Identifying Semantic Components From Cross-Language Variation, Structured Lexical Sources, and Corpora: a Review of Current Literature

UC Berkeley Language and Cognition Lab
Jongmin J. Baek

August 5, 2016

Contents

Chapter 1

Overview

Distributional semantics makes a compelling case of how meaning arises from statistical distributions of data. So far, most accounts of distributional semantics have focused on English words and using vectors to represent those words. However, that landscape is quickly changing. Recent efforts on distributional semantics may be divided into at least four clusters: using more complicated mathematics to represent words and compositions of words, leveraging high-quality structured knowledge bases, combining multiple languages, and going beyond corpus training data to visual, auditory, and olfactory training data, to giving the machine multiple modalities, so to speak. What emerges is an exciting hodgepodge of ideas that look ripe to advance further and intertwine together. In what follows, I have listed papers of interest with a brief summary of their methods and, if particularly promising, an extended discussion of their formalizations and limitations.

Chapter 2

Algebraic Manipulations of Distributional Semantic Representations

2.1 Semantic Compositionality through Recursive Matrix-Vector Spaces (2012, Stanford)

Socher, Huval, Manning, & Ng discuss a method to assign a "vector and a matrix to every node in a parse tree: the vector captures the inherent meaning of the constituent, while the matrix captures how it changes the meaning of neighboring words or phrases." In this way, the matrix-vector recursive neural network, henceforth called MV-RNN, can learn "compositional vector representations for phrases and sentences of arbitrary syntactic type and length".

2.1.1 Formalization

First, a sentence or a phrase is parsed into a binary parse tree of arbitrary depth. Then a composition function $f : \mathbb{R}^{2n} \to \mathbb{R}^n$ is applied to pairs of constituents in a parse tree. More specifically, this function f is

$$f_{A,B}(a, b) = f(Ba, Ab) = g(W \begin{bmatrix} Ba \\ Ab \end{bmatrix})$$

where g is some nonlinearity function such as tanh or sigmoid, A, B are matrices for single words, a, b are vectors for single words, $W \in \mathbb{R}^{n \times 2n}$ is a matrix that maps both transformed words back into $\mathbb{R}^n$. The kicker is that f takes a pair of n-dimensional vectors and maps them into one n-dimensional vector, and therefore each pair of nodes in a parse tree can be collapsed into one vector that can then be recursively recombined. In this way, f can be recursively applied to a parse tree to finally yield one n-dimensional vector that (hopefully) captures

syntactic structure and corresponding semantics of the sentence.
We can do something similar with the word matrices. For computing phrase matrices, we define the function

$$f_M(A, B) = W_M \begin{bmatrix} A \\ B \end{bmatrix}$$

where $W_M \in \mathbb{R}^{n \times 2n}$, so the output of the function is another $n \times n$ matrix, just like the input matrices.
Now we describe the learning algorithm. We want to train for the vector representation of each phrase. For that, we need an objective function. Our objective function is a softmax classifier that predicts class distribution over some classes, such as sentiment or relationship. The class labels go in a matrix W^{label}. The corresponding error function for each sentence s and its corresponding tree t, $E(s, t, \theta)$, minimizes the sum of cross-entropy errors at all nodes of the tree. With a regularization parameter λ the gradient of our overall objective function J becomes

$$\frac{\partial J}{\partial \theta} = \frac{1}{N} \sum_{(x,t)} \frac{\partial E(x, t; \theta)}{\partial \theta} + \lambda \theta$$

where $\theta = (W, W_M, W^{label}, L, L_M)$ are our model parameters. L, L_M are just the sets of all word vectors and word matrices, respectively.

2.1.2 Performance

- MV-RNN outperforms most comparable models in a adverb-adjective pair sentiment classification task. "It shows that the idea of matrix-vector representations for all words and having a nonlinearity are both important. The MV-RNN which combines these two ideas is best able to learn various compositional effects." The MV-RNN especially performs well in negation tasks, as it allows negation to completely shift the sentiment with respect to an adjective.

- MV-RNN can capture logical compositionality. It can learn propositional operators such as NOT and AND.

- MV-RNN outperforms similar models in sentiment prediction for sentences of arbitrary length. Its accuracy is 79.0, while the closest competitors' accuracy is 77.7. *However*, it is interesting to note that it failed to classify examples that require structured real-life knowledge, of the sort likely to be found in FrameNet. It wrongly classified as positive the review "Director Hoffman, his writer and Kline's agent should serve detention." It also wrongly classified as negative "A bodice-ripper for intellectuals."

2.1.3 Limitations

It has some limitations. The model has an ungodly number of parameters ($O(nd^2)$), where n is the number of words and d is the dimension of the word

vector. Therefore it has a tendency to overfit and consumes a lot of resources. The paper in the following chapter alleviates this problem by sharing matrices across words, so naturally words that serve similar functions, i.e. some prepositions, may share matrices.

2.2 Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank (2013, Stanford)

Socher, Perelygin, Wu, Chuang, Manning, Ng & Potts introduce a Sentiment Treebank and, more relevant to our interests, the Recursive Neural Tensor Network. "Recursive Neural Tensor Networks (RNTN) take as input phrases of any length. They represent a phrase through word vectors and a parse tree and then compute vectors for higher nodes in the tree using the same tensor-based composition function." They solve the problem of MV-RNN as discussed in the previous chapter, that "the number of parameters becomes very large and depends on the size of the vocabulary. It would be cognitively more plausible if there were a single powerful composition function with a fixed number of parameters."

2.2.1 Performace

- In a scale-of-1-to-5 semtiment classification task, RNTN outperforms MV-RNN by a small margin.
- In a full setnence binary sentiment classification task, RNTN pushes the state of the art from 80% to 85.4%.
- With sentences that have a 'X but Y' structure, RNTN obtains an accuracy of 41% compared to 37% of MV-RNN.
- Correctly classifies negation of a positive sentences 71.4%, negation of a negative sentence 81.8%.

2.3 GloVe: Global Vectors for Word Representation (2014, Stanford)

Jeffrey Pennington, Richard Socher, and Christopher Manning talk about Global Vectors. Global Vectors combine the collocation matrix LSA model and local context window neural embedding model such as Mikolov et al (2013a), combining the best of both worlds. The LSA model uses the statistical information more efficiently, while being poor at word analogy tasks; the neural embedding model is just the opposite. Global vectors take advantage of the good parts while eschewing the bad parts.

2.3.1 Formalization

The cost function is

$$J = \sum_{i;j=1}^{V} f(X_{ij})(w_i^T \tilde{w_j} + b_i + \tilde{b_j} - logX_{ij})^2$$

Where f is a weighting function for each word-word coocurrence pair, and the expression inside the squared power shows the desire to minimize the difference between the dot product of the words, plus some bias, minus the log of the coocurrence statistic. For a discussion of why this is used, please refer to the paper. f prevents rare coocurrences from severely distorting the cost function, and the example function used is

$$f(x) = \begin{cases} (x/x_{max})^\alpha & x < x_{max} \\ 1 & \text{otherwise} \end{cases}$$

where $x_{max} = 100$ and $\alpha = 3/4$ is chosen empirically.

2.4 Distributional Models of Preposition Semantics (2003, Stanford & U. of Edinburgh)

Colin Bannard and Timothy Baldwin present a method to compare prepositions modeled with distributional semantics. Their main contribution is to show that prepositions, long considered to be semantically vacuous becuase of their promiscuity, can feasibly be modeled with distributional semantics. First, they make a distinction between transitive and intransitive prepositions, and treat them differently; for example, the token "up" may be used transitively or intransitively, such as "time is up" or "it is up to you". Then the Pearson correlation is found between different prepositions to calculate inter-preposition similarity. It concludes with the note that preposition semantics cannot work without considering valence.

Chapter 3

Structured Knowledge Bases and Distributional Semantics

3.1 RC-NET: A General Framework for Incorporating Knowledge into Word Representations (2014, Microsoft Research)

Xu et al. take issue with the fact that statistical distributional semantics do not take into account structured knowledge bases such as WordNet and FrameNet. They propose a "novel framework to take advantage of both relational and categorical knowledge to produce high-quality word representations. This framework is built upon the skip-gram model (Mikolov et al. 2013a), in which we extend its objective function by incorporating the external knowledge as regularization functions."

- "To leverage relational knowledge, we define a corresponding regularization function by inheriting the similar idea from a recent study on multi-relation model (Bordes et al. 2013), which characterizing the relationships between entities by interpreting them as translations in the low-dimensional embeddings of the entities."

- "To incorporate the categorical knowledge, we define another regularization function by minimizing the weighted distance between those words with the same attributes."

- "Then, we combine these two regularization functions with the original

objective function of the skip-gram model."

3.2 Retrofitting Word Vectors to Semantic Lexicons (2015, CMU)

Manaal Faruqui et al. generalize Xu et al.'s ideas to work with a broader set of knowledge bases.

3.3 Single or Multiple? Combining Word Representations Independently Learned from Text and WordNet (2016, University of the Basque Country)

Josu Gokioetxea, Eneko Agirre, and Aitor Soroa explore a way to learn distributed representations independently from corpora and from structured knowledge bases like WordNet, then use simple concatenation or averaging methods to combine them. Notably, such simple methods outperform more sophisticated methods like Canonical Correlation Analysis (CCA) or retrofitting.

3.4 Leveraging Frame Semantics and Distributional Semantics for Unsupervised Semantic Slot Induction in Spoken Dialogue Systems (2014, CMU)

Yun-Nung Chen, William Yang Wang, and Alexander I. Rudnicky combine distributional semantics and frame semantics into one, creating a model that can induce the frame-semantic slots after hearing human speech in an unsupervised fashion. The pipeline goes this way:

- Generate text from the speech files using a generic automatic speech recognition algorithm.
- Use a FrameNet-trained statistical probabilistic semantic parser to generate initial frame-semantic parses.
- Adapt these initial frame-semantic parses to the semantic slots of the target semantic space, leveraging distributed word representations. The last step "distinguishes between generic semantic concepts and domain-specific concepts." This is because we may want to focus on a specific domain, for example for a conversation about restaurant recommendation. To do this, we must rank slot candidates and use one ranked highest. The ranking algorithm uses "(1) the normalized frequency of each slot candidate in

the corpus, since slots with higher frequency may be more important. (2) the coherence of slot-fillers corresponding to the slot... the coherence of the corresponding slot-fillers can help measure the prominence of the slots because they are similar to each other." (585)

3.5 Event Extraction and Classification by Neural Network Model (2016, National Taiwan Normal U.)

Bamfa Ceesay and Wen-Juan Hou present a single model to label semantic roles, then predict event types and subtypes using neural embeddings. The pipeline is:

- Transform 1-grams or 2-grams into feature vectors. Features consist of probability distributions over phrase type, grammatical function, position, voice, FrameNet and VerbNet semantic roles or verb types.
- Train a neural net with the input being a 1-gram or 2-gram and the output being an "event nugget": event type, event subtype, event realis, event mention. The paper discusses nine such events: Business, Conflict, Justice, Life, etc.
- After training, the model can predict, with about 70% accuracy, the event nugget given the feature vector of a 1-gram or a 2-gram.

Chapter 4

Multilingual Distributional Semantics

4.1 Bilingual Word Embeddings for Phrase-Based Machine Translation (2013, Stanford)

Will Zou, Richard Socher, Daniel Cer, and Christopher Manning present a method to embed both English and Madarin into one embedding space. They achieve this by using an objective function that is a weighted sum of a monolingual objective function and a translation equivalence objective functino.

- Use some off-the-shelf monolingual objective function. The paper uses Collobert et al. (2008)'s formulation, the Context Objective:

$$J_{CO}^{(c,d)} = \sum_{w^r \in V_R} \max(0, 1 - f(c^w, d) + f(c^{w^r}, d))$$

  Where f is an objective function defined by the neural network, w^r is a word chosen in a random subset V_R of the vocabulary, and c^{w^r} is a context window of w^r. So this model contrasts the score of when a word is placed in its correct context, versus when a random word is placed in the same context.

- Align English and Mandarin corpora using the Berkeley Aligner (Liang et al., 2006). Formally, they use the following equation to find starting word embeddings:

$$W_{t-init} = \sum_{s=1}^{S} \frac{C_{ts} + 1}{C_t + s} W_s$$

  where S is the number of possible target language words that may be aligned with the source word, C_{ts} is the number of times when word t in the target and word s in the source are aligned in the training text, and C_t

denotes the total number of counts of word t that appeared in the target language. Laplace smoothing is then applied to this function.

- Form alignment matrices $A_{en \to ch}$ and $A_{ch \to en}$. For the former, each row is a Chinese word and each column is an English word. An element $a_i j$ denotes the number of times the ith Chinese word was aligned with the jth English word. The latter matrix is defined similarly. Then our Translation Equivalence Objective is

$$J_{TEO-en \to ch} = ||V_{ch} - A_{en \to ch} V_{en}||^2$$
$$J_{TEO-ch \to en} = ||V_{en} - A_{ch \to en} V_{ch}||^2$$

- Finally, we optimize for a weighted combined objective during training:

$$J_{CO-ch} + \lambda J_{TEO-en \to ch}$$
$$J_{CO-en} + \lambda J_{TEO-ch \to en}$$

- After training, we can visualize the embedding space using t-SNE (van der Maateen 2008).

Figure 1: Overlaid bilingual embeddings: English words are plotted in yellow boxes, and Chinese words in green; reference translations to English are provided in boxes with green borders directly below the original word.

4.2 Massively Multilingual Word Embeddings (2016, CMU & U. of Washington)

Waleed Ammar, Gergoe Mulcaire et al. use pairwise parallel dictionaries and monolingual corpus data to create a single embedding space for more than fifty languages. They use three methods, multiCluster, multiCCA, and multiQVEC-CCA, of which multiQVEC-CCA performs best.

4.2.1 MultiCluster

- Using the parallel dictionary, map each word-language pair to a multilingual cluster. More precisely, each cluster is a connected graph such that

the each node is a word-language pair and the each edge is an entry in the dictionary that indicates translational equivalence.

- Give each cluster an arbitrary unique ID. Replace each instance of the word in each corpus with the corresponding cluster ID.
- With the resulting corpus of ID's, use the standard skipgram model of Mikolov et al. (2013a) to assign a vector to each cluster just as you would a normal, monolingual corpus.

4.2.2 MultiCCA

- Using monolingual corpora, learn monolingual embeddings for each language separately. For languages n and m, we write E^n and E^m as the embeddings.
- Use canonical correlation analysis (CCA) to estimate linear projections from the ranges of the monolingual embedding spaces, yielding a bilingual embedding space. More formally, the optimization problem is

$$\text{MAX}\{\text{Corr}(T_{m\to m,n}E^m(u), T_{n\to m,n}E^n(v))\forall(u,v)\in D^{m,n}\}$$

where $T_{m\to m,n}, T_{n\to m,n} \in \mathbb{R}^{d\times d}$ are linear projections from the language n's embedding space to the bilingual embedding space, and from the language m's embedding space to the bilingual embedding space. $D^{m,n}$ is a bilingual dictionary.

Chapter 5

Multimodal Distributional Semantics

5.1 Multimodal Distributional Semantics (2014, University of Trento & L3S Research Center

A 47-page survey on the topic. "For all its successes, distributional semantics suffers of the obvious limitation that it represents the meaning of a word entirely in terms of connections to other words... (this is) deeply problematic, an issue that is often referred to as the *symbol grounding problem*."

5.2 Grounded Compositional Semantics for Finding and Describing Images with Sentences (2014, Stanford & Google)

Richard Socher, Andrej Karpathy, Quoc Le, Christopher Manning and Andrew Ng introduce the Dependency Tree Recursive Neural Network (DT-RNN) model to "embed sentences into a vector space in order to retrieve images that are described by those sentences." (207) While normal RNN models, using binary constituency trees, are sometimes very effective at this task, they inevitably capture a lot of the syntactic information of a sentence, which is not desirable when wanting to retrieve an image with a sentence description. For example, "the man is on the bike" and "the bike has the man on it" should give roughly similar vector representations, which normal RNN models fail to do. On the other hand, DT-RNNs naturally focus on a sentence's actions and its agents, because the root usually splits to its NP and VP.

5.3 Distributional Semantics from Text and Images (2011, Free University of Bolzano & University of Trento)

Bruni, Tran, and Baroni present a distributional semantic model that learns not only from words but also from images. This model captures qualitatively different aspects of meaning; while traditional distributional semantic models that learns only from corpus data usually capture abstract meaning, this model captures visual meaning. They go on so far as to say "we consider this (very!) preliminary evidence for an integrated view of semantics where the more concrete aspects of meaning derive from perceptual experience, whereas verbal associations mostly account for abstraction."

5.4 Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception (2015, Cambridge)

Douwe Kiela and Stephen Clark discuss a way to give machines ears. Just as visual multi-modal models use "bag of visual words" (BoVW) representations, Kiela and Clark use "bag of audio words" (BoAW) representations. They use the online search engine Freesound to download audio files tagged with keywords. What results is a model that finds auditorily similar nearest neighbors in an almost poetic way.

Auditory				Linguistic			
navy	language	gossip	dinner	navy	language	gossip	dinner
army	mouth	maid	meal	army	word	news	lunch
aviation	man	guest	lunch	military	words	newspaper	wedding
plane	father	elevator	writer	vessel	literature	cute	meal
jet	adult	danger	breakfast	sunk	dictionary	sexy	breakfast
cannon	throat	corridor	couch	ship	tongue	mirror	cocktail
monster	motor	water	dawn	monster	motor	water	dawn
orchestra	engine	stream	summer	zombie	vehicle	droplets	dusk
demon	rain	bath	child	demon	automobile	salt	sunrise
guitar	beach	river	victor	dragon	car	cold	moon
beast	boat	bathroom	morning	beast	motorcycle	sunlight	night
pilot	car	rain	garden	creatures	truck	milk	misty

5.5 Grounding Semantics in Olfactory Perception (2015, Cambridge)

Douwe Kiela and Stephen Clark discuss a way to give machines nostrils. Just as visual multi-modal models use "bag of visual words" (BoVW) representations, Kiela and Clark use "bag of chemical compounds" (BoCC) representations. They use the Sigma-Aldrich Fine Chemicals flavors and fragrances catalog (SAFC), the largest public olfactory databse, to find 137 smells and their 11,152 associated chemical compositions.

apple	bacon	brandy	cashew
pear	smoky	rum	hazelnut
banana	roasted	whiskey	peanut
melon	coffee	wine-like	almond
apricot	mesquite	grape	hawthorne
pineapple	mossy	fleshy	jam
chocolate	**lemon**	**cheese**	**caramel**
cocoa	citrus	grassy	nutty
sweet	geranium	butter	roasted
coffee	grapefruit	oily	maple
licorice	tart	creamy	butterscotch
roasted	floral	coconut	coffee

Table 6: Example nearest neighbors for BoCC-SVD representations.