Anotação de Entidades Mencionadas na área do *Gaming*

Rita Silva[1], *Vera Cabarrão*[2,3], *Sara Mendes*[1]

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

Abstract

This paper aims to analyse the effects of including gaming entities in the performance of the NER system, for the English language and in a machine translation industrial context of customer support content. To identify and classify gaming entities (by the Named Entity Recognition (NER) model), three new categories were created and added to the already used annotation typology: GAME NAME, GAME FEATURE and GAME CURRENCY. A set of reference annotations (gold standard) was also developed, allowing not only the training of the NER system but also the evaluation of its performance and accuracy in a more objective way, namely by counting the number of entities that the system identifies and categorises correctly. In the scope of this work, 6618 sentences from 7 gaming clients were manually annotated, constituting the gold standard which was then used to train and evaluate the NER system. The objective of the experiments was to assess whether the existing NER system improved its performance when trained with the gold standard created specifically for the gaming domain and if it could handle the new gaming categories added to the typology by identifying and categorizing them correctly. The results of both experiments were auspicious and positive, demonstrating the relevance of greater investment in domain-specific entity recognition, namely in the context of customer service text processing.

Keywords: Named Entities, Named Entity Recognition, *gaming*, annotation.

Palavras-chave: Entidades Mencionadas, Reconhecimento de Entidades Mencionadas, *gaming*, anotação.

1. Introdução

O presente artigo foi realizado em colaboração com a Unbabel, uma *start-up* portuguesa que oferece serviços de tradução quase em tempo real, combinando tradução automática (TA) com pós-edição humana. A plataforma de tradução da Unbabel permite às empresas fornecer serviços de apoio ao cliente multilingues, eliminando barreiras comunicacionais e linguísticas ao conjugar inteligência artificial com editores humanos em tempo real, produzindo traduções de alta qualidade de forma rápida, eficiente, acessível e inteligente, que permitem melhorar igualmente a satisfação dos seus clientes.

O Processamento Automático de Línguas Naturais (PLN) é considerado uma disciplina dentro da área da Inteligência Artificial (IA) que utiliza algoritmos para determinação de propriedades da língua natural, de modo a que os computadores possam compreender o que os seres humanos escrevem ou dizem. O Reconhecimento de Entidades Mencionadas (REM) é uma das principais tarefas do PLN e consiste na análise de uma frase ou de um segmento de texto de forma a encontrar entidades que possam ser classificadas em categorias específicas (Chiticariu et al., 2010), como nome, organização, produtos, horas, unidades monetárias, entre outras. As Entidades Mencionadas (EM) são unidades textuais de natureza variada e com uma forte componente referencial (De Oliveira, 2010), importantes para o bom conhecimento e compreensão de um texto. Estas podem corresponder a nomes próprios (e.g. "Rita Silva", "João", "Santos"), expressões temporais (e.g. duração: "10 dias úteis"; datas: "21/06/2021"; épocas festivas: "Ano Novo", entre outras) e numéricas (unidades monetárias:

"100€" ou "100 euros"; referências: "REF 575838543"; números de cartões de crédito, entre outros). A tradução de EM requer abordagens e métodos diferentes daqueles que são usados na tradução de outros tipos de elementos textuais, na medida em que as EM são bastante complexas, encontrando-se em vários contextos textuais, em diferentes posições na frase, muitas vezes não sendo possível desambiguá-las através do contexto (da Silva Romao, 2007). Logo, a tradução incorreta de entidades mencionadas pode afetar a integridade morfossintática geral das frases e interferir na desambiguação do significado das palavras no texto de partida, conduzindo a uma tradução pouco natural e ambígua (de difícil compreensão) ou à necessidade de uma extensa pós-edição (Babych et al., 2003).

Tendo em conta que a Unbabel trabalha com empresas de várias áreas, surgiu a necessidade de incluir entidades mencionadas de domínios específicos na tipologia de EM usada na empresa. Assim, o trabalho apresentado e discutido neste artigo consistiu no tratamento de EM da área do *gaming*. No entanto, já foram realizados outros trabalhos cujo alvo eram entidades dos domínios de entrega de comida e turismo (Menezes, no prelo). Em ambos os projetos, o objetivo foi contribuir para um melhor desempenho do sistema de REM. Este foi retreinado e testado com novos dados anotados e com novas entidades, de forma a verificar a viabilidade da inclusão de entidades mencionadas de domínios específicos.

A indústria do *gaming* é das maiores do mundo e encontra-se em constante desenvolvimento. Esta engloba atualmente uma grande variedade de produtos e, consequentemente, as entidades mencionadas desta área revelam uma grande heterogeneidade. Desta forma, este trabalho visou contribuir para a melhoria do desempenho do modelo de REM da Unbabel ao incluir entidades desta área (até então inexistentes) na sua tipologia.

2. Sistemas de REM

Com a revolução digital surgiram novas formas de comunicação e partilha de conhecimentos. Consequentemente, houve um aumento no número de publicações, documentos e de dados diversos. Neste contexto, os sistemas de processamento de dados permitem não só armazenar uma grande quantidade de dados, mas também tratá-los e aproveitá-los. O objetivo destes sistemas é essencialmente extrair e estruturar a informação, de modo a desenvolver e a explorar automaticamente o conhecimento. A tarefa de extração de informação, formalizada no final dos anos 80, tenta responder a esta necessidade de tratamento de informação, concentrando-se no reconhecimento de pedaços de informação em textos e relacionando-os uns com os outros (Nouvel et al, 2016). Neste contexto, o PLN, enquanto abordagem computadorizada de análise e tratamento de texto, tem como principais tarefas a separação do texto em frases e/ou *tokens*, o reconhecimento de entidades mencionadas, a anonimização de dados e a normalização da formatação. O reconhecimento e a extração de entidades mencionadas são úteis para recolher informação textual, sendo utilizados em várias áreas do conhecimento.

Nos últimos anos, os sistemas de reconhecimento e extração automática de entidades mencionadas tornaram-se numa área de investigação, com um número considerável de estudos sobre o desenvolvimento destes sistemas. Os sistemas de REM têm sido analisados e apresentados em diversos fóruns, com relevo para a ACL (*Association for Computational Linguistic*), o MUC (*Message Understanding Conference*), o CoNLL (*Conference on Computational Natural Language Learning*) e o HAREM (Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas).

A Associação para a Linguística Computacional (ACL) trabalha com questões computacionais envolvendo a língua humana, no campo da linguística computacional (ou processamento de língua natural) e tem promovido várias conferências com o objetivo de reunir os vários intelectuais da área (e subáreas) da linguística computacional na discussão de vários temas, possibilitando assim a partilha de experiências com uma grande comunidade de especialistas.

As MUC constituíram uma série de conferências (sete no total) dedicadas à avaliação em extração de informação, que se realizaram entre 1987 e 1998. A sexta conferência de avaliação do MUC, organizada em

1995, foi o primeiro evento de avaliação a incluir uma tarefa independente de avaliação de sistemas de REM (Cardoso et al, 2006). Para o MUC, a tarefa REM teve como objetivo classificar as entidades mencionadas nas seguintes categorias e tarefas (Cardoso et al, 2006): entidades de nomes próprios (ENAMEX); expressões temporais (TIMEX); e expressões numéricas (NUMEX). A categoria ENAMEX era composta pelas etiquetas PERSON (pessoa), ORGANIZATION (organização) e LOCATION (local); a TIMEX pelas etiquetas TIME (hora) e DATE (data); e NUMEX pelas etiquetas MONEY (moeda) e PERCENT (percentagem).

Por sua vez, a CoNLL (Conference on Computational Natural Language Learning) teve como objetivo promover a avaliação em diversas áreas do PLN. O primeiro evento remonta a 1999, mas foram os eventos de 2002 e de 2003 que se focaram no reconhecimento de entidades, para encorajar a investigação em sistemas de REM independentes da língua (Carvalho et al, 2012).

A iniciativa para a realização de um evento de avaliação de REM centrado na língua portuguesa surgiu em 2005. Deste evento saiu a primeira avaliação conjunta de sistemas de REM, denominada HAREM (Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas).

Pondo em evidência o dinamismo e a relevância da investigação neste domínio, notem-se vários trabalhos de investigação e projetos conduzidos quer pela comunidade académica quer pela indústria para desenvolver os sistemas de REM e alargar o escopo de entidades a ser reconhecidas em diversos contextos. Dois exemplos mais recentes destas iniciativas são o sistema Core NLP[1], desenvolvido pela Universidade de Stanford, e o sistema Spacy[2].

Na área do *gaming*, foco deste trabalho, os sistemas de REM são utilizados essencialmente para extrair diferentes tipos de entidades mencionadas e classificar documentos textuais a partir de textos heterogéneos na internet. A maioria das abordagens dentro deste domínio baseiam-se no projeto de investigação europeu *Realizing and Applied Gaming*.

Ecosystem[3] (RAGE). Este é um projeto que pretende criar um mecanismo de transferência de tecnologia e de saber-fazer para acelerar a inovação e o crescimento dos *Serious Games* ou *Applied Games* na Europa. Este tipo de jogos, aplicados ou sérios (tradução literal), são mais direcionados para fins educativos, de investigação científica, planeamento urbano, entre outros, e não tanto para puro entretenimento. No entanto, o grande desafio do RAGE é a organização e tratamento de informação devido essencialmente ao grande e constante fluxo de documentos com diferentes formatos publicados diariamente na internet. Para superar este desafio, tem sido proposto a utilização de um sistema de REM (Tamla et al., 2020). No entanto, neste trabalho apresenta-se uma abordagem diferente. Pretende-se aplicar o sistema de REM à área do *gaming* e da tradução automática, mais especificamente à tradução de textos (e-mails, perguntas frequentes – FAQs) de serviços de apoio ao cliente de vários tipos de jogos direcionados para o entretenimento.

2.1 REM em Tradução Automática

A tradução de entidades mencionadas, incluindo nomes próprios, expressões temporais e numéricas, é muito importante no processamento multilingue de Língua Natural, nomeadamente na Tradução Automática (TA), requerendo abordagens e métodos diferentes dos que são usados na tradução de outros tipos de palavras. As entidades mencionadas são bastante complexas: existem em grande número, em vários contextos textuais e posições na frase; muitas são específicas de um domínio ou língua, não se encontrando em dicionários bilíngues; nem sempre é possível distinguir, através do contexto, dois (ou mais) tipos de entidades diferentes (Al-Onaizan et al., 2002). Assim, as entidades mencionadas podem gerar problemas aos sistemas de tradução automática, conduzindo frequentemente a uma tradução pouco natural e ambígua ou à necessidade de uma extensa pós-edição (Babych et al., 2003).

[1] https://stanfordnlp.github.io/CoreNLP/
[2] https://spacy.io/
[3] http://rageproject.eu/ https://cordis.europa.eu/project/id/644187

Nem sempre é fácil para um sistema de REM determinar se uma palavra ou sequência de palavras é realmente uma entidade mencionada, ou, até mesmo, classificar uma EM, especialmente em frases ou contextos ambíguos.

Por exemplo, "Abrantes" pode referir-se a um nome próprio (apelido, por exemplo) ou a um local/cidade, só sendo possível determinar se se trata de um ou outro caso em contexto. No entanto, compreender e interpretar o contexto em que uma entidade mencionada pode aparecer, desambiguar o significado de palavras e frases e determinar relações entre palavras são das tarefas mais complexas para a IA.

Adicionalmente, certos formatos de texto, pontuação, símbolos específicos, emojis, entre outros, podem dificultar a tarefa do sistema de REM e obrigar a abordagens diferentes em termos de processamento.

(1) S.E.C. (ORG) chief Mary Shapiro (NAME) to leave Washington (LOC) in December.
(frase retirada de Data Community DC, 2013)

Esta frase, com três entidades mencionadas, mostra bem a complexidade na classificação e reconhecimento de entidades. Primeiramente, S.E.C. é um acrónimo para *Securities and Exchange Commission*, que se trata de uma organização. Para além disso, as duas palavras "Mary Shapiro" indicam uma única pessoa e a palavra Washington, neste caso, é um local e não um nome. Note-se também que o *token* "chief" não está incluído na etiqueta NAME, embora pudesse estar. Assim, nesta frase é difícil perceber se "S.E.C. chief Mary Shapiro" é uma única entidade ou várias, sendo que neste último caso é necessária a utilização de mais do que uma etiqueta (Data Community DC, 2013).

Assim, uma das primeiras tarefas será definir os critérios de delimitação de EM e da sua classificação, até porque os sistemas de reconhecimento de entidades mencionadas dependem fortemente dos exemplos com que são treinados, pelo que precisam de ser constantemente aperfeiçoados. Se o sistema não estiver treinado para reconhecer e categorizar alguma(s) da(s) entidade(s) presente(s) no texto, estas, ao serem consideradas pelo sistema de tradução automática, podem ser incorretamente traduzidas, ou podem até ser completamente distorcidas pelo sistema. Estas incorreções levam a uma tradução pouco natural ou até mesmo incompreensível, o que afeta, sem dúvida, a qualidade do texto de chegada produzido. Para evitar ou reduzir a ocorrência destes erros de tradução, o sistema REM é treinado com uma vasta quantidade de dados anotados manualmente com vários tipos de entidades mencionadas.

Apesar dos recentes avanços nesta área tão essencial do PLN, os sistemas de REM ainda têm muito espaço para evoluir nas diversas áreas em que são aplicados. A indústria e a academia podem e devem continuar a trabalhar para a constante melhoria e aperfeiçoamento deste tipo de sistemas. O presente trabalho visa precisamente contribuir para a melhoria do sistema de REM no contexto da TA.

2.2 REM e domínios

Atualmente, os profissionais e investigadores de várias áreas utilizam diferentes métodos e sistemas de reconhecimento de entidades mencionadas: o REM baseado em regras; o REM baseado na aprendizagem automática (utilizado na Unbabel); e o híbrido (Mansouri et al, 2008).

O REM manual, baseado em regras, centra-se na extração de nomes, utilizando um conjunto de regras que procuram reconhecer padrões regulares nas EM. Geralmente, estes sistemas englobam gramática (por exemplo, categorias morfossintáticas), sintaxe (por exemplo, ordem de palavras) e características ortográficas (por exemplo, letras maiúsculas) em combinação com dicionários disponíveis (Mansouri et al, 2008).

Os modelos de aprendizagem automática, por sua vez, utilizam modelos estatísticos e algoritmos de aprendizagem automática para identificação e classificação de entidades, procurando padrões e relações no texto de forma automática. Estes modelos podem ser subdivididos em dois: os modelos de aprendizagem automática supervisionados e os de aprendizagem automática não supervisionados. Os modelos de aprendizagem supervisionados têm intervenção humana e recorrem a anotações douradas, ou seja, um conjunto

de dados anotados manualmente e utilizados para treino do sistema de REM (este aprende com base nas anotações manuais). Esta coleção dourada (o conjunto de anotações) serve também como referência para a medição e avaliação do desempenho dos modelos de REM. Estes modelos exigem a preparação destes dados de treino categorizados para construir um modelo estatístico, não podendo alcançar um bom desempenho sem uma grande quantidade de dados (Mansouri et al, 2008). Já os modelos não supervisionados aprendem através de padrões no texto. Estes não utilizam dados anotados manualmente, não havendo a necessidade de preparar dados anotados para treinar e construir um modelo eficiente (Mansouri et al, 2008).

Já o modelo híbrido, como refere o nome, consiste numa junção entre o modelo baseado em regras e o de aprendizagem automática.

Os sistemas de REM têm diversas aplicações em várias áreas como a tradução, a Biomedicina e a História e ao nível da anonimização dos dados pessoais (PII).

Na área da tradução, mais especificamente em tradução automática, as entidades mencionadas criam problemas graves de tradução. Desta forma, os sistemas de REM permitem uma melhor identificação e consequente tradução de entidades mencionadas.

Já na biomedicina, à medida que o conhecimento dentro da área foi crescendo e, consequentemente, a literatura (por exemplo, documentação e publicações científicas), aumentou a necessidade de se utilizarem ferramentas robustas a fim de organizar, recuperar e fazer curadoria de toda a informação. Desta forma, é clara a utilidade de um sistema de reconhecimento de entidades mencionadas enquanto passo importante e necessário para progressos significativos na área (Huang et al., 2019). Entre os benefícios da implementação de sistemas REM nesta área estão: a maximização da independência e competitividade, com o aumento do desempenho relativamente aos sistemas de organização de informação já existentes; a extração de diferentes entidades biomédicas como genes, fenótipos, drogas, doenças, entidades químicas, entre outros. Neste contexto, o REM serve de padrão aos investigadores para comparar técnicas e possibilita aos profissionais encontrar entidades em grandes quantidades de texto (Leaman et al., 2008).

Por sua vez, em áreas como a História, os sistemas REM permitem a identificação e extração de topónimos assim como de informação espacial mencionada em conjuntos de textos históricos. Consequentemente, ajudam a ultrapassar alguns dos maiores desafios envolvidos no tratamento de *corpora* históricos. Possibilitam igualmente a aplicação da análise de informação espacial, assim como da cartografia dos espaços, com sistemas de informação geográfica, permitindo novas formas de identificação e análise de referências a nomes de locais em *corpora* literários e históricos (Won et al., 2018).

3. Metodologia

A indústria do *gaming* é das maiores do mundo e tem vindo a ganhar cada vez mais importância e relevância no universo das tecnologias. Esta engloba atualmente uma grande variedade de produtos que vão desde os vários tipos de videojogos (para as mais diversas plataformas, desde o computador ao *smartphone*) até aos fabricantes de equipamentos e acessórios. Adicionalmente, envolve uma cadeia enorme de profissionais, desde criadores de jogos a fabricantes, contando inclusivamente com jogadores profissionais e competições de alto nível. A grande diversidade que esta área apresenta reflete-se na natureza das suas entidades mencionadas, havendo, portanto, a necessidade de uma abordagem especial e de um tratamento específico e cauteloso das entidades mencionadas de *gaming*.

3.1 Identificação de entidades de *gaming*

Primeiramente, foi realizada uma análise exaustiva de um *corpus* composto por 9532 dados textuais (em inglês) de empresas da área do *gaming*, relacionados com o apoio ao cliente (*chats* e FAQs). O objetivo desta análise foi identificar entidades específicas da área em estudo tendo em conta vários fatores como a frequência e a tipologia. Depois do levantamento de todas as ocorrências de entidades mencionadas no *corpus*, procedeu-se

à divisão das entidades em grupos e à definição de etiquetas para cada grupo de entidades. Desta forma, foram criadas três novas etiquetas, que não integravam a tipologia de Entidades Mencionadas da Unbabel: GAME NAME, GAME FEATURE e GAME CURRENCY. Posteriormente, elaborou-se uma lista de diretrizes de forma a integrar as novas etiquetas de *gaming* na tipologia de REM da Unbabel. Por último, foi realizada a anotação de entidades por cada cliente utilizando uma ferramenta de anotação, criando assim uma coleção dourada. No total, foram anotadas 6167 frases de 6 clientes da área do *gaming* com produtos bastante diversificados: três são jogos educacionais e de fantasia, um relaciona-se com a área do desporto e dois constituem jogos realísticos e de guerra. Estas três novas etiquetas foram acrescentadas à tipologia da Unbabel e divergem das restantes na sua natureza. Tendo em conta a sua complexidade e ambiguidade, foi necessário elaborar um conjunto de diretrizes com o objetivo de clarificar e auxiliar na identificação destas entidades.

A etiqueta GAME NAME aplica-se apenas a nomes gerais de jogos e às suas edições especiais.

(2) I bought the Premium Edition of war gameX 3 on game platform$_{\text{ORG}}$ some time ago.
[Premium Edition of war gameX 3]$_{\text{GAME NAME}}$

Já a etiqueta GAME FEATURE aplica-se a determinados componentes e elementos dos jogos, portanto, inclui elementos que influenciam diretamente a jogabilidade e outras mecânicas de jogo, tais como pacotes especiais, mapas de jogo, *boosters* (simplificam a jogabilidade), recompensas, personagens, cartas de jogadores, missões ou níveis e modos de jogo. Estes últimos podem restringir ou alterar o comportamento das ferramentas disponíveis e/ou estabelecer regras de jogo diferentes. Assim, um jogo que inclui vários modos terá configurações diferentes para cada um deles e isto irá, consequentemente, alterar a forma como uma determinada parte do jogo é jogada.

(3) In January I spent 19 coinX to purchase 3 ballX, 2 of them were then used immediately on level 4721.

[ballX]$_{\text{GAME FEATURES}}$ [19 coinX]$_{\text{GAME CURRENCY}}$ [4721]$_{\text{GAME FEATURES}}$

(4) A family member sold my players in sports gameF MX and deleted my club without my authorization.

[MX]$_{\text{GAME FEATURES}}$ [sports gameF]$_{\text{GAME NAME}}$

(*MX* é um modo de jogo no jogo sports gameF).

A etiqueta GAME CURRENCY aplica-se à moeda específica usada num jogo. Geralmente, a moeda utilizada nos jogos engloba uma variedade de nomes, símbolos e difere de jogo para jogo. Em alguns jogos pode-se falar em 2 tipos de moeda: uma que é obtida durante o próprio processo do jogo (por exemplo, subindo de nível ou completando uma missão) e outra obtida através da troca de dinheiro real do jogador (por exemplo, para comprar funcionalidades e elementos de jogo).

(5) According to your account you already received gameX $ 2,500,00.

[gameX $ 2,500,00]$_{\text{GAME CURRENCY}}$

3.1.1 Casos ambíguos e dificuldades

A escolha e definição de etiquetas para cada grupo de entidades apresentou algumas dificuldades. As etiquetas criadas são abrangentes e muitas vezes revestem-se de ambiguidades e incertezas. Nem sempre é possível distinguir claramente, através do contexto, dois tipos de entidades diferentes e, consequentemente, associá-las a uma determinada etiqueta. Para reduzir as possíveis ambiguidades de uma etiqueta, foi necessário analisar com cuidado todos os contextos em que essa mesma entidade ocorria para depois os listar, definir e registar. No entanto, algumas etiquetas são mais problemáticas do que outras. Das 3 etiquetas definidas, GAME NAME é a que apresenta mais ambiguidades, uma vez que pode ser fácil confundir as edições especiais de um jogo com pacotes de expansão, que devem integrar a categoria GAME FEATURE. Para evitar erros é importante, então, saber que os pacotes de expansão são, geralmente, conjuntos de objetos que se podem comprar e adicionar ao jogo e que consistem unicamente em conteúdo adicional que afeta a jogabilidade (a maioria deles requer o jogo original para funcionar).

(6) Are you talking about the real game 4 Dress Pack?

[real game 4]$_{\text{GAME NAME}}$ [Dress Pack]$_{\text{GAME FEATURES}}$

Dentro da etiqueta GAME FEATURE foi também necessário identificar e clarificar o conceito de modos de jogo, uma vez que são facilmente confundíveis com edições especiais. Assim, foi criada uma lista de modos de jogo para cada cliente e foi elaborada uma definição pormenorizada do conceito de modos de jogo.

3.2. Tarefa de anotação

Após a definição das etiquetas, seguiu-se a anotação integral de entidades no *corpus*, ou seja, não apenas a anotação com as novas etiquetas de *gaming* criadas no âmbito deste trabalho (GAME NAME, GAME FEATURE e GAME CURRENCY), mas também com as restantes etiquetas já existentes na tipologia da Unbabel. Estas representam cerca de 27 EM que se dividem entre as 3 categorias apresentadas na conferência MUC, nomeadamente Enamex, Numex e Timex. Destacam-se, por exemplo, entidades como NAME (nome próprio), PRS (produtos e serviços), ORG (organizações), PHONE (número de telefone) e EMAIL (email).

Como se verifica na tabela 1, no total, foram anotadas 6618 frases de 7 clientes da área do *gaming* com conteúdos bastante diversificados. Relativamente às entidades, nas 6618 frases anotadas, há 178 entidades que se inserem na etiqueta GAME CURRENCY; 1489 na etiqueta GAME FEATURE; e 312 na GAME NAME.

Cliente 1	827
Cliente 2	660
Cliente 3	1254
Cliente 4	1631
Cliente 5	1152
Cliente 6	643
Cliente 7	451
Total	6618

Tabela 1: Número de frases anotadas por cada cliente

4. Avaliação dos sistemas de REM

Após a anotação, os dados foram utilizados em duas experiências diferentes. Na primeira experiência, pretendeu-se aferir se o sistema de REM já existente melhorava o seu desempenho ao ser treinado com a coleção dourada de clientes de *gaming*, um domínio bastante específico. As etiquetas alvo nesta experiência foram as de Produtos e Serviços e Organizações (PRS/ORG) e nomes próprios (NAME). Na segunda experiência, pretendeu- se testar se o sistema de REM conseguia lidar com as novas categorias de *gaming*, identificando-as e categorizando-as corretamente. Os resultados do sistema de REM foram avaliados segundo uma métrica padrão de desempenho (*Standard Performance Metric*): *Precisão, Cobertura, Medida-F e Exatidão* (Makhoul et al., 1999). Esta avaliação consiste na comparação do desempenho do sistema de REM face às anotações da coleção dourada, que são tidas como referência. Precisão corresponde à proporção de entidades que o sistema de REM devia ter identificado. Já a cobertura representa a proporção de entidades corretamente identificadas, mas incorretamente categorizadas pelo sistema. A Medida-F, por sua vez, corresponde à média harmónica da precisão e da cobertura, ou seja, à correlação entre o número de entidades identificadas e o número de entidades categorizadas corretamente. Por último, a Exatidão representa o desempenho geral do modelo, correspondendo à proporção de classificações corretas de todas as classificações que o sistema faz. Estas métricas são calculadas com as seguintes fórmulas:

$$Precisão = \frac{TP}{TP + FP}$$

$$Cobertura = \frac{TP}{TP + TN}$$

$$Medida - F = 2 * \frac{Precisão\ x\ Cobertura}{Precisão + Cobertura}$$

$$Exatidão = \frac{TP + TN}{TP + FP + FN + TN}$$

Nestas fórmulas, TP corresponde a *verdadeiros positivos*, ou seja, o número de entidades identificadas e categorizadas de forma correta pelo sistema REM (tendo como base de comparação as anotações humanas). Já FP significa *falsos positivos* e corresponde ao número de entidades que o sistema não etiquetou corretamente (por exemplo, o sistema identifica uma entidade como NAME em vez de ORG, a etiqueta correta). TN corresponde a *verdadeiros negativos* que são o número de entidades, palavras ou expressões corretamente não identificadas e categorizadas pelo sistema. FN corresponde a *falsos negativos*, ou seja, o número de entidades que o sistema não conseguiu identificar, mas que deveria ter identificado.

5. Análise dos resultados

5.1 Distribuição dos dados

Um dos focos da tarefa de anotação foi perceber se o sistema de REM melhora a sua prestação nas categorias NAME e PRS/ORG e se as novas etiquetas criadas para os clientes de *gaming* são corretamente identificadas, podendo, assim, contribuir para melhorar a qualidade das traduções.

Nas 6618 frases anotadas, há 730 entidades que se inserem na etiqueta NAME; 895 na etiqueta PRS/ORG; 78 na etiqueta GAME CURRENCY; 1489 na etiqueta GAME FEATURE; e 312 na etiqueta GAME NAME. No entanto, como se pode observar na tabela 2, há uma grande variação na distribuição de entidades por etiqueta.

Olhando, por exemplo, para o número de entidades etiquetadas como GAME FEATURE, o cliente 3 apresenta um número bastante elevado (1008), mas o cliente 2 já apresenta um número muito baixo (31). Esta distribuição de entidades por etiqueta mostra a heterogeneidade dos dados anotados por cliente (e do domínio em si). Esta heterogeneidade e complexidade verificou-se durante todo o processo de criação das novas etiquetas de *gaming*, na tarefa de anotação e nos próprios resultados do modelo de REM (discutidos na secção seguinte).

	GAME CURRENCY	GAME FEATURES	GAME NAME	NAME	PRS/ORG
Cliente 1	45	129	24	5	95
Cliente 2	13	31	30	27	102
Cliente 3	53	1008	132	25	153
Cliente 4	40	188	58	286	255
Cliente 5	18	64	44	204	228
Cliente 6	9	69	24	89	62
Cliente 7	12	53	5	21	81

Tabela 2: Número de Entidades identificadas e categorizadas pelo sistema em cada cliente de *gaming*

5.2 Resultados

No que diz respeito à primeira experiência, que visava avaliar se o sistema REM melhora o seu desempenho com a adição de mais dados anotados com entidades mencionadas, os resultados mostram que o novo modelo não difere muito do anterior quanto à medida-F nas entidades de produtos e serviços/organizações e nomes próprios. É visível uma ligeira descida de 0,1% nas etiquetas PRS/ORG, e um aumento de 0,4% na etiqueta NAME. Em ambas, a cobertura é ligeiramente mais baixa (73,9% e 93,2%) do que a precisão (74,8% e 94,4%). Estes resultados evidenciam que apesar de ainda haver entidades que o sistema não está a identificar, visível nos valores de cobertura mais baixos principalmente nas entidades de PRS/ORG, as que consegue identificar estão, na sua maioria, corretas (privilegia-se a precisão com valores mais elevados).

Estes resultados põem em evidência que a inclusão dos dados anotados com EM de *gaming* teve um impacto positivo no desempenho do sistema REM, uma vez que ao serem consideradas mais entidades, o problema de classificação aumenta em complexidade, pelo que a estabilidade do desempenho dos modelos é um aspeto a destacar. Uma das hipóteses para a diferença de 1% entre modelos pode dever-se ao facto de, apesar de terem sido anotados bastantes dados (ver tabela 1), o número de entidades de PRS/ORG e NAME não foi muito elevado (ver tabela 2), principalmente se tivermos em conta que o modelo original foi treinado com milhares de exemplos por entidade.

	TP	FP	FN	COBERTURA	PRECISÃO	Medida-F (modelo anterior)	Medida-F (modelo novo)
PRS/ORG	1382	248	464	73,9±2,55	74,8±1,46	79,9	79,5
NAME	640	37,5	46,6	93,2±2,13	94,4±1,01	93,7	93,8

Tabela 3: Resultados das etiquetas NAME e ORG/PRS (todos os dados anotados).

	TP	FP	FN	COBERTURA	PRECISÃO	Medida-F
GAME FEATURES	165	43.8	107	60.4±4.46	79.1±4.10	68.4
GAME NAME	58	12.9	17.2	77.2±8.16	81.7±5.71	79.2
GAME CURRENCY	24.1	5.7	14.1	62.6±13.2	80.4±11.4	69.5

Tabela 4: Resultados das etiquetas GAME NAME, GAME FEATURES, GAME CURRENCY (apenas os dados de *gaming*).

No que diz respeito à segunda experiência, testar a inclusão das novas entidades do domínio específico de *gaming*, os resultados mostram que ao contrário do que se verificou com as etiquetas PRS/ORG, cuja diferença entre a cobertura e a precisão era pouco significativa, estas 3 novas entidades apresentam uma maior discrepância entre estas medidas de avaliação do desempenho do sistema. A cobertura é significativamente mais baixa do que a precisão em todas as categorias, o que significa que o sistema não está a conseguir identificar uma parte importante das entidades a classificar, ainda que as que consegue identificar sejam maioritariamente identificadas e categorizadas corretamente. A categoria/etiqueta que apresenta melhores resultados é a GAME NAME, com uma cobertura de 77% e uma precisão de 82%, o que resulta numa medida-F de 79%. De seguida, surge GAME CURRENCY com a medida de exatidão mais elevada, medida-F de 69%, e, por último, GAME FEATURE, com o valor mais baixo de medida-F, 68%.

Tendo em conta a complexidade e diversidade destas entidades, nomeadamente o facto de surgirem em contextos bastantes diversificados nas frases e de variarem de cliente para cliente ainda que dentro da área de *gaming*, os resultados obtidos são bastante promissores. Estes mostram que ao acrescentar novas etiquetas à tarefa do modelo de REM, o que geralmente deteriora os resultados, os valores de classificação ou se mantêm ou melhoram 1%.

Ao observar a matriz de confusão em baixo (tabela 5), é notório que o sistema confunde algumas etiquetas de *gaming* entre si, bem como com outras etiquetas de fora deste domínio, nomeadamente NAME e PRS/ORG. Na matriz de confusão abaixo, é possível observar que a etiqueta GAME FEATURES é muitas vezes confundida com a etiqueta NAME, mas que as principais trocas se verificam com as etiquetas GAME NAME e PRS/ORG.

		Entidades identificadas e categorizadas pelo modelo de REM da Unbabel					
		PRS/ORG	NAME	GAME FEATURES	GAME CURRENCY	GAME NAME	Não identificadas
Categorização de referência (coleção dourada)	GAME NAME	94	0	22	3	580	10
	GAME CURRENCY	2	0	14	241	0	40
	GAME FEATURES	28	6	1654	31	17	333

Tabela 5: Matriz de confusão

Como se pode verificar, para a etiqueta GAME NAME, o sistema classificou incorretamente 94 entidades como PRS/ORG e 22 como GAME FEATURE. Esta última etiqueta é bastante ambígua não só porque o nome de um jogo é facilmente confundível com títulos de edições especiais e pacotes de expansão de jogos (que devem integrar a categoria GAME FEATURE), mas também porque o jogo é sempre o produto de uma empresa e muitos têm o nome dessa mesma empresa (ou um nome similar). Além disso, o facto de incluir nomes de personagens de jogos pode levar facilmente à confusão com a etiqueta GAME NAME.

Estes dados vêm, então, confirmar o que foi referido anteriormente relativamente à ambiguidade destas novas etiquetas (especialmente GAME NAME e GAME FEATURE) devido à pluralidade de produtos oferecidos pelos 7 clientes cujos dados foram anotados. Um anotador humano, por vezes, tem dificuldade em distinguir claramente, através do contexto, dois tipos de entidades diferentes e classificá-las corretamente. Tal mostra que a desambiguação do significado de uma entidade é uma tarefa particularmente desafiante para um sistema automático. Assim, pode-se concluir que, apesar de estes resultados serem promissores, é necessário anotar mais dados de *gaming* com estas etiquetas e voltar a treinar o modelo de REM para se poder verificar se o aumento de informação de treino permite atingir números de medida-F próximos dos observados para as categorias de EM já consideradas na tipologia da Unbabel que se verificam na tabela 4 para as etiquetas NAME e PRS/ORG.

6. Conclusões

O principal objetivo deste trabalho consistiu em analisar os efeitos da definição de entidades mencionadas específicas do domínio de *gaming* na tipologia de EM da Unbabel e no desempenho do modelo de REM em inglês já existente na empresa. Assim, para este propósito foram identificadas entidades presentes num *corpus* de textos de clientes da área do *gaming*, procedendo-se depois à categorização das entidades recolhidas, à definição de novas etiquetas para cada classe identificada e à criação de uma coleção dourada por meio de uma tarefa de anotação manual dos dados. Após a anotação, os dados foram utilizados tanto para aferir se o sistema de REM já existente melhorava o seu desempenho com entidades previamente consideradas pelo modelo,

tendo- se analisado o desempenho na classificação das entidades de NAME e PRS/ORG, ao ser treinado com a coleção dourada criada pela anotação de textos de clientes de *gaming*, como para testar se o sistema de REM conseguia lidar com as novas categorias de *gaming* identificadas (GAME NAME, GAME FEATURE e GAME CURRENCY), identificando-las e categorizando-las corretamente.

Os resultados finais de ambas as experiências são bastante promissores. Relativamente à primeira tarefa, os resultados mostram que o modelo de REM da Unbabel é bastante estável, não sendo significativamente afetado pela complexificação do problema de classificação decorrente da introdução de novos tipos de EM a classificar. Relativamente às EM previamente consideradas na tipologia da Unbabel, os resultados evidenciam que, apesar de ainda haver entidades que o sistema não está a identificar, principalmente nas entidades de PRS/ORG, as que consegue identificar estão, na sua maioria, corretas. Os resultados da segunda experiência revelam, mais uma vez, que o modelo de REM é eficiente, mas que, ao contrário do que se verificou na experiência anterior com as etiquetas PRS/ORG, as novas etiquetas de *gaming* apresentam uma maior discrepância entre a cobertura e a precisão, com o modelo a evidenciar um pior desempenho em termos de precisão, facto que aponta para a possível necessidade de aumentar a quantidade de dados de treino disponíveis para estas EM. Note-se, no entanto, que, ainda que nem todas as entidades estejam a ser identificadas pelo modelo de REM, a grande maioria das que são identificadas são corretamente classificadas.

Em termos de trabalho futuro, a ambiguidade da etiqueta GAME FEATURE observada põe em evidência a relevância de se trabalhar numa análise mais fina da mesma, uma vez que é muito mais abrangente do que as restantes. O facto de incluir, por exemplo, nomes de personagens de jogos pode levar facilmente à confusão com a etiqueta GAME NAME. Relativamente à ambiguidade entre as etiquetas GAME NAME e PRS/ORG, a confusão entre estas duas etiquetas poderá ser reduzida através da anotação de mais dados de treino para o sistema.

7. Referências

Al-Onaizan, Yaser, and Kevin Knight. Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002.

Babych, Bogdan, and Anthony Hartley. Improving machine translation quality with automatic named entity recognition. *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*. 2003.

Cardoso, Nuno Francisco Pereira Freire. *Avaliação de sistemas de reconhecimento de entidades mencionadas.* (2006).

Carvalho, Wesley Seidel. *Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina*. Diss. Universidade de São Paulo, 2012.

Chiticariu, Laura, et al. Domain adaptation of rule-based annotators for named-entity recognition tasks. *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010.

Da Silva Romão, Luís Carlos. *Reconhecimento de Entidades Mencionadas em Língua Portuguesa: Locais, Pessoas, Organizações e Acontecimentos*. Diss. MSc Thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2007.

Data Community DC. An Introduction to Named Entity Recognition in Natural Language Processing - Part 1. 2013, disponível em https://www.datacommunitydc.org/blog/2013/04/a-survey-of-stochastic-and-gazetteer-based-approaches-for- named-entity-recognition [consultado em 07/06/2021

De Oliveira, Diogo Correia. *Extraction and classification of named entities*. Diss. Master Thesis, University of Technia de Lisboa, 2010.

Huang, Ming-Siang, et al. *Revised JNLPBA corpus: A revised version of biomedical NER corpus for relation extraction task*. arXiv preprint arXiv:1901.10219 (2019).

Leaman, Robert, and Graciela Gonzalez. BANNER: an executable survey of advances in biomedical named entity recognition. *Biocomputing 2008*. 2008, pp. 652-663.

Luis Menezes. *Named Entities for Anonymization and Machine Translation. A case study on the importance of Named Entities for customer support*. Dissertação de Mestrado FLUL/Unbabel (no prelo).

Makhoul, John, et al. Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE* 88.8 (2000), pp. 1338-1353.

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

Mansouri, Alireza, Lilly Suriani Affendey, and Ali Mamat. Named entity recognition approaches. *International Journal of Computer Science and Network Security* 8.2 (2008), pp. 339-344.

Nouvel, Damien, Maud Ehrmann, and Sophie Rosset. *Named entities for computational linguistics*. ISTE, 2016.

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

Tamla, Philippe, Florian Freund, and Matthias L. Hemmje. *Supporting Named Entity Recognition and Document Classification in a Knowledge Management System for Applied Gaming*. KEOD. 2020.

Won, Miguel, Patricia Murrieta-Flores, and Bruno Martins. Ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora. *Frontiers in Digital Humanities 5* (2018): 2.