Procedia Computer Science
Volume 88, 2016, Pages 538–543
7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

Weak Semantic Map of the Russian Language: Preliminary Results

Alexei V. Samsonovich
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Moscow, Russian Federation
Alexei.Samsonovich@gmail.com

Abstract
Weak semantic maps are characterized by definite semantics associated with their dimensions, while the notion of a dissimilarity metric may not be applicable to them. Among the variety of approaches to their construction, weak semantic maps can be derived from semantic relations, such as synonymy and antonymy, using the technique developed by Samsonovich and Ascoli in previous works. This paper presents a weak semantic map constructed using this technique, based on the Microsoft Word Russian Thesaurus. General characteristics of the map are described. The new map is compared to the maps reported earlier, that were constructed using the Microsoft Word English, French, German and Spanish thesauri. Differences and similarities are noted that may be due to differences between languages or between particular corpora. Potential applications of Russian weak semantic maps are discussed.

Keywords: Semantic space; cognitive map; affective space; appraisal; sentiment analysis; semantic dimensions.

1 Introduction

A *semantic map*, also known as *semantic cognitive map* or *semantic space*, is a system for representation of knowledge, which is based on a metric space that captures certain semantic aspects of representations (e.g., the meaning of words). Semantic maps can be divided into *weak* and *strong* maps (Samsonovich, Goldin & Ascoli, 2010). Coordinates in a weak semantic map represent definite semantic features, or *semantic dimensions*, while the distance between two symbolic representations on a weak semantic map tells little about their semantic proximity. In contrast, strong semantic maps implement *dissimilarity metrics*: they separate meanings by a geometrical distance proportional to the dissimilarity. At the same time, dimensions of a strong semantic map are usually hard to interpret.

Examples of weak semantic maps include databases constructed using human ranking and psychometric data, such as the well-known affective space models (e.g., Osgood et al., 1957, 1975; Russell, 1980; Rubin, 1980; Plutchik, 1982; Bradley & Lang, 1999; Lövheim, 2012). Another kind of weak semantic maps, that are considered here, are constructed from relational data available in various

 Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.08.001

semantic networks: e.g., WordNet (Fellbaum, 1998, 2005) or the Microsoft Office Thesaurus. Here we focus on the statistical-mechanic approach to their construction developed by Samsonovich and Ascoli (2007, 2008, 2010, 2012, 2013, 2014). Alternative approaches to weak semantic map construction from linguistic corpora include Laplacian embedding (Belkin & Niyogi, 2003; Kunegis et al., 2010) and eigen-spaces, such as Sentic Net (Cambria & Hussain, 2012). This work describes a new weak semantic map of the Russian language constructed from a dictionary of synonyms and antonyms.

2 Materials and Methods

The Microsoft Office 2013 Thesaurus was used as the source corpus in this work. The method of dictionary extraction was described in previous works (Samsonovich & Ascoli, 2007). The initial number of dictionary entries was 401837. After truncating the initial semantic net to its connected core, the number of dictionary entries was reduced to 42636. The rule of truncation was that each word in the core must have at least two connections, of which at least one must be an antonym connection. All synonym and antonym connections were symmetrized. The number of symmetric connections in the core was 124087, of which 35752 were antonym connections and 88335 were synonym connections. Finally, connection weights were normalized by the synonym-antonym ratio.

The energy function optimization was performed in Matlab R2015b on iMac, using a first-order optimization algorithm available in the built-in function *fminunc* of the Matlab optimization package.

2.1 Method of Semantic Map Construction

A weak semantic map is constructed from a set of dictionary entries (words and phrases) associated with each other by synonym-antonym relations. Positioning of each dictionary entry on the weak semantic map is defined by the global minimum of the energy function (details can be found in Samsonovich & Ascoli, 2010):

$$H(x) = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}\left(x_i, x_j\right) + \frac{1}{4}\sum_{i=1}^{N}|x_i|^2, \qquad x_i = (x_i^1 \dots x_i^D).$$

Here the energy function H includes the dot product of the vectors x of the coordinates of dictionary entries on the map. W is the synonym-antonym connection weights, with positive entries for synonyms and negative for antonyms. N is the number of words, and D is the map dimension. The global minimum of H was determined using the Matlab optimization toolbox.

3 Results

The ends of sorted along the first two principal components (PC1 and PC2) lists of 42636 Russian words, together with their Google translations, are given in Table 1. Semantics of the dimensions PC1 and PC2 based on these data can be approximately identified as "*beginning, increase*" for PC1 and "*approval, joy, safety*" for PC2. Only two words appear to be inconsistent with the rest: "cooling" in PC1+ and "close" in PC2+. Overall, there is no perfect match with the concepts of valence, arousal, or dominance, usually associated with main dimensions of weak semantic maps (e.g., Osgood et al., 1957, 1975). Nevertheless, strong correlations with maps in other languages are seen in Table 2.

The maximum-spread projection of the constructed semantic map is shown in Figure 1 A. Each blue dot represents a dictionary entry (a word or phrase). For comparison, the previously constructed weak semantic map of English (15783 words) is also shown in its maximum-spread projection (Figure 1 B). The mapping of 30+30 words with extreme values of the second coordinate on the Russian map (Table 1, columns PC2+ and PC2-) to the English map was done using Google Translate (however, four inaccurate translations were manually corrected).