Automatic Creation of a Conceptual Base for Portuguese using Clustering Techniques

FCTUC FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE DE COIMBRA

Hugo Gonçalo Oliveira & Paulo Gomes
{hroliv,pgomes}@dei.uc.pt

Cognitive & Media Systems Group
CISUC, University of Coimbra, Portugal

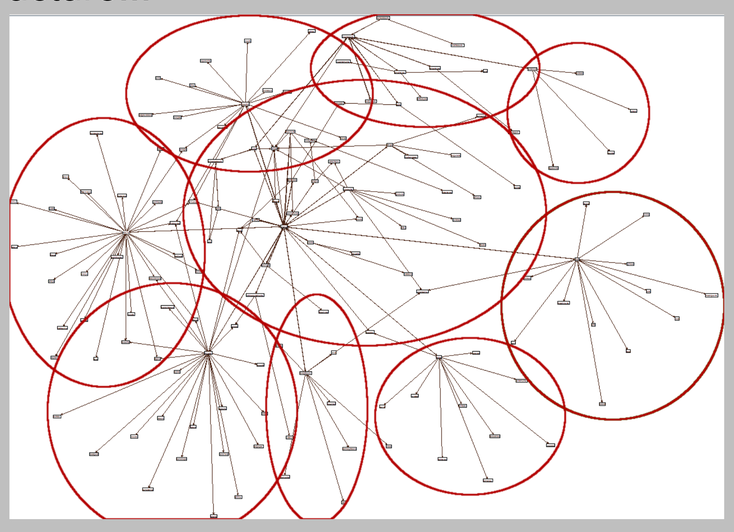
cognitive & media systems

Introduction

- Today's applications demand better access to knowledge on words and their meanings.
- **Lexical ontologies** are typically handcrafted, so...
  - Creation and maintenance involves much human effort
  - Automatic construction from text provides: less intensive labour, easier maintenance and higher coverage
  - As a trade-off for lower, still acceptable, correction
- In information extraction from text...
  - Systems output relational triples relating terms, *a* RELATED_TO *b*
  - Which does not handle ambiguity, since a word may have different meanings.
  - Alternative: synset-based structure (e.g. WordNet [1])
  - Synsets describe concepts as a group of synonymous words
- Research goals:
  - **Automatic creation of a broad-coverage thesaurus for Portuguese.**

Synset discovery

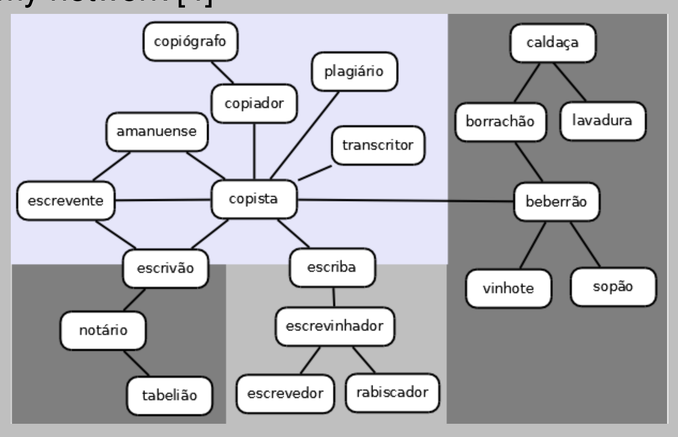
- Synonymy networks extracted from dictionaries tend to have a clustered structure...



- Following [2]...
  1. Split the original network into sub-networks
  2. Calculate the frequency-weighted adjacency matrix F of each sub-network;
  3. $F_{ij} = F_{ij} + F_{ij} * \delta$, $-0.5 < \delta < 0.5$;
  4. Run MCL [3], with $\gamma = 1.6$, over F for 30 times;
  5. Use the (hard) clustering from each run to create P, a matrix with the probabilities of each pair of words in F belonging to the same cluster;
  6. 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.

Experimentation

- Nouns in:
  - PAPEL's synonymy network [4]



  - TeP [5] synsets
  - OpenThesaurus (OT) synsets
- Resulting thesaurus:
  1. CLIP: clustered PAPEL
  2. CleP: clustered TeP synonymy network[a]
  3. ClOT: clustered OT synonymy network
  4. TePOT: TeP + OT[b]
  5. TOP: TeP + OT + CLIP
  6. TOPcl: CleP + ClOT + CLIP

[a] Synonymy instances extracted from the synsets.
[b] Synsets from both thesaurus maximizing $Jaccard(A, B) = A \cap B / A \cup B$, were merged.

Thesaurus comparison

	Words			Synsets		
	Quantity	Ambiguous	Most ambig.	Quantity	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

Thesaurus overlaps

- Despite being all Portuguese broad-coverage resources, TeP, OT and PAPEL are more complementary than overlapping.

	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

Manual synset validation

- CLIP' and TOP' only consider synsets with 10 or less words.

	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%

Concluding remarks

- Clustering a dictionary synonymy network is a suitable alternative for establishing synsets, but...
  - Since word sense divisions are most of the times artificial
  - Trade-off needed to increase the usability of resources
  - **Future**: append the probability of inclusion of each word in a synset
- In **Onto.PT** [6], dictionaries, thesaurus and corpora are being exploited for the creation of a Portuguese lexical ontology
- **Next**: Moving from term-based triples to synset-based triples [7]

Acknowledgements

Hugo Gonçalo Oliveira is supported by FCT scholarship grant SFRH/BD/44955/2008. We would like to thank Hernani Costa and the KDigg team for their participation in project discussions and in the synset validation.

References

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

[2] D. Gfeller, J.-C. Chappelier, and P. D. L. Rios, "Synonym Dictionary Improvement through Markov Clustering and Clustering Stability," in *Proc. of International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, pp. 106–113, 2005.

[3] S. M. van Dongen, *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, The Netherlands, 2000.

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

[5] B. C. Dias-Da-Silva and H. R. de Moraes, "A construção de um thesaurus eletrônico para o português do Brasil," *ALFA*, vol. 47, no. 2, pp. 101–115, 2003.

[6] H. Gonçalo Oliveira and P. Gomes, "Onto.pt: Automatic construction of a lexical ontology for portuguese," in *Proc. 5th European Starting AI Researcher Symposium (STAIRS)*, 2010.

[7] H. Gonçalo Oliveira and P. Gomes, "Towards the automatic creation of a wordnet from a term-based lexical network," in *Proc. ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, 2010.