Descoberta de Synsets Difusos com base na Redundância em vários Dicionários

Discovering Fuzzy Synsets from the Redundancy across several Dictionaries

Fábio Santos
CISUC, Departamento de Engenharia Informática
Universidade de Coimbra, Portugal
fasantos@student.dei.uc.pt

Hugo Gonçalo Oliveira
CISUC, Departamento de Engenharia Informática
Universidade de Coimbra, Portugal
hroliv@dei.uc.pt

Resumo

Numa wordnet, conceitos são representados através de grupos de palavras, vulgarmente chamados de synsets, e cada pertença de uma palavra a um synset representa um diferente sentido dessa mesma palavra. Mas como os sentidos são entidades complexas, sem fronteiras bem definidas, para lidar com eles de forma menos artificial, sugerimos que synsets sejam tratados como conjuntos difusos, em que cada palavra tem um grau de pertença, associado à confiança que existe na utilização de cada palavra para transmitir o conceito que emerge do synset. Propomos então uma abordagem automática para descobrir um conjunto de synsets difusos a partir de uma rede de sinónimos, idealmente redundante, por ser extraída a partir de várias fontes, e o mais abrangentes possível. Um dos princípios é que, em quantos mais recursos duas palavras forem consideradas sinónimos, maior confiança haverá na equivalência de pelo menos um dos seus sentidos. A abordagem proposta foi aplicada a uma rede extraída a partir de três dicionários do português e resultou num novo conjunto de synsets para esta língua, em que as palavras têm pertenças difusas, ou seja, *fuzzy synsets*. Para além de apresentar a abordagem e a ilustrar com alguns resultados obtidos, baseamo-nos em três avaliações – comparação com um tesauro criado manualmente para o português; comparação com uma abordagem anterior com o mesmo objetivo; e avaliação manual – para confirmar que os resultados são positivos, e poderão no futuro ser expandidos através da exploração de outras fontes de sinónimos.

Palavras chave

wordnets, synsets, fuzzy clustering, rede léxico-semântica, sinónimos, confiança, dicionários

Abstract

In a wordnet, concepts are typically represented as groups of words, commonly known as synsets, and each membership of a word to a synset denotes a different sense of that word. However, since word senses are complex entities, without well-defined boundaries, we suggest to handle them less artificially, by representing them as fuzzy objects, where each word has its membership degree, which can be related to the confidence on using the word to denote the concept conveyed by the synset. We thus propose an approach to discover synsets from a synonymy network, ideally redundant and extracted from several broad-coverage sources. The more synonymy relations there are between two words, the higher the confidence on the semantic equivalence of at least one of their senses. The proposed approach was applied to a network extracted from three Portuguese dictionaries and resulted in a large set of fuzzy synsets. Besides describing this approach and illustrating its results, we rely on three evaluations – comparison against a handcrafted Portuguese thesaurus; comparison against the results of a previous approach with a similar goal; and manual evaluation – to believe that our outcomes are positive and that, in the future, they might my expanded by exploring additional synonymy sources.

Keywords

wordnets, synsets, fuzzy clustering, lexical-semantic network, synonyms, confidence, dictionaries

1 Introdução

Wordnets são bases de conhecimento léxico-semântico, inspiradas na Wordnet de Princeton (Fellbaum, 1998), a primeira, que definiu este modelo de recurso lexical. Uma wordnet agrupa as palavras de uma língua em conjuntos de sinónimos, normalmente chamados de *synsets*, que representam as possíveis lexicalizações de um conceito nessa língua. A ambiguidade lexical, ou seja, a possibilidade de usar a mesma palavra para transmitir diferentes significados, pode ser representada no modelo da wordnet através da presença da mesma palavra em diferentes synsets, relativos a cada um dos seus sentidos. Ao

mesmo tempo, um synset pode incluir um conjunto de palavras que partilhem o mesmo significado. No entanto, na realidade os sentidos não são objectos discretos, mas sim estruturas complexas, sem fronteiras bem definidas (Kilgarriff, 1996), ou seja, ainda que claramente útil ao processamento da língua, esta representação acaba por ser artificial.

Existem actualmente inúmeras wordnets para as mais variadas línguas (ver, por exemplo, Bond & Paik (2012)), e até línguas para as quais há mais de uma wordnet disponível. Para a língua portuguesa, existem pelo menos seis wordnets (ver Gonçalo Oliveira et al. (2015)), construídas por equipas independentes, com licenças diferentes, e seguindo abordagens distintas, cada uma com as suas vantagens e desvantagens. Por exemplo, relativamente a wordnets livres para esta língua, a OpenWN-PT (de Paiva et al., 2012) e a PULO (Simões & Guinovart, 2014) têm ainda uma cobertura limitada ao nível de lemas, sentidos e tipos de relação. No entanto, estão as duas alinhadas com a WordNet de Princeton e, indirectamente com outras wordnets. Isto não só trás benefícios ao nível do processamento multilingue, como permite complementar o conhecimento de cada um destes recursos com informação noutras wordnets (nomeadamente relações, definições ou exemplos). Por outro lado, a Onto.PT (Gonçalo Oliveira & Gomes, 2014) é maior que as anteriores, o que se deve essencialmente à exploração de vários recursos, criados de origem para o português, através da sua abordagem automática de construção, a ECO. Além disso, a Onto.PT abrange um leque de tipos de relação mais alargado que a maior parte das wordnets. Uma limitação, relacionada com a sua construção automática, é que ela não se encontra alinhada com nenhuma outra wordnet. Outra, será o facto da Onto.PT ser potencialmente menos fiável que as demais wordnets, nomeadamente daquelas cuja criação é completamente manual ou que, apesar de tirarem partido de abordagens semi-automáticas, têm uma integração de conteúdos mais controlada.

Para balancear a segunda limitação referida, pretendemos criar uma wordnet com uma cobertura comparável à da Onto.PT, mas onde sejam associadas uma ou várias medidas que transmitam a confiança em cada uma das decisões tomadas na sua criação, incluindo a associação de palavras em synsets ou a ligação de synsets através de uma relação semântica, ambas realizadas em passos da abordagem ECO. O resultado será uma wordnet de grande cobertura que, ao mesmo tempo, será suficientemente flexível para permitir ao utilizador escolher a porção que deseja utilizar, através da aplicação de um ponto de corte na confiança – a escolha por uma porção maior da wordnet englobará tendencialmente conteúdos com confianças mais baixas, enquanto que porções menores terão, em teoria, uma maior fiabilidade. As medidas de confiança podem ainda ser relevantes para a desambiguação do sentido das palavras (Navigli, 2009).

Apresentamos aqui o primeiro passo para a construção do novo recurso, nomeadamente a descoberta de grupos de sinónimos em que a pertença de cada palavra tem um valor associado, que deverá indicar a confiança relativamente à palavra transmitir o mesmo significado que as outras palavras no synset. Para calcular o valor da pertença, propomos tirar partido da redundância presente em redes de palavras relacionadas, obtidas a partir de diferentes fontes, nomeadamente dicionários e wordnets livres. No caso deste artigo, explorou-se para este fim a versão actual do CARTÃO (Gonçalo Oliveira et al., 2011), uma rede léxico-semântica extraída automaticamente a partir de três dicionários da língua portuguesa. No CARTÃO, as palavras estão relacionadas através de um conjunto de relações semânticas, ainda que os seus diferentes sentidos não sejam tratados. Sendo um synset um grupo de sinónimos, esta análise foca-se nas relações de sinonímia, ainda que não se descarte completamente a utilização de outros tipos. Assim, como numa rede de sinonímia as palavras estão ligadas através da relação de sinonímia, a identificação de aglomerados de palavras nestas redes (*clusters*) pode ser aproximada precisamente à descoberta de synsets.

Neste artigo, depois de descrevermos algum trabalho relacionado (secção 2), o que inclui a revisão de alguns algoritmos de *clustering* e de abordagens para descoberta de grupos de palavras relacionadas, propomos uma abordagem para a descoberta dos synsets difusos a partir de redes léxico-semânticas (secção 3). Apresentam-se depois os resultados da aplicação desta abordagem à rede CARTÃO, que resulta num conjunto de synsets difusos para o português (secção 4). Seguem-se alguns números relativos à avaliação dos resultados obtidos, automaticamente, contra os conteúdos de um tesauro referência, criado manualmente, e também através da classificação manual de pares de sinonímia. Os resultados obtidos são colocados lado-a-lado com aqueles obtidos através de uma abordagem anterior (Gonçalo Oliveira & Gomes, 2011) que tinha o mesmo objectivo e que também tinha sido aplicada ao CARTÃO (secção 5). Por fim, antes

de concluir, apresentam-se os resultados de uma nova experiência em que, para além de relações de sinonímia, as relações de hiperonímia também foram consideradas no cálculo das pertenças, o que levou a uma evolução positiva destes valores (secção 6).

2 Trabalho Relacionado

O principal objectivo do trabalho apresentado neste artigo é a identificação de agrupamentos (*clusters*) de sinónimos numa rede de palavras. Dadas as características da relação de sinonímia, estes *clusters* poderão depois ser aproximados a synsets. Para tal, pretende-se definir um algoritmo de *clustering* que, para calcular semelhanças, considere a estrutura da rede e, eventualmente, outras propriedades das palavras envolvidas (por exemplo, relações), que deverão ser tidas em conta no cálculo do valor das pertenças, a nossa confiança. A primeira parte desta secção descreve alguns dos algoritmos que ponderamos utilizar para atingir este objectivo. Na segunda parte, são apresentados alguns trabalhos em que abordagens de *clustering* foram aplicadas precisamente à descoberta de grupos de palavras sinónimas ou relacionadas, utilizadas para descrever conceitos.

2.1 Clustering em grafos

A tarefa de *clustering* tem como objectivo identificar, de forma automática e não supervisionada, agrupamentos de instâncias semelhantes, de acordo com um conjunto de dados a seu respeito e com uma métrica de semelhança sobre esses dados. Entre os vários algoritmos para esta tarefa (Xu & Wunsch, 2005), de acordo com o tipo de partição realizada, há três grandes grupos:

- *Clustering* hierárquico (*hierarchical clustering*): o resultado é uma partição hierárquica onde grupos de instâncias se organizam num estrutura em árvore, cuja raíz será um cluster com todas as instâncias e em que cada instância é uma folha;
- *Clustering* rígido (*hard clustering*): o resultado é uma partição rígida, em que cada instância está contida em um e um só *cluster*;
- *Clustering* difuso (*fuzzy clustering*): o resultado é uma partição em que a mesma instância pode pertencer a mais do que um *cluster*, com diferentes graus de pertença.

A nossa abordagem tem como requisito que o algoritmo actue sobre um grafo (de palavras), e que realize uma partição difusa, cujas pertenças sejam baseadas na confiança que há na associação das instâncias (palavras) aos *clusters* (synsets). De forma a escolher a abordagem a seguir na descoberta de synsets difusos, foi analisado um conjunto de algoritmos de *clustering*, que se apresentam de seguida.

O algoritmo Fuzzy C-Means (FCM) (Bezdek, 1981) é uma abordagem clássica para a descoberta de *clusters* difusos. É a variante difusa do algoritmo K-means (Hartigan & Wong, 1979) onde, dados k pontos aleatórios (centróides), classifica cada instância com a classe do centróide mais próximo. Este cálculo pode ser repetido por várias iterações, até haver convergência. No caso específico do FCM, cada instância pode pertencer a todos os clusters identificados, sendo o grau de pertença calculado com base na sua distância para os respectivos centróides.

Um tipo específico de algoritmos de *clustering* inclui aqueles que representam o domínio do problema como um grafo (ver Schaeffer (2007)), que será a forma óbvia de ver as redes de sinonímia. Ao contrário do FCM, em que é necessário indicar o número de *clusters* pretendidos e a sua posição inicial, nos algoritmos de clustering sobre grafos, o número de *clusters* vai depender essencialmente da estrutura do grafo. Olhando especificamente para aqueles que foram aplicados a problemas no âmbito do processamento de linguagem natural (PLN), destacamos o Markov Clustering (MCL) e o Chinese Whispers (CW), ambos baseados em passeios aleatórios pelo grafo (vulgo, *random walks*).

O MCL (van Dongen, 2000) parte da ideia que os caminhos aleatórios tendem a concentrar-se dentro de um mesmo subgrafo denso e não a saltar entre diferentes subgrafos através de ligações esparsas. O CW (Biemann, 2006) é uma variante do MCL que simplifica as operações do algoritmo anterior, sendo por isso mais eficiente. Inicialmente, para um grafo não direccionado com ou sem pesos, é atribuída uma classe distinta a cada nó. Depois, a cada iteração, os nós podem assumir a classe do vizinho que lhe transmitir maior força, o que se repete até haver estabilidade.

Outro algoritmo também aplicado a problemas de PLN é o Clustering by Committee (CBC, Lin & Pantel (2002)). Este algoritmo começa por encontrar conjuntos de instâncias designados por comités (*committees*), dispersos no espaço. Cada comité é constituído por instâncias que pertencem necessariamente a uma classe, que o comité acaba por definir. As restantes instâncias são de-

pois associadas aos comités mais próximos. Sempre que é realizada uma associação, são removidas todas as características comuns entre os comités e as instâncias que lhe foram associadas, o que permite que nas iterações seguintes essas instâncias possam ser associadas a outros comités.

Nesta análise, verificámos que nenhum dos algoritmos analisados ia ao encontro dos nossos objectivos. Por exemplo, apesar de ter uma utilização bastante generalizada, o FCM requer que o número de *clusters* seja um parâmetro dado inicialmente, quando o que pretendemos é que esta decisão seja tomada de forma automática pelo algoritmo. Para além disso, a posição inicial dos centróides é aleatória, o que torna o algoritmo não determinístico.

Dado que o nosso domínio são redes lexicais, faria sentido optar por um algoritmo que actue sobre grafos e que tire partido da sua estrutura. No entanto, nenhum dos dois algoritmos analisados dentro desta categoria descobre clusters difusos, e nem sequer permite que uma instância pertença a mais do que um cluster. Mesmo quando há nós instáveis, uma decisão acaba por ser tomada relativamente à sua pertença a um cluster que, devido aos caminhos aleatórios, pode não ser sempre o mesmo em diferentes iterações. Ou seja, nem o MCL nem o CW são determinísticos, ainda que o problema seja minimizado em grafos pesados e de maior dimensões (Biemann, 2006). Dada a sua relevância para este trabalho, acrescentamos ainda que, sob o ponto de vista da complexidade temporal, o CW é apresentado como uma variante mais eficiente do MCL (Biemann, 2006), precisamente por ser mais agressivo e considerar apenas o vizinho que transmite mais força e não os restantes. Isto reflecte-se numa complexidade temporal de $\mathcal{O}(s.|E|)$ para o CW, enquanto que para o MCL é $\mathcal{O}(s.|V|^2)$, em que s é o número de iterações, $|E|$ é o número de arcos e $|V|$ o número de vértices do grafo.

Sobre o CBC, que será determinístico, não foi desenhado para operar sobre grafos, ainda que uma adaptação seja possível. No entanto, acaba por sofrer de outros problemas semelhantes. Além disso, apesar de permitir a associação da mesma instância a vários *clusters*, após a sua associação a um comité, são removidas da instância todas as características em comum com esse comité, sendo a verdadeira semelhança com outros comités corrompida.

Apesar de nenhum destes algoritmos satisfazer os nossos requisitos, a abordagem que propomos na secção 3 acaba por combinar características dos algoritmos aqui revistos.

2.2 Descoberta de grupos de palavras

A tarefa de desambiguação do sentido das palavras (em inglês, *word sense disambiguation*) (Navigli, 2009) tem como objectivo associar a ocorrência de uma palavra, em contexto, ao seu sentido mais adequado, dentro de um repositório de sentidos (por exemplo, um dicionário). Para o inglês, é comum utilizar-se a WordNet (Fellbaum, 1998) ou, de forma a cobrir mais conhecimento sobre o mundo, uma extensão deste, como a BabelNet (Navigli & Ponzetto, 2012).

Uma tarefa próxima, é a indução dos sentidos das palavras (em inglês, *word sense induction*) (Nasiruddin, 2013). Aí, não existe um repositório e os sentidos são descobertos de forma normalmente não supervisionada, através da análise de semelhanças entre palavras, tendo em conta os contextos em que ocorrem e as relações em que estão envolvidas.

O nosso trabalho está ligado à indução do sentido das palavras, porque queremos identificar precisamente os sentidos possíveis de cada palavra e os sinónimos de cada um, de forma automática, recorrendo simplesmente a uma rede léxico-semântica, onde as palavras são identificadas apenas pela sua ortografia e classe gramatical e não existe divisão entre sentidos. Há alguma relação com o trabalho de Lin & Pantel (2002), em que o algoritmo CBC foi usado para descobrir conceitos, representados através de palavras que co-ocorrem frequentemente em texto e partilham um conjunto de relações sintácticas. Por isso, estes agrupamentos vão para além de grupos de sinónimos. Considere-se por exemplo o conceito com melhor qualidade descoberto por Lin & Pantel (2002), "arma de fogo", que inclui as seguintes palavras: *handgun, revolver, shotgun, pistol, rifle, machine gun, sawed-off shotgun, submachine gun, gun, automatic pistol, ...* Para além do CBC, o algoritmo MCL foi também utilizado para detectar ambiguidades (Dorow et al., 2005). Mais precisamente, ao extrair uma rede de co-ocorrências a partir de um texto, as palavras ambíguas correspondem a vértices mais instáveis, ou seja, que ligam dois subgrafos densos.

Enquanto que os trabalhos anteriores exploram texto corrido, há outros que, tal como nós, usam redes de sinonímia extraídas precisamente de dicionários para identificar grupos de palavras sinónimas. Por exemplo, Gfeller et al. (2005) propõem uma forma de solucionar uma limitação do algoritmo MCL: não permitir que uma palavra seja incluída em mais do que um cluster. Para tal, o MCL é executado várias vezes, com ruído estocástico aleatório, de forma a identificar em que diferentes *clusters* os vértices mais instáveis

da rede aparecem. Estes vértices corresponderão, mais uma vez, a palavras ambíguas ou que poderão necessitar de ser desambiguadas. Um procedimento inspirado no anterior foi também aplicado ao português, na descoberta automática de synsets (Gonçalo Oliveira & Gomes, 2010). Contudo, este procedimento, que poderá aumenta o não-determinismo do MCL, resultou numa maioria de synsets demasiado grandes para que tivesse utilidade efectiva.

A ideia de utilizar conjuntos difusos para representar conceitos também não é nova. A nosso ver, ela vai ao encontro da ideia de que os sentidos das palavras não têm fronteiras muito bem definidas (Kilgarriff, 1996). Neste âmbito, Velldal (2005) apresenta uma abordagem para descobrir, a partir de texto corrido, conjuntos de palavras que podem ajudar a descrever conceitos, e em que as suas semelhanças contextuais são usadas como graus de pertença. O resultado é que, dada uma palavra (por exemplo, *cavalo*), é possível observar diferentes conjuntos difusos, cada um correspondente a um dos seus possíveis sentidos (por exemplo, meio de transporte – *carro (0.97), autocarro (0.80), barco (0.72), ...* – ou animal – *pássaro (0.86), cão (0.83), gato (0.80), ...*). Há também quem tenha atribuído graus de pertença de palavras a synsets, com base em vários julgamentos humanos (Borin & Forsberg, 2010).

No que diz respeito à criação de uma wordnet com medidas de confiança associadas, que é o nosso objectivo a longo prazo, existe para a língua inglesa trabalho na extensão da WordNet para outros domínios através de associações difusas (Araúz et al., 2012). Isto inclui não só um grau de pertença das palavras a synsets, mas também um valor difuso para o estabelecimento de relações entre synsets.

Num trabalho anterior, aplicamos uma abordagem simplista na descoberta de synsets difusos para o português, também a partir de redes de sinonímia extraídas de dicionários (Gonçalo Oliveira & Gomes, 2011). O algoritmo aplicado assume que cada palavra é um *cluster* potencial, que pode atrair nós semelhantes. Para obter as pertenças, é calculado o cosseno entre cada palavra e cada uma das outras, representadas pelo seu vector na matriz de adjacências da rede, que tem 1 nas adjacências e 0 nas restantes palavras. Esta foi a primeira abordagem para a descoberta automática de synsets difusos para o português que, contudo, originou mais uma vez, em média, synsets demasiado grandes, cuja utilização seria impraticável, pelo menos sem a aplicação de um ponto de corte, que se tornou obrigatório, e cujas pertenças nem sempre faziam muito sentido.

Após analisar melhor a abordagem, identificamos uma das causas do último problema, que seria a utilização dos vectores de adjacência completos, ao calcular o cosseno. Estando perante uma matriz esparsa, a maior parte das entradas é 0, ou seja, a pertença de palavras com muitas ligações (muito ambíguas ou com muitos sinónimos) é penalizada perante as outras, por as primeiras terem menos entradas nulas. Para além disso, apesar da abordagem anterior permitir a exploração de várias fontes de sinoníma, ela acabava por não explorar suficientemente a redundância de informação para reforçar a decisões tomadas. Um dos objectivos da abordagem proposta neste artigo é também melhorar o trabalho anterior. Assim, para além da realização de avaliações automática e manual, sempre que possível, foi feita uma comparação com os resultados obtidos anteriormente.

3 Abordagem Proposta

Como nenhum dos algoritmos revistos vai ao encontro dos nossos requisitos, propomos uma abordagem que combina características de mais do que um algoritmo. Para descobrir um conjunto de synsets difusos a partir de uma rede léxico-semântica, a abordagem proposta tem dois passos principais:

1. Identificação de um conjunto de centróides, onde as palavras já têm uma ligação forte e partilham semelhanças;

2. Cálculo dos graus de pertença, com base na proximidade de cada palavra aos centróides.

No nosso caso, os centróides são nada mais nada menos que *clusters* base, identificados a partir da estrutura do grafo e onde não há sobreposição. De certa forma, podem ser vistos como uma estrutura inicial, tal como os comités no CBC, que será numa segunda fase aumentada. Para a sua identificação, contudo, deve ser utilizado um algoritmo eficiente que tire partido da estrutura do grafo, tal como o CW.

No segundo passo, os graus de pertença de cada palavra são calculados com base na semelhança entre as características (palavras relacionadas) da palavra que são relevantes para o *centróide* e as palavras do próprio centróide, o que de certa forma se assemelha ao cálculo das pertenças no FCM. No entanto, não será necessário realizar novas iterações, precisamente porque cada *centróide* já incluirá palavras com um elevado grau de proximidade.

Formalizando, a abordagem proposta é aplicada a uma rede de sinonímia $G = (P, R)$, onde P é o conjunto de palavras e R é o conjunto de pares de sinonímia. A rede G pode ser representada através de uma matriz de adjacências $A(|P| \times |P|)$, onde $A_{ij} = \omega_{ij}$, um peso que reflecte o número de vezes que um par de sinónimos, $R(P_i, P_j)$, ocorre nas fontes utilizadas (dicionários, por exemplo). O peso máximo, m, é portanto uma constante, igual ao número de fontes utilizadas.

No primeiro passo, o algoritmo de *clustering* aplicado resulta num conjunto de clusters centróide C. No segundo, o valor de pertença da palavra P_i ao centróide C_k, $\mu(P_i, C_k)$, é calculado através da equação 1, onde T é o conjunto de palavras relevantes para o cálculo, ou seja, todas as palavras do centróide C_k e a palavra P_i, que pode ou não estar no centróide (ver equações 2). A multiplicação do denominador por m serve apenas para normalizar o valor da pertença no intervalo $[0-1]$. No final, o número de synsets difusos é igual ao número de clusters base.

$$\mu(P_i, C_k) = \frac{\sum_{j=0}^{|C_k|} A_{i[C_kj]}}{m \times |T|} \quad (1)$$

$$\begin{aligned} &T = \{C_k \cup P_i\}, \text{ou seja} \\ &|T| = \{|C_k|, P_i \in C_k\} \vee \{|C_k| + 1, P_i \notin C_k\} \end{aligned} \quad (2)$$

A abordagem é ilustrada com auxílio do grafo na figura 1, centrado na palavra *banana*. Em português europeu, esta palavra tanto pode ser o nome de uma fruta, como pode ter o sentido figurado de uma pessoa sem iniciativa. Suponha-se que o grafo é extraído a partir de três dicionários ($m = 3$) e que o algoritmo CW identifica os dois *clusters* centróide representados na tabela 1.

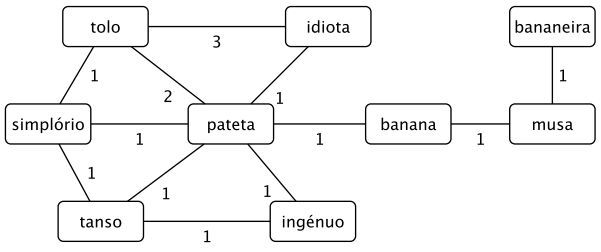

Figura 1: Rede de sinonímia com palavras e pesos das ligações.

Para calcular o valor de pertença de *banana* ao centróide C_A, devem ser consideradas as ligações às palavras *musa* e *bananeira*, ou seja, apenas 1. Este número é dividido por $3 \times |T|$, em que T incluir as palavras relevantes, $T = \{musa, bananeira, banana\}$. Portanto,

C_A	*musa*, *bananeira*
C_B	*banana*, *pateta*, *idiota*, *tolo*, *simplório*, *tanso*, *ingénuo*

Tabela 1: Centróides descobertos a partir da rede da figura 1, com o algoritmo Chinese Whispers.

$\mu(banana, C_A) = \frac{1}{9}$. Para o cálculo da pertença da palavra *banana* ao centróide C_B, as características relevantes são o número de ligações com todas as palavras do centróide C_B, apenas 1, para a palavra *pateta*. Considera-se ainda que cada palavra tem o número máximo de "ligações" a si própria, por isso, neste caso, como $banana \in C_B$, soma-se 3 ao número de ligações relevantes. Ou seja, o numerador será 4, e assim, $\mu(banana, C_B) = \frac{4}{21}$.

C'_A	*bananeira* (0.666), *musa* (0.666), *banana* (0.111)
C'_B	*pateta* (0.476), *tolo* (0.428), *idiota* (0.333), *simplório* (0.285), *tanso* (0.285), *ingénuo* (0.285), *banana* (0.190), *musa* (0.041)

Tabela 2: Synsets difusos com pertenças calculadas com base nos clusters discretos da tabela 1

4 Descoberta de synsets difusos para o português

Esta secção apresenta os resultados da aplicação da abordagem proposta à rede léxico-semântica CARTÃO, que se começa por descrever, seguida de uma visão numérica dos resultados e, por fim, de exemplos ilustrativos, com alguns dos synsets difusos obtidos.

4.1 Rede Léxico-Semântica

A rede léxico-semântica utilizada para a descoberta de synsets difusos foi o CARTÃO (Gonçalo Oliveira et al., 2011), disponível gratuitamente, e extraída de forma automática a partir de três dicionários da língua portuguesa, com base em padrões textuais nas suas definições. Para ajudar a caracterizar esta rede, a tabela 3 apresenta algumas das suas propriedades numéricas, mais propriamente para os sub-grafos de sinonímia entre substantivos (N), verbos (V), adjetivos (Adj) e advérbios (Adv).

Para cada sub-grafo, é indicado o número de vértices (palavras, $|P|$) e arestas distintas (relações de sinonímia, $|R|$), o grau médio dos vértices ($\overline{deg}(P)$) – equação 3 calcula o grau de um vértice – o coeficiente médio de *clustering* ($\overline{CC}$) – equação 4 calcula este coeficiente para

um vértice, sendo que $viz(P_i)$ representa o conjunto dos vizinhos do vértice P_i – o número de componentes conectadas[1] ($|Comp|$), e o número de palavras da componente maior ($|P|_{mc}$).

$$deg(P_i) = |R(P_i, P_j)| : P_i, P_j \in P \quad (3)$$

$$CC(P_i) = \frac{2.|R(P_j, P_k)|}{|viz(P_i)|.(|viz(P_i)| - 1)} : P_j, P_k \in viz(P_i) \quad (4)$$

Tal como outros investigadores (por exemplo, Gfeller et al. (2005)), também nos apercebemos que estes sub-grafos, extraídos de dicionários, são constituídos por uma grande componente, e várias pequenas. Além disso, os coeficientes de *clustering* são comparáveis aos de outras redes de pequeno mundo (em inglês, *small-world networks*), em que a distância média entre dois vértices é curta. Comparando os três sub-grafos, o sub-grafo dos verbos possui um grau médio mais elevado, o que significa que os verbos terão mais sinónimos e/ou serão mais ambíguos e/ou vagos. Também se observa que o sub-grafo de advérbios é significativamente mais pequeno que os demais, por isso acabou por não ser utilizado nas experiências apresentadas nas próximas secções.

4.2 Propriedades dos synsets descobertos

Ao correr o algoritmo proposto no CARTÃO, obtemos um conjunto com quase 15 mil synsets difusos C', a que chamamos CLIP 2.0, com as propriedades apresentadas na tabela 4 para cada categoria gramatical, nomeadamente: número de palavras (#pals), média de sentidos por palavra ($\overline{sents}$), palavra com mais sentidos (max(#sents)), total de synsets (#synsets), média de palavras por synset ($\overline{|synset|}$), synsets com apenas duas palavras ($|synset| = 2$), synsets com mais de 25 palavras ($|synset| > 25$), e tamanho do maior synset (max($|synset|$)). Na mesma tabela, incluem-se as mesmas propriedades para a nossa abordagem anterior, onde foi utilizada uma versão anterior do CARTÃO (originalmente Padawik (Gonçalo Oliveira & Gomes, 2011), depois rebaptizado como CLIP (Gonçalo Oliveira, 2013)), e onde, durante a geração original, foi aplicado um ponto de corte sobre as pertenças (θ) de 0,01. Ainda na mesma tabela, apresentam-se as propriedades dos synsets no tesauro TeP 2.0 (Maziero et al., 2008), criado manualmente para o português do Brasil.

Quando comparado com o CLIP 1.0, parece haver menos ruído, mesmo sem a aplicação de nenhum ponto de corte no CLIP 2.0. Isto porque existem menos synsets, em média mais pequenos para os nomes e adjetivos, e de tamanho comparável para os verbos. As palavras são também menos ambíguas. No TeP, o número médio de palavras por synset é mais baixo, tal como o número médio de sentidos por palavra, o que já era esperado, não só pelo TeP ter sido criado manualmente, mas também devido à nossa abordagem difusa, e pelo maior grau de cobertura do nosso tesauro. Recordamos, no entanto, que pode ser aplicado um ponto de corte às pertenças dos synsets difusos, de modo que estes fiquem pequenos e, tendencialmente, mais confiáveis. Por outro lado, no TeP o número de synsets de verbos e adjetivos é mais do dobro, e ligeiramente mais baixo para os substantivos. No entanto, os nossos synsets cobrem quase o dobro das palavras do TeP (cerca de 70 mil contra 40 mil), mais propriamente um número próximo de verbos, ligeiramente superior de adjectivos, e mais do dobro de substantivos. O número inferior de synsets de verbos e adjetivos pode, por um lado, indicar que o CLIP 2.0 não cobre tantos sentidos quanto o TeP mas, por outro, que o CLIP 2.0 agrupará significados mais próximos, que muitas vezes nem fará sentido separar. Esta capacidade está relacionada com a chamada "ambiguação" do sentido das palavras (Dolan, 1994).

4.3 Alguns resultados

A tabela 5 ilustra os resultados obtidos através de uma selecção manual de palavras polissémicas da língua portuguesa e alguns dos synsets difusos que as incluem, organizados de acordo com o conceito que transmitem (frequentemente clarificado pela palavra com maior pertença) e onde as palavras são apresentadas por ordem decrescente do grau de pertença. Numa observação global, tanto a constituição dos synsets como os graus de pertença parecem fazer sentido.

5 Avaliação

Nesta secção, os resultados obtidos são avaliados, primeiro através da sua confirmação no TeP, aqui usado como recurso dourado por ter sido criado manualmente e, segundo, manualmente. Os resultados de cada avaliação são comparados com os mesmos resultados obtidos para o CLIP 1.0.

[1] Uma componente de um grafo é um subgrafo no qual todos os pares de vértices estão ligados através de pelo menos um caminho, sem que estejam ligados a mais nenhum vértice do grafo.

POS	$\|P\|$	$\|R\|$	$\overline{deg}$(G)	$\overline{CC}$	$\|Comp\|$	$\|P_{mc}\|$
N	43,724	65,127	2.98	0.21	5,812	28,734
V	10,380	26,266	5.06	0.25	362	9,549
Adj	31,014	17,368	3.57	0.23	2,049	12,343
Adv	1,271	1,296	2.04	0.18	160	819

Tabela 3: Propriedades dos sub-grafos de cada categoria gramatical na rede CARTÃO.

	Cat	**Palavras**			**Synsets**				
		#pals	$\overline{sents}$	max(#sents)	#synsets	$\overline{\|synset\|}$	$\|synset\| = 2$	$\|synset\| > 25$	max($\|synset\|$)
	N	43.721	1,92	42	9.881	8,49	4.147	632	554
CLIP 2.0	**V**	10.380	3,15	54	1.438	22,76	289	370	500
	Adj	17.368	2,28	44	3.571	11,.07	1.530	367	322
	N	39.354	7,78	46	20.102	15,23	3,885	3,756	109
CLIP 1.0	**V**	11.502	14,31	42	7.775	21,17	307	2,411	89
($\theta = 0,01$)	**Adj**	15.260	10,36	43	8.896	17,77	1,326	2,157	109
	N	17.158	1,71	21	8.254	3,56	3.079	0	21
TeP 2.0	**V**	10.827	2,08	41	3.978	5,67	939	48	53
	Adj	14.586	1,46	19	6.066	3,50	3.033	19	43

Tabela 4: Propriedades numéricas dos synsets.

Além disso, procuramos validar os graus de pertença através da observação do seu valor para pares de sinonímia confirmados/correctos e não confirmados/incorrectos.

5.1 Comparação com um tesauro criado manualmente

Como o TeP é um tesauro criado manualmente para o português, temos alguma confiança nos seus conteúdos. Para além disso, foi desenvolvido de forma completamente independente do CARTÃO. Daí ter sido o TeP a nossa primeira opção para verificar a qualidade dos synsets descobertos.

Para facilitar a comparação, transformou-se cada conjunto de synsets descobertos num conjunto de pares de sinonímia, que seriam depois confirmados no TeP. Considera-se que um par de sinonímia, $R(w_a, w_b)$, é um conjunto de duas palavras que pertencem ao mesmo synset C_x, ou seja, $R(w_a, w_b) \rightarrow \exists C_x : w_a \in C_x \wedge w_b \in C_x$. Então, para cada par presente nos tesauros descobertos, verificou-se se existia pelo menos um synset no TeP que contivesse as duas palavras. A tabela 6 apresenta a proporção de pares confirmados para os synsets de cada categoria gramatical, não só para os resultados da abordagem actual, mas também para o CLIP 1.0.

Como, considerando todos os pares, a proporção de pares confirmados é muito baixa, na mesma tabela apresenta-se a evolução desse número para diferentes pontos de corte aplicados às pertenças (θ) – ao aplicar um ponto de corte, descartam-se de cada synset todas as palavras cuja pertença é inferior ao valor do corte. Nomeadamente, para os pontos de corte 0,105, 0,225 e 0,510, é apresentado: o número de pares do tesauro (Total); o número de pares com ambas as palavras no TeP (PalavrasNoTeP) e respectiva proporção relativamente ao número total; e o número de pares confirmados no TeP (ParNoTeP) e respectiva proporção relativamente ao número de pares com palavras cobertas. Ainda a este respeito, a figura 2 mostra, para o CLIP 1.0 e 2.0, a evolução das proporções PalavrasNoTeP (Palavras) e ParNoTeP (Pares). Para referência, o TeP inclui 51.533 pares de substantivos, 89.456 de verbos, e 51.645 de adjectivos.

É possível verificar que, tal como esperado numa medida de confiança, a proporção de pares confirmados aumenta para pontos de corte mais elevados, quer para o CLIP 1.0 como para o CLIP 2.0. No entanto, para cortes superiores a 0,6, mais de 90% dos pares do CLIP 2.0 são confirmados, enquanto que o CLIP 1.0 nunca chega a 80% de pares confirmados. Curioso também é a proporção de pares com ambas as palavras no TeP, que desce de forma mais consistente no CLIP 1.0 do que no 2.0. Aliás, a partir de um certo ponto, o CLIP 2.0 modifica mesmo a sua tendência e o número de pares desse tipo deixa de descer. Ou seja, se tomarmos o TeP como referência absoluta, estes números levam-nos a crer que a abordagem aqui proposta não só resulta em synsets mais coerentes, mas a uma medida de confiança mais fiel.

No entanto, apesar de mais confiável que um recurso criado de forma automática, o TeP está longe de ser uma referência absoluta. Aliás, TeP e CARTÃO são recursos, até certo ponto, complementares, não só relativamente a lemas, mas também a pares de sinonímia (veja-se a comparação realizada em Gonçalo Oliveira et al. (2011)). Para além de ter sido criado de forma manual, o TeP foca-se no português do Bra-

Palavra	Conceito	Synsets difusos
pasta	mistura	*mistura*(0.333), *amálgama*(0.127), *mescla*(0.111), *matalotagem*(0.079), *anguzada*(0.079), *comistão*(0.079), *misto*(0.079), *landoque*(0.079), *salsada*(0.0758), *confusão*(0.0758), *enovelamento*(0.063), *cacharolete*(0.063), *macedónia*(0.063), *mexedura*(0.063), *caldeação*(0.063), *mixagem*(0.063), *pasta*(0.063), *angu*(0.063), *amalgamação*(0.063), *comistura*(0.063), *impurezas*(0.063), *mistão*(0.063), *estricote*(0.063), *fusão*(0.045), *temperamento*(0.03), *pot-pourri*(0.015), *imisção*(0.015), *cocktail*(0.015), *ensalsada*(0.015), *envolta*(0.015), *agrupamento*(0.015), *baralha*(0.015), *marinhagem*(0.015), *salgalhada*(0.015), *misturada*(0.015), *miscelânea*(0.015), *têmpera*(0.015), *imperfeição*(0.015), *conjunto*(0.015), *combinação*(0.015), *logro*(0.015), ...
	dinheiro	*dinheiro*(0.28), *bufunfa*(0.069), *caroço*(0.053), *tutu*(0.042), *pataco*(0.037), *bagalhoça*(0.037), *guines*(0.037), *cobre*(0.032), *pecúnia*(0.032), *gaita*(0.032), *cacique*(0.032), *pílula*(0.026), *morubixaba*(0.026), *pila*(0.026), *cacau*(0.026), *arame*(0.026), *calombo*(0.026), *patacaria*(0.026), *gimbo*(0.026), *maco*(0.026), *bubão*(0.026), *chelpa*(0.026), *roço*(0.026), *levação*(0.026), *íngua*(0.026), *vénus*(0.021), *verdinha*(0.021), *mondrongo*(0.021), *pírula*(0.021), *dindim*(0.021), *trocado*(0.021), *curaca*(0.021), *pataca*(0.021), *massaroca*(0.021), *bagalho*(0.021), *carcanhol*(0.021), *pilim*(0.021), *encórdio*(0.021), *teca*(0.021), *coronel*(0.021), *matambira*(0.021), *mussuruco*(0.021), *cinco-réis*(0.021), *metal*(0.021), *cunques*(0.021), *janda-cruz*(0.021), *boro*(0.021), *cum-quibus*(0.021), *bilhestres*(0.021), *calique*(0.021), *parrolo*(0.021), *zerzulho*(0.021), *caronha*(0.021), *nhurro*(0.021), *baguines*(0.021), *pecuniária*(0.021), *pecunia*(0.021), *marcaureles*(0.021), *china*(0.021), *fanfa*(0.021), *dieiro*(0.021), *influente*(0.021), *guino*(0.021), *grana*(0.02), *tostão*(0.01), *riqueza*(0.01), ...
planta	vegetal	*vegetal*(0.667), *plantas*(0.667), *planta*(0.111)
	plano	*plano*(0.379), *projecto*(0.23), *tenção*(0.207), *desígnio*(0.207), *traçado*(0.161), *propósito*(0.161), *intenção*(0.149), *pressuposto*(0.138), *intento*(0.138), *prospecto*(0.126), *desenho*(0.126), *planta*(0.126), *programa*(0.115), *traça*(0.0.115), *mente*(0.092), *risco*(0.089), *resolução*(0.089), *prospeto*(0.08), *arquitectura*(0.08), *ideia*(0.078), *pressuposição*(0.069), *traçamento*(0.069), *prepósito*(0.069), *presuposto*(0.069), *intuito*(0.067), *vista*(0.067), *alçado*(0.057), *planificação*(0.057), *design*(0.057), *pranta*(0.057), *esboço*(0.055), *planejamento*(0.045), *fundição*(0.046), *gizamento*(0.046), *caruru*(0.046), *aspecto*(0.044), *medida*(0.044), *fim*(0.044), *vontade*(0.044), *desejo*(0.044), ...
sede	centro	*centro*(0.6), *núcleo*(0.4), *sensório*(0.333), *foco*(0.333), *club*(0.267), *sede*(0.222), *âmago*(0.222), *meio*(0.167), *coração*(0.167), *metrópole*(0.111), *escol*(0.056), *pólo*(0.056), *clube*(0.056), *umbigo*(0.056), *cérebro*(0.056), *fundo*(0.056), *gema*(0.056), *cadeira*(0.056), *casco*(0.056), *aglomeração*(0.056), *grupo*(0.056), *empório*(0.056), *essência*(0.056), *casino*(0.056), ...
	secura	*sede*(0.429), *secura*(0.333), *sequidão*(0.286), *seda*(0.238), *cerdas*(0.19), *sieda*(0.19), *seeda*(0.19), *aridez*(0.083), *centro*(0.083), *cerda*(0.042), *foco*(0.042), *impassibilidade*(0.042), *mortalha*(0.042), *cadeira*(0.042), *núcleo*(0.042), *diocese*(0.042), *ambição*(0.042), *impaciência*(0.042), *banco*(0.042), *apetite*(0.042), *avidez*(0.042), *ânsia*(0.042), *insensibilidade*(0.042), *capital*(0.042), *polidipsia*(0.042), *luxo*(0.042), *frieza*(0.042), *seta*(0.042), *magreza*(0.042)
	impaciência	*impaciência*(0.533), *frenesi*(0.467), *rabujice*(0.267), *despaciência*(0.267), *farnesia*(0.267), *inquietação*(0.222), *sofreguidão*(0.167), *pressa*(0.167), *desespero*(0.111), *nervosismo*(0.111), *ansiedade*(0.111), *exaltação*(0.111), *cócegas*(0.111), *freima*(0.111), *freimaço*(0.056), *formigueiro*(0.056), *precipitação*(0.056), *agastamento*(0.056), *impertinência*(0.056), *sofreguice*(0.056), *sede*(0.056), *inguinação*(0.056), *ira*(0.056), *furor*(0.056), *excitação*(0.056), *prurido*(0.056), *fúria*(0.056),...
verde	cor verde	*verde*(0.274), *virente*(0.137), *verdejante*(0.137), *relvoso*(0.118), *gramíneo*(0.098), *esmeraldino*(0.098), *prásino*(0.098), *desassazonado*(0.098), *viridente*(0.098), *ervoso*(0.098), *verdoso*(0.098), *ecológico*(0.078), *dessazonado*(0.078), *graminoso*(0.078), *viridante*(0.078), *herboso*(0.078), *porráceo*(0.078), *viçoso*(0.055), *inoportuno*(0.037), *fresco*(0.037), *esverdeado*(0.037), ...
	amador	*inexperiente*(0.917), *noviço*(0.067), *novato*(0.067), *inexperto*(0.417), *novel*(0.267), *ingénuo*(0.267), *inocente*(0.267), *principiante*(0.133), *novo*(0.133), *viçoso*(0.133), *matumbo*(0.067), *incompetente*(0.067), *amador*(0.067), *verde*(0.067), *moço*(0.067), *bisonho*(0.067), *ingênuo*(0.067), ...
limpar	tornar limpo	*limpar*(0.262), *purificar*(0.126), *enxugar*(0.098), *expurgar*(0.066), *mundificar*(0.06), *desinfectar*(0.06), *purgar*(0.055), *secar*(0.055), *depurar*(0.049), *mirrar*(0.049), *lavar*(0.049), *descontaminar*(0.044), *despoluir*(0.038), *desinçar*(0.038), *virginizar*(0.038), *esburgar*(0.038), *dessecar*(0.038), *assear*(0.038), *luir*(0.038), *varrer*(0.038), *esmirrar*(0.033), *desensopar*(0.033), *desenxovalhar*(0.033), *absterger*(0.033), *tamisar*(0.027), *virginalizar*(0.027), *desparasitar*(0.027), *vassourar*(0.027), *desenxamear*(0.027), *emundar*(0.027), *desecar*(0.027), *desempestar*(0.027), *desenodoar*(0.027), *desenfarruscar*(0.027), *perlavar*(0.027), *detergir*(0.027), *achicar*(0.027), ...
	podar	*desramar*(0.778), *escamondar*(0.556), *mondar*(0.556), *limpar*(0.25), *petelar*(0.083), *desgalhar*(0.083), *derramar*(0.083), *alveitarar*(0.083), *carpir*(0.083), *capinar*(0.083), *corrigir*(0.083)
	peneirar	*joeirar*(0.533), *escrivar*(0.333), *utar*(0.267), *acrivar*(0.267), *outar*(0.267), *peneirar*(0.111), *limpar*(0.111), *tamisar*(0.056), *crivar*(0.056), *cirandar*(0.056), *brocar*(0.056)
	roubar	*ripar*(0.533), *bifar*(0.467), *ripançar*(0.4), *surrupiar*(0.267), *palmar*(0.267), *surripiar*(0.222), *furtar*(0.111), *limpar*(0.111), *pifar*(0.056), *raspar*(0.056), *arrancar*(0.056), *puxar*(0.056)
estimar	apreciar	*apreciar*(0.444), *valorar*(0.333), *estimar*(0.333), *avaliar*(0.333), *cotar*(0.222), *valorizar*(0.222), *admirar*(0.222), *ponderar*(0.19), *considerar*(0.143), *amar*(0.095), *discernir*(0.095), *julgar*(0.095), *equacionar*(0.048), *ustir*(0.048), *trutinar*(0.048), *estranhar*(0.048), *qualificar*(0.048), *apreçar*(0.048), *gostar*(0.048), *desfrutar*(0.048), *adular*(0.048), *conhecer*(0.048), *recensear*(0.048), *aquilatar*(0.048), *numerar*(0.048), *desejar*(0.048), *sentir*(0.048), *reputar*(0.048), ...
	avaliar	*avaliar*(0.625), *aquilatar*(0.375), *quilatar*(0.292), *apreçar*(0.292), *equacionar*(0.208), *almotaçar*(0.208), *conceituar*(0.208), *aderar*(0.208), *julgar*(0.185), *estimar*(0.148), *apreciar*(0.148), *pesar*(0.111), *conhecer*(0.111), *louvar*(0.111), *calcular*(0.111), *ajuizar*(0.074), *quantiar*(0.074), *aferir*(0.074), *computar*(0.074), *aperfeiçoar*(0.074), *ponderar*(0.074), *reputar*(0.074), *cotar*(0.037), *valorar*(0.037), *arbitrar*(0.037), *mensurar*(0.037), *qualificar*(0.037), *contrastar*(0.037), *orçar*(0.037), *montar*(0.037), *taxar*(0.037), *apurar*(0.037), *discernir*(0.037), *examinar*(0.037), *tomar*(0.037)

Tabela 5: Synsets difusos de palavras polissémicas no CLIP 2.0.

sil, e acaba por não cobrir várias palavras da língua portuguesa, nem alguns sentidos das palavras que inclui, principalmente aqueles menos comuns. Esta será mesmo a principal razão para a proporção de pares confirmados ser muito baixa quando não é aplicado qualquer ponto de corte.

5.2 Avaliação manual

Devido às limitações já referidas do TeP, decidimos efectuar uma avaliação adicional, desta vez manual, seguindo as mesmas regras que na avaliação feita ao CLIP 1.0, detalhada em Gonçalo Oliveira & Gomes (2011) e Gonçalo Oliveira (2013). Mais precisamente, esta avaliação passou pelas seguintes fases:

1. Remoção (automática) dos synsets de todas as palavras que não ocorrem nos corpos acessíveis a partir do serviço AC/DC (Santos & Bick, 2000);
2. Selecção (automática) apenas dos synsets onde todas as palavras têm uma frequência superior a 100, nos mesmos corpos;
3. Escolha (automática), de n pares de palavras, sendo que cada par tem duas palavras provenientes do mesmo synset;
4. Classificação manual de cada par como sinónimos (correto) ou não (incorreto).

Os dois primeiros passos foram feitos para tornar a avaliação mais rápida e focada em palavras conhecidas, por serem frequentes. No terceiro passo, optamos por gerar três conjuntos aleatórios: 150 pares de nomes, 150 pares de verbos e 150 pares de adjetivos. No quarto passo, cada par foi classificado por dois avaliadores humanos, de forma independente, a quem foi sugerido a consulta de dicionários na rede, em caso de dúvida. A tabela 7 apresenta os resultados obtidos por avaliador e a sua concordância κ, assim como os resultados da avaliação manual do CLIP 1.0, mas apenas para os nomes, tal como apresentada em Gonçalo Oliveira (2013). Apresentam-se ainda as médias das medidas de pertença dos pares classificados como correctos por ambos os avaliadores ($\overline{\mu_c}$), pares onde não houve concordância entre avaliadores ($\overline{\mu_d}$), e pares classificados como incorrectos por ambos ($\overline{\mu_i}$). Não nos foi possível recuperar os dados de avaliação manual do CLIP 1.0, o que não nos permite fazer a análise dos graus de pertença para a abordagem anterior.

Embora exista margem para melhorias, a proporção de pares correctos foi, uma vez mais, superior ao mesmo valor no CLIP 1.0. Nota-se que os verbos são a categoria com mais pares incorretos, provavelmente por ser também o sub-grafo com maior grau médio, ou seja, maior número de ligações por vértice (ver tabela 3), o que dará origem a mais confusão.

A média das pertenças de palavras em pares classificados como corretos ($\overline{\mu_c}$), incorretos ($\overline{\mu_{!c}}$) e discordantes ($\overline{\mu_d}$) têm um comportamento consistente para todas as categorias. Ou seja, o seu valor é mais elevado para os pares classificados como correctos por ambos os avaliadores, seguidos pelos pares em que não houve concordância e pelos pares classificados como incorrectos por ambos.

A título de exemplo, apresentam-se na tabela 8 alguns pares de palavras presentes no mesmo synset (Pal_1 e Pal_2), a pertença de cada uma ao synset (μ_1 e μ_2), e a classificação do par (sinónimos possíveis ou não?) por cada um dos avaliadores (Class_A e Class_B).

6 Utilização de relações de hiperonímia

Os resultados apresentados anteriormente constituíram a primeira experiência na aplicação da abordagem proposta a relações de sinonímia extraídas a partir de três dicionários da língua portuguesa. No entanto, relações de outros tipos podem também transmitir informação relevante no cálculo da pertença (confiança) das palavras a synsets. Nesta secção apresenta-se uma das primeiras experiências onde, para além da utilização das relações de sinonímia, da mesma forma que o anteriormente relatado, as relações de hiperonímia são também consideradas no cálculo da pertença. Este tipo de relação foi escolhido não só por ter muitas instâncias no CARTÃO (cerca de 115 mil, 95 mil distintas), mas principalmente por indicar uma generalização/especificação. Ou seja, hipónimos partilham um conjunto de características com os seus hiperónimos, e por isso podem considerar-se semanticamente próximos. Há mesmo várias medidas para o cálculo de similaridade semântica com base nestas relações, na WordNet (por exemplo, Resnik (1995) ou Leacock & Chodorow (1998)).

Há, no entanto, que distinguir casos em que, nos dicionários, a relação é apresentada como sendo de sinonímia (equivalência) de casos em que é apresentada como hiperonímia (digamos, semelhança). Como num synset todas as palavras devem partilhar um significado, a primeira deve ter mais peso. Além disso, a nosso ver, quando uma palavra não está numa relação de sinonímia com nenhuma das palavras de um synset, ela simplesmente não deve pertencer a esse

Cat	Corte (θ)	Pares CLIP 1.0 ($\theta = 0,01$) Total	PalavrasNoTeP	ParNoTeP	Pares CLIP 2.0 Total	PalavrasNoTeP	ParNoTeP
N	0.000	664.559	293.970 (44.2%)	25.893 (08.8%)	2.317.478	1.081.018 (46.6%)	30.407 (02.8%)
	0.105	126.287	27.251 (21.6%)	10.466 (38.4%)	279.882	126.422 (45.2%)	16.588 (13.1%)
	0.225	74.639	11.726 (15.7%)	6.061 (51.7%)	62.607	23.141 (37.0%)	7.331 (31.7%)
	0.510	51.698	5.296 (10.2%)	2.988 (56.4%)	7.012	1.362 (19.4%)	1.127 (82.7%)
V	0.000	399.614	241.886 (60.5%)	33.818 (14.0%)	1.385.293	1.008.012 (72.8%)	49.476 (04.9%)
	0.105	44.688	16.856 (37.7%)	7.839 (46.5%)	28.378	18.101 (63.8%)	8.654 (47.8%)
	0.225	21.019	6.614 (31.5%)	3.871 (58.5%)	5.443	3.353 (61.6%)	2.519 (75.1%)
	0.510	11.528	2.819 (24.5%)	1.587 (56.3%)	794	423 (53.3%)	375 (88.7%)
Adj	0.000	346.076	212.104 (61.3%)	222.96 (10.5%)	1.149.294	685.983 (59.7%)	27.902 (04.1%)
	0.105	52.005	21.211 (40.8%)	8.446 (39.8%)	33.296	17.044 (51.2%)	7.885 (46.3%)
	0.225	26.283	9.203 (35.0%)	4.927 (53.5%)	9.722	4.420 (45.5%)	3.128 (70.8%)
	0.510	16.222	4.643 (28.6%)	2.621 (56.5%)	2.319	822 (35.4%)	754 (91.7%)

Tabela 6: Confirmação de pares de sinonímia no TeP.

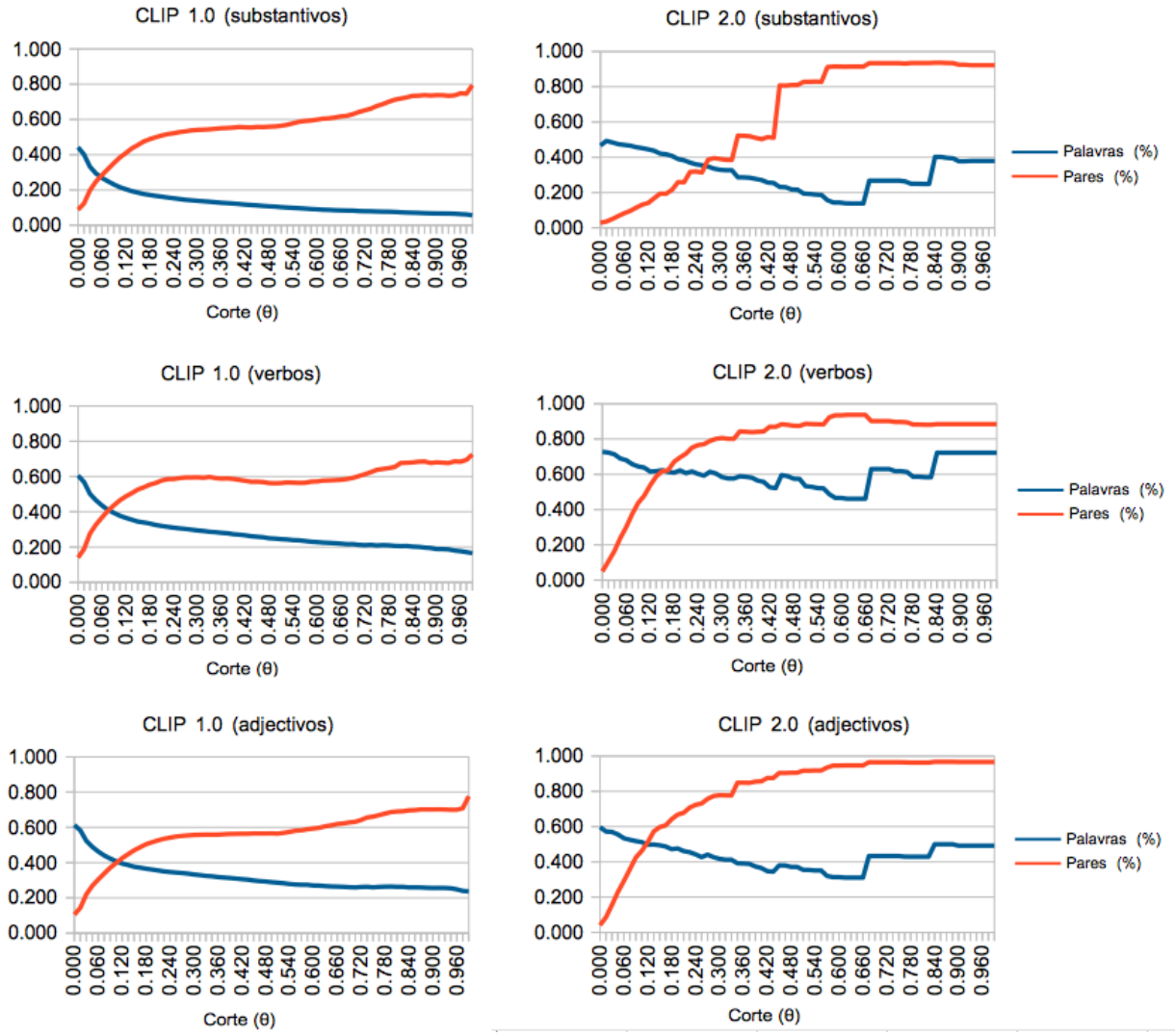

Figura 2: Proporção entre pares do CLIP 1.0 e CLIP 2.0 confirmados no TeP e pares com ambas as palavras cobertas pelo TeP.

Cat	CLIP 1.0 ($\theta = 0,01$) Correctos	κ	CLIP 2.0 Correctos	κ	$\overline{\mu_c}$	$\overline{\mu_d}$	$\overline{\mu_i}$
N	75.0%	0.43	84.7-88.0%	0.75	0.30	0.26	0.22
V	N/A	N/A	68.7-68.7%	0.65	0.25	0.18	0.17
Adj	N/A	N/A	74.7-77.3%	0.74	0.25	0.19	0.15

Tabela 7: Resultados da avaliação manual e média dos graus de pertença para cada classe de pares.

Class$_A$	**Class**$_B$	**Cat**	**Pal**$_1$	μ_1	**Pal**$_2$	μ_2
×	×	N	*sede*	0,429	*diocese*	0,042
✓	×	N	*necessidade*	0,055	*mando*	0,111
✓	✓	N	*vaqueiro*	0,555	*ganadeiro*	0,444
×	×	ADJ	*natal*	0,416	*patriótico*	0,066
✓	×	ADJ	*húmido*	0,055	*hídrico*	0,055
✓	✓	ADJ	*patriótico*	0,666	*nacionalista*	0,555
×	×	V	*vigiar*	0,083	*civilizar*	0,166
✓	×	V	*dormir*	0,133	*roncar*	0,583
✓	✓	V	*remodelar*	0,833	*reformular*	0,111

Tabela 8: Exemplos de pares de sinónimos e sua classificação manual pelos avaliadores.

synset, mesmo que seja um hipónimo ou hiperónimo de alguma. A relação de hiperonímia propriamente dita, estabelecida entre dois synsets (generalização e especialização) será integrada numa fase posterior deste trabalho.

Com base nas considerações anteriores, as relações de hiperonímia vão apenas aumentar a pertença em casos em que uma palavra está em relações de sinonímia com algumas das palavras do synset base, mas também em relações de hiperonímia com outras dessas palavras. Estes casos serão, acreditamos, situações em que, na própria linguagem, a utilização do hipónimo e do hiperónimo se confundem e acabam por ser usadas para referir o mesmo. Ao mesmo tempo, num dicionário que já inclua uma relação de sinonímia entre duas palavras, não serão consideradas relações de hiperonímia entre as mesmas palavras. Assim, o peso vindo de cada fonte nunca pode ser superior a 1. Devemos acrescentar que, como nos dicionários utilizados as relações de hiperonímia se estabelecem apenas entre substantivos, esta experiência foi aplicada somente a palavras desta categoria gramatical.

Para confirmar rapidamente que a consideração dos hiperónimos desta forma alterava as pertenças da forma desejada, aproveitamos os dados da avaliação manual anterior. A tabela 9 apresenta os valores médios das pertenças de pares de palavras do mesmo synset, primeiro, sem a consideração das relações de hiperonímia e, segundo, quando as relações de hiperonímia são consideradas com um peso que é metade dos das relações de sinonímia. Ou seja, para calcular a pertença, antes de aplicar a equação 1, a matriz de adjacências da rede, A, é alterada, de forma a que, sempre que haja uma relação de hiperonímia entre duas palavras $H(P_i, P_j)$, se também existir pelo menos uma relação de sinonímia, $R(P_i, P_j)$, a ligação entre as palavras é reforçada, $A_{ij} += 0.5$.

De forma a observar a evolução nas pertenças médias, a tabela 9 apresenta também a diferença entre o valor destas antes ($peso = 0,0$) e depois de considerar as relações de hiperonímia

Peso hiperonímia	$\overline{\mu_c}$	$\overline{\mu_d}$	$\overline{\mu_i}$
0,0	0,29960	0,26132	0,21957
0,5	0,30368	0,26234	0,22096
Diferença	0,00408	0,00102	0,00139
Ganho	0,01362	0,00390	0,00633

Tabela 9: Diferenças e ganhos nas pertenças médias de pares de sinonímia correctos, discordantes e incorrectos.

($peso = 0,5$), e mostra ainda o ganho em cada média (equação 5).

$$Ganho = \frac{Valor_{nova} - Valor_{anterior}}{Valor_{anterior}} \quad (5)$$

Como apenas os casos em que existiam relações de hiperonímia era valorizados, e não havia mais nenhuma alteração, os valores das pertenças ou se mantinham, ou aumentavam. Ou seja, o ganho seria zero ou positivo. Na tabela verifica-se que, apesar do ganho ser sempre positivo, é ligeiramente superior nos casos em que ambos os anotadores concordaram que as duas palavras do par eram sinónimos, ou seja, tornou o valor das pertenças um pouco mais fiel a uma medida de confiança.

Com base nos valores obtidos, decidimos começar a utilizar também as relações de hiperonímia no cálculo das pertenças aos synsets difusos. A título de exemplo, a tabela 10 apresenta três synsets difusos antes e depois de serem consideradas as relações de hiperonímia.

7 Conclusões e trabalho futuro

Com vista à descoberta de conceitos, descritos por conjuntos de palavras com pertenças variáveis, apresentamos uma nova abordagem para a descoberta de synsets difusos através de redes léxico-semânticas. Esta abordagem tira partido da redundância em redes extraídas a partir de várias fontes, neste caso dicionários, por isso o valor da pertença pode, de certa forma, quantificar a confiança na utilização da palavra para se referir ao conceito que emerge do synset.

Antes	**Depois**
ramada(0.67), *ramagem*(0.52), *rama*(0.52), *enramada*(0.29), *ramosidade*(0.24), *arramada*(0.19), *fronde*(0.19), *parreira*(0.13), *latada*(0.083), *França*(0.042), *ramaria*(0.042), ***folhagem*(0.042)**	*ramada*(0.67), *ramagem*(0.52), *rama*(0.52), *enramada*(0.29), *ramosidade*(0.24), *arramada*(0.19), *fronde*(0.19), *parreira*(0.13), *latada*(0.083), ***folhagem*(0.063)**, *França*(0.042), *ramaria*(0.042)
panfleto(0.83), *libelo*(0.83), *querela*(0.11), ***folheto*(0.11)**	*panfleto*(0.83), *libelo*(0.83), ***folheto*(0.17)**, *querela*(0.11)
apelido(0.46), ***nome*(0.46)**, ***alcunha*(0.40)**, *cognome*(0.31), ***epíteto*(0.23)**, ***sobrenome*(0.23)**, *designação*(0.17), *denominação*(0.17), *qualificação*(0.15), ...	***nome*(0.48)**, *apelido*(0.46), ***alcunha*(0.41)**, *cognome*(0.31), ***sobrenome*(0.25)**, ***epíteto*(0.24)**, *designação*(0.17), *denominação*(0.17), *qualificação*(0.15), ...

Tabela 10: Exemplos de synsets difusos com pertenças das palavras antes e depois de considerar as relações de hiperonímia.

A abordagem proposta diferencia-se de uma abordagem anterior para o mesmo fim por ser realizada em dois passos e por considerar apenas as adjacências relevantes para o cálculo das pertenças de cada palavra a um synset. Isto diminuiu o ruído e tornou o valor das pertenças mais facilmente interpretável, o que se confirma não só pela avaliação manual de ambas as abordagens, mas também pela comparação do valor das pertenças de diferentes pares de palavras. Como esperado numa medida de confiança, pares de palavras que devem estar no mesmo synset (sinónimos) têm em média uma pertença superior a pares que, de acordo com anotadores humanos, não são sinónimos.

Ainda assim, apesar dos resultados positivos, os valores da avaliação mostram que há ainda muita margem de melhoria. Por exemplo, enquanto cerca de 88% dos pares de substantivos pertencentes ao mesmo synset são efectivamente sinónimos, para os verbos, este número desce para 68%. Nos próximos passos a realizar neste âmbito, pretendemos realizar novas experiências para averiguar a melhor forma de considerar outros tipos de relação. Por exemplo, uma ideia a seguir é que palavras sinónimas devem estar relacionadas da mesma forma com as mesmas palavras (por exemplo, tanto *carro*, como *automóvel* devem ser hipónimos de *veículo* e ter como partes *roda* ou *motor*). Por outro lado, pretendemos aplicar esta abordagem a outras fontes de sinonímia, que permitirão não só ampliar o recurso, mas também reforçar a medida de confiança. Entre os recursos candidatos encontram-se outras wordnets livres, como o próprio TeP (Maziero et al., 2008), a OpenWN-PT (de Paiva et al., 2012) ou a PULO (Simões & Guinovart, 2014).

O recurso resultante deste trabalho será uma wordnet para a língua portuguesa, criada de forma automática, e em que haverá valores de confiança associados a algumas das decisões tomadas, incluindo não só a inclusão de palavras em synsets, como também o estabelecimento de relações entre synsets, que será uma das próximas fases do trabalho. Acreditamos que este recurso, a ser disponibilizado em breve, possa ser de grande utilidade para aqueles que procuram uma wordnet para o português em que o balanço entre cobertura e confiança possa ser personalizado de acordo com as necessidades da aplicação.

Apesar de ser possível realizar um exercício de alinhamento da versão actual do recurso a outra wordnet, uma prática cada vez mais comum, isso não é uma das nossas preocupações actuais, como não foi para o Onto.PT. Isto porque, a cada versão, não só os conteúdos, mas a própria estrutura do recurso podem ser substancialmente alterados. Por exemplo, para além da exploração de diferentes recursos, os vários passos da abordagem ECO podem ser implementados de forma diferente e levar a diferenças ao nível das fronteiras dos synsets e da granularidade dos sentidos de cada palavra. Ou seja, para cada nova versão, seria necessário realizar um novo alinhamento, quer devido à aplicação de diferentes implementações de cada passo da abordagem ECO, ou simplesmente à utilização de diferentes recursos. Para minimizar este trabalho, seria necessário definir um núcleo fixo de synsets que se manteriam estáveis de versão para versão, ou então esperar que o recurso atinja uma fase menos experimental.

Agradecimentos

Este trabalho foi parcialmente realizado no âmbito do projecto ConCreTe – *Concept Creation Technology*.

The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

Referências

Araúz, P. León, J. Gómez-Romero & F. Bobillo. 2012. A fuzzy ontology extension of wordnet and eurowordnet for specialized knowledge. Em *Proceedings of Terminology and Knowledge Engineering Conference* TKE 2012, Madrid, Spain.

Bezdek, James C. 1981. *Pattern recognition with fuzzy objective function algorithms*. Norwell, MA, USA: Kluwer Academic Publishers.

Biemann, Chris. 2006. Chinese Whispers: An efficient graph clustering algorithm and its application to natural language processing problems. Em *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing* TextGraphs-1, 73–80. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses. Em *Proceedings of the 6th Global WordNet Conference* GWC 2012, 64–71.

Borin, Lars & Markus Forsberg. 2010. From the people's synonym dictionary to fuzzy synsets - first steps. Em *Proceedings of LREC 2010 workshop on Semantic relations. Theory and Applications*, 18–25. La Valleta, Malta.

Dolan, William B. 1994. Word sense ambiguation: clustering related senses. Em *Proceedings of 15th International Conference on Computational Linguistics* COLING'94, 712–716. Morristown, NJ, USA: ACL Press.

van Dongen, Stijn Marinus. 2000. *Graph clustering by flow simulation*: University of Utrecht. Tese de Doutoramento.

Dorow, Beate, Dominic Widdows, Katarina Ling, Jean-Pierre Eckmann, Danilo Sergi & Elisha Moses. 2005. Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. Em *Proceedings of MEANING-2005, 2nd Workshop organized by the MEANING Project*, Trento.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database (language, speech, and communication)*. The MIT Press.

Gfeller, David, Jean-Cédric Chappelier & Paulo De Los Rios. 2005. Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. Em *Proceedings of International Symposium on Applied Stochastic Models and Data Analysis* ASMDA 2005, 106–113. Brest, France.

Gonçalo Oliveira, Hugo. 2013. *Onto.pt: Towards the automatic construction of a lexical ontology for portuguese*: University of Coimbra. Tese de Doutoramento. http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_PhdThesis2012.pdf.

Gonçalo Oliveira, Hugo, Leticia Antón Pérez, Hernani Costa & Paulo Gomes. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática* 3(2). 23–38.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2010. Automatic creation of a conceptual base for Portuguese using clustering techniques. Em *Proceedings of 19th European Conference on Artificial Intelligence (ECAI 2010)*, 1135–1136. Lisbon, Portugal: IOS Press.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2011. Automatic Discovery of Fuzzy Synsets from Dictionary Definitions. Em *Proceedings of 22nd International Joint Conference on Artificial Intelligence* IJCAI 2011, 1801–1806. Barcelona, Spain: IJCAI/AAAI.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2014. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation* 48(2). 373–393.

Gonçalo Oliveira, Hugo, Valeria de Paiva, Cláudia Freitas, Alexandre Rademaker, Livy Real & Alberto Simões. 2015. As wordnets do português. Em Alberto Simões, Anabela Barreiro, Diana Santos, Rui Sousa-Silva & Stella E. O. Tagnin (eds.), *Linguística, Informática e Tradução: Mundos que se Cruzam*, vol. 7(1) (OSLa: Oslo Studies in Language 1), 397–424. University of Oslo.

Hartigan, J. A. & M. A. Wong. 1979. A K-means clustering algorithm. *Applied Statistics* 28. 100–108.

Kilgarriff, A. 1996. Word senses are not bona fide objects: implications for cognitive science, formal semantics, NLP. Em *Proceedings of 5th International Conference on the Cognitive Science of Natural Language Processing*, 193–200.

Leacock, Claudia & Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. Em Christiane Fellfaum (ed.), *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, 265–283. Cambridge, Massachusetts: The MIT Press.

Lin, Dekang & Patrick Pantel. 2002. Concept discovery from text. Em *Proceedings of 19th International Conference on Computational Linguistics* COLING 2002, 577–583.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo & Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, 390–392.

Nasiruddin, Mohammad. 2013. A state of the art of word sense induction: A way towards word sense disambiguation for under resourced languages. *TALN/RECITAL 2013* .

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2). 1–69.

Navigli, Roberto & Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250.

de Paiva, Valeria, Alexandre Rademaker & Gerard de Melo. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. Em *Proceedings of 24th International Conference on Computational Linguistics* COLING (Demo Paper), 353–359.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. Em *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* IJCAI'95, 448–453. San Francisco, CA, USA.

Santos, Diana & Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em *Proceedings of 2nd International Conference on Language Resources and Evaluation* LREC 2000, 205–210.

Schaeffer, Satu Elisa. 2007. Graph clustering. *Computer Science Review* 1(1). 27–64.

Simões, Alberto & Xavier Gómez Guinovart. 2014. Bootstrapping a portuguese wordnet from galician, spanish and english wordnets. *Advances in Speech and Language Technologies for Iberian Languages* 8854. 239–248.

Velldal, Erik. 2005. A fuzzy clustering approach to word sense discrimination. Em *Proceedings of 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Denmark.

Xu, Rui & D. Wunsch, II. 2005. Survey of clustering algorithms. *Transactions on Neural Networks* 16(3). 645–678. doi:10.1109/TNN.2005.845141.