Decision Support Systems 53 (2012) 695–703

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Experiments with a differential semantics annotation for WordNet 3.0

Dan Tufiş *, Dan Ştefănescu

Research Institute for Artificial Intelligence Romanian Academy, Calea "13 Septembrie", no.13, Bucharest 5, 050711, Romania
Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania

ARTICLE INFO

Available online 23 May 2012

Keywords:
Connotative meaning
Differential semantics
Sentiment analysis
WordNet

ABSTRACT

This article reports on the methodology and the development of a complementary information source for the meaning of the synsets of Princeton WordNet 3.0. This encoded information was built following the principles of the Osgoodian differential semantics theory and consists of numerical values which represent the scaling of the connotative meanings along the multiple dimensions defined by pairs of antonyms (factors). Depending on the selected factors, various facets of connotative meanings come under scrutiny and different types of textual subjective analysis may be conducted (opinion mining, sentiment analysis).

1. Introduction

A connotation is a subjective cultural and/or emotional association that some word or phrase carries, in addition to the word or phrase dictionary (explicit or literal) meaning, which is its denotation. Frequently, connotation and subjectivity are considered synonymic (although they are not). Connotation of a word is intrinsically subjective, referring to emotional responses commonly associated with its referent (that to which it refers). For instance, the word *home* means "the place where one lives", but by connotation, also suggests something good i.e. security, family, love and comfort; the word *murder* means "unlawful premeditated killing of a human being by a human being" but, by connotation, it also suggests bad things such as nastiness, mercilessness, disorder… Connotation has more to do with word meanings, while subjectivity is more about phrases/sentences meaning; subjectivity is on an upper layer and builds on the connotations of constituents.

According to "Semantic Differential" theory [13], the connotative meaning of most adjectives can be, both qualitatively and quantitatively, differentiated along a scale, the ends of which are antonymic adjectives. Such a pair of antonymic adjectives is called a *factor*. The intensive experiments Osgood and his colleagues [13] made with their students outlined that most of the variance in the text judgment was explained by only three major factors: the evaluative factor (e.g. good-bad/good-evil), the potency factor (e.g. strong-weak), and the activity factor (e.g. active-passive).

Kamps and Marx [9] implemented a WordNet-based method in the spirit of the theory of semantic differentials and proposed a method to assess the "attitude" of arbitrary texts. In their approach, a text unit is regarded as a bag of words and the overall scoring of the sentence is obtained by combining the scores for the individual words of the text. Depending on the selected factors, various facets of subjective meanings come under scrutiny.

The inspiring work of Kamps and Marx still has several limitations. The majority of researchers working on subjectivity agree that the connotation (prior subjectivity) load of a given word is dependent on the senses of the respective word ([1,4,11,22] and many others); yet, in Kamps and Marx's model (KMM, henceforth), because they work with words and not word-senses, the sense distinctions are lost, making it impossible to assign different scores to different senses of the words in case. Going up from the level of word to the level of sentence, paragraph or entire text, the bag of words approach can easily fail in the presence of valence shifters such as negation [15]. In order to cope with this problem, the text under investigation needs a minimal level of sentence processing, required for the identification of the structures that could get under the scope of a valence shifter [17]. Compare, for instance, the following sentences: "John is clever and one of the most useful employees" versus "John is clever but not among the most useful employees". The use of negation in the second sentence turned the positive judgment in the first sentence into one that could justify John's discharge.

For dealing with irony or sarcasm, processing requirements go beyond sentence level and detecting the discourse structure of the text might be necessary.

On the other hand, although the adjectives make up the obvious class of subjectivity words, the other open class categories have significant potential for expressing subjective meanings.

In our models, unlike KMM, the building block is the word sense, thus allowing us to assign different connotation values to different senses of a word. This was possible by using an additional source of information besides the WordNet [7] itself, namely the SUMO/MILO ontology [12]. Moreover, we considered all word classes contained in WordNet, not only adjectives.

From this point of view, our work, although adopting a different approach, shares objectives with other wordnet-based methods such as SentiWordNet [6,2] and WordNet Affect [22].

* Corresponding author. Tel.: +40 213188103; fax: +40 213188142.
E-mail addresses: tufis@racai.ro (D. Tufiş), danstef@racai.ro (D. Ştefănescu).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.05.026

2. Base definitions

Let us begin with some definitions, slightly modified, from KMM. We will progressively introduce new definitions to serve our extended model.

Definition 1. Two words w_α and w_β are *related* if there exists a sequence of words (w_α w_1 w_2…w_i… w_β) so that each pair of adjacent words in the sequence belong to the same synset. If the length of such a sequence is $n+1$ one says that w_α and w_β are *n-related*.

Two words may not be related at all or may be related by many different sequences, of various lengths. In the latter case, one would be interested in their minimal path-length.

Definition 2. Let MPL(w_i, w_j) be the partial function:

$$MPL\left(w_i, w_j\right) = \begin{cases} n, \text{the smallest } n \text{ when } w_i \text{ and } w_j \text{ are } n\text{—related} \\ \text{undefined, otherwise} \end{cases}$$

Kamps and Marx [9] showed that *MPL* is a distance measure that can be used as a metric for the semantic relatedness of two words. Observing the properties of the *MPL* partial function, one can quantify the relatedness of an arbitrary word w_i to one or the other word of a bipolar pair. To this end, KMM introduced another partial function as in Definition 3.

Definition 3. Let *TRI* (w_i, w_α, w_β), with $w_\alpha \neq w_\beta$ be:

$$TRI\left(w_i, w_\alpha, w_\beta\right) = \begin{cases} \dfrac{MPL(w_i, w_\alpha) - MPL\left(w_i, w_\beta\right)}{MPL\left(w_\alpha, w_\beta\right)}, & \text{if MPLs defined} \\ \text{undefined}, & \text{otherwise} \end{cases}$$

When defined, ***TRI*(w_i, w_α, w_β)** is a real number in the interval [−1, 1]. The words w_α and w_β are the antonymic words of a factor, while w_i is the word of interest for which *TRI* is computed. If one takes the negative values returned by the partial function ***TRI*(w_i, w_α, w_β)** as an indication of w_i being more similar to w_α than to w_β and the positive values as an indication of w_i being more similar to w_β than to w_α, then a zero value could be interpreted as w_i being neutrally related with respect to w_α and w_β. This is different from being unrelated.

Definition 4. If w_α-w_β is a factor used for the computation of relatedness of w_i to w_α and w_β, the proper function $\boldsymbol{TRI^*}_{\mathbf{w\alpha\text{-}w\beta}}\ (\boldsymbol{w_i})$ returns a value outside the interval [−1, 1] when w_i is unrelated to the w_α-w_β factor:

$$TRI^*_{W\alpha-W\beta}(w_i) = \begin{cases} TRI\left(w_i, w_\alpha, w_\beta\right), iff\ TRI\left(w_i, w_\alpha, w_\beta\right) \\ 2, \text{otherwise} \end{cases}$$

Given a factor w_α-w_β, for each word w_i in WordNet that can be reached on a path from α to β, the function $\boldsymbol{TRI^*}_{\mathbf{w\alpha\text{-}w\beta}}\ (\boldsymbol{w_i})$ computes a score number, which is proportional to the distances from w_i to w_α and to w_β. The set of these words defines the coverage of the factor—COV(w_α, w_β).

Our experiments show that the coverage of the vast majority of the factors, corresponding to the same POS category, is the same. From now on, we will use LUC (Literal Unrestricted[1] Coverage) to designate this common coverage. The Table 1 gives coverage figures for each of the POS categories in Princeton WordNet 3.0 (PWN 3.0).

Table 1
LUC statistics according to the POS of the literals in PWN 3.0.

Class	Factors	LUC
Adjectives	199	4402 (20.43%)
Nouns	106	11,964 (10.05%)
Verbs	223	6534 (56.66%)
Adverbs	*199*	*1291 (28.81%)*

The PWN structuring does not allow us to compute TRI* scores for adverbs using this approach, but more than half of the total number of adverbs (63.11%) are derived from adjectives. For those adverbs, we transferred the score values from their correspondent adjectives in the LUC set and we used the adjectival factors.

The results reported for adjectives by Kamps and Marx are consistent with our findings. They found 5,410 adjectives that were in the coverage of the factors they investigated (WordNet 1.7). For PWN 2.0, the total number of covered adjectives is 5,307. The difference in numbers might be explained by the fact that the two compared experiments used different versions of the Princeton WordNet.

3. Introducing word-sense distinctions

KMM defines a factor as a pair of words with antonymic senses (but does not specify which senses). We generalize the notion of a factor to a pair of synsets. In the following, we will use the colon notation to specify the sense number of a literal that licenses the synonymy relation within a synset. Synonymy is a lexical relation that holds not between a pair of words but between specific senses of those words. That is, the notation {literal$_1$:n$_1$ literal$_2$:n_2 … literal$_k$:n_k} will mean that the meaning given by the sense number n$_1$ of the literal$_1$, the meaning given by sense number n$_2$ of the literal$_2$ and so on are all pair-wise synonymous. The term *literal* is used to denote the dictionary entry form of a word (lemma).

Antonymy is also a lexical relation that holds between specific senses of a pair of words. The synonyms of the antonymic senses, taken pairwise, definitely express a semantic opposition. Take for instance the antonymic pair <rise:1 fall:2>. The senses of these two words belong to the synsets {rise:1, lift:4, arise:5, move up:2, go up:1, come up:6, uprise:6} and {descend:1, fall:2, go down:1, come down:1}. The pair <rise:1 fall:2> is explicitly encoded as antonymic. However, there is a conceptual opposition between the synsets to which the two word senses belong, that is between any pair of the Cartesian product: {rise:1, lift:4, arise:5, move up:2, go up:1, come up:6, uprise:6}⊗{descend:1, fall:2, go down:1, come down:1}. This conceptual opposition is even more obvious in this example, as the pairs <go up:1 go down:1> and <come up:1 come down:1> are also explicitly marked as antonymic.

Definition 5. An *S*-factor is a pair of synsets (S_α, S_β) for which there exist w_i^α:$s_i^\alpha \in S_\alpha$and w_j^β:$s_j^\beta \in S_\beta$so that w_i^α:s_i^α and w_j^β:s_j^β are antonyms and *MPL*(w_i^α, w_j^β) is defined. S_α and S_β have opposite meanings, and we consider that *MPL*(S_α, S_β) = *MPL*(w_i^α, w_j^β).

The example we discussed above showed that the semantic opposition of two synsets may be reinforced by multiple antonymic pairs. Because of the way MPL is defined, choosing different antonymic pairs might produce different values for *MPL*(S_α, S_β). That is why, wherever the case, we need to specify the antonymic pair which defines the *S*-factor.

Based on the definition of the coverage of a factor <w_i^α,w_j^β>, one may naturally introduce the notion of coverage of a *S*-factor—<S_α, S_β>: the set of synsets containing the words in COV<w_i^α,w_j^β>. The coverage of an *S*-factor <S_α, S_β> will be onward denoted by SCOV<S_α, S_β>.

Since the word-relatedness and MPL definitions ignore the word senses, it might happen that the meaning of some synsets in the coverage of an *S*-factor has little (if anything) in common with the semantic field defined by the respective *S*-factor. More often than not,

[1] In the following we will gradually introduce several restrictions, thus justifying the acronym used here.