Discovering Fuzzy Synsets from the Redundancy in Different Lexical-Semantic Resources

Hugo Gonçalo Oliveira, Fábio Santos
CISUC, Department of Informatics Engineering
University of Coimbra, Portugal
hroliv@dei.uc.pt, fasantos@student.dei.uc.pt

Abstract

Although represented as such in wordnets, word senses are not discrete. To handle word senses as fuzzy objects, we exploit the graph structure of synonymy pairs acquired from different sources to discover synsets where words have different membership degrees that reflect confidence. Following this approach, a wide-coverage fuzzy thesaurus was discovered from a synonymy network compiled from seven Portuguese lexical-semantic resources. Based on a crowdsourcing evaluation, we can say that the quality of the obtained synsets is far from perfect but, as expected in a confidence measure, it increases significantly for higher cut-points on the membership and, at a certain point, reaches 100% correction rate.

Keywords: wordnets, synonyms, word senses, fuzzy synsets, clusters

1. Introduction

Wordnets are lexical-semantic knowledge bases, modelled after Princeton WordNet (Fellbaum, 1998), where words are grouped with their synonyms, in the so-called synsets, a representation of natural language concepts by their possible lexicalisations. As natural language is ambiguous, different words might have the same meaning, and the same word might be in more than one synset, one for each of its senses. But word senses are not discrete (Kilgarriff, 1996). They are complex and overlapping structures and their representation as crisp objects does not reflect the human language. A more realistic approach would handle word senses with uncertainty and represent concepts as fuzzy synsets. In fact, this idea is not new. Fuzzy memberships of words to synsets have been obtained from manual judgements (Borin and Forsberg, 2010) and there have been approaches for representing WordNet as an ontology with fuzzy synsets and relations (Araúz et al., 2012).

The wordnet model has been adopted by many languages (see e.g. Bond and Paik (2012)) and there are some with more than one wordnet, including Portuguese, for which there are six wordnets, created by independent teams, following different approaches, and with different licenses (Gonçalo Oliveira et al., 2015). But, in opposition to English, where Princeton WordNet can be seen as a standard resource, the open Portuguese wordnets are still in an early development stage and all of them have their strengths and limitations, either in terms of coverage or correction. For instance, OpenWN-PT (de Paiva et al., 2012) and, especially, PULO (Simões and Guinovart, 2014), still have a low coverage of words, senses and relation types. They both have a controlled expansion and are aligned with Princeton WordNet where additional information can be obtained. On the other hand, Onto.PT (Gonçalo Oliveira and Gomes, 2014) is the largest, but has more reliability issues, because it is created automatically, from the exploitation of dictionaries and other textual sources. Its creation follows the ECO approach, where relations, synsets and their boundaries, as well as relation attachments, are discovered automatically in three steps: (i) extraction, where semantic relations of different types, held between words, are acquired from textual resources; (ii) clustering, where groups of synonymous words (synsets) are discovered; and (iii) ontologising, where the word arguments of the extracted relations are attached to the most suitable synsets discovered.

To minimise the aforementioned limitations, we aim to create a new Portuguese wordnet, following the lines of ECO, but where a confidence degree is assigned to each decision made. This would enable the creation of a wide-coverage lexical-semantic resource and, at the same time, let users control the portion to use, by applying different cut-points, depending on their tolerance to lower coverage or reliability. The assigned measures might also be relevant for other tasks, such as word sense disambiguation (Navigli, 2009).

The first step towards the new wordnet is a kind of word sense induction (Nasiruddin, 2013), where synsets, discovered in an unsupervised fashion, will include words with fuzzy memberships that should reflect confidence on their usage to convey the synset meaning. Since, in ECO, synsets are discovered from synonymy networks acquired directly from available lexical resources, word memberships may rely on evidence taken from the structure of the synonymy connections and their redundancy.

The remaining of this paper starts with a brief reference to related work on fuzzy clustering and on the discovery of word clusters, followed by a description of the proposed algorithm for discovering fuzzy synsets. After running this algorithm on a large synonymy network, a wide-coverage fuzzy thesaurus is obtained, here presented, analysed and illustrated. Towards higher coverage and a fair amount of redundancy, seven Portuguese lexical-semantic resources were exploited, including dictionaries, thesauri and wordnets. A crowdsourcing evaluation, then described, shows that we are heading towards the right direction: the quality of the original synsets is far from perfect, but it increases significantly for higher cut-points on the membership, as expected in a confidence measure. These results also shown an improvement towards previous approaches, either in the convergence towards higher correction or on the higher coverage of words and senses. We conclude by drawing some future directions for this work, whose main goal, we recall to be the creation of a fuzzy wordnet for Portuguese, freely available for usage by the community.

2. Related Work

A classic algorithm for fuzzy clustering is the Fuzzy C-Means (FCM) (Bezdek, 1981), where elements are clustered according to their distance to k centroids, improved iteratively. But FCM requires both the desired number of clusters and the initial centroids as input, which is unknown in our case.

There is related work on the automatic discovery of concept signatures, described by overlapping (Lin and Pantel, 2002) or fuzzy (Velldal, 2005) word clusters. However, although words in the same cluster have a strong relation, they are not exclusively synonyms, and thus a cluster cannot be seen as a wordnet synset.

There are other clustering algorithms that exploit the structure of a graph to find groups of related vertices. Some are based on random walks, such as Markov Clustering (van Dongen, 2000) or Chinese Whispers (CW) (Biemann, 2006), a more efficient alternative. Both of the previous have been applied to different NLP problems, such as word sense discrimination (Dorow et al., 2005), synonymy networks organisation (Gfeller et al., 2005), language identification (Biemann, 2006) or synset discovery (Gonçalo Oliveira and Gomes, 2010).

Although the clustering approach proposed in this paper is based on running one of the previous algorithms on synonym network, it has also inspiration from FCM and Clustering By Committee (Lin and Pantel, 2002).

3. Proposed Approach

Our previous approach to the discovery of fuzzy synsets from synonymy networks (Gonçalo Oliveira and Gomes, 2011) had some limitations. Briefly, each node was a potential cluster that could attract neighbour nodes, depending on the similarity of their adjacencies. Based on their overlap, some clusters ended up being merged, which resulted in very large synsets, impractical if a cut-point was not applied, as well as many word senses. Moreover, redundancy was not exploited efficiently and the membership of highly connected words was penalised, because the whole adjacency vector was considered when computing similarities.

Though inspired by the previous, we propose an alternative approach for the discovery of fuzzy synsets from synonymy networks with two steps: (i) centroid discovery; (ii) fuzzy memberships computation. It is applied to a weighted synonymy network $N = (W, P)$, where W is a set of words and P a set of synonymy pairs. N can be represented as an adjacency matrix $A(|W| \times |W|)$, where $A_{ij} = \omega_{ij}$, a weight that reflects the number of times a synonymy pair, $P(W_i, W_j)$, occurs in the exploited sources. The maximum weight m is a constant, which, in the case of this work, can be equal to the total number of synonymy sources used.

In the first step, an efficient graph clustering algorithm, like CW, is applied. The result is a set of centroids, where words are structurally related, with some similarities to the committees of Lin and Pantel (2002). It may be represented as a partition matrix C with $|W|$ rows, one for each word, and columns that represent hard clusters, used as centroids.

In the second step, the membership value of word W_i to centroid C_k, $\mu(W_i, C_k)$, is computed by equation 1, where $A[C_k]_j$ contains the weight of the connection between W_i and W_j. This is close to computing the memberships in FCM, but it is done only once, because the centroid words are already strongly connected.

$$\mu(W_i, C_k) = \frac{\sum_{j=0}^{|C_k|} A[C_k]_j}{|C_k|} \tag{1}$$

The proposed algorithm is illustrated with the weighted subgraph of figure 1, where two senses of the Portuguese word *canudo* arise: a tube/pipe, or, more informally, a diploma. Suppose that CW identifies two hard clusters, in table 1. To compute the membership of *canudo* to the fuzzy cluster C'_A, the weights of the connections between this word and words in C_A are summed and divided by the size of C_A. Since there is one connection of weight 2 between *canudo* and *diploma*, $\mu(canudo, C_A) = \frac{2}{4} = 0.5$. For computing the membership of *canudo* to C'_B, the three connections between this word and words in C_B are considered (*bica*, *tubo* and *cano*), plus the word *canudo* itself, which belongs to C_B, so $\mu(canudo, C_B) = \frac{3+5+2+m}{6}$. If the network were extracted from five resources, $m = 5$ and $\mu(canudo, C_B) = \frac{15}{6} = 2.5$. For convenience, memberships may be normalised in the $[0, 1]$ interval, if they are divided by m.

Table 2 shows the fuzzy synsets computed from the centroids in table 1 and the network in figure 1. The word *diploma* should not be in C'_B, but it has a weak membership and may be removed if a cut-point, for instance, $\theta = 0.35$ is applied.

C_A	*diploma*, *título*, *certidão*, *certificado*
C_B	*canudo*, *bica*, *tubo*, *cano*, *canal*, *ducto*

Table 1: Hard clusters (centroids) discovered from the network in figure 1

C'_A	*diploma*(3.0), *título*(2.25), *certidão*(3.0), *certificado*(3.25), *canudo*(0.5)
C'_B	*canudo*(2.5), *bica*(1.83), *tubo*(3.17), *cano*(2.33), *canal*(2.67), *ducto*(2.17), *diploma*(0.33)

Table 2: Fuzzy synsets obtained from the hard clusters of table 1 and memberships based on the network in figure 1

4. Experimentation

This section describes the application of the proposed approach to the synonymy networks acquired from seven open Portuguese lexical-semantic resources and describes the obtained results, while comparing them with those of previous approaches.

4.1. Exploited resources

The synonymy network used in this work was acquired from the following seven resources:

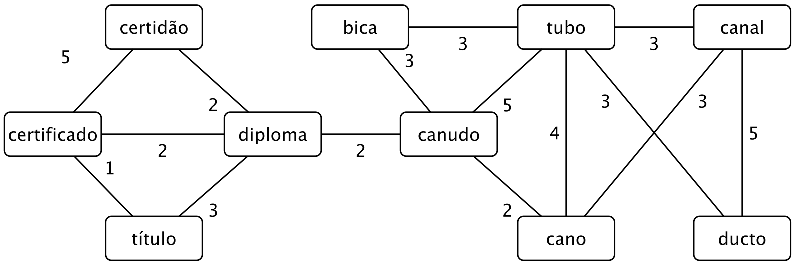

Figure 1: Network with nodes and their weights

- Synonymy relations of the lexical-semantic network PAPEL (Gonçalo Oliveira et al., 2008), automatically extracted from a Portuguese dictionary;
- Synonymy relations extracted from two other Portuguese dictionaries, Dicionário Aberto (Simões et al., 2012) and Wiktionary.PT, with the same grammars as PAPEL, and currently included in the lexical-semantic network CARTÃO (Gonçalo Oliveira et al., 2011);
- Two synonym thesauri for Portuguese: TeP (Maziero et al., 2008), handcrafted for Brazilian Portuguese, and OpenThesaurus.PT[1], used in writing processors;
- Two Portuguese wordnets: OpenWordNet-PT (de Paiva et al., 2012) and PULO (Simões and Guinovart, 2014).

Table 3 characterises each of the previous resources by the synonymy relations acquired from them and the number of involved words, according to their part-of-speech – nouns, verbs or adjectives. The subnetworks for the nouns (N), verbs (V) and adjectives (A) used, obtained after merging all of the previous, is characterised in table 4, which displays the number of vertices ($|W|$), edges ($|P|$), the average vertex degree ($\overline{deg}(N)$), the number of connected components ($\#CC$), and the size of the largest component ($|W_{lc}|$). As others, we have noticed that these subgraphs have one large and several smaller components. $\overline{CC}$s are comparable to those of small-worlds networks. Moreover, the verb subnetwork has a higher $\overline{deg}(N)$, which means that verbs have more synonyms or are more ambiguous.

POS	$\|W\|$	$\|P\|$	$\overline{deg}$(N)	$\#CC$	$\|W_{lc}\|$
N	61,129	130,998	4.286	9,046	38,098
V	12,632	109,184	17.287	410	11,693
A	24,295	79,558	6.549	3,203	16,310

Table 4: Properties of the synonymy network.

4.2. Results in numbers

The result of applying the fuzzy clustering approach to the synonymy network of the seven resources is a fuzzy thesaurus for Portuguese. It was named CLIP 2.1, after CLIP 2.0 (Santos and Gonçalo Oliveira, 2015), where the same approach was applied to the synonymy network of three dictionaries (CARTÃO), and after CLIP 1.0[2], the result of our previous approach in the same three dictionaries (Gonçalo Oliveira and Gomes, 2011). The properties of the previous fuzzy thesauri are displayed in table 5. Those include the number of words and average number of word senses, number of synsets and their average size, synsets of size 2, larger than 25, and the size of the largest synset. For the sake of practicality, the properties of CLIP 1.0 were obtained with a cut-point $\theta = 0.01$. In the same table, the properties of TeP 2.0 (Maziero et al., 2008) were included. TeP is a synonym thesaurus, handcrafted for Brazilian Portuguese based on information in dictionaries, whose synonymy network ended up being included in CLIP 2.1.

Since they have used the same data, CLIP 1.0 and CLIP 2.0 provide a nice comparison between the current and the previous approach for fuzzy synset discovery from synonymy networks. CLIP 2.0 has more nouns and adjectives, but less verbs. Those differences are mostly related to the need of applying a cut-point to CLIP 1.0, otherwise it would become even more impractical. Differences are clearer in the average number of senses and synset sizes, both substantially higher for CLIP 1.0, which suggests that CLIP 1.0 is noisier and points out the limitations of the previous approach. Those numbers can be compared with TeP's, handcrafted, and thus a possible reference. On the other hand, the average number of senses and synset size in CLIP 2.0 are closer to TeP's. The main differences are on the number of words and large synsets. CLIP 2.0 has substantially more words, for a similar number of synsets, which are larger. But we recall that no cut-point was applied to CLIP 2.0 nor CLIP 2.1, and its application could minimise the previous sign of noise.

The new thesaurus, CLIP 2.1, is clearly the largest, which was expected because more resources were exploited in its creation. It has more than twice the number of words in TeP, which it includes, substantially more words than CLIP 1.0, but still less signs of noise than the latter. It should also be noticed that the average number of senses and the synset size are higher for the verbs in all the thesauri. This is particularly salient in TeP and, as a consequence, also in CLIP 2.1. To some extent, this situation was predictable by the properties of the verb synonymy network (see table 4), which has a significantly higher degree than the others. A possible interpretation is that Portuguese verbs are more ambiguous and have more synonyms.

[1] Currently available from `http://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt`

[2] CLIP 1.0 had been originally baptised as Padawik, but was later renamed as CLIP (Gonçalo Oliveira, 2013).

Resource	Nouns		Verbs		Adjectives	
	#Words	#Relations	#Words	#Relations	#Words	#Relations
PAPEL	25,553	41,663	7,634	18,869	11,975	21,722
Dicionário Aberto	26,254	26,675	7,502	12,539	9,611	11,762
Wiktionary.PT	16,370	18,980	4,463	6,910	6,639	9,077
TeP	17,149	103,066	8,280	178,912	14,568	103,290
OpenThesaurus.PT	6,110	21,946	2,856	12,836	3,747	16,262
OpenWordNet-PT	20,568	24,660	2,858	4,891	3,332	4,246
PULO	3,078	4,762	1,203	2,953	876	1,376

Table 3: Synonymy networks of the exploited resources.

	POS	Words			Synsets				
		#	$\overline{senses}$	max(#senses)	#	$\overline{size}$	size=2	size>25	max(size)
CLIP 2.1	**N**	61,124	2.11	35	13,735	9.40	6,431	877	548
	V	12,632	5.29	54	1,126	59.31	327	279	1,689
	A	24,295	2.45	44	4,827	11.92	2,364	414	720
CLIP 2.0	**N**	43,721	1.92	42	9,881	8.49	4,147	632	554
	V	10,380	3.15	54	1,438	22.76	289	370	500
	A	17,368	2.28	44	3,571	11.07	1,530	367	322
CLIP 1.0 ($\theta = 0.01$)	**N**	39,354	7.78	46	20,102	15.23	3,885	3,756	109
	V	11,502	14.31	42	7,775	21.17	307	2,411	89
	A	15,260	10.36	43	8,896	17.77	1,326	2,157	109
TeP 2.0	**N**	17,158	1.71	21	8,254	3.56	3,079	0	21
	V	10,827	2.08	41	3,978	5.67	939	48	53
	A	14,586	1.46	19	6,066	3.50	3,033	19	43

Table 5: Properties of the discovered synsets (CLIP 2.1), followed by the same properties in a thesaurus discovered by the same approach but from only three dictionaries (CLIP 2.0), by a previous approach in the same three dictionaries (CLIP 1.0), and a handcrafted synonym thesaurus (TeP 2.0).

4.3. Examples

For three selected polysemic Portuguese words, table 6 shows the fuzzy synsets where they have the highest memberships, manually organised according to the transmitted meanings. At a first look, both synsets and memberships make sense. An evaluation is reported in the next section.

5. Evaluation

To assess the quality of the discovered synsets, evaluation samples were prepared for CLIP 2.1. Similarly to CLIP 2.0 (Santos and Gonçalo Oliveira, 2015) and CLIP 1.0's evaluation (Gonçalo Oliveira and Gomes, 2011; Gonçalo Oliveira, 2013), to make manual classification faster, words not occurring in AC/DC (Santos and Bick, 2000), a large collection of Portuguese corpora, were first removed, and then synsets with at least one unfrequent word ($frequency < 10$) were discarded. The evaluation samples of CLIP 2.1 contained 240 pairs of nouns, 150 of verbs and 150 of adjectives from the same synset, organised in sets of ten, and deployed to the Crowdflower platform[3], where Portuguese-speaking contributors living in Portuguese-speaking countries manually classified each pair either as possible synonyms or not. A set of examples was provided and contributors were advised to resort to online dictionaries in order to cover as many word senses as possible. Each pair was classified by two judges. In the end, 59% of the noun pairs, 46% verb and 55% adjective pairs were classified as correct. The agreement rates were respectively 87%, 85% and 75%.

These numbers are a hint on the quality of the original synsets and show that there is still much room for improvement. But they do not consider the membership values.

To confirm whether the memberships made sense as a confidence measure, the behaviour of the previous results was analysed for increasing cut-points θ – if the membership of one of the words in the pair is below θ, the pair is ignored. Table 7 shows this evolution in CLIP 2.0 and CLIP 2.1, respectively for nouns, verbs and adjectives, and also for the nouns of CLIP 1.0. For this purpose, samples of previous manual evaluations of the other thesauri were used. For CLIP 1.0, there was only a sample of noun pairs available, more precisely, 400. For CLIP 2.0, there were 150 noun, 150 verb, and 150 adjective pairs available. In both samples, each pair had been labelled, independently, by two judges To enable comparison, the membership degrees of CLIP 2.0 and CLIP 2.1 were normalised in the interval $[0, 1]$, the same where CLIP 1.0's memberships fall. This might still not be enough for a fair comparison, because the fuzzy memberships of each thesaurus were computed by a different measure. Also, since CLIP 2.1 exploits more resources, most of its words have very low memberships in the interval $[0, 1]$.

Figure 7 plots the same evolution of table 7 and confirms that the membership behaves as expected: the proportion of correct pairs increases for higher cut-points. Without a cut-point, CLIP 2.1 is clearly the less accurate thesaurus. But an important difference should be mentioned here, due to its probable negative impact in the most recent results: while CLIP 1.0's and CLIP 2.0's samples had been la-

[3] https://crowdflower.com/

Word	Meaning	Fuzzy synsets
plano	plan	*risco*(0.924), *esboço*(0.848), *traçado*(0.747), *plano*(0.696), *desenho*(0.57), *traço*(0.557), *delineamento*(0.519), *rascunho*(0.456), *debuxo*(0.43), *risca*(0.418), *esquema*(0.392), *linha*(0.38), *planta*(0.367), *borrão*(0.367), *bosquejo*(0.329), *programa*(0.317), *traça*(0.316), *projeto*(0.278), ...
	intent	*propósito*(3.385), *intenção*(3.308), *intento*(3.231), *desígnio*(2.385), *tenção*(2.385), *fim*(1.929), *finalidade*(1.846), *plano*(1.285), *intuito*(1.143), *objetivo*(1.077), *resolução*(1.071), *pressuposto*(1.071), *destino*(1.071), *mira*(1.0), *vista*(0.929), *programa*(0.857), ...
	strategy	*táctica*(3.0), *tática*(3.0), *estratégia*(1.0), *manobra*(0.667), *plano*(0.333), *planos*(0.333), *habilidade*(0.333), *regime*(0.333), *política*(0.333)
	plain	*planície*(1.053), *planura*(1.053), *chã*(0.772), *várzea*(0.719), *planalto*(0.684), *vale*(0.614), *prado*(0.614), *plaino*(0.596), *campo*(0.561), *campina*(0.526), *chapada*(0.491), *vargem*(0.421), *rechã*(0.421), *achada*(0.404), *varga*(0.404), *altiplano*(0.368), *rechão*(0.368), *platô*(0.351), *altoplano*(0.351), *rechano*(0.351), *varja*(0.228), *plano*(0.224), *veiga*(0.193), *chanura*(0.158), *val*(0.158), *chada*(0.123), *pasto*(0.121), ...
selar	to stamp	*selar*(2.2), *estampilhar*(1.75), *sigilar*(1.5), *portear*(1.5), *franquiar*(1.25), *marcar*(0.4), *carimbar*(0.2)
	to seal	*lacrar*(2.5), *cerar*(2.5), *selar*(1.333), *encerrar*(0.333)
	to end	*acabar*(1.967), *morrer*(1.833), *concluir*(1.817), *terminar*(1.633), *expirar*(1.383), *completar*(1.383), *findar*(1.367), *finalizar*(1.333), *rematar*(1.283), *perfazer*(1.217), *ultimar*(1.2), *fechar*(1.148), *finar*(1.133), *fenecer*(1.067), *encerrar*(1.0), *enfenecer*(1.0), *consumar*(0.902), *cerrar*(0.883), *arrematar*(0.867), *epilogar*(0.82), *desfechar*(0.82), *trancar*(0.803), *vencer*(0.803), *liquidar*(0.8),...
frio	cold	*gélido*(2.0), *gelado*(2.0), *glacial*(1.5), *congelado*(1.278), *frio*(1.21), *enregelado*(1.111), *regelado*(1.111), *frígido*(0944), *paralisado*(0.778), *álgido*(0.778), *algente*(0.722), *inerte*(0.722), *solidificado*(0.444), *fria*(0.167), *cortante*(0.158), ...
	insensitive	*insensível*(1.667), *indiferente*(1.333), *frio*(1.308), *apático*(0.846), *impassível*(0.77), *imperceptível*(0.667), *dessecado*(0.583), *empedernido*(0.5), *cruel*(0.462), *duro*(0.462), *passivo*(0.417), *seco*(0.385), *insensitivo*(0.333), *desapegado*(0.308), ...
	downcast	*desanimado*(1.87), *desalentado*(1.87), *frio*(1.435), *esfriado*(1.435), *descorçoado*(1.435), *esmorecido*(1.435), *abatido*(1.25), *alicaído*(1.174), *gelado*(1.125), *desacorçoado*(1.087), *descoroçoado*(1.087), *sucumbido*(1.087), *descorajado*(1.087), *desacoroçoado*(1.087), *desencorajado*(1.087), *caído*(1.167), *desmoralizado*(1.167), *acabrunhado*(1.083), *arreado*(1.083), *amarasmado*(1.083), *deprimido*(0.609), ...

Table 6: Fuzzy synsets of polysemic words.

	CLIP 1.0	**CLIP 2.0**			**CLIP 2.1**		
θ	**N**	**N**	**V**	**A**	**N**	**V**	**A**
0.000	74.3%	86.8%	68.5%	75.8%	59.2%	46.3%	55.2%
0.025	76.9%	86.1%	68.5%	75.8%	60.7%	50.0%	58.8%
0.050	78.1%	85.8%	71.1%	79.1%	71.4%	58.7%	68.5%
0.075	79.8%	85.4%	74.2%	87.2%	77.5%	67.0%	73.9%
0.100	81.1%	86.3%	80.9%	90.3%	83.1%	65.9%	77.1%
0.125	83.1%	86.7%	90.7%	97.4%	87.2%	65.8%	85.2%
0.150	83.5%	86.2%	94.1%	97.3%	87.1%	70.0%	86.4%
0.175	84.0%	86.3%	92.3%	99.0%	88.6%	84.8%	88.9%
0.200	85.1%	85.8%	91.3%	99.0%	88.6%	88.6%	92.3%
0.225	85.1%	86.9%	97.2%	100.0%	86.8%	90.5%	95.5%
0.250	84.3%	87.4%	97.1%	100.0%	87.5%	90.0%	100.0%
0.300	83.9%	89.5%	97.1%	100.0%	84.6%	90.0%	100.0%
0.350	83.8%	98.2%	96.7%	100.0%	90.0%	100.0%	100.0%
0.400	83.1%	98.2%	96.7%	100.0%	100.0%	100.0%	100.0%
0.450	83.6%	100.0%	96.4%	100.0%	100.0%	100.0%	100.0%
0.500	84.4%	100.0%	95.0%	100.0%	100.0%	–	100.0%
0.550	84.8%	100.0%	95.0%	100.0%	100.0%	–	100.0%
0.600	85.0%	100.0%	100.0%	100.0%	100.0%	–	–

Table 7: Evolution of correct synonymy pairs while increasing the cut-point in different fuzzy thesauri.

belled by a controlled group of human judges, with some expertise, CLIP 2.1's were labelled by less experienced crowdsourcers. On the other hand, CLIP 2.0, created with the same approach as CLIP 2.1 and with pairs labelled in a similar process to CLIP 1.0, has a higher correction rate for nouns, even without the application of any cut-point. Despite the previous differences, the correction rate of CLIP 2.1 and CLIP 2.0 increase faster than for CLIP 1.0. For CLIP 2.1, it reaches 100% with θ between 0.25, for adjectives, and 0.4, for nouns, while, the nouns of CLIP 1.0 never reach 100%. Even if the impact of the different kind of judges is ignored and we consider that, until a certain point, CLIP 2.1 has a lower correction rate than CLIP 2.0, its higher coverage should be highlighted here as an advantage over the other two.

Though risking not having a representative sample, as a complementary exercise, we used the manually classified CLIP 1.0 pairs to assess CLIP 2.1. Without any cut-point,

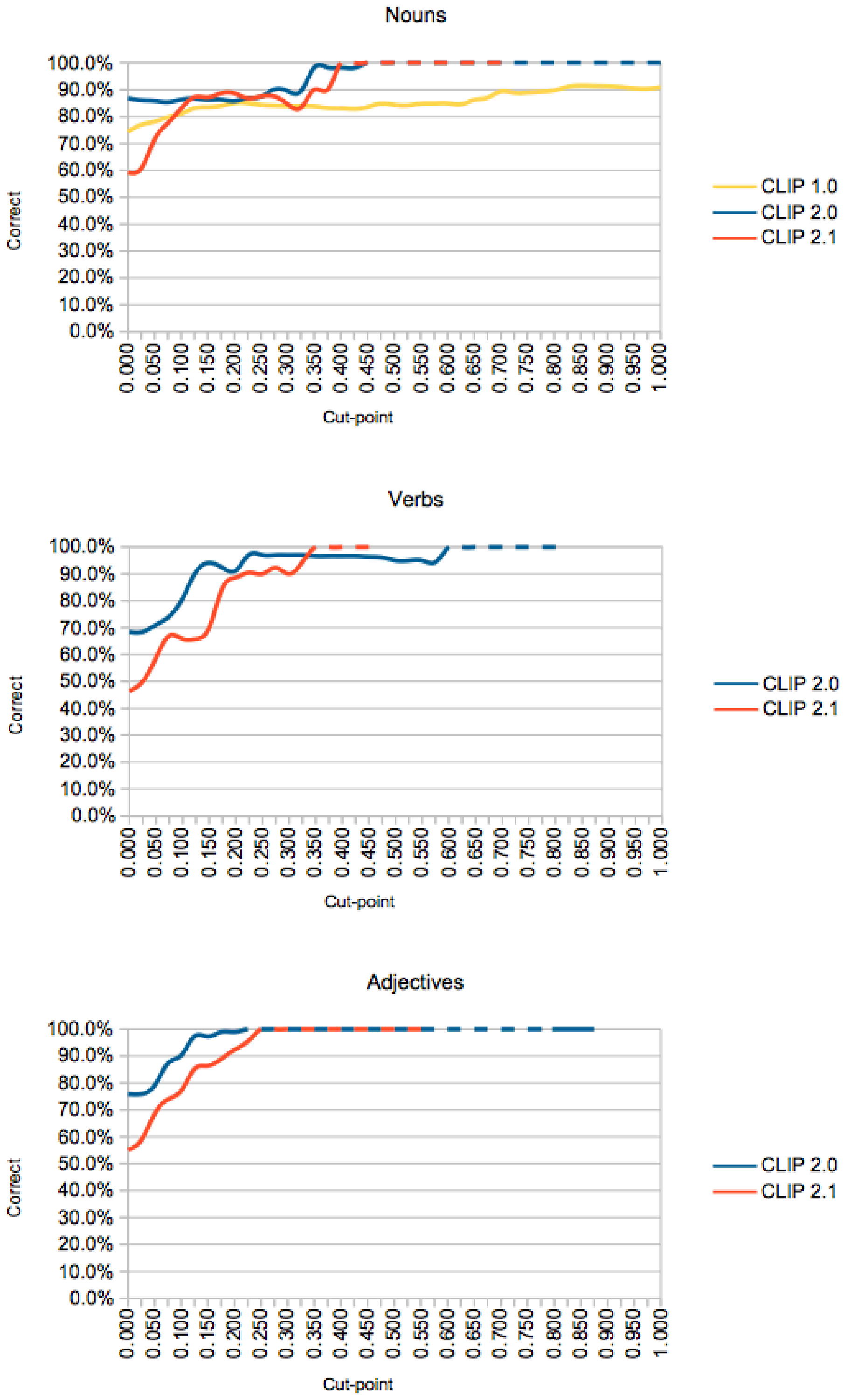

Figure 2: Plots of the evolution of correct synonymy pairs while increasing the cut-point in different fuzzy thesauri.

76% of the pairs were correct, which is slightly higher than for CLIP 1.0 (74%). This proportion reached 100% for $\theta = 0.425$, which is close to that of the crowdsourcing-based evaluation.

6. Conclusion and Further Work

An approach for discovering fuzzy synsets from (ideally redundant) synonymy networks was proposed in this paper and its application to a network acquired from seven Portuguese lexical-semantic resources was described. Based on the properties of the resulting synsets, we can say that they are less noisier than those of a previous approach and have a wider coverage of words, because more resources were exploited. Also, after analysing the behaviour of the memberships for different cut-points, we concluded that the current degrees are better-suited as a confidence measure.

Future lines concerning the improvement of these results should explore other semantic relations, besides synonymy, for computing memberships, though with a lower weight than synonymy. For instance, if several words in a synset share a relation with another, their membership may increase. This should include relations such as hypernymy and meronymy, as well as antonymy.

Moreover, to continue our path towards a fuzzy Portuguese wordnet, created automatically, and following the ECO (Gonçalo Oliveira and Gomes, 2014) approach, relations of other types will be integrated. The selection of the proper synset attachments should also be fuzzy, and thus have a degree that, among other kinds of evidence, may consider the current synset memberships. Further work on the creation of this fuzzy wordnet, currently dubbed cONTO.PT, is described in (Gonçalo Oliveira, 2016). Language resources developed on the scope of cONTO.PT are available from `http://ontopt.dei.uc.pt/`, under the menu item cONTO.PT.

Acknowledgments

The work was partially developed in the scope of project ConCreTe. 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.

Bibliographical References

Araúz, P. L., Gómez-Romero, J., and Bobillo, F. (2012). A fuzzy ontology extension of WordNet and EuroWordnet for specialized knowledge. In *Proceedings of Terminology and Knowledge Engineering Conference*, TKE 2012, Madrid, Spain, June.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Biemann, C. (2006). Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of 1st Workshop on Graph Based Methods for Natural Language Processing, New York City*, TextGraphs-1, pages 73–80. ACL Press.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference*, GWC 2012, pages 64–71.

Borin, L. and Forsberg, M. (2010). From the people's synonym dictionary to fuzzy synsets - first steps. In *Proceedings of LREC 2010 workshop on Semantic relations. Theory and Applications*, pages 18–25, La Valleta, Malta.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An open Brazilian wordnet for reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

Dorow, B., Widdows, D., Ling, K., Eckmann, J.-P., Sergi, D., and Moses, E. (2005). Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. In *Proceedings of MEANING-2005, 2nd Workshop organized by the MEANING Project*, Trento, February.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Gfeller, D., Chappelier, J.-C., and De Los Rios, P. (2005). Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. In *Proceedings of International Symposium on Applied Stochastic Models and Data Analysis*, ASMDA 2005, pages 106–113, Brest, France.

Gonçalo Oliveira, H. and Gomes, P. (2010). Automatic creation of a conceptual base for Portuguese using clustering techniques. In *Proceedings of 19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 1135–1136, Lisbon, Portugal, August. IOS Press.

Gonçalo Oliveira, H. and Gomes, P. (2011). Automatic Discovery of Fuzzy Synsets from Dictionary Definitions. In *Proceedings of 22nd International Joint Conference on Artificial Intelligence*, IJCAI 2011, pages 1801–1806, Barcelona, Spain, July. IJCAI/AAAI.

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

Gonçalo Oliveira, H., Santos, D., Gomes, P., and Seco, N. (2008). PAPEL: A dictionary-based lexical ontology for Portuguese. In *Proceedings of 8th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2008)*, volume 5190 of *LNCS/LNAI*, pages 31–40, Aveiro, Portugal, September. Springer.

Gonçalo Oliveira, H., Antón Pérez, L., Costa, H., and Gomes, P. (2011). Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática*, 3(2):23–38, December.

Gonçalo Oliveira, H., de Paiva, V., Freitas, C., Rademaker, A., Real, L., and Simões, A. (2015). As wordnets do português. In Alberto Simões, et al., editors, *Linguística, Informática e Tradução: Mundos que se Cruzam*, OSLa: Oslo Studies in Language, pages 397–424. University of Oslo.

Gonçalo Oliveira, H. (2013). *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. Ph.D. thesis, University of Coimbra.

Gonçalo Oliveira, H. (2016). Groundwork for the automatic creation of a fuzzy portuguese wordnet. In *Proceedings of 12th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2016)*, page to be published.

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

Lin, D. and Pantel, P. (2002). Concept discovery from text. In *Proceedings of 19th International Conference on Computational Linguistics*, COLING 2002, pages 577–583.

Maziero, E. G., Pardo, T. A. S., Felippo, A. D., and Diasda-Silva, B. C. (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392.

Nasiruddin, M. (2013). A state of the art of word sense induction: A way towards word sense disambiguation for under resourced languages. In *Proceedings of Traitement Automatique des Langues Naturelles and Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, TALN/RECITAL 2013.

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

Santos, D. and Bick, E. (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, LREC 2000, pages 205–210.

Santos, F. and Gonçalo Oliveira, H. (2015). Descoberta de synsets difusos com base na redundância em vários dicionários. *Linguamática*, 7(2):3–17, December.

Simões, A. and Guinovart, X. G. (2014). Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In *Advances in Speech and Language Technologies for Iberian Languages*, volume 8854 of *LNCS*, pages 239–248.

Simões, A., Álvaro Iriarte Sanromán, and Almeida, J. J. (2012). Dicionário-Aberto: A source of resources for the Portuguese language processing. In *Proceedings of 10th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2012)*, volume 7243 of *LNCS*, pages 121–127, Coimbra Portugal, April. Springer.

van Dongen, S. M. (2000). *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

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