Page Ranking WordNet Synsets: An Application to Opinion Mining

Eleftherios Matios

Department of Computational Linguistics

4. November 2010

Table of Contents

Goal

Rank terms in respect to how strongly they posses the semantic property of "positivity" and "negativity".

Method

Transform WordNet into a graph and apply PageRank to it.

Motivation

"positivity" and "negativity" are to properties that are of central importance in sentiment analysis (the discipline that deals with the analysis of text in regards to opinion-related properties (ORPs)

WordNet

WordNet is a lexical database for the English language. Words are organized in synsets and divided into nouns, verbs, adjectives and adverbs.

Example

S: (n) **tree** (a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms)

Special graph structure:

- Binary relation $s_i \triangleright s_k$ connects nodes
- *"a term belonging to synset s_k occurs in the gloss of synset s_i"*
- Result is a directed graph
- Relations can be obtained from eXtended WordNet

Hypotheses

- the $s_i \triangleright s_k$ relation transmits the semantic properties from s_i to s_k.
- different senses of the same term have different ORPs

Example

S: (n) **good** (benefit) *"for your own good"; "what's the good of worrying?"*

good $\triangleright$ benefit

Table of Contents

Definitions

- $G =< N, L >$ is a directed graph with N being its set of nodes and L its set of directed links
- $|N| \times |N|$ is the *adjacency matrix* of G, such that $W_0[i, j] = 1$ iff there is a link from n_i to n_j and 0 otherwise
- $B(i)$ denotes the *backward neighbours* of n_i as the set $\{n_j | W_0[j, i] = 1\}$
- $F(i)$ denotes the *forward neighbours* of n_i as the set $\{n_j | W_0[i, j] = 1\}$
- W is the *row-normalized adjacency matrix* of G, such that $W[i, j] = \frac{1}{|F(i)|}$ iff $W_0[i, j] = 1$ and $W_0[i, j] = 0$ otherwise

Algorithm (1)

- Input: the row-normalized Matrix W
- Output: a vector $a =< a_i, \ldots, a_{|N|} >$
- two indipendet rankings have to be computed for positivity and negativity

The vector is computed iteratively with the formula:

$$a_i^{(k)} \leftarrow \alpha \sum_{j \in B_{(i)}} \frac{a_j^{(k-1)}}{|F(j)|} + (1-\alpha)e_i$$

- $a_i^{(k)}$ measures the score of n_i for positivity or negativity in the k-th iteration
- e_i is a constant such that $\sum_i e_{i=1}^{|N|} = 1$
- α is a control parameter between 0 and 1

Algorithm (2)

As a vector:

$$a^{(k)} = \alpha a(k-1)W + (1-\alpha)e$$

- the assumption is, that a node n_i has a high score when it has many high-scoring backward neighbours, with only few forward neighbours each
- a node n_j passes its score to its forward neighbours $F(j)$ but this score is equally divided between the members of $F(j)$
- e_i is used to “smooth” those scores as to avoid that scores get trapped in cliques with backward neighbours but no forward neighbours
- the algorithm runs until it reachis a stable state

eXtended WordNet

- based on WordNet 2.0
- nodes that participate in the $s_i \triangleright s_k$ relation are connected
- automatically generated $\Longrightarrow$ noise

Micro-WNOp

- used as a benchmark for PageRank
- 1,105 WordNet synsets, each with a triplet of scores for positivity, negativity and neutrality
- representative of WordNet in regards to part of speech, but not ORPs
- generated by randomly selecting 100 terms of each category (positive, negative, neutral)

Table of Contents

Parameter Tweaking (1)

Experiments are done with different values for e_i

- **e1:** all values are set to $\frac{1}{|N|}$, this is used as the baseline
- **e2:** non-null e_i scores for synsets that contain the adjective good (bad), null scores for everything else
- **e3:** non-null e_i scores for synsets that contain at least one of the following adjectives good, excellent, positive, forunate, correct, superior
- **e4:** utilizes Senti-WordNet, a ressource in which every synsets is assigned a triplet of scores for positivity, negativity and neutrality, e_i values for each synset are propotional to those Senti-WordNet scores
- **e5:** sames as **e4** with a newer version of Senti-WordNet

Parameter Tweaking (2)

The second parameter that allows tweaking is α

- α determines the contribution of $a^{(k)}$ and e
- $\alpha = 0$ makes $a^{(k)}$ coincide with e, thus disregards the contribution of PageRank
- $\alpha = 1$ makes discards e and makes $a^{(k)}$ dependent on the topology of the graph, resulting in an "unbiased" ranking.
- α is optimized by iterating 101 times through the algorithm and incrementing α by 0.1 every time and then picking the best result

Effectiveness measure (1)

For a pair of nodes (n_i, n_j)

n_i can either

- preceede n_j: $(n_i \preceq n_j)$
- succed n_j: $(n_i \succeq n_j)$
- be tied with n_j: $(n_i \approx n_j)$

Effectiveness measure (2)

Rankings are evaluated by computing the *p-normalized Kendall τ distance* between preditcion and Micro-WNOp rankings. Defined as:

$$\tau_p = \frac{n_d + p \cdot n_u}{Z}$$

- n_d: number of discordant pairs (inverted ordering)
- n_u: number of pairs ordered in the gold standard and tied in the
- p: penalization attributed to every pair, set to $p = \frac{1}{2}$, equal to the propability of guessing
  $\Longrightarrow$ no gain from assigning ties randomly
- Z: normalization, equal to number of ordered pairs in gold standard, make the range of $\tau_p = [0, 1]$

Evaluation (1)

		Positivity	Negativity
e	PageRank?	τ_p	τ_p
e1	before	.500	.500
	after	.596 (-0.81%)	.549 (9.83%)
e2	before	.500	.500
	after	.467 (-6.67)	. 502 (0.31%)
e3	before	.500	.500
	after	.471 (-5.79%)	0.45 (-0,92)
e4	before	.349	296
	after	**.349** (-6.75)	**.284** (-4.31%)
e5	before	.400	.407
	after	.380 (-4.88%)	.393 (-3.45%)

Values of τ_p between predicted ranking and gold standard rankings (smaller is better) with different e_i vectors

Evaluation (2)

Sets

Training and Testing was done on Micro-WNOp, which was divided into three parts

- **Common:** 110 synsets, used for aligning evaluation criteria
- **Group1:** 496 synsets, the validation set
- **Group2:** 499 synsets, independently evaluated from Group1, used as a test set

- for positivy, all rankings produced with pagerank are better than the baseline
- for negativity, Senti-WordNet based vectors outperformed everything else

Evaluation (3)

- e4 performs the best
- key to good performance is a combination of positivity flow and internal source of score e_i
- however improvement comes through an already high-quality ressource
- **but** Senti-WordNet was built by a semi-supervised learning Method, that uses the e2 vector as its trainings data
  so it was not necessarily to be expected that e4 would outperform e2

Table of Contents

Conclusion and Outlook

Proof-of-concept

The paper can be seen as a proof-of-concept for the applicability of random-walk algorithms to the determination of semantic properties on a synset

Outlook

this model can be applied to other categorizational task, in which semantic properties of terms have to be compared (i.e. membe rship in a domain)

References I

Andrea Esuli and Fabrizio Sebastiani.
Sentiwordnet: A publicly available lexical resource for opinion mining.
In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 2006.

Andrea Esuli and Fabrizio Sebastiani.
Pageranking wordnet synsets: An application to opinion mining.
In *Proceedings of the ACL*, 2007.