Finding Compositional Rules for Determining the Semantic Orientation of Phrases

António Paulo Santos[1], Carlos Ramos[1], and Nuno Marques[2]

[1]GECAD, Institute of Engineering - Polytechnic of Porto, Portugal
[2]DI-FCT, Universidade Nova de Lisboa, Monte da Caparica, Portugal
pgsa@isep.ipp.pt, csr@isep.ipp.pt, nmm@fct.unl.pt

Abstract. The semantic compositionality principle states that the meaning of an expression can be determined by its parts and the way they are put together. Based on that principle, this paper presents a method for finding the set of compositional rules that best explain the *positive*, *negative*, and *neutral* semantic orientation (SO) of two-word phrases, in terms of the SO of its words. For instance, the phrase "*fake contract*" has a negative SO. A corpus was built for evaluating the proposed method and several experiences are reported. We also use the conditional probability as a reliability measure of the compositional rules. The reliability of the learned rules ranges from 60.44% for verb-noun phrases to 100% for adjective-adjective phrases.

Keywords: semantic orientation, compositional rules, sentiment analysis

1 Introduction

The *semantic orientation* (SO) or *polarity* of a word indicates the direction the word deviates from the norm for its semantic group or lexical field [1]. This notion can be generalized to bigger text units (e.g. a word, a sentence, a document). Past work has shown that the SO of two-word phrases has an important role in the determination of the SO of bigger text units [2–4]. Determining the SO of two or more sequences of words (e.g. determining if "*fake contract*" is *positive*, *negative*, or *neutral*) has been researched following different approaches [2, 4–9]. Promising approaches rely on the principle of *semantic compositionality*, implicitly [6, 8] or explicitly [3, 4, 7, 9, 10]. Generally, the SO composition is modeled by a set of rules in several works [6, 7, 10, 11]. These rules are often created manually and defined based on human knowledge of language (human experts). An alternative to manual rules may be to learn the rules from data as we propose in this paper.

This paper is organized as follows. Section 2 describes the proposed approach for finding the compositional rules and describes the resources required to apply it. Section 3 reports the results of applying the previous approach to a set of two-word phrases. We close with the conclusion in section 4.

adfa, p. 1, 2011.

2 SO Compositionality

A popular approach for modeling the SO composition of two-word phrases is through a set of rules manually created [3, 7, 10, 11]. For example, a possible hand-coded rule could be:

$$\text{ADJ}^{-}\ \text{N}^{0} \rightarrow \text{Phrase}^{-}$$

Meaning that a two-word phrase composed by a “negative adjective” followed by a “neutral noun”, should be classified as negative, such in the “*fake*$^{-}$ *contract*^{0}” phrase. This rule is intuitive. However, this is not always the case. Furthermore, there may be many rules to consider. Our approach instead is to find the best compositional rules from annotated corpus (section 2.1), with help of a lexicon (section 2.2) and applying the algorithm described in section 2.3. Unfortunately, to the best of our knowledge, no corpus currently exists to train and evaluate two-word phrases. So, we built a corpus as described next.

2.1 Corpus of Two-word Phrases

The corpus of two-word phrases[1] was built by the following steps: First, we collected a set of sentences. We collected all the available Portuguese news headlines on *Wikinews*, in three categories.[2] There were a total of 4,646 news headlines, 982 of them in the *economy and business* category, 2,946 in the *politics and conflicts category* and 718 in the *health* category.

The second step was to apply a part-of-speech tagger[3] to each sentence and extract the two-word phrases using the extraction patterns shown in **Table 1**. On that table, the first five patterns were used in [2], and the remaining were defined by us. These patterns were chosen because we believe that they are the most relevant to study the semantic orientation of two words phrases in Portuguese language, present in our corpus.

Table 1. Statistics of the extracted phrases (N = noun, ADJ = adjective, ADV adverb, VRB = verb, *prp* = preposition, *art* = article, *num* = numerals)

Pattern	#phrases	Sem. Orientation		
		-	0	+
ADJ N	433	71	300	62
ADV ADJ	38	9	13	16
ADJ ADJ	56	4	50	2
N ADJ	1,753	496	860	397
ADV VRB	240	37	183	20
N prp N	2,112	703	961	448
VRB (prp|art|num)? N	3,147	999	1,151	997
Total:	7,779	2,319	3,518	1,942

[1] https://github.com/i000313/phd.polarity.compositionalRules

[2] https://pt.wikinews.org/ - Wikinews in Portuguese

[3] http://opennlp.sourceforge.net/ - OpenNLP POS Tagger 1.5.2

The third step was to manually classify each extracted phrase as *negative* (-), *neutral* (0) or *positive* (+), according the majority polarity of that phrase in the corpus. The total number of annotated phrases is shown in **Table 1** and **Table 2** shows some examples of them. During this third and manual step, we also reviewed the part-of-speech tags assigned to the phrases on the second step. Phrases with wrong part-of-speech were discarded.

Table 2. Example of extracted phrases with their semantic orientation annotated by a human

Pattern	Example of Extracted Phrases
ADJ N	*[possível aliança]*$^+$ (possible alliance), [*novas regras*]^0 (new rules)
N prp N	[*acordo de cessar-fogo*]$^+$ (cease-fire agreement)
VRB (prp|art|num)? N	[*sobrevive a tiro*]$^+$ (survives shooting), [*acusado de corrupção*]$^-$ (accused of corruption)

2.2 Lexicon

Our approach relies on a lexicon of words, where each entry is associated with its part-of-speech tag and it is associated with a *positive* (+), *negative* (-), or *neutral* (0) a priori semantic orientation. Therefore, we built a lexicon with 21,826 nouns lemmas, 10,500 adjectives lemmas, 1,111 adverbs lemmas, and 8,789 verbs lemmas. The lexicon was built, as described in [12].

2.3 Learning the Compositional Rules

As input, we need a *set of two-word phrases* manually annotated and a *polarity lexicon*. Let *Phrases* = $\{p_1, p_2, \ldots, p_m\}$ represent the set of two-word phrases, where each two-word phrase p_i is a sequence of two words. Each two-word phrase p_i should be manually classified as *positive*, *negative*, or *neutral*, and each word should be tagged with its part-of-speech. Let *Lexicon* = $\{w_1, w_2, \ldots, w_n\}$ represent the polarity lexicon, where each word w_j is associated with its part-of-speech (i.e. *noun*, *adjective*, *adverb*, and *verb*) and SO (i.e. *positive*, *negative*, and *neutral*).

The first step is to assign the SO found in the lexicon to $word_1$ and $word_2$ of each p_i phrase. If a word is not found in the lexicon, we should assign an “Unknown” value to that word (represented in this paper by “?”). This last procedure will allow us to have compositional rules capable of dealing with incomplete information. After the assignment of the SO to the words of p_i phrase, a count is incremented. The count represents the number of times we saw each $POS_1^{SO}\ POS_2^{SO} \rightarrow phrase^{SO}$ combination (or rule), where $phrase^{SO}$ represents the SO assigned by human experts to the phrase p_i and POS_1^{SO}, POS_2^{SO} represents the SO assigned to $word_1$ and $word_2$ based on the lexicon. For example, in our study, after performing this step for all the two-word phrases, we counted the N^0 ADJ^0 → Phrase^0 combination, 267 times. This means that a *neutral* noun followed by a *neutral* adjective (according to the lexicon) was classified as *neutral* by human experts.

The second step is to compute the reliability of each rule. The reliability of each $POS_1^{SO1}\ POS_2^{SO2} \rightarrow phrase^{SO}$ rule is the percentage of phrases that contain POS_1^{SO1} POS_2^{SO2} also contain $phrase^{SO}$. In other words, this percentage represents the probability of being assigned the SO to *phrase* given that it was assigned the SO1 and SO2 to the part-of-speech POS_1 and POS_2, respectively. It can be seen as an estimate of the conditional probability $Prob(\mathrm{Phrase}^{\mathrm{SO}}|\mathrm{POS}_1^{\mathrm{SO1}}\ \mathrm{POS}_2^{\mathrm{SO2}})$. It is computed as follows:

$$reliability = \frac{count(Phrase^{SO} \text{ and } POS_1^{SO1}\ POS_2^{SO2})}{count(POS_1^{SO1}\ POS_2^{SO2})}$$

For example, in our study, the N^0 ADJ^0 → $phrase^0$ rule had a probability of 77% (267/345). This value means that 77% of the phrases composed by a *neutral* noun followed by a *neutral* adjective, where classified as *neutral* by humans experts. Of the remaining 23% phrases composed of a *neutral noun* followed by a *neutral adjective*, 17% of them were classified as *positive* such as “[$parto^0$ $normal^0$]$^{+}$” (*natural birth*), and 5% were classified as *negative* such as “[$programa^0$ $nuclear^0$]$^{-}$“ (*nuclear plan*).

The third step is to select the rules with highest reliability, among the rules with the same left-hand side. For example, from the three rules below, the algorithm automatically chooses the first one. This selection is necessary because these rules are conflicting and incompatible. Two or more rules are incompatible if they have the same left-hand side, but a different right-hand side.

N^0 ADJ^0 → 0 [reliability = 77.39%]
N^0 ADJ^0 → - [reliability = 5.22%]
N^0 ADJ^0 → + [reliability = 17.39%]

The rules learned can be used as a probabilistic model to classify unseen examples.

3 Compositional Rules – Reliability Evaluation

Applying the approach described in the previous section, we got a set of compositional rules which best describe the SO of the phrases annotated by human experts. **Table 3** illustrates some of those rules. The rules are grouped by part-of-speech pattern. Each rule is associated with a reliability value (the conditional probability). For example, for the pattern “ADJ N”, the cell “++” (row: + and column: +) containing the “$^{+}$100” value represents the rule:

ADJ^+ N^+ → + [reliability = 100%]

This rule says that 100% of the phrases with a positive adjective followed by a positive noun were classified as positive by human experts.

In **Table 3**, “n.f.” stands for *not found*, meaning that a certain phrase doesn’t exist in the corpus. For instance, in our corpus there are no occurrences of phrases composed by a *positive* ADJ followed by a *negative* N. **Table 4** shows the average reliability value of the learned rules, for each extraction pattern. Among these patterns the hardest SO to predict is the SO of two-word phrases formed by a VRB, followed by an optional “prp|art|num”, followed by a N, since the conditional probability value is the lowest. The most frequent cases are neutral verbs followed by neutral nouns that together convey a non-neutral SO. E.g. “*[levado para hospital]*$^{-}$” (taken to hospital), “[sai do hospital]+” (leaves the hospital), “*meter água*” (expression that can be used literally to mean “taking on water” or idiomatically to mean “to screw up”).

Table 3. Compositional rules learned for each part-of-speech pattern and their reliability values (in percentage). (?)=word not found in the lexicon. (n.f)=phrase not found in the corpus

ADJ N

		N			
		+	-	0	?
ADJ	+	$^{+}100$	n.f.	$^{+}100$	$^{+}71$
	-	$^{-}100$	$^{-}100$	$^{-}100$	$^{-}67$
	0	$^{+}75$	$^{-}86$	$^{0}87$	$^{0}89$
	?	$^{+}50$	$^{-}75$	$^{0}94$	$^{0}94$

ADV ADJ

		ADJ			
		+	-	0	?
ADV	+	$^{+}100$	n.f.	$^{+}100$	$^{+}71$
	-	n.f.	$^{-}100$	$^{-}100$	n.f.
	0	$^{+}85$	$^{-}89$	$^{0}92$	$^{0}90$
	?	$^{+}67$	$^{-}71$	$^{0}75$	$^{0}89$

N ADJ

		ADJ			
		+	-	0	?
N	+	$^{+}84$	$^{-}80$	$^{0}61$	$^{+}57$
	-	$^{-}57$	$^{-}86$	$^{-}90$	$^{-}89$
	0	$^{0}49$	$^{-}88$	$^{0}77$	$^{0}68$
	?	$^{+}66$	$^{-}85$	$^{0}81$	$^{0}64$

ADV VRB

		VRB			
		+	-	0	?
ADV	+	n.f.	n.f.	$^{+}100$	$^{+}100$
	-	n.f.	n.f.	n.f.	$^{-}100$
	0	$^{0}75$	$^{0}59$	$^{0}91$	$^{0}59$
	?	$^{0}100$	$^{-}75$	$^{0}100$	$^{0}100$

N prp N

		N			
		+	-	0	?
N	+	$^{+}78$	$^{+}62$	$^{+}65$	$^{+}65$
	-	$^{-}85$	$^{-}97$	$^{-}94$	$^{-}95$
	0	$^{+}48$	$^{-}67$	$^{0}72$	$^{0}70$
	?	$^{0}49$	$^{-}67$	$^{0}76$	$^{0}80$

VRB (prp|art|num)? N

		N			
		+	-	0	?
VRB	+	$^{+}77$	$^{+}47$	$^{+}46$	$^{0}51$
	-	$^{-}55$	$^{-}70$	$^{-}65$	$^{-}64$
	0	$^{+}63$	$^{-}77$	$^{0}64$	$^{0}65$
	?	$^{+}68$	$^{-}74$	$^{+}39$	$^{0}42$

Table 4. Average reliability of the learned rules, for each part-of-speech pattern, considering only words present in the lexicon (+/-/0) VS considering all words (+/-/0/?)

Pattern	Average reliability of: +/-/0	Average reliability of: +/-/0/?
ADJ N	93.50%	85.87%
ADV ADJ	95.14%	86.85%
ADJ ADJ	100%	91.43%
N ADJ	74.67%	73.88%
ADV VRB	87.18%	81.25%
N prp N	74.22%	73.13%
VRB (prp|art|num)? N	62.67%	60.44%

4 Conclusions

This paper presents an approach for learning compositional rules from data, which explains the semantic orientation of two-word phrases, considering the a priori semantic orientation of its words and makes three important contributions:

- First, it describes a method to find the set of compositional rules that best describes a set of two-word phrases. Among other things, results have shown that the SO of some part-of-speech patterns are considerably more compositional than others. For instance, the SO of phrases formed by a noun followed by an adjective are much more compositional than phrases formed by a verb followed by a noun.
- Second, it empirically gives evidence that the proposed reliability measure is indeed a good estimator for measure of SO rule quality.
- Finally, it introduces a new corpus consisting of real online data, which will be useful for future studies and research in this area. Namely this corpus can present a first tool for researchers that want to research the semantic dependencies among words and their role for extracting, namely, semantic orientation at sentence level.

Acknowledgments

António Paulo Santos is supported by the FCT grant SFRH/BD/47551/2008 and supported by Department of Informatics of FCT/Universidade de Lisboa.

References

1. Lehrer, A.: Semantic Fields and Lexical Structure. North-holl. Linguist. Ser. 11, 225 (2008).
2. 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. Association for Computational Linguistics, Morristown, NJ, USA (2002).
3. Moilanen, K., Pulman, S.: Sentiment composition. In: Proceedings of the Recent Advances in Natural Language Processing International Conference. pp. 378–382 (2007).
4. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1201–1211 (2012).
5. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). pp. 347–354 (2005).
6. Polanyi, L., Zaenen, A.: Contextual valence shifters. Comput. attitude Affect text Theory Appl. 20, 1–10 (2006).
7. Choi, Y., Cardie, C.: Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP ’08. p. 793. Association for Computational Linguistics, Morristown, NJ, USA (2008).
8. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Comput. Linguist. 37, 267–307 (2011).
9. Yessenalina, A., Cardie, C.: Compositional Matrix-Space Models for Sentiment Analysis. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 172–182 (2011).
10. Klenner, M., Petrakis, S., Fahrni, A.: Robust Compositional Polarity Classification. In: Proceedings of the International Conference RANLP 2009. pp. 180–184 (2009).
11. Shaikh, M.A.M., Prendinger, H., Mitsuru, I.: SenseNet : A Linguistic Tool to Visualize Numerical-Valance Based Sentiment of Textual Data. In: Proc. ICON- 2007 5th Int’l Conf. on Natural Language Processing. pp. 147–152. , Hyderabad, India (2007).
12. Santos, A.P., Ramos, C., Marques, N.C.: Determining the Polarity of Words through a Common Online Dictionary. In: Antunes, L. and Pinto, H.S. (eds.) 15th Portuguese Conference on Artificial intelligence. pp. 649–663. Springer Berlin Heidelberg, Berlin, Heidelberg (2011).