ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-1135

Automatic Creation of a Conceptual Base for Portuguese using Clustering Techniques

Hugo Gonçalo Oliveira[1] and **Paulo Gomes**[2]

Abstract. When a semantic network is based on triples relating terms, ambiguity arises as a problem, so we have used these triples to identify clusters, which can be seen as synsets. We report the results of this approach on a synonymy network extracted from a dictionary and additional tests involving manually created thesaurus. Part of the resulting synsets were also evaluated by human subjects.

1 INTRODUCTION

In order to understand the information conveyed by natural language, today's applications demand better access to knowledge on words and their meanings, commonly structured in lexical ontologies, such as Princeton WordNet [2]. However, since this kind of resource is most of the times handcrafted, their creation and maintenance involves much human effort. So, the automatic construction of such resources from text arises as an alternative, providing less intensive labour, easier maintenance and allowing for higher coverage.

Typical systems on the automatic acquisition of knowledge from text (e.g. [5, 4]) output relational triples relating terms, *a* RELATED_TO *b*. Still, since a word may have different meanings, this representation does not handle ambiguity. Therefore, we have to move on to a structure based on concepts, which, as in WordNet, can be represented as synsets. Besides being an alternative for dealing with ambiguity, synsets describe concepts as a group of synonymous words bringing together natural language and knowledge engineering in a suitable representation for the Semantic Web.

We present an experimentation towards the automatic creation of a broad-coverage thesaurus for Portuguese. After noticing that the network established by the synonymy instances (*a* SYNONYM_OF *b*) extracted from a dictionary had a clustered structure, we followed [3], who developed a procedure based on the Markov Clustering algorithm (MCL) [8] to identify clusters, which can be seen as synsets. MCL assigns each word to one cluster but, if it is ran several times with random stochastic noise (δ), synsets are determined by the probability of each pair of words belonging to the same cluster.

2 EXPERIMENTATION

Experiments were made with the nouns[3] in existing freely available lexical resources for Portuguese: PAPEL [4], whose synonymy instances establish a synonymy network, and two handcrafted thesaurus, TeP [1] and OpenThesaurus.PT[4] (OT), used as reference for comparison with the synsets obtained automatically, for testing the clustering procedure, and for further thesaurus augmentation.

Furthermore, other thesaurus were created with the later resources: clustering on PAPEL (CLIP), clustering on TeP's (CleP) and OT's (ClOT) network, where a triple was established from each pair of words in the same synset[5], TeP and OT combined (TePOT), where each synset O_i in OT was merged with the TeP synset T_i which maximised $Jaccard(O_i, T_i)$[6], TeP, OT and CLIP merged (TOP) and finally, clustering on a network with the triples of TeP, OT and PAPEL (TOPcl).

2.1 Procedure

The synset discovery procedure is slightly different from [3]'s[7] and has the following stages: (i) split the original network into sub-networks, such that there is no path between two elements in different sub-networks, and calculate the frequency-weighted adjacency matrix F of each sub-network; (ii) add stochastic noise to each entry of F, $F_{ij} = F_{ij} + F_{ij} * \delta$; (iii) run MCL (with $\gamma = 1.6$) on F for 30 times; (iv) use the clustering obtained by each run to create a new matrix P with the probabilities of each pair of words in F belonging to the same cluster; (v) create the clusters based on P and on a given threshold $\theta = 0.2$. If $P_{ij} > \theta$, i and j belong to the same cluster; (vi) in order to clean the results, remove: (a) big clusters, B, if there is a group of clusters $C = C_1, C_2, ...C_n$ such that $B = C_1 \cup C_2 \cup ... \cup C_n$; (b) clusters completely included in other clusters. Besides improving synonymy representation in dictionaries, this procedure should homogenise synonymy representation in thesaurus. Also, if it is applied over a synonymy network extracted from several broad-coverage resources, it should be possible to obtain a richer conceptual base, with a broader coverage of the lexicon.

2.2 Results

For each thesaurus, Table 1 presents the quantity of words, words belonging to more than one synset (ambiguous), the number of synsets where the most ambiguous word occurs, the quantity of synsets, the average synset size (number of words), and the size of the biggest synset. Looking at TeP and OT against CleP and ClOT, clustering

[1] CISUC, University of Coimbra, Portugal, hroliv@dei.uc.pt, supported by FCT scholarship grant SFRH/BD/44955/2008
[2] CISUC, University of Coimbra, Portugal, pgomes@dei.uc.pt
[3] The procedure can be applied to words of other categories but, for experimentation purposes, only nouns were used.
[4] http://openthesaurus.caixamagica.pt/

[5] For instance, the TeP synset (*trabalho*, *emprego*, *serviço*) was transformed into: *trabalho* SYNONYM_OF *emprego*, *trabalho* SYNONYM_OF *serviço*, *emprego* SYNONYM_OF *serviço*.
[6] $Jaccard(A, B) = A \cap B / A \cup B$
[7] For the network we used, the obtained clusters were closer to the desired results if $-0.5 < \delta < 0.5$. Also, in MCL, we used frequency-weighted adjacency matrixes F, where each element F_{ij} corresponds to the number of times a triple denoting synonymy between i and j occurs, thus strengthening the probability that two words appearing frequently as synonymous belong to the same cluster.

seems to merge several synsets into bigger ones. Also, due to the probabilities involved in the clustering procedure, words tend to occur in more synsets. Still, ClOT has less ambiguous words than OT, due to its small size. From the original resources, PAPEL is the one with most words, however, after clustering, it is organised in less, but much bigger, synsets than TeP, which strengthens both our beliefs.

Clustering TeP, OT and PAPEL results on a thesaurus with really big synsets and very ambiguous words, which is not very practical. So, it should be better to merge manually created thesaurus by some other procedure. Nevertheless, experiments with different values of γ in MCL should be made to confirm the later assumption.

	Words			Synsets		
	Quant.	Ambig.	Most amb.	Quant.	Avg. size	Biggest
TeP	17,158	5,867	20	8,407	3.51	21
OT	5,819	442	4	1,872	3.37	14
CleP	17,158	8,484	37	4,039	19.2	174
ClOT	5,819	103	5	1,450	4.14	41
CLIP	23,741	12,196	47	7,468	12.57	103
TePOT	18,443	6,119	17	8,041	3.89	37
TOP	30,554	13,294	21	9,960	6.6	277
TOPcl	30,554	15,289	73	7,319	22.85	288

Table 1. (Noun) thesaurus comparison.

2.3 Evaluation

Each pair of thesaurus was compared after considering a reference thesaurus R and a thesaurus to be compared T. For each synset $R_i \in R$, $M \in T$ is the synset maximising $c_i = Jaccard(R_i, T_j), T_j \in T$. The overlap of R in T is then given by the sum of coefficients c_i, $O = \sum_{i=0}^{|R|} c_i$.

As Table 2 shows, the original resources have low overlaps. We can say that, despite being broad-coverage resources, all the three cover significantly different parts of lexicon. This can also be noticed by looking at the number of words in TeP, in PAPEL and in one of the thesaurus with both of them. So, as [7] and [6] suggest, these resources are more complementary than overlapping and are thus not well suited for a gold standard evaluation. On the other hand, a resource created from them all will be considerably richer than each one alone. The overlaps of OT and ClOT over each other are considerably high, which means that a thesaurus very close to OT could be established from its synonymy network. The same does not happen for TeP, possibly due its big size and the ambiguity of its words, but a deeper analysis will be needed for clearer conclusions.

R/*T*	TeP	OT	CleP	ClOT	CLIP	TePOT	TOP	TOPcl
TeP	100	**17.6**	**38.9**	14.5	**17.9**	92.3	79.9	30.7
OT	**39.7**	100	**17.1**	79.8	**22.9**	66.5	52.0	25.9
CleP	**65.2**	9.6	100	9.5	19.5	63.8	55.9	60.6
ClOT	38.2	**93.7**	19.0	100	24.8	67.1	52.0	31.0
CLIP	**17.9**	**10.3**	13.9	9.9	100	19.2	65.1	52.4
TePOT	92.7	22.9	38.6	19.5	18.7	100	85.7	33.9
TOP	63.9	15.3	27.5	13.0	42.4	68.4	100	49.5
TOPcl	30.6	8.9	37.2	9.5	50.3	33.9	66.0	100

Table 2. Overlaps (%) of each thesaurus over a reference thesaurus

Manual validation of the synsets was also performed for CLIP and TOP. Random samples of approximately 50 synsets[8] from each thesaurus were generated and given to ten reviewers who classified each synset as: correct (1), if, in some context, all the words of the synset could have the same meaning, or incorrect (0), if at least one word of the synset could not have the same meaning as the others. The reviewers were advised to look for the possible meanings of each word in different dictionaries. Still, if they did not know how to classify the synset, they had a third option, N/A (2).

In the end, 519 synsets of CLIP and 480 of TOP were validated. Besides the average validation results and the agreement rates, Table 3 contains the results considering only synsets of ten or less words (CLIP' and TOP'). The precision results are improved if the automatically created thesaurus is merged with the ones created manually, and also when bigger synsets are ignored. The latter are worst because they tend to bring together more than one concept sharing at least one word.

	Sample	Correct	Incorrect	N/A	Agreement
CLIP	519 sets	65.8%	31.7%	2.5%	76.1%
CLIP'	310 sets	81.1%	16.9%	2.0%	84.2%
TOP	480 sets	83.2%	15.8%	1.0%	82.3%
TOP'	448 sets	86.8%	12.3%	0.9%	83.0%

Table 3. Results of manual synset validation.

[8] A synset could not be in more than one sample and, to minimise the validation effort, synsets with more than 50 words were not included.

3 CONCLUDING REMARKS

We have shown that, using clustering techniques, it is possible to create a thesaurus suitable to be used as a conceptual base for a lexical resource. We are aware that the word sense divisions present in dictionaries and lexical ontologies are most of the times artificial, but this trade-off is often needed to increase the usability of computational broad-coverage lexical resources. Nevertheless, using the proposed clustering procedure, it should be possible to append the probability of inclusion of each word in a synset, which would be a first approach to handle the referred problem, without usability losses.

REFERENCES

[1] Bento Carlos Dias-Da-Silva and Helio Roberto de Moraes, 'A construção de um thesaurus eletrônico para o português do Brasil', *ALFA*, **47**(2), 101–115, (2003).

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

[3] David Gfeller, Jean-Cédric Chappelier, and Paulo De Los Rios, 'Synonym Dictionary Improvement through Markov Clustering and Clustering Stability', in *Proc. Intl. Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, pp. 106–113, (2005).

[4] Hugo Gonçalo Oliveira, Diana Santos, and Paulo Gomes, 'Relations extracted from a portuguese dictionary: results and first evaluation', in *Local Proc. 14th Portuguese Conference on Artificial Intelligence (EPIA)*, (2009).

[5] Marti A. Hearst, 'Automatic acquisition of hyponyms from large text corpora', in *Proc. 14th conference on Computational Linguistics*, pp. 539–545, Morristown, NJ, USA, (1992). ACL.

[6] Diana Santos, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros, and Rosário Silva, 'O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL', in *Actas do XXV Encontro Nacional da Associação Portuguesa de Linguística*, (2009). forthcomming.

[7] J. Teixeira, L. Sarmento, and E. Oliveira, 'Comparing verb synonym resources for portuguese', in *Computational Processing of the Portuguese Language, 9th International Conference, Proc. (PROPOR 2010)*, (2010).

[8] S. M. van Dongen, *Graph Clustering by Flow Simulation*, Ph.D. dissertation, University of Utrecht, The Netherlands, 2000.