The original publication is available at: Springer
DOI: 10.1007/978-3-642-28885-2_27

The role of language registers in polarity propagation

António Paulo Santos[1], Hugo Gonçalo Oliveira[2], Carlos Ramos[1], and Nuno C. Marques[3]

[1] GECAD, Institute of Engineering - Polytechnic of Porto, Portugal
[2] CISUC, University of Coimbra, Portugal
[3] DI-FCT, Universidade Nova de Lisboa, Monte da Caparica, Portugal
pgsa (at) isep.ipp.pt, hroliv (at) dei.uc.pt, csr (at) isep.ipp.pt, nmm (at) di.fct.unl.pt

Abstract. This paper presents work on the automatic creation of a polarity lexicon based on a lexical-semantic network. During this work, we noticed that the language registers of a relation should be considered in polarity propagation. After analysing the possible registers and performing some experiments, our intuition was confirmed – there are registers that should invert the transmitted polarity (irony), while others always transmit negative polarity (pejorative, disparaging).

Keywords: sentiment analysis, opinion mining, polarity, lexical-semantic resources, language registers

1 Introduction

Polarity lexicons (e.g. SentiWordNet [1], for English, or SentiLex [10], for Portuguese) are useful resources for sentiment analysis, and their creation is one of the main research lines in this area. They consist of a set of lexical items (words or expressions) and the typical sentiment they transmit, usually referred to as polarity. Common values for polarity are positive, neutral and negative.

Most approaches on the automatic construction of sentiment lexicons fall in one of two categories: corpora-based approaches (e.g. [3][11][4][6]), that explore the co-occurrence of words in large collections of texts; and dictionary or wordnet-based approaches (e.g. [5][7][9][8]), that exploit information provided by lexical-semantic resources.

In this work, while applying a polarity propagation algorithm [8] to PAPEL [2], a lexical-semantic network for Portuguese, we noticed that there were "erroneous" synonymy relations that were originating propagation errors. We analysed these relations and verified that some of them had additional language information fields, typically found in dictionaries, namely: domain, register and variant. Also, we observed that some of the registers could change the simple polarity propagation. After performing several experiments, we confirmed that these registers provide valuable information for polarity propagation. Therefore, when available, they should be handled properly.

We start this paper by describing, briefly, the polarity propagation algorithm used. Then, we present PAPEL, with special focus to the information fields its relations contain, and to the synonymy relations, which were the ones exploited in this work. Before concluding, we present the performed experiments, which confirmed our assumptions regarding the correct treatment of the registers.

2 Polarity Propagation

The polarity propagation algorithm used in this work [8] sees a dictionary (or a lexical-semantic network) as a graph, where lexical items are the nodes, and the edges connecting the items represent the semantic relations. It starts with a small set of seed words, for instance, five positive and five negative, manually labelled. Then, the algorithm visits every lexical item in the graph by a breadth-first traversal. The polarity of the lexical items is iteratively propagated to the unlabelled items. While most relations tend to preserve polarity (e.g. synonymy) others invert polarity (e.g. antonymy).

3 PAPEL

PAPEL [2] is a public domain lexical-semantic network, automatically extracted from a proprietary dictionary. PAPEL 2.0 contains about 97,000 lexical items – 55,900 nouns, 24,000 verbs, 21,000

adjectives and 1,400 adverbs – and about 198,000 connections between them. The latter denote semantic relations and are represented as triples, with the following structure:

```
arg1 RELATION_NAME arg2     domain;register;variant
```
(e.g. `divertimento SINONIMO_N_DE alegria`)

A triple indicates that one sense of the lexical item in the first argument (`arg1`) is related to one sense of the lexical item in the second argument (`arg2`) by means of a relation identified by `RELATION_NAME`. In PAPEL, the name of the semantic relation defines the part-of-speech of its arguments. Furthermore, some of the triples have additional information fields, typically found in dictionaries.

3.1 Additional fields

In this work, we used PAPEL 2.0[4]. In this version, some of the triples have the following additional language information fields, also obtained from the dictionary definitions, that provide information about the way words are, or may be used:

- **Register**: the situation or context where the definition holds.
- **Domain**: the specific sphere of knowledge where the definition is common or valid.
- **Variant**: the Portuguese variant where the definition applies.

Table 1 shows some of the possible values for these fields, their meaning and the number of triples of PAPEL 2.0 in which they occur (Occurrences). Whereas domain and variant do not seem relevant for polarity propagation, some of the registers might be, as we will show in the next section.

Table 1. Information fields, some of their possible values and number of occurrences

Field	Value	Meaning	Occurrences
Domain	bot.	botany	6,397
	zool.	zoology	2,515
	medic.	medicine	2,256
	quim.	chemistry	1,001
	mus.	music	669
Register	fig.	figurative	7,357
	pop.	popular	2,072
	coloq.	informal	747
	pej.	pejorative	431
	depr.	disparaging	251
	vulg.	vulgarism	75
	cal.	slang	31
	irón.	ironic	17
Variant	Bras.	Brazil	2,092
	reg.	regionalism	1,900
	Ang.	Angola	386
	Moçamb.	Mozambique	247

3.2 Synonymy in PAPEL

PAPEL contains several semantic relations. In this work we exploited the synonymy relation. In PAPEL 2.0, there are 79,161 relational triples denoting synonymy – 37,452 between nouns, 21,465 between verbs, 19,073 between adjectives and 1,171 between adverbs.

On the context of polarity propagation, synonymous lexical items tend to have the same polarity. However, some of the registers presented in the previous section might change the typical polarity of two lexical items connected by synonymy. After analysing different triples with different registers, we divided the register values into two types, according to their ability of changing polarity transmitted in synonymy relations:

- **Polarity keepers**: registers that preserve the regular behaviour of the synonymy relation, in terms of transmitted polarity.
- **Polarity modifiers**: registers that have the ability of changing the regular behaviour of the synonymy relation, in terms of polarity transmitted.

[4] Available through http://www.linguateca.pt/PAPEL/

While registers like figurative or informal (see examples 1 and 2 below) belong to the first type, based on observation, we assume that irony, pejorative and disparaging registers belong to the second (see examples 3, 4 and 5 below). In section 4, the former assumption is the main goal of our experimentation.

(1) `caro SINONIMO_ADJ_DE salgado ;fig;`

(2) `excelente SINONIMO_ADJ_DE porreiro ;coloq;`

(3) `punição SINONIMO_N_DE recompensa ;irón;`

(4) `concubina SINONIMO_N_DE fêmea ;pej;`

(5) `enfraquecer SINONIMO_V_DE efeminar ;depr;`

Still, irony behaves differently than pejorative and disparaging registers. The presence of irony in a synonymy relation means that the connected lexical items are only synonymous in an ironic context, whereas in most typical contexts they have an opposite meaning. Therefore, in order to propagate the typical polarity, we can handle ironic synonymy relations as antonymy. On the other hand, pejorative and disparaging synonymy relations always transmit negative polarity. So, for the former registers, if the propagated polarity is positive or neutral, it becomes negative. Otherwise, the negative polarity is preserved.

There are other interesting registers in PAPEL which we thought about exploiting for polarity propagation. For instance, vulgarism (see examples 6 and 7 below) and slang (see examples 8 and 9 below) are usually associated with negative polarities (as 6 and 8), which suggests that they belong to the polarity modifiers group. However, even though less common, there are as well positive relations with these registers (as 7 and 9).

(6) `insignificância SINONIMO_DE merdice ;vulg;`

(7) `erecção SINONIMO_DE tesão ;vulg;`

(8) `excremento SINONIMO_N_DE trampa ;cal;`

(9) `força SINONIMO_N_DE tusa ;cal;`

4 Experiments

So far, our assumptions regarding the impact of language registers in polarity propagation are the following:

1. Ironic relations invert the propagated polarity
2. Pejorative relations always propagate negative polarity
3. Disparaging relations always propagate negative polarity

The goal of our experimentation is thus to confirm the former assumptions. We assess the performance of polarity classification when, in propagation, the target registers in the synonymy triples are handled differently. Therefore, we ran the propagation algorithm with different numbers of seeds, obtained from the manually annotated SentiLex-PT01 [10], a public sentiment lexicon for Portuguese with 6,321 adjectives, 3,585 of which have their polarity manually classified. The algorithm only stops when there are no more nodes to visit.

The results shown in table 2 were computed after comparing the automatically labelled adjectives in PAPEL with the manually labelled adjectives in SentiLex-PT01. Also, the results were measured following three different criteria:

1. Ignoring all triples with registers that would be exploited (iron, pej, depr);
2. Considering all the triples of the target register in the graph construction but ignoring their extra information during the polarity propagation. The triples of all the other exploited registers are ignored;
3. Considering all the triples, and handling the target register properly. The triples of all the other exploited registers are ignored.

Table 2. Average results, according to the handled registers (Register = target registers; $\overline{Class}$ = average number of classified words; $\overline{Eval}$ = average number of evaluated adjectives; $\overline{Acc}$ ± SD = average accuracy ± standard deviation).

Registers	**Seeds**	Criteria 1			Criteria 2			Criteria 3		
		$\overline{Class}$	$\overline{Eval}$	$\overline{Acc}$ $\pm SD$	$\overline{Class}$	$\overline{Eval}$	$\overline{Acc}$ $\pm SD$	$\overline{Class}$	$\overline{Eval}$	$\overline{Acc}$ $\pm SD$
iron	3	30,783	2,308	57.67 ±11.0	30,781	2,307	57.63 ±11.0	30,781	2,306	**57.73** ±10.9
	12	29,857	2,315	64.64 ±7.9	29,854	2,312	64.39 ±7.8	29,855	2,315	**64.73** ±7.9
pej	3	"	"	"	30,851	2,311	57.60 ±10.9	30,852	2,310	**58.73** ±11.0
	12	"	"	"	29,978	2,322	64.80 ±8.0	30,049	2,331	**65.14** ±7.3
depr	3	"	"	"	30,829	2,305	62.31 ±04.2	30,841	2,307	**62.81** ±3.9
	12	"	"	"	29,865	2,319	65.20 ±06.6	29,871	2,323	**65.48** ±6.4
All	3	30,783	2,308	57.67 ±11.0	30,898	2,308	62.12 ±4.2	30,917	2,312	**63.77** ±3.7
	12	29,857	2,315	64.64 ±7.9	29,977	2,322	65.13 ±6.5	30,045	2,338	**66.03** ±5.9

Table 2 shows the average number of classified, evaluated words, and accuracy for each criteria, obtained while handling the target registers differently. Each combination "target register(s)+seeds" was performed for 10 runs varying the seed words. We present the accuracies for three seeds (one positive, one negative and one neutral), and 12 (four positive, four negative and four neutral), which confirms that, as expected, the number of seeds is proportional to accuracy.

As for the target registers, using only irony (iron) addresses our first assumption, only pejorative (pej) addresses the second assumption, only disparaging (depr) our third assumption, and, finally, using the three exploited registers (All) intends to show the overall performance improvement.

All our assumptions were confirmed by these experiments, as the proper handling of each and all the registers lead to higher accuracies. When the three exploited registers are handled according to our assumptions, there is an improvement on accuracy: about 1.5% when using three seeds and 0.9% when using 12 seeds. Even though the improvements might not look significant, we must have in mind that the number of target registers in synonymy triples is almost residual – 11 irony registers, 159 pejorative, and 88 disparaging, in a total of 79,000 triples. Also, as the current version of SentiLex only contains adjectives, these were the only evaluated words, which are always about 7.5% of the classified words.

5 Concluding remarks

We confirmed that information about the language register of semantic relations might be valuable for polarity propagation. Since synonymy relations connect lexical items with the same meaning, it is expectable that both items have the same (typical) polarity. However, we noticed that some registers might change the transmitted polarity and should thus be handled properly.

After identifying the registers that modify polarity, we did some experiments with a polarity propagation algorithm in a lexical-semantic network and confirmed that our intuition was true. Therefore, once available, information about language registries, more precisely irony, pejorative and disparaging markers, should be exploited in the automatic construction of polarity lexicons.

Acknowledgements

António Paulo Santos is supported by the FCT grant SFRH/BD/47551/2008. Hugo Gonçalo Oliveira is supported by the FCT grant SFRH/BD/44955/2008 co-funded by FSE.

References

1. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation. pp. 417–422. LREC'06 (2006)
2. Gonçalo Oliveira, H., Santos, D., Gomes, P.: Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. Linguamática 2(1), 77–93 (May 2010)
3. Hatzivassiloglou, V., Mckeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of 35th Annual Meeting of the Association for Computational Linguistics (ACL'97). pp. 174–181. Association for Computational Linguistics, Madrid, ES (1997)
4. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of HTML documents. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 1075–1083 (2007)

5. Kamps, J., Mokken, R.J., Marx, M., de Rijke, M.: Using WordNet to measure semantic orientation of adjectives. In: Proceedings of the 4th Intl. Conference on Language Resources and Evaluation. LREC'04, vol. IV, pp. 1115–1118. ELRA, Paris, France (2004)
6. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 355–363. EMNLP'06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006)
7. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th Intl. conference on Computational Linguistics. pp. 1267–1373. COLING'04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
8. Paulo-Santos, A., Ramos, C., Marques, N.: Determining the polarity of words through a common online dictionary. In: Progress in Artificial Intelligence, Proceedings of 15th Portuguese Conference on Artificial Intelligence (EPIA'11). LNCS, vol. 7026, pp. 649–663. Springer (2011)
9. Rao, D., Ravichandran, D.: Semi-supervised polarity lexicon induction. In: Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09). pp. 675–682. Athens, Greece (2009)
10. Silva, M., Carvalho, P., Costa, C., Sarmento, L.: Automatic expansion of a social judgment lexicon for sentiment analysis. Tech. rep., Faculdade de Ciências da Universidade de Lisboa (2010)
11. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 417–424. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)