EXPLORING SEMANTIC ORIENTATION OF ADVERBS

S. B. Potemkin (potemkin@philol.msu.ru)

G. E. Kedrova (kedr@philol.msu.ru)

Lomonosov Moscow State University, Moscow, Russia

Sentiment analysis often relies on a semantic orientation lexicon of positive and negative words. Determining the semantic orientation of words is necessary for correct estimation of the content of statements in the media, Internet, in the writings and speech. Qualitative adverbs expressing evaluation, intensity, direction of action are important as the modifiers of the main sentence predicate. In this paper we propose a method for extracting seed set of adverbs from a collection of pairs of antonym. A model based on the representation of a set of synonyms from the Russian lexicons as a graph, and determination the semantic orientation of the adverbs concerning three main dimensions of the semantic differential also demonstrated. The assessment of performance of the method in comparison with the dictionary data shows effectiveness of the method obtained.

Key words: adverb, semantics, semantic orientation, sentiment analysis.

Introduction

Nowadays, the availability of resources for Natural Language Processing (NLP) remains a hot topic, in particular for Russian especially due to the lack of comprehensive semantic resources, despite efforts made to provide a freely-available Russian WordNet [1]. Ability to establish relativity, similarity, or semantic distance between words and concepts is the basis of computational linguistics. This paper deals with measuring of distance within the syntactic category of adverbs. This set of words is crucial for some applications because adverbs modify or clarify the meaning of other words (verbs, nouns, adjectives). The adverbs are of particular interest to determine the semantic orientation of syntagma containing a main word and its modifier (adverb). Measuring the semantic distance or similarity between the English words most often is based on WordNet [2], and almost exclusively on taxonomic relationships established in this database. So such approach is applicable only to the syntactic categories of nouns and verbs.

The aim of this paper is to extract a list of semantically oriented adverbs and develop the measure of proximity based on dictionaries of synonyms. The article is structured as follows. In Section 1 the problem of extracting the seed set of semantically oriented adverbs from the lexicon of Russian antonyms is discussed. In Section 2 we describe the previously proposed measures of semantic distance between words, as well as an elementary way to map synonyms onto a graph. In Section 3 the

basic characteristics of the subjective understanding of the meaning and the measures based on the distance in a graph of synonyms are discussed. Finally, Section 4 presents some results and conclusions. Additionally, we explore the use of visualization techniques to gain insight into the results obtained.

1. Extracting the seed set of adverbs

A number of approaches have been proposed for creating semantic orientation lexicons in English, most of them are computationally expensive and rely on significant manual annotation and large corpora. Particularly, the General Inquirer [3] created in the beginning of the last century is used as the gold standard for assessment the quality of new-generated lexicons. For Russian language there is no open-source and reliable lexicon with positively and negatively marked entries. We propose some approaches to generate a broad coverage semantic orientation lexicon for Russian adverbs which includes both individual words and multi-word adverbial expressions using only dictionaries of antonyms and synonyms, requiring a small amount of manual pruning and database processing.

First of all we have analyzed a list of antonyms collected from published dictionaries of antonyms [4, 5]. This list contains 7,300+ antonymous pairs (adjectives, nouns, verbs, adverbs and prepositions as well). The semantically oriented words were manually extracted from this list and arranged in 2 separate lists — positive (1,859) and negative (2,229) words. This seed lexicon could be compared with the GI lexicon which contains orientation labels for only about 3,600 entries.

Next step was to extend our seed lexicon to obtain a broad coverage of different texts under consideration concerning sentiment analysis. Automatic approaches to create (English) semantic orientation lexicon and, more generally, approaches for word-level sentiment annotation can be grouped into two kinds: (1) those that rely on manually created lexical resources—most of which use WordNet; and (2) those that rely on text corpora [6]. As a lexical source we use a structured list of Russian synonyms collected from a number of published and Internet-available dictionaries such as [7] and others (11 sources). List of synonyms contains ~600,000 word-pairs including ~10,000 pairs of adverbs. All synonyms $\{s(w_i)\}$ of each seed word w_i receives the same semantic orientation as w_i. The number N of occurrences of a synonym $s(w_i)$ in the extended set contributed by different seed-words w_i, $(i=1\ldots N)$ indicates the confidence of semantic orientation. After manual pruning we have got a list of positively marked (5990, including 731 adverbs) and negatively marked (6853, including 592 adverbs) words. Since the most part of Russian adverbs could be derived as the short form singular neutral or short form plural adjective (3135) the list of semantically orientated adverbs could be expanded.

2. Measures of distance

A number of distance or similarity measures exist for English based (completely or partially) on WordNet. In particular, such measure is defined as the number of edges

of the path through the taxonomic relations (IS-A, Part-of, or WordNet's hyponymy relation). In [8] the concept of bond length was extended for all relations in WordNet by their clustering in the horizontal (synonyms) or vertical (hyponymy) direction and assigning a penalty for changing the direction of the path motion. Overview of five measures and evaluation of their effectiveness using the associations between the words is given in [9]. Exclusive usage of hyponymy delimits the measure of distance or similarity only to the syntactic categories of nouns and verbs, as hyponymy relations in WordNet are established only for these grammatical categories. Therefore, such measures could not be applied to adjectives and adverbs.

The semantic distance between the words could be determined in the similar way as the definition adopted in graph theory [10]. The simplest approach is just to gather all the words from the Dictionary of synonyms and to link each member of a synonymous group with its dominant word as indicated in the Dictionary. Let *G(W,S)* be the undirected graph, with *W* the set of nodes being all the words from the Dictionary with associated part-of-speech, *S* — the set of edges connecting each member of synonymous group with its dominant word. Every group of synonymous words could be connected to each other and form a clique in *G* graph. A path *P* is the sequence of nodes connected by edges of *G* and geodesic is the shortest path between two nodes. Geodesic distance, $D(w_i,w_j)$ between two words w_i and w_j is the length (number of edges) of the shortest path between w_i and w_j. If there is no path between w_i and w_j, the distance between them is infinity. The minimal path-length defines a *metric* on the set of synonyms. All axioms of the metric space are fulfilled in this case. Usually synonymous groups comprises the words of the same grammatical category and entire graph *G* is decomposed into disjoint sub-graphs or networks for nouns, verbs, adjectives and adverbs. (Fig. 1). In each network exists a maximal connected component that contains 70–90% of all nodes of the graph constructed from the Dictionary of synonyms. Maximum component in the class of Russian adverbs contains about 8500 words. The words in this connected component could be analyzed using the metric defined by the length of geodesics.

3. Semantic orientation of adverbs

Classical work on the measurement of emotional or affective values in texts is the theory of semantic differential by Charles Osgood. Word meaning in cognitive psychology, is "a strictly psychological one: those cognitive states of human language users which are necessary antecedent conditions for selective encoding of lexical signs and necessary subsequent conditions in selective decoding of signs in messages." [11]. Semantic differential method was applied mainly to the adjectives measured in such dimensions as *active/passive*, *good/bad*, *positive/negative*, *beautiful/ugly*, etc. Each pair of bipolar adjectives is a factor or an axis in the method of semantic differential. Application of factor analysis to extensive empirical material gave an unexpected result: most of the variance in judgment could be explained by only three major factors including the *evaluative* factor (e. g., *positive/negative*); the *potency* factor (e. g., *strong/weak*); and the *activity* factor (e. g., *active/passive*). Among these three factors, the evaluative factor has the strongest relative weight for determining the semantic orientation.

Turning to the selected Russian adverbs, we note that the vast majority of adverbs is matched with the words of other parts of speech primarily with the adjectives (*cheerful — cheerfully // бодрый — бодро*, *brutal — brutally // жестокий — жестоко)*, so that the semantic differential can be naturally extended to motivated adverbs, which bear semantic meaning and, accordingly, deliver the information on their semantic orientation. All three pairs of bipolar adverbs *negatively/positively* (*плохо/хорошо*); *weakly/strongly (слабо/сильно)*, *passively/actively (пассивно/активно)* are contained in the maximal component of the sub-graph of synonymous adverbs G_{adv}. One can assume that the distance to *positively (хорошо)* is a measure of positive assessment of an adverb. However, it is easy to show that this measure is in fact rather controversial.

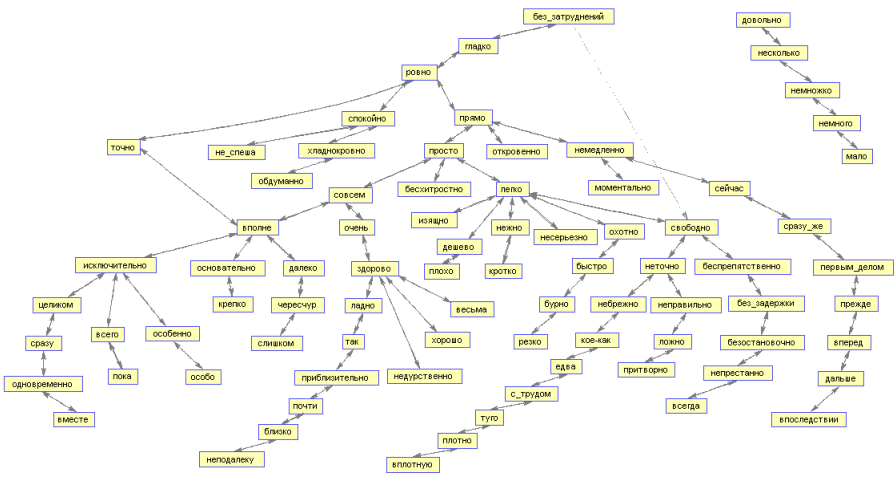

Fig. 1. A fragment of the maximum connected component subgraph of adverbs, $G_{adv.}$ One cannot select a consistent spanning tree

A striking example of this is that the words *positively (хорошо)* and *negatively (плохо)* are closely related through the path of synonyms. There is a sequence of only 5 words in English (*negatively, hardly, tightly, thoroughly, comprehensively, soundly, positively*), and 6 words in Russian (*плохо, дешево, легко, просто, совсем, очень, здорово, хорошо*) — see Fig. 1 — connecting opposites, each pair of words in this sequence is certainly synonymous (at least in one of their meanings). Thus, we find that *d(positively, negatively)* = 6; *d(хорошо, плохо)* = 7. Despite the fact that the adverb *positively (хорошо)* and *negatively (плохо)* have opposite meanings, they are closely related by synonymy path. Of course, this is not due to any error in the Dictionary of synonyms. Partial explanation lies in the wide use of two Russian adverbs *хорошо* (625 ipm), *плохо* (187 ipm) [12]. The other source of uncertainty is the fact that a spanning tree probably could not be chosen perfectly, e. g. the path from *неподалеку (not far from)* to *вплотную (close to)* is 12 arcs length while their meanings are very similar. In addition, we observe 'shift of meaning' while travelling by the path of polysemic

synonyms, i. e. the left arc of path A — B1/B2 — C connects A with B1 meaning while the right arc connects A with the other one, B2. Here we assume that B1/B2 do have some sema in common (if these are not the pure homonyms which could be filtered out automatically). Nevertheless due to the fact that both words *хорошо*, *плохо* are members of the maximum connected component of G_{adv} sub-graph, we can consider not only the shortest distance from any adverb to "*positively*", but the shortest distance to its antonym, "*negatively*". This idea is concretized [13] in the definition of EVA function, which allows to measure the relative distance from the word of two opposites, "*positively*" and "*negatively*":

$$EVA\ (w) = (d\ (w, neg) - d\ (w, pos)) / d\ (neg, pos).$$

Under the assumption that there is no word "worse than *negatively*" or "better than *positively*" the values of EVA lie in the interval [-1,1], for example, the word "honestly" is evaluated by function EVA (*honestly*) gives a value of 1 as follows EVA (*honestly*) = (d (*honestly*, *neg*) — d (*honestly, pos*)) / d (*pos, neg*) = (8-2) / 6 = 1,. The measures for other Osgood's dimensions is defined similarly. For the potency factor the function: POT (w) = (d (w, *weakly*)-d (w, *strongly*)) / d (*strongly, weakly*) is defined; for the activity factor the function: ACT (w) = (d (w, *passively*)-d (w, *actively*)) / d (*actively, passively*) is defined. This fact allows to define measures for any two words belonging to the maximal connected component of the adverbs subgraph.

An assumption on the boundary position of words *negatively/positively* is not entirely justified. Intuitively, *perfectly (превосходно)* is better than positively, *disgustingly (отвратительно)* is worse than negatively. Bearing this in mind and using the geometry of a triangle with vertices {w, *pos, neg*}, we redefine the function of EVA, namely:

$$EVA_1(w) = (d(w, neg) - d(w, pos))*(d(w, neg)+d(w, pos))/d^2(neg, pos).$$

The values of EVA_1 sometimes are beyond the interval [−1,1]. Similarly, we can redesign POT(w) and ACT(w).

For English adjectives (and motivated adverbs) there exists the source for assessing the measure constructed above in comparison with the independently obtained answers to the "General Inquirer" [11], which contains a set of words to assess three Osgood's factors. Word lists were obtained from the Stanford political dictionary, where each of the 3,000 most frequent common words were assessed by three or more experts concerning each Osgood's factor. Thus 765 positive and 873 negative words for the assessment factor were obtained, 1,474 strong and 647 weak word for the potency factor and 1,568 active and 732 passive words for activity factor. Comparison of results obtained with the General Inquirer gave the values of 70–80 % of matches, depending on what words were considered as neutral in terms of EVA function.

In the absence of available data for content analysis we used the Russian dictionaries of antonyms as an independent source. Antonymous pair is a pair of words (or rather, the specific meanings of words), one opposed to the other on semantic grounds, such as *hot — cold fast — slow, present — absent*. We suggest that adverbs belonging to pair of antonyms lie on the "opposite sides" of the entire set of adverbs.

Methods of multidimensional scaling deliver a mapping of multidimensional space with the defined distance between individual points $d(w_i, w_j)$ onto a space of smaller dimension, namely the plane (Fig. 2). Figure 2a, b shows that the pairs of antonyms lie near the diameters of the set of adverbs. For a more profound study of the structure of the space of adverbs we have constructed chains of synonyms connecting antonyms pairs within the sub-graph G_{adv}.

Chain in Fig. 2a is a consistent result, i. e. the chain of synonyms passes on the periphery of the set of adverbs and the distances between the synonyms do not exceed the distance between the antonyms. Unfortunately, the situation is not always as favorable. In Fig. 2b pair of antonyms is close to the diameter, but the chain of synonyms is not at the periphery of the set, but lays in the central part of the set, alternates its direction, and the distances between synonyms is often greater than the distance between antonyms. Probably it is necessary to determine more accurate distance between the words and to choose correctly the axes of the adverb space using the principal components method. These new axes should not coincide Osgood's dimensions.

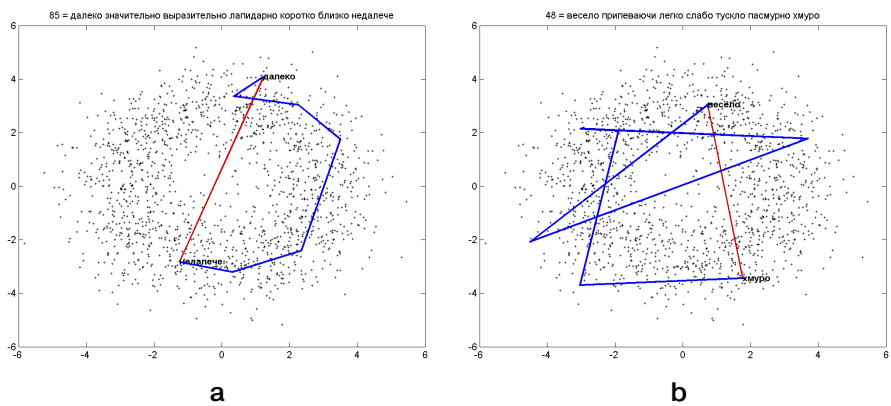

Fig. 2. Two chains of synonyms, joining antonymous pairs of adverbs. a) Left — a consistent path; b) Right — an inconsistent result

4. Discussion and conclusions

In this paper we define a measure of the distance between adverbs using synonyms graph. It seems obvious that the choice of similarity measure, or distance largely depends on the type of the problem. The choice of distance measure on the grounds of synonyms is connected with the goal of determining the semantic orientation of adverbs. In contrast to Osgood's semantic differential associated with the reaction of people on the stimulus — words presented, or the possible emotional impact of words, this model is based solely on the lexical material and is intended to represent relatively objective meanings which are fixed in Dictionaries.

Some inadequate results (as in Fig. 2b) probably arise from the inadequate dimension (3 axes) of Osgood's space.

Further studies will determine the semantic orientation of sentences or the whole text on the basis of the orientation of its constituent words. Our method allows to evaluate other classes of words such as nouns, adjectives and verbs, but this extension will require a significant increase of calculations and special methods for processing large data sets, since an algorithm for computing shortest paths requires $O(n^3)$ operations, where n is the number of words in graph G(W,S).

References

1. *Aleksandrova Z. E.* 2005. Russian Synonyms Dictionary [Slovar' Sinonimov Russkogo Iazyka].
2. *Azarova I. V., Mitrofanova O. A., Sinopal'nikova A. A.* 2003. Computational Thesaurus of Russian Language of the kind of WordNet [Komp'iuternyi Tezaurus Russkogo Iazyka Tipa WordNet]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2003" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2003") : 43–50.
3. *Budanitsky A., Hirst G.* 2001. Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures. Workshop on WordNet and Other Lexical Resources. Second meeting of the NAACL.
4. *Hirst G., St-Onge D.* 1998. Lexical Chains Representations of Context for the Detection and Correction of Malapropisms". WordNet. An Electronic Lexical Database.
5. *Kamps J., Marx M., Robert J., Mokken M.* 2004. Using WordNet to Measure Semantic Orientations of Adjectives. Proceedings of the 4th International Conference on Language Resources and Evaluation, IV : 1115–1118.
6. *Mohammad S.* et.al. 2009. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009).
7. *Osgood C. E., Succi G. J., Tannenbaum P. H.* 1957. The Measurement of Meaning.
8. *Potemkin S. B.* 2008. Semantic Distance in the Linguistic Database WorldNet [Semanticheskoe Rasstoianie v Lingvisticheskoi Baze Dannykh WorldNet]. Materialy 10 Mezhdunarodnoi Konferentsii "Kognitivnoe Modelitovanie v Lingvistike" (Proc. of the X International Conference "Cognitive Modelling in Linguistics").
9. *Sharov S. 2003.* Frequency Dictionary [Chastotnyi Slovar'], available at: http://www.artint.ru/projects/frqlist.asp
10. *Stone P. J.* 1997. Thematic Text Analysis: New Agendas for Analyzing Text Content. Text Analysis for the Social Sciences.
11. *Vvedenskaia L. A.* 2004. Russian Antonyms Dictionary [Slovar' Antonimov Russkogo Iazyka].
12. *L'vov M. R.* 2006. Russian Antonyms Dictionary [Slovar' Antonimov Russkogo Iazyka].
13. *WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series.1998.*