May). Domain adaptation for named entity recognition using crfs. In LREC 2016, pp 560-565.
Wang, Z., Mayhew, S., & Roth, D. (2019). Cross-lingual ability of multilingual bert: An empirical study. arXiv preprint arXiv:1912.07840.
Wen, C., Chen, T., Jia, X., & Zhu, J. (2021). Medical Named Entity Recognition from Un-labelled Medical Records based on Pre-trained Language Models and Domain Dictionary. *Data Intelligence*, *3* (3), 402-417.
Wu, S., & Dredze, M. (2019). Beto, Bentz, becas: The surprising cross-lingual effectiveness of BERT. 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) arXiv:1904.09077.