Proposta recebida em Abril 2023 e aceite para publicação em Junho 2023.

Avaliação no Desafio de Identificação de Personagens

Evaluation in DIP, the Character Identification Challenge in Portuguese

Roberto Willrich
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina

Diana Santos
Linguateca & ILOS, Universidade de Oslo

Resumo

A primeira edição do Desafio de Identificação de Personagens (DIP) foi uma avaliação conjunta de soluções computacionais para a identificação de personagens em textos literários, bem como a extração de características destas personagens e seus relacionamentos. Para esta avaliação, foi necessária a definição de uma metodologia de avaliação, incluindo a seleção de métricas adequadas ao problema da identificação de personagens em textos literários. Este artigo apresenta uma panorâmica de avaliação na área de identificação de personagens em textos literários, assim como as escolhas concretas que foram realizadas pela comissão organizadora do DIP. Estas escolhas resultaram na definição da metodologia de avaliação do DIP. O uso da metodologia de avaliação proposta é ilustrado pela avaliação da solução candidata submetida ao DIP. Ao final, são apresentadas críticas e sugestões de melhorias à metodologia de avaliação proposta.

Palavras chave

estudos literários computacionais, identificação de personagens, avaliação conjunta, métricas de avaliação

Abstract

The first edition of the Character Identification Challenge (DIP) was a joint evaluation of computational solutions for the identification of characters in literary texts, as well as the extraction of characteristics and relationships of these characters. For this evaluation, it was necessary to define an evaluation methodology, including the selection of appropriate metrics for the problem of identifying characters in literary texts. This article surveys the evaluation methods employed in character identification in literary works and presents the concrete choices done in DIP. These choices resulted in the definition of the DIP evaluation methodology. The use of the proposed evaluation methodology is illustrated by the evaluation of the candidate solution submitted to DIP. We end the paper with some critical remarks and suggestions for improvement.

Keywords

computational studies of literature, shared task, character identification, evaluation metrics

1. Introdução

As personagens, suas características (como género e profissão), relacionamentos familiares, profissionais, e suas interações, são elementos importantes para o estudo de vários tipos de textos literários, como romances e contos. Por exemplo, como apontado por Labatut & Bost (2019), no contexto da análise da narrativa, ou análise narratológica, vários trabalhos realizam a extração manual das personagens e suas interações. Esta operação manual é muito dispendiosa, principalmente caso seja realizada em um corpus muito grande. Neste sentido, diversas iniciativas buscam aplicar técnicas do domínio da inteligência artificial ou outras para a identificação de personagens em textos literários.

Esta identificação de personagens apresenta uma série de desafios. O primeiro é como classificar uma entidade no texto como personagem. Muitos trabalhos (Finkel et al., 2005; Elson et al., 2010; Lee & Yeung, 2012) consideram como personagem apenas pessoas mencionadas no texto. Outros trabalhos, como Valls-Vargas et al. (2014), consideram, além de pessoas, personagens de outras classes, como animais, objetos antropomórficos e criaturas fantásticas. O segundo desafio é que textos literários geralmente utilizam várias formas de mencionar uma mesma personagem, e não apenas pelo seu nome completo, prenome, ou uso apenas do sobrenome (acompanhado ou não da forma de um título pessoal) (de Does et al., 2017). Além destes, textos literários podem se referir a uma personagem de diversas outras formas, como diminutivos do nome, pseudônimos (alcunhas), uso apenas do título pessoal, pronomes pessoais, e descrições nominais (*o lavrador*, *a tal menina*, *o seu professor*). Esta característica torna a identificação de

DOI: 10.21814/lm.15.1.398

personagens em texto literário um grande desafio. Santos et al. (2023) abordam diversos destes desafios em profundidade.

Visando contribuir para o desenvolvimento da área de identificação de personagens, propomos o primeiro Desafio de Identificação de Personagens (DIP)[1] (Santos et al., 2022a). Trata-se de uma avaliação conjunta de soluções computacionais para identificação de personagens de textos literários, bem como a extração de características destas personagens e de seus relacionamentos. O DIP foi organizado de forma conjunta pela Linguateca,[2] NUPILL[3] da Universidade Federal de Santa Catarina (UFSC), Universidade Estadual do Maranhão (UEMA) e Universidade de Oslo (UiO).

Este artigo tem por objetivo apresentar a metodologia de avaliação definida para aferir a qualidade das soluções submetidas ao DIP. Para tal, foi necessária a adoção de métricas de avaliação adequadas para aferir a qualidade de soluções de identificação de personagens, suas características e relacionamentos. De forma a ilustrar o uso da metodologia adotada no DIP, este trabalho apresenta a avaliação da solução submetida à primeira edição do DIP.

O restante deste artigo está organizado na forma que segue. A Seção 2 apresenta um enquadramento teórico acerca das principais etapas envolvidas no processo de identificação de personagens e das métricas clássicas de avaliação. A Seção 3 descreve diversas propostas de solução para identificação (semi-)automatizadas de personagens, suas características e relacionamentos de diferentes tipos. Esta descrição foca principalmente nos processos e métricas de avaliação adotados por estes trabalhos. Em seguida, as Seções 4 e 5 detalham nossa experiência na primeira edição do DIP e a metodologia de avaliação adotada. De forma a ilustrar o uso desta metodologia, a Seção 6 apresenta uma avaliação do sistema participante do DIP. Na sequência, a Seção 7 apresenta uma análise crítica da metodologia adotada, e aponta melhorias a serem realizadas. Finalmente, a Seção 8 apresenta as conclusões e as perspectivas futuras deste trabalho.

2. Enquadramento teórico

Esta seção visa revisar alguns conceitos importantes para o entendimento do problema de identificação de personagens literárias, bem como apresentar as métricas clássicas de avaliação de técnicas de extração de conhecimento envolvendo personagens, suas características, relações pessoais e interações.

2.1. Processo de Identificação de Personagens

As técnicas de identificação de personagens em geral dividem o processo nas seguintes etapas (Labatut & Bost, 2019):

- **Detecção de menções a personagens**
  nesta etapa normalmente são usadas técnicas de Processamento de Linguagem Natural (PLN) do tipo Reconhecimento de Entidades Mencionadas (NER, *Named Entity Recognition*). Estas técnicas permitem identificar e classificar as entidades mencionadas em um texto escrito. Um desafio nesta etapa é classificar corretamente que uma entidade mencionada no texto literário seja classificada como personagem.
- **Co-identificação de personagens**
  esta etapa, também chamada de resolução de correferência, tem por objetivo identificar o conjunto de menções utilizadas no texto que co-identificam uma mesma personagem. A co-identificação de personagem é uma tarefa desafiadora, especialmente dada a multiplicidade de formas que o autor utiliza no texto para se referir às personagens.
- **Extração de características das personagens**
  uma vez identificadas as personagens, alguns trabalhos tentam extrair características destas personagens, como seu tipo (principal/secundária), género, profissão/ocupação/estatuto social, e faixa etária (p.e., bebê, jovem, adulto e idoso). Para isto, podem ser utilizadas técnicas de PLN.
- **Extração de relações entre personagens**
  esta etapa visa extrair os diversos tipos de relacionamentos entre personagens. No domínio de literatura, o objetivo é extrair relações que representem traços de personagens e elementos-chave da narrativa (Chu et al., 2021). Estas relações podem ser do tipo familiares (p.e., pai, mãe, filho, tio, etc.), de sentimentos (p.e., ama, odeia, amigo, inimigo, amante), profissionais (p.e., patrão/empregado, amo/escravo), ou outros tipos relacionados à narrativa (p.e., aliado, membro do clã, traído, assassino).

[1] https://www.linguateca.pt/DIP
[2] https://www.linguateca.pt/
[3] https://nupill.ufsc.br/

- **Extração de interações entre personagens**
  existem diversas iniciativas (incluindo Elson et al. (2010), Lee & Yeung (2012) e Agarwal et al. (2013)) que tentam extrair automaticamente as interações de diálogo entre personagens. O objetivo destes trabalhos é a extração da rede social (também chamada de rede conversacional) da obra, que representa as personagens por nodos e as suas interações por arestas ligando as personagens.

2.2. Métricas de Avaliação

A identificação de personagens literárias é de fato um processo de extração de conhecimento a partir de textos em linguagem natural (obras literárias). Para a avaliação de tais técnicas, é possível utilizar métricas clássicas, como: precisão (*precision*); revocação (*recall*), também chamada de abrangência, sensibilidade ou cobertura; acurácia (*accuracy*), também chamada de acerto; e a medida-F (*f-score* ou *f-measure*), que conjuga as duas primeiras. Estas métricas clássicas podem ser adaptadas para avaliar tarefas do tipo classificação ou de recuperação/recolha de informação.

2.2.1. Métricas do contexto de recuperação de informação

Em geral, os trabalhos relacionados ao DIP adotam as métricas clássicas aplicadas ao contexto da recuperação de informação. Neste contexto, a precisão (P) é determinada pela razão entre o número de instâncias relevantes (*IRel*) que foram recuperadas e todas as instâncias recuperadas (*IRec*), conforme Equação 1.

$$P = \frac{\{IRel\} \cap \{IRec\}}{\{IRec\}} \tag{1}$$

Instância no contexto de identificação de personagens pode se referir às personagens em si, ou então alguma outra informação, por exemplo, um certo fato, como um relacionamento entre duas personagens (por exemplo, Pedro é irmão de Maria).

A revocação R, por sua vez, é a fração de instâncias relevantes que são recuperadas com êxito, e calculada conforme Equação 2.

$$R = \frac{\{IRel\} \cap \{IRec\}}{\{IRrel\}} \tag{2}$$

A *medida-F* é determinada pela média harmônica de precisão e revocação, conforme expressa na Equação 3.

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

2.2.2. Métricas do contexto de classificação

Já no contexto de classificação, estas métricas são quantificadas usando os termos verdadeiros-positivos (VP), verdadeiros-negativos (VN), falsos-positivos (FP) e falso-negativos ou falsos-negativos (FN). As equações 4, 5, 6 definem a acurácia, precisão, revocação no contexto de classificação. Para cálculo da medida-F é utilizada a Equação 3.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{4}$$

$$P = \frac{VP}{VP + FP} \tag{5}$$

$$R = \frac{VP}{VP + FN} \tag{6}$$

2.2.3. Métricas de avaliação de resolução de correferências

Como visto na Seção 2.1, a resolução de correferências, no contexto da identificação de personagens, é o processo que permite co-identificar os diversos termos e expressões, chamadas de menções, pelos quais uma personagem (uma mesma entidade mencionada) pode ser referenciada no texto literário. O conjunto de menções a uma mesma personagem no texto é denominado de conjunto de equivalência. Em outras palavras, um conjunto de equivalência contém o grupo de menções no texto que são semanticamente equivalentes e se referem à mesma personagem.

Em geral, a avaliação de técnicas de resolução de correferências é baseada em métricas que permitem comparar os conjuntos de equivalência obtidos por um sistema automático com os conjuntos de equivalência anotados manualmente, chamada aqui de coleção dourada.

Existem diversas métricas de avaliação da resolução de correferências (Pradhan et al., 2014). A primeira métrica de avaliação de sistemas de resolução de correferências foi a MUC (Vilain et al., 1995). Esta métrica mede a eficiência da ligação entre as menções através da Medida-F MUC.

Afim de ilustrar a métrica MUC, considere aqui o seguinte exemplo de ligações de correferência de personagens de uma obra presentes

na coleção dourada: <Pedro – Pedro da Silva, Pedro da Silva – Dr. Silva, Maria – Maria da Silva, Maria da Silva – Mariazinha, Mariazinha – Sra. Silva>. Neste caso, são identificados dois conjuntos de equivalências, definido como K = {Pedro, Pedro da Silva, Dr. Silva}, {Maria, Maria da Silva, Mariazinha, Sra. Silva }.

Também considere as seguintes ligações de correferências preditas por um sistema de identificação de personagens: <Pedro – Pedro da Silva, Dr. Silva – Mariazinha, Maria – Maria da Silva, Maria da Silva – São Paulo>. Neste caso, são preditos três conjuntos de equivalências S = {Pedro, Pedro da Silva}, {Dr. Silva, Mariazinha}, {Maria, Maria da Silva, São Paulo}.

Analisando as ligações de correferência acima, observa-se que o sistema em análise identificou corretamente 2 ligações, e 3 são incorretas, de um total de 5 ligações da coleção dourada. Neste caso, a revocação será de $\frac{2}{5} = 0,4$, enquanto que a precisão será de $\frac{2}{4} = 0,5$. Finalmente, pode-se determinar a medida-F MUC utilizando a equação 3, que resulta em 0,44.

A medida-F MUC tem diversas limitações para avaliação da resolução de correferências (Luo, 2005): ela ignora menções únicas (sem outras menções que co-identifiquem a mesma entidade); ela não permite comparar efetivamente o desempenho de sistemas, pois esta medida favorece sistemas que produzem poucas entidades, e com isto pode resultar medidas-F mais altas para sistemas de pior desempenho.

Surgiram algumas propostas para sanar as limitações do MUC. Uma das primeiras resultou na métrica *B-cubed* (Bagga & Baldwin, 1998), onde a precisão e revocação de um sistema são calculadas com base na precisão e revocação obtidas para cada menção i, conforme formalizado nas equações 7 e 8. Nestas equações, K_i é o conjunto de equivalência da coleção dourada com todas as menções de uma entidade mencionada, que inclui a menção i, e S_i é o conjunto de equivalência predita por um sistema com todas as menções de uma entidade mencionada, que inclui a menção i. Com base nestas definições, a precisão de uma menção i é determinada pela razão entre o número de menções corretas da entidade referenciada por i (presentes em K_i e também na saída do sistema S_i) contendo a menção i e o número de menções em S_i. A revocação de uma menção i é determinada pela razão entre o número de menções corretas da entidade em S_i sobre o número de menções em K_i.

$$P_i = \frac{|S_i \cap K_i|}{|S_i|} \tag{7}$$

$$R_i = \frac{|S_i \cap K_i|}{|K_i|} \tag{8}$$

A medida-F *B-cubed* é calculada com base nas precisões e revocações obtidas pela soma ponderada das medidas calculadas para cada personagem, como formalizado nas equações 9 e 10. O peso associado a cada entidade dependeria da tarefa a ser cumprida pelo sistema. No caso do DIP, de extração de informação, o peso poderia igual para todas as entidades (personagens), ou então personagens mais importantes poderiam ter pesos maiores.

$$P = \frac{\sum_{i=1}^{N_k} w_i \times P_i}{|K|} \tag{9}$$

$$R = \frac{\sum_{i=1}^{N_k} w_i \times R_i}{|K|} \tag{10}$$

A título de exemplo, e usando os mesmos conjuntos K e R da métrica MUC, a precisão usando *B-cubed* resultaria em:

$$\begin{aligned} P_{Pedro\ da\ Silva} = \frac{2}{2}; P_{Pedro} = \frac{2}{2} \\ P_{Dr.\ Silva} = \frac{1}{2}; P_{Mariazinha} = \frac{1}{2} \\ P_{Maria\ da\ Silva} = \frac{2}{4}; P_{Maria} = \frac{2}{4} \\ P_{Sao\ Paulo} = \frac{0}{4}; \\ P = \frac{1}{7}\left(\frac{2}{2} + \frac{2}{2} + \frac{1}{2} + \frac{1}{2} + \frac{2}{4} + \frac{2}{4}\right) \end{aligned} \tag{11}$$

A métrica *B-cubed* também tem suas limitações. Como apontado por Luo (2005), o *B-cubed* calcula a precisão e revocação de menções comparando entidades contendo a menção e portanto uma entidade pode ser usada mais que uma vez. Motivado pelas limitações das métricas *MUC* e *B-cubed*, Luo (2005) propõe a métrica medida-F CEAF (*Constrained Entity-Alignment*), que é calculada com base no melhor mapeamento um-a-um entre as correferências da coleção dourada e da saída gerada pelo sistema.

3. Trabalhos Relacionados

Existem diversas iniciativas que tratam da identificação de personagens literárias. A maior parte dos trabalhos citados aqui visam, além da identificação de personagens, a extração das chamadas redes sociais, ou também chamadas de redes

conversacionais. Estas redes são derivadas a partir das interações de diálogo entre personagens, e são representadas por grafos, onde os nodos representam as personagens e as arestas indicam a existência de diálogo entre duas personagens.

Esta seção apresenta as principais iniciativas de identificação de personagens literárias, detalhando principalmente as métricas adotadas no processo de avaliação. A Tabela 1 sumariza as principais características das iniciativas analisadas, em termos de tipo de informação extraída, os corpora (*datasets*) usados nos experimentos, e as métricas adotadas. As métricas citadas nesta tabela são aquelas adotadas pelas iniciativas para avaliar os resultados das etapas da identificação de personagens apresentadas na Seção 2.1. Alguns trabalhos listados utilizam outras métricas para análise dos grafos gerados, que não são apresentadas por estarem fora do escopo deste artigo.

3.1. Extração de redes sociais

Elson et al. (2010) propõem um método para extração de redes sociais de texto literários. Os textos considerados foram 60 romances britânicos do século XIX. Neste trabalho, os nodos da rede social são ponderados de acordo com o nível de interação da personagem representada pelo nodo com as outras personagens, e as arestas são ponderadas de acordo com o nível de interação entre duas personagens.

Neste trabalho, os autores utilizaram a ferramenta Stanford NER tagger (Finkel et al., 2005) para extrair menções classificadas como pessoas ou organizações. Em seguida, as menções são agrupadas (clusterizadas) em correferentes para a mesma entidade (pessoa ou organização). Para identificar a personagem que está interagindo em cada caso de discurso direto, os autores utilizaram uma técnica de aprendizagem baseada em regras e estatísticas proposta por Elson & McKeown (2010).

A avaliação da proposta foi realizada considerando quatro romances. Para cada livro, foram selecionados aleatoriamente 4 a 5 capítulos. Estes capítulos foram anotados manualmente as personagens e as interações existentes. Para calcular o acordo entre os anotadores, cada anotador atribuiu "sim" ou "não" para cada par de personagens participando de uma interação. A partir disto, foi determinado o coeficiente de concordância de Kappa entre os anotadores.

Para avaliação, este trabalho adotou as métricas precisão, revocação e medida-F na determinação das conversações. Os autores também realizaram extrações de características das redes conversacionais, como o número de personagens e número de personagens que fala, o grau médio do grafo e variação da densidade do grafo, entre outros.

3.2. Extração de redes sociais de pessoas e lugares

Lee & Yeung (2012) também propõem um método para extração de redes sociais de textos literários. Neste trabalho, redes sociais, de forma similar às redes conversacionais adotadas por Elson & McKeown (2010), são grafos onde: os nodos representam personagens do tipo pessoa e suas localizações (nomes dos lugares); arestas entre personagens representam a existência de interação pessoal entre elas; e finalmente arestas entre personagens e lugares são presentes se a pessoa esteve fisicamente presente na localização.

Lee & Yeung (2012) também utilizaram a ferramenta Stanford NER tagger, agora para extração das entidades do corpus adotado (traduções em inglês dos 5 primeiros livros do Velho Testamento) classificadas como pessoas e nomes geográficos. Para resolução de correferência, foi utilizado o sistema *Stanford Deterministic Coreference Resolution* (Lee et al., 2011) e o método de agrupamento proposto por Elson & McKeown (2010) para associar as entidades mencionadas com suas diversas menções.

Para avaliação do método proposto, Lee & Yeung (2012) adotaram uma coleção dourada, onde as arestas personagem-personagem e personagem-localização foram anotadas manualmente. Além disso, o processo de avaliação considerou apenas as personagens principais (mencionadas ao menos 10 vezes no corpus). O trabalho também utilizou as métricas precisão, revocação e medida-F para avaliar de forma independente o reconhecimento das entidades mencionadas, o conjunto de arestas personagens-personagens, e o conjunto de arestas personagem-lugar.

3.3. Extração da rede social de Alice no país das maravilhas

Agarwal et al. (2013) apresentam um procedimento para extração da rede social da obra *Alice no país das maravilhas*. No trabalho, a rede social é definida por um grafo onde os nodos representando personagens e arestas indicam eventos sociais (evento em que duas personagens interagem de forma deliberada e consensual). Para esta tarefa, foi adotado um sistema pré-treinado com um corpus de notícias usando *tree kernel* e SVM (*Support Vector Machines*).

Iniciativa	Informação extraída	Corpus	Métricas
(Elson et al., 2010)	Rede de interações entre personagens	4 romances	Precisão Revocação Medida-F
(Lee & Yeung, 2012)	Rede de interação entre personagens e lugares	5 livros do Velho Testamento	Precisão Revocação Medida-F
(Agarwal et al., 2013)	Rede de interações entre personagens	1 livro	Acurácia Precisão Revocação Medida-F
(Valls-Vargas et al., 2014)	Personagens e seus tipos	Contos folclóricos (269 sentenças)	Acurácia Precisão Revocação
(Groza & Corde, 2015)	Personagens, seus tipos, relacionamentos e papéis	7 contos populares	Acurácia Precisão Revocação Medida-F
(Vala et al., 2015)	Grafo de correferência de personagens	88 obras	Precisão Revocação Medida-F
(Chaturvedi et al., 2017)	Personagens e seus relacionamentos	50 sentenças de romances	Precisão Revocação Medida-F
(Dekker et al., 2019)	Redes de interações de personagens e seus relacionamentos	40 romances	Precisão Revocação Medida-F
(Łajewska & Wróblewska, 2021)	Personagens do tipo pessoa	13 romances	Precisão Revocação Medida-F
(Chu et al., 2021)	Personagens, características e relações	Triplas compiladas dos infoboxes de 142 wikis da fandom.com	Precisão Revocação Medida-F HITs@k MRR
(Srinivasan & Power, 2022)	Personagens e seus tipos	20 sumários de estórias de ficção	Precisão Revocação Medida-F

Tabela 1: Sumário das iniciativas analisadas

Agarwal et al. (2013) avaliam a detecção de eventos sociais e extração da rede social usando as métricas clássicas de acurácia, precisão, revocação e medida-F. Além disso, para avaliar a rede social, são usadas métricas aplicadas à análise de redes sociais para comparar a rede extraída e a rede dourada.

3.4. Identificação de personagens com o sistema *Voz*

citevalls2014toward propõem um método que usa Raciocínio Baseado em Casos (CBR, *Case-Based Reasoning*) para identificar personagens em textos usando o sistema *Voz*. Este sistema primeiro extrai entidades mencionadas do texto e em seguida determina um vetor de características (*feature-vector)* para cada uma delas usando informações linguísticas e conhecimento externo. Este trabalho também propõe uma métrica para determinar a similaridade entre cada entidade com aquela da base de casos (*case-base*), que é utilizada para inferir se a entidade é uma personagem ou não. Este sistema utiliza as ferramentas Stanford CoreNLP[4] para segmentar o texto em sentenças e anotá-las com diversas informações.

[4] https://github.com/stanfordnlp/CoreNLP

Para a realização da avaliação, Valls-Vargas et al. (2014) utilizaram uma coleção de contos folclóricos russos traduzidos para o inglês. Como operação manual, foram retiradas no texto diálogos e passagens onde o narrador dialoga com o leitor diretamente. Ao todo, o corpus foi formado de 269 sentenças segmentadas pela Stanford CoreNLP. Este corpus contém diferentes tipos de personagens (humanos, animais, e objetos antropomórficos e criaturas fantásticas). Para avaliação, foram anotados manualmente 1122 entidades no corpus, sendo 614 personagens e 507 não são personagens. Como métricas, os autores utilizaram acurácia, precisão e revocação.

3.5. Identificação de personagens em contos populares

Groza & Corde (2015) propõem uma técnica de identificação de personagens em contos populares, além da extração dos relacionamentos entre estas personagens e seus papéis no desenvolvimento da estória. A extração do conhecimento dos contos é obtida pela combinação de um módulo de PLN, baseado na ferramenta de engenharia de textos GATE (Bontcheva et al., 2004), e na inferência de ontologia do domínio de contos populares. Esta ontologia é processada usando a OWLAPI (Horridge & Bechhofer, 2011) para a geração das classes de personagens para a GATE. O corpus de contos é analisado visando popular esta ontologia e anotar cada conto com as entidades mencionadas que foram identificadas.

Groza & Corde (2015) não apresentam explicitamente uma definição de personagens de contos populares, mas a ontologia de domínio define diferentes tipos de personagens, como pessoa (com diversas subclasses, como homem, mulher, menino, menina, rei, princesa, etc.), animal (com subclasses do tipo urso, pássaro, cão, e outros), planta, e entes sobrenaturais (como gigante).

Neste trabalho, a técnica de identificação de personagens foi testada usando sete contos populares. Para determinação da acurácia, os autores escolheram manualmente cerca de 3 personagens por conto, totalizando 20 personagens. Estas personagens foram classificadas como personagem principal ou secundária. Groza & Corde (2015) adotaram as métricas acurácia, precisão, revocação e medida-F no contexto de tarefas de classificação, onde: VP representa o número de sentenças que são encontradas tanto no conjunto manualmente anotado como no conjunto de teste; VN representa o número de sentenças que não estão no conjunto anotado manualmente, nem no conjunto de teste; FP representa o número de sentenças que estão no conjunto de teste e não no conjunto manualmente anotado; e FN representa o número de sentenças que estão no conjunto manualmente anotado, mas não no conjunto de teste.

3.6. Técnica de extração de grafos de correferência

Vala et al. (2015) propõem um pipeline de oito estágios para detectar personagens literárias e construir um grafo de correferência. Neste grafo, os nodos representam menções e as arestas entre menções indicam que as menções co-identificam uma mesma personagem. Este trabalho também utiliza as ferramentas Stanford CoreNLP tanto para NER quanto para resolução de correferência.

Para a avaliação, Vala et al. (2015) utilizaram dois conjuntos de dados (*dataset*): uma coleção manualmente anotada de 58 obras com a lista completa de todas as personagens e suas possíveis menções. O segundo conjunto é formado por 30 romances e a lista de suas personagens obtidas do site Sparknotes[5], sendo que foram adicionados manualmente as listas das possíveis menções que co-identificam cada personagem (i.e, os conjuntos de equivalência).

Para avaliação, foram considerados os conjuntos de equivalência gerados pelo sistema proposto, anotada por $E = \{E_1, \ldots, E_n\}$, onde E_i é o conjunto de equivalência contendo as possíveis menções para uma personagem i. Esta lista é confrontada com os conjuntos de equivalência K da coleção dourada contendo todas as menções corretas para cada personagem. Para avaliação, os autores formalizaram o problema em como determinar uma combinação bipartida máxima (*maximum bipartite matching*). Para precisão, a combinação é a medida da pureza de um conjunto de equivalência, E_i, em relação às menções da coleção dourada, K_j: $1 - \frac{|E^i - K^j|}{|E^i|}$. A revocação é definida de forma mais flexível da combinação, com o objetivo de medir se uma personagem K_j foi identificada. Esta combinação é medida como a seguinte função binária: 1 se $E_i \cap K_j \neq \emptyset$, e 0 caso contrário. O trabalho também utiliza a métrica medida-F com base nas definições de precisão e revocação anteriores.

[5] https://www.sparknotes.com/

3.7. Aprendizagem não supervisionada para extração de relações entre personagens

Chaturvedi et al. (2017) utilizam o pipeline BookNLP[6] para obter as tags POS (*Part Of Speech*), análises de dependência, resolução de correferências, e identificação das personagens principais. Em seguida, cada sentença envolvendo personagens é representada por um vetor de características que é associado a um estado latente. Para este último, o problema é tratado como uma tarefa de aprendizado não supervisionado, e os estados latentes são aprendidos utilizando suposições Markovianas para extração do fluxo de informações entre sentenças individuais.

Para avaliação, Chaturvedi et al. (2017) utilizaram 50 sentenças manualmente anotadas, e a métrica adotada foi a média da medida-F.

3.8. Avaliação de ferramentas para identificação de personagens e extração de redes sociais

Dekker et al. (2019) citam que extração de personagens no domínio da literatura é desafiador, pois os nomes não seguem as mesmas "regras" do mundo real, como apontado por de Does et al. (2017). Os autores citam que a resolução de correferência também é um desafio no domínio da literatura, devido à variedade de alcunhas que uma personagem pode receber.

Dekker et al. (2019) avaliam quatro ferramentas NER para identificação de personagens e redes sociais em romances, aferindo o desempenho destas ferramentas frente aos desafios apontados. As ferramentas NER avaliadas foram BookNLP[7], Stanford NER,[8] Illinois tagger[9] e IXA-Pipe-NERC[10].

Para a avaliação das ferramentas, Dekker et al. (2019) utilizaram um corpus de 40 romances clássicos e modernos, obtidos do projeto Gutenberg ou comprados em formato ebook. A avaliação foi realizada com base em uma coleção dourada composta de anotações realizadas manualmente em 10 romances (a partir de 300 sentenças em média selecionadas de cada livro). Durante o processo de identificação manual das personagens, foram adotadas as seguintes regras: ignorar pronomes genéricos (p.e., você, ele); ignorar frases exclamativas (p.e., Por Cristo!); ignorar sintagmas nominais genéricos (p.e., Mário não sabia o que dizer **ao mago**); e incluir personagens não humanas.

Para a avaliação foram utilizadas as métricas de precisão, revocação e medida-F. Os experimentos realizados por Dekker et al. (2019) demonstraram que a ferramenta BookNLP superou o desempenho das demais.

3.9. Anotação de Protagonistas

Łajewska & Wróblewska (2021) propõem a *protagonistTagger*, uma ferramenta para marcação (*tagging*) de personagens do tipo pessoa. O método implementado por esta ferramenta compreende duas etapas: (1) reconhecimento de entidades mencionadas (NER) da classe pessoa; e (2) desambiguação de entidades nomeadas (NED, *Named Entity Disambiguation*). O trabalho utilizou o modelo de linguagem pré-treinada oferecida pela biblioteca em código aberto SpaCy.[11]

Para a realização de experimentos foram considerados um conjunto de dados composto por 1300 sentenças de 13 romances clássicos. O processo de criação deste corpus foi o seguinte: (1) obtenção de um corpus inicial com os textos dos romances sem anotações; (2) obtenção de uma lista com nomes completos de todos os protagonistas (usando um parser da Wikipedia), que são considerados como etiquetas (*tags*) pré-definidas; (3) reconhecimento das entidades mencionadas da classe pessoa usando um modelo NER pré-treinado usando textos manualmente anotados; (4) e uso de um algoritmo de emparelhamento para anotar cada entidade nomeada da categoria pessoa a uma das etiquetas pré-definidas em (2).

Łajewska & Wróblewska (2021) citam que um dos casos mais difíceis de tratar no processo de emparelhamento são os diminutivos e alcunhas. Para tratar este problema, o *protagonistTagger* considera uma lista completa de diminutivos com mais de 3300 diferentes formas de nomes. Outro desafio foi a identificação de entidades mencionadas pelo seu sobrenome. No caso, para identificar se a menção ao sobrenome se refere à toda a família, ou a uma única personagem, foi considerada a palavra precedente ao sobrenome. Se é um título pessoal (por exemplo, *Mr., Mrs., Ms. ou Miss*) a citação identifica uma única pessoa.

Em Łajewska & Wróblewska (2021), os autores utilizaram as métricas clássicas de precisão, revocação e medida-F para avaliar o algoritmo de emparelhamento da etapa (4).

[6] https://github.com/booknlp/booknlp
[7] https://github.com/booknlp/booknlp
[8] https://nlp.stanford.edu/software/CRF-NER.shtml
[9] https://github.com/kordjamshidi/illinois-cogcomp-nlp
[10] https://github.com/ixa-ehu/ixa-pipe-nerc
[11] https://spacy.io/

3.10. KnowFi

Chu et al. (2021) propõem o método *KnowFi* para extração de conhecimento de textos de ficção longos. Este método combina a aprendizagem neural aprimorada por BERT (*Bidirectional Encoder Representations for Transformers*) com o algoritmo de aprendizado com seleção e agregação criteriosa de passagens de texto. Além disso, *KnowFi* determina os tipos semânticos das entidades usando SpaCy, e assim provendo um tipo para as menções (pessoa, nacionalidade/religião, evento, etc.).

Para a avaliação da extração de relações entre personagens, Chu et al. (2021) criaram o corpus LoFiDo *Long Fiction Documents*, composto de triplas Sujeito-Predicado-Objeto (SPO) compiladas dos *infoboxes* de 142 wikis de comunidades fãs em fandom.com. Este corpus especifica 64 relações, como inimigo, amigo, aliado, religião, arma, etc. Para avaliação, foi criada uma coleção dourada checada manualmente.

Chu et al. (2021) utilizaram as métricas clássicas de média de precisão, de revocação e de medida-F. Além destas, foram utilizadas as métricas HITs@k, para avaliar com que frequência um resultado correto aparece entre as k principais extrações por par entidade-relação, e a métrica MRR (*Mean Reciprocal Rank*) para obter a classificação recíproca média da primeira extração.

Neste trabalho foram realizadas duas avaliações: automática e manual. A avaliação manual foi necessária devido à baixa eficiência demonstrada pela avaliação automática, quando foi utilizada uma coleção dourada incompleta, gerada automaticamente. Na avaliação manual, foram selecionadas as top 100 extrações do resultado sobre as quais foram realizadas avaliações manuais para checar a correção de cada uma.

3.11. Extração de Personagens e seus tipos

Srinivasan & Power (2022) propõem uma solução para extração e classificação de personagens a partir de sumários da estória. A extração de personagens e a resolução de correferências foi realizada através da ferramenta StanfordNLP. Usando as tags POS a partir do texto anotado, foram identificados os nomes próprios. Ao final, é aplicado um conjunto de heurísticas para eliminação de entidades que não são personagens, considerando, entre outros, o fato que a entidade não é sujeito de nenhuma sentença, não tem um nome pessoal ou título pessoal, ou cujo nome não é reconhecido pela base de dados WordNet.

A solução proposta permite classificar uma personagem como protagonista, antagonista ou personagem auxiliar. Para esta classificação foram utilizados algoritmos de aprendizado supervisionados. Além do tipo, também foram analisados os pronomes pessoais para a identificação do género das personagens, que foi confirmado através de regras heurísticas.

Para a avaliação, este trabalho adotou um corpus formado por 20 sumários de estórias de ficção que ao total fazem menção a 218 personagens. Como métricas de avaliação, Srinivasan & Power (2022) utilizaram a precisão, revocação, e medida-F para o contexto de sistemas de classificação (Seção 2.2.2). Elas foram calculadas para duas classes: personagens e não-personagens. Para avaliação da extração de personagens, os autores consideraram *VP* aquelas instâncias que foram identificadas corretamente, *FP* aquelas instâncias que foram incorretamente identificadas como personagens, *VN* aquelas instâncias que foram corretamente identificadas como não-personagens, e *FN* aquelas instâncias que a abordagem falhou na identificação delas como personagem apesar delas serem de fato personagens.

Para avaliação da classificação das personagens, os autores também usaram as médias de precisão, revocação e medida-F entre os três tipos de personagens (protagonista, antagonista e suporte).

3.12. Outros sistemas

Conhecemos vários outros sistemas que descrevem tarefas relacionadas com o DIP, mas não os descrevemos aqui por não descreveram a forma como são avaliados: por exemplo Jayakumar et al. (2022) extraem redes sociais dinâmicas, Santos & Freitas (2019) criam redes para a literatura lusófona.

4. O DIP

Muito brevemente, visto que já foi apresentado em várias outros contextos (Santos et al., 2022b,a, 2023), o DIP, Desafio de Identificação de Personagens, é uma tarefa que pretende desenvolver e avaliar programas, que, para uma dada obra literária em português, consiga identificar:

- as personagens nessa obra;
- as relações familiares entre personagens.[12]

[12] Especificamente as seguintes: *mãe, pai, filho, filha, neto, neta, avó, avô, irmã, irmão, cunhado, cunhada,*

As personagens são representadas por um conjunto de menções (ou nomes) que lhes correspondem, pelo seu género, e pela sua profissão/ocupação/estatuto social. É possível que uma personagem tenha várias profissões ao longo da história, ou que nenhuma seja mencionada.

Para este desafio, foram criados dois conjuntos de dados compostos de 200 obras literárias de domínio público, disponibilizados no formato TXT e PDF. Estes conjuntos de dados estão disponíveis no site do DIP.[13] O desafio também definiu um formato de representação dos dados extraídos que seriam extraídos pelas soluções participantes do desafio. Esta padronização foi necessária para simplificar o processo de avaliação.

5. Metodologia de Avaliação Adotada

Para o processo de avaliação, foi necessário criar uma Coleção Dourada (CD). Para tal, todas as personagens, suas diversas menções, atributos e relações de 38 obras foram anotadas manualmente.[14]

No processo de avaliação, os resultados das soluções seriam confrontados com aqueles disponíveis na CD. Para aplicar métricas mais tradicionais, e porque não tínhamos ideia das possíveis interligações entre as diferentes subtarefas, escolhemos avaliar separadamente as cinco subtarefas do DIP, a saber:

- reconhecimento da entidade do tipo personagens mencionados na obra;
- resolução de correferência, para a identificação da conjunto de menções que identificam unicamente uma personagem;
- classificação do género da personagem entre masculino (M), feminino (F) e ambos os sexos (A);
- identificação das possíveis profissões/ocupações/estatutos sociais mencionados na obra;
- extração das relações familiares (um conjunto pré-determinado) entre as personagens

Para avaliar estas cinco tarefas, definimos cinco medidas parcelares, e consideramos a medida final como a média aritmética das cinco medidas (ou das quatro, se a obra avaliada não inclui qualquer relação familiar):

AI avaliação da identificação, usando a medida-F sobre o conjunto de todos os nomes identificados pelo sistema, e coligidos na CD

ACI avaliação da co-identificação, expandindo todos os conjuntos, e usando a medida-F sobre esses conjuntos

AG avaliação do género, entrando em conta com a co-identificação

APOES avaliação da profissão, ocupação e estatuto social, usando uma medida-F adaptada

AR avaliação das relações, também usando uma medida-F

As seções que seguem descrevem como essas medidas são calculadas. Para ilustrar estas métricas, considere os seguintes dados anotados em uma obra fictícia da CD:

- Co-identificações de uma personagens, seus gêneros e profissões: {Bento – Padre Bentinho – Dom Casmurro, M, advogado}, {Capitu – Capitolina, F,}, {Dona Fortunata, F,}, {Thomaz, M, escravo}, {Doutor João da Costa – João da Costa, M, médico};
- Relações entre personagens: (Bento, marido, Capitu),(Dona Fortunata, mãe, Capitu).

Também considere que um sistema de identificação de personagens extraia da mesma obra as seguintes informações:

- Co-identificações de uma personagens, seus gêneros e profissões: {Dom Casmurro, M, escravo - advogado}, {Bento, Padre Bentinho, M}, {Capitu, M,}, {São Paulo, M, }, {Thomaz, M, escravo};
- Relações entre personagens: (Bento, marido, Capitu),(Thomaz, primo, Capitu).

5.1. Avaliação da identificação (AI)

Para avaliar a identificação de personagens com esta métrica, é calculada inicialmente a precisão P_i e revocação R_i de cada obra i usando as equações 12 e 13. Nestas equações S_i é o conjunto de menções a personagens identificadas por um sistema para uma obra i, e K_i é o conjunto de menções presentes na CD para esta mesma obra. A precisão para uma obra i então determinada pela razão entre o número de menções corretamente identificadas pelo sistema nesta obra e o

primo, prima, tio, tia, sobrinho, sobrinha, bisavó, bisavô, bisneto, bisneta, nora, genro, sogro/a, mulher, marido, padrinho, madrinha, compadre, comadre, afilhado e *afilhada.*

[13] https://www.linguateca.pt/aval_conjunta/dip/colecao.html

[14] Disponíveis em https://www.linguateca.pt/aval_conjunta/dip/nova_colecao_dourada/

número total de menções identificadas pelo sistema de uma entidade. Por sua vez, a revocação para uma obra i é determinada pela razão entre o número de menções corretamente identificadas pelo sistema e o número de menções realmente presentes na obra.

$$P_i = \frac{|S_i \cap K_i|}{|S_i|} \tag{12}$$

$$R_i = \frac{|S_i \cap K_i|}{|K_i|} \tag{13}$$

No exemplo ilustrativo, temos as seguintes menções identificadas:

- K_1 = {Bento, Padre Bentinho, Dom Casmurro, Capitu, Capitolina, Dona Fortunata, Thomaz, Doutor João da Costa, João da Costa};
- S_1 = {Dom Casmurro, Bento, Padre Bentinho, Capitu, São Paulo, Thomaz};

Neste caso, $P_1 = \frac{5}{6}$ e $R_1 = \frac{5}{9}$. Sendo assim, a medida-F será de 0,67.

Com base na precisão e revocação de cada obra i, é determinada a média da medida-F_i usando a Equação 3. Ao final, a nossa métrica AI é determinada pela média das Medidas-F obtidas no conjunto de obras N utilizadas para avaliação, conforme Equação 14.

$$AI = \frac{\sum_{i=1}^{N} Medida_{F_i}}{N} \tag{14}$$

Durante o ensaio do DIP e do processo de avaliação, os organizadores e participantes se depararam com diversos desafios, alguns já citados na literatura, e outros. Um caso que tivemos de tratar foi a existência — inesperadamente, em muitas obras — de personagens com o mesmo nome, quer totalmente (dois Franciscos), quer parcialmente (ou seja o Sr. João da Esquina, e o Dr. João Semana podem ser ambos apenas tratados por João em contextos específicos).

Embora estes casos — sobretudo os de sobreposição total — sejam praticamente impossíveis de identificar automaticamente, quisemos que a CD incluísse o número certo (e as formas certas) das personagens, para podermos caracterizar a literatura que era o nosso objeto de estudo.

Assim, a primeira passagem para a avaliação de identificação é a "tradução" de nomes semelhantes em nomes diferentes, para depois aplicar a medida-F, como já indicado.

5.2. Avaliação da co-identificação (ACI)

Para avaliação da co-identificação das personagens, criamos a métrica ACI, que se trata de uma medida-F obtida comparando-se pares de co-identificação das personagens.

Para determinar a ACI, inicialmente o conjunto de co-identificação de cada personagem é convertida em pares de co-identificação. Para tal, inicialmente as menções de cada personagem são organizadas em ordem alfabética. A primeira menção identificará a personagem, e cada par de co-identificação da personagem relaciona esta primeira menção com cada outra do conjunto de menções. Para considerar personagem referenciadas com apenas uma menção, é utilizado o termo ZERO para formar o par com esta menção única. No exemplo ilustrativo, os pares de co-identificação da CD seriam: (Bento – Dom casmurro), (Bento – Padre Bentinho), (Capitolina – Capitu), (Dona Fortunata – ZERO), (Thomaz – ZERO), (Doutor João da Cota – João da Costa). E os pares identificados pelo sistema seriam: (Dom Casmurro – ZERO), (Bento – Padre Bentinho), (Capitu – ZERO), (São Paulo – ZERO), (Thomaz – ZERO).

Na métrica ACI proposta, a precisão e revocação são obtidas confrontando os pares de co-identificação de cada obra i considerando apenas os pares da CD de personagens corretamente identificadas pelo sistema. No exemplo anterior, a CD indica a existência de 5 personagens, e o sistema identificou 3 delas. Uma personagem é identificada pelo sistema se ao menos uma das menções que co-identificam uma personagem na CD esteja presente no resultado. Assim, apenas os seguinte pares da CD serão considerados: (Bento, Dom casmurro), (Bento, Padre Bentinho), (Capitolina – Capitu), (Thomaz – ZERO).

No exemplo ilustrativo, segundo as equações 1 e 2), $P_1 = \frac{2}{5}$ e $R_1 = \frac{2}{4}$. Sendo assim, a medida-F será de 0,44.

A exemplo da métrica AI, a medida ACI será calculada como a média da medida-F obtida considerando as obras da CD.

5.3. Avaliação do género (AG)

Para avaliar a identificação do género de uma personagem proposta pelo sistema, consideramos as suas diferentes menções, e confirmamos com o género presente na CD para cada menção. Para tal, para cada obra i presente na CD é criado um conjunto de pares <menção, género>, designado por KD_i.

Para cada personagem j proposta pelo sistema, usamos as menções que também se encontram na CD e descartamos as não existentes, formando assim o conjunto KS_j. KD_j corresponde aos pares com as mesmas menções na CD.

O resultado da avaliação usa a seguinte função:

$$AG_j = \begin{cases} -1, & \text{se } \bigcup \text{gen}(KD_j) = \{M, F\} \\ -1, & \text{se } \text{gen}(KS_j) \neq \bigcup \text{gen}(KD_j) \\ 1, & \text{se } \text{gen}(KS_j) = \bigcup \text{gen}(KD_j) \\ 0, & \text{se } MF \in \bigcup \text{gen}(KD_j) \end{cases} \quad (15)$$

E o valor para uma obra é a média de AG_j naquela obra, ou seja:

$$AG_i = \frac{\sum_{j=1}^{|numpers|} AG_j}{|numpers|} \quad (16)$$

Ao contrário do que se poderia pensar, esta é a subtarefa cuja avaliação é mais original, devido às várias possibilidades que tivemos de contemplar.

Explanando em português a equação 15, se a personagem contiver géneros distintos segundo a CD (estiver marcada com MF, ambos), não recebe pontuação (note-se que não houve nenhum caso destes nas obras do DIP).

Se o sistema tiver proposto géneros incongruentes, ou seja, tanto M como F, baseados na separação de diferentes nomes de uma mesma personagem, segundo a CD, em personagens diferentes, também conta como -1.

Além disso, recebe 1 ou -1 conforme o género concorde com o da CD ou não.

Atente-se no seguinte exemplo ilustrativo: A partir da CD são extraídos os seguintes conjuntos: {(Bento, M), (Padre Bentinho, M), (Dom Casmurro, M), (Capitu, F), (Capitolina, F), (Thomaz, M), (José Bento, M)}. Já o sistema hipotético identificou as seguintes três personagens com seus géneros: {(Bento|Padre Bentinho|Dom Casmurro|Capitu, M), (Capitolina, F), (Thomaz|José Bento, M)}.

A primeira recebe -1 (porque na CD corresponde a diferentes personagens com géneros diferentes), e as duas segundas recebem 1. O valor de AG será de (-1+2)/3= 0,333.

A métrica AG é determinada pela Equação 17, onde $|G|$ é o número de obras na CD, e AG_i é a avaliação da qualidade da identificação do género das personagens da obra i.

$$AG = \frac{\sum_{i=1}^{|numpers|} AG_i}{|G|} \quad (17)$$

De notar que as nossas escolhas obedeceram ao seguinte princípio: não devemos penalizar, ou avaliar, uma mesma questão mais do que uma vez. Por isso aceitamos (e consideramos corretas, do ponto de vista do género) personagens incorretas, por exemplo juntando várias diferentes, desde que o seu género seja semelhante.

5.4. Avaliação das profissões, ocupações e estatutos sociais (POES)

Neste caso, que podemos considerar um caso de classificação múltipla (pode haver mais de uma POES para uma personagem), também usamos a medida-F, mas apenas relativa às personagens que existem na CD. Ou seja, não penalizamos nem premiamos a atribuição de POES a personagens não existentes. Isto é mais uma aplicação do princípio mencionado na seção anterior. Essa penalização terá sido feita por altura da avaliação da identificação.

Para avaliar a identificação dos POES das personagens, e ao mesmo tempo considerando a existência de diferentes menções identificando cada personagem, consideramos o POES atribuído a cada menção. Para tal, para cada obra i presente na CD são criados dois conjuntos de pares <menção, POES>, anotados por $POESG_i$ e $POESS_i$. $POESG_i$ mantém os pares identificando os POES de cada menção da obra i presente na CD (K_i) e cujas menções também foram identificadas pelo sistema (S_i). $POESS_i$ mantém os pares das menções presentes em K_i, mas agora com os POES identificados pelo sistema.

No exemplo ilustrativo, a partir da CD são extraídos o conjunto $POESG_1$ = {(Bento, advogado), (Padre Bentinho, advogado), (Dom Casmurro, advogado), (Capitu, ZERO), (Thomaz, escravo)}. Já para o sistema hipotético do exemplo $POESS_1$ = {(Bento, ZERO), (Padre Bentinho, ZERO), (Dom Casmurro, escravo), (Dom Casmurro, advogado), (Capitu, ZERO), (Thomaz, escravo)}.

Confrontando os dois conjuntos anteriores, pode-se calcular a precisão e revocação de cada obra. No exemplo, a precisão seria de 3/6 e a revocação de 3/5, e assim a medida-F APOE da obra será de 0,55.

A medida POES da solução é determinada pela média da medida-F POES das obras da CD.

5.5. Avaliação das relações (AR)

A avaliação das relações familiares, anotada por AR, foi de longe a que exigiu um processamento mais complicado, porque foi preciso converter — ou alinhar — os identificadores das personagens propostos pelo sistema para os identificadores na CD, antes de poder avaliar as relações familiares entre as personagens.

É depois necessário expandir as relações para todas as formas possíveis de exprimir a mesma relação, entrando em conta com o género da personagem. Assim, se `X neta Y`, adicionar-se-á `Y avó X` se Y for do género feminino, e Y avô X se for do género masculino.

Após estas duas operações, usamos simplesmente a medida-F sobre o conjunto das relações (expandido).

Há três questões que devemos levantar:

- A contagem faz-se sempre sobre pares de relações (a relação proposta pelo sistema e a usa inversa). Mas há um caso apenas, o de `X viúva Y`, em que não há relação inversa. Esse caso deveria contar a dobrar para ser igualmente premiado ou penalizado.
- Não fazemos uma expansão transitiva, apenas reflexiva. Ou seja, de `A irmão B` e `B irmão C` não concluímos, para efeitos de avaliação, que `A irmão C`. Nem de `D mãe E` e `E mãe F` obtemos `D avó F`. (Mas veja-se Mota & Santos (2023) para outros processamentos e conclusões.)
- No caso de não haver nenhuma relação familiar na CD,[15] não calculamos esta medida de avaliação.

A fim de ilustrar o cálculo da medida AR, considere novamente o exemplo ilustrativo desta seção. Os pares de relação da CD são os seguintes (considerando a menção que identifica a personagem como descrita na Seção 5.2): {(Bento, marido, Capitu)}. E o sistema identificou as seguintes relações: {(Bento, marido, Capitu),(Thomaz, primo, Capitu)}. Neste caso, a precisão será 1/2 e revocação será 1/1.

5.6. Exemplos concretos

Exemplos de aplicação de cada uma destas medidas com base numa resposta fictícia sobre a obra *Dom Casmurro*, de Machado de Assis,[16] podem ser consultados na página associada à avaliação do DIP.[17]

obra	AI	ACI	AG	APOES	AR
001	0,596	0,913	1,000	0,000	0,000
002	0,690	0,750	1,000	0,133	0,000
004	0,711	0,688	0,818	0,370	0,308
005	0,479	0,611	1,000	0,250	0,000
006	0,800	0,854	0,941	0,182	0,222
025	0,455	0,520	0,882	0,143	0,000
026	0,405	0,683	0,647	0,304	0,286
030	0,773	0,708	0,800	0,094	0,250
032	0,704	0,871	0,778	0,200	0,000
033	0,535	0,182	0,909	0,178	0,133
037	0,683	0,427	0,789	0,189	0,073
043	0,660	0,646	0,850	0,391	0,338
047	0,854	0,660	1,000	0,143	0,632
051	0,667	0,889	1,000	0,222	0,000
054	0,642	0,690	0,857	0,581	0,235
055	0,588	0,727	1,000	0,235	–
064	0,628	0,685	1,000	0,189	0,000
072	0,561	0,780	0,889	0,405	0,000
075	0,598	0,672	0,882	0,362	0,145
096	0,560	0,868	0,750	0,200	0,242
099	0,716	0,484	1,000	0,386	0,231

Tabela 2: Avaliação do PALAVRAS no DIP.

6. Avaliação do sistema participante

Apenas um sistema participou no DIP, o PALAVRAS-DIP (ver Bick (2023)), sobre o qual apresentamos aqui as medidas que alcançou em outubro de 2022. Fazendo a média sobre as várias obras, em 21 obras, obteve um AI médio de 0,634, um ACI médio de 0,681, um AG médio de 0,895, e um APOES médio de 0,246. Como uma obra na CD não tinha nenhumas relações, a média de AR, feita sobre 20 obras, foi de 0,155.

Na Figura 1 mostramos as diferentes classificações para a identificação. A obra em que obteve melhores resultados na métrica AI, 0,854, foi a obra 47: *A escrava Isaura*, de Bernardo Guimarães. A obra em que funcionou pior, com 0,405, foi a obra 26: *Simá*, de Lourenço Amazonas.

A Figura 2 contém as diferentes classificações para a co-identificação (ou avaliação de correferência). A obra em que obteve melhores resultados, 0,913, foi a obra 1: *Miss Kate*, de Cosme

[15] Na primeira edição do DIP, isso apenas aconteceu para a obra 55: *O bom crioulo* de Adolfo Caminha.

[16] No âmbito do DIP, foram disponibilizadas as soluções para quatro obras: *Dom Casmurro* e *As Pupilas do Senhor Reitor*, de Júlio Dinis, na apresentação da tarefa; e *Quincas Borba* de Machado de Assis e *Dramas da Corte* de Alberto Osório de Castro, como resultado do ensaio.

[17] https://www.linguateca.pt/aval_conjunta/dip/avaliacao.html

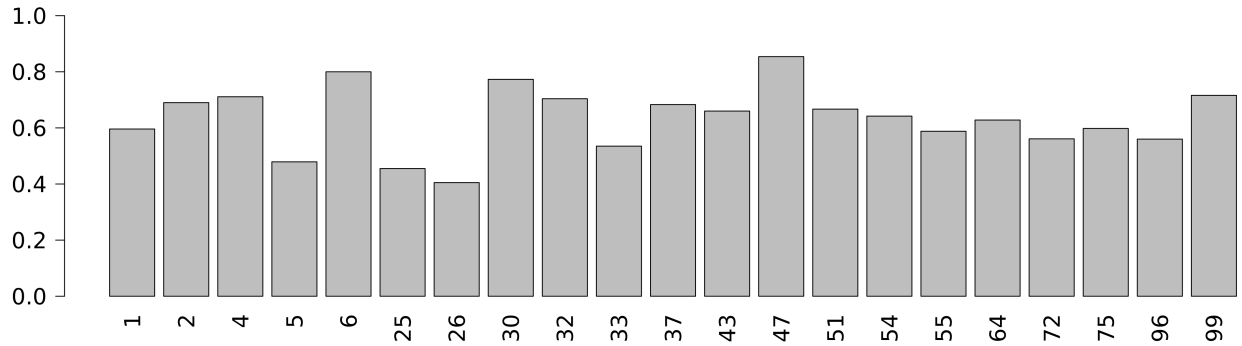

Figura 1: Resultados da avaliação da identificação

Velho, e a pior, com apenas 0,182, foi a 33: *A ermida de Castromino*, de Teixeira de Vasconcellos. Estes casos representam diferentes situações: na primeira obra, com poucas (21) personagens, a maior parte das correferências são variações dum mesmo nome, tendo ou não formas de tratamento. A segunda obra, com 33 personagens, é talvez o caso da obra com mais erros provenientes do reconhecimento ótico de carateres.

A Figura 3 contém as diferentes classificações para a avaliação do género. É claramente a tarefa em que o PALAVRAS tem melhor desempenho, e que podemos também claramente considerar a tarefa mais fácil. Mesmo assim houve algumas obras com uma pontuação fraca, como 0,647 para a obra 26: *Simá*, de Lourenço Amazonas, e 0,750 para a obra 99 *Amar, verbo intransitivo*, de Mário de Andrade.

A Figura 4 contém as diferentes classificações para a avaliação da profissão, ocupação e estatuto social.

Para a avaliação de APOES, a melhor pontuação foi de 0,581 na obra 54: *O Barão de Lavos*, de Abel Botelho. A pior, de zero, refere-se à obra 1, *Miss Kate*, de Cosme Velho. Esta obra ilustra uma das fragilidades, ou problemas, da forma de fazer a avaliação da profissão, ocupação e estatuto social no DIP: o facto de não haver uma normalização, ou ontologia das várias formas de definir esses termos. Por exemplo, a personagem Tiburtino Mendes está descrita na coleção dourada como "sextanista da faculdade de Medicina," e na resposta do sistema como "estudante." A resposta está evidentemente certa, e é até talvez mais útil de um ponto de vista de leitura distante, em que quereríamos juntar todos os estudantes e não distinguir o ano ou a faculdade em que estudam.

Mas para conseguir que a avaliação automática considerasse a resposta do sistema como certa, teríamos de fazer um trabalho considerável de padronização e definição do que se consideraria correto ao nível do campo semântico "descrição profissional."

Finalmente, a Figura 5 contém as diferentes classificações para a avaliação da extração de relações.

Esta foi a área em que o PALAVRAS teve pior desempenho, e que foi aliás mencionada pelo seu autor como ainda não estando suficientemente desenvolvida na altura da participação no DIP. Não admira, portanto, que não tenha conseguido identificar relações em oito obras. O melhor resultado foi de 0,632 para a obra 47: *A escrava Isaura*, de Bernardo Guimarães.

Se quisermos apreciar o desempenho global do PALAVRAS em relação a cada obra, e fazendo portanto a média aritmética das cinco (ou quatro) medidas, vemos na Figura 6 que o desempenho varia entre um máximo de 0,6578 e um mínimo de 0,3874 para os já mencionados *A escrava Isaura* e *A ermida de Castromino*.

Colocámos, na figura, numa cor diferente as obras escritas por autoras. Na figura 7 parece que as obras escritas por autoras tiveram pior classificação global, mas a diferença não é estatisticamente significativa.

6.1. Nova tentativa do PALAVRAS

Para ver se estas medidas podem contribuir para uma comparação racional de diferentes sistemas, pedimos novos resultados ao PALAVRAS em abril de 2023, sabendo que o autor tinha melhorado o sistema desde o DIP, que relembramos, ocorreu de 15 a 17 de setembro de 2022. Os novos resultados encontram-se na Tabela 3.

O valor médio de AI, ACI, AG e APOES, calculado sobre 21 obras, foi de 0,633, 0,702, 0,934 e 0,262. Quanto à avaliação de relações, a média sobre as 20 obras é de 0.195.

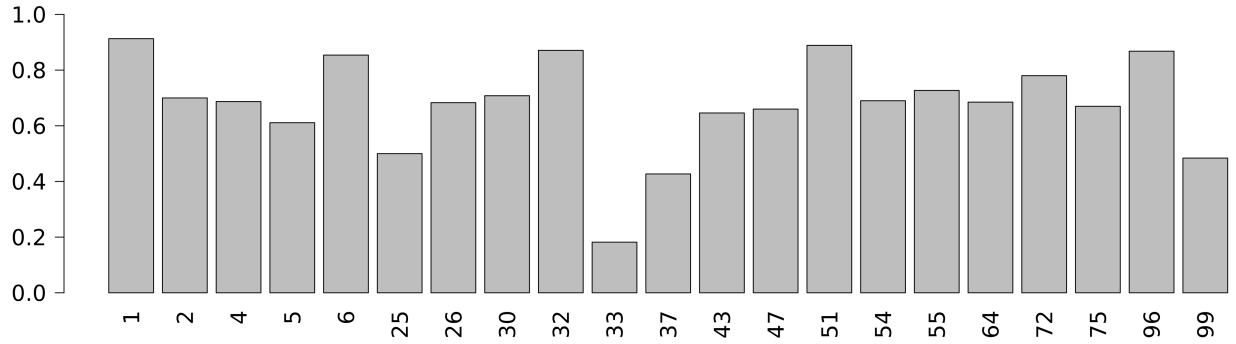

Figura 2: Resultados da avaliação da correferência, ou co-identificação

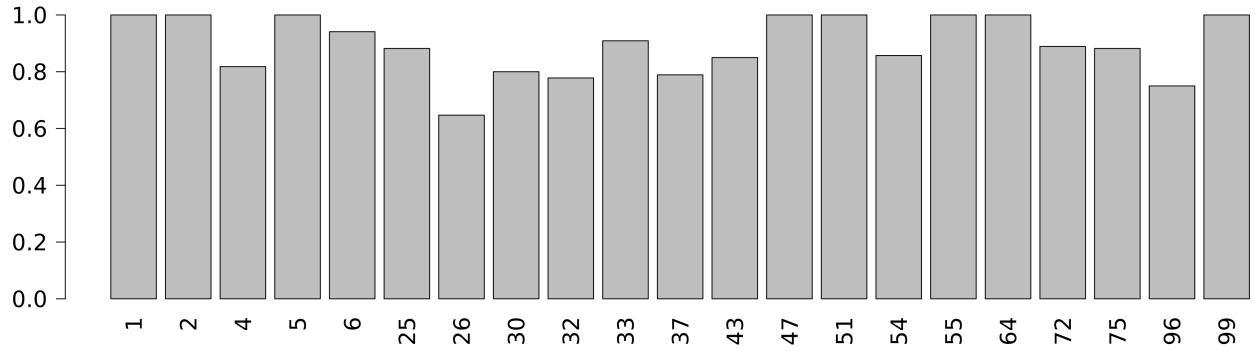

Figura 3: Resultados da avaliação do género

obr	AI	ACI	AG	APO	AR
001	0,485	**0,942**	0,857	0,000	0,000
002	0,647	**0,824**	1,000	0,091	0,000
004	**0,777**	0,585	**0,911**	0,303	0,093
005	**0,514**	0,514	1,000	0,080	**0,400**
006	0,724	0,659	0,852	0,053	**0,308**
025	0,446	0,485	**0,913**	**0,182**	0,000
026	0,404	**0,706**	**0,789**	**0,379**	0,286
030	0,765	**0,825**	**0,939**	**0,211**	**0,429**
032	**0,765**	0,730	**0,833**	**0,205**	0,000
033	**0,567**	**0,591**	**1,000**	0,114	**0,174**
037	0,626	**0,489**	**0,909**	**0,246**	**0,087**
043	**0,686**	**0,712**	0,822	**0,400**	0,314
047	0,796	**0,976**	1,000	**0,196**	0,381
051	**0,833**	0,889	1,000	**0,750**	0,000
054	0,614	**0,820**	**0,938**	0,439	0,000
055	**0,632**	0,462	1,000	**0,364**	–
064	0,567	**0,716**	1,000	0,167	0,000
072	**0,650**	0,762	**0,920**	0,333	**0,222**
075	0,446	0,543	**0,933**	0,324	0,114
096	**0,687**	**0,939**	**1,000**	**0,421**	**0,846**
099	0,656	**0,577**	1,000	0,243	**0,242**

Tabela 3: Avaliação do PALAVRAS em maio de 2023: a negrito estão os casos em que melhorou

De notar que, tal como na avaliação conjunta propriamente dita, tivemos de fazer alguns ajustes à CD para não penalizar novas menções que não tinham sido encontradas pelos revisores humanos.

Embora não vejamos grande diferença no desempenho do sistema, convém também sublinhar que o PALAVRAS não seguiu completamente as diretivas do DIP, em duas questões que podem ter consequências negativas nas medidas de avaliação: normalizou os nomes das profissões, ocupações e estatutos sociais – por exemplo, *creada* passou para *criada*; e também os (poucos) casos em que um nome de relação familiar era usado como forma de tratamento (ou seja, *viuva Maria* para *viúva Maria*, ou *mãi Joana* para *mãe Joana*).

Poderíamos — ou talvez deveríamos mesmo, para um medição mais justa do desempenho do sistema — fazer uma avaliação mais relaxada, para realmente identificar, e premiar, os casos que apenas são devidos a falta de normalização na CD.

7. Avaliação crítica da metodologia

Há várias razões para sermos críticos em relação à metodologia de avaliação usada no primeiro DIP, embora tenha sido por nós proposta com a me-

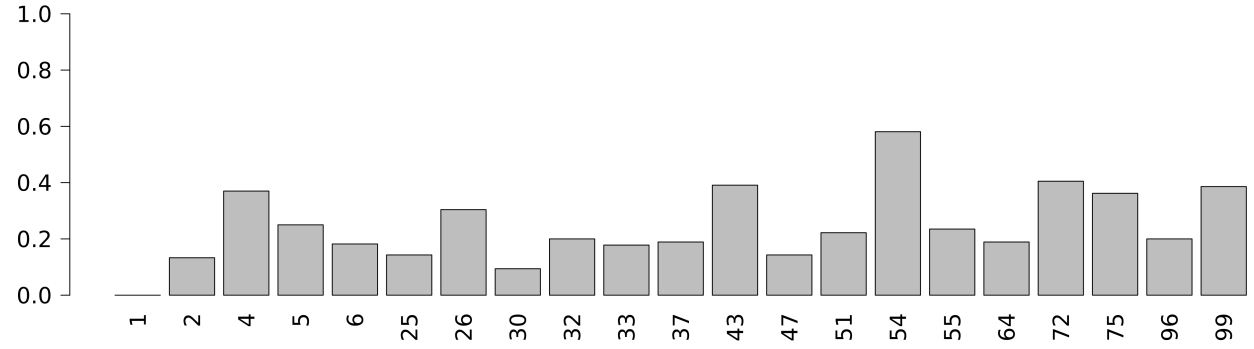

Figura 4: Resultados da avaliação da profissão, ocupação e estatuto social

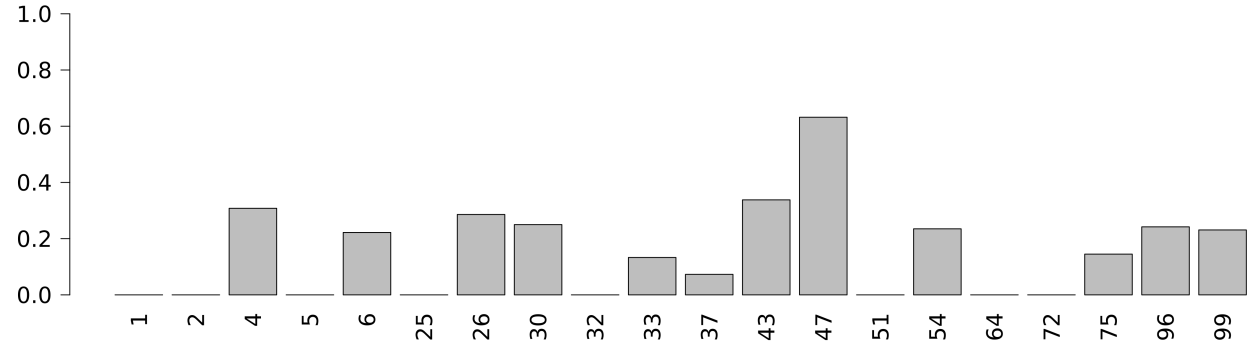

Figura 5: Resultados da avaliação da extração de relações familiares

lhor das intenções. Em primeiro lugar, o termos dividido a avaliação em cinco avaliações distintas impede uma visão global da dificuldade da tarefa, e da relação entre as diferentes subtarefas. Ou seja, é plausível que o facto de que uma personagem tenha muitos nomes torne a deteção do seu género, ou da sua profissão, mais difícil. Ou que uma personagem muito frequente tenha mais formas de ser mencionada do que outra que só aparece duas vezes. Contudo, todas são tratadas da mesma maneira.

Queremos com isto dizer que as medidas de avaliação abstraem da dificuldade intrínseca de uma dada obra, e de uma dada personagem de uma dada obra.

Idealmente deveríamos pesar a avaliação de um sistema pela dificuldade da obra e da personagem: uma obra com cinquenta personagens cada uma delas com mais do que uma forma de serem mencionadas é certamente mais complicada do que outra apenas com três personagens só com uma forma de serem identificadas. Contudo, ambas as obras contam igualmente.

A mesma coisa se passa para personagens com nomes genuinamente diferentes, e que mudam de profissão ao longo de uma obra. O acerto de um sistema sobre elas conta o mesmo que os casos de uma personagem que só aparece uma vez, e que portanto só tem uma forma de ser mencionada.

Deveríamos ser capazes de medir a dificuldade de identificação das personagens de uma obra, com base na coleção dourada, e depois avaliar um sistema entrando em conta com isso. Contar pouco os casos fáceis, e dar mérito aos casos difíceis.

Em segundo lugar, o uso de medidas separadas para as subtarefas pode ser enganador, visto que elas não medem, por exemplo, a capacidade global de identificar profissões de um dado sistema. E isto porque apenas nos atemos às profissões das personagens bem identificadas.

Considerando uma situação em que há dez personagens com 10 profissões, e que o sistema apenas identifica duas dessas personagens, e as profissões estão certas, terá uma medida 1 na APOES, mas só encontrou duas das 10 profissões que um leitor estaria interessado em encontrar.

Parece-nos que o problema da avaliação do DIP como a concebemos é que não toma (excepto na medida do género), as personagens certas como ponto de partida.

O ideal, parece-nos agora, seria ter uma medida por personagem. Quantas suas formas foram identificadas e juntas? Quantas suas profissões e relações com outras personagens foram identificadas? Quantas personagens faltam? Quantas são espúrias?

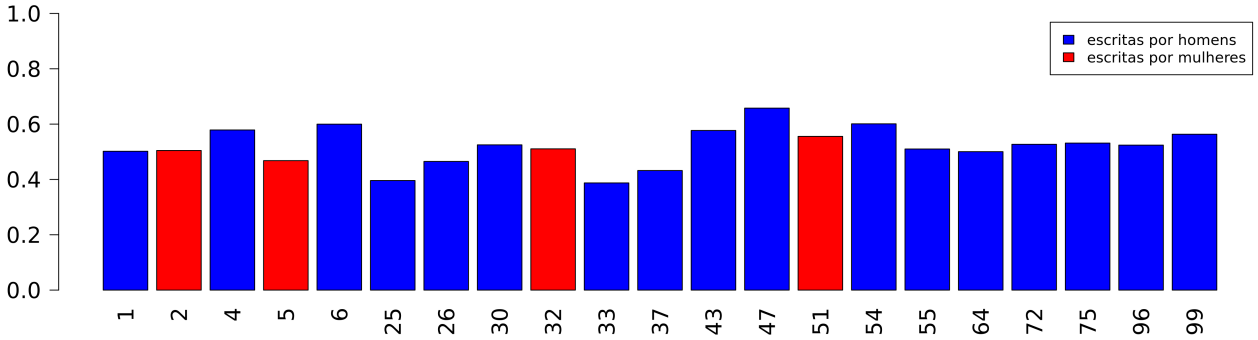

Figura 6: Resultados da avaliação total

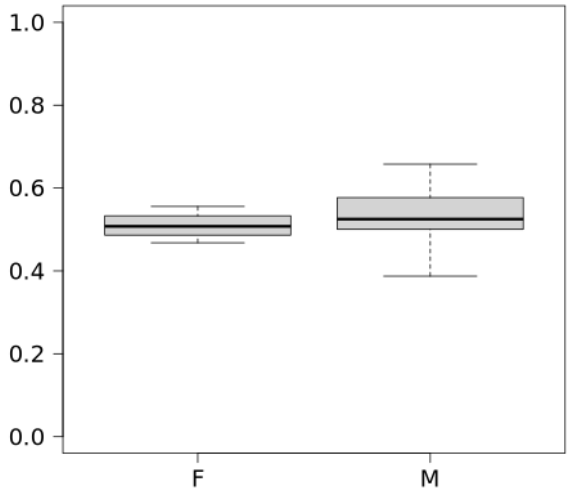

Figura 7: Avaliação total pelo género do autor

Em vez de concentrar a avaliação em pares de nomes correferentes, como fazemos em ACI, ou a pares profissão nome, em APOES.

No entanto, definir uma métrica que consiga pontuar todos os casos possíveis de uma forma satisfatória não é fácil, e terá de ficar para trabalho futuro.

8. Conclusões

A tarefa do DIP é um pouco diferente das tarefas tratadas na literatura que revisamos, porque junta várias tarefas (como identificação e correferência), não se aplica aos próprios textos (como é o caso da marcação de protagonistas) mas apenas extrai essa informação da obra completa. Embora também compare/avalie um tipo de redes de personagens, são diferentes das usadas nos trabalhos acima descritos, porque não dependem da interação ou da co-ocorrência das personagens no texto, são apenas representações das suas relações familiares.

Seja como for, existem muitos pontos de contacto com os trabalhos mencionados anteriormente, e todos focam o estudo de personagens em textos literários.

Neste artigo, apresentámos detalhadamente as métricas de avaliação usadas no primeiro DIP, os resultados com elas obtidos pelo único sistema participante, o PALAVRAS-DIP, no DIP e seis meses mais tarde, e fizemos várias críticas no sentido de esclarecer os potenciais problemas, e futuramente desenvolver outras medidas.

Agradecimentos

Agradecemos vivamente à Cristina Mota a bateria de testes que desenvolveu para testar os programas de avaliação, e aos outros organizadores, participantes e observadores do DIP, pelo retorno dado em variadas ocasiões.

Agradecemos também a Rebeca Schumacher e Cristina Mota os seus comentários que nos permitiram melhorar o texto.

Agradecemos à FCCN – Fundação para a Computação Científica Nacional (Portugal), o alojamento da Linguateca nos seus servidores, e ao UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais.

Referências

Agarwal, Apoorv, Anup Kotalwar & Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on Alice in wonderland. Em *6th International Joint Conference on Natural Language Processing*, 1202–1208.

Bagga, Amit & Breck Baldwin. 1998. Algorithms for scoring coreference chains. Em *1st International Conference on Language Resources and Evaluation (LREC)*, vol. 1, 563–566.

Bick, Eckhard. 2023. Extraction of literary character information in Portuguese. *Linguamática* 15(1). 31–40. 10.21814/lm.15.1.397.

Bontcheva, Kalina, Valentin Tablan, Diana Maynard & Hamish Cunningham. 2004. Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering* 10(3–4). 349–373. 10.1017/S1351324904003468.

Chaturvedi, Snigdha, Mohit Iyyer & Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. Em *AAAI Conference on Artificial Intelligence*, vol. 31 1, 3159–3165. 10.1609/aaai.v31i1.10982.

Chu, Cuong Xuan, Simon Razniewski & Gerhard Weikum. 2021. KnowFi: Knowledge extraction from long fictional texts. Em *3rd Conference on Automated Knowledge Base Construction*, on line. 10.24432/C51S38.

Dekker, Niels, Tobias Kuhn & Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science* 5. e189.

de Does, Jesse, Katrien Depuydt, Karina Van Dalen-Oskam, Maarten Marx et al. 2017. Namescape: Named entity recognition from a literary perspective. Em *CLARIN in the Low Countries*, chap. 30, 361–370. Ubiquity Press London.

Elson, David & Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. Em *AAAI Conference on Artificial Intelligence*, vol. 24 1, 1013–1019. 10.1609/aaai.v24i1.7720.

Elson, David K, Nicholas J. Dames & Kathleen McKeown. 2010. Extracting social networks from literary fiction. Em *48th Annual Meeting of the Association for Computational Linguistics*, 138–147.

Finkel, Jenny Rose, Trond Grenager & Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. Em *43rd Annual Meeting of the Association for Computational Linguistics*, 363–370. 10.3115/1219840.1219885.

Groza, Adrian & Lidia Corde. 2015. Information retrieval in falktales using natural language processing. Em *IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 59–66. 10.1109/ICCP.2015.7312606.

Horridge, Matthew & Sean Bechhofer. 2011. The OWL API: A Java API for OWL ontologies. *Semantic Web* 2(1). 11–21.

Jayakumar, Archana, Vedica Rao, AS Rohit Kumar, Prithwijit Banerjee & Roopa Ravish. 2022. Analyzing the development of complex social systems of characters in a work of literary fiction. Em *3rd International Conference for Emerging Technology (INCET)*, 1–7. 10.1109/INCET54531.2022.9824015.

Labatut, Vincent & Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys* 52(5). 1–40. 10.1145/3344548.

Łajewska, Weronika & Anna Wróblewska. 2021. Protagonists' tagger in literary domain–new datasets and a method for person entity linkage. *arXiv preprint arXiv:2110.01349* .

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu & Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. Em *15th Conference on Computational Natural Language Learning: Shared task*, 28–34.

Lee, John & Chak Yan Yeung. 2012. Extracting networks of people and places from literary texts. Em *26th Pacific Asia Conference on Language, Information, and Computation*, 209–218.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. Em *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 25–32.

Mota, Cristina & Diana Santos. 2023. Pais, filhos, e outras relações familiares no DIP. *Linguamática* 15(1). 41–53. 10.21814/lm.15.1.402.

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng & Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. Em *52nd Annual Meeting of the Association for Computational Linguistics*, 30–35. 10.3115/v1/P14-2006.

Santos, Diana & Cláudia Freitas. 2019. Estudando personagens na literatura lusófona. Em *XII Symposium in Information and Human Language Technology and Collocates Events (STIL)*, 48–52.

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2022a. Introduction to DIP: goal, setup, resources and results. Apresentação. https://www.linguateca.pt/aval_conjunta/dip/apr_encontro/DIPpresentation.pdf.

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2023. DIP: Desafio de identificação de personagens: objectivo, organização, recursos e resultados. *Linguamática* 15(1). 3–30. 10.21814/lm.15.1.399.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022b. Identifying literary characters in Portuguese: Challenges of an international shared task. Em *Computational processing of the Portuguese language, (PROPOR)*, 413–419. 10.1007/978-3-030-98305-5_39.

Srinivasan, Vardhini & Aurelia Power. 2022. Character extraction and character type identification from summarised story plots. *Journal of Computer-Assisted Linguistic Research* 6. 19–41. 10.4995/jclr.2022.17835.

Vala, Hardik, David Jurgens, Andrew Piper & Derek Ruths. 2015. Mr. Bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 769–774. 10.18653/v1/D15-1088.

Valls-Vargas, Josep, Santiago Ontañón & Jichen Zhu. 2014. Toward automatic character identification in unannotated narrative text. Em *7th Intelligent Narrative Technologies Workshop*, 38–44.

Vilain, Marc, John D Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. Em *6th Message Understanding Conference (MUC)*, 45–52.