A expansão de conjuntos de co-hipónimos a partir de colecções de grandes dimensões de texto em Português

Luís Sarmento

Faculdade de Engenharia da Universidade do Porto
las@fe.up.pt

Resumo. Neste artigo iremos apresentar dois métodos para a expansão de conjuntos de co-hipónimos usando exclusivamente informação extraída a partir de uma colecção de texto em português de grandes dimensões. Os métodos baseiam-se na hipótese de que é possível explorar com sucesso a enorme redundância de informação existente em tais colecções recorrendo a algoritmos relativamente simples. Estes métodos operam de uma forma análoga ao conhecido sistema Google Sets, e num dos casos são alcançados tempos de execução muito reduzidos. Iremos enquandrar os dois métodos desenvolvidos numa estratégia mais ampla de construção de recursos léxico-semânticos para a língua portuguêsa e iremos posicioná-los relativamente a trabalhos realizados para outras línguas. Serão apresentados detalhadamente os algoritmos desenvolvidos, e para cada um deles serão apresentados e discutidos os resultados experimentais, comparando as suas limitações e vantagens. Abordaremos em seguida algumas questões relativas à avaliação deste género de métodos e destacaremos a necessidade de desenvolver recursos para esse efeito. Serão em seguida discutidas algumas limitações que derivam da indeterminação associada co-hiponímia e alguns dos problemas intrínsecos às abordagens que apresentamos. Terminaremos apresentando possibilidades de trabalho futuro.

1 Introdução

Em linguística é comum distinguir dois tipos de relações entre as palavras: relações paradigmáticas e relações sintagmáticas. Segundo [6] as relações paradigmáticas estabelecem-se entre palavras que partilham um determinado paradigma, seja este semântico, gramatical, morfológico ou outro, possuindo assim um série de características em comum que as permitem agrupar num determinado conjunto. Por exemplo, poderemos dizer que os pares (vermelho, azul), (cor, vermelho), (bom, mau), (ouvir, falar) agrupam palavras relacionadas por um derminado tipo de relação paradigmática. Por outro lado, as relações sintagmáticas são aquelas que se estabelecem entre palavras que co-ocorrem ao nível de um sintagma, podendo não existir qualquer tipo de relação paradigmática entre elas. Por exemplo "sentar" – "cadeira" ou "viagem – avião" estão relacionados sintagmáticamente, visto que co-ocorrem normalmente no interior de um determinado sintagma (ex: "sentar numa cadeira" ou "viagem de avião"). Note-se que em ambos os tipos de relações existe

obviamente uma forte ligação semântica entre as palavras, sendo que por isso nem sempre é simples destinguir entre os dois tipos de relações.

Considerando apenas as relações paradigmáticas do tipo semânticas, podemos enumerar as relações de sinonímia, de antonímia (ou mais genericamente de contraste), de hiponímia e de meronímia. Estas são relações habitualmente empregues na construção de rede léxicais do tipo WordNet, pelo que, dada a inexistência de um recurso deste género para o Português, e sendos tais recursos extremamente difíceis e custosos de produzir manualmente, torna-se particularmente interessante estudar métodos (semi-)automáticos para a identificação e compilação deste tipo de relações.

De todas as relações paradigmáticas semânticas anteriormente enumeradas, a relação de hiponímia reveste-se de particular interesse pois é a base da organização geral de um recurso como o WordNet. A hiponímia é a relação estabelecida entre um elemento e a classe mais geral onde esse elemento se inclui. Por exemplo "vermelho" é hipónimo de "cor", ou "cão" é hipónimo de "mamífero". A hiponímia possui uma uma relação inversa denominada hiperonímia, pelo que tudo o que for enunciado para a hiponímia aplica-se em sentido inverso à hiperonímia. Podemos então afirmar que "cor" é hiperónimo de "vermelho" e "mamífero" é hiperónimo de "cão". A relação de hiponímia (e logo a de hiperonímia) apresenta a propriedade transitiva, isto é, se "cão" é hipónimo de "mamifero" e "mamífero" é hipónimo de "animal", então "cão" é hipónimo de "animal".

Da relação de hiperonímia deriva-se a relação de co-hiponímia, ou seja a relação que se estabelece entre as palavras que possuem um hiperónimo comum. Assim, o conjunto (vermelho, azul, amarelo) está ligado por co-hiponímia, já que todos os seus elementos possuem (pelo menos) um hiperónimo comum ("cores primárias"). De uma forma aproximada, pode-se afirmar que a co-hiponímia une elementos "semelhantes" segundo um determinado critério, por vezes implícito ou até desconhecido à partida.

No resto do artigo, começaremos por justificar o interesse no desenvolvimento de mecanismos automáticos para a pesquisa de co-hipónimos, nomeadamente no seu interesse para a posterior pesquisa de relações de hiperonímia/hiponímia e também para acelerar processos de enriquecimento semi-automático de recursos léxico-semânticos. De seguida, apresentaremos alguns trabalhos relacionados, com os quais tentaremos comparar a nossa aproximação. Serão em seguida apresentadas 2 técnicas que permitem o alargamento de conjuntos de co-hipónimos, partindo de um conjunto reduzido de elementos semente, e analisados os seus resultados.

2 Motivação

A pesquisa e o alargamento automático de conjuntos de co-hipónimos possui um grande interesse prático, nomeadamente como forma de auxiliar a construção de recursos léxico-semânticos tais como por exemplo o WordNet. A obtenção automática de co-hipónimos poderá acelerar bastante o processo de adição de novas palavras, ou novas ligações ao recurso, já que permitirá, a partir de um conjunto de palavras conhecidas e já incluídas no recurso (ex: frutos como "morango", "banana", "maçã"), obter novos candidatos que partilhem directa ou transitivamente as mesmas relações de hiponímia (ex: "laranja", "cereja", "pêssego", "ananás", "melão", etc.).

Apesar de esta aproximação exigir sempre a validação manual, todo o processo de compilação de novos elementos e de descoberta de relações beneficia de uma aceleração considerável.

Em segundo lugar, e tal como demonstrado em [7] a obtenção de grupos de co-hipónimos potencia a obtenção dos respectivos hiperónimos (podem ser obtidos vários hiperónimos relacionados). É sabido que a determinação de relações de hiperonímia em corpora pode ser feita com algum sucesso para elementos frequentes usando certos padrões léxico-sintácticos tais como por exemplo "(um|uma) [HIPERÓNIMO] como (alo) [HIPÓNIMO]" [4]. Contudo, para elementos (neste caso os hipónimos) mais raros a probabilidade de ocorrência dos referidos padrões reduz-se substancialmente, sendo por isso muito dificil encontrar evidência significativa da relação hiperonímia-hiponímia em causa. A obtenção alternativa de conjuntos de co-hipónimos permite aliviar este problema, quer por propagação automática das relações de hiperonímia já conhecidas ao restantes elementos do conjuntos de co-hipónimos, quer por fusão das evidências de hiperonímia recolhidas individualmente para cada um dos elementos do conjunto de co-hipónimos.

3 Trabalho Relacionado

Este trabalho foi inicialmente inspirado num sistema experimental fornecido pelo motor de pesquisa Google denominado Google Sets e que se encontra disponível em http://labs.google.com/sets. Este sistema recebe como parâmetros de entrada um conjunto de elementos fornecidos pelo utilizador e tenta expandir esse conjunto até a um máximo de 15 elementos "semelhantes". Apesar de não existir muita informação disponível acerca do modo de funcionamento concreto do Google Sets - nem uma avaliação do seu desempenho - este parece apoiar-se essencialmente na grande quantidade de dados que o Google tem nas suas bases de dados, sobre os quais aparenta utilizar alguma técnica de agrupamento ou de processamento "data-driven" para a obtenção dos elementos semelhantes. O sistema parece ter sido preparado apenas para lidar com inglês, não respondendo actualmente a pedidos de expansão de conjuntos em português. Por inspecção visual, o funcionamento deste sistema parece ser muito positivo, embora com a limitação do tamanho máximo do conjunto resultado a 15 elementos, o que talvez possa ser considerado uma indicação indirecta dos limites de precisão da aproximação adoptada.

Em [1] é descrita uma técnica que através de agrupamentos sucessivos das co-occorências entre palavras se consegue obter evidência de relações sintagmáticas e paradigmáticas em alemão. Os autores mostram como, através de agrupamentos sucessivos das co-ocorrências calculadas sobre um corpus de mais de 100 milhões de palavras, se consegue separar as relações sintagmáticas das paradigmáticas, e como numa segunda se torna fase possível discriminar as relações de hiperonímia e de co-hiponímia. A avaliação dos resultados foi realizada comparando os resultados obtidos com informação acerca das relações presente no GermaNet e com pares de palavras relacionadas compilados manualmente.

Em [8] foi calculada a matriz de co-ocorrências com uma janela de comprimento 2 para 1 milhão de palavras do British National Corpus, removendo palavras-função e

ignorando a anotação morfo-sintáctica. Em seguida usou-se esta informação para, dada uma palavra "semente" e aplicando a medida de distância vectorial "city-block", obter o conjunto de palavras semelhantes. Desta forma, o autor foi capaz de obter palavras relacionada paradigmaticamente, já que a pesquisa efectuada permite encontrar palavras que possuam perfís de co-ocorrência semelhantes, o que segue indirectamente a definição de relação paradigmática. Os resultados obtidos foram comparando com o desempenho humano na tarefa de selecção de sinónimos do TOEFL, tendo sido reportados resultados equiparáveis. Na nossa opinião a avaliação realizada pelo autor é pouco rigorosa não permitindo tirar conclusões acerca do desempenho efectivo deste método. Para além disso, esta aproximação não permite distinguir qual o tipo de relação semântica efectivamente obtida entre as palavras.

Um aproximação mais sofisiticada encontra-se descrita em [5] onde através da utilização de um *parser* se construem tripletos de relações gramaticais da forma (palavra1, relação, palavra2). Foram processados vários corpora de texto jornalístico em língua inglesa totalizando cerca de 100 milhões de palavras dos quais foi possível gerar 56.5 milhões de tripletos. Usando uma medida derivada da Informação Mútua sobre a lista de tripletos, foi computada a semelhança entre todos os pares de nomes, verbos e adjectivos/adverbos, para obter um *tesaurus* composto por vários conjuntos de semelhantes (ou em alguns casos sinómimos). Os resultados obtidos foram comparados com a informação contida no WordNet e o *tesaurus* Roget. Os autores concluiram que esta aproximação permite construir *tesaurus* em larga escala, embora não permita diferenciar entre os diferentes sentidos das palavras.

Outro trabalho relevante, embora com uma orientação ligeiramente diferente do nosso objectivo, vem descrito em [10] onde se descreve um sistema que recorre à pesquisa na Web através do motor de pesquisa Altavista para obter sinónimos. O sistema utiliza uma bateria de pesquisas na Web e pondera os resultados obtidos usando a medida PMI-IR (adaptada da medida Pontwise Mutual Information) para obter um medida de "sinonímia" entre palavras. Esta aproximação é extremamente interessante já que compensa os problemas da escassez de dados que normalmente afectam a performance de certos métodos quando lidam com palavras raras, utilizando a imensidão da Web. O autor reporta que os resultados deste sistema na tarefa de identificação de sinónimos do TOEFL foram superiores à média obtidos por humanos.

4 Descoberta de Co-hipónimos Usando o BACO

O objectivo do nosso trabalho consiste na exploração de métodos para expansão de um conjunto de co-hipónimos, que se baseiem apenas em informação inferida a partir de uma colecção de texto de grandes dimensões. Por outras palavras, pretendemos que, a partir de um conjunto de elementos fornecidos como entrada, o sistema seja capaz de encontrar numa colecção de documentos de texto outros elementos que lhes sejam semelhantes. Por exemplo, para um conjunto de entrada como (vermelho, verde, azul) espera-se ser capaz de encontrar várias outras palavras que se refiram a nomes de cores. Com este objectivo experimentamos dois métodos diferentes para expandirmos listas de co-hipónimos, cada um dos quais explorando uma propriedade

heurística distinta associada à co-hiponímia. Iremos em seguida descrever cada um dos métodos.

4.1 Pesquisa Usando Contextos Léxicais de 3 Palavras

O primeiro método consiste na observação de que os co-hipónimos deverão ocorrer em contextos léxicais muito semelhantes, já que pela própria definição de co-hiponímia possuem um hipónimo comum do qual herdam a maioria das propriedades e por isso também as ligações léxicais que podem estabelecer com outros elementos. Por exemplo, se pesquisarmos num corpus de grandes dimensões contextos onde ocorre a palavra "morango" poderemos quase certamente encontrar algo como "[com compota de] morango" ou "[um batido de] morango". Inversamente, se no mesmo corpus pesquisarmos quais as palavras que ocorrem nos contextos léxicais "[com compota de] X", ou "[um batido de] X" parece natural que, se estes contextos léxicais ocorrerem, a palavra X terá grande probabilidade de se referir a um fruto, ou eventualmente a outro alimento.

Seguindo esta ideia decidimos implementar uma pesquisa sobre a base de texto BACO [9] que permite efectuar pesquisas rápidas sobre contextos léxicais. O BACO (BAse de Co-Ocorrência) é uma base de dados gerada a partir de colecção web WPT03 (ver http://poloxldb.linguateca.pt/) e possui informação relativa a mil milhões de palavras provenientes de textos da web portuguesa. Esta informação encontra-se armazenada em várias tabelas diferentes, cada uma optimizada para um determinado tipo de pesquisa. Para além de permitir a pesquisa ao nível da frase, o BACO possui também tabelas que armazenam sequências de 2, 3 e 4 palavras, vulgarmente conhecidas com n-gramas. As tabelas de n-gramas são particularmente apropriadas para a pesquisa dos contextos léxicais onde ocorre uma dada palavra (neste caso contextos até 3 palavras), ou, inversamente, para a pesquisa de palavras que ocorrem num determinado contexto.

No sentido de máximizar a informação de contexto, decidimos realizar as pesquisa sobre a tabela de 4-gramas, permitindo assim a pesquisa de contextos com comprimento de 3 palavras. A tabela de 4-gramas do BACO possui o seguinte esquema: wpt_4_gramas(p_1, p_2, p_3, p_4, f, d). A informação armazenada nesta tabela é uma sequência de quatro palavras (p_1, p_2, p_3 e p_4), o número de vezes que essa sequência foi encontrada (f), e o número de documentos (d) da colecção WPT03 nos quais a referida sequência ocorre. Durante toda a experimentação, e para acelerar o processo de testes, foi utilizado um sub-conjunto da tabela de 4-gramas que se refere apenas de 370 mil documentos, cerca de 25% do total dos documentos disponíveis no BACO e contendo 92.835.207 tuplos. A tabela 4-gramas completa possui cerca de 273 milhões de tuplos, mas não foi usada nestas experiências por razões de desempenho e de espaço em disco.

Muito suncintamente, o algoritmo para a expansão de hipónimos consiste numa primeira fase, em pesquisar a tabela de 4-gramas para encontrar os contextos léxicais nos quais ocorrem os exemplos de entradas, e, numa segunda fase, é realizada uma pesquisa complementar, sendo desta vez pesquisadas as palavras que ocorrem nos referidos contextos. O contexto léxical considerado nestas experiências é constituido pelas 3 palavras anteriores à palavra fornecida como semente. Este não o único

contexto possível havendo pelo menos mais três. Contúdo se entrarmos em consideração que a pesquisa de co-hipónimos se refere à pesquisa de nomes e que, tal como demonstrado nos exemplos anteriores, procuramos especialmente ocorrências em que o nome ocorre como modificador no interior de um sitagma nominal, então a escolha das três palavras anteriores parece adequar-se melhor à pesquisa dessas estruturas. Apesar disso, não temos dados concrectos que nos permitam afirmar isto peremptoriamente e a investigação desta questão será alvo de futuro trabalho.

Foram considerados relevantes apenas os contextos com os quais ocorrem um determinado número mínimo de elementos do conjunto inicial, para que os referidos contextos sejam suficientemente correlacionados com o conjunto exemplo. Por outro lado, um contexto léxical que co-ocorra com demasiadas palavras (para além do conjunto inicial) é considerado demasiado genérico pelo que também é invalidado. De todos os candidatos a co-hipónimos assim obtidos removem-se aqueles que são artigos, preposições, e outras palavras ou siglas muito frequentes e que não podem ser à partida considerados co-hipónimos válidos. De notar que são apenas pesquisados candidatos com um única palavra, o que é certamente uma limitação desta implementação. No entanto, se considerarmos que um recurso como o WordNet é constituido maioritariamente por palavras simples, podemos ainda assim considerar que se trata de uma ajuda para a construção de recursos semelhantes.

O algoritmos em pseudo-código será o seguinte:

0: Incialização
seja S ={s_1,s_2,... s_n} o conjunto de co-hipónimos semente (pelo menos 1)
seja L o número minimo de co-hipónimos que um dado contexto têm de cobrir
seja M o número máximo de palavras com que um contexto válido pode co-ocorrer
seja P o conjunto de palavras proíbidas (artigos, preposições, etc.)
1: para cada s_i elemento de S
pesquisar na tabela 4-gramas o contexto $c_k = (p_1,p_2,p_3)$ para os quais $p_4=s_i$
recolher c_k no conjunto de contextos C, incrementando a sua contagem
2: para cada ck do conjunto C cuja contagem seja igual superior a L
pesquisar na tabelas de 4-gramas lista de [p_4] para os quais $(p_1,p_2,p_3) = c_k$
se tamanho da lista [p_4] < M então contexto é válido
adicionar cada elementos da lista p_4 ao conjunto resultado R, incrementando a sua contagem
3: Retornar R, o conjunto de candidatos a co-hipónimo. Ordenar por contagem excluindo os elementos pertencentes aos conjuntos P e S.

Na seguinte tabela, apresentam-se os resultados da execução deste algoritmo para várias configurações de conjuntos “semente” iniciais. O valor L indica o número de elementos do conjunto inicial com os quais um determinado contexto léxical tem de co-ocorrer para ser considerado válido e o valor #C indica o número de contextos léxicais distintos encontrados que verificam tal condição. Os valores entre parêntesis indicam o número de contextos léxicais válidos que existem em comum entre o elemento proposto e os elementos do conjunto inicial (as reticências entre os candidatos apresentados indicam séries de candidatos considerados correctos mas omitidos por questões de brevidade). Os elementos em itálico indicam ocorrências erradas.

#	Conjunto inicial	L	#C	Resultado
1	amarelo, vermelho, azul	3	44	verde (26), branco (22), preto (19), cinza (14), castanho (14), ... violeta (6), prata (5), *escuro* (4), dourado (4), *fe* (4), ... *pele* (4), *cores* (3), ... *liso* (2), *carvalho* (2), marrom (2), ... *terra* (2), *iluminado* (2), *54* (2), *brasil* (2), *pobre* (2)
2	granito, mármore, basalto	3	3	*betão* (2), *vidro* (2), *papel* (2), *pedra* (2), *madeira* (2), *material* (2)
3	whiskey, rum, gin	2	6	vodka (4), vinho (3), tequila (3), porto (3), cerveja (2), licor (2), sumo (2), coca-cola (2), verdelho (2), whiskie (2), tinto (2), aguardente (2), conhaque (2), *jack* (2), *neoplast* (2), uisque (2), água (2), *coca* (2), *plástico* (2), champanhe (2), cachaça (2), champagne (2)
4	porto, braga, aveiro	3	210	coimbra (141), lisboa (137), *vila* (126), *castelo* (115), leiria (110), viseu (110),... almada (51), guimarães (49),... *cidade* (5) ... régua (2), *avaliação* (2), *recrutamento* (2), *municípios* (2), *editorial* (2), gorazde (2), *gás* (2), *coliseu* (2), alvor (2), inhambane (2)

Tabela 1. Resultados usando o método da pesquisa por contextos léxicais de 3 palavras

4.2 Breve Análise de Resultados

Numa breve análise a estes resultados, ressaltam as seguintes observações:

1. Há uma grande variância entre o número de contextos válidos encontrados para cada conjunto inicial e consequentemente entre o número de candidatos a co-hipónimos retornados. Nitidamente, para certos conjuntos os dados são mais esparsos. No caso particular do conjunto (graníto, mármore, basalto) foram encontrados apenas 3 contextos válidos em contraste com o conjunto (porto, braga, aveiro) onde foram encontrados 210 contextos válidos.
2. Na maior parte dos casos análisados, os elementos do topo da lista retornada são de facto alguns do co-hipónimos esperados. Contúdo, em certos casos verifica-se a presença de co-hipónimos mais distantes e que por isso não correspondem ao esperado, como foi o caso do conjunto 2 para o qual, como já mencionado, não foi aparentemente possível encontrar contextos léxicais válidos em número suficiente. Reduzindo o parâmetro L para 2, isto é reduzindo a exigência relativamente à especificidade do contextos léxicais válidos, o resultado torna-se extremamente ruidoso.
3. Em alguns casos é possível encontrar na lista de candidatos não só co-hipónimos, mas também os hiperónimos (“cores”, ”material”, “cidade”).
4. Dada a limitação deste método em apenas permitir a pesquisa de uma palavra, alguns candidatos apresentados são de facto apenas parte de um eventual candidato

multi-palavra. Por exemplo, nos conjunto 3 o candidato "jack" provavelmente se refere a "jack daniels", assim como "coca" se deverá referir a "coca cola".

5. Este método apresenta algum sensibilidade relativamente a certos problemas de ambiguidade muito típicos da língua portuguesa. Por exemplo, nos resultado da lista (whiskey, rum, gin) encontramos "plástico" o que parece um resultados perfeitamente estemporâneo. Após inspecção mais detalhadas verificamos que se a causa deste resultado advém de ter sido considerado válido o padrão "uma garrafa de X" que é instânciavel não só para bebidas – ex: "uma garrafa de tequila" - como também para materiais – ex: "uma garrafa de plástico".

Como observações mais gerais podemos ainda afirmar que este método é relativamente elegante do ponto de vista algorítmico pois a sua implementação decorre directamente da definição de co-hipónimos adoptada. Apesar disso, apresenta graves problemas de eficiência e não é facilmente escalável. Os exemplos que apresentamos demoraram entre 10 e 30 segundos para o caso de conjuntos iniciais contendo elementos pouco frequentes, como é o caso do conjunto 2 e 3 ou mais de 5 minutos para conjuntos com elementos muito frequentes, como o conjunto 4. Finalmente, o algoritmo não apresenta nenhum mecanismo de controlo automático relativamente ao parâmetro L, que tem de ser por enquanto controlado manualmente.

4.3 Pesquisa por Co-ocorrência de Co-hipónimos em Coordenação

O segundo método desenvolvido baseia-se na observação simples de que os co-hipónimos são muitas vezes referidos no contexto de uma coordenação. Por exemplo, "as minhas cores preferidas são o vermelho e o azul" ou "os materiais usados foram mármore, ferro e granito". Estas coordenações, que aparentam ser frequentes, sugerem ser possível extrair informação útil acerca da co-hiponímia. Foram por isso selecionados manualmente para uma janela de 4 palavras 12 padrões que estivessem tipicamente associados a coordenações. Na seguinte tabela apresentam-se os padrões selecionados, sendo que X e Y indica a posição dos supostos co-hipónimos.

Formula	P1 X P3 Y	X P2 Y P4	X P2 P3 Y
Padrões	, X e Y , X ou Y , X , Y	X , Y , X , Y e X, Y ou	X e o Y / X e a Y X ou o Y / X ou a Y X , o Y / X , a Y

Tabela 2. Padrões de 4 átomos associados à co-hiponímia.

Refira-se que esta lista não é de forma alguma exaustiva e a validação destes padrões não passou de algumas experiências realizadas sobre a base de dados. A pesquisa directa de tais padrões sobre a tabelas de 4-gramas revelou-se muito pouco eficiente já que muitos dos átomos associados aos padrões de coordenação (i.e a virgula, o "e", o "ou", etc) são muito frequentes, não permitindo por isso tirar grande vantagem dos índices de pesquisa associados às tabelas. Por esse motivo resolveu-se criar uma tabela auxiliar contendo apenas uma sub-selecção dos tuplos da tabela de 4-gramas que correspondem aos padrões apresentados em cima. A seguinte tabela

indica o número de tuplos que foi possível recolher para cada um dos referidos padrões:

Padrão	Tuplos recolhidos
, X e Y	179415
, X ou Y	25.203
, X , Y	399.013
X , Y ,	428.746
X , Y e	202.619
X, Y ou	28.941
X e o Y	112.746
X e a Y	153.477
X ou o Y	6.824
X ou a Y	13.083
X , o Y	207.068
X , a Y	271.152
Total	**2.028.287**

Tabela 3. Número de tuplos recolhidos para cada um dos padrões utilizados

Através desta tabela, a pesquisa de co-hipónimos de um derminado elemento torna-se na simples tarefa de pesquisar elementos nas posições X e Y com os quais o elemento base co-ocorra na posição complementar. Ao contrário do método anterior, este método não possui nenhum parâmetro de qualidade que permita filtrar elementos ruidosos durante o processo de pesquisa. Na verdade, o único critério que foi considerado para tentar julgar a qualidade dos candidatos é o número de padrões com os quais o candidato e os elementos do conjunto inicial co-ocorrem. Desta forma, associado a cada candidato é possível associar um valor que pode variar entre 1 e 24 por cada elemento exemplo fornecido. O valor 1 ocorre quando o candidato a co-hipónimo co-ocorre apenas uma vez com um padrão numa determinada posição (X ou Y), e o valor 24 ocorre quando o candidato co-ocorre nos 12 padrões em ambas as posições (X e Y).

Para permitir uma comparação entre os dois métodos, foram repetidas as experiências com os mesmo conjuntos iniciais. A próxima tabela apresenta resumidamente os resultados obtidos (as reticências entre os candidatos apresentados indicam séries de candidatos considerados correctos mas omitidos por questões de brevidade).

#	Conjunto inicial	Resultado
1	amarelo, vermelho, azul	verde (48), preto (39), branco (38), laranja (28), rosa (23), cinza (18), castanho (18), violeta (13), cinzento (11), negro (11), lilás (11), *cor* (11), … cores (6), … transparente (4), azulão (4), *champanhe* (4), *sol* (4), *céu* (3), castanha (3), mediterrâneo (3), alaranjado (3), *camisa* (3), *claro* (3), púrpura (3), âmbar (3)…
2	Granito, mármore, basalto	*madeira* (9), pedra (8), calcário (7), *bronze* (7), *cimento* (6), *vidro* (5), xisto (5), *cantaria* (4), *tijoleira* (4), ardósia (3), arenito (3), barro (3), gesso (3), calcários (3), travertino (3), *tabaco* (2), *ouro* (2), *cellano* (2), (2), quartzo (2),…

3	whiskey, rum, gin	Vodka (8), conhaque (3), tequila (3), rumpi (2), *tabaco* (2), *limão* (2), calvados (2), *creme* (2), bourbon (2), *curaçau* (2), *anseios* (2), *açúcar* (2), brandy (2),…
4	Porto, braga, aveiro	lisboa (31), coimbra (28), leiria (24), gaia (23), viseu (22), setúbal (22), évora (21), guimarães (21), guarda (19), *minho* (19),… *algarve* (16),… *madeira* (14), …*portugal* (13), … cidade (13), … *sporting* (12), *benfica* (12) … (várias centenas de candidatos)

Table 4. Resultados do método de pesquisa por contextos de coordenações

4.4 Breve Análise de Resultados

A primeira observação que resulta da aplicação deste método é a de que este gera muito mais candidatos que o anterior. De facto, para todos os conjuntos exemplo foi possível obter várias dezenas de candidatos.

Mais uma vez também, os candidatos do topo da lista podem ser considerados correctos, havendo no entanto uma situação na qual isto não se verfica claramente, nomeadamente para o conjunto (granito, mármore, basalto). Neste caso, e tal como já verificado na experiência com o método anterior, os resultados apresentam alguns candidatos diferentes daqueles que se esperaria (tipos de rocha), o que, como iremos discutir em seguintes secções, advém de algumas ambiguidades intrínsecas à relação de co-hiponímia, e das limitações inerentes aos métodos apresentados no tratamento de dados esparsos.

Verifica-se também que grande parte dos candidatos que se podem considerar errados, são de facto palavras relacionadas sintagmaticamente (isto é pertencentes contexto comum). Este problema não era tão notório no método anterior. Por outro lado, pelo que nos foi dado a observar nestas experiências, este método é muito mais resistente a certos problemas da ambiguidade como os os verificados na situação "garrafa de gin" vs. "garrafa de plástico". Esta robustez é perfeitamente compreensivel já que ocorrência de coordenações entre "gin" e "plástico" parece absolutamente improvável pois são conceitos que ocorrem em dimensões quase ortogonais. Apesar da sua simplicidade, este método apresenta a enorme vantagem de ser capaz de eliminar estas ambiguidades.

Um ponto positivo deste método é a sua velocidade de execução. Pelo facto de se ter reduzido o espaço de pesquisa de cerca de 92 milhões de tuplos para a cerca 2 milhões de tuplos, a pesquisa de candidatos é executada em menos de 5 a 10 segundos, para as mesmas condições de hardware. Este desempenho permite utilizar este método como auxílio em tempo-real na construção de recursos lexico-semânticos. As grandes desvantagens deste método estão relacionadas com a inexistência de um parâmetro eficiente para controlo da qualidade dos candidatos obtidos, e com a dificuldade de filtrar as relações sintagmáticas.

5 O Problema da Avaliação dos Resultados

Antes de avançar com a discussão dos resultados, há um ponto que é importante abordar e que se prende com a avaliação dos resultados obtidos por estes dois métodos. Até agora temos vindo a apresentar resultados mas em nenhum dos casos fornecemos medidas concretas desempenho (como a Precisão ou a Abragência) calculadas relativamente a um padrão. De facto, a avaliação rigorosa dos resultados obtidos é problemática para o nosso caso, facto pelo qual os comentários acerca das qualidades ou defeitos dos métodos apresentados se baseiam na simples inspecção visual dos resultados.
O problema da avaliação de métodos de obtenção de palavras semelhantes a partir de texto foi já estudado anteriormente [3] tendo sido sugeridas duas técnicas:

1. A utilização de um padrão semântico, tal como um *tesaurus* ou outro recurso léxico-semântico já existente, para poder verificar se os métodos são capazes de "descobrir" correctamente relações comparando-as com as que já são conhecidas e que estão já codificadas explicitamente.
2. A utilização de um dicionário de definições lingua corrente a partir do qual é possível inferir se as relações descobertas são válidas. A ideia é para que dois conceitos / palavras relacionadas as suas definições no dicionário deverão apresentar um elevado grau de sobreposição.

Infelizmente, para o caso da lingua portuguesa não conhecemos recursos - como *tesauros* ou dicionários de definições - publicamente disponíveis para o efeito. Assim sendo restam-nos duas alternatives: (i) executar a avaliação manual dos resultados obtidos ou (ii) construir padrões especializados através da compilação e organização de informação de várias fontes. A primeira opção é sempre viável, mas não pode ser considerada uma metodologia satisfatória a longo prazo pois a subjectividade e flutuação temporal de critérios impede a correcta avaliação da evolução dos métodos e a sua comparação. A construção de um recurso destinado a permitir a avaliação destes métodos, apesar de poder parecer um trabalho excessivo parece também ser a única forma de conseguir realizar avaliações consistentes a médio e longo prazo. Como tal, e depois do desenvolvimento destas experiências, foi tomada a decisão de planear um tal recurso que sirva de padrão para a avaliação de métodos de pesquisa de co-hipónimos e que deverá ser utilizado em futuros desenvolvimento desta linha trabalho.

6 Discussão

Apesar de não termos efectuado uma avaliação efectiva dos métodos apresentados é útil fazer algumas observações. Em primeiro lugar, ambos os métodos foram capazes de expandir os conjuntos iniciais com co-hipónimos razoavelmente relevantes, em particular para conjuntos contendo elementos bastante frequentes. Como se pode verificar, os topos das listas expandidas são na maior parte dos casos co-hipónimos que se podem considerar válidos. Contudo, o mesmo já não ser verifica para

conjuntos iniciais cujos elementos não ocorram tão frequentemente, como foi o caso do conjunto (granito, mármore, basalto). A título ilustrativo, e considerando toda a colecção WPT03, a palavra "basalto" ocorre apenas 460 vezes em 339 documentos, enquanto que a palavra "amarelo" ocorre 18.186 em 12.336 documentos e a palavra "aveiro" ocorre 147.851, em 65.601 documentos. Estes valores reflectem a natural dificuldade em se lidar com dados esparsos, o que sugere que melhores desempenho obrigarão à eventual utilização de técnicas de suavização de dados.

Em segundo lugar, convém referir que apesar destas diferenças nos valores das frequências das palavras semente, ambos métodos fazem uso de medidas de qualidade (ou de limiares de filtragem) baseados no número de co-ocorrências distintas entre os candidatos e os contextos léxicais / padrões pesquisados, dependendo por isso da frequência apenas indirectamente. A grande vantagem de basearmos as nossas medidas de qualidade no número de co-ocorrência distintas, e não na frequência de cada uma dessas co-ocorrência, é podermos evitar parte dos problemas associados à duplicação de documentos que existe naturalmente em colecções web. A existência de documentos duplicados enviesa todos os parâmetros derivados do valor da frequência, pelo que a utilização deste parâmetro pode ser problemática.

Um ponto que ficou por explorar prende-se com a utilização de medidas de qualidade alternativas, também baseadas no número de co-ocorrências, mas que fossem capazes de ponderar a importância relativa de uma determinada co-ocorrência. De entre estas medidas, destaca-se a Informação Mútua [2] cuja aplicação poderia permitir a identificação dos contextos léxicais mais discriminantes, reduzindo o impacto dos contextos muito genericos que por co-ocorrerem com um elevado número de palavras também geram muitos candidatos inválidos.

Outra questão que não foi abordada, relaciona-se com o tamanho do contexto léxical utilizado neste métodos. Por razões de viabilidade, todas as pesquisas centram-se num contexto léxical de 3 palavras. Os resultados obtidos sugerem que este contexto parece ser suficientemente informativo para a tarefa a que nos proposemos, mas não foi possível recolher qualquer evidência de qual seria a evolução do desempenho dos métodos com o aumento (para 4 ou 5 palavras) ou com a redução (para 2 ou mesmo 1 palavra) do contexto léxical. E mesmo considerando manter o tamanho do contexto léxical em 3 palavras no caso do primeiro método, ficou também por explorar o impacto da escolha da posição da janela em torno dos candidatos: foram consideradas as 3 palavras anteriores ao candidato, mas o que aconteceria se fossem as 3 palavras posteriores? Ou se se considerassem as duas palavras anteriores e uma posterior?

Contúdo, a exploração sistemática destas questões teóricas exige, como já referido anteriormente, o desenvolvimento de técnicas e recursos objectivos para a avaliação de desempenho, pelo que essa deverá ser uma prioridade futura.

7 Indeterminação Instrínseca à Co-hiponímia

Em alguns dos resultados obtidos, ambos os métodos apresentaram resultados que fogem um pouco ao que seria de esperar: parte dos candidatos retornados não tinham aparentemente o nível de especialização compatível com o conjunto semente,

parecendo assim um pouco deslocados dos restantes co-hipónimos. Uma desta situações foi particularmente visível para o conjunto (granito, mármore, basalto) em que se verificou a presença de candidatos que nada tinham a ver com rochas, tal como intuitivamente se esperaria, nomeadamente "madeira", "vidro" ou "betão". Estes resultados estão relacionados com uma característica intrinseca da noção de co-hiponímia que é a multiplicidade de hiperónimos comuns que podem unir os referidos co-hipónimos. Neste caso, apesar de implicitamente termos assumido que o hiperónimo natural fosse "rocha", não é menos verdade que a partir dos mesmo exemplos também seria possível inferir como um hiperónimo válido "materiais de construção". Poderiamos eventualmente arranjar ainda outros hiperónimos comuns mais ou menos especializados, ou com um maior ou menor grau de sobreposição com os hipónimos apresentados, e que serviriam igualmente para agrupar estes 3 co-hipónimos. Note-se que os hiperónimos alternativos não estão relacionados entre sí apenas por especialização, já que a sua relação pode ser também de sobreposição parcial.

Estas características de indeterminação associadas à co-hiponímia dificultam a formulação do problema da pesquisa de co-hipónimos e, de certa forma, alastram o problema também à pesquisa dos vários hiperónimos possíveis. Evidentemente que o problema assim colocado torna-se muito difícil de resolver pelo que se torna necessário encontrar meios alternativos para conseguir refinar as noções de proximidade entre os co-hipónimos, ou para tentar identificar os contextos semânticos mais restritos em que eles se podem conjugar (ex: "natureza", "construção", "escultura", "cantaria", etc.). Como trabalho futuro, pretende-se vir a estudar formas de separar os vários contextos semânticos associados a um grupo de palavras, mesmo que não seja possível determinar com exactidão o que cada um desses contexto semânticos pode ser. Admite-se que sendo possível discriminar / separar contextos semânticos alternativos (quaisquer que estes sejam), será também possível melhorar as técnicas de pesquisa de co-hipónimos pela selecção apenas dos contextos semânticos interessantes, algo que os métodos apresentados são incapazes de fazer correctamente.

Conclusões

Neste artigo apresentamos dois métodos alternativos para a expansão de conjuntos de co-hipónimos. Verificamos qu e, com técnicas relativamente simples e fazendo uso das quantidades massivas de texto que agora estão disponíveis, é de facto possível obter resultados satisfatórios na pesquisa de elementos semelhantes aos fornecidos como exemplo ao sistema. Os dois métodos apresentam diferentes níveis de robustez relativamente a ambiguidades típicas do português e à possibilidade de ruído proveniente de palavras relacionadas sintagmaticamente com os elementos do conjunto inicial. Foram apontadas as actuais limitações associadas às possibilidades reais de avaliação deste género de sistemas e concluiu-se que é necessário desenvolver urgentemente recursos para esse fim. Foram também levantadas algumas questões relacionadas com a necessidade de utilização de técnicas de suavização de dados que poderão melhorar o desempenho dos métodos de pesquisa semelhantes aos

apresentados em situações onde os dados são mais esparsos. Foi também apontado interesse em explorar medidas de qualidade alternativas, como por exemplo a Informação Mútua, no sentido de ajudar a descriminar ligações mais relevantes entre os exemplos e os seus contextos léxicais. Reflectiu-se também sobre a necessidade de desenvolver métodos capazes de discriminar contextos semânticos, para lidar mais eficientemente com os problemas de indeterminação associados à própria noção de co-hiponímia.

Como conclusão final podemos dizer que, mesmo considerando todas limitações destes métodos, e em particular o facto de nesta implementação apenas se lidar com palavras simples, este métodos demonstraram ter potencial para serem utilizados na construção ou na expansão de recursos léxico-semânticos para a língua portuguesa.

Referências

1. Chris Biemann, Stefan Bordag, Uwe Quasthoff: Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurences. In: Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation, Lisboa, Portugal (25 May 2004).
2. Kenneth Church, William Gale, Patrick Hanks, Donald Hindle: Using statistics in lexical analysis. In: Uri Zernik (ed.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. New Jersey: Lawrence Erlbaum (1991) pp. 115-164
3. Gregory Grafenstette: Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. Corpus processing for lexical acquisition. MIT Language, Speech and Communication Series. MIT Press Cambridge, MA, USA (1996) pp. 205 – 216.
4. Marti A. Hearst: Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics. Nantes, France (1992) pp. 539 - 545.
5. Dekang Lin: Automatic Retrieval and Clustering of Similar Words. In: Proceedings of COLING-ACL 1998, Montreal, Vol. 2 (1998) pp. 768–773.
6. M. Lynne Murphy: Semantic Relations and the Lexicon: antonymy, synonymy and other paradigms. University Press, Cambridge (2003)
7. Patrick Pantel and Deepak Ravichandran: Automatically Labeling Semantic Classes. The Proceedings of HLT-NAACL. Boston, MA (2004) pp. 321-328
8. Reinhard Rapp: The computation of word associations: comparing syntagmatic and paradigmatic approaches. In: Proceedings of the 19th international conference on Computational linguistics. Taipei, Taiwan (2002)
9. Luís Sarmento and Luís Cabral: BACO – A large database of text and co-occurrences. In preparation (2005).
10. Peter D. Turney: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the 12th European Conference on Machine Learning. Lecture Notes in Computer Science; Vol. 2167 (2001) pp. 491-502