Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia

Hugo Gonçalo Oliveira*, Hernani Costa, Paulo Gomes
hroliv@dei.uc.pt, hpcosta@student.dei.uc.pt, pgomes@dei.uc.pt

Cognitive and Media Systems Group
Centro de Informática e Sistemas
Universidade de Coimbra, Portugal

Resumo Este artigo apresenta um sistema para a aquisição automática de relações semânticas a partir de texto em português, o que pode ser visto como um passo central na construção automática de recursos léxico-semânticos. O sistema foi aplicado à Wikipédia, actualmente uma enorme fonte de conhecimento livre. Os resultados obtidos e a sua avaliação são discutidos, as actuais limitações referidas e são ainda apresentadas várias ideias para futuras melhorias.

Abstract This paper presents a system for the automatic acquisition of semantic relations from Portuguese text, which can be seen as core step in the automatic construction of lexico-semantic resources. The system was applied to Wikipedia, currently a huge and free source of knowledge. The obtained results are shown and their evaluation is discussed together with the current limitations and cues for further improvement.

1 Introdução

A realização de tarefas, cada vez mais comuns, onde é necessário compreender as interacções entre as palavras e os seus significados, tal como a resposta automática a perguntas, a tradução automática ou a recuperação de informação, levou à criação de recursos semânticos computacionais de larga cobertura, como as ontologias lexicais, de onde se destaca, para o inglês, a WordNet de Princeton [9]. No entanto, a construção e a manutenção deste tipo de recurso envolve muito trabalho intensivo, realizado por humanos. De forma a contornar este problema, têm nas últimas décadas surgido várias propostas para, a partir de texto, extrair automaticamente conhecimento léxico-semântico que pode ser utilizado para criar ou para ampliar uma ontologia lexical.

Estas abordagens têm sido aplicadas a diferentes tipos de texto, e conhecimento léxico-semântico vem sendo extraído a partir de recursos estruturados, como os dicionários [6] [17] [12], ou não estruturados, como os corpos [13] [5] [10]. Se por um lado há vantagens em utilizar dicionários, por estes se encontrarem já estruturados em palavras e significados e ainda

* Financiado pela bolsa FCT SFRH/BD/44955/2008

por utilizarem um vocabulário simples, quase previsível, este tipo de recurso contém conhecimento limitado, é normalmente estático e nem sempre se encontra disponível para fins de investigação. Por outro lado, hoje em dia é possível encontrar muito texto pela Web, praticamente acerca de qualquer assunto, mas cujo processamento não é tão simples devido à existência de menos restrições sintácticas e à utilização de vocabulário mais variado e mais ambíguo. Um terceiro tipo de recurso, que podemos considerar semi-estruturado, é a enciclopédia, onde existem também entradas para diferentes entidades, mas cujas descrições são mais extensas, podendo ser encaradas como texto de corpos. Além disso, o conteúdo das enciclopédias não se limita a informação sobre as palavras e inclui mais conhecimento sobre o mundo e saber humano.

Assim, também devido à sua disponibilidade na Web, no últimos anos tornou-se frequente a utilização de enciclopédias, como a Wikipédia[1], para extrair informação. Tendo em conta a sua construção colaborativa, este recurso é uma enorme fonte de informação em permanente evolução. Para o inglês, a Wikipédia foi já utilizada numa grande quantidade de tarefas, onde destacamos a extracção de relações taxonómicas [14] e de outras relações léxico-semânticas, com vista ao enriquecimento da WordNet [19]. A utilidade da Wikipédia na extracção de conhecimento léxico-semântico é apontada por [21], que implementaram um interface para o acesso programático a este recurso e também ao Wikcionário. Além disso, a descrição de alguns trabalhos que utilizam a Wikipédia para extrair conceitos, relações, factos e descrições pode encontrar-se em [15]. Também para o português a Wikipédia se revelou ser um importante recurso, por exemplo no apoio à identificação de entidades mencionadas (EM) [4].

O trabalho aqui descrito enquadra-se num projecto que tem como objectivo final a construção automática de uma ontologia lexical para o português onde, entre outros recursos, a Wikipédia é também explorada. Mais precisamente, são extraídas relações semânticas a partir dos resumos da versão portuguesa da Wikipédia de forma a obter informação que pode ser utilizada para criar um novo recurso ou para enriquecer recursos lexicais já existentes, como o PAPEL [11], uma rede lexical extraída automaticamente a partir de um dicionário.

Começamos por apresentar as fases do nosso sistema que se baseia num conjunto de gramáticas semânticas, onde estão presentes padrões textuais indicadores de relações. De seguida descrevemos a experimentação realizada que inclui: a extracção de triplos a partir da Wikipédia; a análise dos resultados; a avaliação manual de uma amostra de resultados; a análise dos principais padrões textuais que originaram triplos; e uma proposta para avaliação automática, cuja utilidade não foi contudo comprovada. Por fim concluímos ao apontar algumas limitações actuais do sistema e referimos ideias para trabalho futuro.

2 Extracção automática de relações semânticas

O sistema de extracção de relações semânticas que estamos a desenvolver é constituído por vários módulos (ver figura 1) e está centrado num conjunto de

[1] http://wikipedia.org

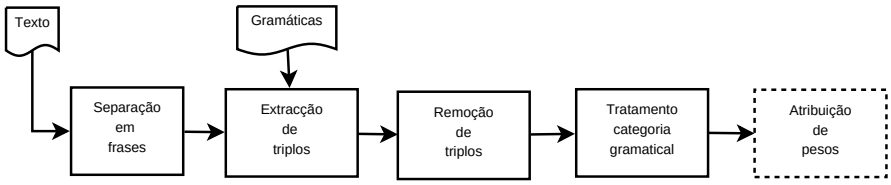

Figura 1. Os módulos do sistema de extracção.

gramáticas semânticas, construídas com base em padrões que indicam relações em texto escrito em português. Até ao momento, o sistema extrai relações de sinonímia, hiperonímia, parte, causa e finalidade. Exemplos destas relações e de alguns dos padrões ou palavras chave utilizados na sua extracção podem encontrar-se na apresentação dos resultados obtidos, mais propriamente na tabela 1 e na tabela 3.

Como o sistema está preparado para analisar texto frase a frase, o primeiro módulo prepara o texto fornecido, separando-o em frases. Na fase de extracção, cada frase é analisada e é obtida uma uma árvore de derivação por gramática. Em cada árvore o sistema procura por nós que identificam um padrão, dentro dos quais poderão existir nós identificadores dos argumentos de uma relação, cujo conteúdo serão termos, ou enumerações de termos, a ser combinados num triplo relacional. Por exemplo, ao encontrar os nós HIPERONIMO e HIPONIMO, o sistema vai extrair o triplo *hiper* HIPERONIMO_DE *hipo*, em que *hiper* e *hipo* são respectivamente os conteúdos de HIPERONIMO e HIPONIMO.

Na versão actual das gramáticas optámos por extrair relações entre termos compostos, ou seja, se um termo ocorrer modificado por um adjectivo (p.e. *computador pessoal*) ou por uma preposição (p.e. *sistema de controlo*) é extraído dessa forma, e pode dar origem a um termo com várias palavras (p.e. *movimento de massa exclusivo das regiões vulcânicas*). Futuramente, após avaliar a relevância destes termos, será possível tomar decisões relativamente à sua manutenção, possibilitando também uma melhor organização do recurso.

Ainda na fase de extracção, o sistema tira partido de dois padrões léxico-sintácticos, `[N ADJ]` e `[N de|do|da|com|para N]`, para obter relações de hiperonímia a partir de termos compostos. Por exemplo, a partir dos termos *computador pessoal* e *sistema de controlo* são extraídos respectivamente os triplos *computador* HIPERONIMO_DE *computador_pessoal* e *sistema* HIPERONIMO_DE *sistema_de_controlo*. Neste tipo de extracção, o segundo padrão mencionado não é aplicado se o primeiro `N` se tratar de uma palavra sem conteúdo (p.e. *tipo*, *forma*) ou que implique uma relação de parte (p.e. *parte*, *membro*, *grupo*, *conjunto*) e não de hiperonímia, chamadas, no contexto da análise de dicionários, cabeças vazias (do inglês *empty heads*)[6].

Para identificar as categorias gramaticais das palavras é previamente realizada a análise morfo-sintáctica de cada frase, utilizando um modelo para o *pos-tagger* fornecido no pacote OpenNLP[2], treinado com o Bosque, uma

[2] http://opennlp.sourceforge.net/

parte do treebank Floresta Sintá(c)tica [1] completamente revista por linguistas. No entanto, as gramáticas contêm essencialmente padrões lexicais e apoiam-se nas categorias gramaticais apenas para identificar adjectivos. Além da análise morfo-sintáctica, cada palavra da frase é lematizada, também recorrendo a um modelo do OpenNLP a que foi acrescentado um pequeno conjunto de regras para passagem de plural para singular.

Após a extracção, triplos cujos argumentos estejam numa lista de palavras não pretendidas (essencialmente *stopwords*) são removidos. A penúltima fase possibilita remover triplos ou alterar o nome da sua relação com base na categoria gramatical dos seus argumentos, baseando-se numa especificação onde, para cada nome de relação extraída, poderá existir um segundo nome de acordo com a categoria gramatical dos seus argumentos. Se pretendido, é também possível lematizar os argumentos dos triplos, com base no lema obtido anteriormente.

Estamos ainda a ponderar a inclusão de uma fase em que os triplos recebem pesos de acordo não só com a frequência com que foram extraídos, mas também com o valor de métricas distribucionais calculadas na Web ou numa colecção de documentos, tal como [5] ou [20] sugerem. A este respeito verificamos [7] que a qualidade de triplos de hiperonímia e também de parte está correlacionada com o valor de algumas métricas distibucionais em corpos, como o LSA e o coeficiente de Jaccard. Os pesos poderão depois ser utilizados para eliminar triplos pouco relevantes ou cuja probabilidade de estarem correctos seja baixa.

3 Experimentação com a Wikipédia

Para testar o nosso sistema optámos por aplicá-lo a resumos da versão portuguesa da Wikipédia. Estes foram escolhidos por descreverem em poucas palavras o conteúdo de cada artigo, tendo por isso a informação mais relevante nele contida e menos variações ao nível da estrutura.

3.1 Preparação

Cedo verificámos que grande parte dos conteúdos da Wikipédia são demasiado específicos para serem centrais na construção de uma ontologia lexical, como é o caso de artigos sobre personalidades, organizações ou épocas históricas. Devido a este problema procuramos uma forma de filtrar resumos associados a EM, o que levaria também a uma diminuição da quantidade de texto a processar.

Para tal, utilizamos os resumos disponibilizados pelo projecto DBpedia [2] e a taxonomia definida no seu âmbito. Com vista à construção de uma base de conhecimento, a DBpedia mapeia a Wikipédia numa taxonomia onde a cada artigo é atribuído um ou vários tipos de alto nível, como por exemplo *Person*, *Place*, *Organization*, *MeanOfTransportation*, *Device* ou *Species*, além de tipos mais específicos como *Writer*, *Airport*, *SoccerClub*, *Bird*, *Automobile* ou *Weapon*.

Ainda que a atribuição de tipos não se encontre disponível para a versão portuguesa da Wikipédia, há uma correspondência entre os identificadores das entradas nas várias línguas que dizem respeito ao mesmo assunto. Por isso

utilizamos os tipos atribuídos às entradas da versão inglesa para filtrar da versão portuguesa entradas do tipo *Person*, *Place*, *Organization*, *Event*, entre outros associados a EM. Apesar de haver várias entradas da Wikipédia portuguesa sem correspondência, cerca de 30% dos 368.521 resumos originais foi removido, perfazendo as 494.187 frases a que chamaremos o conjunto de resumos A.

Ainda assim, ficamos com muito texto que não nos interessava processar, de onde destacamos entradas acerca de geografia portuguesa e brasileira. Por isso, à custa de perdermos entradas interessantes e que só existem na Wikipédia portuguesa, optamos por diminuir ainda mais o conjunto de frases mantendo apenas entradas que, através da taxonomia, conseguimos confirmar pertencerem aos tipos: *Species*, *AnatomicalStructure*, *ChemicalCompound*, *Disease*, *Currency*, *Drug*, *Activity*, *Language*, *MusicGenre*, *Colour*, *EthnicGroup* e *Protein*. Além disso, apesar de vários resumos serem constituídos por duas ou três frases, optámos por utilizar apenas a primeira frase de cada um. Desta forma ficamos com um total de 37.898 frases para processar, que constituem o conjunto B, aquele que mais exploramos nesta experiência.

3.2 Resultados

As quantidades de triplos extraídos a partir de ambos os conjuntos (A e B), antes (Total) e depois (S/rep) de remover triplos repetidos, são apresentadas na tabela 1 juntamente com alguns exemplos. Para a hiperonímia separamos os triplos extraídos a partir da análise de termos compostos (TC) dos triplos extraídos através da identificação de padrões textuais.

Relação	**Extraídos A**		**Extraídos B**		**Exemplos**
	Total	S/rep	Total	S/rep	
Hiperonímia TC	711.954	390.492	24.367	16.228	(*desordem*,*desordem_cerebral*) (*átomo*,*átomo_de_carbono*)
Hiperonímia	149.845	144.839	31.254	29.563	(*desporto*,*automobilismo*) (*estilo_de_música*,*folk*)
Sinonímia	25.816	25.518	11.872	11.862	(*inglês_antigo*,*anglo-saxão*) (*estupro*,*violação*)
Parte	12.093	11.485	1.321	1.287	(*jejuno*,*instestino*) (*rolas*,*columbidae*)
Finalidade	13.277	12.992	777	743	(*amoxicilina*,*tratamento_de_infecções*) (*construção*, *terracota*)
Causador	5.854	5.740	559	520	(*parasita*, *doença*) (*doença_neuromuscular*,*fadiga*)

Tabela 1. Resultados totais da extracção em triplos.

Verifica-se que, de ambos os conjuntos, foi extraído um grande número de relações de hiperonímia através da análise de padrões textuais. Isto explica-se porque muitos resumos começam com a construção `[X é um Y]`, resultando em X HIPERONIMO_DE Y. Além disso, há frases com

uma enumeração no lugar de X, o que dá imediatamente origem a uma relação de hiperonímia por cada termo enumerado. Por exemplo, a frase *A heroína ou diacetilmorfina é uma droga* dá origem a: *droga* HIPERONIMO_DE *heroína*, *droga* HIPERONIMO_DE *diacetilmorfina*, *heroína* SINONIMO_DE *diacetilmorfina* e *diacetilmorfina* SINONIMO_DE *heroína*.

Outras curiosidades estão relacionadas com o âmbito dos triplos extraídos. As relações de hiperonímia atribuem essencialmente um género, espécie ou ordem a plantas, animais ou outros seres vivos. As relações de finalidade associam normalmente problemas de saúde às suas terapêuticas, e as relações de causa também se estabelecem muitas vezes entre problemas de saúde, suas causas e efeitos. Já as relações de sinonímia são por vezes estabelecidas entre termos na variante europeia e na variante brasileira do português, como por exemplo em *marrom* SINONIMO_DE *castanho* ou *esófago* SINONIMO_DE *esôfago*. Além disso, muitas das frases de onde são extraídas relações de sinonímia são iniciadas pela enumeração de uma grande quantidade de sinónimos. O caso extremo desta situação é a frase iniciada por: *Bagre-bandeira, bagre-cacumo, bagre-de-penacho, bagre-do-mar, bagre-fita, bagre-mandim, bagre-sari, bandeira, bandeirado, bandim, pirá-bandeira, sarassará, sargento ou bagre-bandeirado ... é um peixe da família dos ariídeos....*

3.3 Avaliação manual

A primeira abordagem à avaliação dos nossos resultados foi feita manualmente, através da classificação de um grupo de triplos seleccionado aleatoriamente de acordo com a escala proposta em [10], que sugere a classificação de triplos em quatro grupos: correctos (3); com uma preposição ou um adjectivo que deixam um dos argumentos estranho e impede o triplo de estar correcto (2); correcto, mas demasiado geral ou específico para ter utilidade (1); incorrecto (0).

Assim, foram inicialmente geradas 12 amostras aleatórias com 85 triplos extraídos a partir do conjunto A, classificadas cada uma por dois revisores. Para confirmarem a qualidade dos triplos, os revisores foram aconselhados a procurar na Web, incluindo a própria Wikipédia, por informação acerca das entidades envolvidas. A utilização desta escala permitiu por um lado identificar triplos que, devido a algum problema com as regras das gramáticas, deu origem a argumentos incompletos, e por outro identificar triplos que apesar de estarem correctos, não têm grande utilidade prática, principalmente no âmbito de uma ontologia lexical. Nesta categoria, encontram-se triplos que indicam subdivisões geográficas (p.e. *sub-região_estatística_portuguesa* PARTE_DE *região_do_alentejo*), relacionados com épocas históricas (p.e. *tragédia_de_1892* CAUSADOR_DE *crise_política*), entre outros demasiado específicos (p.e. *romancista_brasileiro* PARTE_DE *academia_brasileira_de_letras*, *escola* HIPERONIMO_DE *escola_de_música_Juilliard*).

Além disso, utilizamos a especificação de relações utilizada no PAPEL [12] para tratar o nome de cada triplo de acordo com as categorias gramaticais dos seus argumentos. No entanto, verificamos que, essencialmente devido a limitações

do *pos-tagger*, mas também devido ao género de texto processado, a grande maioria dos triplos cujo nome era alterado devido à categoria de um, ou ambos, os argumentos não ser substantivo, estava incorrecto. Optamos então por prosseguir a avaliação utilizando apenas relações cujos argumentos eram identificados como substantivos.

Tendo isto em conta, foram gerados novos dados para teste com triplos do conjunto B. Para tal, utilizamos 663 triplos que já tinham sido classificados na primeira avaliação e se mantinham no conjunto B, aos quais juntamos mais 12 amostras aleatórias, com cerca de 90 triplos cada uma, avaliadas da mesma forma que as primeiras.

Os resultados da segunda avaliação encontram-se na tabela 2, onde as proporções apresentadas somam as avaliações dos dois revisores, a que juntamos a concordância exacta entre ambos (CcEx) e a concordância relaxada (CcRel), em que os valores 1 e 3 foram considerados correctos e 0 e 2 incorrectos, atendendo a que estes resultados poderiam vir a ser utilizada noutro âmbito e os triplos classificados com 1 também estão correctos.

Um dado saliente ao comparar os resultados obtidos com o conjunto A com os obtidos com o conjunto B é a diferença do número de triplos classificados com 1. Em termos de proporção, este número decresceu de 39% para 22% do total de triplos. Também em proporção, houve um aumento dos triplos classificados com 3, o que se verifica principalmente nos triplos de finalidade e causa, aproximadamente 2 e 1,5 vezes mais. As melhorias dever-se-ão ao conjunto B ser mais restrito, com uma construção mais próxima e onde existirá menor ambiguidade. Ainda assim, cerca de um quarto dos triplos de causa e finalidade e um quinto dos triplos de parte continua completamente errado, o que estará essencialmente relacionado com a ambiguidade de alguns padrões utilizados.

Na tabela 2 verifica-se ainda uma maior concordância na divisão entre triplos correctos e incorrectos, essencialmente por se tratar de uma divisão mais objectiva, onde não entra a subjectividade de avaliar a utilidade efectiva de um triplo numa ontologia lexical. Por exemplo, vários triplos de hiperonímia extraídos através de termos compostos não acrescentam muito à base de conhecimento (p.e. *equipa* HIPERONIMO_DE *equipa_de_seis_jogadores*), mas esta classificação é bastante sensível ao critério do revisor.

Há no entanto um ponto em que esta avaliação piorou, mais propriamente na proporção de triplos de hiperonímia classificados com 2. Isto acontece porque a proporção de frases sobre espécies aumentou e muitas destas espécies são identificadas por duas palavras. Por exemplo, na frase *O Iriatherina werneri é uma espécie de peixe de aquário*, o *pos-tagger* não conhece as duas palavras da entidade *Iriatherina werneri*, o que leva o sistema a não interpretar a entidade como um substantivo modificado e, por isso, a extrair um triplo com um argumento incompleto, *peixe_de_aquário* HIPERONIMO_DE *werneri*.

3.4 Eficiência dos padrões

Além de avaliar a qualidade dos triplos extraídos, também nos pareceu interessante fazer o levantamento dos padrões ou palavras chave que davam

Relação	Avaliados	3(%)	2(%)	1(%)	0(%)	CcEx(%)	CcRel(%)
Hiperonímia TC	323	35,0	4,2	42,1	18,7	57,3	82,7
Hiperonímia	322	57,5	33,8	1,6	7,1	89,8	93,1
Sinonímia	286	85,7	7,3	0,4	6,6	90,0	91,6
Parte	268	44,2	26,7	8,4	20,7	63,1	78,4
Finalidade	264	53,0	16,5	4,0	26,5	71,2	82,2
Causador	244	41,8	24,6	7,8	25,8	61,5	79,5

Tabela 2. Resultados da avaliação manual de triplos.

origem a mais triplos. A esses dados, que para o conjunto B se encontram na tabela 3, juntamos informação acerca da classificação obtida na avaliação manual por triplos extraídos através destes padrões. Neste caso, apenas considerámos triplos onde a avaliação de ambos os revisores era concordante. Dentro dos padrões que levam à extracção de mais triplos incorrectos, destacamos `[usado|utilizado]` que, quando seguido de `[em|no|na]` pode não indicar a relação de finalidade, mas sim um local onde um objecto é utilizado, como em *O Ariary malgaxe é a moeda usada em Madagáscar*. Outro padrão bastante ambíguo parece ser `[inclui|incluem]`. Por outro lado, a utilização do padrão `é um género de` apenas levou à extracção de triplos de hiperonímia correctos.

Relação	Padrão	Extraídos	Avaliados			
			3	2	1	0
Hiperonímia	*termo composto*	24.367	72	7	75	32
Hiperonímia	`é uma espécie de`	15.824	54	96	0	0
Hiperonímia	`é um|uma`	11.865	94	11	1	17
Hiperonímia	`é um género de`	2.402	24	0	0	0
Sinonímia	`ou`	4.886	154	2	0	2
Sinonímia	`também conhecido|a|os|as por|como`	3.016	60	4	0	4
Parte	`inclui|incluem`	471	34	0	2	15
Parte	`grupo de`	158	17	3	1	0
Finalidade	`utilizado|a|os|as para|como|em|no|na`	376	71	16	1	20
Finalidade	`usado|a|os|as para|como|em|no|na`	237	41	3	1	4
Causador	`causado|a|os|as`	165	27	11	1	10

Tabela 3. Triplos extraídos e sua qualidade de acordo com o padrão utilizado.

3.5 Proposta para validação automática

Como é sabido, ainda que seja provavelmente a forma mais confiável de avaliação, a avaliação manual de relações semânticas é um trabalho moroso e cansativo, além de ser muitas vezes subjectivo por mais critérios que sejam definidos. Isto confirma-se pelas taxas de concordância que obtivemos na nossa avaliação manual. Ainda que tenhamos utilizado duas formas para medir a concordância,

nem sempre é fácil distinguir entre as várias classificações de uma escala. Por exemplo, além da subjectividade existente ao decidir a utilidade de um triplo, a distinção entre a classificação 1 e 2 pode não ser muito clara, já que o triplo pode ser muito geral, ou específico, exactamente por lhe faltar um modificador. Além disso, este tipo de avaliação não é facilmente repetível, o que não se passaria se existisse um método automático para avaliar a qualidade dos resultados. Com isto em mente, surgiu a nossa primeira abordagem a uma avaliação automática.

Uma das formas que vem sendo comum para validar, de forma automática, dados resultantes da extracção de informação passa por tirar partido da enorme quantidade de informação disponível na Web. No caso específico da validação de triplos semânticos, uma alternativa seria procurar por frases em que a relação entre ambos os argumentos está explícita através de padrões textuais. Isto é feito por exemplo em [12], mas sobre um corpo de notícias.

Seguindo estas ideias, a validação automática dos triplos extraídos no âmbito deste trabalho teria por base a aplicação de quatro métricas vulgarmente utilizadas para avaliar, na Web, a semelhança entre dois termos [3], mais precisamente: WebJaccard (1), WebOverlap (2), WebPMI (4) e WebDice (3). Nestas equações, $P(X)$ refere-se ao número de páginas em que o termo X ocorre e $P(X \cap Y)$ é o número de páginas em que X e Y co-ocorrem. Na equação 4, N deveria ser o total de páginas indexadas no motor de pesquisa que, não sendo calculável, poderá ser aproximado a 10^{10} [3].

$$WebJaccard(X,Y) = \frac{P(X \cap Y)}{P(X) + P(Y) - P(X,Y)} \tag{1}$$

$$WebOverlap(X,Y) = \frac{P(X \cap Y)}{\min\left(P(X), P(Y)\right)} \tag{2}$$

$$WebDice(X,Y) = \frac{2 * P(X \cap Y)}{P(X) + P(Y)} \tag{3}$$

$$WebPMI(X,Y) = \log_2 \left(\frac{P(X \cap Y)}{P(X) * P(Y)} * N \right) \tag{4}$$

As medidas acima referidas são normalmente utilizadas no cálculo da semelhança distribucional entre dois termos, ou seja, a semelhança dos termos com base nas suas ocorrências e vizinhanças, e, ainda que termos relacionados tenham habitualmente distribuições semelhantes, estas métricas não têm nenhuma relação semântica específica em vista. Sendo assim, inspirados por [16], para aplicarmos estas métricas à validação de triplos semânticos, deverá ser incluído também um padrão textual frequente indicador da relação, ou seja $X = XR$, $Y = RY$ e $X \cap Y = XRY$, sendo R o padrão. A tabela 4 contém padrões que podem ser utilizados para cada relação, depois de observar aqueles que mais frequentemente extraíram triplos (tabela 3). Curiosamente os padrões que extraem mais triplos indicam a relação inversa, ou seja, por exemplo, para validar o triplo t_1 $RELACAO$ t_2, $X = t_2$ e $Y = t_1$.

O primeiro passo foi calcular estas métricas para cada triplo avaliado manualmente em que a classificação fosse concordante para ambos os revisores. Para cada triplo e padrão relativo à sua relação (ver versão simplificada na

tabela 4), calculamos as métricas com base no Google. Logo aí verificamos que obtínhamos valores apenas para uma pequena quantidade de triplos (20% dos concordantes), porque os restantes nunca co-ocorriam com o padrão escolhido. Isto é compreensível, tendo em conta que termos semanticamente relacionados podem co-ocorrer de várias formas ou, por outras palavras, cada relação semântica pode ser traduzida numa enorme quantidade de padrões textuais. Outras limitações estão relacionadas com a própria pesquisa do Google, que não é suficientemente versátil para englobar um grande número de expressões. Além disso, ao procurar por um termo flexionado, o Google não consegue procurar por termos com o mesmo lema, o que limita as pesquisas deste tipo.

Ainda assim, passamos ao passo seguinte onde pretendíamos verificar se existia uma correlação entre os valores obtidos com as métricas para cada tipo de relação e a avaliação humana. Contudo, devido aos factores já referidos, a que acrescentamos a pouca quantidade de triplos disponíveis para esse cálculo, obtivemos sempre valores de correlação baixa, que nunca ultrapassavam os 20%, mesmo transformando a escala da avaliação manual numa escala apenas com 0s e 1s (semelhante à considerada para o cálculo da concordância relaxada).

No futuro pretendemos continuar a nossa busca por um método de validação automática para este trabalho e queremos ainda experimentar estas métricas em corpos para os quais exista um interface de pesquisa mais versátil, como o serviço AC/DC [8].

Relação	**Padrão indicador (R)**
Hiperonímia	`é|são um|uma`
Sinonímia	`também conhecido|conhecida|chamado|chamada|designado|designada de|por|pela`
Parte-de	`tem|possui|engloba|abrange|inclui|têm um|uma|vários|alguns|`
Causa	`devido|derivado|derivada|causado|causada|resultado|efeito|consequência a|ao|à|por|pelo|pela| de|do|da`
Finalidade	`usado|usada|utilizado|utilizada|através|objectivo|finalidade|intuito|serve no|na|para|de o|a|um|uma`

Tabela 4. Triplos extraídos e sua qualidade de acordo com o padrão utilizado.

4 Discussão e trabalho futuro

Apresentamos neste artigo o nosso sistema de extracção de relações semânticas a partir de texto não estruturado escrito em português e a sua aplicação a resumos da Wikipédia. O conhecimento extraído, já estruturado, pode ser de grande utilidade no aumento de recursos lexicais para a nossa língua. Nesse contexto, seria interessante realizar uma análise à quantidade de conhecimento extraído que ainda não se encontra no recurso em causa, uma pouco à imagem do que Hearst [13] fez para a WordNet.

Como se pode observar pelos resultados da avaliação, há ainda um longo caminho a percorrer e o sistema tem várias limitações, não só relacionadas com a ambiguidade e com a enorme possibilidade de formas para indicar uma

relação semântica, mas também relacionadas com o *pos-tagger* utilizado e o lematizador, que quando não reconhecem uma palavra procuram inferir a sua categoria gramatical com base em probabilidades e o seu lema com base em regras. Torna-se por isso, para já, impossível obter triplos cujos argumentos estejam lematizados, pois correríamos o risco de deteriorar a sua qualidade. Procuraremos ultrapassar esta limitação com a utilização de outro *pos-tagger* ou analisador morfológico.

Apesar de termos encontrado uma forma de filtrar quase todas as EM, através da taxonomia da DBpedia, haverá ainda várias entradas relevantes para o nosso recurso que ocorrem apenas na Wikipédia portuguesa e estão, desta forma, a ser filtradas sem necessidade. Por isso continuaremos em busca de uma filtragem mais adequada às nossas necessidades, e que poderá tirar partido de outra informação disponível na Wikipédia.

Além de questões já referidas ao longo da descrição da experimentação, e de experiências com métricas de semelhança distribucional, algo que também queremos realizar no futuro é definir um método para aferir a relevância de relações de hiperonímia obtidas através da análise de termos compostos. Por um lado, há uma pequena parte de triplos que podem ser obtidos desta forma e que não estão correctos (p.e. *bola de berlim* não é uma *bola* e *pé de atleta* não é um *pé*) e por outro, os triplos correctos nem sempre têm grande utilidade, tal como discutido na secção 2. Logo, este método terá em conta do numero de ocorrências e utilizações dos vários átomos do termo composto em colecções de documentos.

Numa fase posterior do trabalho pretendemos vir a integrar um conjunto de triplos extraídos da Wikipédia, também de forma automática, numa ontologia lexical ao estilo da WordNet mas para o português. À semelhança do que foi feito por [18], os termos serão associados, ou darão origem, a synsets, e os triplos passar-se-ão a estabelecer entre synsets. Este tipo de estruturas são uma forma aceitável de lidar com a ambiguidade e, além disso, permitirão a inferência de novas relações. Uma primeira abordagem a este problema, onde são utilizados recursos lexicais para o português, é descrita em [11].

Há ainda a acrescentar que, futuramente, pretendemos disponibilizar os resultados deste trabalho para toda a comunidade que trabalhe com o processamento computacional da língua portuguesa.

Referências

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: um treebank para o português. In: Gonçalves, A., Correia, C.N. (eds.) Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001). pp. 533–545. APL, Lisboa (2001)
2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia – a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (Setembro 2009)
3. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. In: Proc. 16th International conference on World Wide Web (WWW'07). pp. 757–766. ACM, New York, NY, USA (2007)

4. Cardoso, N.: REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In: Mota, C., Santos, D. (eds.) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas, pp. 195–211. Linguateca (2008)
5. Cederberg, S., Widdows, D.: Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: Proc. 7th Conference on Computational Natural Language Learning (CoNLL). pp. 111–118. Association for Computational Linguistics, Morristown, NJ, USA (2003)
6. Chodorow, M.S., Byrd, R.J., Heidorn, G.E.: Extracting semantic hierarchies from a large on-line dictionary. In: Proceedings of the 23rd annual meeting on Association for Computational Linguistics. pp. 299–304. Association for Computational Linguistics, Morristown, NJ, USA (1985)
7. Costa, H., Gonçalo Oliveira, H., Gomes, P.: The impact of distributional metrics in the quality of relational triples. In: Proc. ECAI Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010) (2010), no prelo
8. Costa, L., Santos, D., Rocha, P.A.: Estudando o português tal como é usado: o serviço AC/DC. In: The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009) (2009)
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
10. Freitas, M.C.: Elaboração automática de ontologias de domínio: discussão e resultados. Ph.D. thesis, Pontifícia Universidade Católica do Rio de Janeiro (2007)
11. Gonçalo Oliveira, H., Gomes, P.: Towards the automatic creation of a wordnet from a term-based lexical network. In: Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing (2010), no prelo
12. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. Linguamática 2(1), 77–93 (Maio 2010), nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009
13. Hearst, M.A.: Automated discovery of wordnet relations. In: [9], pp. 131–151 (1998)
14. Herbelot, A., Copestake, A.: Acquiring ontological relationships from wikipedia using RMRS. In: Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies (2006)
15. Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from wikipedia. Intl. Journal of Human-Computer Studies (Maio 2009)
16. Oliveira, P.C.: Probabilistic Reasoning in the Semantic Web using Markov Logic. Master's thesis, Universidade de Coimbra, Faculdade de Ciências e Tecnologia, Departmento de Engenharia Informática (2009)
17. Richardson, S.D., Dolan, W.B., Vanderwende, L.: Mindnet: Acquiring and structuring semantic information from text. In: Proc. 17th Intl. Conf. on Computational Linguistics (COLING). pp. 1098–1102 (1998)
18. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In: Proc. Advances in Web Intelligence 3rd Intl. Atlantic Web Intelligence Conference (AWIC). pp. 380–386. Springer (2005)
19. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. Data Knowledge Engineering 61(3), 484–499 (2007)
20. Wandmacher, T., Ovchinnikova, E., Krumnack, U., Dittmann, H.: Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In: Third Australasian Ontology Workshop (AOW 2007). CRPIT, vol. 85, pp. 61–69. ACS, Gold Coast, Australia (2007)
21. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proc. 6th Intl. Language Resources and Evaluation (LREC'08). Marrakech, Morocco (2008)