Extracção de Relações Semânticas de Textos em Português Explorando a DBpédia e a Wikipédia

Exploring DBpedia and Wikipedia for Portuguese Semantic Relationship Extraction

David S. Batista David Forte Rui Silva Bruno Martins
Mário J. Silva
Instituto Superior Técnico and INESC-ID, Lisboa, Portugal
{david.batista,david.forte,rui.silva,bruno.g.martins,mjs}@ist.utl.pt

Resumo

A identificação de relações semânticas, expressas entre entidades mencionadas em textos, é um passo importante para a extracção automática de conhecimento a partir de grandes colecções de documentos, tais como a Web. Vários trabalhos anteriores abordaram esta tarefa para o caso da língua inglesa, usando técnicas de aprendizagem automática supervisionada para classificação de relações, sendo que o actual estado da arte recorre a métodos baseados em *string kernels* (Kim et al., 2010; Zhao e Grishman, 2005). No entanto, estas abordagens requerem dados de treino anotados manualmente para cada tipo de relação, além de que os mesmos têm problemas de escalabilidade para as dezenas ou centenas de diferentes tipos de relações que podem ser expressas. Este artigo discute uma abordagem com supervisão distante (Mintz et al., 2009) para a extracção de relações de textos escritos em português, a qual usa uma técnica eficiente para a medição de similaridade entre exemplares de relações, baseada em valores mínimos de dispersão (i.e., *min-hashing*) (Broder, 1997) e em dispersão sensível à localização (i.e., *Locality-Sensitive Hashing*) (Rajaraman e Ullman, 2011).

No método proposto, os exemplos de treino são recolhidos automaticamente da Wikipédia, correspondendo a frases que expressam relações entre pares de entidades extraídas da DBPédia. Estes exemplos são representados como conjuntos de tetragramas de caracteres e de outros elementos representativos, sendo os conjuntos indexados numa estrutura de dados que implementa a ideia da dispersão sensível à localização. Procuram-se os exemplos de treino mais similares para verificar qual a relação semântica que se encontra expressa entre um determinado par de entidades numa frase, com base numa aproximação ao coeficiente de Jaccard obtida por *min-hashing*. A relação é atribuída por votação ponderada, com base nestes exemplos. Testes com um conjunto de dados da Wikipédia comprovam a adequabilidade do método proposto, tendo sido extraídos 10 tipos diferentes de relações, 8 deles assimétricos, com uma pontuação média de 55.6% em termos da medida F_1.

Palavras chave

Extracção de Relações, Extracção de Informação

Abstract

The identification of semantic relationships, as expressed between named entities in text, is an important step for extracting knowledge from large document collections, such as the Web. Previous works have addressed this task for the English language through supervised learning techniques for automatic classification. The current state of the art involves the use of learning methods based on *string kernels* (Kim et al., 2010; Zhao e Grishman, 2005). However, such approaches require manually annotated training data for each type of semantic relationship, and have scalability problems when tens or hundreds of different types of relationships have to be extracted. This article discusses an approach for distantly supervised relation extraction over texts written in the Portuguese language, which uses an efficient technique for measuring similarity between relation instances, based on minwise hashing (Broder, 1997) and on locality sensitive hashing (Rajaraman e Ullman, 2011).

In the proposed method, the training examples are automatically collected from Wikipedia, corresponding to sentences that express semantic relationships between pairs of entities extracted from DBPedia. These examples are represented as sets of character quadgrams and other representative elements. The sets are indexed in a data structure that implements the idea of locality-sensitive hashing. To check which semantic relationship is expressed between a given pair of entities referenced in a sentence, the most similar training examples are searched, based on an approximation to the Jaccard coefficient, obtained through min-hashing. The relation class is assigned with basis on the weighted votes of the most similar examples. Tests with a dataset from Wikipedia validate the suitability of the proposed method, showing, for instance, that the method is able to extract 10 different types of semantic relations, 8 of them corresponding to asymmetric relations, with an average score of 55.6%, measured in terms of F_1.

Keywords

Relation Extraction, Information Extraction

1 Introdução

Em Extracção de Informação (EI) e Processamento de Linguagem Natural (PLN), a tarefa de extracção de relações consiste em detectar e classificar relações semânticas, em colecções de documentos textuais. Por exemplo, na frase ***Brooklyn*** *é um dos 62 condados do estado americano de* ***Nova Iorque***, a relação semântica *localizado-em* encontra-se expressa entre dois nomes de locais. Alguns domínios de aplicação incluem a detecção de relações de interação entre pares de proteínas, ou entre genes e doenças, na literatura biomédica (Bunescu e Mooney, 2005a; Kim et al., 2010; Zhou e Zhang, 2007), a detecção de diferentes tipos de associações entre entidades mencionadas em textos jornalísticos, tais como as relações entre pessoas e os seus locais de nascimento, ou entre pessoas e as organizações a que pertencem (Hachey, Grover e Tobin, 2012), ou ainda a detecção de relações semânticas entre pares de expressões nominais em geral (Hendrickx et al., 2010). Ao longo dos anos, têm sido propostas diferentes abordagens para resolver a tarefa de extracção de relações. Em particular, os métodos baseados em regras aplicam regras linguísticas para capturar os padrões tipicamente usados para expressar relações (Brin, 1999). Os métodos baseados em características, por outro lado, transformam exemplos das relações semânticas a extrair em conjuntos de características linguísticas, como por exemplo características lexicais, sintácticas e/ou semânticas, capturando a semelhança entre vectores de características através de técnicas de aprendizagem automática supervisionada (GuoDong et al., 2005). Os trabalhos mais recentes na área envolvem a utilização de métodos de aprendizagem baseados em *string kernels*, quer explorando *kernels* para representar sequências (Bunescu e Mooney, 2005a), numa tentativa de capturar padrões sequenciais dentro de frases dos textos, ou *kernels* específicos para árvores ou para grafos no geral, por forma a aprender funções de classificação relacionadas com os padrões em estruturas resultantes de uma análise sintáctica (Nguyen, Moschitti e Riccardi, 2009; Bunescu e Mooney, 2005b). Os métodos baseados em *string kernels* são superiores aos métodos baseados em características, no sentido de melhor contornarem o facto de os dados de treino serem tipicamente muito esparsos, ou no sentido de proporcionarem uma exploração eficiente de espaços de características muito grandes. Porém, estes métodos são computacionalmente exigentes quando se considera um número elevado de classes, ou sempre que é preciso manipular conjuntos de treino grandes, tornando assim difícil a sua aplicação em problemas reais.

Neste trabalho, propomos explorar uma abordagem diferente para a extracção automática de relações semânticas, com base na pesquisa pelos kNN exemplos de treino mais próximos, como forma de fazer a classificação, aproveitando um método eficiente baseado em valores mínimos de funções de dispersão como forma de medir a similaridade entre relações, para diferentes tipos de relações semânticas. O método proposto é avaliado na tarefa específica de extracção de relações em textos escritos em português, sendo os exemplos de treino extraídos automaticamente da Wikipédia, com base nas relações entre pares de entidades que se encontram explicitamente codificados na DBPédia. Desta forma, através de supervisão distante (Hoffmann, Zhang e Weld, 2010), contornamos a dificuldade em coleccionar exemplos de treino anotados.

Num trabalho anterior explorámos já a ideia de pesquisar pelos kNN exemplos de treino mais próximos, como forma de abordar a extracção de relações de textos em inglês, de forma supervisionada e usando conjuntos de dados bem conhecidos na área (Batista et al., 2013). Com este trabalho pretendemos agora testar a eficiência desta abordagem num cenário envolvendo supervisão distante e textos em português.

Desta forma, realizámos experiências exaustivas com diferentes configurações do método de classificação proposto, variando o tamanho das assinaturas de *min-hash*, assim como o número de exemplos mais próximos considerados para a classificação. Experiências com um conjunto de dados da Wikipédia portuguesa comprovam a adequabilidade do método proposto. Os melhores resultados correspondem a uma macro-média de 55.6% em termos da medida F_1, quando se consideram 10 tipos diferentes de relações semânticas, 8 delas correspondendo a relações semânticas assimétricas.

O resto deste artigo está organizado da seguinte forma: a Secção 2 apresenta trabalhos relacionados importantes. A Secção 3 descreve o método proposto, detalhando a recolha automática de exemplos de treino a partir da Wikipédia e da DBPédia, descrevendo a representação considerada para os exemplos dos diferentes tipos de relações, e apresentando a abordagem de dispersão sensível à localização. A Secção 4 apresenta a avaliação experimental do método proposto. Finalmente, a Secção 5 resume as nossas principais conclusões e apresenta orientações possíveis para trabalho futuro.

2 Trabalho Relacionado

Extrair relações semânticas entre expressões nominais (por exemplo, entre nomes de entidades como pessoas, locais ou organizações), tal como mencionadas nos textos, é um passo crucial na compreensão da linguagem natural, com muitas aplicações práticas. Vários autores têm proposto técnicas de aprendizagem automática para abordar o problema, por exemplo formulando o mesmo como uma tarefa de classificação binária (i.e., uma tarefa de classificação supervisionada binária, definida sobre exemplares de candidatos a relações entre pares de expressões nominais, em que os exemplares são classificados como membros de uma classe *exemplares_relacionados* ou de uma classe *exemplares_não_relacionados*).

Apesar do exemplo anterior se focar no caso de relações binárias entre pares de expressões nominais, a discussão é facilmente estendida à extracção de relações considerando n tipos diferentes de relações semânticas entre entidades.

Entre as abordagens anteriores relevantes incluem-se os trabalhos de autores que adoptaram métodos baseados em vectores de características e aprendizagem supervisionada, ou métodos baseados em *string kernels*. A maior vantagem de métodos baseados em *string kernels* reside no facto destas soluções permitirem explorar um espaço de características muito grande (i.e., frequentemente exponencial ou, nalguns casos, infinito) em tempo computacional polinomial, sem a necessidade de representar explicitamente os vectores de características. No entanto, os métodos de classificação baseados em *string kernels* são apesar de tudo muito exigentes em termos de requisitos computacionais, perante problemas que envolvam um número elevado de classes ou grandes colecções de dados.

Dado um conjunto de exemplos positivos e negativos sobre um dado tipo de relação semântica, os métodos baseados em características começam por extrair características sintácticas e semânticas dos textos, utilizando-as como pistas para decidir se as entidades numa dada frase se encontram relacionadas ou não. As características sintácticas extraídas das frases incluem tipicamente:

1. As próprias entidades em si.
2. A categoria semântica das duas entidades (e.g. pessoa, local ou organização).
3. A sequência de palavras entre as entidades.
4. O número de palavras entre as entidades.
5. O caminho entre as duas entidades numa árvore de análise sintáctica (i.e., *parse tree*).

As características semânticas podem incluir, por exemplo, o caminho entre as duas entidades numa estrutura resultante de uma análise de dependências entre os constituintes da frase. Muito embora a análise de dependências possa ser vista como puramente sintáctica, muitos trabalhos anteriores nesta área argumentam que a mesma está próxima de uma representação semântica.

As várias características são apresentadas a um classificador sob a forma de um vector de características. Vários algoritmos de aprendizagem supervisionada, como máquinas de vectores de suporte ou modelos baseados em regressão logística, assim como conjuntos de características diferentes, têm sido explorados na literatura (Zhou e Zhang, 2007; Kambhatla, 2004).

Os métodos baseados em características têm a limitação de envolverem escolhas heurísticas, sendo muitas vezes necessário proceder à selecção de características numa base de tentativa e erro, por forma a maximizar o desempenho. Para solucionar o problema da selecção automática de um conjunto adequado de características, foram desenvolvidos *kernels* especializados para a tarefa de extracção de relações, tirando partido de representações ricas dos exemplos de treino, e explorando estas representações ricas de uma forma exaustiva e implícita.

Bunescu e Mooney (2005a) apresentaram um *kernel* de subsequências, que trabalha com representações dos exemplos baseadas em sequências esparsas, combinando palavras e etiquetas morfológicas (i.e., *POS tags*) por forma a capturar as palavras no contexto em torno das expressões nominais envolvidas nas relações. Três *kernels* de subsequências são usados para calcular a similaridade entre os exemplares, isto é, entre as instâncias de uma dada relação, ao nível das palavras, ou seja, comparando as sequências de palavras que ocorrem (i) antes e entre, (ii) entre, e (iii) entre e depois das expressões nominais envolvidas nas relações. Um *kernel* combinado é então produzido pela soma dos três *sub-kernels* anteriores. Os autores avaliaram a sua abordagem na tarefa de extrair interações proteicas no corpus AImed (i.e., um conjunto de textos constituído por resumos de publicações no MEDLINE[1]), concluindo que os *kernels* de subsequências, em conjunto com classificadores baseados em máquinas de vectores de suporte, melhoram a qualidade dos resultados, em comparação com um sistema baseado em regras. Além disso, Bunescu e Mo-

[1] `http://ipubmed.ics.uci.edu/`

oney (2005a) também argumentaram que representações mais ricas para as palavras no contexto, com base nas etiquetas morfológicas das palavras e nos tipos de entidades envolvidos nas relações, podem levar a resultados melhores com o *kernel* de subsequências do que com uma abordagem baseada num *kernel* de árvores de dependências, tal como proposto por Culotta e Sorensen (2004).

Zelenko, Aone e Richardella (2003) descreveram uma abordagem para extracção de relações que usa um *kernel* para a comparação de representações das frases baseadas em estruturas sintácticas pouco profundas. O *kernel* é desenhado para calcular a similaridade entre duas árvores de análise sintáctica (i.e., *shallow parse trees*), aumentadas com as entidades sob as quais incide a relação, em termos da soma ponderada do número de sub-árvores que são comuns entre duas representações. Estes autores avaliaram a sua abordagem numa tarefa de extracção de relações dos tipos *pessoa-filiação* e *organização-localização*, notando no entanto que o método proposto é vulnerável a erros na geração das árvores de análise sintáctica.

Culotta e Sorensen (2004) descreveram ainda uma versão modificada do *kernel* anterior utilizando representações baseadas em árvores de dependências, em que um *kernel* para a comparação de conjuntos de palavras também é usado como forma de compensar erros na análise sintáctica. Na proposta de Culotta e Sorensen, cada nó da árvore de dependências contém informação rica, como a identidade da palavra, a etiqueta morfológica, o tipo de sintagma (nominal, verbal, etc.), ou o tipo de entidade. Usar uma representação estruturada mais rica pode levar a um ganho de desempenho, em comparação com a utilização de abordagens baseadas simplesmente em sacos de palavras (i.e., *bags of words*). Uma versão refinada foi posteriormente proposta por Zhao e Grishman (2005), usando nós compostos para integrar informações de diferentes fontes sintácticas (por exemplo, informação lexical, informação resultante de uma análise sintáctica, e informação resultante de uma análise de dependências). Desta forma, os erros de processamento que ocorram ao nível das representações podem ser superados por informações provenientes de outros níveis.

Airola et al. (2008) introduziram um *kernel* denominado de *all-dependency-paths*, usando uma representação baseada num grafo direccionado com arestas ponderadas, que combinam dois subgrafos desconexos. Temos assim que, uma estrutura representa a árvore de dependências de uma frase, e uma outra estrutura representa a ordem sequencial das palavras. Bunescu e Mooney (2005) apresentaram ainda uma abordagem alternativa que utiliza a informação concentrada no caminho mais curto na árvore de dependências entre as duas entidades. Estes autores argumentam que o caminho mais curto, entre as duas expressões nominais, numa árvore resultante da análise de dependências, codifica informação suficiente para extrair relações.

Estudos recentes continuam a explorar combinações ou extensões dos métodos baseados em *kernels* descritos anteriormente (Kim et al., 2010). No entanto, a maioria das abordagens propostas são avaliadas em conjuntos de dados diferentes, pelo que não é possível ter uma ideia clara sobre qual a abordagem que é efectivamente melhor. Os conjuntos de dados habitualmente usados na língua inglesa incluem versões do corpus das avaliações efectuadas no contexto do programa ACE (Hachey, Grover e Tobin, 2012), conjuntos de dados construídos a partir da Wikipédia (Culotta, McCallum e Betz, 2006), e subconjuntos de publicações na MEDLINE (Bunescu e Mooney, 2005a). Algumas avaliações conjuntas de sistemas computacionais para o processamento de linguagem natural também abordaram especificamente problemas de extracção de relações semânticas. Um destes eventos foi a tarefa no SemEval 2010 em *Multi-way Classification of Semantic Relations Between Pairs of Nominals* (Hendrickx et al., 2010).

Além dos métodos baseados em aprendizagem supervisionada para a extracção de relações semânticas num dado domínio fechado, importa também referir alguns trabalhos recentes que se focaram na extracção de relações num contexto aberto, tais como o sistema TextRunner[2], o sistema ReVerb[3], o sistema OLLIE[4] ou o SOFIE[5]. Algumas das técnicas elementares que estão na base destes vários sistemas são descritas num artigo de revisão sobre a àrea de *Open Information Extraction* (OIE) (Etzioni et al., 2008).

Os sistemas de OIE procuram identificar um conjunto aberto de relações, operando com base na análise de padrões frequentes em grandes colecções de dados, e muitas vezes usando regras produzidas por peritos humanos. Estes sistemas não requerem assim dados de treino, podendo contemplar a extracção de relações em domínios como a Web (Fader, Soderland e Etzioni, 2011). No entanto, abordagens independentes do domínio produzem geralmente resul-

[2] `http://www.cs.washington.edu/research/textrunner/`
[3] `http://reverb.cs.washington.edu/`
[4] `https://github.com/rbart/ollie`
[5] `http://www.mpi-inf.mpg.de/yago-naga/sofie/`

tados de pior qualidade, e por regra também não normalizam as relações extraídas, tendo este passo de ser tratado à posteriori (Soderland e Mandhani, 2007). Por exemplo, qualquer aplicação usando os resultados destes sistemas deve lidar com homonímia e sinonímia entre as expressões que codificam as relações, assim como com problemas associados à polissemia e sobreposição das relações. Por este motivo, defendemos que deverá ser preferível a extracção de relações semânticas através de exemplos, pelo menos em domínios específicos de aplicação.

Como forma de contornar a dificuldade associada à anotação manual de dados de treino, alguns autores investigaram ainda paradigmas alternativos para a extracção de relações em textos, baseados em supervisão distante, bootstraping (Pantel e Pennacchiotti, 2006) ou outros métodos (Riedel et al., 2013). Por exemplo Mintz et al. (2009), Krause et al. (2012), ou Riedel, Yao e McCallum (2010) usaram a Freebase[6], i.e. uma base de dados estruturada de informação semântica cobrindo milhares de relações entre entidades, como forma de construir os exemplos de treino. Para cada par de entidades referenciado numa relação da Freebase, os autores procuram por frases num corpus que contenham as entidades, usando posteriormente estas frases como exemplos de treino de um extractor de relações baseado num classificador tradicional. Desta forma, os autores combinam as vantagens dos métodos supervisionados para a extracção de relações, com as vantagens dos métodos não-supervisionados para extracção de informação.

Temos, por exemplo, que nas experiências efectuadas por Mintz et al. (2009), os autores utilizaram um classificador de máxima entropia combinando atributos lexicais (e.g., sequências de palavras e as etiquetas morfológicas correspondentes) e sintácticos (e.g., dependências entre as entidades envolvidas na relação). Os resultados mostraram que a metodologia de supervisão distante, com base no Freebase, lhes permitiu extrair 10.000 exemplares de 102 tipos diferentes de relações, com uma precisão de 67.6%. Indo além de recursos como o Freebase, outros autores propuseram ainda a utilização das *infoboxes* da Wikipédia, ou alternativamente de informação proveniente do projecto DBPédia, onde uma rede semântica é extraída automaticamente a partir das *infoboxes* da Wikipédia (Auer et al., 2007), como forma de construir os exemplos de treino (Blessing e Schütze, 2010; Hoffmann, Zhang e Weld, 2010; Wu e Weld, 2010).

[6] `http://www.freebase.com/`

Em relação à extracção de relações de textos em português, temos que alguns resultados foram reportados no contexto da tarefa piloto ReRelEM (Freitas et al., 2008), a qual se realizou no âmbito do segundo evento de avaliação conjunta HAREM (Mota e Santos, 2008) com o objectivo de avaliar o reconhecimento e classificação de relações semânticas ao nível de pares de entidades mencionadas num dado documento. Aquando do evento de avaliação HAREM, a coleção de documentos usada na medição da qualidade dos resultados incluía um total de 12 documentos com 4.417 palavras e 573 entidades mencionadas. Na mesma, encontravam-se descritas 6,790 relações entre pares de entidades manualmente anotadas (i.e., 1.436 relações de identidade, 1.612 relações de ocorrência, 1.232 relações de colocação, e 2.510 relações de outros tipos). Desde então, a colecção de documentos foi estendida e disponibilizada online para outros investigadores (i.e., existem agora 24 tipos de relações semânticas diferentes, e um total de 7.847 entidades mencionadas anotadas na colecção, sendo que 3.776 se encontram relacionadas entre si num total de 4.803 relações manualmente anotadas). No entanto, não temos conhecimento de outros estudos em extracção de relações desde textos, que tenham feito uso desta colecção estendida de documentos e anotações.

Os três sistemas diferentes que participaram na tarefa ReRelEM optaram pelo reconhecimento de diferentes tipos de relações, sendo assim difícil tirar conclusões sobre os seus méritos relativos. Por exemplo o sistema SEI-Geo apenas reconhece relações de inclusão entre entidades do tipo local, realizando esta tarefa através do mapeamento dos pares de entidades para com uma ontologia onde relações semânticas deste tipo já se encontram expressas (Chaves, 2008).

O sistema SeRELeP usa regras heurísticas (e.g., na frase *a Brigada Militar de Porto Alegre ocorre em Porto Alegre*, o sistema iria reconhecer uma relação semântica do tipo ocorrência através de uma regra que indica que se houver uma entidade mencionada do tipo local (e.g., *Porto Alegre*), cujo sintagma seja parte do sintagma de uma entidade mencionada do tipo acontecimento ou organização (e.g., *Brigada Militar de Porto Alegre*), então a entidade inserida é marcada como relacionada com a entidade do tipo local) sobre os resultados do analisador sintáctico PALAVRAS (Bick, 2000), reconhecendo relações de identidade, ocorrência e inclusão entre as entidades mencionadas (Bruckschen et al., 2008).

Finalmente, o sistema REMBRANDT obteve os melhores resultados na tarefa, tendo conseguido uma medida F_1 de 45.02%, utilizando

heurísticas básicas de relacionamento entre entidades com base nas suas unidades, nas suas categorias, e nas ligações das respectivas páginas da Wikipédia (e.g., as entidades mencionadas que tenham sido emparelhadas com uma mesma página da Wikipédia são anotadas como sendo idênticas) (Cardoso, 2008).

Além dos esforços realizados no contexto da tarefa ReRelEM, temos que existem também alguns outros trabalhos anteriores que abordaram a tarefa da extracção automática de relações, a partir de textos em português. Por exemplo Oliveira, Costa e Gomes (2010) apresentaram um sistema simples que utiliza padrões léxico-sintácticos para extrair relações de 5 tipos (i.e., *sinonímia*, *hiperonímia*, *parte-de*, *causa* e *finalidade*) entre termos compostos (i.e., termos modificados por adjectivos ou por preposições), a partir de resumos de artigos da versão portuguesa da Wikipédia, desta forma obtendo informação que pode ser utilizada para criar ou estender redes lexicais ao estilo da WordNet.

García e Gamallo (2011) compararam o impacto de diferentes tipos de características linguísticas (i.e., conjuntos de lemas e de etiquetas morfo-sintácticas, padrões léxico-sintácticos, e dependências sintácticas) na tarefa de extracção de relações, através de aprendizagem automática e utilizando uma técnica de supervisão distante que segue o método geral de Mintz et al. (2009) para a recolha de exemplos de treino desde *infoboxes* e textos da Wikipédia. Os autores avaliaram modelos baseados em máquinas de vectores de suporte com diferentes conjuntos de características, através de experiências com dados da versão portuguesa da Wikipédia, e com o foco na extracção de relações do tipo *ocupação* entre pessoas e actividades profissionais.

Resultados preliminares mostraram que as características baseadas nos padrões léxico-sintácticos obtêm uma maior precisão do que as características baseadas em conjuntos de palavras ou dependências sintácticas, muito embora a combinação de diferentes tipos de características ajude a atingir um compromisso entre a precisão e cobertura, melhorando assim sob o desempenho de modelos que apenas utilizem um único tipo de características. As experiências dos autores mostraram também que modelos que usem padrões léxico-sintácticos baseados apenas nos contextos do meio (i.e., que apenas utilizem palavras que ocorram entre os pares de entidades relacionadas) têm um melhor desempenho do que modelos que utilizem todos os contextos (i.e., a informação proveniente de uma janela de palavras ocorrendo antes, entre, e após as entidades mencionadas). Curiosamente, os autores também mencionam que a revisão manual de algumas instâncias (i.e., aquelas que foram posteriormente utilizadas para medir a qualidade das extrações produzidas) revelou que o método de supervisão distante tem uma precisão de cerca de 80% aquando da recolha automática dos exemplos de treino.

Num outro trabalho relacionado, Gamallo, Garcia e Fernández-Lanza (2012) reportam uma abordagem de domínio aberto para a extracção de relações em vários línguas (i.e., os autores abordaram a tarefa de extrair relações entre pares de entidades mencionadas em textos provenientes das versões em inglês, espanhol, português ou galego da Wikipédia), fazendo uso de um analisador sintáctico multilingue baseado em regras. Especificamente, temos que o método de extracção proposto envolve três etapas, nomeadamente (i) a análise de dependências, em que cada frase do texto de entrada é processada com a ferramenta TreeTagger[7] por forma a atribuir etiquetas morfo-sintáticas às palavras, e é de seguida analisada do ponto de vista das dependências com um *parser* multilingue proposto anteriormente e denominado DepPattern[8], (ii) o encontrar das cláusulas constituintes, onde sob cada frase analisada os autores descobrem as cláusulas verbais e os seus constituintes, incluindo as suas funções (e.g., sujeitos, objectos diretos, atributos e complementos preposicionais), e (iii) a aplicação de regras de extracção, onde alguns padrões são usados sobre as cláusulas constituintes para extrair as relações. Temos, por exemplo, que a regra de extracção mais simples é aplicada sobre as cláusulas que contenham apenas um sujeito e um objecto directo. Nestes casos, os dois componentes são os argumentos da relação, enquanto que a expressão verbal corresponde ao tipo da relação. Infelizmente, Gamallo et al. apenas reportam resultados com este método para o caso da língua inglesa, com avaliações iniciais apontando para uma precisão de cerca de 68%.

O trabalho apresentado neste artigo segue a ideia de utilizar supervisão distante como forma de realizar a tarefa de extracção de relações para textos em português, usando especificamente frases da Wikipédia onde co-ocorram entidades relacionadas na DBPédia. Propomos também uma forma diferente de classificar relações semânticas, baseada num método eficiente para a pesquisa por instâncias similares.

[7] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[8] `http://gramatica.usc.es/pln/tools/deppattern.html`

3 Método Proposto

A abordagem proposta para a classificação de uma relação entre duas entidades mencionadas numa frase, de acordo com o seu tipo semântico, é baseada na ideia de encontrar as relações mais semelhantes numa determinada base de dados de relações exemplo previamente anotadas. O procedimento corresponde essencialmente ao desenvolvimento e aplicação de um classificador baseado na votação ponderada dos kNN vizinhos mais próximos, onde cada exemplar de relação tem um peso correspondente à sua semelhança para com a relação a ser classificada. Os exemplares de relações mais similares têm, por consequência, um peso maior na votação, do que os exemplares que são mais dissimilares.

A representação considerada para cada relação exprime-se essencialmente em termos de tetragramas de caracteres, considerando especificamente as palavras que ocorrem:

1. Entre as duas entidades que constituem a relação binária, isto é, entre as duas subsequências correspondentes aos nomes de entidades que são relacionados.
2. Numa janela de três palavras ocorrendo antes da primeira entidade, e entre as entidades envolvidas na relação.
3. Entre as entidades e numa janela de três palavras após a segunda entidade.

Esta representação segue essencialmente a observação de Bunescu e Mooney (2005a) de que uma relação entre duas entidades é geralmente expressa utilizando apenas palavras que aparecem em um de três padrões básicos, nomeadamente antes-e-entre (i.e., palavras antes e entre as duas entidades envolvidas na relação), entre (i.e., apenas as palavras entre as duas entidades), e entre-e-depois (i.e., palavras que ocorrem entre e depois das duas entidades).

Além dos tetragramas de caracteres, também consideramos palavras correspondentes a preposições, verbos e padrões léxico-sintácticos relacionais, que ocorram nas mesmas janelas textuais consideradas para os tetragramas de cáracteres, extraídos com um modelo de etiquitação morfológica desenvolvido com o pacote OpenNLP[9] e treinado com os dados do corpus CINTIL (Branco e Silva, 2006). As preposições e os verbos são extraídos directamente com base nas etiquetas morfológicas. Os padrões relacionais correspondem a uma regra inspirada no sistema de OIE ReVerb (Fader, Soderland e Etzioni, 2011), em que se extraem sequências de palavras formadas por um verbo seguido de uma preposição, ou de um verbo, seguido de vários nomes, adjectivos ou advérbios, e terminando numa preposição.

A cada tetragrama ou palavra, em cada um dos três grupos (isto é, nos grupos antes-e-entre, entre, e entre-e-depois das entidades envolvidas na relação), é atribuído um identificador único. A semelhança entre duas relações pode ser medida através do coeficiente de similaridade de Jaccard entre cada conjunto de identificadores únicos globais, associados às representações.

Muito embora a maioria dos métodos anteriores para extracção de relações usem representações baseadas em palavras individuais, pensamos que a utilização de tetragramas de caracteres pode trazer algumas vantagens, nomeadamente no melhor lidar com problemas de variabilidade lexical. Também experimentámos utilizar outras representações para as relações, utilizando por exemplo n-gramas de palavras, depois de lematizar o texto. No entanto, observámos que a representação descrita neste secção consegue o melhor compromisso entre a precisão do classificador e o desempenho computacional.

3.1 Geração Automática de Exemplos

A Wikipédia, na sua versão portuguesa para o nosso caso em particular, é um ponto de partida ideal para o desenvolvimento de extractores automáticos de relações, pois trata-se de um recurso abrangente que contém um conjunto muito diversificado de conteúdos aprofundados. Na Wikipédia, além de descrições textuais para conceitos e entidades relevantes, em diferentes domínios do conhecimento, há também informação estruturada sob a forma de *infoboxes*, i.e., tabelas criadas manualmente que apresentam, sob a forma de atributos e valores, factos importantes sobre muitos dos artigos da Wikipédia. Projectos como a DBPédia exploraram a construção automática e a disponibilização de redes de conhecimento derivadas de factos expressos nas *infoboxes* das páginas da Wikipédia em várias línguas, incluindo o português (Auer et al., 2007).

Uma vez que os mesmos factos são frequentemente expressos tanto no texto dos artigos da Wikipédia como nas *infoboxes*, e consequentemente também em recursos como a DBPédia, temos então que combinando as relações da DBPédia com frases constituintes dos artigos na Wikipédia, onde as entidades envolvidas co-ocorram, podemos coleccionar grandes volumes de dados de treino para extractores de relações,

[9] `http://opennlp.apache.org/`

que muito embora sejam ruidosos podem ser úteis dado o seu grande volume (Mintz et al., 2009).

Por exemplo, o artigo da Wikipédia portuguesa sobre o artista *Otis Redding* contém a frase *Otis Redding nasceu na pequena cidade de Dawson, Georgia*. Simultaneamente, a *infobox* deste artigo contém o atributo *origem = Dawson, Georgia* e, consequentemente, a rede da DBPédia contem uma associação do tipo *origem* entre as entidades *Otis Redding* e *Georgia*. Ao combinar a informação da DBPédia com frases dos artigos na Wikipédia, como no exemplo apresentado, podemos gerar dados de treino para um extractor de relações do tipo *origem-de*. Estes dados são muito ruidosos, já que alguns atributos da DBPédia podem não encontrar correspondências em frases da Wikipédia, enquanto outros podem surgir em frases em que as entidades co-ocorrem, mas em que a verdadeira relação não está a ser expressa no texto. No entanto, argumentamos que o grande volume de dados, possível de ser extraído desta forma, compensa o ruído presente nas anotações.

O procedimento geral usado na construção automática da base de dados de exemplos para o extractor de relações é desta forma o seguinte:

1. Recolhem-se da DBPédia todas as relações expressas entre conceitos (i.e., páginas da Wikipédia) correspondentes a pessoas, locais ou organizações. De cada uma destas relações, mantém-se informação sobre as duas entidades que estão relacionadas, e a classe semântica do relacionamento;
2. Para cada relação entre um par de entidades, tal como extraída na primeira etapa, analisamos o texto dos dois artigos da Wikipédia portuguesa correspondentes;
3. O texto dos artigos da Wikipédia é segmentado nas frases constituintes;
4. As frases são filtradas, de modo a manter somente aquelas em que co-ocorrem as duas entidades envolvidas na relação. Este passo de filtragem considera pequenas variações nos nomes das entidades, tal como usados na DBPédia e no nome do artigo da Wikipédia, aquando do mapeamento para com o texto das frases. Desta forma podemos melhorar a abrangência do método proposto. Por exemplo, consideram-se além dos nomes originais, as sequências de caracteres até à primeira vírgula ou parêntesis, dado que muitos conceitos da Wikipédia são desambiguados através da inclusão de mais informação no nome – por exemplo, a página da Wikipédia correspondente ao estado da Georgia nos EUA, é identificada pelo sequência de caracteres *Georgia_(Estados_Unidos)*, embora seja de esperar que muitas frases apenas se refiram a este estado pelo nome de *Georgia*.
5. As frases que resultam da etapa de filtragem anterior são mantidas como exemplares de um determinado tipo de relação semântica.

Após a execução do procedimento descrito acima, vamos obter muito exemplos dos vários tipos de relações semânticas que se encontram codificados na DBPédia, os quais foram por sua vez derivados da informação nas *infoboxes* da Wikipédia. Uma vez que muitos destes tipos de relações correspondem a ligeiras variações de um mesmo conceito semântico (e.g., *locatedInArea* ou *subRegion* são variações de um mesmo conceito que se pode generalizar para *localizado-em*), procedemos a um agrupamento/generalização manual dos diferentes tipos de relações presentes na DBPédia, tendo finalmente obtido um conjunto de dados contendo 10 tipos de relações diferentes, tal como ilustrado na Tabela 1. Importa referir que as associações entre os oito primeiros tipos de relações na Tabela 1 são orientadas (i.e., estes tipos de relações devem ser consideradas como assimétricas), enquanto que as associações entre os últimos dois tipos (i.e., relações do tipo *parceiro* e *não-relacionado/outros*) são simétricas.

Tal como referido atrás, temos que um pequeno sub-conjunto dos exemplos de treino gerados automaticamente foi posteriormente revisto manualmente, por forma a construir uma colecção para a avaliação de resultados. Durante este processo de revisão manual, e ainda que de uma forma muito informal, verificou-se que o método de supervisão distante tem uma precisão de cerca de 80% na atribuição de tipos de relações que se encontrem realmente expressos nas frases, aquando da recolha automática dos exemplos de treino. Este resultado está em concordância com o trabalho anterior de García e Gamallo (2011). Importa no entanto referir que, também aquando

Relação	**Núm. Exemplos**
local-de-enterro-ou-falecimento	6.726
influenciado-por	147
pessoa-chave-em	355
localizado-em	46.236
origem-de	23.664
antepassado-de	266
parte-de	5.142
sucessor-de	496
parceiro	128
não-relacionado/outros	6.441

Tabela 1: Tipos de relações considerados.

Tipo de Relação	Exemplos de Instâncias de Relações
local-de-enterro-ou-falecimento	**Camilo Pessanha** morreu no dia 1 de Março de 1926 em **Macau**, devido ao uso excessivo de Ópio. Classe DBPédia : **deathPlace** Direcção : (entidade1,entidade2)
	Corisco foi enterrado em Jeremoabo, na **Bahia**. Classe DBPédia : **placeOfBurial** Direcção : (entidade2,entidade1)
influênciado-por	O som inicial do **U2** foi influenciado por bandas como **Television** e Joy Division. Classe DBPédia : **influencedBy** Direcção : (entidade1,entidade2)
	Rubem Fonseca escreveu os contos "Chegou o Outono", "Noturno de Bordo"e "Mistura" baseado na linguagem de **Machado de Assis**. Classe DBPédia : **influenced** Direcção : (entidade2,entidade1)
pessoa-chave-em	**Magic Circle Music** foi fundada pelo baixista do Manowar **Joey DeMaio** em 2005. Classe DBPédia : **foundedBy** Direcção : (entidade1,entidade2)
	A **Microsoft** foi fundada em 1975 por **Bill Gates** e Paul Allen. Classe DBPédia : **keyPerson** Direcção : (entidade2,entidade1)

Tabela 2: Exemplos para alguns dos diferentes tipos de relações consideradas.

Tipo de Relação	Padrões Relacionais Extraídos
origem	nasceu em; começou a; competiu em; nasceu a; foi formado em;
influenciado-por	é inspirada por; combinando com; apareceu em; influenciou ministérios de; foi influenciado por;
local-de-enterro-ou-falecimento	nasceu em; morreu em; faleceu em; visconde com; morreu de;
parceiro	é casado com; foi casado com; casou com; casou-se com; compete ao lado;

Tabela 3: Os 5 padrões relacionais mais frequentes para alguns dos tipos de relações semânticas.

do processo de revisão manual, foram detectados vários problemas ao nível da segmentação das frases provenientes dos artigos da Wikipédia (e.g., é comum observar frases que incluem, no seu início ou no final, palavras provenientes do título da secção imediatamente antes da frase). Na construção da colecção manualmente revista para a avaliação de resultados, todos os problemas detectados foram corrigidos.

A Tabela 2 mostra alguns exemplares das diferentes classes de relações consideradas após a generalização, mostrando ainda a classe da relação originalmente expressa na DBPédia, assim como a direcção do relacionamento.

Por outro lado a Tabela 3 mostra alguns dos padrões relacionais mais frequentemente associados a algumas das relações semânticas consideradas, dando assim uma ideia dos valores das características léxico-sintácticas que foram usadas nas representações dos exemplares de relações.

3.2 Pesquisa por Relações Similares

Importa observar que uma abordagem simplista para encontrar os exemplares de relações mais similares entre si, numa base de dados de tamanho N, envolve o cálculo da similaridade entre N^2 pares de exemplares. Este procedimento torna-se rapidamente difícil de escalar para valores grandes de N. Apesar de a tarefa ser paralelizável, é necessário baixar a complexidade $O(N^2)$ para alcançar uma boa escalabilidade. Desta forma, o desenho de operações de pré-processamento adequadas, que facilitem os cálculos de similaridade entre exemplares, assume uma importância relevante. No nosso método, isto é feito pelo cálculo de uma aproximação ao coeficiente de similaridade de Jaccard, obtida através de uma técnica baseada em valores mínimos de funções de dispersão (i.e., *min-hash*), e utilizando ainda uma técnica de dispersão sensível à localização (*Locality-Sensitive Hashing (LSH)*) para encontrar rapidamente as kNN relações mais similares.

A técnica de *min-hash* foi apresentada no trabalho seminal de Broder (1997; Broder et al. (2000)), onde os autores descrevem uma aplicação bem sucedida na detecção de páginas Web duplicadas. Dado um vocabulário Ω de tamanho D (ou seja, o conjunto de todos os elementos representativos usados nas descrições dos exemplares de relações), e dois conjuntos de elementos, S_1 e S_2, onde $S_1, S_2 \subseteq \Omega = \{1, 2, \ldots, D\}$

temos que o coeficiente de similaridade de Jaccard, entre os dois conjuntos de elementos, é dado pela razão entre o tamanho da intersecção de S_1 e S_2, sobre o tamanho da sua união:

$$\begin{aligned} \mathrm{J}(S_1, S_2) &= \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \\ &= \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S1 \cap S_2|} \end{aligned} \quad (1)$$

Os dois conjuntos são mais semelhantes entre si quando o coeficiente de similaridade de Jaccard está perto de 1, e mais dissimilares quando o coeficiente de Jaccard é próximo de 0.

Para pares de conjuntos grandes, calcular eficientemente os tamanhos dos conjuntos resultantes da sua união e intersecção é computacionalmente exigente, uma vez que o número total de elementos a analisar é enorme. No entanto, suponhamos que uma permutação aleatória π é realizada sobre o vocabulário Ω, ou seja:

$$\pi : \Omega \longrightarrow \Omega \ , onde \ \Omega = \{1, 2, \ldots, D\}. \quad (2)$$

Notando que o coeficiente de Jaccard corresponde à razão entre o número de elementos que ocorre simultaneamente em S_1 e S_2, sobre o número de elementos que ocorre em pelo menos um dos conjuntos, temos que um argumento elementar de análise de probabilidades pode mostrar que:

$$\Pr\left(\min(\pi(S_1)) = \min(\pi(S_2))\right) = \mathrm{J}(S_1, S_2) \quad (3)$$

Após a criação de k permutações independentes dos elementos pertencentes a Ω, pode-se estimar a medida de similaridade $\mathrm{J}(S1, S2)$ de forma eficiente e não tendenciosa, como uma distribuição amostral de uma variável aleatória binomial:

$$\begin{aligned} \hat{\mathrm{J}}(S_1, S_2) &= \frac{1}{k} \sum_{j=1}^{k} \mathrm{um}(\min(\pi_k(S_1)) \\ &= \min(\pi_k(S_2))) \end{aligned} \quad (4)$$

Na fórumula acima a função um() devolve o valor de 1 quando para uma dada permutação o elemento mínimo dos dois conjuntos é igual, e o valor de 0 caso contrário.

$$\mathrm{Var}(\hat{\mathrm{J}}(S_1, S_2)) = \frac{1}{k} \mathrm{J}(S_1, S_2)\left(1 - \mathrm{J}(S_1, S_2)\right) \quad (5)$$

Na implementação do mecanismo de *min-hash*, cada uma das permutações independentes corresponde a um valor de uma função de dispersão, no nosso caso considerando 32 *bits* de armazenamento. Cada uma das permutações independentes k está assim associada a uma função de *hash* polinomial $h^k(x)$ que mapeia os membros de Ω para valores distintos. Para qualquer conjunto S, tomarmos os k valores de $\mathrm{h}^k_{min}(S)$, ou seja, o membro de S com o valor mínimo de $h^k(x)$. O conjunto de k valores é referido como a assinatura *min-hash* de um exemplo.

A pesquisa eficiente pelos kNN vizinhos mais próximos é implementada através de uma técnica simples de dispersão sensível a localização, que utiliza as assinaturas de *min-hash* para comprimir as representações das relações em assinaturas pequenas (ou seja, para gerar assinaturas pequenas, do conjunto de todos os tetragramas de caracteres, preposições, verbos, e padrões relacionais, ocorrendo antes-e-entre, entre, e entre-e-depois das entidades envolvidas na relação), ao mesmo tempo preservando a similaridade esperada de qualquer par de instâncias. Esta técnica utiliza L tabelas de dispersão diferentes (ou seja, na nossa implementação, usamos L estruturas de dados persistentes construídas com a biblioteca MapDB[10]), cada uma correspondendo a um n-tuplo das assinaturas *min-hash*, a que nos referimos aqui como uma banda. No momento da classificação, calculamos a assinatura *min-hash* da relação a ser classificada, e de seguida consideramos qualquer relação de exemplo que se encontre associada a um mesmo contentor da estrutura de dados, para qualquer uma das bandas *min-hash*, como uma relação candidata a pertencer ao conjunto das kNN mais similares. Verificamos apenas os pares candidatos, utilizando as assinaturas *min-hash* completas para aproximar o coeficiente de similaridade de Jaccard. Desta forma, podemos evitar as comparações de similaridade com todas as relações na base de dados de exemplos. O Capítulo 3 do livro de Rajaraman e Ullman (2011) descreve o uso da assinaturas *min-hash* com técnicas baseadas em dispersão sensível a localização, em aplicações relacionadas com a pesquisa por itens semelhantes.

Um esboço completo do método de classificação proposto é assim o seguinte. Começando por analisar o conjunto de frases envolvido na indexação dos exemplares de treino:

1. Extraem-se conjuntos de tetragramas de caracteres, preposições, verbos, e padrões relacionais das *substrings* que ocorrem antes-e-entre, entre, e entre-e-depois das entidades envolvidas na relação, para cada relação em cada frase de um determinado conjunto de textos de exemplo. Os exemplos são previamente recolhidos da Wikipédia, tal como explicado na subsecção anterior, correspondendo a frases onde co-ocorrem pares de en-

[10] `http://www.mapdb.org/`

tidades relacionados na DBPédia.

2. As assinaturas *min-hash* são extraídas a partir dos conjuntos gerados na primeira etapa.

3. As assinaturas são divididas em bandas, e os exemplares de relações que estas representam são indexados em L diferentes tabelas de dispersão, com base nos valores presentes nas bandas das assinaturas.

Na classificação de relações, e para verificar se uma dada relação semântica está ou não descrita numa frase, seguem-se os seguintes passos:

1. Começamos pela extracção dos tetragramas de caracteres, preposições, verbos, e padrões de relacionamento, a partir das subsequências que ocorrem antes-e-entre, entre, e entre-e-depois das entidades envolvidas.

2. Gera-se uma assinatura *min-hash* a partir do conjunto gerado no primeiro passo.

3. As relações de exemplo com pelo menos uma banda idêntica no índice construído na fase de indexação são consideradas como candidatas, e a sua semelhança para com a relação a classificar é, então, estimada usando as assinaturas *min-hash* completas.

4. Os exemplos mais semelhantes são mantidos numa lista de prioridades, de onde posteriormente se podem extrair os kNN exemplares mais semelhantes.

5. Os kNN exemplares mais semelhantes são analisados, e a classe semântica da relação é atribuída com base numa votação, ponderada pelo valor de similaridade, entre as classes presentes nos kNN exemplares mais semelhantes.

4 Avaliação Experimental

O método de extracção de relações aqui proposto foi avaliado com base em frases da Wikipédia e de duas formas distintas, nomeadamente:

1. Deixando de fora da fase de indexação, na base de dados de exemplos, as relações correspondentes a uma pequena parte dos dados gerados automaticamente a partir da Wikipédia, correspondente a um conjunto de relações de exemplo verificadas manualmente quanto à sua exactidão.

2. Deixando de fora da fase de indexação 25% dos exemplos de cada classe semântica.

	Conjuntos de Dados		
	Treino	Teste	Total
# Frases	97.363	625	97.988
# Palavras	2.172.125	14.320	2.186.445
# Classes	10	10	10
# Instâncias	89.054	547	89.601
# Entidades únicas	70.716	838	71.119
Média palavras/frase	22,42	24,12	22.43
StDev. palavras/frase	11.39	11.00	11.39
Média instâncias/classe	8.905,4	54,7	8.960,1
StDev. instâncias/classe	14,109.33	64.18	14.172,38

Tabela 4: Caracterização estatística dos conjuntos de dados usados nas diferentes experiências.

A Tabela 4 apresenta uma caracterização estatística do sub-conjunto dos dados que se encontra manualmente verificado (i.e., a coluna assinalada como teste), assim como do sub-conjunto dos dados para os quais não temos anotações manuais (i.e., a coluna assinalada como treino, correspondendo às relações de exemplo que são indexadas nos testes relacionados com o primeiro método experimental), e para o conjunto completo de exemplares de relações. A anotação manual consistiu em verificar se, de facto, as frases que estavam a ser geradas através do processo automatizado correspondem verdadeiramente a exemplos válidos de um tipo semântico de uma relação em particular. O conjunto de dados completo usado nas nossos experiências encontra-se disponibilizado online[11].

Realizámos experiências com representações diferentes das relações (por exemplo, utilizando apenas tetragramas de caracteres, ou usando tetragramas e as características derivadas de etiquetas morfológicas), e também com diferentes parâmetros no método de classificação baseado em assinaturas *min-hash*, através da variação do número de vizinhos mais próximos que foi considerado (i.e., 1, 3, 5 ou 7), variando o tamanho das assinaturas *min-hash* (i.e., 200, 400, 600 ou 800 números inteiros) e o número de bandas LSH considerado (i.e., 25 ou 50 bandas). Importa notar que, ao usar b bandas LSH, cada uma com r valores, temos que a probabilidade de as assinaturas *min-hash* de dois conjuntos S_1 e S_2 concordarem em todas os valores de pelo menos uma banda, gerando-se assim um par candidato, é de $1-(1-\mathrm{J}(S_1,S_2)^r)^b$. Com 50 bandas e uma assinatura *min-hash* de tamanho 600, cerca de um em cada mil pares com uma similaridade até 85% vai deixar de se tornar um par candidato através do método LSH e, por consequência, vai ser um falso negativo. Especificamente com estes parâmetros, as relações com uma similaridade abaixo de 85% são muito susceptíveis de ser

[11] `http://dmir.inesc-id.pt/project/DBpediaRelations-PT_01_in_English`

descartadas através do método LSH, o que pode contribuir para a confiança em uma classificação correta (ou seja, estamos de certa forma a trocar precisão por abrangência, nos parâmetros considerados para a indexação).

Como medidas de avaliação, utilizámos principalmente as macro-médias da precisão (P), abrangência (A), e da medida F_1 sobre todos os tipos de relações, excepto o tipo *não-relacionado/outro*. Usamos assim a macro-média das pontuações sob 18 classes de relações semânticas, uma vez que temos duas direcções possíveis para 8 tipos semânticos das relações inferidas a partir DBPédia (i.e., as relações *parceiro* e *não-relacionado/outro* são bidireccionais).

A Tabela 5 apresenta os resultados obtidos para diferentes representações e parâmetros de indexação, quando se considera o conjunto de dados com as anotações manuais. A Tabela 6 apresenta os resultados obtidos para diferentes parâmetros de indexação com o conjunto completo de características, sob 25% do conjunto completo de relações de cada classe. Os resultados mostram que o método usando supervisão distante, juntamente com a técnica de classificação proposta, permite extrair relações com uma exactidão razoável. Também podemos verificar que os valores das diferentes métricas de avaliação são ligeiramente inferiores no caso dos testes com os 25% do conjunto total de exemplos. Isto indica que os resultados medidos com a colecção manualmente anotada podem ser encarados como um limite superior a uma aproximação da verdadeira exactidão do sistema.

Os resultados da Tabela 5 indicam também que a combinação de tetragramas de caracteres, verbos, preposições, e padrões relacionais, proporciona um melhor desempenho de identificação e classificação. Os resultados sugerem ainda que a utilização dos cinco ou sete primeiros vizinhos, em vez de apenas o exemplo mais semelhante, resulta num aumento de desempenho.

A Tabela 7 apresenta resultados individuais por classe sob 25% dos exemplares de cada relação, considerando as características de representação e indexação que obtiveram o melhor desempenho nos resultados das Tabelas 5 e 6. Isto corresponde a uma configuração com:

- Tetragramas de caracteres, verbos, preposições e padrões relacionais para representar as relações semânticas;
- Assinaturas *min-hash* com tamanho = 800;
- Número de bandas no método LSH = 25;
- Os sete vizinhos mais próximos;

Além dos resultados para a configuração regular de classificação de relações de acordo com os tipos e com a direcção, também apresentamos resultados para uma avaliação em que se ignoram as direcções das relações, bem como os resultados obtidos para a classe correspondente ao tipo *não-relacionado/outro*. Finalmente, esta tabela apresenta também uma avaliação global dos resultados obtidos através da medida de exactidão, a qual mede a porção de classificações corretas, dando assim uma maior importância aos tipos de relações com maior número de ocorrências no corpus, o que não acontece com as macro-médias. Os resultados mostram que classes como *origem-de* e *parte-de* são relativamente fáceis de identificar e classificar, enquanto que classes como *influenciado-por* ou *sucessor-de* são muito mais difíceis de identificar e classificar correctamente. Note-se por exemplo que, para a classe correspondente a *influenciado-por*, o conjunto de dados indexado contém apenas 110 relações de exemplo, enquanto que o conjunto de exemplos usado para a medição de resultados nesta classe tem apenas 35 exemplares de relações.

5 Conclusões e Trabalho Futuro

A utilização de técnicas de Extracção de Informação como forma de suportar a criação de bases de conhecimento, em larga escala, a partir de repositórios de documentos de texto, tais como a Web ou como colecções de textos jornalísticos, é objecto actual de estudo intenso. No entanto, as melhores abordagens existentes, para a extracção de relações semânticas, não são facilmente transponíveis para línguas ou domínios diferentes. Temos ainda que os métodos supervisionados requerem grandes quantidades de dados anotados, e têm uma complexidade computacional elevada. Por outro lado, as técnicas independentes de domínio apresentam resultados de baixa precisão e não normalizam as relações.

Neste artigo foi proposta uma abordagem de supervisão distante para a classificação de relações extraídas de textos escritos em português, suportada por dados extraídos da Wikipédia e da DBPédia, e baseada na medição de similaridade entre as relações a classificar e relações armazenadas numa base de dados de relações de exemplo. No método proposto, os exemplos de treino são recolhidos automaticamente a partir da Wikipédia, correspondendo a frases que expressam relações entre pares de entidades extraídas da DBPédia. Estes exemplos são representados como assinaturas *min-hash* de conjuntos de elementos, originalmente contendo

Características	Min Hash	1 kNN P	1 kNN A	1 kNN F_1	3 kNN P	3 kNN A	3 kNN F_1	5 kNN P	5 kNN A	5 kNN F_1	7 kNN P	7 kNN A	7 kNN F_1
Tetragramas	200/25	0.492	0.400	0.441	0.627	0.426	0.507	0.716	0.423	0.532	0.724	0.429	0.539
	200/50	0.489	0.400	0.440	0.625	0.425	0.506	0.716	0.423	0.532	0.726	0.430	0.540
	400/25	0.476	0.405	0.438	0.559	0.418	0.478	0.724	0.434	0.543	**0.736**	**0.443**	**0.553**
	400/50	0.474	0.405	0.437	0.557	0.423	0.481	0.715	0.434	0.540	0.731	0.441	0.550
	600/25	0.609	0.435	0.508	0.645	0.437	0.521	0.688	0.440	0.537	0.663	0.440	0.529
	600/50	0.583	0.435	0.498	0.646	0.437	0.521	0.686	0.433	0.531	0.719	0.441	0.547
	800/25	0.545	0.426	0.478	0.610	0.430	0.504	0.651	0.434	0.521	0.640	0.442	0.523
	800/50	0.541	0.423	0.475	0.611	0.432	0.506	0.652	0.436	0.523	0.643	0.444	0.525
Tetragramas e Verbos	200/25	0.476	0.414	0.443	0.628	0.437	0.515	0.713	0.429	0.536	0.718	0.432	0.539
	200/50	0.474	0.414	0.442	0.628	0.437	0.515	0.713	0.429	0.536	0.718	0.432	0.539
	400/25	0.499	0.417	0.454	0.563	0.430	0.488	0.725	0.437	0.545	0.729	0.442	0.550
	400/50	0.497	0.417	0.453	0.565	0.436	0.492	0.674	0.440	0.532	0.729	0.443	0.551
	600/25	0.580	0.425	0.491	0.640	0.442	0.523	0.669	0.439	0.530	0.728	0.435	0.545
	600/50	0.553	0.425	0.481	0.641	0.442	0.523	0.724	0.439	0.547	0.728	0.441	0.549
	800/25	0.549	0.424	0.479	0.615	0.433	0.508	0.720	0.443	0.549	**0.736**	0.441	**0.551**
	800/50	0.549	0.424	0.479	0.615	0.433	0.508	0.712	**0.447**	0.549	0.731	0.438	0.548
Tetragramas, Verbos e Preposições	200/25	0.477	0.403	0.437	0.628	0.431	0.511	0.720	0.432	0.540	0.723	0.438	0.546
	200/50	0.478	0.404	0.438	0.628	0.431	0.511	0.666	0.432	0.524	0.670	0.438	0.530
	400/25	0.522	0.431	0.472	0.574	0.432	0.493	0.732	0.446	0.554	0.731	0.442	0.551
	400/50	0.522	0.431	0.472	0.578	0.441	0.500	0.679	0.446	0.538	0.732	0.445	0.554
	600/25	0.581	0.427	0.492	0.630	0.432	0.513	0.673	0.446	0.536	0.677	0.441	0.534
	600/50	0.554	0.427	0.482	0.631	0.432	0.513	0.726	0.439	0.547	0.731	0.442	0.551
	800/25	0.548	0.426	0.479	0.616	0.435	0.510	0.721	**0.449**	0.553	**0.733**	0.447	**0.555**
	800/50	0.545	0.423	0.476	0.620	0.446	0.519	0.721	0.445	0.550	0.732	0.446	0.554
Tetragramas, Verbos, Preposições e Padrões Relacionais	200/25	0.472	0.404	0.435	0.629	0.436	0.515	0.724	0.436	0.544	0.723	0.440	0.547
	200/50	0.474	0.404	0.436	0.575	0.436	0.496	0.671	0.436	0.529	0.670	0.440	0.531
	400/25	0.521	0.429	0.471	0.572	0.429	0.490	0.730	0.443	0.551	0.731	0.441	0.550
	400/50	0.521	0.429	0.471	0.573	0.436	0.495	0.680	0.447	0.539	**0.732**	0.444	0.553
	600/25	0.579	0.423	0.489	0.628	0.429	0.510	0.673	0.446	0.536	0.678	0.437	0.531
	600/50	0.552	0.423	0.479	0.629	0.428	0.509	0.728	0.446	0.553	0.731	0.438	0.548
	800/25	0.547	0.423	0.477	0.616	0.433	0.509	0.715	0.445	0.549	0.723	0.444	0.550
	800/50	0.544	0.420	0.474	0.618	0.439	0.513	0.716	0.444	0.548	0.731	**0.449**	**0.556**

Tabela 5: Resultados para diferentes representações das relações e parâmetros de indexação.

Características	Min Hash	1 kNN P	1 kNN A	1 kNN F_1	3 kNN P	3 kNN A	3 kNN F_1	5 kNN P	5 kNN A	5 kNN F_1	7 kNN P	7 kNN A	7 kNN F_1
Tetragramas, Verbos, Preposições e Padrões Relacionais	200/25	0.448	0.353	0.395	0.460	0.345	0.394	0.492	0.331	0.396	0.487	0.325	0.390
	200/50	0.450	0.354	0.396	0.459	0.347	0.395	0.489	0.332	0.395	0.507	0.328	0.398
	400/25	0.440	0.350	0.390	0.448	0.344	0.389	0.468	0.328	0.386	0.479	0.320	0.384
	400/50	0.439	0.351	0.390	0.445	0.343	0.387	0.465	0.327	0.384	0.483	0.321	0.386
	600/25	0.461	0.358	0.403	0.466	0.353	0.401	0.482	0.337	0.397	0.469	0.324	0.383
	600/50	0.461	**0.360**	0.404	0.463	0.353	0.401	0.490	0.340	0.401	0.492	0.329	0.394
	800/25	0.446	0.358	0.397	0.462	0.350	0.398	0.492	0.338	0.401	**0.516**	0.333	**0.405**
	800/50	0.445	0.358	0.397	0.453	0.349	0.394	0.484	0.336	0.397	0.510	0.333	0.403

Tabela 6: Resultados obtidos sob 25% do conjunto completo de instâncias de cada classe.

tetragramas de caracteres assim como outros elementos representativos, e indexados numa estrutura de dados que implementa a ideia de dispersão sensível a localização. Para verificar qual a relação semântica que se encontra expressa entre um determinado par de entidades, são procurados os kNN exemplos de treino mais similares, e a relação é atribuída com base numa votação ponderada. Testes com um conjunto de dados da Wikipédia comprovam a adequabilidade do método proposto, sendo que o mesmo é, por exemplo, capaz de extrair 10 tipos diferentes de relações semânticas, oito deles correspondendo a tipos de relações assimétricos, com uma pontuação média de 55.6% em termos da medida F_1.

Apesar dos resultados interessantes, há também muitos desafios em aberto para trabalho futuro. Temos, por exemplo, que a maioria dos métodos baseados em *kernels*, do actual estado-da-arte, exploram semelhanças entre representações de relações baseadas em grafos, derivados simultâneamente de informações lexicais e de estruturas resultantes de uma análise sintáctica e de dependências (Nguyen, Moschitti

Relação	**Direcção**	**Instâncias** (treino/teste)	**Assimétricas** P	A	F_1	**Simétricas** P	A	F_1
local-de-enterro-ou-falecimento	(e1,e2)	4.788/1.596	0.802	0.595	0.683	0.806	0.574	0.671
	(e2,e1)	257/85	0.375	0.035	0.065			
influenciado-por	(e1,e2)	84/28	0.000	0.000	0.000	0.000	0.000	0.000
	(e2,e1)	26/9	1.000	0.111	0.199			
pessoa-chave-em	(e1,e2)	106/35	0.500	0.086	0.146	0.233	0.079	0.117
	(e2,e1)	161/53	0.200	0.113	0.145			
localizado-em	(e1,e2)	33.639/11.213	0.916	0.929	0.922	0.924	0.922	0.923
	(e2,e1)	1.038/346	0.395	0.087	0.142			
origem-de	(e1,e2)	16.784/5.594	0.723	0.806	0.807	0.733	0.908	0.811
	(e2,e1)	965/321	0.664	0.567	0.612			
antepassado-de	(e1,e2)	151/50	0.471	0.800	0.593	0.545	0.727	0.623
	(e2,e1)	49/16	0.000	0.000	0.000			
parte-de	(e1,e2)	2.590/863	0.541	0.544	0.543	0.680	0.576	0.623
	(e2,e1)	1.267/422	0.574	0.275	0.372			
sucessor-de	(e1,e2)	117/39	0.400	0.051	0.091	0.541	0.161	0.248
	(e2,e1)	255/85	0.359	0.165	0.226			
parceiro	—	96/32	—	—	—	0.600	0.188	0.286
não-relacionado/outros	—	4.831/1.610	—	—	—	0.767	0.543	0.636
Macro-médias	—	—	0.516	0.333	0.405	0.583	0.468	0.494
Exactidão	—	—		0.813			0.834	

Tabela 7: Resultados obtidos individualmente para cada classe e direcção de relacionamento, sob 25% do conjunto completo de instâncias de cada classe.

e Riccardi, 2009). Estudos recentes têm proposto métodos baseados em assinaturas *min-hash* para comparar grafos (Teixeira, Silva e Jr., 2012). Para trabalho futuro, seria interessante experimentar a aplicação destes métodos na tarefa de extracção de relações em textos, usando desta forma representações ricas para os exemplos de relações, baseadas em grafos.

Desde o trabalho seminal de Broder (1997) sobre a utilização de assinaturas *min-hash* para a detecção de páginas Web duplicadas, ocorreram desenvolvimentos teóricos e metodológicos consideráveis, em termos da aplicação deste tipo de abordagens. Para trabalho futuro, gostaríamos de avaliar a abordagem b-bit *minwise hashing* de Li e König (2010) para melhorar a eficiência de armazenamento, experimentar com a extensão proposta por Chum, Philbin e Zisserman (2008) para aproximar medidas de similaridade entre histogramas de valores, e com uma abordagem em duas etapas semelhante à do sistema de desambiguação de entidades KORE (Hoffart et al., 2012), onde documentos textuais são representados por frases chave, que por sua vez são representados como conjuntos de n-gramas.

Finalmente gostaríamos de realizar experiências, com o método proposto neste artigo, sobre outras colecções de dados, de forma a avaliar a técnica de extracção de relações em textos de outros géneros, tais como artigos técnicos, textos literários, documentos jurídicos, etc. Em particular, seria interessante aplicar o método proposto à colecção de textos do ReRelEM, como forma de validar o método de supervisão distante proposto neste artigo. Gostaríamos assim de experimentar com dados derivados de um mapeamento das classes da DBPédia para com as classes do ReRelEM, permitindo assim a validação dos resultados.

Ainda no que se refere a experiências com outros tipos de dados, importa referir que embora apenas tenhamos feito algumas experiências iniciais relacionadas com a utilização do método proposto na extracção de relações em textos provenientes de outros domínios (e.g., usando frases da Wikipédia como dados de treino e tentando extrair relações em frases provenientes de textos jornalísticos, posteriormente observando a qualidade dos resultados de um modo informal), importa referir que os textos da Wikipédia constituem um género muito específico, onde existem determinados padrões que são frequentemente usados como forma de expressar relações (e.g., as relações do tipo *origem-de* são tipicamente expressas à custa de um padrão em que a primeira menção a um determinado nome de pessoa é seguida do local e data de nascimento, entre parêntesis). Em textos de outros domínios, o método proposto vai muito provavelmente obter resultados diferentes em termos da qualidade

das extracções, sendo que os nossos testes iniciais apontam no sentido de ser difícil vir a usar frases da Wikipédia como forma de aprender bons extractores de relações para outros domínios.

Agradecimentos

Este trabalho foi suportado pela Fundação para a Ciência e Tecnologia (FCT), através do projecto com referência PTDC/EIA-EIA/109840/2009 (SInteliGIS), assim como através dos projectos com referências PTDC/EIA-EIA/115346/200912 (SMARTIES), UTA-Est/MAI/0006/2009 (REACTION), e através do financiamento plurianual do laboratório associado INESC-ID com a referência PEst-OE/EEI/LA0021/2013. O autor David Batista foi também suportado pela bolsa de doutoramento da FCT com referência SFRH/BD/70478/2010.

Referências

Airola, Antti, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, e Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. Em *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, e Zachary Ives. 2007. DBpedia: a nucleus for a web of open data. Em *Proceedings of the International Conference on the Semantic Web and of the Asian Conference on the Semantic Web*.

Batista, David S., Rui Silva, Bruno Martins, e Mário J. Silva. 2013. A minwise hashing method for addressing relationship extraction from text. Em *Proceedings of the International Conference on Web Information Systems Engineering*.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press Aarhus.

Blessing, Andre e Hinrich Schütze. 2010. Fine-grained geographical relation extraction from wikipedia. Em *Proceedings of the International Conference on Language Resources and Evaluation*.

Branco, António e João Ricardo Silva. 2006. A suite of shallow processing tools for portuguese: Lx-suite. Em *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations*.

Brin, Sergey. 1999. Extracting patterns and relations from the world wide web. Em *Proceedings of the International Workshop on The World Wide Web and Databases*.

Broder, Andrei. 1997. On the resemblance and containment of documents. Em *Proceedings of the Conference on Compression and Complexity of Sequences*.

Broder, Andrei, Moses Charikar, Alan M. Frieze, e Michael Mitzenmacher. 2000. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3).

Bruckschen, Mírian, José Guilherme Camargo de Souza, Renata Vieira, e Sandro Rigo, 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. Linguateca.

Bunescu, Razvan e Raymond Mooney. 2005a. Subsequence kernels for relation extraction. Em *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Bunescu, Razvan C. e Raymond J. Mooney. 2005b. A shortest path dependency kernel for relation extraction. Em *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Cardoso, Nuno. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em *Actas do Encontro do Segundo HAREM*.

Chaves, Marcírio, 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo Geo-Ontologias para Reconhecimento de Relações Entre Locais: a participação do SEI-Geo no Segundo HAREM. Linguateca.

Chum, Ondrej, James Philbin, e Andrew Zisserman. 2008. Near duplicate image detection: min-hash and TF-IDF weighting. Em *Proceedings of the British Machine Vision Conference*.

Culotta, Aron, Andrew McCallum, e Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. Em *Proceedings of*

the Conference of the North American Chapter of the Association of Computational Linguistics.

Culotta, Aron e Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Etzioni, Oren, Michele Banko, Stephen Soderland, e Daniel S. Weld. 2008. Open information extraction from the web. *Communication of the ACM*, 51(12):68–74.

Fader, Anthony, Stephen Soderland, e Oren Etzioni. 2011. Identifying relations for open information extraction. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota, 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo Relaçoes semânticas do ReRelEM: além das entidades no Segundo HAREM. Linguateca.

Gamallo, Pablo, Marcos Garcia, e Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. Em *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP.*

García, Marcos e Pablo Gamallo. 2011. Evaluating various linguistic features on semantic relation extraction. Em *Proocedings of the Conference on Recent Advances in Natural Language Processing.*

GuoDong, Zhou, Su Jian, Zhang Jie, e Zhang Min. 2005. Exploring various knowledge in relation extraction. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Hachey, Ben, Claire Grover, e Richard Tobin. 2012. Datasets for generic relation extraction. *Natural Language Engineering*, 18(1).

Hendrickx, Iris, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, e Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. Em *Proceedings of the International Workshop on Semantic Evaluation.*

Hoffart, Johannes, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, e Gerhard Weikum. 2012. Kore: keyphrase overlap relatedness for entity disambiguation. Em *Proceedings of the International Conference on Information and Knowledge Management.*

Hoffmann, Raphael, Congle Zhang, e Daniel S Weld. 2010. Learning 5000 relational extractors. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Kambhatla, Nanda. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Posters and Demonstrations.*

Kim, S., J. Yoon, J. Yang, e S. Park. 2010. Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, 11(107).

Krause, Sebastian, Hong Li, Hans Uszkoreit, e Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. Em *Proceedings of the International Conference on The Semantic Web.*

Li, Ping e Christian König. 2010. b-Bit minwise hashing. Em *Proceedings of the International Conference on World Wide Web.*

Mintz, Mike, Steven Bills, Rion Snow, e Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics and of the International Conference on Natural Language Processing of the Asian Federation of Natural Language Processing.*

Mota, Cristina e Diana Santos. 2008. *Desafios na avaliação onjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.* Linguateca.

Nguyen, Truc-Vien T., Alessandro Moschitti, e Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Oliveira, Hugo Gonçalo, Hernani Costa, e Paulo Gomes. 2010. Extracção de conhecimento léxico-semântico a partir de resumos da wikipédia. *Actas do II Simpósio de Informática.*

Pantel, Patrick e Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Rajaraman, Anand e Jeffrey Ullman, 2011. *Mining of Massive Datasets*, capítulo 3. Finding Similar Items. Cambridge University Press.

Riedel, Sebastian, Limin Yao, Benjamin M. Marlin, e Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. Em *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*.

Riedel, Sebastian, Limin Yao, e Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. Em *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*.

Soderland, Stephen e Bhushan Mandhani. 2007. Moving from textual relations to ontologized relations. Em *Proceedings of the AAAI Spring Symposium on Machine Reading*.

Teixeira, Carlos, Arlei Silva, e Wagner Jr. 2012. Min-hash fingerprints for graph kernels: A trade-off among accuracy, efficiency, and compression. *Journal of Information and Data Management*, 3(3).

Wu, Fei e Daniel S Weld. 2010. Open information extraction using wikipedia. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zelenko, Dmitry, Chinatsu Aone, e Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*.

Zhao, Shubin e Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zhou, Guodong e Min Zhang. 2007. Extracting relation information from text documents by exploring various types of knowledge. *Information Processing and Management*, 43(4).