DOI: 10.5433/1981-8920.2016v21n2p245

UM ESTUDO BIBLIOGRÁFICO SOBRE LIGAÇÃO DE ENTIDADES

UN ESTUDIO BIBLIOGRÁFICO SOBRE CONEXIÓN DE ENTIDADES

Eduardo Habib Bechelane Maia*
Marcello Peixoto Bax**

RESUMO:
Introdução: Ligação de Entidades (LE) é um importante tópico de pesquisa que tem atraído recentemente muita atenção de pesquisadores. Na tarefa de LE, menções textuais encontradas em linguagem natural são ligadas à sua entrada correspondente em uma base de conhecimento. Essa tarefa é desafiadora devido a problemas como variação de nomes, ambiguidade das entidades ou porque a entidade mencionada pode não existir na base de conhecimento.
Objetivo: Apresentar os problemas relacionados à LE, suas aplicações típicas, bem como sintetizar suas principais abordagens no contexto da ligação de conceitos.
Metodologia: Pesquisa de levantamento junto à literatura vigente, para descrição detalhada do estado da arte das abordagens em LE, bem como para a sistematização e categorização das abordagens identificadas.
Resultados: A maior parte dos trabalhos propostos para a LE divide esse processo em duas etapas: reconhecimento e ligação de entidades. No entanto, novas propostas têm unificado estas etapas em um único processo.
Conclusão: Apesar de mais complexas, as novas abordagens em LE permitem capturar a dependência entre as decisões de Ligação e de Reconhecimento de Entidades minimizando erros e inconsistências. As avaliações deveriam ocorrer em bases de dados unificadas, considerando a dificuldade de comparar resultados de bases de dados distintas devido à influência que estas exercem nos resultados obtidos.

Palavras-chave: Processamento de Linguagem Natural. Ligação de Entidades. Revisão de Literatura.

* Doutorando em Biotecnologia na Universidade Federal de São João Del Rei. Atua no Departamento de Informática, Gestão e Design – Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) Divinópolis – MG – Brazil.E-mail: habib@div.cefetmg.br
** Doutor em Informática, Análise de Sistemas e Tratamento de Sinal pela Universidade de Montpellier II, França. Professor na Escola de Ciência da Informação – ECI – Universidade Federal de Minas Gerais (UFMG). E-mail: bax@eci.ufmg.br

1 INTRODUÇÃO

O Processamento de Linguagem Natural estuda os problemas da geração e compreensão automática da informação armazenada em linguagem natural. A informação pode ser encontrada em bases de dados estruturadas ou não estruturadas. Extrair a informação de uma base de dados não estruturada envolve o processamento da linguagem natural de forma a produzir um texto semi-estruturadoe armazenado para consultas posteriores (Rao, McNamee e Dredze, 2013).

Segundo Dai *et al.* (2012), em um sistema de recuperação ideal, ousuário deverealizar uma pesquisa e receber os resultados agrupados de acordo com as diferentes entidades/conceitos referenciadospelos termos utilizados.Esses resultados recebidos serão mais úteis ao usuário, se o resultado da pesquisa possuir informações realmente relacionadas ao tema pesquisado e de importância para o usuário.

Realizar uma pesquisa e retornar o resultado adequadoé cada vez mais difícil sobretudo devidoao aumento do tamanho das bases onde são realizadas. Na Internet, por exemplo, atualmente existem pouco mais de 1 bilhão de páginas[1] e esse número tem aumentado exponencialmente. A interação entre as pessoas nas redes sociais também acaba produzindo um grande volume de dados. No Twittermais de 7 mil *tweets* são criados a cada segundo[2], mais de 400 mil por minuto.

Com o crescente volume de dadossendo gerados, têm surgidoum novo campo de pesquisa chamado Ciência dos Dados (*Data Science*). Trata-se de uma atividade orientada à análise de dados e tarefa de LE, tratada neste artigo, pode ser vista como um dos componentes mais fundamentais da "caixa de ferramentas" de um cientista de dados. O surgimento de grandes comunidades de compartilhamento de conhecimento, como a Wikipedia[3],tem incentivado o desenvolvimento de técnicas automáticas de extração de informação que

[1]http://www.internetlivestats.com/
[2]http://www.internetlivestats.com/twitter-statistics/
[3]A Wikipedia possui mais de 5 milhões de verbetes publicados e mais de 39 milhões de páginas.https://en.wikipedia.org/wiki/Wikipedia:Size_ of_Wikipedia.

facilitam a construção de novas bases de conhecimento estruturadas e que são, por isso, mais legíveis mecanicamente. Shen, Wang e Han (2015), destacam que essas bases construídas automaticamente contêm informações ricas sobre entidades do mundo real e suas relações mútuas. Alguns exemplosde bases estruturadas são:PubMed (SAYERS et al., 2009), YAGO (FABIAN, GJERGJI e GERHARD, 2007), DBpedia (BIZER et al., 2009), EntrezGene (MAGLOTT et al., 2005), Freebase (BOLLACKER et al. 2008), KnowItAll (ETZIONI et al., 2005), Probase (WU et al., 2012) e HPRD (PERI et al., 2010). Construir essas bases de conhecimento automaticamente a partir de outras fontes de informação não é uma tarefa simples, e é uma das aplicações da LE conhecida como “povoamento de bases de conhecimento”, apresentadana Seção 4.1 deste artigo.

Somado ao grande volume de dados produzidos atualmente, existem dificuldades normais inerentes ao reconhecimento da linguagem natural. Assim, uma das grandes dificuldadesna pesquisa em documentos, seja na web, seja em alguma outra base de dados, é que tanto as pesquisas quanto as informações, muitas vezes estão armazenadas em linguagem natural. Processar essas informações é, então, uma tarefa de grande complexidade. Nesse processamento, a ambiguidade é uma das maiores dificuldades, pois existem muitos termos que podem ser referenciados pelo mesmo nome. Assim, ao realizar uma pesquisa, o primeiro obstáculo se relaciona à forma de reconhecer a entidade a ser compreendida e à interpretação da mesma. Por exemplo, como um sistema poderia, automaticamente, identificar os elementos que devem ser considerados importantes em um texto e como representar corretamente as informações contidas nele, já que os elementos identificados podem ser ambíguos?

Essa tarefa de identificar uma palavra ou uma frase que referencia uma entidade particular é chamada de Reconhecimento de Entidade Nomeada (REN). O Reconhecimento de Entidades é uma das principais tarefas de processamento de linguagem natural e tem sido amplamente aplicada em vários campos, tais como extração de informação, tradução automática e assim por diante.

Uma Entidade Nomeada (EN) é a unidade básica de informação textual, bem como a base da compreensão do texto. Reconhecer entidades no meio de textos em linguagem natural de maneira eficiente pode contribuir, e muito, no contexto da Web Semântica. A Web semântica tem como objetivo relacionar os significados de diferentes palavras e, assim, conseguir atribuir um sentido às publicações na Internet de modo que ele seja explícito tanto para as pessoasquanto para o computador. Seu objetivo final écriar uma Web em que os computadores possam entender o conteúdo dos documentos existentes de forma que sejam capazes de decidir a sua relevância para ser utilizado como parte da resposta à pergunta formulada por quem procura uma informação.Entretanto, existe uma tarefa crítica nesse contexto que é ligar as Entidades Nomeadas a uma base de conhecimento onde existem mais informações sobre o termo reconhecido. Essa tarefa é chamada de Ligação de Entidades[4] (LE). É também conhecida como Reconhecimento e Ligação de Entidades[5] ou simplesmente Resolução de Entidades[6].

O restante desse artigo está organizado da seguinte maneira. Na Seção 2 será realizada uma definição mais formal da tarefa de LE. Alguns problemas relacionados à LE serão introduzidos na Seção 3. Na Seção 4, são apresentadas algumas situações em que a LE pode ser aplicada. Na Seção 5 são apresentadas abordagens que representam o estado da arte em LE, enquanto que na Seção 6 uma conclusão desse estudo é apresentada.

2 LIGAÇÃO DE ENTIDADES (LE)

Antes de iniciar a discussão sobre o estado da arte em LE, é necessário realizar uma introdução básica sobre o tema.

Larson (2010) define LE como a tarefa de reconhecer entidades em um texto escrito em linguagem natural, atribuir a elas um *id* único e ligá-las a uma base de conhecimento.

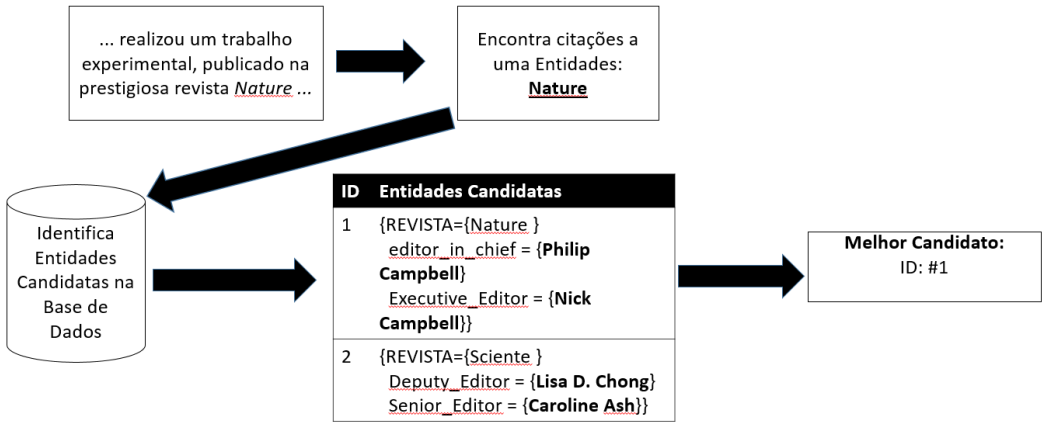
[4] *EntityLinking*.
[5] *EntityRecognitionandLinking*.
[6] *EntityResolution*.

Segundo Nural*et al.* (2013), dada uma entidade e um texto ao seu redor, LE é a tarefa de eliminação de ambiguidades e ligação da entidade a uma base de conhecimento como, p.ex, a Wikipedia ou DBpedia. Huber (2012) diz ainda que a LE pode ser considerada como parte da tarefa de Reconhecimento de Entidades. As duas tarefas potencializam-se mutuamente setratadas juntas, obtendo resultados superiores aqueles obtidos quando executadas isoladamente.

Shen, Wang e Han (2015) definemformalmente o problema de LE da seguinte forma:

> Dada uma base de conhecimentocontendo um conjunto de entidades E, e uma coleção de textos em que é identificado um conjunto de menções a entidades M, o objetivo da ligação de entidades é mapear cada menção à uma entidade $m \in M$ com a sua entidade correspondente$e \in E$ na base de conhecimento. Pode ocorrer de alguma das entidades identificadas no texto não existirem na base. Essas menções são definidas como menções sem ligação e são representadas por NIL. Portanto, se a entidade *e* para a menção à entidade *m* não existe na base, $e \nexists E$, o sistema de ligação de entidades deve gerar um identificador para a entidade mencionada na base de conhecimentos e colocar o seu valor como NIL. A Figura 1 ilustra o processo de ligação de entidades (SHEN, WANG e HAN, 2015, 444).

Figura 1 - Ligação de Entidades



Pode-se verificar na Figura 1 quea tarefa de ligação de entidades é precedida pelo processo de reconhecimento de entidades. Assim, após a identificação dos termos citados no texto, são detectadas entidades candidatas na base. Em seguida, é escolhida, entre as entidades candidatas, aquela que melhor representa otermo identificado no texto para que possa ser feita a ligação entre esse termo e a entidade da base.

Existem três principais desafios na tarefa deligação de entidades (Dai, Hong-Jie*et al.*, 2012; Rao, McNamee e Dredze, 2013):

1) **Variação de nomes**:a mesma entidade pode ser referenciada por mais de um nome: Por exemplo, Petrobrás e Petróleo Brasileiro Sociedade Anônima referenciam a mesma entidade, que é a Empresa Brasileira Petrobrás.
2) **Ambiguidade**:o mesmo nome pode referenciar mais de uma entidade.Assim, uma mesma palavra pode levar a diferentes entidades dependendo do contexto em que ela aparece. Por exemplo, imagine que o usuário pesquise pelo termo “Sol” na Wikipedia. Ao realizar essa pesquisaele poderia querer receber informações sobre as seguintes entidades que são referenciadas exatamente pelo mesmo nome e existem na Wikipedia:
    a. A estrela do nosso sistema solar.
    b. A Cerveja de origem Mexicana
    c. A estação de metrô, em Madri.
    d. O Jornal Semanal Português
    e. Etc.
3) **Ausência da entidade**:muitas vezes uma entidade mencionada pode não existir na base de conhecimento, o que impossibilita que a ligação seja realizada.

Assim, percebe-se que a tarefa de LE é uma tarefa difícil de tratar, pois além da ambiguidade naturalmente existente em linguagens naturais, existem muitas variações possíveis no nome de um mesmotermo e a potencial impossibilidade de ligação com a base de dados.

O tratamento do problema da LEé comumente separado em dois estágios, sendo eles: (1) REN, onde as entidades são reconhecidas e desambiguadas e (2) a posterior ligação a uma base de conhecimento.

Anastacio, Martins e Calado (2013), dividem o processo de ligação de entidades em ainda mais etapas (cinco), que são:

1. ***Query Expansion*:** Uma mesma Entidade pode ser referenciada no texto por diversos nomes diferentes. Portanto, a maioria dos sistemas aplicam técnicas de expansão que tentam identificar outros nomes utilizados no mesmo documento, mas que referenciam a mesma entidade.
2. ***Candidate Generation***:Essa etapa seleciona as Entidades Candidatas da base de conhecimento que correspondem à entidade mencionada no texto, baseada em similaridade de strings. Por exemplo, a estrutura de hyperlinks da Wikipedia é muito utilizada para que os algoritmos possam obter nomes alternativos para uma mesma entidade. Esses nomes podem ser obtidos a partir de páginas de desambiguação e redirecionamento, ancoras existentes nos textos, etc.
3. ***Candidate Ranking:***Nesse módulo é realizada a classificação das entidades candidatas de acordo com a probabilidade de serem a referência correta.
4. ***Candidate Validation:*** Esse módulo decide se o elemento mais bem ranqueado é resultante do fato dele ser o elemento correto ou devido ao fato dele não existir na base de conhecimento. Anastacio, Martins e Calado (2013) citam ainda que abordagens comumente utilizadas para tomar essa decisão são a criação de uma pontuação mínima para que a Entidade Candidata possa ser considerada a entidade correta ou a utilização de um esquema de votação.

5. ***NILentityresolution:*** Se o sistema decide que uma dada Entidade não possui correspondente na base de conhecimento, essa etapa deve gerar um identificador correspondentes a todas as referências a essa entidade em particular e colocar o valor NIL.

Embora essas propostas de divisões do problema em várias etapas o tornem mais fácil de tratar, pois cada etapa pode ser otimizada separadamente, essa abordagem possui desvantagens, segundo Luo*et al*. (2015):

- Erros causados em uma etapa anterior serão propagados para a etapa da LE e como são feitas em momentos diferentes e uma etapa opera independente da outra, esses erros não são corrigíveis;
- Uma etapa não se beneficia de informações disponíveis em outra etapa.

Embora a maioria dos trabalhos em ligação de entidades tenha optado por dividir a tarefa em várias, atualmente já existem alguns trabalhos recomendando que o reconhecimento e a ligação sejam tratadas em conjunto (Luo et al., 2015; Guo, Chang e Kiciman, 2013; Sil e Yates, 2013 e Pu et al., 2010).Como atividades complementares, uma pode reforçar a outra, melhorandoo resultado final.

Pode-se perceber, pelo processo de ligação de entidades descrito que, embora não seja o foco desse artigo abordar o reconhecimento de entidades, será inevitável que o assunto seja tratado, dada a importância dele em todo o processo. Entretanto, como o objetivo principal aqui é apresentar uma revisão bibliográfica específica sobre ligação de entidades, para o aprofundamento sobre técnicas, métodos e ferramentas envolvidos no reconhecimento de entidades pode-se consultar diversos *surveys* sobre o assunto como, p.ex., Nadeau e Satoshi (2007), Marrero*et al*. (2009), Kaur e Vishal (2010) e Sharnagat (2014).

Com vistas à uma contextualização mais exaustiva do problema, na próxima seção serão apresentados alguns problemas já tratados na literatura e relacionados àligação de entidades.

3 PROBLEMAS RELACIONADOS

Existem problemas correlatos à LE e que, devido à similaridade,podem auxiliar na criação de soluções de LE. Nas próximas subseções alguns desses problemas serão detalhados.

3.1 Resolução des EntidadesCorreferentes (*EntityCoreferenceResolutionProblem*)

Correferência é quando duas ou mais expressões linguísticas no mesmo texto fazem referência à mesma entidade externa. A resolução do problema de correferência trata a situação em que duas ou mais expressões de um mesmo texto se referem à mesma entidade (Pessoa, Local, Organização etc). Por exemplo, supondo que em um texto sejam identificadas as entidades *Microsoft´s CIO* e *Jim Dubois*, e como essas duas entidades referenciam a mesma entrada na base de conhecimento (Jim Dubois, CIO da Microsoft), elas são correferentes. No caso do problema da ligação de entidades, a identificação de correferências permite detectar quais entidades devem ser ligadas à mesma entrada na base de dados. Diversos autores, como por exemplo, Jiang (2009) e Han e Zhao (2009) apresentaram soluções em que aplicam correferência na desambiguação de nomes. Huynh, Thinh e Thi (2013) apresentam um estudo que aplica técnicas de resolução do problema da correferênciaemligação de entidades.

3.2 Desambiguação (*Word-SenseDisambiguation- WSD*)

Em linguagem natural, muitas vezes uma palavra possui diversos possíveis significados. WSD refere-se ao problema de identificar qual é o significado de uma determinada palavra empregada no texto. Existem alguns trabalhos que fazem revisão de literatura do problema WSD como os trabalhos de Chandra, Ganesh e Snajay (2014) e Bayaty et al. (2015). Moro *et al*. (2014) e Moro *et al*. (2015) utilizam ideias advindas do problema WSD na LE.

3.3 Ligação de Registros (*Record Linkage- RL*)

A ligação de registros é tarefa de encontrar registros que se referem à mesma entidade em bases de dados diferentes. Importante quando se realiza a fusão de bases de dados que compartilham registrosque podem ter ou não identificadores em comum. Pode-se pesquisar por técnicas e métodos empregados na solução desse problema nos estudos de Elmagarmid (2007), Brizan, Guy e Tansel (2015) e Christen (2012). Esse mesmo problema é chamado por alguns outros autores de *ObjectIdentification* (CULLOTA e MCCALLUNM, 2005), *Data Deduplication* (SARAWAGI e BHMIDIPATY, 2002; CHRISTEN, 2012), *CleaningApplication* (DASU e JOHNSON, 2003), *Merge-purge* (HERNÁNDEZ e STOLFO, 1998), *EntityIdentification*(LENZERINI, 2002), *Data Matching* (BILENKOet al., 2005), *ReferenceReconciliation* (DONG, HALEVY e MADHAVAN, 2005) e *EntityResolution* (BENJELLOUNet al., 2009), (EFTHYMIOU*et al.*, 2015), (WANG*et al.*, 2014) e (KANG*et al.*, 2008).

No campo da Biomedicina, problema similar a esse é chamado de *TermIdentification* (KRAUTHAMMER& NENADIC, 2004). Outro nome pelo qual esse problema também é bastante referenciado é *EntityDisambiguation* (DAI, TSAI e HSU, 2011; NGUYEN, HIEN e TRU, 2008; LAEK, PETER, 2012; SANTOS, IVO e BRUNO, 2015).

4 APLICAÇÕES

A LE pode ser aplicada em diversas tarefas. Dai et al. (2012) descrevem aplicações em*Knowledge Base Population* e *Normalização de Genes*. Shen, Wang e Han (2015) citam ainda as tarefas de *Extração de Informações*,*Análise de Conteúdo*,*QuestionAnswering*. Olieman e Nack (2014)utilizam a LE na tarefa de *Formação de Grupos.* Nas próximas subseções essas tarefas serão detalhadas.

4.1 Povoamento de Bases de Conhecimento (*Knowledge Base Population - KBP*)

No mundo atual, a todo momento novos dados são gerados e se tornam disponíveis aos usuários. O objetivo do povoamento de bases de conhecimento é promover a descoberta de informações sobre entidades e posteriormente utilizar essas informações para incrementar uma base de dados estruturada. Assim, a tarefa de KBP leva em consideração a necessidade de se coletar informações sobre uma determinada entidade que está espalhada em um conjunto grande de documentos e posteriormente utilizá-la para incrementar uma Base de Dados pré-existente. Isto requer a capacidade de identificar os documentos relevantes e integrar atributos proveniente destes documentos, que podem ser redundantes, complementares ou podem estar em conflito com outros atributos existentes. Nesse contexto, a LE pode ser considerada como uma subtarefa importante para o incremento de uma base de conhecimento, pois ao realizar a tarefa de LE, caso existam informações sobre a entidade citada na base de dados, ela é ligada. Caso a entidade citada não exista na base de dados, é executada a tarefa de *Slot Filling* que é responsável por coletar informações relativas a certos atributos de uma determinada entidade e povoar a base de conhecimentos (JI eGRISHMAN, 2011).

4.2 Normalização de Genes (*Gene Normalization - GN*)

Os bancos de dados de genes contêm informações relacionadas à genes, tais como sequências, produtos, informações sobre suas interações e localização.

A ideia da tarefa de GN é a detecção de identificadores únicos em uma base de dados de genes mencionados na literatura científica, onde essas menções se referem a uma entrada específica no banco de dados.

Existe um esforço considerável da comunidade científica internacional, chamado BioCreative (LEITNER et al., 2010), que promove o desenvolvimento e a avaliação de sistemas de Extração de Informação aplicadas na área de mineração de textos biomédicos.

Já existem bancos de dados de artigos biomédicos online, como o PubMed (SAYERS et al., 2009), que auxiliam em pesquisas na literatura médica. Frisch et al. (2009) criaram uma ferramenta chamada LitInspector que realiza a mineração de textos automaticamente na base de dados do PubMed. Os autores citam ainda a existência de outras ferramentas com o mesmo propósito do LitInspector que são: LitMiner (MAIER et al., 2005), PubGene (JENSSEN et al., 2001), iHOP (HOFFMAN e VALENCIA, 2005), EBIMed (REBHOLZ-SCHUHMANN et al., 2007) e PolySearch (CHENG et al., 2008). Segundo Frisch et al. (2009), essas ferramentas oferecem diversas estratégias de pesquisa na base de dados do PubMed como, por exemplo, detecção automática de genes e codificação de cores em palavras-chave.

4.3 Extração de Informação

Extração de informação (EI) é a tarefa de extrair automaticamente informações estruturadas a partir de documentos não estruturados e/ou semi-estruturados e que, na maioria dos casos, estão em linguagem natural. As entidades e relações extraídas automaticamente geralmente são ambíguas. Assim, ligá-las a uma base de dados é uma boa forma de fazer a desambiguação entre elas e tornar mais fácil a utilização futura da informação extraída.

Lin e Etzioni (2012)propuseram uma abordagem para a realização da LE, em milhões de extrações de informações realizadas em documentos da Web, com a Wikipedia.

4.4 Recuperação de Informação

O Objetivo da Recuperação de Informação (RI) é realizar a recuperação automática de informações associadas a um determinado conjunto de documentos. Essa tarefa consiste na pesquisa por informações em documentos e/ou base de dados. As pesquisas podem ser baseadas em metadados ou em indexação de textos e podem ser realizadas em banco de

dados, sejam eles relacionais e isolados ou em uma base de conhecimento como a Web, Wikipedia, etc. Atualmente existe, ainda, a tendência de se realizar pesquisa semântica, ao invés de pesquisas unicamente baseadas em palavras-chave.

Shen, Wang e Han (2015) afirmam que a pesquisa baseada na semântica das entidades se beneficia da LE, uma vez que a pesquisa também precisa realizar o processo de desambiguação das entidades mencionadas na consulta e que são, inerentemente, ambíguas. Esse processo de desambiguação permite a identificação da semântica correta de uma entidade mencionada, o que é essencial para que a consulta possa ser realizada de forma mais eficiente.

4.5 Análise de Conteúdo

Para Bardin (2011), análise de conteúdo é um conjunto de técnicas manuais ou automatizadas que visam a análise das comunicações de forma a obter, por procedimentos sistemáticos e objetivos de descrição do conteúdo das mensagens, alguns indicadores, quantitativos ou não, que permitam produzir inferências válidas e confiáveis destas mensagens.

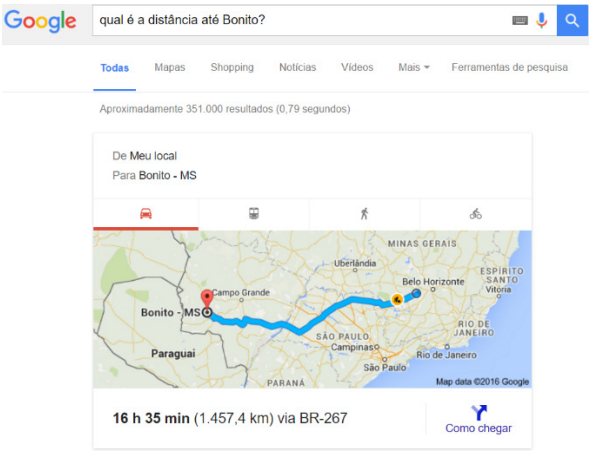
A análise do conteúdo de um texto pode se beneficiar da LE, pois ao ligar uma entidade a uma base de conhecimento, essa entidade passa pelo processo de desambiguação, tornando mais fácil correlacionar a entidade a um determinado tópico.

4.6 Question Answering

Question Answering se preocupa com a construção de sistemas que respondam automaticamente a perguntas feitas por seres humanos em uma linguagem natural. Um grande exemplo de sistemas desse tipo são as máquinas de busca como o Google. Ao digitar no Google, por exemplo, “qual é a distância até Bonito?”, a ferramenta, para responder a pergunta realiza um processo de identificação de Entidades e desambiguação da pergunta

realizada. Ao desambiguar a entidade Bonito, o Google associa essa entidade à uma cidade turística localizada no estado de Mato Grosso do Sul, embora existam cidades com o mesmo nome nos estados da Bahia, Pará e Pernambuco. Em seguida, o Google recupera as informações referentes a localização atual do usuário e da cidade pesquisada e traça uma rota com a distância e o tempo médio de viagem até lá, conforme Figura 2.

Figura 2 - Resposta à pergunta Qual é a distância até Bonito?



4.7 Formação de Grupos

Segundo Cheong (2010), a aprendizagem baseada em grupos cria um ambiente no qual os alunos podem praticar, incrementar e melhorar suas habilidades sociais, tais como liderança, comunicação, social e habilidades de resolução de conflitos. Entretanto, o modo como os grupos são formados influencia na maneira como os seus membros interagem, e, assim, afetam os resultados da experiência de aprendizagem. Tendo isso em vista, apenas colocar os alunos em equipes e criar tarefas para serem realizadas em grupos não irá necessariamente resultar em alunos que irão se desenvolver mais e que terão suas habilidades melhoradas. Olieman (2013) cita que grupos malformados podem utilizar o tempo disponível de forma improdutiva, não possuir membros com habilidades essenciais para o bom desenvolvimento do trabalho proposto ou possuir pessoas com personalidade incompatíveis.

Muitas vezes, a alternativa utilizada para a formação de grupos é deixar que os próprios grupos se formem de forma autônoma. Embora essa abordagem diminua a possibilidade de que pessoas com personalidades incompatíveis fiquem no mesmo grupo, ela não impede que os grupos sejam malformadosou que possuam membros com habilidades desbalanceadas.

Visando minimizar esses problemas, alguns trabalhos como os propostos por Kyprianidou et al. (2011), Craig et al. (2010) e Ounnas (2010) têm sugerido novas abordagens que visam formar grupos de trabalho de forma mais eficiente.

Olieman e Nack (2014) propõem uma abordagem que trata o problema da formação de grupos utilizando técnicas de LE.

Na próxima seção serão apresentadas algumas abordagens, técnicas e soluções que representam o estado da arte para a solução do Problema de LE.

5 ABORDAGENS DE LIGAÇÃO DE ENTIDADES

A estatística é muito usada para tratar problemas de construção de modelos estocásticos e, com isso, prever o comportamento de processos randômicos.

Assim como a maioria dos métodos de previsão estatística os métodos supervisionados baseiam-se na utilização dos dados históricos a partir de uma série temporal. Desse modo, uma solução supervisionada deve tomar decisões baseadas na experiência de exemplos anteriores, que são analisados e modelados em uma representação do processo, que é, então, utilizado para tentar prever o comportamento futuro.

Especuladores em bolsas de valores utilizam frequentemente modelos baseados em preços históricos de uma ação para tentar antecipar o comportamento de um determinado ativo e, com isso, tentar prever o seu valor futuro.

As soluções supervisionadas acrescentam aos dados não rotulados, rótulos com algum significado para a solução do problema. Por exemplo, um rótulo para um vídeo poderia informar que tipo de ação estaria sendo realizada

no vídeo, um rótulo para uma foto poderia informar que existe um avião na foto, rótulos para artigos poderiam informar quais os tópicos do artigo e rótulos para redes sociais poderiam informar qual é o sentimento de um tweet, por exemplo.

Após a obtenção de um conjunto de dados rotulados e a sua utilização no treinamento do algoritmo proposto, pode-se aplicar modelos de aprendizagem automática ao conjunto de dados para se prever um possível rótulo para os dados apresentados.

Os dados rotulados geralmente são obtidos solicitando que pessoas analisem os dados não rotulados e criem rótulos manualmente. Embora existam trabalhos que façam essa geração de forma automática e com menor custo, como Nuralet al. (2013), Challi, Hasan e Joty (2009) e Zheng et al. (2014), produzir automaticamente rótulos de qualidade aceitável para serem utilizados em modelos supervisionados é uma tarefa desafiadora (HUANG et al., 2014), o que faz com que a geração manual seja menos sujeita a erros, embora mais custosa.

Como produzir rótulos manualmente torna a obtenção dos dados rotulados mais custosa do que a simples utilização dos dados não rotulados, existem soluções que trabalham com os dados brutos, ou seja, sem rótulos. As soluções não supervisionadas utilizam dados não rotulados, que podem ser obtidos facilmente em todos os ambientes. Exemplos de dados não rotulados incluem fotos, textos, áudio, vídeos, dados de redes sociais, etc. Não existe nada que identifique ou explique o dado não rotulado. Ele apenas contém os dados e nada mais.

Soluções semi-supervisionadas tentam combinar dados não rotulados e rotulados em modelos integrados.

Existem muitas abordagens que tratam o problema da LE e que utilizam métodos supervisionados, não supervisionados e semi-supervisionados. Nas próximas subseções serão apresentadas soluções que utilizam essas abordagens para a resolução do problema de LE.

5.1 Métodos supervisionados

Existem muitos trabalhos que demonstram a utilização de algoritmos supervisionados para a LE utilizando, dentre outras abordagens, Markov Logic Networks (MLNs), Modelos baseados em Entropia Máxima (Maximum Entropy Models - ME), Modelos vetoriais, Support Vector Machines (SVM), Frame Semantics, Metric Learning, Learning toRank, Campos Aleatórios Condicionais (Conditional Random Fields - CRF) e Gradient Tree Boosting.

As próximas subseções irão apresentar alguns trabalhos que utilizam a aplicação de abordagens supervisionadas para a solução do problema da LE.

5.1.1 Modelos baseados em MarkovLogic Networks (MLNs)

A Rede Lógica de Markov é um modelo probabilístico que aplica a Rede de Markov em lógica de primeira ordem permitindo, assim, a inferência baseada na máxima entropia. Em uma MLN as probabilidades são atribuídas de forma a maximizar a entropia.

Dai, Tsai e Hsu (2011) utilizam MLN para abordarem a ligação de genes de uma entidade mencionada a um ID da base de dados EntrezGene (Maglott et al., 2005), que é um banco de dados de genes muito popular. No modelo proposto, as MLNs são utilizadas para unificar o estágio de desambiguação e de detecção dos genes que não existem na base de dados. Pela abordagem proposta são obtidas informações de contexto das entidades reconhecidas para a realização da desambiguação, bem como as restrições ao vincular a entidade mencionada à uma entrada na base de conhecimento. Assim, uma menção a uma entidade somente pode ser ligada a uma entrada no banco de dados quando a menção existir na base de conhecimento, ou seja, se ela não for NIL. Outra vantagem, segundo os autores, de se empregar MLN no modelo de desambiguação para a LE é que é fácil modelar dependências arbitrárias entre entidades como, por exemplo, a saliência de uma determinada entidade num determinado cenário.

5.1.2 Modelos Baseados na Entropia Máxima

Modelos Baseados na Entropia Máxima são abordagens em que, dado um conjunto de atributos e de dados para treinamento, aprende-se o peso dos atributos e os classifica. Nos modelos de entropia máxima, deve-se buscar a maximização da entropia, o que indicará que o resultado escolhido é o que tem a maior probabilidade de ser o esperado (de acordo com os dados de treinamento).

Compeau e Pevzner (2015) definem a entropia pela fórmula:

$$H(p_1,p_2, \dots , p_n) = \sum_{i=1}^{n} \quad p_i \log_2 p_i$$

Maximizar a Entropia garante que para cada característica pi, o valor esperado de pi é igual ao valor empírico de pi, de acordo com os dados de treinamento.

Sil e Yates (2013) apontam que como os métodos de LE geralmente utilizam como entrada os termos identificados por um sistema de Reconhecimento de Entidades, erros nesse sistema serão propagados para a ligação à base de conhecimento. Assim, qualquer entidade que não for detectada corretamente pelo sistema de reconhecimento, não será ligada corretamente à base de dados, fazendo com que sejam gerados mais/menos links ou ainda links incorretos para a base de dados. Alguns erros comuns que se propagam para a etapa de LE é a separação incorreta de palavras, cujo significado deveria ser interpretado em conjunto (e formariam, portanto, uma única entidade), em duas entidades; a junção de palavras que deveriam ser interpretadas de forma separada; e a desconsideração de alguma palavra em uma entidade composta por mais de um termo. Para solucionar esse problema Sil e Yates (2013) propuseram um sistema integrado de reconhecimento e LE que realiza o re-ranqueamento das entidades mencionadas e dos links para entidades candidatas fornecidas por modelos comprovadamente eficientes de Reconhecimento e LE. Utilizam, para isso, uma abordagem baseada na Entropia Máxima de forma que seja realizado o re-ranqueamento de um conjunto de entidades e citações a entidades. A abordagem proposta é mais

eficiente porque permite capturar a dependência entre as decisões de ligação e de reconhecimento de termos. Além disso, ela produz mais candidatos, segundo os autores, pois captura as dependências entre as entidades e, com base na correlação entre elas, verifica-se a possibilidade de identificar quais as mais prováveis de serem selecionadas, diminuindo a ocorrência de erros como os citados acima.

5.1.3 Modelo Vetorial

O modelo Vetorial é um modelo algébrico para a representação de documentos textuais e consultas como vetores de identificadores. Esse modelo é composto por um espaço vetorial com t dimensões, onde cada dimensão representa um determinado termo.

Segundo Baeza-Yates e Ribeiro-Neto (2013) a principal vantagem desses modelos é que eles propõem que casamentos parciais sejam possíveis. Para isso, é realizada a atribuição de pesos não binários aos termos de indexação que, por sua vez, são utilizados para definir o grau de similaridade entre cada documento armazenado no sistema e a consulta do usuário.

Olieman e Nack (2014) propõem a utilização de um modelo vetorial que dê suporte a professores na formação de grupos de trabalho, problema conhecido como Computer-Supported Group Formation (CSGF) (Ounnaset al., 2009). O método proposto seleciona as entidades candidatas para cada tema e ranqueia elas de acordo com a probabilidade de que o termo se refira à entidade candidata selecionada. Em seguida, ocorre um re-ranqueamento utilizando um modelo Vetorial, no qual as entidades são representadas por parágrafos que as mencionam na Wikipedia, em um determinado contexto. O método proposto produz perfis de alunos com base em: (1) questionários previamente preenchidos pelos mesmos; (2) acesso ao perfil dos estudantes no LinkedIn ou a um website pessoal; (3) acesso a um portfólio de projetos de cada participante; (4) acesso ao registro acadêmico dos estudantes. Com base nesses dados, o método foi capaz de realizar a vinculação dos alunos a uma grande quantidade de temas na Wikipeda. Embora a proposta tenha

conseguido associar corretamente os estudantes aos conhecimentos que possuem, não foi possível definir automaticamente, embora os autores tenham tentado, o nível de conhecimento de cada estudante sobre um determinado tópico.

5.1.4 Modelos Baseados em Support Vector Machine (SVM)

Máquinas de vetores de suporte são modelos supervisionados cujo principal objetivo é a análise de dados e posterior reconhecimento de padrões. São muito indicadas para analisar os dados utilizados para classificação e análise de regressão. As SVMs baseiam-se na ideia de separação dos exemplos positivos e negativos por uma margem grande o que faz do SVM um classificador linear binário não probabilístico. Assim, com os dados utilizados para o treinamento, cada entrada será tratada em uma de 2 categorias (positiva e negativa). Um modelo SVM mapeia os exemplos, então, como pontos no espaço de forma que os exemplos de categorias distintas estejam divididos por um espaço claro e tão amplo quanto possível.

Bunescu, Razvan C., e Pasca (2006) desenvolveram um método de desambiguação de entidades que estão ligadas a um dicionário, mapeando nomes próprios para suas possíveis entidades correspondentes. Para tanto, o método proposto detecta quando um nome próprio se refere a uma entidade existente no dicionário e realiza a desambiguação entre nomes múltiplos que podem ser referenciados pela mesma entidade, utilizando um SVM e a taxonomia da Wikipedia.

Benton et al. (2014) desenvolveram o Slinky, uma ferramenta de LE cuja proposta é ser flexível, modular, rápida e precisa. O Slinky utiliza o processamento paralelo de consultas e entidades candidatas utilizando o aprendizado em cascata para alcançar um desempenho rápido e preciso na LE. A ferramenta utiliza um modelo de aprendizado SVM em sua abordagem de ranqueamento. O Slinky funciona da seguinte maneira: (1) o pré-processador de consultas extrai as características de cada consulta (Menção, documento, etc) com base exclusivamente na consulta. (2) São geradas

entidades candidatas para a consulta na base de conhecimento. (3) Ocorre a atribuição de uma pontuação de forma independente para cada candidato e os candidatos com pontuação negativa são eliminados. Esse algoritmo é implementado como um classificador linear binário e são utilizados múltiplos classificadores em pipeline. (4) O algoritmo de ranqueamento ordena os candidatos restantes incluindo os candidatos NIL. Ele também funciona como um classificador linear, mas considera conjuntamente todos os candidatos remanescentes, permitindo a utilização de mais recursos. Nos testes realizados pelos autores, a ferramenta proposta se comportou significantemente melhor em comparação com o modelo proposto por Raykar, Krishnapuram e Yu (2014).

5.1.5 Frame Semantics

A ideia básica de Frame-Semantics é de que não é possível entender o significado de uma palavra sem compreender o conhecimento essencial relacionado àquela palavra. Assim, um "Quadro semântico" é um conjunto de conceitos que se relacionam e considera-se que sem o conhecimento de qualquer um desses conceitos, não é possível entender os outros por completo.

Collin, Charles e John (1998) propuseram o projeto Framenet[7]. Esse projeto foi construído utilizando a Teoria do "Quadro Semântico". Os desenvolvedores do projeto definem um "Quadro Semântico" como "a descrição de um tipo de evento, relação ou entidade e os seus participantes.". A tarefa de criação de rótulos no Framenet envolve:

- Dada uma sentença, identifica-se os elementos que podem compor um frame;
- É realizada a desambiguação desses elementos;

[7]https://framenet.icsi.berkeley.edu

- São identificados os papéis semânticos dos elementos identificados dentro dos frames.

Nuralet al. (2013) destacam que a maior parte dos estudos em LE referem-se à ligação de Pessoas, Localização e Organizações (PLO) a uma base de dados. Por isso, desenvolveram uma técnica baseada em Frame Semantics que visa melhorar a LE em problemas que não envolvem esses 3 tipos de entidades.

A criação de rótulos no projeto proposto por Nuralet al. (2013), é realizada da seguinte forma:

1. O texto é rotulado de acordo com o projeto Framenet, utilizando um sistema de rotulagem automática;
2. O tópico principal é identificado para cada elemento semântico;
3. É realizada a LE utilizando o DBpedia Spotligh, conforme proposto por Bizer et al. (2009);
4. Os links potenciais são mapeados em sinônimos cognitivos do Word Net (Miller et al., 1990);
5. Aplica-se Selectional Preferences para determinar qual é a relação sintática mais adequada entre verbos e substantivos e, com isso, determinar a validade das entidades.

Em experimentos, Nuralet al. (2013), demonstraram que o método proposto levou a uma melhoria da precisão e do F-Score na LE não PLO.

5.1.6 Metric Learning

Abordagens baseadas em *Metric Learning* tentam agrupar entidades dentro da mesma classe. Esse tipo de abordagem, tenta realizar medidas de similaridades e dissimilaridades entre entidades. As entidades semelhantes podem ser, então, agrupadas de acordo com a sua afinidade.

Witten e Frank (2005) explicam que em uma abordagem baseada em métrica, cada nova instância é comparada com as existentes utilizando

métricas de distância e a distância mais próxima é utilizada para a inserção da nova instância na mesma classe.

Li, Yang e Jiebo (2015) propuseram um método que faz a LE de vídeo à Wikipedia gerando entidades candidatas a partir do título de vídeos do Youtube utilizando, para isso, Metric Learning. A ideia do método proposto é demonstrar que pode-se utilizar conteúdo visual como forma de tornar mais eficiente a LE de vídeo. Pela abordagem proposta 30 frames são extraídos do vídeo a ser ligado e depois comparados com imagens das Entidades Candidatas na base de dados. Assim, os frames extraídos podem auxiliar na escolha da entidade que melhor representa o vídeo.

5.1.7 Learning toRankBasedModels

Learning toRank (L2R) ou Machine-Learned Ranking (MLR) é a aplicação do aprendizado de máquina na construção de sistemas de ranqueamento para sistemas de informação. Baeza-Yates e Ribeiro-Neto (2013) explicam que, para a aplicação de L2R, pode-se utilizar qualquer algoritmo de aprendizado de máquina e alimentá-lo com dados de treinamento que contenham informação de ranqueamento que permitam que ele “aprenda” um ranking dos resultados de maneira análoga aos algoritmos supervisionados para a classificação de textos.

Anastacio, Martins e Calado (2013) apresentam um estudo sobre desambiguação de entidades que funciona da seguinte forma:

1. Primeiramente são realizadas 2 expansões nas entidades. A primeira procura por acrônimos para as referências a entidades nomeadas e a segunda procura por menções expandidas para a mesma entidade;
2. Em seguida, é realizado o ranqueamento das entidades candidatas. Os autores utilizaram 5 diferentes algoritmos do estado da arte baseados em L2R, para a construção do modelo de ranqueamento candidato. Os Algoritmos utilizados foram: Ranking SVM (Joachims, 2002) e Ranking Perceptron (Collins e Duffy, 2002), Pairwise L2R e o ListNet (Cao et al.,

2007), Coordinate Ascent (Metzler e Bruce Croft, 2007) e o Ada Rank (Xu e Li, 2007);

3. Por fim, as referências nulas são agrupadas de forma que mesmo referências que possuam nomes diferentes, mas apontem para uma mesma entidade são ligadas ao mesmo identificador.

Através de experimentos visando determinar qual algoritmo de ranqueamento é melhor dentre os cinco estudados, Anastacio, Martins e Calado (2013) verificaram que, na base de dados utilizada durante os experimentos, nenhum dos cinco algoritmos se destacou muito em relação ao outro.

5.1.8 Modelos baseados em Campos Aleatórios Condicionais (CustomRandomFields – CRF)

Um Campo Aleatório Condicional é um modelo estocástico muito utilizado para rotular e segmentar sequencias de dados ou extrair informações de documentos. Esses modelos também são conhecidos como Campos Aleatórios de Markov, pois são uma generalização das cadeias de Markov substituindo o espaço-índice unidimensional do tempo por um espaço-índice mais genérico. Esses modelos possuem um conjunto de variáveis aleatórios contendo propriedades de Markov descritas por um grafo não direcionado.

Luo et al. (2015) abordam o problema da LE através da construção de um modelo que otimiza ao mesmo tempo duas etapas:(1)o Reconhecimento e a Desambiguação de Entidades e;(2)a Ligação a uma base de conhecimento. Otimizar essas duas etapas em conjunto torna a solução mais complexa e custosa. Entretanto, os autores argumentam que essa é uma abordagem promissora na resolução do problema de LE, pois existe uma dependência mútua entre essas duas etapas e tratando-as em conjunto é possível otimizar o processo de LE por completo. Assim, é proposto o JERL (Joint Entity Recognition and Linking), uma abordagem que reconhece o quanto essas duas etapas são dependentes e, com isso, é possível identificar as dependências mútuas entre elasde forma a tornar possível preverde maneira mais eficiente o quão coerente são suas saídas. A abordagem proposta utiliza um modelo semi-

CRF para tratar completamente a otimização de todo o processo de LE. Segundo Luoet al. (2015), os modelos semi-CRF relaxam as hipóteses de Markov entre as palavras e ajustam a distribuição de limites de segmentação diretamente. A abordagem proposta pelo JERL vai além e estende o modelo semi-CRF de forma que ele realiza o tratamento também da distribuição de entidades e as dependências mútuas nessas segmentações.

Durrett e Klein (2014) implementam um modelo que realiza simultaneamente as tarefas de Resolução de Correferência, Reconhecimento de Entidades Nomeadas e a LE, utilizando, para isso, Campos Aleatórios Condicionais. As variáveis no modelo proposto armazenam as decisões sobre prioridade, tipo semântico e LE para cada menção. Fatores unários codificam as características que são comumente empregadas na resolução de cada tarefa de forma isolada. Para a execução da LE e NER em conjunto, é realizado o mapeamento entre tipos semânticos das entidades reconhecidas e a semântica da Wikipedia, que são descritas por caixas de informação, categorias e textos. A correferência interage com as outras tarefas de uma forma mais complexa, através de fatores que incentivam a consistência de tipos semânticos e links de entidades em todos os arcos de correferência. A proposta apresentada utiliza a propagação de crenças (Loopy Belief Propagation) como forma de realizar a inferência aproximada.

5.1.9 GradientTreeBoosting

Yang e Chang (2015) propuseram um framework chamado Structured Multiple Additive Regression Trees (S-MART), que aplica a tarefa de LE em tweets. O framework S-MART generaliza a implementação do Multiple Additive Regression Trees (MART) para problemas de aprendizado estruturado. O MART é uma implementação do método da árvore de gradiente (*Gradient tree boosting*) para a mineração de dados preditivo. O S-MART não otimiza diretamente os parâmetros do modelo. Ao invés disso, ele se aproxima da pontuação da solução ótima adicionando interativamente modelos de árvores de regressão. S-MART combina a não linearidade e eficiência dos modelos baseados em árvore com a previsão estruturada, levando a uma família de

novos algoritmos. Um novo algoritmo de inferência é proposto para lidar com a tarefa de LE em Tweets. Esse algoritmo foi proposto especialmente para estruturas que não se sobrepõem, com o objetivo de evitar a atribuição de entidades conflitantes. Os testes executados em grandes conjuntos de dados demonstram que o S-MART supera em mais de 10% os algoritmos do estado da arte utilizados para vencer o desafio Named Entityr Ecognitionand Linking (NEEL) do ano de 2014.

5.1.10 Outros métodos supervisionados

Milne e Witten (2008) desenvolveram um sistema baseado em aprendizado de máquina que, assim como Mihalcea e Csomai (2007), realiza a LE a artigos da Wikipedia. Eles utilizam os links encontrados em artigos da Wikipedia para realizar o treinamento do algoritmo proposto. Para isso, partem do pressuposto de que todos os links foram adicionados manualmente e, portanto, realizam a ligação correta de forma a representar algo que explicita o sentido correto da entidade, uma vez que já consumiram algum esforço humano para a realização dessa ligação. Com isso, eles demonstram que pode-se estabelecer relações entre conceitos e que eles podem ser utilizados para verificar o relacionamento semântico entre entidades. Utilizam ainda, ao contrário de Mihalcea e Csomai (2007), o contexto em que a entidade aparece como forma de realizar a correta ligação com a Wikipedia.

Huynh, Thinh e Thi (2013) propõem um sistema de LE chamado MACH (MAchine learning-based with Coreference and Heuristic). No sistema proposto é adicionado ao processo de aprendizado proposto por Milne e Witten (2007) a utilização de relações de correferência e a identifição de menções inexistentes na base de dados. Além disso, o método proposto vale-se de algumas heurísticas que utilizam palavras em torno de uma entidade mencionada para filtrar entidades candidatas indesejáveis.

5.2 Métodos não supervisionados

Conforme explicado no início dessa seção, as soluções não supervisionadas utilizam dados não rotulados para a solução de um determinado problema.

Zheng et al. (2014) destacam que os métodos supervisionados requerem que dados sejam rotulados manualmente, mas nos bancos de dados biomédicos esses rótulos ainda não estão disponíveis. Portanto, é necessário que sejam desenvolvidas soluções não supervisionadas para esses modelos. Lai et al. (2009), apresentaram seis métodos de desambiguação para serem utilizados em textos biomédicos como forma de descobrir o ID correto em citações ambíguas a genes. Os seis métodos de desambiguação utilizados foram;*"Parenthesis Support", "Chromosome Location Support", "Dictionary Matching Support", "Gene Ontology Support", "Protein-Protein Interaction Support" e "TissueSupport"* e podem ser explicados resumidamente da seguinte forma:

- *Parenthesis Support*: Se um gene é seguido por outro gene entre parênteses, os 2 genes são tratados como sinônimos
- *Chromosome Location Support*: Quando se tem uma citação ambígua de um cromossomo, caso exista no contexto onde o cromossomo foi citado o local onde ele é encontrado, essa localização pode ser utilizada para a desambiguação.
- *Dictionary Matching Support*: Pode ser utilizado para verificar o quão frequentemente um determinado ID do gene aparece no dicionário. O ID que aparece mais frequentemente é, então, o que será selecionado.
- *Gene Ontology Support*: Utiliza-se a base de dados do NCBI (Sayers et al., 2009), que inclui os IDs de um gene e seus respectivos termos ontológicos. Em seguida, procura-se no texto por cada gene ambíguo e baseado no número de termos encontrados, é selecionado o ID ambíguo que mais vezes aparece. Foi implementado ainda uma melhoria

utilizando o algoritmo de distância de Smith-Waterman (Smith e Waterman, 1981).

- *Protein-Protein Interaction (PPI) Support*: Utiliza-se a interação entre genes e proteínas como forma de identificar corretamente o ID do gene na base de dados. Assim, dado um gene ambíguo, é utilizada a proteína com a qual o gene interage para realização da desambiguação e a ligação com o ID correto.
- *Tissue Support*: A base de dados *Human Protein Reference Database* (HPRD)contém informações do tecido do corpo humano onde um determinado gene age. Assim, ao encontrar um gene ambíguo, caso a informação do tecido onde o gene atua seja citada no texto, ela pode ser utilizada para a realização da desambiguação.

Em seus experimentos Lai et al. (2009) conseguiram uma melhoria significativa de 21,5% utilizando uma combinação dos seis métodos de desambiguação descritos acima na normalização de genes utilizando a base de dados Bio Creative II (LEITNER et al., 2010).

Zhenget al. (2014) propuseram uma abordagem não supervisionada de inferência coletiva para ligar entidades em textos não estruturados da literatura biomédica. Segundo os autores, métodos de inferência coletivos tratam o problema da LE através da maximização da concordância entre o texto do documento de referência e o contexto das entidades da base de conhecimento. Assim, em sua proposta, os autores definem pesos baseados na entropia para as relações da base de conhecimento e os incorporam em um processo de ranqueamento de candidatos em duas etapas como forma de realizar a LE. Os autores propuseram a utilização de estruturas ricas em ontologias para realizar a Ligação das Entidades o que tornou possível abandonar a tarefa manual e tediosa da criação de rótulos. A vantagem é que existem muitas ontologias relacionadas e publicadas na Web pela comunidade, como o BioPOrtal, por exemplo.

Kalloubi, Nfaoui e Beqqali (2014a e 2014b) propuseram um sistema para extrair automaticamente entidades nomeadas, desambiguá-las e ligá-las a uma

base de conhecimentos utilizando uma abordagem baseada na centralidade de grafos e em Linked Open Data (LOD). O sistema proposto utiliza a base de dados DBpedia combinada com centralidade de grafos para realizar a desambiguação e reconhecer automaticamente entidades nomeadas no twitter e ligá-las a recursos LOD como forma de enriquecer o conteúdo publicado no microblog.

Guo et al. (2011) propuseram uma abordagem não supervisionada baseada em grafos para realizar a LE. Na abordagem proposta, os nós do grafo correspondem ao contexto em torno da entidade mencionada e as entidades candidatas na base de conhecimento enquanto que as arestas representam a dependência entre os nós. Assim, procura-se pelo contexto nos nomes dos nós e os candidatos são procurados nos artigos da base de conhecimento. Utiliza-se a medida da conectividade de grafos, para determinar a importância de cada nó candidato, como forma de atribuir o nó mais importante como referência para a entidade mencionada. No modelo proposto, um grafo G = (V; E) é construído e corresponde ao contexto onde o nome alvo aparece. A abordagem apresenta um método de desambiguação baseado nas medidas de conectividades de saída e de entrada do grafo. Para a medida de conectividade de saída, o conjunto de nós do grafo é formado pelos nomes que são mencionados no contexto e nos artigos dos respectivos candidatos. Um vértice é criado ligando uma entidade a uma menção toda vez que um nome de referência é exibido na descrição artigo. Por fim, a referência candidata com mais vértices de saída é considerada a mais importante nesse grafo e a entidade correspondente a este nó artigo é então selecionada para ser ligada a entidade mencionada. Para a medida de conectividade de entrada, o conjunto de nós consiste dos nomes das entidades candidatas e das entidades correspondentes ao contexto. Existe um vértice que liga um nome de nó candidato a um contexto do artigo quando ele possui esse nome. A entidade com o maior número de arestas de entrada é considerada a mais importante e a entidade mencionada é, então, atribuída a entidade correspondente a esse nó.

Mihalcea e Csomai (2007) implementaram um conjunto de técnicas de extração automática de “palavras chave”, sua desambiguação e posterior ligação com artigos da Wikipédia. O algoritmo proposto é integradoa um sistema chamado de Wikify!.Nesse sistema, dado um documento de entrada cujas entidades devem ser ligadas à Wikipedia, são identificados os conceitos importantes do texto e esses conceitos são ligados à páginas da Wikipedia. O sistema proposto funciona em quatro passos:

1. Pré-processamento do documento de entrada, onde ocorre a separação entre o texto e as Tags HTML;
2. Determinação de palavras chave, onde é identificado cada n-grama que aparece no documento e no vocabulário controlado, que é formado pelos títulos dos artigos da Wikipedia e suas variações morfológicas que são utilizadas cinco ou mais vezes;
3. Ranqueamento da palavra chave, onde é realizada a desambiguação e um cálculo da probabilidade de ligá-lo à Wikipedia, o que é chamado de *key phraseness*. Assim todas as frases em um novo documento cuja *keyphrase* excede um determinado limite são escolhidas como *keyphrases*. Se um determinado termo é usado como link para um número suficiente de artigos da Wikipedia, eles realizam a ligação sempre que é encontrado em outros documentos, independentemente do contexto;
4. A estrutura do documento é reconstruída e os links para a Wikipedia são adicionados.

Ferragina e Scaiella (2010) desenvolveram o Tagme, que é um sistema de wikificação especializado em textos curtos e que possui uma API disponível para ser utilizada livremente. Primeiramente, o sistema Tagme procura no texto por menções que estejam de acordo com um vocabulário pré-definido, que é composto pelos títulos dos artigos, âncoras e redirecionamentos da Wikipidia.

Cada menção é, então, associada a um conjunto de entidades candidatas sobre as quais se faz necessário realizar um processo de desambiguação. A desambiguação explora a estrutura de grafos da Wikipedia, de acordo com a medida de similaridade introduzida por Milne e Witten (2008). Assim, são consideradas a quantidade de links de entrada entre duas páginas. A desambiguação utilizada no sistema Tagme contabiliza todas as possíveis ligações entre menções e entidades de forma a manter armazenadas a quantidade de ligações existentes. Segundo um estudo realizado por Cornolti et al. (2013), o sistema Tagme possui o melhor desempenho entre os sistemas de wikificação disponíveis ao público.

Shen et al. (2013), criaram um framework baseado em Grafos chamado Kauri. O sistema proposto liga coletivamente todas as menções a entidades nomeadas, em todos os tweets de um determinado usuário, à artigos da wikipedia, através da modelagem dos tópicos de interesse. A abordagem proposta realiza, para isso, a ligação coletiva das entidades nomeadas de todos os tweets postados por um usuário considerando que cada usuário possui um tópico de interesse principal. O framework proposto unifica duas categorias de informação: (1) informações locais do próprio tweet e (2) informações de interesse dos usuários entre tweets.

Liu et. al (2013) propuseram um novo método de inferência coletiva que integra três tipos de similaridades: (1) *mention-mentionsimilarity*, (2) *entity-entitysimilarity* e (3) *entity-mentionsimilarity*. *Entity-mentionsimilarity* é utilizada para modelar a compatibilidade local, enquanto *entity-entitysimilarity* e *mention-mentionsimilarity* são combinadas de forma a modelar uma consistência global. O modelo proposto dá preferência a configurações onde menções similares possuem entidades similares com alto grau de compatibilidade local. Com isso, o modelo mapeia simultaneamente um conjunto de menções de um tweet com as suas entidades apropriadas.

Zuo et al. (2014) construíram um modelo que se baseia em uma estratégia de desambiguação que explora a utilização de *bagging* de múltiplos classificadores de ranqueamento. Cada classificador opera em uma amostragem randômica composta por subconjuntos de termos vizinhos da

entidade mencionada. Os termos escolhidos para fazer parte da amostragem são os mais promissores na determinação do contexto da menção. A probabilidade de uma entidade candidata ser referenciada é implementada de forma similar ao proposto por Lin e Etzioni (2012). Finalmente, com base nos termos da amostragem e na probabilidade da entidade candidata ser referenciada, cada classificador propõe uma lista ranqueada de entidades candidatas e a menção é ligada à entidade que é mais bem classificada por todos os classificadores. Para tornar o algoritmo o mais eficiente, ele explora as decisões anteriores de desambiguação sempre que possível. Assim, se uma menção ocorrer várias vezes no mesmo documento, a decisão de desambiguação anterior é utilizada, sem executar novamente o algoritmo. Os autores demonstram que a utilização dos termos vizinhos à menção leva a uma melhor captura do seu contexto.

Lin e Etzioni (2012) propuseram uma técnica eficiente para a realização da LE, em milhões de extrações realizadas na Web, com a Wikipedia. O objetivo do trabalho é criar uma base de dados útil contendo fatos gerais. Eles realizam o agrupamento do contexto em torno das entidades como forma de auxiliar na etapa de desambiguação na próxima interação do algoritmo proposto. Os autores utilizam a contagem de entidades esperadas como forma de detectar erros sistemáticos e demonstraram que, cerca de um terço das menções às entidades detectadas durante os experimentos não possuíam correspondentes na Wikipedia.

5.3 Métodos semi-supervisionados

Métodos semi-supervisionados apresentam abordagens que combinam a utilização de dados rotulados e não rotulados em uma solução integrada. Dessa maneira, dados com e sem rótulos são utilizados para criar hipóteses. As soluções semi-supervisionadas estão no meio do caminho entre as supervisionadas e as não supervisionadas.

Huang et al. (2014) propõem um modelo semi-supervisionado de regularização em grafos para a realização de inferência coletiva como forma de detecção de desambiguação e posterior realização da wikificação dos tweets.

Assim, as relações entre múltiplas menções a entidades são analisadas simultaneamente, o que leva ao ranqueamento de cada vértice ao mesmo tempo. Como um simples tweet pode não prover evidências suficientes para identificar menções e inferir o conceito a que ele referencia devido à falta de informações, a abordagem proposta considera os princípios de compatibilidade local, correferência e relação semântica como forma de incorporar evidências locais e globais a partir de vários tweets.

Perera et al. (2016) propuseram uma abordagem que realiza a ligação implícita de entidades. Ligar entidades de forma implícita é uma tarefa desafiadora, uma vez que a menção à entidade não contém o seu nome. Os autores do estudo demonstraram que, para um determinado subconjunto de *tweets* sobre filmes e livros selecionados por eles, verificou-se que 21% dos *tweets* sobre filmes faziam menções implícitas, enquanto que 40% das menções à livros eram feitas de forma implícita. Com base nesses dados, o modelo proposto pelos autores faz a LE considerando, para isso, o seu contexto e significado real. Ao fim do estudo, é demonstrada a importância de se realizar a exploração do contexto de uma entidade na tarefa de realizar a ligação implícita de suas menções. Durante o processo de LE, para suprir a ausência do nome de uma entidade, Perera et al. (2016) utilizam outras características da mesma para identificá-la como o autor de um livro no qual o filme é baseado, ou a utilização do diretor de um filme para fazer a referência. Uma determinada menção a entidade pode, entretanto, referenciar entidades distintas em momentos diferentes. A citação do diretor Steven Spielberg, por exemplo, em 2009 pode se referir ao filme “Transformers: A vingança dos derrotados”, enquanto que, em 2011, a mesma menção pode se referir ao filme “Transformers: O lado oculto da lua”. Um exemplo citado por Pereraet al. (2016) demonstra dois tweets sobre o filme gravidade: *`New Sandra Bullock astronaut lost in space movie looks absolutely terrifying,'* e *`ISRO sends probe to Mars for less money than it takes Hollywood to send a woman to space.'*. O primeiro *tweet* cita explicitamente uma atriz do filme, o que permite que a informação seja extraída de uma base de dados que correlaciona autores e filmes. O segundo *tweet* não cita nenhuma entidade relacionada ao filme, mas

outros *tweets* que ocorrem em um período temporal próximo a ele e citam explicitamente o filme gravidade citam também ISRO, *'Woman to space'* e *'less money'*. Assim, a grande inovação desse estudo é que ele considera também o contexto temporal, o que permite que a desambiguação da entidade mencionada seja realizada de forma mais eficiente.

Ganeaet al. (2016) propuseram um modelo gráfico probabilístico que aborda a desambiguação coletiva de entidades através da propagação de crenças em ciclos (*Loopy Belief Propagation*) como forma de realizar uma inferência aproximada. Algoritmos de propagação de crença em ciclos, permitema realização de inferência em modelos gráficos, como em redes Bayesianas e *Markov Random Fields*. O modelo proposto é baseado em estatísticas empíricas que exigem um procedimento simples e rápido de treinamento e que depende de poucos parâmetros. Essa simplicidade, permite que o algoritmo seja executado rapidamente, mesmo em grandes quantidades de dados, o que torna ele propício para aplicações em tempo real. Para executar a modelagem, os autores consideram que uma entidade depende de sua menção, das palavras vizinhas locais para a identificação do contexto e de outras entidades que são encontradas no mesmo documento. A abordagem proposta consiste de duas partes principais: (1) a probabilidade de uma entidade candidata ser a escolhida, dada a sua menção e o seu contexto e (2) a distribuição das entidades candidatas correspondentes a todas as menções no documento.

Yao e Sun (2014) propõem um novo método para reconhecer e normalizar nomes de celulares em fóruns da internet. Para a realização dessa tarefa, o modelo proposto gera as possíveis variações de nomes de telefones baseado em convenções existentes. O nome de um aparelho móvel geralmente é formado pelo fabricante do telefone e pelo nome ou número do modelo. A geração prévia de possíveis nomes permite que se obtenham um grande número de exemplos para a realização do treinamento do algoritmo proposto com um baixo custo de anotação manual. Um desafio ao minerar dados em fóruns de internet é que, além da variação inerente da linguagem natural, ainda podem ocorrer abreviações e erros de digitação. Visando

contornar esse problema, os autores utilizam *Brown Clustering*, que é um algoritmo de clusterização hierárquica que agrupa as palavras com significado e função sintática parecidos. Essas palavras são agrupadas no conjunto C, que forma o conjunto de nomes candidatos. Após a execução desse algoritmo, Yeao e Sun (2014) propõem a geração de exemplos para o modelo CRF (Conditional Random Field) proposto de forma semi-automática e de modo que o esforço manual para a criação de rótulos seja mínimo. Assim, cria-se dois conjuntos (P e N) de identificadores de telefones. No conjunto P são armazenados os nomes formais dados aos aparelhos, como, por exemplo, Motorola Moto X, Moto X, etc. No conjunto N, são armazenados um conjunto de palavras que não representam nomes de telefones, mas que podem ser facilmente identificadas, como por exemplo, firmware, debug e update. De posse dos 2 conjuntos, são selecionadas as sentenças que possuam pelo menos uma entidade que pertença ao conjunto P ou ao conjunto N.Além disso, a sentença não pode possuir nenhumelemento que seja resultado da operação$C \setminus (P \cup N)$. Os autores propõem, então, a utilização de seis regras para o modelo CRF proposto, sendo que: duas delas são léxicas, duas são gramaticais e duas estão relacionadas aos nomes dos telefones. Por fim, se um determinado nome for indicado como possível candidato a nome de telefone ele é, então, mapeado para o nome formal.

Em trabalhos recentes, Moro et al. (2014a) e Moro et al. (2014b), propuseram o Babelfy, um sistema que explora uma nova abordagem unificada baseada em grafos para WSD e LE. O sistema proposto aproveita a informação estrutural conjunta fornecida por uma grande rede semântica, tanto a nível lexicográfico quanto a nível enciclopédico. Assim, pela proposta dos autores, o conhecimento lexicográfico obtido durante o processo de WSD também é utilizado para ajudar a tarefa de EL, enquanto que a informação enciclopédica obtida durante a LE auxilia no processo de desambiguação em WSD.

Phelan, McCarthy e Smyth (2009) utilizaram o *Twitter* para recomendar notícias em tempo real aos usuários com base em suas preferências e nas

notícias mais atuais do microblog. Eles descrevem uma nova técnica de recomendação de notícias que aproveita dados obtidos em tempo real como base para a classificação e recomendação de artigos, utilizando o *Twitter* para classificar uma coleção de feeds RSS com base nos tweets do usuário e de seus amigos ou dos tweets mais recentes realizados pelos usuários do microblog.

6 CONCLUSÃO

Nesse *survey* foi realizada uma descrição detalhada do problema conhecido comoLE, bem como foram levantados os problemas relacionados à essa tarefa e as situações em que a LE podem ser aplicadas.Por fim, as abordagens que representam o estado da arte nessa tarefa foram apresentadas, dividindo-as em supervisionadas, não supervisionadas e semi-supervisionadas.

Pôde-se perceber,durante esse estudo,que a maior parte dos trabalhos propostos para a LE dividem esse processo em etapas. Notadamente, a maior parte dos trabalhos dividem esse processo em dois (Reconhecimento e Ligação de Entidades), e utilizam como entrada os nomes identificados por um sistema de Reconhecimento de Entidades. Entretanto, embora essa abordagem simplifique o problema, por permitir que se otimize cada etapa individualmente,erros na identificação de um termo em uma das etapas serão propagados para as etapas seguintes.Com isso, têm surgido novas propostas em que os processos de Reconhecimento e Ligação de Entidades são tratados como um único processo. Essa abordagem torna o problema mais complexo, mas parece ser promissora, pois permite capturar a dependência entre as decisões de Ligação e de Reconhecimento de Entidades e, com isso, melhorar a decisão final, tanto no reconhecimento quanto na LE, evitando, assim a propagação de erros ou que as duas etapas criem saídas inconsistentes.

Verificou-se ainda que, embora existam muitas abordagens diferentes para a LE, muitas vezes elas são avaliadas utilizando bases de dados

diferentes, o que torna difícil compará-las, pois os resultados são muito dependentes da base de dados utilizada na avaliação.

Por fim, espera-se que esse estudo possa ser útil para a realização de pesquisas em LE.

REFERÊNCIAS

ANASTÁCIO, I., MARTINS, B., & CALADO, P. **Supervised learning for linking named entities to knowledge base entries.** In Proceedings of Text Analysis Conference (TAC 2011), 2011.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação:** conceitos e tecnologia das máquinas de busca. São Paulo: Bookman Editora, 2013.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. **The berkeleyframenet project.** In: PROCEEDINGS OF THE 17TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, Association for Computational Linguistics, v. 1, p. 86-90, August, 1998.

Bayaty, Z. A., Boshra, F., & Joshi, S. (2015, May). Empirical comparative study to supervised approaches for WSD problem: Survey. In Humanitarian Technology Conference (IHTC2015), 2015 IEEE Canada International (pp. 1-7). IEEE.

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., &Widom, J. (2009). Swoosh: a generic approach to entity resolution. The VLDB Journal—The International Journal on Very Large Data Bases, 18(1), 255-276.

Benton, A., Deyoung, J., Teichert, A., Dredze, M., Van Durme, B., Mayhew, S., & Thomas, M. (2014). Faster (and better) entity linking with cascades. In NIPS Workshop on Automated Knowledge Base Construction.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S., Adaptive name matching in information integration. Intelligent Systems, IEEE, 18(5), 16-23, 2005

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. Web Semantics: science, services and agents on the world wide web, 7(3), 154-165.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250). ACM.

Brizan, D. G., &Tansel, A. U. (2015). A. Survey of Entity Resolution and Record Linkage Methodologies. Communications of the IIMA, 6(3), 5.

Bunescu, R. C., &Pasca, M. (2006, April). Using Encyclopedic Knowledge for Named entity Disambiguation. In Proceedings of European Chapter of the Association for Computational Linguistics (EACL 2006) (Vol. 6, pp. 9-16).

Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F., & Li, H. (2007, June). Learning to Rank: from pairwise approach to listwise approach. In Proceedings of the 24th International Conference on Machine Learning (pp. 129-136). ACM.

Chali, Y., Hasan, S. A., &Joty, S. R. (2009, August). Do automatic annotation techniques have any impact on supervised complex question answering?. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP 2009). Conference Short Papers (pp. 329-332). The 47rd Annual Meeting of the Association for Computational Linguistics.

Chandra, G., &Dwivedi, S. K. (2014, December). A Literature Survey on Various Approaches of Word Sense Disambiguation. In Computational and Business Intelligence (ISCBI 2014) 2nd International Symposium on (pp. 106-109). IEEE.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., &Wishart, D. S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic acids research, 36(suppl 2), W399-W405.

Cheong, C. (2010). From group-based learning to cooperative learning: A metacognitive approach to project-based group supervision. Informing Science: the International Journal of an Emerging Transdiscipline, 13(1), 73-86.

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. Knowledge and Data Engineering, IEEE Transactions on, 24(9), 1537-1555.

Collins, M., & Duffy, N. (2002, July). New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. InProceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 263-270). Association for Computational Linguistics.

Compeau, P, Pevzner, P.. Bioinformatics Algorithms: An Active-Learning Approach. 2nd Edition. Active Learning Publishers. 2015. Volume 1. 384 páginas.

Cornolti, M., Ferragina, P., &Ciaramita, M. (2013, May). A framework for benchmarking entity-annotation systems. In Proceedings of the 22nd international conference on World Wide Web (pp. 249-260). International World Wide Web Conferences Steering Committee.

Craig, M., Horton, D., & Pitt, F. (2010, November). Forming reasonably optimal groups:(frog). In Proceedings of the 16th ACM international conference on Supporting group work (pp. 141-150). ACM.

Culotta, A., & McCallum, A. (2005, October). Joint deduplication of multiple record types in relational data. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 257-258). ACM.

Dai, H. J., Tsai, R. T. H., & Hsu, W. L. (2011, November). Entity Disambiguation Using a Markov-Logic Network. In Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2011) (pp. 846-855).

Dai, H. J., Wu, C. Y., Tsai, R., & Hsu, W. (2012). From entity recognition to entity linking: a survey of advanced entity linking techniques. In Proceedings of the 26th Annual Conference of the Japanese Society for Artificial Intelligence (pp. 1-10).

Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning (Vol. 479). John Wiley & Sons.

Dong, X., Halevy, A., &Madhavan, J. (2005, June). Reference reconciliation in complex information spaces. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (pp. 85-96). ACM.

Durrett, G., & Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. Transactions of the Association for Computational Linguistics, 2, 477-490.

Efthymiou, V., Stefanidis, K., &Christophides, V. (2015, October). Big data entity resolution: From highly to somehow similar entity descriptions in the Web. In Big Data (Big Data), 2015 IEEE International Conference on (pp. 401-410). IEEE.

Elmagarmid, A. K., Ipeirotis, P. G., &Verykios, V. S. (2007). Duplicate record detection: A survey. Knowledge and Data Engineering, IEEE Transactions on,19(1), 1-16.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., ... & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, 165(1), 91-134.

Fabian, M. S., Gjergji, K., & Gerhard, W. (2007). Yago: A core of semantic knowledge unifying wordnet and wikipedia. In 16th International World Wide Web Conference, WWW (pp. 697-706).

Ferragina, P., &Scaiella, U. (2010, October). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1625-1628). ACM.

Frisch, M., Klocke, B., Haltmeier, M., &Frech, K. (2009). LitInspector: literature and signal transduction pathway mining in PubMed abstracts.Nucleic acids research, 37(suppl 2), W135-W140.

Ganea, O. E., Horlescu, M., Lucchi, A., Eickhoff, C., & Hofmann, T. (2016). Probabilistic Bag-Of-Hyperlinks Model for Entity Linking. In Proceedings of the 25th International Conference on World Wide Web.

Goker, A., & Davies, J. (Eds.). (2009). Information retrieval: searching in the 21st century. John Wiley & Sons.

Guo, S., Chang, M. W., &Kiciman, E. (2013). To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In Proceeding of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2013) (pp. 1020-1030)

Guo, Y., Che, W., Liu, T., & Li, S. (2011, November). A Graph-based Method for Entity Linking. In Proceedings of International Joint Conference on Natural Language Processing (pp. 1010-1018).

Han, X., & Zhao, J. (2009, November). Named entity disambiguation by leveraging wikipedia semantic knowledge. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 215-224). ACM.

Hernández, M. A., &Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. Data mining and knowledge discovery, 2(1), 9-37.

Hoffmann, R., & Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics, 21(suppl 2), ii252-ii258.

Huang, H., Cao, Y., Huang, X., Ji, H., & Lin, C. Y. (2014, June). Collective Tweet Wikification based on Semi-supervised Graph Regularization. In Association for Computational Linguistics (ACL) (pp. 380-390).

Huber, Torsten. "Entity Linking-A Survey of Recent Approaches." (2012).

Hunter, L., & Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed?. Molecular cell, 21(5), 589-594.

Huynh, H. M., Nguyen, T. T., & Cao, T. H. (2013, November). Using coreference and surrounding contexts for entity linking. In Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on (pp. 1-5). IEEE.

Jenssen, T. K., Lægreid, A., Komorowski, J., &Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression.Nature genetics, 28(1), 21-28.

Ji, H., &Grishman, R. (2011, June). Knowledge base population: Successful approaches and challenges. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 1148-1158). Association for Computational Linguistics.

Jiang, L., Wang, J., An, N., Wang, S., Zhan, J., & Li, L. (2009, December). Grape: A graph-based framework for disambiguating people appearances in web search. In Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on (pp. 199-208). IEEE.

Joachims, T. (2002, July). Optimizing search engines using clickthrough data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 133-142). ACM.

Kalloubi, F., El Habib, N., & El Beqqali, O. (2014b, November). Graph based tweet entity linking using DBpedia. In Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on (pp. 501-506). IEEE.

Kalloubi, F., Nfaoui, E. H., & El Beqqali, O. (2014a, May). Named entity linking in microblog posts using graph-based centrality scoring. In Intelligent Systems: Theories and Applications (SITA-14), 2014 9th International Conference on (pp. 1-6). IEEE.

Kang, H., Getoor, L., Shneiderman, B., Bilgic, M., &Licamele, L. (2008). Interactive entity resolution in relational data: A visual analytic tool and its evaluation. Visualization and Computer Graphics, IEEE Transactions on, 14(5), 999-1014.

Kaur, D., & Gupta, V. (2010). A survey of named entity recognition in english and other indian languages. IJCSI International Journal of Computer Science Issues, 7(6), 1694-0814.

Krauthammer, M., &Nenadic, G. (2004). Term identification in the biomedical literature. Journal of biomedical informatics, 37(6), 512-526.

Kyprianidou, M., Demetriadis, S., Tsiatsos, T. andPombortsis, A. 2011. Group formation based on learning styles: can it improve students' teamwork? Educational Technology Research and Development. 60, 1 (Sep. 2011), 83–110.

Laek, I., &Vojta, P. (2012, December). Context Aware Named Entity Disambiguation. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on (Vol. 1, pp. 402-408). IEEE.

Lai, P. T., Bow, Y. Y., Huang, C. H., Dai, H. J., Tsai, R. T. H., & Hsu, W. L. (2009, August). Using contextual information to clarify gene normalization ambiguity. In Information Reuse & Integration, 2009. IRI'09. IEEE International Conference on (pp. 1-5). IEEE.

Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L. A., & Valencia, A. (2010). An overview of BioCreative II. 5. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 7(3), 385-399.

Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246). ACM.

Li, Y., Yang, X., & Luo, J. (2015). Semantic Video Entity Linking Based on Visual Content and Metadata. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4615-4623).

Lin, T., &Etzioni, O. (2012, June). Entity linking at web scale. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (pp. 84-88). Association for Computational Linguistics.

Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., & Lu, Y. (2013). Entity Linking for Tweets. In Association for Computational Linguistics (ACL 2013) (1) (pp. 1304-1311).

Luo, G., Huang, X., Lin, C. Y., &Nie, Z. Joint named entity recognition and disambiguation. In Proceedings of the Empirical Methods in Natural Language Processing. 2015.

Maglott, D., Ostell, J., Pruitt, K. D., &Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. Nucleic acids research, 33(suppl 1), D54-D58.

Maier, H., Döhr, S., Grote, K., O'Keeffe, S., Werner, T., de Angelis, M. H., & Schneider, R. (2005). LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts. Nucleic acids research, 33(suppl 2), W779-W782.

Marrero, M., Sánchez-Cuadrado, S., Lara, J. M., &Andreadakis, G. (2009). Evaluation of named entity extraction systems. Advances in Computational Linguistics, Research in Computing Science, 41, 47-58.

Metzler, D., & Croft, W. B. (2007). Linear feature-based models for information retrieval. Information Retrieval, 10(3), 257-274.

Mihalcea, R., &Csomai, A. (2007, November). Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 233-242). ACM.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database*. International journal of lexicography, 3(4), 235-244.

Milne, D., & Witten, I. H. (2008, October). Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 509-518). ACM.

Moro, A., Cecconi, F., &Navigli, R. (2014b, October). Multilingual word sense disambiguation and entity linking for everybody. In Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272 (pp. 25-28). CEUR-WS. org.
Moro, A., Raganato, A., &Navigli, R. (2014a). Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics, 2, 231-244.

Nadeau, D., &Sekine, S. (2007). A survey of named entity recognition and classification. LingvisticaeInvestigationes, 30(1), 3-26.

Nguyen, H. T., & Cao, T. H. (2008, July). Named entity disambiguation on an ontology enriched by Wikipedia. In Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on (pp. 247-254). IEEE.

Nural, M. V., Miller, J. A., &Arpinar, I. B. (2013, September). Improving Entity Linking Performance Using Frame Semantics. In Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on (pp. 56-63). IEEE.

Olieman, A. M., e Nack, F. "Mastery Profiling through Entity Linking to Support Project Team Formation in Higher Education". 6th International Conference on Computer Supported Education (CSEDU14), 2014.

Ounnas, A. 2010. Enhancing the Automation of Forming Groups for Education with Semantics. PhD. Thesis. University of Southampton.

Ounnas, A., Davis, H. C., & Millard, D. E. (2009). A Framework for Semantic Group Formation in Education. Educational Technology & Society, 12(4), 43-55.

Perera, S., Mendes, P. N., Alex, A., Sheth, A. P., &Thirunarayan, K. "Implicit Entity Linking in Tweets". Extended Semantic Web Conference, 2016.

Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., ... & Rashmi, B. P. (2004). Human protein reference database as a discovery resource for proteomics. Nucleic acids research, 32(suppl 1), D497-D501.

Phelan, O., McCarthy, K., & Smyth, B. (2009, October). Using twitter to recommend real-time topical news. In Proceedings of the third ACM conference on Recommender systems (pp. 385-388). ACM.

Pu, K. Q., Hassanzadeh, O., Drake, R., & Miller, R. J. (2010, October). Online annotation of text streams with structured entities. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 29-38). ACM.

Rao, D., McNamee, P., &Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In Multi-source, Multilingual Information Extraction and Summarization (pp. 93-115). Springer Berlin Heidelberg.

Raykar, V. C., Krishnapuram, B., & Yu, S. (2010, July). Designing efficient cascaded classifiers: tradeoff between accuracy and cost. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 853-860). ACM.

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., &Stoehr, P. (2007). EBIMed—text crunching to gather facts for proteins from Medline. Bioinformatics, 23(2), e237-e244.

Santos, J. T. L., Anastacio, I. M., & Martins, B. E. (2015). Named Entity Disambiguation over Texts Written in the Portuguese or Spanish Languages. Latin America Transactions, IEEE (Revista IEEE America Latina), 13(3), 856-862.

Sarawagi, S., &Bhamidipaty, A. (2002, July). Interactive deduplication using active learning. In Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 269-278). ACM.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., &Feolo, M. (2009). Database resources of the national center for biotechnology information. Nucleic acids research, 37(Database issue), D5.

Sharnagat, Rahul. "Named Entity Recognition: A Literature Survey." unpublished (2014).

Shen, W., Wang, J., Luo, P., & Wang, M. (2013, August). Linking named entities in tweets with knowledge base via user interest modeling. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 68-76). ACM.

Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. Knowledge and Data Engineering, IEEE Transactions on, 27(2), 443-460.

Sil, A., & Yates, A. (2013, October). Re-ranking for joint named-entity recognition and linking. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 2369-2374). ACM.

Smith, T. F. e W., M. S. (1981). "Identification of Common Molecular Subsequences" (PDF). Journal of Molecular Biology 147: 195–197. doi:10.1016/0022-2836(81)90087-5. PMID 7265238.

Wang, T., Kou, Y., Shen, D., Liu, H., & Yu, G. (2014, September). SIER: An Efficient Entity Resolution Mechanism Combining SNM and Iteration. In Web Information System and Application Conference (WISA), 2014 11th (pp. 238-241). IEEE.

Witten, I. H. e Frank, E. Data Mining: practical machine learning tools and techniques. 2ª Edição – (2005). Morgan Kaufmann series in data management systems. ISBN: 0-12-088407-0

Witten, I., & Milne, D. (2008, July). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA (pp. 25-30).
Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012, May). Probase: A probabilistic taxonomy for text understanding. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 481-492). ACM.

Xu, J., & Li, H. (2007, July). Adarank: a boosting algorithm for information retrieval. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 391-398). ACM.

Yang, Y., & Chang, M. W. S-MART: Novel Tree-based Structured Learning Algorithms Applied to Tweet Entity Linking. In Proceedings of the 7th International Joint Conference on Natural Language Processing (IJCNLP 2015), The 53rd Annual Meeting of the Association for Computational Linguistics, p. 504–513, 2015.

Yao, Y., & Sun, A. "Product name recognition and normalization in internet forums." Special Interest Group Symposium on Information Retrieval in Practice (SIGIR Industry Track). 2014.

Zheng, J. G., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., & Ji, H. (2015). Entity linking for biomedical literature. BMC medical informatics and decision making, 15(Suppl 1), S4.

Zuo, Z., Kasneci, G., Gruetze, T., &Naumann, F. (2014). BEL: Bagging for Entity Linking. In Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014). (pp. 2075-2086).

Title

A bibliographic study on entity linking

Abstract:
Introduction: Linking Entities (LE) is an important research topic that has recently attracted great attention from researchers. In the LE tasks, textual mentions found in natural language are linked to their corresponding entry in a knowledge database. This task is challenging due to problems such as name variation, entity ambiguity, or because the entity mentioned may not exist in the knowledge database.
Goals: To present the problems related to LE, its typical applications, as well as to synthesize their main approaches in the context of concept linkage.
Methodology: Survey research with the current literature, for a detailed description of the state of the art of the approaches in LE, as well as for the systematization and categorization of the identified approaches.
Results: Most of the studies in LE divide this process into two stages: recognition and linking of entities. However, new proposals have unified these steps into a single process.
Conclusion: Although more complex, the new LE approaches allow us to capture the dependence between Entity Liaison and Entity Recognition decisions, minimizing errors and inconsistencies. Evaluations should occur in unified databases, considering the difficulty of comparing different database results due to the influence they have on testing outputs.

Keywords: Natural Language Processing. Entity Linking. Literature Review.

Titulo

Un estudio bibliográfico sobre conexión de entidades

Resumen:
Introducción: Las entidades de conexión (CO) es un importante tema de investigación que recientemente ha llamado mucho la atención de los investigadores. En la tarea LE, referencias textuales que se encuentran en lenguaje natural están vinculados a su entrada correspondiente en una base de conocimientos. Esta tarea es

un reto debido a problemas como cambios en los nombres de organizaciones o ambigüedad ya que la entidad no puede existir en la base de conocimientos.
Objetivo: Presentar los problemas relacionados con la LE, sus aplicaciones típicas, así como sintetizar sus principales enfoques en el contexto de los conceptos de unión.
Metodología: Encuesta La investigación realizada por la literatura actual, para indicar la descripción detallada de los procedimientos de la técnica en LE, así como la sistematización y categorización de los enfoques identificados.
Resultados: La mayoría de los trabajos propuestos para LE dividir este proceso de dos etapas: reconocimiento y entidades de unión. Sin embargo, nuevas propuestas han unificado estos pasos en un solo proceso.
Conclusión: Aunque más complejas, nuevos enfoques en LE permiten capturar la dependencia entre las decisiones de la conexión y el reconocimiento de entidades minimizando errores e inconsistencias. Las evaluaciones deben ocurrir en la base de datos unificada, teniendo en cuenta la dificultad de comparar los resultados de diferentes bases de datos debido a la influencia que ejercen sobre los resultados.

Palabras-clave: Procesamiento del Lenguaje Natural. Conexión de Entidades. Revisión de la Literatura.

Enviado em: 17.07.2016.
Aceito em: 20.11.2016.