Automatic Discovery of Fuzzy Synsets from Dictionary Definitions

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

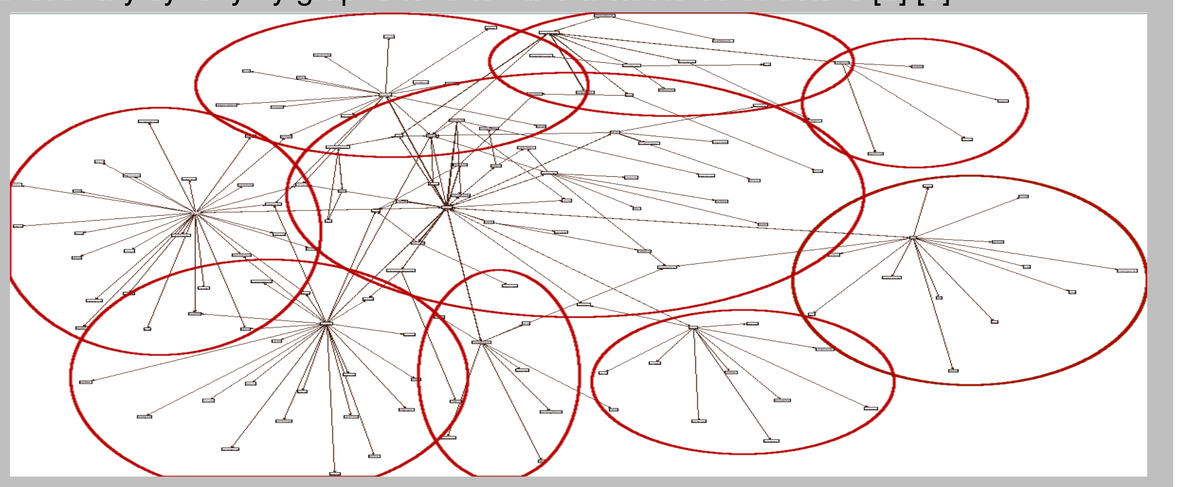
cognitive & media systems

Introduction

- Wordnets (as Princeton WordNet [1]) play a crucial role in several NLP tasks (e.g. WSD, QA, determination of similarities ...)
  - **synsets**: groups of synonymous word senses:
    - A: (machine.1)
    - B: (computer.1, computing_machine.1, computing_device.1, data_processor.1)
    - C: (chip.7, microchip.1, microprocessor_chip.1)
  - semantic relations between synsets
    - A hypernym-of B, C part-of B, ...
- From a linguistic point of view...
  - Word senses are not discrete
  - Typically complex and overlapping
  - Sense division is usually artificial
- Goal:*reality-usability trade-off*
  - Exploit redundancy in dictionaries
  - Deal with uncertainty
  - **Fuzzy synsets**

From dictionary definitions to fuzzy synsets

- **Extraction of synpairs**
  - **mind**, n: *brain, head, intellect*
    - (*brain*, *mind*) (*head*, *mind*) (*intellect*, *mind*)
  - **computer**, n: *the same as computing machine*
    - (*computing_machine*, *computer*)
- **Discovery of fuzzy synsets in synonymy graphs**
  - Synonymy graph $G = (N, E)$, with $|N|$ nodes and $|E|$ edges, $E \subset N^2$
  - $w_a \in N$ is a word.
  - $E(w_a, w_b, \sigma_{ab})$ means synpair (w_a, w_b) was extracted σ_{ab} times.
  - $\vec{w_a}$ = adjacency vector of w_a
  - Dictionary synonymy graphs tend to have a clustered structure [2] [3]



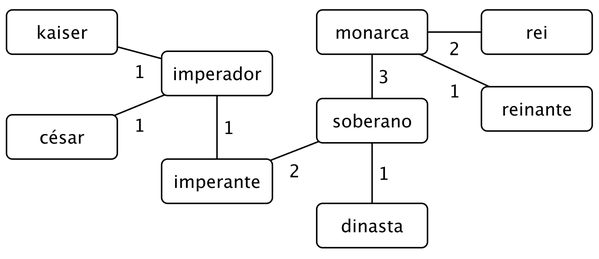
- Discovery of fuzzy synsets:
  1. Matrix M ($|N| \times |N|$)
  2. $M_{ij} = sim(\vec{w}_i, \vec{w}_j)$ (expression 1)
  3. Normalise columns M_j, $\sum_{k=0}^{|M_j|} M_{jk} = 1$
  4. Each row M_i is a cluster F_i with the words $w_j : M_{ij} > 0$, $\mu_{F_i}(w_j) = M_{ij}$.
  5. $\forall(F_i, F_j : F_i \cup F_j = F_j \wedge F_i \cap F_j = F_i) \;\; F_k = F_j \wedge \forall(w_l \in F_j)\mu_{F_k}(w_l) = \mu_{F_i}(w_l) + \mu_{F_j}(w_l)$

$$sim(\vec{x}, \vec{y}) = cos(\vec{x}, \vec{y}) = \frac{\vec{x}.\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=0}^{|N|} pmi(x, w_i) \times pmi(y, w_i)}{\sqrt{\sum_{i=0}^{|N|} pmi(x, w_i)^2 \times \sum_{i=0}^{|N|} pmi(y, w_i)_i^2}} \quad (1)$$

$$pmi(x, y) = \frac{\frac{\sigma_{xy}}{S}}{\frac{\sum_{j=0}^{|N|} \sigma_{xj}}{S} \times \frac{\sum_{i=0}^{|N|} \sigma_{iy}}{S}} \times df, \quad S = \sum_{i=0}^{|N|} \sum_{j=0}^{|N|} \sigma_{ij}, \quad df(x, y) = \frac{\sigma_{xy}}{\sigma_{xy} + 1} \times \frac{min\left(\sum_{j=0}^{|N|} \sigma_{xj}, \sum_{i=0}^{|N|} \sigma_{iy}\right)}{min\left(\sum_{j=0}^{|N|} \sigma_{xj}, \sum_{i=0}^{|N|} \sigma_{iy}\right) + 1}$$

A fuzzy thesaurus for Portuguese

- Synpairs extracted from:
  - Dicionário da Língua Portuguesa (through PAPEL [4])
  - Dicionário Aberto (DA) [5]
  - Portuguese Wiktionary
- Example



$synset_1$	μ	$synset_2$	μ
kaiser (*kaiser*)	1.0	reinante (*regnant*)	1.0
césar (*caeser*)	1.0	rei (*king*)	1.0
imperador (*emperor*)	0.95	monarca (*monarch*)	0.86
imperante (*dominant*)	0.57	soberano (*sovereign*)	0.85
dinasta (*dynast*)	0.24	dinasta (*dynast*)	0.76
soberano (*sovereign*)	0.15	imperante (*dominant*)	0.43
monarca (*monarch*)	0.14	imperador (*emperor*)	0.05

Polysemic words

Word	Concept	Fuzzy synsets
pasta	money	*arame*(0.6774), *zerzulho*(0.6774), *metal*(0.6774), *carcanhol*(0.6774), *pecunia*(0.6774), *bagarote*(0.6774), *pecuniária*(0.6774), *cunques*(0.6774), *matambira*(0.6774), *bagalho*(0.6774), *cacau*(0.6774), *calique*(0.6774), *níquel*(0.6774), *mussuruco*(0.6774), *massaroca*(0.6774), *baguines*(0.6774), *pastel*(0.6774), *dieiro*(0.6774), *pilim*(0.6774), *gimbo*(0.6735), *chelpa*(0.6735), *pecúnia*(0.6735), *patacaria*(0.6735), *pataco*(0.6347), *bagalhoça*(0.62), *bago*(0.6181), *cobre*(0.6173), *jimbo*(0.5953), *guines*(0.5903), *pasta*(0.5657), *maquia*(0.5243), *grana*(0.5226), *painço*(0.517), *jibungo*(0.517), *numerário*(0.5145), *dinheiro*(0.5139), *fanfa*(0.4617), *posses*(0.4604), *finanças*(0.4425), *ouro*(0.4259),...
	file	*diretório*(1.0), *dossier*(0.9176), *pasta*(0.1118), ...
	mixture	*amálgama*(0.09279), *dossier*(0.08130), *landoque*(0.05162), *angu*(0.04271), *pot-pourri*(0.03949), *marinhagem*(0.03722), *mosaico*(0.03648), *cocktail*(0.03480), *mixagem*(0.02688), *cacharolete*(0.02688), *macedónia*(0.02688), *comistão*(0.02374), *colectânea*(0.02317), *anguzada*(0.02205), *caldeação*(0.02108), *mistura*(0.02032), *moxinifada*(0.01976), *imisção*(0.01917), *massamorda*(0.01845), *pasta*(0.01827), *farragem*(0.01779), *matalotagem*(0.01397), *misto*(0.01280), *salsada*(0.01262), *ensalsada*(0.01050)
	briefcase	*maleta*(0.0759), *saco*(0.0604), *maco*(0.054), *fole*(0.0154), *pasta*(0.0128), ...
cota	mother	*mamãe*(0.8116), *mamã*(0.8116), *nai*(0.7989), *malúrdia*(0.7989), *darona*(0.7989), *mamana*(0.7989), *velha*(0.7989), *mãe-de-famílias*(0.7989), *ti*(0.7989), *mare*(0.6503), *naia*(0.5549), *uiara*(0.5549), *genetriz*(0.5549), *mãe*(0.5221), *madre*(0.2749), *cota*(0.2407), ...
	father	*palúrdio*(0.6458), *dabo*(0.6458), *genitor*(0.6458), *painho*(0.6458), *benfeitor*(0.6458), *papai*(0.6183), *papá*(0.6169), *tatá*(0.4934), *pai*(0.3759), *primogenitor*(0.3543), *velhote*(0.2849), *velho*(0.2817), *cota*(0.1463), *progenitor*(0.08416015), *ascendente*(0.062748425)
	quota	*colecta*(0.6548), *quota*(0.5693), *contingente*(0.309), *pagela*(0.2304), *prestação*(0.1723), *cota*(0.1655), *mensalidade*(0.0908), *quinhão*(0.0605),...

From a fuzzy to a simple thesaurus

- Remove from fuzzy synsets all words with $\mu_{F_i}(w_j) < \theta$

θ	Noun words				Noun synsets				
	Total	Ambig.	Avg(senses)	Max(senses)	Total	Avg(size)	$size = 2$	$size > 25$	$max(size)$
0.025	39,350	21,730	3.18	18	13,344	9.39	3,921	576	80
0.05	39,288	17,585	1.86	9	12,416	5.89	4,224	119	62
0.075	38,899	12,505	1.44	7	12,086	4.64	4,878	47	58
0.1	38,129	8,447	1.26	6	11,748	4.10	5,201	34	58
0.15	35,772	4,198	1.12	4	11,044	3.64	5,248	16	58
0.25	30,266	1,343	1.04	3	9,830	3.22	5,095	10	58
0.5	22,203	0	1.0	1	8,004	2.77	5,011	3	47

- Comparison with Portuguese public thesauri
  - TeP 2.0 [6]
  - OpenThesaurus.PT (OT.PT)

Thesaurus	POS	Words				Synsets				
		Quant.	Ambig.	Avg(senses)	Max(senses)	Quant.	Avg(size)	$size = 2$	$size > 25$	$max(size)$
OT.PT	N	6,110	485	1.09	4	1,969	3.38	778	0	14
	V	2,856	337	1.13	5	831	3.90	226	0	15
	Adj	3,747	311	1.09	4	1,078	3.80	335	0	17
TeP 2.0	N	17,158	5,805	1.71	20	8,254	3.56	3,079	0	21
	V	10,827	4,905	2.08	41	3,978	5.67	939	48	53
	Adj	14,586	3,735	1.46	19	6,066	3.50	3,033	19	43
Padawik (fuzzy)	N	39,354	24,343	7.78	46	20,102	15.23	3,885	3,756	109
	V	11,502	10,411	14.31	42	7,775	21.17	307	2,411	89
	Adj	15,260	10,636	10.36	43	8,896	17.77	1,326	2,157	109
Padawik ($\theta = 0.075$)	N	38,899	12,505	1.44	7	12,086	4.64	4,878	47	58
	V	11,070	5,717	1.76	7	4,198	4.63	1,189	14	49
	Adj	14,964	6,644	1.69	6	5,666	4.45	1,980	11	46

- Manual validation of noun synsets

Classification	Synsets (sample = 440 × 2 sets)		Synpairs (sample = 440 × 2 sets)	
Correct	646	(73.4%)	660	(75.0%)
Incorrect	231	(26.3%)	218	(24.8%)
N/A	3	(0.3%)	2	(0.2%)
Agreement	364	82.7%	366	83.2%

- Thesaurus T, Synsets $S_i \in T$
- Synpairs (w_a, w_b) : $w_a \in S_i \wedge w_b \in S_i$
- Two human judges
- Correct = in some context, all the words in the synset/synpair might have the same meaning.

Acknowledgements

This work was developed in the scope of the project Onto.PT [7].

We would like to thank all the people involved in the synset validation and Leticia Antón Pérez for developing the Wiktionary parser. Hugo Gonçalo Oliveira is supported by FCT, grant SFRH/BD/44955/2008, co-funded by FSE.

References

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

[2] D. Gfeller, J.-C. Chappelier, and P. D. L. 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.

[3] E. Navarro, F. Sajous, B. Gaume, L. Prévot, S. Hsieh, T. Y. Kuo, P. Magistry, and C. R. Huang, "Wiktionary and nlp: Improving synonymy networks," in *Proc. 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, (Suntec, Singapore), pp. 19–27, ACL, 2009.

[4] H. Gonçalo Oliveira, D. Santos, and P. Gomes, "Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação," *Linguamática*, vol. 2, no. 1, pp. 77–93, 2010.

[5] A. Simões and R. Farinha, "Dicionário Aberto: Um novo recurso para PLN," *Vice-Versa*, September 2010.

[6] E. G. Maziero, T. A. S. Pardo, A. D. Felippo, and B. C. Dias-da-Silva, "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)*, pp. 390–392, 2008.

[7] H. Gonçalo Oliveira and P. Gomes, "Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese," in *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, IOS, 2010.