# Capítulo 3

# Extração de Informação

*Daniela Barreiro Claro*
*Joaquim Santos*
*Marlo Souza*
*Renata Vieira*
*Vládia Pinheiro*

Publicado em: 26/09/2023
Atualizado em: 16/04/2026



## 3.1 Introdução

A Extração de Informação (EI) é desenvolvida com o objetivo de se obter informação estruturada de dados não-estruturados (Jurafsky; Martin, 2023; Konstantinova, 2014).

Os primeiros trabalhos a debruçarem-se sobre o problema remontam à década de 1970, com a aplicação de gramáticas formais e *parsers* sintáticos para a estruturação de informação em domínios como prontuários médicos (Sager, 1978; Sager et al., 1987) e textos jornalísticos (DeJong, 1979). A comunidade científica demonstrou grande interesse pela área nas décadas posteriores devido à sua utilidade prática, seu foco no processamento de dados reais, suas tarefas bem-definidas e a facilidade de mensurar a qualidade dos resultados em comparação com o desempenho humano na mesma tarefa (Cowie; Lehnert, 1996).

Para autores como Eisenstein (2019) e Jurafsky; Martin (2023), a EI é normalmente dividida em diversas tarefas de interesse, com foco no tipo de informação a ser extraída do texto. Entre as mais comumente citadas na literatura estão o Reconhecimento de Entidades Nomeadas (REN), a Extração de Relações (ER) e a Extração de Eventos (EE).

O Reconhecimento de Entidades Nomeadas (REN) consiste em identificar e classificar entidades mencionadas em textos através de designadores rígidos como nomes próprios, expressões temporais e espécies biológicas (Nadeau, 2007). Esse é considerado por alguns como um primeiro passo na análise semântica de um texto (Santos; Cardoso, 2007a), pois permite identificar as entidades às quais se faz referência nele.

A Extração de Relações (ER), também chamada de extração de informação tradicional ou somente extração de informação, por sua vez, diz respeito à identificação de relacionamentos semânticos entre duas ou mais entidades, ou seja, identificar "quem fez o que para quem e quando". Ananiadou; Mcnaught (2005) a definem como o processo de extrair fatos (em nossa terminologia, relacionamentos) a partir de uma fonte textual e representá-los a partir de um gabarito (em inglês, *template*). As *relações* são elementos essenciais para o entendimento da informação relatada no texto e sua identificação é passo essencial para a estruturação da mesma. Assim, identificar relações entre entidades é tarefa essencial para construção de bases de conhecimento e de grande utilidade na construção de soluções para

a resposta automática a perguntas (em inglês, *query answering*), sumarização, recuperação de informação e mais (Nasar et al., 2021).

A extração de eventos consiste na tarefa de identificação de uma menção a um evento em uma sentença e, se existirem, extração de outras informações sobre o evento. Um evento pode, por sua vez, ser entendido como uma ocorrência específica envolvendo participantes (Consortium, 2005), i.e., algo que acontece e que pode ser descrito como uma mudança de estado da qual participam entidades como agentes. Devido a intrínseca natureza temporal dos eventos, tal problema possui uma natureza mais complexa e costuma possuir tratamento específico.

Assim, nesse capítulo, iniciaremos com um pouco de história da Extração de Informação (EI) e sua evolução para Extração de Informação Aberta, e destacaremos as tarefas de Reconhecimeno de Entidades Nomeadas (REN) e Extração de Relação (ER).

## 3.2 Um pouco de história

Os primeiros trabalhos que abordaram o problema de EI dos quais temos conhecimento surgiram no final da década de 1970. Esses primeiros trabalhos da década de 1970 e 1980 tinham como modelo geral a aplicação de regras para a identificação de informações especificadas em um gabarito. Tais sistemas empregavam analisadores sintáticos (*parsers*) e regras definidas especificamente para o domínio e gênero textual estudado.

Entre esses primeiros trabalhos, estão aqueles de Sager (1978), Sager et al. (1987), de DeJong (1979) e de Cowie (1983). Sager et al. exploraram como identificar informações do estado de saúde de pacientes através dos textos de prontuários médicos. DeJong (1979), por sua vez, descrevem o sistema FRUMP que, a partir de um *parser* e regras de análise conceitual baseadas em uma arquitetura cognitiva proposta pelos autores e no conceito de dependência conceitual de Schank et al. (1973), processavam textos de notícias e realizavam tarefas como sumarização e identificação de papéis semânticos associados aos constituintes da sentença. Cowie (1983), por fim, descreve um sistema que emprega regras simples de segmentação e análise sintática rasa para identificar propriedades de plantas a partir de textos descritivos no campo da botânica. Diferente dos métodos anteriores, o trabalho dos autores se baseia em grande parte no estudo de padrões de descrição das informações a serem identificadas, em detrimento do emprego de *parsers* robustos da língua.

A década de 1990 traz um grande interesse na área de EI com a implementação das conferências MUC (do inglês, *Message Understanding Conference*, ou Conferência de Compreensão de Mensagem), promovidas pela Agência de Projetos de Pesquisa Avançada de Defesa (DARPA, do inglês *Defense Advanced Research Projects Agency*). As conferências MUC, realizadas e financiadas pelo exército americano, representaram um esforço em avançar a tecnologia de EI e consistiam de tarefas de avaliação conjunta de métodos desenvolvidos por pesquisadores para problemas propostos pelos organizadores. As sete conferências realizadas de 1987 a 1997, foram cruciais para definir aspectos centrais da área, como estruturar a tarefa de ER, definindo suas métricas de avaliação, e propor a tarefa de REN (Grishman; Sundheim, 1996).

A partir da MUC-3, em 1991, a conferência passa a ter foco no processamento de textos jornalísticos em detrimento dos relatórios militares utilizados anteriormente (DARPA, 1991). Com a disponibilidade de dados e o incentivo no desenvolvimento de soluções para a tarefa, vemos na década de 1990 o surgimento das primeiras aplicações comerciais de EI, como o JASPER (Andersen et al., 1992)., construído para a agência de notícias Reuters.

A MUC-6, ocorrida em 1995, introduz a tarefa de REN com o intuito de ser uma tarefa de uso prático, independente de domínio e que poderia ser realizada automaticamente em um futuro próximo (Grishman; Sundheim, 1996). Enquanto os trabalhos em REN se avolumaram a partir de sua proposição na MUC-6, trabalhos anteriores como Rau (1991) e Wolinski et al. (1995) já se debruçavam sobre o problema de identificação e classificação de nomes próprios. Desde então, o interesse na tarefa cresceu significativamente e outras conferências de avaliação conjunta têm sido dedicadas a essa tarefa, como a *Automatic Content Extraction* (ACE) e a conferência Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas (HAREM), dedicada exclusivamente à língua portuguesa, com sua primeira edição em 2005 (Santos; Cardoso, 2007a).

Por outro lado, houve um crescimento de abordagens baseadas em dados nesta década, a partir da análise de *corpora*. Tais esforços são impulsionados pelos resultados positivos na área, como o trabalho de Hearst (1992). Logo, métodos baseados em dados passaram também a explorar o emprego de análise estatística e aprendizado de máquina na construção de padrões para a extração de relações (Riloff et al., 1993, 1999; Roark; Charniak, 2000; Soderland et al., 1995)

Não foi somente na extração de padrões que métodos de aprendizado de máquina, em particular aprendizado supervisionado, foram aplicados. A década de 2000 viu a proliferação de métodos supervisionados aplicados à ER (Culotta et al., 2006; Kambhatla, 2004; Zelenko et al., 2003; Zhao; Grishman, 2005) e ao REN (Asahara; Matsumoto, 2003; McCallum; Li, 2003; Sekine, 1998).

Devido à dificuldade de construção de dados para treinamento e padrões para extração, além da pouca adaptabilidade dos sistemas construídos para outros escopos e domínios, nos anos 2000, sistemas baseados em métodos de aprendizado semi-supervisionado, como o DIPRE (Brin, 1998) e Snowball (Agichtein; Gravano, 2000) começaram a aparecer, juntamente com os estudos sobre expansão automatizada de anotações (*bootstrapping*) (Riloff et al., 1999). Também para entidades nomeadas, estudos investigaram como utilizar recursos da Web (Etzioni et al., 2005; Nadeau, 2007) ou *corpora* (Cucchiarelli; Velardi, 2001) para aprender entidades com pouco ou nenhum esforço de anotação.

Buscando superar as dificuldades da limitação de escopo, i.e. das relações-alvo a serem extraídas e categorias de entidades a serem identificadas, ainda restritas à definição de padrões desde a criação dessas tarefas, Banko et al. (2007) propõe a tarefa de extração de informação aberta (EIA), também conhecida como **Open Information Extraction**, OpenIE ou OIE, a qual busca extrair todas as relações possíveis expressas em um texto, sem necessidade de pré-definição de relações e entidades.

Devido ao recente sucesso da aplicação de métodos baseados em redes neurais, em particular *deep learning* e grandes modelos de linguagem, às tarefas de Processamento de Linguagem Natural, uma tendência atual da área se delineou como o estudo de arquiteturas neurais para os problemas de EI e a geração de grandes conjuntos de dados por supervisão fraca. *Surveys* recentes, como (Cui et al., 2018; Konstantinova, 2014; Nasar et al., 2021), nos mostram a evolução da área em direção à aplicação de métodos neurais. Na vertente de geração de dados, vemos o emprego da Wikipédia e Freebase como fontes mais usadas para obter anotações de entidades e relações em textos (Nguyen et al., 2016; Smirnova; Cudré-Mauroux, 2018; Takamatsu et al., 2012).

Porém, toda a tarefa de EI necessita de uma concordância entre as definições de Entidade e Relação. Neste sentido, a próxima seção discute a conceituação de relação adotada neste capítulo, assim como o conceito de entidade.

## 3.3 Conceituação formal: Relação e Entidade

A natureza das relações estudadas na área de Extração de Informação e os critérios para reconhecer sua ocorrência em um texto têm recebido pouca atenção na literatura. Este é um passo importante para estabelecer metodologias adequadas para avaliar os sistemas, bem como para criar conjuntos de dados que possam apoiar a criação de sistemas futuros.

Enquanto as noções de Relação e Entidade são de grande importância e já bem estudadas nas áreas de Computação, Linguística, Ciência da Informação e Filosofia da Linguagem, esses conceitos não são empregados de forma consistente entre as áreas, ou mesmo entre suas subáreas.

### 3.3.1 Entidade

Para Chen (1976), uma entidade é um objeto que pode ser concreto, tal como pessoa, livro, casa ou ainda abstrato, tal como um emprego, um sentimento, uma disciplina. As entidades podem estabelecer relações entre si. Duas ou mais entidades são vinculadas, ou seja conectadas por uma relação[1].

Tradicionalmente em reconhecimento de entidades nomeadas, as entidades consideradas são aquelas referenciadas por um nome próprio, acrescidas das referências temporais e valores que são expressões numéricas. Essas expressões, portanto, geralmente não constituem uma entrada em uma base lexical. Porém a tarefa se expandiu para domínios especializados, onde as entidades de interesse são mais conceituais. No domínio bio-médico por exemplo, podemos ter como exemplo de entidades de interesse, sintomas e tratamento que não são referenciadas por nomes próprios.

### 3.3.2 Relação

Os conceitos de *relação* e *relacionamento* são noções fundamentais que vêm sendo estudadas em áreas como Ciência da Computação, Linguística e Filosofia.

No campo de bancos de dados e modelagem conceitual, Chen (1976) define um relacionamento, no contexto da modelagem de Entidade-Relacionamento, como uma associação entre entidades. Guarino; Guizzardi (2015), por sua vez, estudando a natureza ontológica dos relacionamentos com base na semântica de veridadores (*truthmaker semantics*) (Fine, 2017), postulam relacionamentos como entidades que atuam como veridadores (*thruthmakers*) de alguma proposição relacionando duas ou mais entidades, ou seja, *uma relação mantida entre essas entidades*. Um veridador é um elemento cuja existência torna verdadeira uma proposição particular. Por exemplo, considerando a sentença (1) "*a* é uma maçã", a existência de um objeto denotado pelo nome *a* que por acaso é uma maçã é uma condição suficiente para a verdade da frase (1). Como tal, dizemos que esse objeto é o veridador de (1). Tal definição nos permite adotar critérios ontológicos para validar a existência de relacionamentos a partir da informação relatada em um texto e, por isso, adotaremos tal definição de relacionamento neste capítulo.

O conceito de relações é muito menos consistente na literatura. Ainda na área de modelagem conceitual, Guarino; Guizzardi (2015) definem as relações como proposições para as quais os relacionamentos são veridadores e, portanto, possuem conteúdo proposicional. Assim, podemos entender uma relação como um tipo para entidades como relacionamentos. Ou seja, relações são universais ontológicos que descrevem a natureza dos relacionamentos.

[1] Em nossa terminologia, por um relacionamento.

Xavier et al. (2015), no entanto, argumentam que a noção de relacionamento adotada na área de Extração de Informação é mais geral do que isso, não se limitando àquelas entre objetos e propriedades, mas também àquelas que descrevem ou implicam propriedades de classes gerais como descrito pela sentença (2) "Filósofos são autores de Livros". Assim, para o contexto de EI consideramos relações como tipos de relacionamentos de primeira ou segunda ordem. Isso significa que uma relação é um tipo de relacionamento que existe entre objetos, suas propriedades e classes de objetos ou suas propriedades.

Enquanto os métodos tradicionais de Extração de Informação dependem de um conjunto pré-existente de relações semânticas bem definidas que são relevantes para um domínio específico, a noção de "relação" e "entidade" na literatura da área mais recente, tais como a Extração de Informação Aberta, requer mais aprofundamento por demandar um significado diferente, principalmente com diferente visões de autores. Esta indeterminação terminológica pode trazer problemas para comparar os resultados dos métodos propostos ou para reutilizar os conjuntos de dados criados na área.

As seções seguintes exploram essas duas áreas: Extração de Informação e Extração de Informação Aberta.

## 3.4 Extração de Informação (EI)

A Extração de Informação é caracterizada por obter informação estruturada a partir de textos, sendo entidades ou fatos, i.e. relacionamentos entre entidades, de tipos previamente definidos, conforme exemplo no Quadro 3.1. Métodos com limitação de escopo possuem como uma de suas principais desvantagens a necessidade de intervenção humana para especificar novos fatos a serem extraídos. Esta limitação impede que sistemas de Extração de Informação, doravante denominados de EI tradicional extraiam fatos fora do escopo pré-definido.

Quadro 3.1: Exemplos de relações específicas na EI tradicional

| **Relação Específica** | **Exemplo de Sentença** | **Extração** |
|---|---|---|
| location-of(algo/alguém, local) | Um aluno pode ser encontrado na escola | location-of(aluno, escola) |
| is-a(subclasse, superclasse) | Salvador é uma cidade | is-a(Salvador, cidade) |
| part-of(todo, parte) | Roda é um componente de um carro | part-of(roda, carro) |

Fonte: (Souza; Claro, 2014)

### 3.4.1 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas (REN) consiste na tarefa de identificar e classificar expressões linguísticas, denominadas entidades nomeadas (EN), que referenciam entidades específicas num domínio de discurso, como nomes próprios, expressões temporais e espécies biológicas (Mota et al., 2007; Nadeau, 2007). De uma forma geral, o REN pode ser dividido em duas etapas: a identificação (ou delimitação) da expressão, na qual as palavras que formam a EN são selecionadas; a classificação, em que é atribuída uma categoria semântica à EN.

A classificação das ENs determina os tipos de entidades a serem consideradas e são especificadas a partir do escopo definido previamente para a tarefa. Algumas das categorias mais comumente utilizadas incluem as entidades que referenciam Pessoas Singulares (antropônimos); Coletivas (empresas e organizações) e Lugares (topônimos) (Mota et al., 2007). Para exemplificar tomemos a sentença: "Renata Silva e Maria Costa palestraram na Universidade Federal da Bahia". No exemplo temos três ENs: "Renata Silva", "Maria Costa", "Universidade Federal da Bahia", sendo as duas primeiras correspondentes à categoria semântica Pessoa e a última, à categoria semântica Organização. Entretanto, existem outras categorias de ENs, como as menções a Obras (por exemplo, "Código Da Vinci"); Acontecimentos (por exemplo, "Festa de Santo Antônio"), Tempo (por exemplo, "meio-dia"); Coisa (por exemplo, "barco"), entre outras.

O REN é uma tarefa com grande importância para o Processamento de Linguagem Natural, pois consiste numa primeira tarefa de análise semântica de um texto, com potencial aplicações a diversas tarefas. Por exemplo, em sistemas de perguntas e respostas, as perguntas frequentemente se referem a informações sobre entidades. Também, métodos de identificação de estruturas mais complexas, como eventos ou relações, dependem do bom desempenho do REN como uma etapa de pré-processamento (Socher et al., 2012; Zelenko et al., 2003).

### 3.4.2 Extração de Relações

A tarefa de extração de relações (ou de relacionamentos) (ER) refere-se a identificar relacionamentos entre entidades de um determinado escopo mencionadas em um texto (Jurafsky; Martin, 2023). O escopo, no contexto da ER, refere-se a um conjunto de relações-alvo de um determinado domínio de conhecimento ou aplicação a ser investigado. Por exemplo, o Quadro 3.2 apresenta alguns exemplos de relações no domínio de geografia brasileira. Na descrição das relações, os elementos em negrito referem-se às entidades em um dado relacionamento descrito pelo termo em itálico.

Quadro 3.2: Exemplos de relações no domínio da geografia brasileira.

| **Relação** | **Descrição** | **Exemplo** |
|---|---|---|
| Pertence(Cidade, Unidade Federativa) | Sobre uma **cidade** que está localizada em uma determinada **Unidade Federativa**, dizemos que a primeira *pertence* a esta última. | Pertence(Salvador, Bahia) |
| Tem_Prefeito(Cidade, Pessoa) | Uma **pessoa** que executa a função administrativa de gestão do executivo em nível municipal de uma dada **cidade** é denominada de seu(sua) *prefeito(a)*. | Tem_Prefeito(Salvador, Bruno Reis) |
| Fundação(Cidade, Data) | A **data** em que uma **cidade** foi fundada, é dita sua data de *fundação*. | Fundação(Salvador, 29 de março de 1549) |

Nesse contexto, a delimitação de um escopo ou domínio de interesse, concentra-se na determinação das relações a serem processadas, i.e. nos tipos de relacionamentos de interesse, assim como da natureza das entidades associadas por tais relações.

### 3.4.3 Extração Conjunta de Entidades e Relações

As tarefas de reconhecimento de entidades nomeadas e extração de relações são interdependentes, no sentido de que a definição do escopo a ser estudado delimita tanto as categorias e natureza das entidades a serem extraídas, como também as relações entre essas entidades. Também, note-se que, pelo fato de as relações serem comumente definidas entre entidades de tipo especificado, como o caso da relação *Tem_Prefeito* no Quadro 3.2 que ocorre entre entidades das classes **Cidade** e **Pessoa**, tanto as informações das entidades mencionadas no texto são úteis para a extração de relações, quanto a informação das relações identificadas pode ser útil ao processo de identificação de entidades.

De fato, na literatura recente, existem vários trabalhos que consideram a tarefa de extração conjunta de entidades e relações (ERE, do inglês *Entity and Relation Joint Extraction*), composta das tarefas de REN e ER (Agichtein; Gravano, 2000; Shaowei et al., 2022; Yuan et al., 2021b). Enquanto normalmente abordagens estruturam suas soluções de forma sequencial, usualmente realizando REN inicialmente e, posteriormente, realizando ER, como nos trabalhos de (Hasegawa et al., 2004) e de (Socher et al., 2012), a literatura recente aponta para as vantagens da identificação conjunta ao permitir um melhor aprendizado de restrições para identificação de entidades e relações, c.f. o recente *survey* realizado por (Shaowei et al., 2022) sobre métodos para tal tarefa.

### 3.4.4 Métodos empregados para EI na literatura

Várias abordagens foram adotadas para o problema de EI durante seu desenvolvimento histórico. Enquanto abordagens iniciais privilegiavam métodos ricos em conhecimento, como regras e recursos linguísticos e de conhecimento de mundo, a literatura recente na área privilegia métodos baseados em dados, como o aprendizado de máquina, com o recente emprego de arquiteturas neurais aos problemas.

A seguir faremos uma breve apresentação das abordagens descritas na literatura para os problemas de EI.

#### 3.4.4.1 REN

As abordagens iniciais para REN baseavam-se, majoritariamente, no emprego de regras léxico-sintáticas e consulta a almanaques (*gazetteers*). Tais abordagens dependem da construção de listas de nomes próprios como antropônimos, topônimos etc., e outras palavras, como "Ltda.", "Jr." etc., que auxiliam no processo de identificação e classificação de ENs complexas ou desconhecidas. Essa é, por exemplo, a abordagem empregada por Wolinski et al. (1995) que combina almanaques e regras para a identificação e classificação de ENs. Posteriormente, almanaques foram também empregados em conjunção com métodos baseados em dados, como o trabalho de Florian et al. (2003) que os emprega aliados aos classificadores, enquanto Liu et al. (2019a) os utilizam durante o treinamento de uma rede neural, como um sinal de treinamento (parte da função de perda, ou *loss* em inglês).

Muitos trabalhos debruçaram-se também sobre o problema de construção automática ou semi-automática de almanaques, dos quais os trabalhos de Nadeau (2007), de Riloff et al. (1999) e de Etzioni et al. (2005) são alguns dos mais importantes.

Enquanto as abordagens iniciais para o problema baseavam-se em regras, com a disponibilidade de dados anotados para a tarefa, tais métodos foram rapidamente suplantados por métodos baseados em dados, tais como: os métodos baseados em classificação (Asahara;

Matsumoto, 2003; Sekine, 1998) e classificação sequencial (Bikel et al., 1999; McCallum; Li, 2003).

A redução de REN à tarefa de classificação sequencial merece destaque pelos bons resultados obtidos. Tal redução se dá através de um esquema de codificação do problema que nos permite representar fragmentos textuais e sua classificação como um problema de rotulação ou etiquetação.

Partindo-se do pressuposto de que os fragmentos textuais descrevendo entidades nomeadas são contíguos, podemos codificar a tarefa de delimitação de entidades como classificação sequencial empregando rótulos que descrevem os limites de uma EN, e.g. o esquema BIO com os rótulos **B** (do inglês, *begin*) para designar a palavra inicial de uma EN, **I** (do inglês, *inside*) para designar palavras que fazem parte da EN mas não a iniciam e **O** (do inglês, *outside*) para designar palavras que não pertencem a uma entidade. Da mesma forma, podemos estender nosso esquema de codificação para incluir as classes de interesse. Assim, seguindo o esquema BIO, teremos os rótulos **B-PER** e **I-PER** para descrever entidades da classe **Pessoa**.

A redução do problema de REN à classificação sequencial está ilustrada no Exemplo 3.1.

Exemplo 3.1:

> Renata/B-PER Silva/I-PER e/O Maria/B-PER Costa/I-PER palestraram/O na/O Universidade/B-ORG Federal/I-ORG da/I-ORG Bahia/I-ORG.

Recentemente, destacam-se na literatura abordagens baseadas em redes neurais profundas, com uma grande concentração nos últimos anos em modelos gerativos de linguagem, devido aos resultados positivos obtidos por tais arquiteturas em diversas tarefas.

Na literatura são de grande destaque os modelos recentes BART (Lewis et al., 2020), RoBERTa (Liu et al., 2019c), T5 (Raffel et al., 2020), BERT (Devlin et al., 2019) e GPT-3 (Brown et al., 2020c), conforme descritos no Capítulo Modelos de linguagem.

Similarmente, na língua portuguesa, nas duas edições do HAREM (Mota; Santos, 2008; Santos; Cardoso, 2007b), o primeiro esforço sistemático de desenvolvimento de soluções para a tarefa na língua, a maioria dos sistemas participantes baseava-se em métodos ricos em conhecimento, como regras e almanaques. De fato, nas duas avaliações, somente os sistemas MALINCHE (Solorio, 2007), NEURA (Ferrández et al., 2007) e R3M (Mota, 2008) não se baseavam em regras. Métodos baseados em classificação sequencial se seguiram para a língua portuguesa, como o RELP-CRF (Amaral; Vieira, 2014) baseado em um classificador sequencial. Mais recentemente, abordagens baseadas em redes neurais e modelos de linguagem foram desenvolvidas tornando-se o estado da arte da tarefa na língua. A Tabela 3.1 apresenta o atual estado da arte em português, com base no *corpus* HAREM. A métrica de avaliação apresentada, medida F1, será discutida na Seção Avaliação.

Tabela 3.1: Trabalhos estado da arte no REN em português

| Modelo | Medida F1 |
|---|---|
| BERT-CRF (Souza et al., 2020a) | 83,70% |
| BiLSTM-CRF+FlairBBP (Santos et al., 2019b) | 82,26% |
| LSTM-CRF (Castro et al., 2018) | 76,27% |
| CharWNN (Santos; Guimarães, 2015) | 65,41% |

Souza et al. (2020a) desenvolveram um modelo BERT para o Português com 2,68 bilhões de *tokens* e aplicaram o modelo em um classificador CRF. Santos et al., avaliaram o

impacto do modelo contextualizado Flair Embeddings aplicado a tarefa de REN junto com uma rede neural BiLSTM-CRF. Os autores também desenvolveram um modelo Flair Embeddings para o português, o FlairBBP, treinado com 4,9 bilhões de *tokens* (Santos et al., 2019b). Castro et al. (2018) utilizou uma rede LSTM e um classificador CRF junto com modelos *word embeddings* pré-treinados. Santos; Guimarães (2015) desenvolveram uma rede neural convolucional capaz de capturar características a nível de caracteres e também de incorporar *word embeddings* pré-treinados.

#### 3.4.4.1.1 Reconhecimento de Entidades em Domínios Específicos

O reconhecimento de entidades tem sido aplicado em muitas áreas específicas, como direito, saúde e geologia. Nesses casos há uma demanda de adaptação dos modelos preditivos de acordo com a nova linguagem especializada do domínio e um novo conjunto de rótulos que devem ser aprendidos. Da mesma forma, são necessários novos conjuntos de dados para o processo de aprendizado, uma vez que abordagens de aprendizado de máquina necessitam de exemplos anotados para se chegar a um modelo preditivo eficaz.

Muitos trabalhos endereçam domínios específicos, citamos exemplos em diversas línguas. Para o inglês, uma rede neural BiLSTM-CRF para o domínio biomédico é proposta em (Habibi et al., 2017).

Um conjunto de dados do domínio jurídico em língua alemã é apresentado por Leitner et al. (2019), que empregam redes neurais BiLSTM para a rotulação dos textos. Em (Qiu et al., 2019), uma rede neural BiLSTM-CRF com mecanismo de atenção é aplicada para reconhecer entidades geológicas para a língua chinesa.

Para o português, um *corpus* para detecção de eventos de quedas de pacientes em prontuários eletrônicos é descrito em (Santos et al., 2020a). Os autores usaram uma rede neural BiLSTM-CRF+Flair para gerar um modelo classificador de *tokens*. Um *corpus* no domínio jurídico, tendo categorias específicas como legislação e jurisprudência é proposto por Araujo et al. (2018), que usaram uma rede neural BiLSTM-CRF para criar um primeiro *baseline* para esse *corpus*. Ademais, Consoli et al. (2020) analisam um *corpus* no domínio de geologia usando uma rede neural BiLSTM-CRF com um modelo contextualizado Flair Embeddings.

Desde a terceira edição, este livro conta com um capítulo de reconhecimento de entidades nomeadas no domínio jurídico, o Capítulo Reconhecimento de Entidades Nomeadas no Domínio Legal.

### 3.4.4.2 Extração de Relações

As abordagens iniciais para o problema de ER baseavam-se na definição de gabaritos e regras de extração, com base em informação sintática obtida de analisadores sintáticos rasos ou profundos (Cowie, 1983; Sager, 1978). Tais métodos foram rapidamente suplantados por métodos baseados em dados e padrões obtidos de *corpora*, como os famosos padrões de Hearst (1992) para identificação de relações de hiponímia.

O trabalho de Hearst (1992) se baseou na definição de padrões léxico-sintáticos para expressão de relações de hiponímia e hiperonímia a partir de uma análise de *corpus*. Ao escolher a relação de hiponímia, que ocorre em todo domínio, e padrões gerais baseados em aspectos da língua, como os representados no Quadro 3.3, garante-se a generalização dos padrões obtidos para diversos domínios e aplicações.

Quadro 3.3: Exemplos de Padrões de Hearst para hiponímia

| **Padrão** | **Exemplo** | **Extração** |
|---|---|---|
| $NP$, tais quais $\{NP\}^* NP$ | … países, tais quais França, Brasil e China | Is_a(França, país) |
| $NP\{, NP\}^*$ e outros(as) $NP$ | Contusões, feridas, osso quebrados e outras lesões | Is_a(contusão, lesão) |

Devido à dificuldade de construção manual das regras, os métodos de Riloff et al. (1993), empregam heurísticas para geração de padrões baseadas em informação gramatical, e de Soderland et al. (1995), que se baseia numa semântica de quadros (*frames*) empregando um analisador semântico e medidas de qualidade de identificação de exemplos, baseado no percentual de acerto sobre relacionamentos previamente conhecidos, para identificação de quadros relevantes.

As abordagens baseadas em aprendizado de máquina, hoje as mais comuns e com melhor desempenho na literatura (Konstantinova, 2014; Nasar et al., 2021) dividem-se em abordagens que realizam reconhecimento de entidades e extração de relações de forma conjunta e separada.

Abordagens baseadas na realização de REN e ER de forma separada baseiam-se em um fluxo de processamento em que, em geral, as entidades são identificadas primeiro e a tarefa de ER se reduz a identificar quando uma sentença ou fragmento textual denota uma relação semântica entre duas entidades. Consideremos o Exemplo 3.2, retirado de (Socher et al., 2012):

Exemplo 3.2:

> [Gripe aviária]$_{e1}$ é uma doença infecciosa causada pelo vírus da [influenza tipo a]$_{e2}$

Podemos, então, reduzir o problema de identificar a relação Causa-Efeito($e1$,$e2$) a um problema de classificação textual, identificando se a sentença acima fornece indícios para a expressão da relação de interesse. As soluções propostas na literatura para o problema são variadas e baseadas em diferentes métodos.

Zelenko et al. (2003), por exemplo, propõem funções de *kernel* para árvores sintáticas rasas, i.e. funções que descrevem medidas de similaridade entre tais árvores. Eles empregam tais medidas para treinar um classificador de *perceptron* com votação (*voted perceptron*) sobre relações no domínio de organizações extraídas de um *corpus* de textos jornalísticos. De forma similar, Zhao; Grishman (2005) empregam diferentes funções de *kernel* sobre informações sintáticas relevantes para a identificação de relação e argumentos visando treinar um classificador SVM sobre o *corpus* de ER da conferência ACE.

Culotta et al. (2006), por outro lado, empregam um classificador sequencial baseado em modelos escondidos de Markov para identificação de relações em um texto. Ao restringir sua análise a textos biográficos, os autores reduzem o processo de identificar instâncias de relações à identificação de fragmento textual que delimita o argumento e sua classificação, tarefa para a qual a classificação sequencial já é comumente utilizada. Consideremos o Exemplo 3.3 sobre George W. Bush, retirado de (Culotta et al., 2006):

Exemplo 3.3:

George é filho de $\underbrace{\text{George H. W. Bush}}_{\text{pai}}$ e $\underbrace{\text{Barbara Bush}}_{\text{mãe}}$.

Ao identificar o papel de pai e mãe, os autores conseguem construir a relação Pai(George H. W. Bush, George W. Bush) e Mãe(Barbara Bush, George W. Bush).

Métodos baseados em redes neurais, de forma geral, costumam empregar técnicas de aprendizado de representação (Bengio et al., 2013) para aprender representações do conteúdo semântico dos fragmentos textuais e reduzem o problema de ER à classificação textual. É o caso de Socher et al. (2012), que propõem a MV-RNN, uma rede neural que constrói um espaço de representação baseado em matrizes e vetores com o objetivo de capturar a composicionalidade de sentido de sintagmas e sentenças e os aplica para ER. Similarmente, Zeng et al. (2014) e Wang et al. (2016) empregam redes neurais convolucionais para obter representações vetoriais de sentenças que serão empregadas no processo de classificação quanto à relação expressa pela mesma.

#### 3.4.4.3 Extração Conjunta de Entidades e Relações

Abordagens baseadas em identificação sequencial de entidades e relações possuem desvantagens observadas na literatura. Primeiramente, como a ER é guiada pelas entidades identificadas no processo de REN, a propagação de erros da primeira tarefa pode ter impacto considerável na performance dos sistemas desenvolvidos. Segundo, uma vez que o contexto determinado limita tanto as tarefas de REN, quanto as de ER, existe uma interdependência entre as tarefas. Assim, propostas visando realizar a extração de entidades e relações de forma conjunta começaram a surgir na literatura recente, ganhando certo interesse da comunidade.

As abordagens empregadas para tal tarefa são diversificadas, incluindo desde métodos de aprendizado relacional a redes neurais

Roth; Yih (2007) propõem a utilização de métodos de programação inteira ao problema, baseados na teoria estatística de aprendizado relacional. Os autores utilizam classificadores locais para a identificação de entidades e relações e um classificador global que combina as informações dos classificadores locais em uma predição que maximiza a qualidade da extração, codificada por meio de restrições em programação inteira. Também baseados em modelos estatísticos, Yu; Lam (2010) propõem o uso de modelos gráficos globais para identificação de um descritor de relação e uma segmentação do texto para identificação dos argumentos.

Li; Ji (2014) e Miwa; Bansal (2016), por sua vez, reduzem a tarefa de ERE à classificação sequencial, utilizando redes neurais recorrentes bidirecionais sequenciais e estruturadas com base na estrutura superficial e na árvore de dependências sintáticas da entrada para identificação conjunta de entidades e relações.

## 3.5 Extração de Informação Aberta

A Extração de Informação Aberta (EIA), também conhecida como *Open Information Extraction*, Open IE ou OIE em inglês, é a tarefa de extrair informações estruturadas de documentos sem necessitar da pré-definição do contexto da tarefa, i.e. das relações e tipos de entidade de interesse. A tarefa foi inicialmente proposta pelo trabalho de (Banko et al., 2007) e ganhou popularidade nas últimas décadas devido à sua aplicabilidade para

processar e estruturar o conhecimento a partir de grandes volumes de dados disponíveis na Web, seguindo o paradigma da Web como um *Corpus* (WaC) (Meyer et al., 2003).

A EIA surge visando generalizar a tarefa de Extração de Relações. A principal diferença entre as duas abordagens, porém, reside na dependência da ER de uma especificação prévia do domínio de aplicação, bem como das relações alvo a serem identificadas, que a EIA visa eliminar.

Seguindo o trabalho original de Banko et al. (2007), que propôs o sistema TextRunner, vários métodos e sistemas para EIA foram propostos na literatura (Del Corro; Gemulla, 2013; Fader et al., 2011; Xavier et al., 2015), mas, como observado por Glauber; Claro (2018), os principais avanços na área se concentraram principalmente no idioma inglês.

A EIA para a língua portuguesa tem uma história bastante recente. A partir dos trabalhos de Souza; Claro (2014), Pereira; Pinheiro (2015) e de (Barbosa et al., 2016), têm crescido o número de estudos sobre a tarefa assim como os resultados obtidos por esses estudos, com recentes desenvolvimentos de métodos (Oliveira et al., 2022c; Sena et al., 2017; Sena; Claro, 2019, 2020; Souza et al., 2018), construção do *corpus* (Glauber et al., 2018) e avaliação dos sistemas disponíveis (Glauber et al., 2019a, 2019b; Malenchini et al., 2019).

Embora a área tenha visto um crescimento recente para o desenvolvimento de métodos para línguas como o inglês, principalmente com a aplicação de métodos supervisionados e redes neurais, esses avanços ainda não foram incorporados na literatura sobre EIA para a língua portuguesa. A razão para isso é principalmente a falta de recursos linguísticos disponíveis para orientar o desenvolvimento de pesquisas para a língua. Embora o foco no idioma inglês possa ser devido ao seu uso generalizado em todo o mundo, foi reconhecido pela comunidade científica que esse foco no inglês com suas características particulares pode introduzir algum viés na área (Bender, 2009).

Assim, esta seção aborda EIA para a língua portuguesa, incluindo uma formalização e a evolução das abordagens da área.

### 3.5.1 Formalização

A tarefa de EIA pode ser formalmente definida sendo $X = \langle x_1, x_2, \cdots, x_n \rangle$ uma sentença composta de *tokens* $x_i$. Um extrator EIA é uma função que mapeia $X$ em um conjunto $Y = \langle y_1, y_2, \cdots, y_j \rangle$ como um conjunto de tuplas $y\_i = \langle rel_i, arg1_i, arg2_i, \cdots, argn_i \rangle$, que descrevem as informações expressas na sentença X. Neste capítulo, consideramos que as tuplas estão sempre no formato $y = (arg_1, rel, arg_2)$, onde $arg1$ e $arg2$ são sintagmas nominais, não necessariamente formados por *tokens* presentes em X, e $rel$ é um descritor de um relacionamento entre $arg_1$ e $arg_2$. Não consideraremos extrações formadas por mais de dois argumentos neste capítulo.

### 3.5.2 Abordagens

Os primeiros métodos de EIA empregavam padrões de inspiração linguística para extração, como ArgOE (Gamallo; Garcia, 2015), ou adaptação de métodos para a língua inglesa, como SGS (Souza et al., 2018), InferReVerbPT Sena et al. (2017) e RePort Pereira; Pinheiro (2015). Os trabalhos são principalmente influenciados por métodos baseados no inglês da chamada segunda geração de EIA (Fader et al., 2011).

O primeiro sistema de EIA para o português de que temos conhecimento foi o DepOE (Gamallo et al., 2012). Ele executa a extração aberta multilíngue de triplas (inglês, espanhol, português e galego) usando o analisador sintático de dependências baseado em

regras *DepPattern*. No entanto, nenhuma avaliação ou resultados são relatados para a língua portuguesa. Os autores apresentam somente uma comparação dos seus resultados com *Reverb* na língua inglesa.

Souza; Claro (2014) se propuseram a analisar o conjunto de características mais representativas da língua portuguesa para a identificação de extrações válidas no contexto de EIA, tal qual empregado na língua inglesa com o sistema ReVerb (Fader et al., 2011).

O sistema RePort (Pereira; Pinheiro, 2015), por outro lado, é uma adaptação do ReVerb para a língua portuguesa baseada em análise sintática rasa com regras sintáticas e lexicais. Os autores relatam que suas extrações apresentam grande similaridade com suas correlatas extraídas pelo ReVerb (dos textos traduzidos para o inglês).

O RELP, proposto por Abreu; Vieira (2017), é um sistema aberto de extração de relações que extrai relações entre entidades nomeadas em um domínio de organização aplicando classificação sequencial com CRF (*Conditional Random Fields*). O sistema RelP extrai qualquer descritor de relação que expressa um relacionamento entre pares de entidades nomeadas (Organização, Pessoa ou Lugar), caracterizando-o como uma abordagem híbrida da REN com a EIA.

O InferReVerbPT desenvolvido por Sena et al. (2017) baseia-se numa adaptação do sistema ReVerb para a língua portuguesa, expandindo-o com a extração de relacionamentos implícitos obtidos por inferência por propriedades de simetria e transitividade das relações com inferência transitiva e simétrica. Um classificador SVM foi empregado para realizar a inferência baseado nas propriedades semânticas do verbo central no descritor de relação.

Souza et al. (2018) analisaram que a maior desvantagem dos estudos baseados em recursos linguísticos, como dados anotados, reside na escassez de tais recursos na maioria dos idiomas além do inglês. Assim, para mitigar esse problema, eles propõem um método de classificação de fatos baseado na similaridade de estruturas gramaticais (SGS). Sua abordagem modela estruturas morfosintáticas dos fatos (triplas descrevendo relacionamentos) para identificar padrões de semelhanças que podem ser usados para distinguir entre fatos válidos e inválidos. Eles aplicaram algoritmos de isomorfismo de grafos para detectar subgrafos descrevendo tais padrões.

Um novo sistema de EIA baseado em análise de dependência foi proposto por Gamallo; Garcia (2015), chamado ArgOE. Tal sistema é multilíngue, baseado em heurísticas e utiliza a informação de dependência sintáticas do texto para analisar a estrutura de dependência do verbo, bem como um conjunto de regras para gerar os relacionamentos. A introdução de um Analisador de Dependência em sistemas de EIA focados inteiramente na língua portuguesa foi feita pelos autores Oliveira et al. (2022c). O DptOIE é baseado em análise de dependência e regras elaboradas manualmente. As sentenças são pré-processadas por meio de um tokenizador, um PoS *Tagger* e um analisador de dependências. Os autores propõem um acoplamento de três módulos para tratar casos particulares: conjunções coordenadas, orações subordinadas e aposto.

Com a evolução dos métodos de EIA para a língua inglesa utilizando os modelos neurais, novas abordagens foram propostas também para a língua portuguesa.

O primeiro trabalho que utilizou aprendizado supervisionado com rede neural profunda para o português foi o de Ro et al. (2020) que descreve o sistema **Multi2OIE**. Os autores utilizaram o modelo de linguagem BERT multilíngue (Devlin et al., 2019) para obter representações vetoriais das palavras e reduzem a tarefa de EIA à classificação sequencial, identificado os fragmentos do texto que determinam os argumentos ($arg_1, arg_2$) e o descritor de relação ($rel$). Seu sistema foi capaz de produzir extrações para vários idiomas (inglês, português e espanhol), treinados, entretanto, sobre dados traduzidos do inglês.

Stanovsky et al. (2018) propuseram uma abordagem de EIA para a língua inglesa baseada em triplas. Os mesmos fazem uso de uma classificação sequencial cuja limitação define uma tripla extraída para cada sentença. Este método utiliza uma arquitetura de Redes Neurais Recursivas (RNN) para realizar EIA. A EIA é formulada como uma tarefa de rotulagem de sequências, utilizando estratégias semelhantes às que foram aplicadas anteriormente a tarefas como o Reconhecimento de Entidades Nomeadas. Já os autores em Cui et al. (2018) e Zhang et al. (2017) propõem modelar o problema da EIA como um problema de aprendizado sequência a sequência (*seq2seq*). Eles definem uma estrutura *encoder-decoder* para aprender argumentos e tuplas de relação inicializadas a partir de um sistema de EIA.

Seguindo o trabalho de (Stanovsky et al., 2018), em 2022, Cabral et al. (2022) propuseram **PortNOIE**, uma arquitetura neural para EIA em português que combina representações contextuais de palavras com codificadores neurais para extrair relacionamentos baseado em classificação sequencial iterativa. Diferente de outros métodos de classificação sequencial para EIA, os autores focam na extração de múltiplas triplas de uma mesma sentença.

## 3.6 Avaliação

A avaliação sistemática de sistemas de EI foi estabelecida primeiramente nas conferências MUC, em particular na sua segunda edição, com o estabelecimento de gabaritos-padrão que deveriam ser utilizados por todos os sistemas participantes e a adoção de métricas de qualidade, baseadas naquelas usadas na área de recuperação de informação, que foram abordadas no Capítulo Recuperação de Informação. Para avaliar a tarefa de extração de relações, a MUC-2 estabeleceu como métricas de qualidade do sistema as medidas de precisão e cobertura, também denominada de *Recall* ou Revocação.

A precisão de um sistema reflete a qualidade de suas extrações, i.e., quantas das extrações realizadas estão corretas, dado um *corpus* de teste. A medida de precisão pode ser calculada como:

$$P = \frac{\#(\text{relacionamentos corretamente extraídos})}{\#(\text{relacionamentos extraídos pelo sistema})} \tag{3.1}$$

A cobertura também conhecida como revocação, reflete quão abrangente um sistema é em suas extrações, i.e., quantas das extrações a serem realizadas em um *corpus* de teste, o sistema é capaz de realizar. A medida de cobertura pode ser calculada como:

$$R = \frac{\#(\text{relacionamentos extraídos})}{\#(\text{relacionamentos no } \textit{corpus})} \tag{3.2}$$

Enquanto a MUC-3 adicionou duas novas métricas de avaliação, a saber sobre-geração (*overgeneration*) e sub-geração (*fallout*), tais métricas receberam pouco interesse na literatura. De fato, Lehnert; Sundheim (1991) argumentam que tais métricas foram pouco informativas ou difíceis de calcular para a tarefa de EI e, portanto, abandonadas. Foi também empregado nessa conferência um sistema automático de avaliação disponibilizado às equipes participantes que permitiu uma maior compreensão do modelo de avaliação e, como discutem Lehnert; Sundheim (1991), um avanço qualitativo nos sistemas gerados.

Além das medidas de precisão e cobertura, assim como em tarefas de classificação de texto e recuperação de informação, utilizamos a média harmônica entre essas medidas, chamada medida F1, a fim de condensar a informação contida nas duas. A medida F1 pode ser calculada como:

$$F1 = \frac{2 * P * R}{P + R} \tag{3.3}$$

A avaliação da tarefa de REN segue padrões semelhantes aos aplicados à tarefa de ER. De fato, desde a MUC-6 (Grishman; Sundheim, 1996), as medidas de precisão, cobertura e F1 tem sido usada consistentemente como métricas de avaliação da tarefa de REN em diversos esforços de avaliação, como a CoNNL (Sang; De Meulder, 2003), para a língua inglesa, e das duas edições do HAREM (Gonçalo Oliveira et al., 2008; Santos et al., 2007), com excessão à ACE (Doddington et al., 2004) que apresenta uma combinação da tarefa de REN com reconhecimento de co-referência entre entidades e utiliza um sistema de pontuação próprio.

A avaliação de sistemas de EIA, por sua vez, possui algumas peculiaridades que precisam ser discutidas. Uma vez que a tarefa é postulada por Banko et al. (2007) como a extração de todas as relações identificadas em um dado fragmento textual, sem limitação de domínio de interesse, tal tarefa impõe imensa dificuldade aos esforços de avaliação.

De fato, Glauber et al. (2018) relatam um esforço de anotação de dados para a tarefa em língua portuguesa em que foram identificados por anotadores humanos mais de 400 relacionamentos em um *corpus* de 25 sentenças retiradas de textos jornalísticos e de enciclopédia. Assim, a avaliação de EIA deu-se, em grande parte de seu desenvolvimento e maturação, em conjuntos de dados não anotados, recorrendo a avaliações qualitativas das saídas dos sistemas e comparação direta por humanos das extrações obtidas.

Nesses esforços de avaliação, a precisão do sistema pode ser mensurada a partir da avaliação humana das saídas. Não é possível, entretanto, avaliar medidas como cobertura e F1, dada a inexistência de uma referência do conjunto total de relacionamentos a serem identificados. Assim, os autores da área propuseram diferentes métricas a fim de estimar tais valores, como a métrica rendimento (*yield*) (Fader et al., 2011; Schmitz et al., 2012).

A métrica de rendimento consiste no núemro de extrações válidas, i.e. corretas, de um dado sistema. Como calcular tal medida é, na maioria dos casos, impraticável dada a grande quantidade de extrações realizadas pelos sistemas, ela pode ser estimada a partir da precisão do sistema calculada sobre uma amostra aleatória das extrações realizadas ($P'$). Assim, podemos estimar o rendimento como:

$$Y = P' \cdot \#(\text{extrações realizadas}) \tag{3.4}$$

Foi também explorada a estratégia de criação (semi-)automática de conjuntos de dados usando vários sistemas (Del Corro; Gemulla, 2013), estratégias de supervisão fraca (Smirnova; Cudré-Mauroux, 2018), ou a geração de *corpora* para a tarefa a partir da transformação de anotações de tarefas próximas, como identificação de papéis temáticos (*Semantic Role Labeling*) por (Stanovsky et al., 2018). *Corpora* gerados de forma semi-automática vêm ganhando atenção na literatura recente, particularmente para a língua inglesa, devido a necessidade de dados anotados para se utilizar técnicas de aprendizado de máquina e redes neurais em EIA. *Corpora* como o OIE2016 (Stanovsky et al., 2018), Wire57 (Léchelle et al., 2018) e CARB (Bhardwaj et al., 2019) vêm se tornando *corpora* de referência em língua inglesa para o problema, apesar dos problemas existentes na construção de tais recursos – a não exaustividade das relações anotadas.

Para a língua portuguesa, foram propostas algumas iniciativas para avaliar os sistemas da OIE. Uma avaliação conjunta foi promovida durante o Fórum Ibérico de Avaliação de Línguas (IberLEF) em 2019 (Collovini et al., 2019). A avaliação foi feita usando o *corpus* proposto por Glauber et al. (2018), que é composto por 442 relacionamentos extraídos

de 25 frases de fontes como a seção em português da Wikipédia, o *corpus* CETENFolha, resenhas de filmes do portal Adoro Cinema[2] e o *corpus* Europarl. Apesar desta tarefa ter contemplado quatro cenários de avaliação, a avaliação geral dos sistemas permaneceu consistente nos diferentes cenários, indicando robustez nos resultados da avaliação. No geral, os sistemas DPTOIE (Oliveira et al., 2022c) e Linguakit (Gamallo; Garcia, 2015) tiveram o melhor desempenho, com o Linguakit2 dominando as avaliações de correspondência exata e o DPTOIE as avaliações de correspondências parciais (Collovini et al., 2019).

Outra abordagem de avaliação foi idealizada por (Malenchini et al., 2019). Seu foco foi a avaliação extrínseca dos sistemas de EIA através de sua contribuição na tarefa de respostas automáticas a perguntas. Os autores apresentaram um conjunto de dados de referência (*benchmark*) para avaliação extrínseca de sistemas de EIA em textos de língua portuguesa. Os sistemas que alcançaram os melhores valores na avaliação realizada pelos autores foram os sistemas ArgOE (Gamallo; Garcia, 2015), DependentIE (Glauber et al., 2019a) e DptOIE (Oliveira et al., 2022c).

## 3.7 Considerações finais

Este capítulo descreveu uma visão geral da área de Extração de Informação, apresentando a Extração de Informação Tradicional e a Extração de Informação Aberta. Transversalmente, apresentamos as formalizações necessárias e os conceitos fundamentais para a compreensão da EIA, assim como a avaliação da área e as heranças de outras áreas afins, tais como RI.

Nessa primeira versão, este capítulo descreveu de maneira bem sucinta as abordagens propostas para EI e EIA durante seu desenvolvimento histórico e as abordagens atuais da literatura, como as utilizando modelos de linguagens. Especificamente, a utilização da arquitetura Transformers, descritas no Capítulo Modelos de linguagem para as tarefas de EI e EIA tem sido bastante difundida para a língua inglesa e tem atuado em diversas áreas da PLN.

## Agradecimentos

Agradecemos as colaborações dos autores deste Capítulo e suas indicações, assim como agradecemos a Adriana Pagano e Aline Macohin pela revisão e comentários.

## Referências

ABREU, S. C. DE; VIEIRA, R. Relp: Portuguese open relation extraction. **KO KNOWLEDGE ORGANIZATION**, v. 44, n. 3, p. 163–177, 2017.

AGICHTEIN, E.; GRAVANO, L. **Snowball: Extracting relations from large plain-text collections**. Proceedings of the fifth ACM conference on Digital libraries. **Anais**...2000.

AMARAL, D.; VIEIRA, R. Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática (Braga)**, 2014.

[2] https://www.adorocinema.com/

ANANIADOU, S.; MCNAUGHT, J. **Text Mining for Biology And Biomedicine**. Norwood, MA, USA: Artech House, Inc., 2005.

ANDERSEN, P. M. et al. **Automatic extraction of facts from press releases to generate news stories**. Third Conference on Applied Natural Language Processing. **Anais**...1992.

ARAUJO, P. H. L. DE et al. **LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018. Disponível em: <https://github.com/peluz/lener-br>

ASAHARA, M.; MATSUMOTO, Y. **Japanese named entity extraction with redundant morphological analysis**. Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics. **Anais**...2003.

BANKO, M. et al. **Open Information Extraction from the Web**. Proceedings of the 20th International Joint Conference on Artifical Intelligence. **Anais**...: IJCAI'07.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9909B5C03DA1A3CCFFF4263898B69100?doi=10.1.1.74.5174&rep=rep1&type=pdf>

BARBOSA, G. C. G.; GLAUBER, R.; CLARO, D. B. **Classificação de Relações Abertas Utilizando Features Independentes do Idioma**. Proceedings of the 4th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe). **Anais**...SBC, 2016.

BENDER, E. M. **Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology**. Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? **Anais**...Athens, Greece: Association for Computational Linguistics, mar. 2009. Disponível em: <https://www.aclweb.org/anthology/W09-0106>

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, v. 35, n. 8, p. 1798–1828, 2013.

BHARDWAJ, S.; AGGARWAL, S.; MAUSAM, M. **CaRB: A crowdsourced benchmark for open IE**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...2019.

BIKEL, D. M.; SCHWARTZ, R.; WEISCHEDEL, R. M. An algorithm that learns what's in a name. **Machine learning**, v. 34, p. 211–231, 1999.

BRIN, S. **Extracting patterns and relations from the world wide web**. International workshop on the world wide web and databases. **Anais**...Springer, 1998.

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. (H. Larochelle et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., 2020. Disponível em: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>

CABRAL, B.; SOUZA, M.; CLARO, D. B. **PortNOIE: A Neural Framework for Open Information Extraction for the Portuguese Language**. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, 2022.

CASTRO, P. V. Q. DE; SILVA, N. F. F. DA; SOARES, A. DA S. **Portuguese Named Entity Recognition Using LSTM-CRF**. (A. Villavicencio et al., Eds.)Proceedings of the 13th International Conference on the Computational Processing of the Portuguese Language. **Anais**...2018.

CHEN, P. P. The Entity-Relationship Model - Toward a Unified View of Data. **ACM Trans. Database Syst.**, v. 1, n. 1, p. 9–36, 1976.

COLLOVINI, S. et al. **IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks**. Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. **Anais**...2019. Disponível em: <http://ceur-ws.org/Vol-2421/NER/_Portuguese/_overview.pdf>

CONSOLI, B. S. et al. **Embeddings for Named Entity Recognition in Geoscience Portuguese Literature**. Proceedings of The 12th Language Resources and Evaluation Conference. **Anais**...2020.

CONSORTIUM, L. D. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. **Version**, n. 5.4.3, 2005.

COWIE, J. R. **Automatic analysis of descriptive texts**. First Conference on Applied Natural Language Processing. **Anais**...1983.

COWIE, J.; LEHNERT, W. Information extraction. **Communications of the ACM**, v. 39, n. 1, p. 80–91, 1996.

CUCCHIARELLI, A.; VELARDI, P. Unsupervised named entity recognition using syntactic and semantic contextual evidence. **Computational Linguistics**, v. 27, n. 1, p. 123–131, 2001.

CUI, L.; WEI, F.; ZHOU, M. **Neural Open Information Extraction**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...2018.

CULOTTA, A.; MCCALLUM, A.; BETZ, J. **Integrating probabilistic extraction models and data mining to discover relations and patterns in text**. Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. **Anais**...2006.

DARPA (ED.). **Proceedings of the 3rd Message Understanding Conference (MUC-3)**. San Diego, EUA: Morgan Kaufmann, 1991.

DEJONG, G. Prediction and substantiation: A new approach to natural language processing. **Cognitive Science**, v. 3, n. 3, p. 251–273, 1979.

DEL CORRO, L.; GEMULLA, R. **Clausie: clause-based open information extraction**. Proceedings of the 22nd international conference on World Wide Web. **Anais**...: WWW '13.New York, NY, USA: ACM; ACM, 2013. Disponível em: <http://doi.acm.org/10.1145/2488388.2488420>

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. (J. Burstein, C. Doran, T. Solorio, Eds.)Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. **Anais**...Minneapolis, MN, USA: Association for Computational Linguistics, 2019. Disponível em: <https://doi.org/10.18653/v1/n19-1423>

DODDINGTON, G. et al. **The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation**. (M. T. Lino et al., Eds.)Proceedings of LREC'2004, Fourth International Conference on Language resources and Evaluation (Lisboa, 26-28 May 2004). **Anais**...2004. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf>

EISENSTEIN, J. **Introduction to Natural Language Processing**. [s.l.] The MIT Press, 2019.

ETZIONI, O. et al. Unsupervised named-entity extraction from the web: An experimental study. **Artificial intelligence**, v. 165, n. 1, p. 91–134, 2005.

FADER, A.; SODERLAND, S.; ETZIONI, O. **Identifying Relations for Open Information Extraction**. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. **Anais**...Edinburgh, Scotland, UK.: Association for Computational Linguistics, jul. 2011. Disponível em: <https://www.aclweb.org/anthology/D11-1142>

FERRÁNDEZ, Ó. et al. Tackling HAREM's portuguese named entity recognition task with spanish resources. **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca (http://www. linguateca. pt/aval_conjunta/LivroHAREM/Cap11-SantosCardoso2007-Ferrandezetal. pdf)**, 2007.

FINE, K. Truthmaker semantics. **A Companion to the Philosophy of Language**, p. 556–577, 2017.

FLORIAN, R. et al. **Named entity recognition through classifier combination**. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. **Anais**...2003.

GAMALLO, P.; GARCIA, M. **Multilingual open information extraction**. (F. Pereira et al., Eds.)Portuguese Conference on Artificial Intelligence. **Anais**...Cham: Springer; Springer International Publishing, 2015. Disponível em: <https://doi.org/10.1007/978-3-319-23485-4_72>

GAMALLO, P.; GARCIA, M.; FERNÁNDEZ-LANZA, S. **Dependency-based open information extraction**. Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP. **Anais**...: ROBUS-UNSUP '12.Stroudsburg, PA, USA: Association for Computational Linguistics; Association for Computational Linguistics, 2012. Disponível em: <http://dl.acm.org/citation.cfm?id=2389961.2389963>

GLAUBER, R. et al. **Challenges of an Annotation Task for Open Information Extraction in Portuguese**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018.

GLAUBER, R.; CLARO, D. B. A systematic mapping study on open information extraction. **Expert Systems with Applications**, v. 112, p. 372–387, 2018.

GLAUBER, R.; CLARO, D. B.; OLIVEIRA, L. S. **Dependency Parser on Open Information Extraction for Portuguese Texts - DptOIE and DependentIE on IberLEF**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). **Anais**...http://ceur-ws.org/Vol-2421/: CEUR Workshop Proceedings, a2019.

GLAUBER, R.; CLARO, D. B.; SENA, C. F. DE L. **Towards a Pragmatic Open Information Extraction for Portuguese Text - ICEIS17, InferPortOIE and PragmaticOIE on IberLEF**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). **Anais**...http://ceur-ws.org/Vol-2421/: CEUR Workshop Proceedings, b2019.

GONÇALO OLIVEIRA, H. et al. **Avaliação à medida no Segundo HAREM**. (C. Mota, D. Santos, Eds.)Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. **Anais**...Linguateca, 2008.

GRISHMAN, R.; SUNDHEIM, B. **Message Understanding Conference- 6: A Brief History**. COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. **Anais**...1996. Disponível em: <https://aclanthology.org/C96-1079>

GUARINO, N.; GUIZZARDI, G. **We need to Discuss the Relationship: Revisiting Relationships as Modeling Constructs**. Proceedings of the 27th International Conference on Advanced Information Systems Engineering (CAISE 2015). **Anais**...Springer-Verlag, 2015.

HABIBI, M. et al. Deep learning with word embeddings improves biomedical named entity recognition. **Bioinformatics**, v. 33, n. 14, p. i37–i48, 2017.

HASEGAWA, T.; SEKINE, S.; GRISHMAN, R. **Discovering relations among named entities from large corpora**. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (acl-04). **Anais**...2004.

HEARST, M. A. **Automatic acquisition of hyponyms from large text corpora**. Proceedings of the 14th conference on Computational linguistics-Volume 2. **Anais**...Association for Computational Linguistics, 1992.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. USA: Prentice Hall PTR, 2023.

KAMBHATLA, N. **Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction**. Proceedings of the ACL interactive poster and demonstration sessions. **Anais**...2004.

KONSTANTINOVA, N. **Review of relation extraction methods: What is new out there?** Analysis of Images, Social Networks and Texts: Third International Conference, AIST 2014, Yekaterinburg, Russia, April 10-12, 2014, Revised Selected Papers 3. **Anais**...Springer, 2014.

LÉCHELLE, W.; GOTTI, F.; LANGLAIS, P. WiRe57: A Fine-Grained Benchmark for Open Information Extraction. **arXiv preprint arXiv:1809.08962**, 2018.

LEHNERT, W.; SUNDHEIM, B. A performance evaluation of text-analysis technologies. **AI magazine**, v. 12, n. 3, p. 81–81, 1991.

LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. **Fine-Grained Named Entity Recognition in Legal Documents**. (M. Acosta et al., Eds.)Semantic Systems. The Power of AI and Knowledge Graphs. **Anais**...Cham: Springer International Publishing, 2019.

LEWIS, M. et al. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. (D. Jurafsky et al., Eds.)Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais**...Association for Computational Linguistics, 2020. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.703>

LI, Q.; JI, H. **Incremental joint extraction of entity mentions and relations**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...2014.

LIU, T.; YAO, J.-G.; LIN, C.-Y. **Towards improving neural named entity recognition with gazetteers**. Proceedings of the 57th annual meeting of the association for computational linguistics. **Anais**...a2019.

LIU, Y. et al. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**., b2019. Disponível em: <https://arxiv.org/abs/1907.11692>

MALENCHINI, F. M. et al. **Um Benchmark para Sistemas de Extração de Informação Aberta em Português**. Proceedings of theSymposium in Information and Human Language Technology (STIL 2019). **Anais**...Salvador, Bahia: SBC, out. 2019.

MCCALLUM, A.; LI, W. **Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons**. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. **Anais**...2003.

MEYER, C. F. et al. The world wide web as linguistic corpus. Em: **Corpus Analysis**. [s.l.] Brill Rodopi, 2003. p. 241–254.

MIWA, M.; BANSAL, M. **End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Association for Computational Linguistics, 2016.

MOTA, C. R3M, uma participação minimalista no Segundo HAREM. **quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008**, 2008.

MOTA, C.; SANTOS, D. (EDS.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. [s.l.] Linguateca, 2008.

MOTA, C.; SANTOS, D.; RANCHHOD, E. Avaliação de reconhecimento de entidades mencionadas: princípio de HAREM. **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**, p. 161–175, 2007.

NADEAU, D. **Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision**. tese de doutorado—[s.l.] University of Ottawa, 2007.

NASAR, Z.; JAFFRY, S. W.; MALIK, M. K. Named entity recognition and relation extraction: State-of-the-art. **ACM Computing Surveys (CSUR)**, v. 54, n. 1, p. 1–39, 2021.

NGUYEN, D. B.; THEOBALD, M.; WEIKUM, G. J-NERD: joint named entity recognition and disambiguation with rich linguistic features. **Transactions of the Association for Computational Linguistics**, v. 4, p. 215–229, 2016.

OLIVEIRA, L.; CLARO, D.; SOUZA, M. DptOIE: a Portuguese open information extraction based on dependency analysis. **Artificial Intelligence Review**, v. 56, p. 1–32, dez. 2022.

PEREIRA, V.; PINHEIRO, V. **Report - um sistema de extração de informações aberta para língua portuguesa**. Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...SBC, 2015.

QIU, Q. et al. BiLSTM-CRF for geological named entity recognition from the geoscience literature. **Earth Science Informatics**, v. 12, n. 4, p. 565–579, 2019.

RAFFEL, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, 2020.

RAU, L. F. **Extracting company names from text**. Proceedings the Seventh IEEE Conference on Artificial Intelligence Application. **Anais**...IEEE Computer Society, 1991.

RILOFF, E. et al. **Automatically constructing a dictionary for information extraction tasks**. AAAI. **Anais**...Citeseer, 1993.

RILOFF, E.; JONES, R.; et al. **Learning dictionaries for information extraction by multi-level bootstrapping**. AAAI/IAAI. **Anais**...1999.

RO, Y.; LEE, Y.; KANG, P. **Multi^2OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT**. Findings of the Association for Computational Linguistics: EMNLP 2020. **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.findings-emnlp.99>

ROARK, B.; CHARNIAK, E. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. **arXiv preprint cs/0008026**, 2000.

ROTH, D.; YIH, W. Global inference for entity and relation identification via a linear programming formulation. **Introduction to statistical relational learning**, p. 553–580, 2007.

SAGER, N. Natural language information formatting: the automatic conversion of texts to a structured data base. Em: **Advances in computers**. [s.l.] Elsevier, 1978. v. 17p. 89–162.

SAGER, N.; FRIEDMAN, C.; LYMAN, M. S. **Medical language processing: computer management of narrative data**. [s.l.] Addison-Wesley Longman Publishing Co., Inc., 1987.

SANG, E. F. T. K.; DE MEULDER, F. **Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition**. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. **Anais**...2003. Disponível em: <https://aclanthology.org/W03-0419>

SANTOS, C. N. DOS; GUIMARÃES, V. **Boosting Named Entity Recognition with Neural Character Embeddings**. (X. Duan et al., Eds.)Proceedings of the 5th Named Entity Workshop. **Anais**...Association for Computational Linguistics, 2015.

SANTOS, D.; CARDOSO, N. **Breve introdução ao HAREM**. (D. Santos, N. Cardoso, Eds.)Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. **Anais**...Linguateca, a2007. Disponível em: <http://www.linguateca.pt/LivroHAREM/>

SANTOS, D.; CARDOSO, N. **A golden resource for named entity recognition in portuguese**. Proceeding of the 7th International conference on the computational processing of portuguese. **Anais**...Springer, b2007.

SANTOS, D.; CARDOSO, N.; SECO, N. **Avaliação no HAREM: Métodos e medidas**. (D. Santos, N. Cardoso, Eds.)Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. **Anais**...Linguateca, 2007.

SANTOS, J. et al. **Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition**. Proceedings of the 8th Brazilian Conference on Intelligent Systems. **Anais**...2019.

SANTOS, J.; SANTOS, H. D. P. DOS; VIEIRA, R. **Fall Detection in Clinical Notes using Language Models and Token Classifier**. (A. G. S. de Herrera et al., Eds.)Proceedings of the 33rd IEEE International Symposium on Computer-Based Medical Systems. **Anais**...2020.

SCHANK, R. C. et al. **MARGIE: Memory Analysis Response Generation, and Inference on English.** IJCAI. **Anais**...1973.

SCHMITZ, M. et al. **Open language learning for information extraction**. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. **Anais**...: EMNLP-CoNLL '12.Stroudsburg, PA, USA: Association for Computational Linguistics; Association for Computational Linguistics, 2012. Disponível em: <http://dl.acm.org/citation.cfm?id=2390948.2391009>

SEKINE, S. **Description of the Japanese NE system used for MET-2**. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. **Anais**...1998.

SENA, C. F. L.; CLARO, D. B. InferPortOIE: A Portuguese Open Information Extraction system with inferences. **Natural Language Engineering**, v. 25, n. 2, p. 287–306, 2019.

SENA, C. F. L.; CLARO, D. B. PragmaticOIE: a pragmatic open information extraction for Portuguese language. **Knowl. Inf. Syst.**, v. 62, n. 9, p. 3811–3836, 2020.

SENA, C. F. L.; GLAUBER, R.; CLARO, D. B. **Inference Approach to Enhance a Portuguese Open Information Extraction**. Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 3: ICEIS. **Anais**...INSTICC; SciTePress, 2017.

SHAOWEI, Z. et al. Survey of Supervised Joint Entity Relation Extraction Methods. **Journal of Frontiers of Computer Science & Technology**, v. 16, n. 4, 2022.

SMIRNOVA, A.; CUDRÉ-MAUROUX, P. Relation extraction using distant supervision: A survey. **ACM Computing Surveys (CSUR)**, v. 51, n. 5, p. 1–35, 2018.

SOCHER, R. et al. **Semantic compositionality through recursive matrix-vector spaces**. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. **Anais**...2012.

SODERLAND, S. et al. **CRYSTAL inducing a conceptual dictionary**. Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2. **Anais**...1995.

SOLORIO, T. MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish. **Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Capítulo**, v. 10, p. 123–136, 2007.

SOUZA, E. N. P. DE; CLARO, D. B.; GLAUBER, R. A Similarity Grammatical Structures Based Method for Improving Open Information Systems. **j-jucs**, v. 24, n. 1, p. 43–69, 28 jan. 2018.

SOUZA, E. N. P.; CLARO, D. B. Extração de Relações utilizando Features Diferenciadas para Português. **Linguamática**, v. 6, n. 2, p. 57–65, 2014.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. (R. Cerri, R. C. Prati, Eds.)Proceedings of the 2020 Brazilian Conference on Intelligent Systems. **Anais**...Springer International Publishing, 2020.

STANOVSKY, G. et al. **Supervised open information extraction**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). **Anais**...2018.

TAKAMATSU, S.; SATO, I.; NAKAGAWA, H. **Reducing wrong labels in distant supervision for relation extraction**. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...2012.

WANG, L. et al. **Relation classification via multi-level attention cnns**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...2016.

WOLINSKI, F.; VICHOT, F.; DILLET, B. **Automatic processing proper names in texts**. Proc. Conference on European Chapter of the Association for Computational Linguistics. **Anais**...EACL, 1995.

XAVIER, C. C.; LIMA, V. L. S. DE; SOUZA, M. Open information extraction based on lexical semantics. **Journal of the Brazilian Computer Society**, v. 21, n. 1, p. 1–14, 2015.

YU, X.; LAM, W. **Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach**. Coling 2010: Posters. **Anais**...2010.

YUAN, Y. et al. **A relation-specific attention network for joint entity and relation extraction**. International joint conference on artificial intelligence. **Anais**...International

Joint Conference on Artificial Intelligence, 2021.

ZELENKO, D.; AONE, C.; RICHARDELLA, A. Kernel methods for relation extraction. **Journal of machine learning research**, v. 3, n. Feb, p. 1083–1106, 2003.

ZENG, D. et al. **Relation classification via convolutional deep neural network**. Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. **Anais**...2014.

ZHANG, S.; DUH, K.; VAN DURME, B. **Mt/ie: Cross-lingual open information extraction with neural sequence-to-sequence models**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. **Anais**...2017.

ZHAO, S.; GRISHMAN, R. **Extracting relations with integrated information using kernel methods**. Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05). **Anais**...2005.