Geração de uma Ontologia Lexical para o Português: Métodos e Avaliação: estado da arte

Nuno Alexandre Lopes Seco

December 11, 2006

Contents

List of Figures

List of Tables

Chapter 1

Estado da Arte

1.1 Introdução

Neste capítulo revemos a bibliografia mais relevante na área da Extracção Automática de Ontologias (EAO). A maior parte do trabalho realizado nesta área tem incidido sobre o Inglês, isto é, a partir de um conjunto de textos escritos em Inglês é extraído uma ontologia que representa o conhecimento contido nesses documentos. No entanto, no nosso estudo pretendemos focar o Português para o qual também já existe algum trabalho realizado (embora seja consideravelmente menor). Tendo este facto em conta decidimos dividir o nosso estudo em dois grupos distintos.

1. Línguas Estrangeiras
2. Português

1.1.1 Métodos de Extracção

Dicionários

O primeiro trabalho que focou a extracção de relações taxionómicas do Merriam-Webster Pocket Dictionary é do Amsler (1980) [3], isto é a extracção do hiperónimo da entrada do dicionário referente a um sentido de uma palavra. As relações foram extraídas e desambiguadas manualmente. Embora não o tenha efectivamente implementado Amsler especulou que seria possível definir uma gramática (puramente sintática) que automatizasse o processo.

Outro aspecto interessante discutido no trabalho de Amsler é o facto de este esclarecer que não é possível definir uma entrada sem recorrer a outros conceitos que estão relacionados através meronímia, sinonímia ou hiperonímia. Assim como a utilização de *case-arguments*, tais como instrumento, agente, modo, etc. *Case-arguments* são o conjunto abstracto de categorias de podem ser considerados como argumentos de verbos.

Em de salientar que na tese de Vanderwende [35] esta considera que todos os investigadores posteriores a Amsler têm verificado até certo ponto todos estes aspectos teóricos.

O trabalho de Chodorow et al. [8] apresenta um dos primeiros estudos com o objectivo de extrair automáticamente uma hierarquia lexical a partir de texto. Eles implementam heurísticas que automatizam o processo de identificação de hiperonimos exposto na tese de Amsler Neste caso concreto o texto provem de definições de dicionários em inglês, mais especificamente o *Longman Dictionary of Contemporary English* (LDOCE) e o *Webster's Seventh New Collegiate Dictionary*. Para a extracção da hieraquia é essencial a identificação e consequente extracção do género (o hiperónimo) contido em cada definição relativa a uma palavra. Os autores dão como exemplo uma entrada do dicionário correspondente à palavra CARRO:

car:a *vehicle* moving on wheels.

A palavra *vehicle* corresponde a um hiperónimo da palavra *car*. Já foi demonstrado por [3] que a ocorrência do hiperónimo da palavra a ser definida é sistemático nas definições, tal como é a menção de características que permitem distinguir a palavra em causa dos seus co-hipónimos. De acordo com esta observação aceitamos [**?**] como o hiperónimo e a menção a *wheels* como a caraterística que distingue *car* de outros tipos de *vehicles* como boats ou airplanes. Embora o exemplo dado por Chodrow et al. não seja o melhor visto que ter wheels não é suficiente para distinguir um carro de um mota, pensamos que transmite a ideia de que os dicionários podem ser processados de forma a extrair hierarquias lexicais. Explorando estas características utilzando ferramentas computacionais estes autores extraiem uma hierarquia lexical utilizando os hiperónimos que se encontram nas definições e extraem conjuntos de palavras que estão relacionados entre si por usufruirem das mesam diferenciação (e.g terem *wheels*).

O trabalho deles assenta sobre a observação de que o hipronimo de uma entrada é normalmente a cabeça do sintagma nominal ou verbal. Os autores afirmam que as definições do dicionário utilizado estão expostas de uma forma simples de prevísivel e como tal não necessitam de todo o poder fornecido por um analisador sintático.

O trabalho de Markowitz et al. [21] apresenta-nos um conjunto de padrões linguísticos a que eles chamam de *defining formulae* que nos permitem identificar e extrair relações semânticas das definições do *Webster's Seventh New Collegiate Dictionary*. Os padrões identificados estão associados às relações de hiperónomia (é_um ou instância_de) e meronimia (membro_de ou parte_de). De acordo com os autores a existência da palavra *any* no inicio de uma definição indica a existência de uma relação de hiperonimia entre a palavra a ser definida e a cabeça do sintagma nominal que segue a palavra *any*. Como exemplos temos as seguintes entradas retiradas do dicionário utilizado:

1. **nectar** any delicious drink

2. **rottweiler** any of a breed of tall vigorous black short-haired cattle dogs

Destas entradas as relações *nectar is_a drink* e *rottweiler instance_of breed* podem ser extraídas. Os autores dão ainda relevância à informação que surge entre parenteses nas definições que estão relacionados com o domínio biológico. Eles observam que o nome entre parenteses está capitalizado e que fornece o nome científico do nível a que a palavra a ser definida pertence. Por exemplo:

1. **acarid** any of an order (Acarina) of arachnids including mites and ticks

Relativamente à relação de meronimia os autores identificam o padrão *a member of* como indicativo desta relação. Analog ao que acontecia no caso da hiperonimia a cabeça do sintagma nominal que sucede o padrão refere o a parte ao qual a palavra a ser deifinida pertence. Temos como exemplo:

1. **republican** a member of a political party advocating republicanism

de onde se pode inferir que *republican part_of political party*.

Para além destas relações os autores também identificam padrões que os permitem inferir os tipos de substantivos que determinado adjectivo pode modificar ou que detrminado verbo pode aceitar como argumento. No caso do verbos recorrem novamente à identificação de parenteses na definição como em:

1. **lay** to bring forth and deposit (an egg)

Onde se assumir que o objecto típico do verbo *lay* é *egg*. No caso dos adjectivos os autores sugerem que os adjectivos com definições que começam com os padrões *containing*, *consisting of*, *extending* e *causing* não podem modificar substantivos que referem objecto animados.

Em 1988 surge o trabalho de Ahlswede e Evens [1] . Este trabalho é de facto o primeiro que compara estratégias de derivação sintática com estratégias de string pattern matching no que diz respeito à extracção do hiperónimo de definição.

De modo a testar a estratégia de derivação sintáctica eles utilizaram o derivador implementado por Naomi Sager [28]. A segunda estratégia foi implementada, inicialmente combinando os comandos SED e AWK existentes nos sistemas Unix. Embora não satisfeitos com os resultados obtidos desenvolveram o seu próprio software contendo um analisador morfológico, um vocabulário etiquetado e uma gramática com cerca de 10 regras, este software demorou cerca de 4 semanas a implementar.

No estudo comparativo que fazem consideram que o resultado obtido utilizando o software deles situa-se favorável em relação ao derivador sintáctico. As razões principais que levam a tal conclusão são:

1. demorou 8 pessoas-mês para testar e desnvolver a gramática utilizada no derivador

2. o derivador sintáctico é bastante lento

Alguns autores como Lucretia Vanderwende [35], no entanto, apontam que este estudo foi realizado em 1988 e que desde então a tecnologia tem evoluído permitindo a utilização de derivadores bastante mais eficientes. Outro factor a ter em conta é que os autores deste estudo apenas estão concentrados em extrair o hiperónimo.

O trabalho de Jensen et al. [15] concentra-se na disambiguação do predicativo, isto é, se o predicativo se relaciona com o sujeito, o complento directo ou o complemento indirecto. Considere-se a seguinte frase: *"O João viu o Pedro na colina com o telescópio."*. Podemos dividir esta frase nos seguintes constituintes:

1. Sujeito: O João
2. Verbo: viu
3. Complemento Directo: o Pedro
4. Complemento Indirecto: na Colina
5. Predicativo: com o telescópio

O problema que surge a agora é saber com que elemento é que o predicativo se relaciona, isto é o mesmo que perguntar onde é que está o telescópio; no João, no Pedro ou na colina? Neste exemplo qualquer das hipoteses é igualmente possível de modo que só será possível desambiguar o predicativo considerando o contexto em que a frase aparece. No entanto existem situações em que as relações semânticas entre as palavras do predicativo e dos restantes constituintes da frase poderão indicar o elemento com o qual o predicsativo se relaciona. É precisamente tendo esta premissa em mente que o trabalho de Jensen at al. [15] é realizado. Eles mostram que através da utilização de padrões linguísticos aplicados sobre o *Webster's Seventh New Collegiate Dictionary* conseguem extrair informação semântica relevante para efectuar a desambiguação do predicativo. Neste trabalho os autores focam essencialmente as relações de PARTE_DE e de INSTRUMENTO, com estas relações mostram como conseguem efectuar o processo de desambiguação. Alguns exemplos dos padrões utilizados para cada tipo de relação são:

1. part_of: part of, arises from, end of, member of
2. instrument: for, used for, used to, a means for, ...

No trabalho de [37] é feita menção a algumas das assunções que se têm em mente quando se pretende extrair de conhecimento de um dicionário de uma forma automática. Esta assunções têm a ver com os seguintes pontos:

1. Suficiência — diz respeito ao conhecimento existente num dicionário e se este é suficinete para facilitar o processamento computacional de texto.

2. Extracabilidade — diz respeito à capacidade de extracção de conhecinento das definições contidas no dicionário

3. Bootstrapping — diz respeito ao conhecimento linguistico inicial que é necessário para efectuar a extracção automática de conhecimento do texto das deinfições.

Os autores comentam que os projectos que se baseiam na construção manual de estruturas semânticas (e.g. CYC ou WordNet) têm visões pessimistas em relação aos pontos referente a *Extracabilidade* e *Bootstrapping*. Finalmente os autores apresentam 3 métodos de extracção automática de conhecimento que implementaram. Os 3 métodos discutidos assentam sobre as assunções acima apresentadas, no entanto variam na quantidade de informção inicial que é necessária. O primiero é uma abordagem baseada em co-occorrências que permite aos autores estabelecer associações entre palavras. Este método não necessita de informação linguistica inicial. O segundo método já faz uso de uma gramática e de uma colecção de padrões linguísticos que permitem, para além de outros de outros itens, identificar o género (hiperónimo) e o diferenciado para cada entrada no dicionário. O útltmo método é o que requer mais conhecimento incial e permite aos autores criarem uma estrutura semântica livre de referências circulares. Por exemplo os autores notam que a palavra *trip* está definida em função de *journey* e que *journey* está definido em função de *trip*. Estas referências circulares tornam o conhecimento algo vago e é conviniente remove-las. Os autores partem de um conjunto de 3600 unidades semânticas que correspondem aos vários sentidos das 1200 palavras que são utilizadas para definir o vocabulário controlado do dicionário LDOCE. De seguida o algoritmo analisa as restantes palavras que se encontram no dicionário e para aquelas que são definidas com palavras que já têm uma unidade semântica é gerada a unidade semântica para a entrada em causa. No final de cada ciclo é adicionado ao conjunto de unidades semânticas as novas unidades. O processo repete-se até que todas as palavras tenham sido transformadas em unidades semânticas. De acordo com os autores são precisos quatro iterações até que todas as palavras sejam processadas.

Em 1989 Alshawi [2] estuda outra forma de analisar as definições do LDOCE. A sua preocupação principal é em permitir derivações sintácticas parciais que se adequam mais à realidade das definições que encontramos em dicionários, que na realidade na constituem frases gramaticalmente completas. A gramática utilizada é proposta com base nas definições encontradas no LDOCE daí ser praticamente impossível aplicá-la num domínio de texto livre ou, até, noutro dicionário.

Ann Copestake apresenta-nos o seu trabalho de extracção de hierarquias de dicionários conduzido no âmbito do projecto ACQUILEX[9]. O trabalho dela difere dos anteriores no sentido em que ela tenta automatizar o mais possível o tarefa

de desambiguação dos sentidos das palavras contidas no dicionário. O LDOCE é, mais uma vez, o dicionário utilizado pela autora que se concentra somente nos substantivos. É utilizado o parser desenvolvido por Alshawi [2] para analisar as definições do dicionário. Este parser extrai o hiperónimo da definição de cada entrada no dicionário que normalmente é a cabeça do sintagma nominal contido na definição. O algoritmo desenvolvido por Copestake segue uma estratégia *top-down*, dado uma entrada do dicionário o algoritmo encontra todas as definições que contêm a palavra utilizada no mesmo sentido. Depois para cada uma das entradas recolhidas o algoritmo procura recursivamente as entradas que contêm as anteriores até se chegue a entradas que não são utilizadas em mais definições. Por exemplo, dado a palavra líquido no sentido de uma substância que flui encontramos que na definição da palavra óleo a palavra líquido (no mesmo sentido) corresponde ao hiperónimo da entrada estabelecendo-se assim uma relação de hiperónima entre liquido e óleo. O processe agora repetia-se para o sentido de óleo em causa formando assim uma cadeia de entradas relacionadas por hiperonomia. Obviamente que fulcral a este procedimento esta a desambiguação dos sentidos das palavras nas deifnições. A autora recorre a um conjunto de heurísticas para determinar qual o sentido correcto da palavra em cada definição. O LDOCE tem associado a cada entrada um código, o *box code*, que exprime informação semântica. Os códigos representam conceitos como humano, abstracto, líquido, ... Uma das heurísticas tem em conta essa informação e no exemplo do óleo e líquido acima referenciado serviu para determinar o sentido da palavra líquido na definição de óleo era o pretendido porque a entrada de óleo e a entrada de líquido inicialmente fornecido partilhavam o mesmo *box code*. Tal como no trabalho de Wilks et al. [37], Copestake também faz referência as definições circulares que dificultam e nalguns casos prejudicam a construção da hierarquia. É feito referência que o modelo de herança tem permitir um mecanismo de *overriding* das propriedades dos nós. Por exemplo podemos dizer que uma *lista telefónica* é um *livro*, que uma *autobiografia* é um *livro* e que um *livro* é para *ser lido*. No entanto, não parece estar correcto dizer que uma *lista telefónica* é para *ser lida*. Por este motivo é necessário permitir que certos attributos de determinado conceito possam ser redefinidos em especializações deste. Neste caso diriamos que *lista telefónica* é para *ser referenciado*.

Montemagni et al. [22] reparam que grande parte do trabalho feito tem focado na extracção automática do hperónimo das entradas. No entanto a extracção das características que diferenciam um conceito do seu hiperónimo tem, em comparação, sido bastante reduzido. Assim sendo focam o seu trabalho nesta faceta. Os autores argumentam que a extracção de hiperónimos (que normalmente se encontram no inicio da definição) tem funcionado bem com recurso a pattern-matching, mas que para encontrar o differentia (que não se encontra no inicio da definição) é necessário uma análise estrutural mais profunda. As derivações estruturais em combinação com heurísticas de extracção sobre as derivações estruturais permitem a detecção do "differentia". Um exemplo de uma heurística

que apresentam para a extracção de uma relação de Porpósito é:

> if tile PP with FOR is not a post-modifier of a verb USED, then a PURPOSE relation between the definiendum and the head(s) of the PP can be hypothesized if the nearest noun that the PP post-modifies is the genus term.

Eles notam que a extracção deste tipo de informação nunca seria possível utilizando esquemas simples de "pattern-matching". No entanto admitem que apesar da anàlise sintática ajudar muito nem sempre é sufuciente para a extracção semântica das differentia. Pois existem casos de ambiguidade sintática (ver exemplo do predicativo).

Bruce et al. [6] é mais um dos trabalhos de extracção automática de uma taxonomia através da analise de definições. Eles apontam que dois subproblems têm de ser considerados de modo a se efectuar a extracção automática:

1. definir um algoritmo que identifique o hiperónimo
2. definir um algortimo de desambiguação do hiperónimo

Eles concentram esforços no segundo item e extende um algoritmo anterior também da autoria deles. O dicionário utilizado pelos investigadores foi o LDOCE. Como já foi salientado em trabalhos anteriores o LDOCE associa códigos semânticos e pragmáticos a cada entrada do dicionário assim como informação sobre frequência de utilização. Este códigos estão organizados numa estrutura hierarquica que os autores utilizam para efectuar a desambiguação do hiperónimo. Por exemplo, se uma entrada tem associado determinado código semântico então os autores escolhem a entrada da palavra identificada como sendo o hiperónimo com o mesmo código. Ainda no processo de desambiguação entram o código pragmático e a informação sobre a frequência de utilização da palavra, com cada um deste factores a ter uma contribuição especifica.

Embora o trabalho de Kozima et al. [17] se concetre mais na determinação automática de semelhanças semânticas entre palavras, ele utilizam uma rede semântica extraída a partir do vocabulário fechado do LDOCE para efectuar os cálculos. A metodologia de extracção da reda passa por analisar um conjnuto de cerca de 3000 palavras sementes (que sao utilizadas para definir as restates palavras no dicionário) criando uma estrutura a chamaram de *Paradigme*.

Em 1994 Nancy Ide e Jean Verónis [13] apresentam uma artigo crítico sob a investigação que se tem feito na àrea de extracção de informação a partir de dicionários e questionam os avanços conseguidos nos 15 anos antecedentes. Neste trabalho os autores afirmam que toda a investigação na àrea na tem conseguido mais do que a extracção de pequenas taxonomias de palavras. Como causas do aparente fracasso apontam questões como a variabilidade dos padrões linguísticos utilizados na extracção de informção. A falta de conhecimento do mundo para

distinguir relações diferentes mas manifestadas utilizando o mesmo padrão textual. Por exemplo as frases: *...levar numa carrinha* e *...levar pela alça* necessitam de conhecimento de mundo de que o objecto a ser levado no primeiro caso não faz parte da carrinha, mas que a alça no segundo exemplo já é parte integrante do objecto a transportar. Apontam ainda aspectos de inconsistências entre dicionários o que leva a crer que cada dicionário só poderá conter parte do conhecimento real.

No entanto admitem que o trabalho realizado utilizando dicionários tem contribuído para a compreensão da natureza, do tipo, do papel semântico dos itens léxicos encontrados em dicionários. Embora que o objectivo inicial de conseguir automáticamente uma Ontologia Lexical tenha fracassado.

Outro aspecto salientando pelo mesmo autores em [14] é o facto de nenhuma avaliação ao conhecimento extraído de dicionários ter sido avaliado. Como tal, em [14], pretendem apresentar uma primeira avaliação do conhecimento extraído do dicionário. O trabalho indica que a informação extraída apenas de um dicionário é bastante fraca, no entanto a utilização de vários dicionários parece fornecer resultados bastante mais adequados. Identificam uma série de problemas de que todos os dicionários (utilizados no trabalho deles) sofrem, mas no entanto combinando a informação extraída de todos é possível melhor a conhecimento extraído significativamente. Os autores afirmam que cerca de 50-70 % informação extraída de um só dicionário está de alguma forma distorcida, mas que através da utilização de vários dicionários conseguem diminuir essa distorção para cerca de 4%.

O trabalho de Vanderwende [35] está centrado na interpretação daquilo que ela chama de *Noun Sequences* que traduzimos para sequencia de substantivos. Embora esta área pouco se relaciona com o que nos preocupa neste trabalho, a autora utiliza informação extraída automaticamente de uma dicionário; e é precisamente esse módulo que iremos focar.

O derivador sintáctico utilizado, denominado de MEG Parser (Microsoft English Grammar Parser), ao contrário do que acontece noutros trabalhos não é alterado para lídar com as especificidades léxicas das entradas dos dicionários. O parser não só produz informação sintáctica como também produz informação funcional sobre cada constituinte da frase. Por exemplo, num grupo nominal existira informação indicando o núcleo do grupo assim como os respectivo modificadores caso estes existam.

Vanderwende salienta que muitas das definições são elípticas, isto é, existe a omissão de uma ou mais palavras sem que com isso se perca a transmissão de ideia que é pretendida. O MEG parser por não ser dependente da noção de subcategorização, que alguns parser utilizam para guiar a derivação sintáctica, consegue lidar que definições elípticas.

Outra caraterística do MEG é a sua capacidade de derivações parciais que já tinha sido referenciado em Alshawi como sendo importante. Entende-se como derivações parciais de partes da definição. Por exemplo, em inglês não existe nen-

huma gramática que preveja que numa frase apareça primiero o grupo preposicional e só depois o gupo nominal, no entanto como Vanderwende nota existem definições construídas assim. O MEG parser lida com estas situações efectuando as derivações parciais do grupo preposicional e nominal mas nunca indicando que as duas formam uma frase. Assim é possível ainda

O sistema SESSEMI extrai relações semânticas das derivações sintácticas através da procura e análise de padrões indicativos de determinado tipo de relação. Vanderwende aponta que estes padrões sintácticos são fiáveis visto que são extraídos de dicionários que apesar de tudo são construídos com alguma sistematicidade o que lhes torna fontes de conhecimento adequadas. A mesma estratégia, segundo a autora, não poderia ser aplicado em texto livre.

Estes padrões são aplicados de modo a gerar uma estrutura frame com conteúdo semântico da entrada. As relações extraídas são novamente processadas utilizando um novo padrão VERB-ARGS que analisa os nomes e verbos das relações extraídas.

No trabalho desenvolvido por [24] incide sob a extracção de características diferenciadoras entre conceitos que são co-hipónimos, tentando extrair relações do tipo do utilizado_para e tem_tamanho. Estas relações são de extrema importância pois permitem explicitar informação importante sobre os conceitos. Os dicionários seguem regras de lexicografia que facilita a extracção deste conhecimento, normalmente designado como sendo dicionarios do tipo analítico (ver [4]) argumentando que técnicas estatística baseadas em extracção de ngramas de corpora nunca serão capazes de capturar as caraterísticas differenciadoras das palavras. No entanto tambem salienta, de acordo com [18] que as definições de dicionários estão sempre incompletas ou vaga em relação a determinados detalhes necessários para compreender um conceito. Para evitar a construção manual de regras de extracção utilza o conhecimento sobre relações contidas no FrameNet e no Penn Treebank.

O autor utilza o WordNet como um simples dicinário para este estudo. Todas as definiçoes passam por uma pré-procesamento da em que é acrescentando e transformando-as para que possam ser facilmente interpretadas por parsers. De notar ques esta abordagem contrasta com as anteriores onde é necessário ajustar o parser para processar definições. Por exemplo, no Wordnet a palavra LOCK está definida como *"a fastener.... "*. O autor utilza um conjunto de heurísticas para transformar a definição em *A LOCK is a fastener...* facilitando posteriormente o trabalho do parser. Exsitem alterações deste genero para cada tipo de palavra (nome, verbo, adjectivo, adverbio). O parser utilizado é o Link Grammar Parser que produz relações sintácticias que são depois convertidas para relações semânticas de acordo com mapeamentos pré-definidos pelo autor. Depois de um processo de desambiguação das relações produzidas e dos termos que nelas participam. As relações são pesadas de acordo com uma métrica de *cue validities* (ver [30]) que basicamente determina a importância da relação. As relações, para uma dada entrada no wordnet, são aglomeradas numa estrutura recursiva.

Copora

1.1.2 Métodos de Avaliação

Nesta sub-secção iremos abordar algumas das métricas e estratégias seguidas para avaliar os algoritmos de extracção acima referenciados. Note-se que a avaliação no contexto de EAO normalmente incide sobre o *produto*, isto é, sobre a ontologia que é gerada através de um qualquer processo de extracção. A questão da avaliação, independentemente da disciplina, é essencial em qualquer abordagem cientifica pois permite comparar ideias e seguir os caminhos mais indicados na procura de novo conhecimento científico. Concretizando, e seguindo a divisão proposta proposta por [11], podemos encontrar 4 estratégias de avaliação de EAO:

1. Categoria 1 — utilização de uma *referencia dourada* (e.g. uma ontologia) que permite comparar o resultado obtido com o resultado desejado

2. Categoria 2 — utilização de uma aplicação que necessita de uma ontologia e avaliar os resultados produzidos pela aplicação quando se introduzem modificações na ontologia

3. Categoria 3 — utilização de corpora que incide sobre o mesmo conhecimento representado na ontologia e medir a *cobertura* da ontologia em relação ao corpus

4. Categoria 4 — utilização de avaliadores humanos que efectuam uma avaliação qualitativa da ontologia de acordo com critérios estabelecidos

De seguida apresentamos alguns exemplos que se enquadram nas quatro categorias acima referidas.

Categoria 1

Fortemente ligado a esta categoria estão todos os trabalhos que tentam avaliar quantitativamente a semalhança entre duas ontologias. Pois assumindo a existência de uma referência dourada, neste casa uma ontologia, estes trabalhos tornam-se bastante relevantes. Alguns dos mais destacados são o de [20] e o de [31].

Em [20] é efectuado uma comparação ao nível léxico e taxonómico de duas estruturas ontológicas. Podemos assumir na nossa discussão que uma das ontologia é o produto de um qualquer processo de extracção e a outra uma referência dourada já validada. Ao nível léxico a avaliação é efectuada comparando o vocabulário de uma ontologia com a da outra. Esta avaliação não tem em consideração a organização semântica dos conceitos, apenas pretende averiguar se os termos da ontologia de referência existem na ontologia a validar. Embora seja reconhecidamente uma avaliação muito superficial, poderá ser a primeira a ser

efectuada numa bateria de testes. Para além disso é possível utilizar métricas bem conhecidas pela comunidade científica tais como a *precisão* e *abrangência*. Os autores utilizam ainda a métrica de semelhança de [19] que os permite atribuir um valor aos vocábulos que não são exactamente iguais mas que são muito aproximados. Ao nível taxonómico as duas estruturas podem ser comparadas utilizando como pontos de referência os conceitos que foram identificados como pertencentes a ambas as estruturas. Estes autores propõem que para cada conceito comum se extraia, separadamente, todos os hiperónimos e hipónimos de cada ontologia dando origem a dois conjuntos de conceitos, C_z e C_x, respectivamente. Para cada conceito comum é possível obter um valor numérico que representa a semelhança entre as duas taxonomias parciais ao qual o conceito pertence bastando para tal basta calcular o conhecido coeficiente de Jaccard: $|C_z \cap C_x| \setminus |C_x \cup C_z|$. Poder-se-á então calcular a média dos coeficientes para todos os conceitos de forma a obter-se um valor global da semelhança das duas ontologias. Parece-nos, no entanto, que esta forma de comparação despreza as relações explícitas entre pares de conceitos o que pode de alguma forma enviesar os resultados obtidos. Imagine-se, o caso extremo, que para um qualquer conceito comum os hipónimos numa ontologia correspondem aos hiperónimos na outra e vice-versa. Neste caso a coeficiente obtido terá o valor de 1 indicando semelhança máxima, o que está claramente errado. Uma solução como a que foi proposta por [31] parece ser mais adequada, pois tem em conta a relações utilizadas. O objectivo deste trabalho é extrair um conjunto de triplos da forma <TERMO$_1$, VERBO, TERMO$_2$> ou <TERMO$_1$, PREPOSIÇÃO, TERMO$_2$> de texto (podemos considerar que triplos contendo palavras em comun podem ser ligados dando início uma ontologia). O autor confronta os triplos extraídos com um conjunto de termos que são considerados estatisticamente relevantes e que constituem a referência dourada.

Outro trabalho de comparação de ontologias é apresentado em [11] onde extendem a noção de índice de Rand [26] de forma a poderem comparar duas ontologias. A seguinte formulação é proposta: Dado um conjunto C de conceitos comum a duas ontologias U e V podemos averiguar a sua semelhança utilizando a seguinte formula: $sim(U,V) = 1 - [\sum_{1 \leq i,j \leq n} \mid \delta(U(c_i), U(c_j)) - \delta(V(c_i), V(c_j)) \mid]/[n(n-1)/2]$ onde $U(c_i)$ e $V(c_i)$ devolvem uma referência para o conceito c_i em cada uma das ontologias. A função δ devolve o valor 1 se os conceitos c_i e $c_j$pertencerem à mesma categoria (são hipónimos do mesmo conceito) e zero caso contrário. O denominador da expressão representa o número de combinações possíveis de pares de conceitos de C. Esta fórmula permite avaliar a semelhança estrutural (pelo menos do ponto de vista taxonómico) das duas ontologias. Quanto mais próximo de 1 for o valor de sim então mais semelhantes são as taxonomias.

Categoria 2

Nesta categoria encontramos o trabalho de [25] para determinar o grau de coerência de um conjunto de conceitos de acordo com uma ontologia. O trabalho destes autores incide na área do processamento da fala, onde para cada locução é necessário escolher de entre as várias *hipóteses de interpretação*[1] as que mais se adequam ao contexto. Note-se que entendemos interpretação como a geração de uma estrutura (formal e não ambígua) capaz de representar sentido do discurso [16]. O módulo responsável por esta tarefa no sistema em causa é conhecido como *OntoScore*. A estrutura produzida por este sistema, ao interpretar o discurso, é um grafo que relaciona os termos contidos na dicção. Como cada termo poderá referir-se a conceitos diferentes, devido à ambiguidade da linguagem, o OntoScore tem de considerar todas as combinações de conceitos possíveis. Para cada par de conceitos em cada combinação é encontrado o menor caminho, com base na ontologia, que une os conceitos. Portanto, para cada conjunto de conceitos é gerado um grafo que liga cada conceito a todos os outros com base nas relações encontradas na ontologia. A cada tipo de relação é atribuído um peso entre [0..1] e o OntoScore tem de encontrar a combinação que minimiza a soma dos pesos das relações atribuídas a cada grafo. O grafo com o menor peso é escolhido como sendo a melhor interpretação para o discurso e é comparado com o grafo que foi manualmente criado para o efeito. A comparação é efectuada quantitativamente contando o número de inserções, remoções ou substituições de relações que teriam de ser efectuadas no grafo gerado automaticamente até se obter o grafo gerado manualmente (análogo à medida de [19] para strings). Dado este cenário facilmente conseguimos conceber um método de avaliação de EAO. Consideremos dois métodos de EAO, M_1 e M_2, que ao processar um mesmo conjunto de textos T produz duas ontologias O_1 e O_2, respectivamente. Podemos indirectamente comparar os métodos, M_1 e M_2, comparando os resultados obtidos na aplicação anteriormente descrita quando se utiliza cada uma das ontologias geradas. Se, por exemplo, as interpretações produzidas utilizando a ontologia O_1 são melhores do que as produzidas utilizando a ontologia O_2, então podemos especular que o método M_1 mostra-se mais adequado que o M_2, ou o contrário caso os resultados conseguidos são melhores utilizando a ontologia O_2.

O trabalho de [27] também pode ser considerado como pertencente a esta categoria (embora uma relação com a categoria 4 também possa ser estabelecida). Richardson derivou relações semânticas entre palavras utilizando um dicionário electrónico. As relações extraídas são atribuídas um peso representativo daquilo a que o autor chama de *associação cognitiva*, isto é, a relevância que a relação tem em associar as duas palavras que nela participam. Por exemplo, se se perguntasse *quais as partes de um carro?*, existem certamente componentes dum carro que nos chamam mais à atenção do que outras (motor é certamente mais saliente do que rádio). O peso que Richardson atribui às relações (de meroníma neste caso)

[1] Traduzido do inglês — Speech Recognition Hypotheses.

que envolvem a palavra carro devem reflectir a relevância da parte em causa no contexto de carro. De acordo com o autor este pesos servem também para medir a precisão das relações extraídas, pois podemos assumir que se existem muitas relações com pesos baixos então a precisão do mecanismo de extracção é reduzida.

Categoria 3

Paradigmático desta categoria é o trabalho realizado por [5]. Neste trabalho o autor tenta não só avaliar o léxico utilizado numa ontologia mas também a sua estrutura. O método, constituído por quatro etapas [11], assume a existência de um corpus representativo do domínio a modelar. Numa primeira etapa é gerado um conjunto de clusters/tópicos utilizando uma abordagem probabilística baseada em *expectation maximization* de tal modo que no fim do processo de clustering cada documento é considerado como pertencente um ou mais tópicos. Seguidamente, cada conceito contido na ontologia é representado num vector de termos que incluí o nome utilizado para representar o termo na ontologia assim como os termos do dois hiperónimos imediatos no WordNet. Este "mini-documento" (vector de termos) pode ser utilizado em conjunção como o modelo probabilístico obtido no passo anterior para medir o grau de pertença do conceito aos tópicos extraídos. Assim, é possível medir-se a cobertura do vocabulário utilizado na Ontologia analisando os conceitos que são realmente considerados como pertencentes a pelo menos um tópico. Ainda utilizando os clusters obtidos no passo inicial podemos avaliar, de um modo superficial, a estrutura da ontologia. Isto é, conceitos que são mapeados para os mesmo tópicos têm de estar muito próximos semanticamente na ontologia indicando que a estrutura da ontologia está de acordo com os tópicos extraídos do corpus.

Categoria 4

Nesta categoria encontramos o trabalho de [34] que tem como objectivo auxiliar os engenheiros de conhecimento, de uma qualquer empresa, a escolher a melhor ontologia a ser utilizado num (novo) projecto. A metodologia basicamente resume-se a uma análise humana das ontologias candidatas tendo em conta um conjunto de critérios e objectivos estabelecidos para o projecto. Como este tipo de avaliação envolve questões como:

1. O custo da ontologia
2. A linguagem utilizada para implementar a ontologia
3. O software utilizado para desenvolver a ontologia

parece-nos que detalhar mais esta abordagem não terá grande interesse no âmbito deste trabalho, visto que o intuito é criar uma ontologia de âmbito geral sem ter nenhuma aplicação específica em mente, tal como esta metodologia pressupõe.

Destacamos ainda o trabalho de [12] que introduzem um conjunto de meta-propriedades (e.g., Rigidez, Identidade e Unidade) para estabelcer um conjunto de restrições entre as ligações que podem ser estabelecidas entre conceitos. A primeira propriedade, Rigidez, assenta sobre a noção de essencialismo. Diz-se que uma propriedade de um conceito é essencial se a propriedade é válida para esse conceito. Por exemplo, a propriedade *é duro* é uma propriedade essencial de um martelo mas não de uma esponja. No entanto é possível que algumas esponjas (secas) sejam duras, contrariamente aos martelos que serão e sempre foram duros (caso contrário não serviram para o propósito para o qual foram concebidos). Assim sendo, diz-se que uma propriedade é rígida se esta for essencial para todas as instâncias que têm a propriedade. No exemplo dado, podemos dizer a que a propriedade *é duro* não é rígida, pois existem instâncias, esponjas, que poderão ter essa propriedade apenas por um determinado período de tempo. Em contraste temos a propriedade *é pessoa* que é rígida. Isto é, não é possível que existirem instâncias que só usufruiram desta propriedade durante um período de tempo. Dentro das propriedades não rígidas é feita ainda outra distinção dividindo este grupo em semi-rígidas e anti-rígidas. Uma propriedade é semi-rígida quando existem instâncias para as quais a propriedade é essencial e outras para as quais não é essencial como é o caso da propriedade *é duro* (essencial para o martelo e não essencial para a esponja). Propriedades anti-rígidas são aquelas que são aquelas que nunca serão essenciais para qualquer instância; por exemplo a propriedade *é estudante* normalmente é considerada anti-rígida, pois ser estudante (no sentido mais estrito) é uma propriedade que só afecta as instâncias durante um intervalo de tempo, ou seja, não existem instâncias que têm, sempre tiveram e sempre terão a propriedade *é estudante*. Assumindo que é possível etiquetar todos os conceitos de uma ontologia de acordo com a sua rigidez, podemos empregar regras que nos permitem validar a informação na ontologia. Por exemplo, nunca será possível termos a classe *pessoa* a herdar (através de relação de hiperonímia) da classe *estudante* visto que esta última é anti-rígida e a primeira é rígida. Se tal fosse possível, e como a relação de hiperonímia implica que uma pessoa é um estudante, quando alguma pessoa deixasse ser estudante também teria de deixar de ser considerada pessoa, o que é absurdo.

A segunda meta-propriedade é a Identidade. Esta propriedade prende-se com a capacidade de reconhecer entidades como sendo a mesma (ou diferentes). Esta meta-propriedade torna-se mais fácil de compreender se considerarmos o eixo do tempo. Por exemplo, ao longo do tempo, conseguimos identificar pessoas como as mesmas que conhecemos no passado mesmo que estas tenham mudado de feições. Poderá parecer um pouco paradoxal estar questionar *"se duas entidades são a mesma"*, pois se realmente são a mesma então só existe uma. No entanto, e de acordo com os autores, este tipo de questões estão sempre presentes durante o processo de modelação ontológica. Consideremos um exemplo mais prático e real (tal como os autores afirmam); suponhamos que temos na nossa ontologia a classe *duração temporal* que tem como instâncias *uma hora*, *duas hora*, *trinta*

minutos, etc. Durante o processo de modelação existiu necessidade de criar uma nova classe com o nome *intervalo de tempo* que se refere a intervalos específicos como por exemplo *terça-feira das 14:00 às 15:00* e *sexta-feira das 12:00 às 13:00*, etc. A ligação entre estas duas classes foi estabelecida utilizando uma relação de hiperonímia da seguinte forma *intervalo de tempo* é uma sub-classe de *duração temporal* o que à partida parece fazer sentido. Agora podemos utilizar a noção de identidade para avaliar a decisão tomada. O que torna duas durações temporais iguais é facto de terem a mesma duração, ou seja todas as durações temporais de uma hora são a mesma. Por outro lado, vemos que existem dois intervalos de tempo que claramente não são o mesmo embora tenham a mesma duração. Ora esta situação leva-nos a uma contradição pois de acordo com a intensão da classe *duração temporal* as duas instâncias deviam ser a mesma, mas de acordo com a intensão da classe *intervalo de tempo* elas são diferentes. Esta contradição surge devido à forma relaxada e ambígua como se utiliza a linguagem natural na frase "intervalos de tempo são durações temporais" o que não é verdade. Rigorosamente deveríamos dizer que "intervalos de tempo têm durações temporais" não modelando, portanto, a relação de subclasse mas antes a de **parte de** (duração temporal é **parte de** intervalo de tempo).

A noção de Unidade obriga-nos a pensar no que faz e não faz parte de uma entidade. Diz-se que um conceito manifesta a propriedade de unidade quando as suas instâncias podem ser identificadas como entidades singulares. Por exemplo, o conceito *água* não tem instâncias que possam ser identificadas como entidades isoladas. Por outro lado o conceito *oceano* já manifesta a propriedade de unidade pois tem instâncias que correspondem a entidades isoladas como o Oceano Atlântico. Logo seria paradoxal ter *oceano* como uma subclasse de *água*.**Ainda não percebi bem isto**

Os conceitos da ontologia podem ser etiquetados de acordo com estas propriedades permitindo assim a verificação automática da semântica da ontologia. Várias ferramentas de edição de ontologias já implementam as restrições impostas por esta teoria, alguns exemplos são: ODEClean [10], OntoEdit [33] e Protégé [23]. Recentemente em [36] foi proposto uma metodologia que permite definir, de uma forma automática, estas meta-propriedades para os conceitos de uma ontologia.

Níveis de Avaliação

Para além da divisão anteriormente elaborada em relação às estratégias de avaliação podemos ainda identificar outra forma de categorizar a avaliação; de acordo com determinada característica ou nível da ontologia em que a avaliação incide. Isto é, como uma ontologia é uma estrutura relativamente complexa, poderá ser mais fácil avaliar a ontologia focando apenas alguns desses aspectos ou níveis. Os níveis propostos por [11] são os seguintes:

1. Lexical — A este nível o interesse está em averiguar se todos os conceitos

e os respectivos termos foram incluídos na ontologia.

2. Taxonómico — A componente taxonómica de qualquer ontologia desempenha um papel importante na organização do nosso conhecimento [29]. Assim sendo, existem abordagens que se concentram em validar a estrutura taxonómica da ontologia.

3. Relacional — Apesar das relações anteriores desempenharem um papel importante, estas não são suficientes para representar o conhecimento de forma eficaz (considere o conhecido *"Tennis Problem"* [32] no WordNet), de modo que se torna importante avaliar estas relações.

4. Contextual — No contexto da semantic web uma ontologia poderá fazer parte de uma *floresta* de ontologias que se referenciam mutuamente. Poderá fazer sentido avaliar a ontologia no contexto das outras. Isto é, uma ontologia que contenha conceitos que são frequentemente referenciados em outras ontologias pode ser considerada como autoritária. Significando que o conhecimento nela contida é considerado como correcto pela restante comunidade. Um paralelo pode ser estabelecido com *page rank* do Google.

5. Sintáctica — Normalmente uma ontologia é representada utilizando uma linguagem formal (e.g. OWL, RDF) com uma sintaxe rígida a avaliação a este nível pretende validar a sintaxe da linguagem utilizada na construção da ontologia.

6. Estrutural — A este nível pretende-se averiguar se a ontologia encontra-se correctamente estruturada e documentada. Por exemplo, uma ontologia contendo definições em linguagem natural deverá utilizar uma linguagem pouco ambígua e estar de acordo com a estrutura ontológica em causa.

Mais tarde, em [7], foi acrescentado um outro nível de avaliação que foi chamado de filosófico e que pretende avaliar uma ontologia de acordo com critérios filosóficos.

1.1.3 Aplicações

1.2 Agradecimentos

Este trabalho foi realizado no âmbito do projecto POSI/PLP/43931/2001, financiada pela Fundação para a Ciência e Tecnologia (FCT), co-financiada pelo POSI.

Bibliography

[1] Thomas Ahlswede and Martha Evens. Parsing vs. text processing in the analysis of dictionary definitions. pages 217–224, 1988.

[2] Hiyan Alshawi. Analysing the dictionary definitions. *Computational Lexicography for Natural Language Processing*, pages 153–170, 1989.

[3] Robert Alfred Amsler. The structure of the merriam-webster pocket dictionary. 1980.

[4] J. Ayto. On specifying meaning. *Lexicography: Principales*, pages 89–98, 1983.

[5] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data-driven ontology evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, 2004.

[6] Rebecca Bruce and Louise Guthrie. Genus disambiguation: a study in weighted preference. pages 1187–1191, 1992.

[7] Paul Buitelaar, Philipp Cimiano, Marko Grobelnik, and Michael Sintek. Handouts of the tutorial on ontology leanring from text, 2005.

[8] Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. pages 299–304, 1985.

[9] Ann Copestake. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. 1990.

[10] Mariano Fernández-López and Asunción Gómez-Pérez. The integration of ontoclean in webode. In *EKAW2002 Workshop on Evaluation of Ontology-Based Tools (EON2002)*, 2002.

[11] Janez Groblnik, Marko Grobelnik, and Dunja Mladenic. D1.6.1 ontology evaluation. Technical report, Josef Stefan Institute, 2005.

[12] Nicola Guarino and Christopher Welty. Evaluating ontological decisions with ontoclean. *Commun. ACM*, 45(2):61–65, 2002.

[13] Nancy Ide and Jean Veronis. Machinereadable dictionaries: What have we learned, where do we go. 1994.

[14] Nancy Ide and Jean Véronis. Refining taxonomies extracted from machine-readable dictionaires. *Research in Humanities Computing*, 2:145–170, 1994.

[15] Karen Jensen and Jean-Louis Binot. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Comput. Linguist.*, 13(3-4):251–260, 1987.

[16] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[17] Hideki Kozima and Teiji Furugori. Similarity between words computed by spreading activation on an english dictionary. pages 232–239, 1993.

[18] S. Landau. *Dictionaries: The Art and Craft of Lexicography*. Cambrdge University Press, Cambridge, second edition, 2001.

[19] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.*, 10(8):707–710, February 1966.

[20] Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In *EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 251–263, London, UK, 2002. Springer-Verlag.

[21] Judith Markowitz, Thomas Ahlswede, and Martha Evens. Semantically significant patterns in dictionary definitions. pages 112–119, 1986.

[22] Simonetta Montemagni and Lucy Vanderwende. Structural patterns vs. string patterns for extracting semantic information from dictionaries. pages 546–552, 1992.

[23] Mark A. Musen Natalya Fridman Noy, Ray W. Fergerson. The knowledge model of protégé-2000:combining interoperability and flexibility. 2000.

[24] Thomas Paul O'Hara. *Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions*. PhD thesis, New Mexico State University, 2005.

[25] Robert Porzel and Rainer Malaka. *A Task-Based Framework for Ontology Learning, Population and Evaluation*. IOS Press, July 2005. book chapter.

[26] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

[27] Stephen Richardson. *Determining Similarity and Infering Relations in a Lexical Knowledge Base*. PhD thesis, City University of New York, 1997.

[28] Naomi Sager. Natural language information processing: A computer grammar of english and its applications. 1981.

[29] Nuno Seco. Computational models of similarity in lexical ontologies. Master's thesis, University College Dublin, February 2005.

[30] Edward E. Smith and Douglas Medin. *Categories and concepts*. Harvard University Press, Cambridge, 1981.

[31] Peter Spyns. Evalexon: assessing triples mined from texts. Technical report, 2005.

[32] Mark Stevenson. Combining disambiguation techniques to enrich an ontology. In *Proceedings of the Machine Learning and Natural Language Processing for Ontology Engineering Workshop*, 2002.

[33] Y. Sure, J. Angele, and S. Staab. Ontoedit: Multifaceted inferencing for ontology engineering. *Journal on Data Semantics*, pages 128–152, 2003.

[34] Adolfo Lozano Tello and Asunción Gómez-Pérez. Ontometric: A method to choose the appropriate ontology. *Journal of Database Management*, 15(2):1–18, 2004.

[35] Lucretia H. Vanderwende. *The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries*. PhD thesis, 1995.

[36] Johanna Völker, Denny Vrandecic, and York Sure. Automatic evaluation of ontologies (aeon). In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *Proceedings of the 4th International Semantic Web Conference (ISWC2005)*, volume 3729 of *LNCS*, pages 716–731. Springer Verlag Berlin-Heidelberg, NOV 2005.

[37] Yorick Wilks, Dan Fass, Cheng ming Guo, James E. Mcdonald, Tony Plate, and Brian M. Slator. Machine tractable dictionaries as tools and resources for natural language processing. pages 750–755, 1988.