Information Fusion 14 (2013) 281–287

Contents lists available at SciVerse ScienceDirect

Information Fusion

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

Fusing distributional and experiential information for measuring semantic relatedness

Yair Neuman *, Dan Assaf, Yohai Cohen

Department of Education, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

ARTICLE INFO

Article history:
Received 6 October 2011
Received in revised form 6 December 2011
Accepted 2 February 2012
Available online 23 February 2012

Keywords:
Semantic representation
Semantic relatedness
Cognition
Family resemblance
Semiotics
Interdisciplinary research

ABSTRACT

Models of semantic relatedness have usually focused on language-based distributional information without taking into account "experiential data" concerning the embodied sensorial source of the represented concepts. In this paper, we present an integrative cognitive model of semantic relatedness. The model – semantic family resemblance – uses a variation of the co-product as a mathematical structure that guides the fusion of distributional and experiential information. Our algorithm provides superior results in a set expansion task and a significant correlation with two benchmarks of human rated word-pair similarity datasets.

1. Introduction

Semantic similarity or in its extensive sense "semantic relatedness" involves the degree to which two words are close in their meaning [1–4]. In fact, this notion can be traced back to antiquity and to Aristotle who identified four strategies through which associations are formed in our mind [5]: similarity (e.g., an orange and a lemon), difference (e.g., high versus low), contiguity in time (e.g., sun rise and a rooster's crow), and contiguity in space (e.g., a cup and a spoon). These perceptually grounded associations may provide the basic strata for semantic relatedness as words correspond with concepts that have an embodied base [6–8]. While the sensorimotor aspect of associations has been recognized from antiquity, the availability of huge language corpora has naturally led to the reliance on "language-based distributional data" for building semantic representations and for measuring semantic similarity and relatedness. This approach may computationally solve the problem of measuring semantic relatedness through brute force and huge knowledge sources such as Wikipedia, WordNet, or the New York Times archive [2,9,10]. However, it cannot replace a cognitive economic model of semantic relatedness.

As argued by Perlovsky [11, p. 2099]: "Current engineering approaches attempt to develop computer capabilities for language and cognition separately ... Nature does it differently." In a series of papers [12–15] Perlovsky argues that "Abstract thoughts cannot emerge without language [13, p. 71] and that the logic of this interdependence involves the knowledge instinct (KI) to fit top-down to bottom-up signals. Moreover, he argues that conceptual structures evolve from "vague-to-crisp", and that language that is significantly crisp [13, p. 75] may mediate the emergence of conceptual structures. For instance, while the concept of "dog" is perceptually vague, the use of the sign "dog" to signify the objects of the dogs' family may group them together, despite huge differences between dog instances such as Chihuahua and San Bernard. While presenting a mathematical model of this process, Perlovsky does not directly address the issue of semantic relatedness that may be an interesting test case for his mathematical modeling of language and thought.

In psychology, Andrews et al. [16] criticize the orthogonality of the experiential and distributional dimensions in models of semantic representation and call for an integrated model. Moreover, they provide empirical support for the benefits of an integrative model. Nevertheless, their reliance on hand-crafted semantic feature norms limits the representation of the experiential dimension in two important senses. Theoretically, this model assumes people hold in their mind huge feature sets for each word. It is doubtful whether this assumption is the most economical one. Practically, the norms used by the researchers were hand crafted and cannot be easily and trivially extended to new words and contexts.

There is, however, another source of difficulty in integrating the two dimensions for studying semantic relatedness. Human language as a complex evolutionary system, e.g. [17,18], involves

* Corresponding author.
E-mail address: yneuman@bgu.ac.il (Y. Neuman).

1566-2535/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2012.02.001

the abstraction of signs from their original embodied context through a complex network of connotations. For instance, while the adjective "sweet" is sensorially embedded in our taste, chains of connotations [5] lead it far beyond its source to abstract senses such as "Sweet Dreams" [8]. In other words, the meaning of a word is an emerging phenomenon that cannot be *simply* grounded in experiential data. This difficulty is further elaborated through Wittgenstein's idea of family resemblance.

1.1. Family resemblance

One of the main problems in understanding concept formation is that different instances of a given concept do not share a fixed set of features that may be analytically used for defining it. It implies that feature norms are limited in modeling the resemblance of concepts and the words we use to represent them. For instance, looking for a set of features that uniquely defines game may end in bitter disappointment. After all, what is the set of defining features that uniquely characterizes playing with a teddy bear and playing Chess? This lack of essential features for defining a concept is usually referred to in the context of family resemblance. Wittgenstein coined the term "family resemblance" [19] as a part of his attack on essentialism.

The idea behind "family resemblance" is that instances of a given concept are "united not by a single common defining feature, but by a complex network of overlapping and crisscrossing similarities" [20, p. 121]. The same argument holds for words that represent these concepts. The semantic relatedness of words cannot be simply reduced to feature norms because family resemblance, whether of words or concepts, is an emerging property of the semantic network. This emerging meaning cannot be captured at the micro or the macro level of the network but at the mesoscopic level of the network [21,22]. A model that accounts for family resemblance of words should be located at this mesoscopic level and show how the semantic relatedness is an "emerging" phenomenon.

1.2. Lexical priming and abstraction

In this paper, we address the challenge of integrating the experiential and the distributional dimensions through two cognitive processes: abstraction and lexical priming. Lexical priming [23] involves the idea that every word is mentally primed to occur with particular other words, semantic sets, and pragmatic functions. In this context, semantic associations are formed when a word (or a word sequence) is associated in the mind of a language-user with a semantic set or class, some members of which are also collocates for that user [23, p. 24]. In other words, the distributional dimension of semantic representation may be traced back to the psycho-linguistic process of lexical priming and the way it guides our processing of linguistic data. A measure of lexical priming may be used for constructing the distributional dimension.

The second dimension can be studied through abstraction. Language involves words that refer to concrete entities whose basic meaning is clearly grounded in our sensorimotor experience. On the other hand, the language denotes objects whose meaning cannot be traced to sensorimotor experience. These words such as "God" or "Justice" are clearly more abstract than "doughnut" or "apple". The meaning of words may be associated with their level of abstractness. For instance, by using the Corpus of Contemporary American English (COCA) [24], we found that while the word "love" primes abstract nouns such as "affair" and "affection", it also primes more concrete nouns such as "sweets" and "songs". Based on their level of abstractness, we may better understand the meaning of nouns primed by "love" and group them into sets and pragmatic functions. For instance, from the perspective of pragmatic function, songs and sweets are gifts given by the lover to his or her loved one. In this sense, and combined with lexical priming, the abstractness level of a word may be used as a heuristic cue for building the experiential dimension of a semantic representation. Given the importance of heuristics in human cognition [25], it is worth examining the use of abstractness level as a heuristic for building the experiential dimension of a semantic representation. Nevertheless, it is clear that none of the dimensions is sufficient when used isolated from its complement. Therefore, we present an abstract structure that fuses these sources of information. First, however, we present the way in which we measured abstractness and lexical priming.

2. Measuring abstractness and concreteness

This section describes the way through which we measured the abstractness level of words as a step toward the identification of the experiential dimension and the measurement of semantic relatedness. Concrete words refer to things, events, and properties that we can perceive directly with our senses, such as banana, tree, and sweet. Abstract words refer to ideas and concepts that are distant from immediate perception, such as God, justice, and science. In this section, we describe an algorithm that can automatically calculate a numerical rating of the degree of abstractness of a word on a scale from 0 (highly concrete) to 1 (highly abstract). The algorithm has been developed by Peter Turney and cited in [8,26].

The algorithm is a variation of Turney and Littman's [27] algorithm that rates words according to their semantic orientation and calculates the abstractness of a given word by comparing it to 20 abstract words and 20 concrete words that are used as paradigms of abstractness and concreteness. The abstractness of a given word is the sum of its similarity with 20 abstract paradigm words minus the sum of its similarity with 20 concrete paradigm words. The similarity of words was measured through a Vector Space Model (VSM) of semantics [28].

The MRC Psycholinguistic Database Machine Usable Dictionary [29], which contains 4295 words rated with degrees of abstractness, was used to guide the search for paradigm words.[1] Turney used half of these words to train a supervised learning algorithm and the other half to validate it. On the testing set, the algorithm attains a correlation of 0.81 with the dictionary ratings. This indicates that the algorithm agrees well with human judgments of the degrees of abstractness of words.

For building a list of words rated according to their level of abstraction, Turney used a corpus of 5×10^{10} words (280 gigabytes of plain text) gathered from university websites by a webcrawler[2] and then indexed it with the Wumpus search engine [30].[3] The list was selected from the terms (words and phrases) in the WordNet lexicon.[4] By querying Wumpus, he obtained the frequency of each WordNet term in the corpus and selected all terms with a frequency of 100 or more. This resulted in a set of 114,501 terms. Next he used Wumpus to search for up to 10,000 phrases per term, where a phrase consists of the given term plus four words to the left of the term and four words to the right of the term. These phrases were used to build a word–context frequency matrix **F** with 114,501 rows and 139,246 columns. A row vector in **F** corresponds to a term in WordNet and the columns in **F** correspond to contexts (the words to the left and right of a given term in a given phrase) in which the term appeared.

The columns in **F** are unigrams (single words) in WordNet with a frequency of 100 or more in the corpus. A given unigram is represented by two columns, one marked *left* and one marked *right*.

[1] The dictionary is available at http://www.ota.oucs.ox.ac.uk/headers/1054.xml.
[2] The corpus was collected by Charles Clarke at the University of Waterloo.
[3] Wumpus is available at http://www.wumpus-search.org/.
[4] WordNet is available at http://www.wordnet.princeton.edu/.