Title: Semantic folding

URL: https://en.wikipedia.org/wiki/Semantic_folding

PageID: 50222574

Categories: Category:Computational linguistics, Category:Machine learning, Category:Natural language processing, Category:Semantics

-----

Semantic folding theory describes a procedure for encoding the semantics of natural language text in a semantically grounded binary representation . This approach provides a framework for modelling how language data is processed by the neocortex . [1]

Theory

Semantic folding theory draws inspiration from Douglas R. Hofstadter 's Analogy as the Core of Cognition which suggests that the brain makes sense of the world by identifying and applying analogies . [2] The theory hypothesises that semantic data must therefore be introduced to the neocortex in such a form as to allow the application of a similarity measure and offers, as a solution, the sparse binary vector employing a two-dimensional topographic semantic space as a distributional reference frame. The theory builds on the computational theory of the human cortex known as hierarchical temporal memory (HTM), and positions itself as a complementary theory for the representation of language semantics.

A particular strength claimed by this approach is that the resulting binary representation enables complex semantic operations to be performed simply and efficiently at the most basic computational level.

Two-dimensional semantic space

Analogous to the structure of the neocortex, Semantic Folding theory posits the implementation of a semantic space as a two-dimensional grid. This grid is populated by context-vectors [note 1] in such a way as to place similar context-vectors closer to each other, for instance, by using competitive learning principles. This vector space model is presented in the theory as an equivalence to the well known word space model [3] described in the information retrieval literature.

Given a semantic space (implemented as described above) a word-vector [note 2] can be obtained for any given word Y by employing the following algorithm :

The result of this process will be a word-vector containing all the contexts in which the word Y appears and will therefore be representative of the semantics of that word in the semantic space. It can be seen that the resulting word-vector is also in a sparse distributed representation (SDR) format [Schütze, 1993] & [Sahlgreen, 2006]. [3] [4] Some properties of word-SDRs that are of particular interest with respect to computational semantics are: [5]

high noise resistance: As a result of similar contexts being placed closer together in the underlying map, word-SDRs are highly tolerant of false or shifted "bits".

boolean logic: It is possible to manipulate word-SDRs in a meaningful way using boolean (OR, AND, exclusive-OR) and/or arithmetical (SUBtract) functions .

sub-sampling: Word-SDRs can be sub-sampled to a high degree without any appreciable loss of semantic information.

topological two-dimensional representation: The SDR representation maintains the topological distribution of the underlying map therefore words with similar meanings will have similar word-vectors. This suggests that a variety of measures can be applied to the calculation of semantic similarity , from a simple overlap of vector elements, to a range of distance measures such as: Euclidean distance , Hamming distance , Jaccard distance , cosine similarity , Levenshtein distance , Sørensen-Dice index , etc.

Semantic spaces

Semantic spaces [note 3] [6] in the natural language domain aim to create representations of natural language that are capable of capturing meaning. The original motivation for semantic spaces stems from two core challenges of natural language: Vocabulary mismatch (the fact that the same meaning can be expressed in many ways) and ambiguity of natural language (the fact that the same term can have several meanings).

The application of semantic spaces in natural language processing (NLP) aims at overcoming limitations of rule-based or model-based approaches operating on the keyword level. The main drawback with these approaches is their brittleness, and the large manual effort required to create either rule-based NLP systems or training corpora for model learning. [7] [8] Rule-based and machine learning -based models are fixed on the keyword level and break down if the vocabulary differs from that defined in the rules or from the training material used for the statistical models.

Research in semantic spaces dates back more than 20 years. In 1996, two papers were published that raised a lot of attention around the general idea of creating semantic spaces: latent semantic analysis [9] from Microsoft and Hyperspace Analogue to Language [10] from the University of California . However, their adoption was limited by the large computational effort required to construct and use those semantic spaces. A breakthrough with regard to the accuracy of modelling associative relations between words (e.g. "spider-web", "lighter-cigarette", as opposed to synonymous relations such as "whale-dolphin", "astronaut-driver") was achieved by explicit semantic analysis (ESA) [11] in 2007. ESA was a novel (non-machine learning) based approach that represented words in the form of vectors with 100,000 dimensions (where each dimension represents an Article in Wikipedia). However practical applications of the approach are limited due to the large number of required dimensions in the vectors.

More recently, advances in neural networking techniques in combination with other new approaches (tensors) led to a host of new recent developments: Word2vec [12] from Google and GloVe [13] from Stanford University .

Semantic folding represents a novel, biologically inspired approach to semantic spaces where each word is represented as a sparse binary vector with 16,000 dimensions (a semantic fingerprint) in a 2D semantic map (the semantic universe). Sparse binary representation are advantageous in terms of computational efficiency, and allow for the storage of very large numbers of possible patterns. [5]

Visualization

The topological distribution over a two-dimensional grid (outlined above) lends itself to a bitmap type visualization of the semantics of any word or text, where each active semantic feature can be displayed as e.g. a pixel . As can be seen in the images shown here, this representation allows for a direct visual comparison of the semantics of two (or more) linguistic items.

Image 1 clearly demonstrates that the two disparate terms "dog" and "car" have, as expected, very obviously different semantics.

Image 2 shows that only one of the meaning contexts of "jaguar", that of "Jaguar" the car, overlaps with the meaning of Porsche (indicating partial similarity). Other meaning contexts of "jaguar" e.g. "jaguar" the animal clearly have different non-overlapping contexts.

The visualization of semantic similarity using Semantic Folding bears a strong resemblance to the fMRI images produced in a research study conducted by A.G. Huth et al., [14] [15] where it is claimed that words are grouped in the brain by meaning. voxels , little volume segments of the brain, were found to follow a pattern were semantic information is represented along the boundary of the visual cortex with visual and linguistic categories represented on posterior and anterior side respectively. [16] [17] [18]

Notes

References