

Computational Models of Language Meaning in Context

Edited by

Hans Kamp¹, Alessandro Lenci², and James Pustejovsky³

1 Universität Stuttgart, DE, Hans.Kamp@ims.uni-stuttgart.de

2 University of Pisa, IT, alessandro.lenci@ling.unipi.it

3 Brandeis University, Waltham, USA, jamesp@cs.brandeis.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13462 “Computational Models of Language Meaning in Context”. The seminar addresses one of the most significant issues to arise in contemporary formal and computational models of language and inference: that of the role and expressiveness of distributional models of semantics and statistically derived models of language and linguistic behavior. The availability of very large corpora has brought about a near revolution in computational linguistics and language modeling, including machine translation, information extraction, and question-answering. Several new models of language meaning are emerging that provide potential formal interpretations of linguistic patterns emerging from these distributional datasets. But whether such systems can provide avenues for formal and robust inference and reasoning is very much still uncertain. This seminar examines the relationship between classical models of language meaning and distributional models, and the role of corpora, annotations, and the distributional models derived over these data. To our knowledge, there have been no recent Dagstuhl Seminars on this or related topics.

Seminar 10.–15. November, 2013 – www.dagstuhl.de/13462

1998 ACM Subject Classification 1.2.7 Natural Language Processing

Keywords and phrases formal semantics, distributional semantics, polysemy, inference, compositionality, Natural Language Processing, meaning in context


Digital Object Identifier 10.4230/DagRep.3.11.79

1 Executive Summary

Hans Kamp

Alessandro Lenci

James Pustejovsky

License  Creative Commons BY 3.0 Unported license
© Hans Kamp, Alessandro Lenci, and James Pustejovsky

The term distributional semantics qualifies a rich family of computational methods sharing the assumption that the statistical distribution of words in context plays a key role in characterizing their semantic behavior. Distributional semantic models, such as LSA, HAL, etc., represent the meaning of a content word in terms of a distributed vector recording its pattern of co-occurrences (sometimes, in specific syntactic relations) with other content words within a corpus. Different types of semantic tasks and phenomena are then modeled in terms of linear algebra operations on distributional vectors. Distributional semantic models provide a quantitative correlate to the notion of semantic similarity, and are able to address various lexical semantic tasks, such as synonym identification, semantic classification, selectional preference modeling, and so forth.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Computational Models of Language Meaning in Context, *Dagstuhl Reports*, Vol. 3, Issue 11, pp. 79–116

Editors: Hans Kamp, Alessandro Lenci, and James Pustejovsky



DAGSTUHL
REPORTS

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Distributional semantics has become increasingly popular in Natural Language Processing. Its attractiveness lies in the fact that distributional representations do not require manual supervision and reduce the the a priori stipulations in semantic modeling. Moreover, distributional models generally outperform other types of formal lexical representations, such as for instance semantic networks. Many researchers have also strongly argued for the psychological validity of distributional semantic representations. Corpus-derived measures of semantic similarity have been assessed in a variety of psychological tasks ranging from similarity judgments to simulations of semantic and associative priming, showing a high correlation with human behavioral data.

Despite its successes, no single distributional semantic model meets all requirements posed by formal semantics or linguistic theory, nor do they cater for all aspects of meaning that are important to philosophers or cognitive scientists. In fact, the distributional paradigm raises the question of the extent to which semantic properties can be reduced to combinatorial relations. Many central aspects of natural language semantics are left out of the picture in distributional semantics, such as predication, compositionality, lexical inferences, quantification and anaphora, just to quote a few. A central question about distributional models is whether and how distributional vectors can also be used in the compositional construction of meaning for constituents larger than words, and ultimately for sentences or discourses – the traditional domains of denotation-based formal semantics. Being able to model key aspects of semantic composition and associated semantic entailments represents a crucial condition for distributional model to provide a more general model of meaning. Conversely, we may wonder whether distributional representations can help to model those aspects of meaning that notoriously challenge semantic compositionality, such as semantic context-sensitivity, polysemy, predicate coercion, pragmatically-induced reference and presupposition.

The main question is whether the current limits of distributional semantics represent contingent shortcomings of existing models – hopefully to be overcome by future research –, or instead they point to intrinsic inadequacies of vector-based representations to address key aspects of natural language semantics. To this end, there were five themes addressed by the participants:

1. The problems in conventional semantic models that distributional semantics claims to be able to solve;
2. The promise of distributional semantics linking to multimodal representations
3. The current limitations of distributional semantics theories to account for linguistic compositionality;
4. The absence of any robust first-order models of inference for distributional semantics;
5. The integration of distributional semantic principles and techniques into a broader dynamic model theoretic framework.

2 Table of Contents

Executive Summary

Hans Kamp, Alessandro Lenci, and James Pustejovsky 79

Overview of Talks

Towards a distributionally motivated formal semantics of natural language

Hans Kamp 83

Model Theory and Distributional Semantics

Katrin Erk 83

Implicative uses of evaluative factive adjectives

Lauri Karttunen 83

Formal Semantics and Distributional Semantics: A Survey of Chance and Challenge

Hinrich Schütze 84

Working Groups 84

Statements on Distributional Semantics by the Seminar Participants 85

Dr Strangestats or How I Learned to Stop Worrying and Love Distributional Semantics

Marco Baroni 85

Position statement

Peter Cariani 86

Position statement

Stephen Clark 90

Position statement

Ann Copestake 93

Position statement

Katrin Erk 94

Position statement

Stefan Evert 97

Statements on Distributional Semantics

Sebastian Löbner 98

Putting together the pieces

Louise McNally 98

Incrementality in Compositional Distributional Semantics

Alessandra Zarcone, Sebastian Padó 101

Acquiring conceptual knowledge for semantic interpretation

Massimo Poesio 102

Research overview

Tim Van de Cruys 105

Semantics, Communication, and Probability

Jan van Eijck 107

Position statement <i>Dominic Widdows</i>	108
Norms and Exploitations in Text Meaning and Word Use <i>Patrick Hanks</i>	109
Position statement <i>Anna Rumshisky</i>	110
Position statement <i>Annie Zaenen</i>	112
Panel Discussions	113
Polysemy	113
Inference	113
Compositionality	114
Negation	114
Next Steps	115
Participants	116

3 Overview of Talks

3.1 Towards a distributionally motivated formal semantics of natural language

Hans Kamp (Universität Stuttgart, DE)

License  Creative Commons BY 3.0 Unported license
© Hans Kamp

Formal models of semantics for natural language have proved to be very powerful and useful in their description of linguistic phenomena. To date, there is no distributional semantic model that satisfies the requirements posed by formal semantics or linguistic theory for modeling meaning. Nor does distributional semantics to my knowledge address issues of meaning that are important to philosophers or cognitive scientists. In fact, the distributional paradigm raises the question of the extent to which semantic properties can be reduced to combinatorial relations. Many central aspects of natural language semantics, such as predication, compositionality, lexical inferences, quantification and anaphora, seem left out of the picture. The challenge is to find models that have the explanatory adequacy of formal semantic theories, but which are at the same time able to capture the contextual and distributional nature of language use.

3.2 Model Theory and Distributional Semantics

Katrin Erk (University of Texas – Austin, US)

License  Creative Commons BY 3.0 Unported license
© Katrin Erk

What is the denotation of distributional representations? It seems reasonable to say that purely linguistic data can change our beliefs about the world. But does this also hold for distributional information? After all, distributional data just counts sentential contexts in which words have been observed, and it is not clear how we could derive truth conditions from a distributional vector. But distributional information can do something less: It can provide tentative, uncertain information about similarities between different predicates that have been mentioned in the text. And this, I think, suffices to reduce our uncertainty about which world we are in. We can describe this in a probabilistic semantics setting. We have a prior probability distribution over worlds, which can be updated (in a standard fashion, by Baye's rule) using uncertain distributional information.

3.3 Implicative uses of evaluative factive adjectives

Lauri Karttunen (Stanford University, US)

License  Creative Commons BY 3.0 Unported license
© Lauri Karttunen

Evaluative adjectives such as stupid, smart, lucky, sweet, cruel can take propositional complements as in *John was smart to leave early*. In this construction they are generally considered to be factive: *John was not smart to leave early* is supposed to mean that John did leave early and that it was not a good idea. When one looks at the web, however, one finds easily examples where the intended meaning is not factive, as in *Luckily, I was not*

stupid to send them any money. Amazon MT experiments and corpus studies confirm the existence of this pattern. We will discuss whether it should be seen as a performance error or whether there seems to be a real 'dialect' split among speakers of (American) English. Whatever the analysis, it seems that the pattern is prevalent enough for NLP applications that assign factuality judgments to events to need to take it into account.

3.4 Formal Semantics and Distributional Semantics: A Survey of Chance and Challenge

Hinrich Schütze (LMU München, DE)

License  Creative Commons BY 3.0 Unported license
© Hinrich Schütze

I will first describe what I see as the strengths and weaknesses of distributional models on the one hand and formal models on the other hand. I will then contrast two different types of distributional models, count vector models and deep learning embedding models. The main part of the talk will be about compositionality and about the extent to which distributional and formal semantic models can handle different aspects of compositionality.

4 Working Groups

Participants were assigned to one of four groups, each discussing a specific set of questions related to the seminar topic. The topics are given below.

1. **Polysemy and Vagueness**
 - type coercion, metonymy, complex types,
 - metaphor, figurative language
 - issues of lexical inference (for non-function words)
 - semantic relations
2. **Inference and Reasoning**
 - structural deduction based on the representational syntax, axioms, and inference rules.
 - Inference from a DS perspective: computation over and similarity of vectors?
 - What to do about Quantification
 - But inference is not just deduction; Can DS distinguish between deduction, induction, and abduction?
 - Defeasibility and default logics how do these stand up against the more natural soft constraints given by distributional techniques and probabilistic reasoning.
3. **Compositionality**
 - function application
 - selectional preferences are handled well in DS. What about type shifting?
 - semantic roles,
 - Basic semantics of predication in DS
4. **Modality and Negation**
 - Negation: difficult to handle in DS.
 - Tense: put it on the map
 - Deontic logic
 - Epistemic logic and reasoning about knowledge and beliefs

The group composition was as follows:

Group 1: Stefan Evert, Tim van de Cruys, Patrick Hanks, Sebastian Löbner, Suzanne Stevenson, Alessandra Zarcone

Group 2: Ann Copestake, Ido Dagan, Jan van Eijck, Graeme Hirst, Sebastian Padó, Anna Rumshisky, Dominic Widdows


Group 3: Marco Baroni, Stephen Clark, Katrin Erk, Jerry Hobbs, Alessandro Lenci, Louise McNally, Massimo Poesio

Group 4: Nicholas Asher, Peter Cariani, Hans Kamp, Lauri Karttunen, James Pustejovsky, Hinrich Schütze, Mark Steedman, Annie Zaenen

5 Statements on Distributional Semantics by the Seminar Participants

5.1 Dr Strangestats or How I Learned to Stop Worrying and Love Distributional Semantics

Marco Baroni (University of Trento, IT)

License  Creative Commons BY 3.0 Unported license
© Marco Baroni

I was a teenage generativist. I was raised in fairly observant Chomskyan schools, and I still abide by the program for linguistics as the algorithmic study of human language competence Chomsky laid out 60 years ago. Then, how did I become an adult werelinguist, theoretical semanticist by day, corpus-based, statistics-driven computationalist at night? I don't do corpus-based, statistics-driven distributional semantics because I am, in principle, attracted by or sympathetic towards usage-based, nonsymbolic, inductive approaches to language. I do distributional semantics because at a certain point I discovered that it is the only semantic formalism allowing me to do my job as a linguist. I first felt the need for semantics while writing my master thesis about derivational morphology, where the salience of morpheme boundaries predicts phenomena such as the likelihood that an affix undergoes phonetic reduction, blocking of phonological rules, morphemic-route access in lexical retrieval, etc. But one of the main factors determining, in turn, the salience of morpheme boundaries is semantic transparency, that is, the extent to which the meaning of a derived word is related to the meaning of its stem, (cf. re-decorate vs. recollect). I then started looking around for an approach to semantics that would (i) provide large-scale coverage of the lexicon and (ii) make quantitative predictions about degrees of similarity (or relatedness). The first requirement came from the fact that I needed to account for the often semantically arbitrary sets of stems and derived forms that were subject to specific morpheme salience phenomena. The second requirement derived from the fact that, in all phenomena I looked into, the effect of semantic transparency was never all or nothing, but rather a fuzzy phenomenon with many intermediate cases, so I needed a theory making graded predictions.

Formal approaches to semantics, even those that paid attention to lexical meaning, failed both requirements. The functionalist stuff, while in principle sympathetic to the idea of degrees of similarity, was too awfully fuzzy, not explicit enough to make quantitative predictions, and in any case failing the coverage requirement. Unfortunately, I discovered distributional semantics too late to use it in my morphology work, where I just gave up the idea of accounting for semantic transparency effects, but from when, years later, I discovered

LSA and its cousins, I never found a reason to go back to other approaches to semantics, simply because, from a practical point of view, I still see no alternative to the distributional approach. Something I've learned along the way is that being able to quantify degrees of semantic similarity is not only good for tasks such as assessing the semantic transparency of derived forms or finding near synonyms. Distributional semanticists (including some that will attend this seminar) came up with clever and elegant ideas to account, in terms of semantic similarity, for complex linguistic phenomena such as predicting the selectional preferences of verbs, capturing argument alternation classes or accounting for co-composition effects. And there is ongoing and very promising work (that, I think, will be discussed at the seminar) on dealing with fundamental challenges for distributional semantics such as polysemy or scaling up to phrase and sentence meaning. So, while there is a lot of hard work ahead of us, I'm confident that in a few years we will have empirically successful models of distributional semantics that are not limited to single words in isolation, and, equipped with these new models, we will be able to account for many more linguistic phenomena in terms of semantic similarity.

Still, current distributional semantics is entirely prisoned inside a linguistic cage: all it can tell us (and that's not little) is how similar words, phrases and sentences are to each other. Without a hook into the outside world, all we will be able to do is to measure how similar, say, the sentence "A boy is laughing" is to "A girl is crying", but we will never be able to tell whether either sentence can be truthfully asserted of the current state of the world. While I understand that there is much more to the outside world than this, I think that one first, reasonable step we can take is to explore whether we can connect distributional semantic representations with our visual perception of the world. In concrete, we should aim for a system that, given a picture depicting a scene with, say, a laughing boy, could tell us that A boy is laughing is an appropriate statement describing the scene. Interestingly, state-of-the-art image analysis systems represent images not unlike distributional semantics represents words – that is, images are represented by vectors that record the distribution of a set of discrete feature occurrences in them. So, there is hope, and I think a central goal for distributional semantics in the next few years should be to work on how to develop a common semantic space, where vector-based representations of linguistic expressions, on one side, and objects and scenes, on the other, can be mapped and compared. Given such shared linguistic-visual semantic space, the same similarity scoring techniques we are already using in distributional semantics might be extended to account for referential aspects of meaning: The sentence "A boy is laughing" is truthfully stated of (a picture depicting) a scene if the vector representing the sentence and the vector representing the scene are above a certain threshold of similarity. My colleagues and I are currently working on methods to build the proposed shared linguistic-visual semantic space (and other researchers are also making good progress in this direction). At the seminar, I would like to discuss (among many other things, of course) both concrete ideas about how to construct the common space, and what are linguistically interesting scenarios in which we could make use of it.

5.2 Position statement

Peter Cariani (Harvard Medical School – Newton, US)

License  Creative Commons BY 3.0 Unported license
© Peter Cariani

I think that perhaps I am the outlying point, the wild card here, so I will try to explain myself. I come to questions of meaning acquisition/construction from a naturalistic, pragmatic,

constructivist perspective that is heavily influenced by cybernetics and systems theory; auditory, computational, and theoretical neuroscience; and perceptual and cognitive psychology. I currently teach courses related to the neuropsychology of music, which I think has some deep parallels with the kinds of computational semantics questions we have here before us. I hope that it will be useful to the group have an independent external perspective. My intention is not to distract or detain you from the nuts-and-bolts aspects of the specific questions at hand before us (I'm sure you would all be happy just hashing through the minutiae of your sub-fields, as would I), but I want to try to stand back (since this is the only place I can stand here) and raise broader questions when (if) needed. Here are some of the basic issues I see:

- (I) Statistics vs. structure in interpretation/anticipation
- (II) Why do semantic analysis? Computational tools vs. modeling minds/brains
- (III) Getting pragmatics into computational semantics
- (IV) Implementations: Symbols vs. connectionism vs. something else
- (V) Can computational semantics (ultimately) understand human life?

I. Statistics vs. structure in interpretation/anticipation

A. Statistics-based learning and interpretation. The fundamental issues really go back to old and unresolved debates about the how minds and brains work, i.e. how much of human cognition is driven by the statistics of external input patterns vs. by the internal organization of mental processes. The answer is that both aspects play important roles, that minds/brains are anticipatory systems that register, remember, and act upon external event statistics, albeit very heavily filtered through a powerful mental apparatus that ever attempts to predict the future by constructing highly structured models of the world. What is (could be) the relationship between these two kinds of anticipatory processes, in minds, brains, and machines? For example in this current discourse before us, I see the distributional semantics approach as part of a larger resurgence of associationist psychology that I believe has been fueled in recent decades by the (perceived and real) successes of hidden-Markov models for automatic speech recognition. In the neurosciences, over the past two decades, there has been a blossoming of interest in Bayesian perceptual models and the statistics of natural scenes. These methods have their own practical applications and efficacies, but every powerful information technology eventually becomes a model of minds and brains for some fraction of those who use it. We need to be clear about whether our purpose is to develop computational tools that serve as adjuncts to our own reasoning and meaning-making (e.g. more effective search engines or corpora analyzers) or whether we are trying to model human mental processes of meaning formation. Coming out of the applied mathematics of statistics-based machine learning, it seems to me that distributional semantics tends to view itself as a set of useful techniques (and perhaps the mind as an assemblage of such hacks, as Minsky thinks). The underlying (often tacit) assumptions of these models are that minds (sensory, cognitive systems) do not have strong internal structure and adapt to the statistics of incoming information (in whatever modality or form).

B. Structure-based learning and interpretation. On the other hand, are what I think of as structuralist theories, in the old psychological sense of that term, a la Titchener and Piaget and the Gestaltists, that hold that there is strong dimensional structure to mental processes, and that therefore it is necessary to model those structural constraints if we are to understand and predict human interpretation and to replicate its functionalities in artificial systems. Logic- and model-based approaches to semantics share with structuralist psychology that there are strong constraints (I use the term low dimensional structure), and

that the crux of understanding systematicity and compositionality lies in the underlying sets of basic informational processes (symbols and rules, neural/mental representations and operations) that are operant in logics/models/minds/brains. Clearly both kinds of mechanisms are operant in minds and brains. There is widespread evidence that humans and animals learn the statistics of their surrounds and adapt to them, such that, in the absence of better predictive information, they will produce expectancies based on those statistics. For example, we see this in music perception when listeners are exposed to artificial scales (e.g. Bohlen–Pierce) and come to expect those musical intervals - they adapt to the pitch statistics of their recent experience. However, this statistical prediction is a weak expectancy, and it is easily superseded if there is strong predictive structure in the music (repeating phrases, motifs, sections, rhythmic patterns, etc.). Musical expectancy is a combination of what I call pattern (structure) and frame (statistics)³ I am currently working on neural timing net models for rhythmic pattern expectancy – when there is longer range repeating structure, that dominates; in lieu of longer-range structure, basic event probabilities dominate. In speech perception, I think we only use prior phoneme and word probabilities when signal-to-noise ratios are low – otherwise, when signals are clear, deterministic auditory pattern recognition processes dominate and we can easily achieve 100 accuracy identifying strings of nonsense syllables and words.

II. Developing effective computational tools vs. modeling the mind

I can see already that the different approaches (model-based vs. distributional semantics) have different purposes. The latter can leverage the awesome power of computer statistical analysis over extremely large and varied digital corpora. The former, however, hold out the even greater promise of an eventual theory of how minds make meanings, and if we can solve those hard problems, we shall have much more effective digital analysis technologies. In this discussion we need to be as clear as possible about what our goals are re: computational semantics – otherwise we will discuss and/or argue at cross-purposes.

III. Getting pragmatics into computational semantics

Pragmatic frames should be central to both model-theoretic and distributional semantics. The perceived intended purposes of communications we receive and texts that we interpret play heavily into how we interpret the meaning of the message. In terms of forming interpretive meanings of human and animal communications, I think pragmatics comes first, semantics second, and syntactics third. We humans are already primed heavily by the situational pragmatics to assume the nature of the message (neutral communication, threat, warning, command, question, affective expression, etc.), that in turn bias selection of semantic senses, that rapidly form a conceptual model of the contents of the message. We then do a detailed syntactic analysis if there are unresolved incongruities, if the model doesn't make any sense or if the contents of the message don't comport with the nature of the communication. This has practical implications. The distributional strategy of product (or music) recommendation based on co-occurrences of looks or purchases (those people who looked at this eventually bought that, or those people who like this music also liked that) is useful in that it indicates a correlation that may have a relevant underlying cause. However, the correlation does not inform the prospective buyer of why people liked or bought those things they chose. Really, except for purely imitative buyers/listeners, this is what we want to know, why we should choose one thing over another. We want to choose our music by what we want from it (e.g. happy/sad/interesting/comforting/surprising/nostalgic/arousing/sleep

inducing/meditative/distracting/identity-affirming/etc. music). Music recommendation systems are beginning to do this, but it requires a structural theory of the effects of different kinds of music and musical parameters on internal psychological states. Purposes of actors and pragmatic contexts can be incorporated into model-theoretic accounts (as pragmatist philosophy holds, truth is really efficacy relative to purpose). There are also ways that distributional approaches might incorporate pragmatic observables. I believe that brains encode information in a manner that allows a given event, object, or association to be content-addressable via all of its manifold aspects: pragmatic, semantic, syntactic. This is how we solve Dreyfus' frame problem, we can search by purpose, by effect on the world, by effect on us, by form and bring up those relevant dimensional aspects that we need in a given situation.

IV. Symbols vs. connectionism vs. something else

I am a theoretical-computational auditory neuroscientist and have been dealing with the whole issue of time codes in the brain. The focus of this workshop is not on neural models per se, but in essence I think that the distributional-model-theoretic semantics discussion has many parallels with the symbols vs. connectionism debate twenty years ago. I had a front row seat at the MIT debate between Smolensky, champion of connectionist neural nets, and the tag team Pylyshyn and Fodor, champions of symbolic computations. I was rooting for the neural networks, but symbols easily carried the day. How minds realize universals, abstract categories, systematicity, and compositionality are fundamental problems that neuroscience and psychology need to solve in order to construct an adequate theory of mind. It's a useful heuristic to try to imagine how minds work, i.e. how brains operate to form meanings and interpretations. In terms of neural activity patterns, it appears to me that all of these aspects simultaneously activate in parallel respective sets of neural assemblies (a la Lashley and Hebb) that in effect resonate with each other to different degrees, such that subsets of neural assemblies implementing different interpretations reinforce each other, with different pattern-resonant subsets competing with other subsets. The end result is a parallel-analysis and competitive winner-take-all process, but one in which later, conflicting information can reverse earlier dominant interpretations (defeasible constraints).

V. Can computational semantics (ultimately) understand what it means to be human?

I know this is very philosophical, but maybe it is worth thinking about. I think we should always try to think as far ahead as we can about where (how far) these theories can take us. Even above the questions of the respective efficacies and limitations of model-theoretic vs. distributional semantics that we will hash out here, there are some general questions of the extent to which formal systems (either logic-based models or full-blown psychological models of human minds) can capture private and public meanings. It would seem to me that if we had an adequate theory of the brain, such that we could simulate its information processing aspects properly, including sensorimotor transactions with the environment and embedded internal reward systems, that we could have an adequate model for human meaning. I don't believe that the brain or mental processes involved are necessarily logical in the truth-theoretic sense (e.g. each of us simultaneously holds sets of logically conflicting beliefs; moral and political reasoning is notoriously based on competing modes of thinking that are based on largely complementary imperatives). This begs the question of whether a computer, using only the encoded text resources of the internet, could possibly understand what it is like to be human and to interpret texts in those terms (in terms of the meanings of things).

Or in other words, would the machine need itself to have needs, drives, feelings, friends, enemies, memories to interpret texts in the ways that we do? This sounds like Hubert Dreyfus' frame problem (which I think brains solve by encoding all memories of events and their hedonic outcomes in pragmatic and semantic terms – we have memory that is content-addressable both by semantics–world effect and pragmatics–use effect. The midline dopamine systems encode the internal time– sequences of all the neural events that lead up to reward or punishment, such that both relations between perceived world–events and relations between actions, world–events, and rewards can be predicted). Can all those aspects of our internal structure and our interactions with the world be modeled and simulated, such that the machine will extract a meaning that is similar to one we would produce?

References

- 1 Cariani, P. 1989. On the Design of Devices with Emergent Semantic Functions. Ph.D., State University of New York at Binghamton.
- 2 Cariani, P. A. Delgutte, B. 1996a. Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *J Neurophysiol*, 76, 1698–716. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. *J Neurophysiol*, 76, 1717–34.
- 3 Cariani, P. 2001a. Cybernetics and the semiotics of translation. In: Petrilli, S. (ed.) *Lo Stesso Altro: Athanor: arte, letteratura, semiotica, filosofia*, v. XII, n. 4. Rome: Meltemi.
- 4 Cariani, P. A. 2001b. Neural timing nets. *Neural Netw*, 14, 737–53.
- 5 Cariani, P. 2002. Temporal codes, timing nets, and music perception. *J. New Music Res.*, 30, 107–136.
- 6 Cariani, P. 2011. The semiotics of cybernetic percept-action systems. *International Journal of Signs and Semiotic Systems*, 1, 1–17.
- 7 Cariani, P. 2012. Creating new primitives in minds and machines. In: McCormack, J., D'Inverno, M. (eds.) *Computers and Creativity*. New York: Springer.
- 8 Cariani, P., Micheyl, C. 2012. Towards a theory of information processing in the auditory cortex. In: Poeppel, D., Overath, T., Popper, A. (eds.) *Human Auditory Cortex: Springer Handbook of Auditory Research*. New York: Springer.

5.3 Position statement

Stephen Clark (University of Cambridge, GB)

License  Creative Commons BY 3.0 Unported license
© Stephen Clark

My current research goal is to develop compositional techniques for distributional semantics. This goal is relevant for the scientific enterprise of computational linguistics, since the accounts of distributional semantics currently lack a satisfactory compositional treatment; and also for the engineering enterprise of natural language processing (NLP), since representing the meanings of phrases and larger units in a vector space will allow the calculation of semantic similarity for those phrases and larger units. Calculating semantic similarity is crucial for many NLP tasks and applications. In collaboration with Bob Coecke (Oxford) and Mehrnoosh Sadrzadeh (Queen Mary), I have developed a tensor-based theoretical framework for distributional semantics which applies readily to variants of categorial grammar [5]. The idea is that the syntactic type of a constituent determines its semantic type; for example, the meaning of a transitive verb in English is represented as a 3^{rd} order tensor. Tensors are multi-linear maps in multi-linear algebra; hence the framework encapsulates the old idea

from formal semantics that the meanings of some words and phrases can be represented as functions. The compositional operation which combines the tensors is tensor contraction (just matrix multiplication extended to the multi-linear algebra case). Moreover, since tensors are functions, the combinatory operations of Combinatory Categorical Grammar (CCG) [14] transfer over to the tensor-based framework in a straightforward way, meaning that the framework applies to CCG as well as the context-free pregroup grammars of Lambek used in our original papers [11].

A useful instance of the framework to consider is adjective-noun modification. Since the syntactic type of an adjective in English is N/N , its semantic type is $N \otimes N$; a vector in $N \otimes N$ is a matrix representing a function from the noun space onto the noun space. Composition of an adjective and noun is achieved with matrix multiplication. In fact this is the proposal of [2] (independently conceived), which can be seen as an instance of our more general framework. The framework is currently largely a theoretical framework, with some small-scale attempts at implementation [8, 7, 10], and additional theoretical work building on it [3]. There are a number of practical and theoretical stumbling blocks in the way of a large-scale implementation. Many of these stumbling blocks are fundamental questions relating to natural language semantics, and in particular semantics in context, and hence of relevance to the Seminar.

Questions

What is the sentence space? The theoretical framework assumes a separate vector space for sentences, S , compared with nouns, which live in N . (There may be other spaces corresponding to the basic syntactic types, also, for example PP .) However, the framework only dictates how to compose functions and arguments to deliver a vector in that (assumed) space; it does not place any constraints on what the sentence space should be. This raises the question of whether it makes sense to represent the meanings of sentences in a vector space, and how structured should such a sentence space be? The answer may depend on the application; for example for sentiment analysis, a simple space of positive/negative may suffice. Another way to ask the same question is whether phrases and sentences should live in the same space as nouns (as they do in the neural-network based work of Socher, for example [13]). Making this assumption simplifies the implementation, but it is questionable whether the semantics of sentences can be fully captured in a vector space designed to represent the semantics of nouns.

Should the composed representations be distributional? I make a distinction between a distributed representation – which I take to mean simply vector- (or tensor-) based – and a distributional representation, which I take to mean a representation based on contextual information (as in the classic vector-based representations of word meanings [12, 9]). [1] take the intriguing position that all distributed representations are distributional, including those at the phrase and sentence level. Another alternative is to suppose that the word representations – especially those of nouns – are distributional, but the representations of larger phrases are distributed, without necessarily reflecting the distributional contexts of those phrases in some large, idealized corpus.

Can higher-order tensors be built in practice? Whilst there are machine learning techniques in place for learning higher-order tensors, given some suitable objective function, in practice the task of learning tensors for all word-category pairs in the lexicon is a formidable one. Dimensionality reduction techniques may help, but it is likely that the order of some of the tensors will need reducing in the grammar. For example, syntactic types such as

$((N/N)/(N/N))/((N/N)/(N/N))$ are not uncommon in the output of the CC CCG parser [4], but this would result in an 8-order tensor, with a huge number of parameters.

Do we need both operator and contextual semantics The meaning of a transitive verb, for example, in the framework is a 3rd-order tensor (a function). This is what I am calling operator semantics. But of course a (1st-order) vector can also be built for transitive verbs, in the standard way (which I am calling contextual semantics). We can also build (distributed) representations of the selection preferences of the verb. Are these separate from the operator semantics? Do we need all these representations? Another way to consider this question is whether the proposal of Erk and Pado [6] would benefit from the addition of operator semantics as provided by our compositional tensor-based framework. One area where this question arises is in relative clauses. Here, a verb phrase, which has operator semantics in the framework (represented by a matrix), needs to combine, via the relative pronoun, with the noun, which has contextual semantics (represented by a vector) [3]. Hence there appears to be a typemismatch here. Providing an additional representation for the verb phrase – either its contextual vector, or its selectional preferences – and allowing that to combine with the noun (eg through pointwise multiplication) may solve this problem.

Can logical operators be incorporated into the framework? This question relates to the more general question of whether traditional notions from formal semantics – which could also include quantification and inference – can be incorporated into a vector-space setting. My work is currently less focused on this question, but it is obviously important. A more general question is whether formal semantics is needed in addition to distributional semantics, or whether there is an all-encompassing framework.

References

- 1 M. Baroni, R. Bernardi, and R. Zamparelli. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies*, 2013.
- 2 M. Baroni and R. Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the EMNLP Conference*, pages 1183–1193, Boston, MA, 2010.
- 3 Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. The Frobenius anatomy of relative pronouns. In *Proceedings of the ACL Mathematics of Language workshop*, Sofia, Bulgaria, 2013.
- 4 Stephen Clark and James R. Curran. Widecoverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.
- 5 Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis 36: A Festschrift for Joachim (Jim) Lambek*, 2010.
- 6 Katrin Erk and Sebastian Pado. A structured vector space model for word meaning in context. In *Proceedings of the EMNLP Conference*, Hawaii, 2008.
- 7 Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, 2011.
- 8 Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. Concrete sentence spaces for compositional distributional models of meaning. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS-11)*, pages 125–134, Oxford, UK, 2011.
- 9 Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994.

- 10 Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In Proceedings of the International Conference on Computational Linguistics, Mumbai, India, 2012.
- 11 Joachim Lambek. From Word to Sentence. A Computational Algebraic Approach to Grammar. Polimetrica, 2008.
- 12 Hinrich Schütze. Automatic word sense discrimination. Computational Linguistics, 24(1):97–123, 1998.
- 13 Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1201–1211, Jeju, Korea, 2012.
- 14 Mark Steedman. The Syntactic Process. The MIT Press, Cambridge, MA, 2000.

5.4 Position statement

Ann Copestake (University of Cambridge, GB)

License  Creative Commons BY 3.0 Unported license
© Ann Copestake

The following brief (and unavoidably rushed) notes are partly drawn from the ‘Lexicalised compositionality’ draft paper available from my web page (joint work with Aurélie Herbelot). See also my position paper in the distributional semantics workshop in IWCS 2013.

Theoretical perspective. Distributional semantics is best seen as belonging to usage-based accounts of language. While the philosophical tradition is in many ways difficult and not very helpful as a guide for computational linguists (later Wittgenstein etc), some more modern work seems to provide a better basis: I find Brandom’s approach in ‘Making it explicit’ particularly helpful. Human languages can be used to ‘do logic’, but it doesn’t follow that that’s all language semantics is about. The notion of an ‘Ideal distribution’ in our lexicalised compositionality paper is an attempt to show under what conditions there is a relationship between a distributional account and a model-theoretic account. There is no reason why a notion of an individual (linguistic and/or real world) can’t be combined with a distributional account. This seems essential for modelling quantification, and (probably) also some lexical semantic phenomena such as antonymy. The role of generalization (inheritance) in distributions seems a promising area of investigation: the difference in this regard from previous approaches to lexical semantics is very striking.


Compositional semantics and distributions. I believe it is better to base distributions on a lightweight model of compositional semantics, such as (D)MRS, than on syntax, since there are cases of syntax-semantics mismatches (expletive ‘it’ etc, etc) which compositional semantics is well-equipped to deal with. Similarly, I currently see no need for distributional semantics to redo compositional accounts of tense (e.g., the English auxiliary system). Lightweight models allow for different interpretations of e.g., adjective noun combination, which gives scope for distributional semantics. Distributional semantics is particularly good at semi-compositional situations (cf derivational morphology).

Corpora. If we want to take the idea of psycholinguistic plausibility seriously (and I see this as a major advantage of distributional approaches), then we should work with realistic corpora. Collecting corpora based on an individual’s language experience should be a priority. Failing good notions of situated discourse in corpora, it may make sense to work with corpora

that exemplify one particular language game (or small class of language games) such as Wikipedia.

5.5 Position statement

Katrin Erk (University of Texas – Austin, US)

License  Creative Commons BY 3.0 Unported license
© Katrin Erk

Own previous work: Graded representations for word meaning in context. The topic that first got me interested in distributional models was the problem of representing word senses. Manually annotating documents with dictionary-based word senses is notoriously difficult, and both cognitive linguists [27, 7, 13] and lexicographers [15, 14] have cast doubt on the existence of clear-cut sense boundaries. Distributional models can be used to represent word meaning in context without reference to dictionaries if we compute a separate vector for each occurrence. Then we just have different occurrences that are closer together or further apart in space, without the need to draw sense boundaries. We have proposed a number of models for computing such occurrence representations [10, 11, 20].

Own previous work: Combining logic and distributional semantics. Distributional models have proved incredibly useful at the level of words and of short phrases. So what should be their role in sentence meaning representations? One possibility would be to use compositional distributional approaches to derive vectors for arbitrary sentences. But my hunch is that these vectors will become more and more noisy as phrase length and phrase complexity rise, where by noisy I mean that quite different phrases would receive similar vectors. (Table 1 in [24] seems to hint at something like this.) This is also my answer to Q3, the question about current limitations of compositional distributional semantics. It is my impression that the largest limitation of compositional distributional semantics lies in phrase length and complexity. Instead of pursuing a compositional distributional approach to sentence meaning, we are representing sentence meaning through logical form and are adding distributional similarity information (at both the word and short phrase level) as weighted inference rules [12, 4]. We use Markov Logic Networks [23] to do probabilistic inference on the resulting weighted clause set.

Q1: What can distributional semantics do that conventional semantic models cannot?

To me, the central reason to adopt distributional semantics is gradience, for example the ability to model degrees of similarity in word meanings [10, 26, 8, 28, 22]. A related strength of distributional models is their ability to describe relations between words through an open-ended list of possible phrases rather than through a fixed list of possible relations. Lapata and Lascarides [16] use this idea for logical metonymy. In a corpus-based model, the most likely interpretations for “begin song” that they derive are “sing”, “rehearse”, “write”, “hum”, “play”. Butnariu and Veale [5] use a similar idea for interpreting noun-noun compounds. Another big advantage of distributional models is coverage. They can extract usage-based information (representing a mixture of semantic and pragmatic phenomena) automatically for large numbers of lexical items.

Problems of distributional data. That said, I would like to list some problems of distributional semantics. They have all been discussed before, but I think they still need to be mentioned. The first problem is that we only have a single signal, co-occurrence, caused by a mixture of phenomena. A verb-noun cooccurrence can indicate a selectional preference

or an idiom [2]. Distributional similarity links near-synonyms (cup-mug) and pragmatically connected words (cup-milk) [21]. The second problem is lack of reference: We only have co-occurrence between words, no link between words and objects in the world [3]. The third problem is a reporting bias. Newspaper text tend to report on man bites dog, but not dog bites man [25].

Q5: Distributional semantics and model theory. I think that the time has come to revisit deep semantic analysis in computational linguistics. I have noticed more and more papers that talk about the need to address phenomena that deep semantic analysis is good at, like negation, modals, and implicatives. This happens in particular in textual entailment [1, 19, 18], but also in sentiment analysis [6]. And while it has been stated repeatedly that logic is “brittle”, I think it is not the logic that is brittle, but the inference mechanism and the background information available to the system. One way to address this is to use probabilistic inference, and to add distributional information.

There are currently two main approaches to combining distributional semantics and model theory. We transform distributional similarity to weighted distributional inference rules, and use probabilistic inference. Lewis and Steedman [17], on the other hand, use clustering on distributional data to infer word senses, but use standard first-order inference on the resulting logical forms. The main difference between the two approaches lies in the role of gradience. Lewis and Steedman view weights and probabilities as a problem to be avoided. We believe that the uncertainty inherent in both language processing and world knowledge should be front and center in the inference we do. (Though it is true that probabilistic inference is currently slow and memory-intensive.)

But in both current approaches to integrating distributional information into model-theoretic semantics, one important question is still open: What is the denotation of distributional representations? I have proposed interpreting distributional representations over conceptual structures [9], but that cannot be quite right: Given that distributional data is collected from texts of many speakers, it is not clear whose concepts these are supposed to be.

Here is a new proposal. It seems reasonable to say that purely linguistic data can change our beliefs about the world – that is what language does. But does this also hold for distributional information? After all, distributional data just counts sentential contexts in which words have been observed [3], and it is not clear how we could derive truth conditions from a distributional vector [29]. But distributional information can do something less: It can provide tentative, uncertain information about similarities between different predicates that have been mentioned in the text. And this, I think, suffices to reduce our uncertainty about which world we are in. We can describe this in a probabilistic semantics setting. We have a prior probability distribution over worlds, which can be updated (in a standard fashion, by Bayes’ rule) using uncertain distributional information.

References


- 1 Roy Bar-Haim, Ido Dagan, Iddo Grental, and Eyal Shnarch. Semantic inference at the lexical-syntactic level. In Proceedings of AAAI, Vancouver, Canada, 2007.
- 2 Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.
- 3 Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. Frege in space: A program for compositional distributional semantics. To appear.
- 4 Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. Montague meets Markov: Deep semantics with probabilistic logical form. In Proceedings of *SEM, Atlanta, Georgia, USA, 2013.

- 5 Cristina Butnariu and Tony Veale. A concept-centered approach to nouncompound interpretation. In Proceedings of COLING, 2008.
- 6 Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In Proceedings of EMNLP, Honolulu, Hawaii, 2008.
- 7 D. A. Cruse. Polysemy and related phenomena from a cognitive linguistic viewpoint. In P. Saint-Dizier and E. Viegas, editors, *Computational Lexical Semantics*, pages 33–49. Cambridge University Press, 1995.
- 8 Georgiana Dinu and Mirella Lapata. Measuring distributional similarity in context. In Proceedings of EMNLP, Cambridge, MA, 2010.
- 9 Katrin Erk. Towards a semantics for distributional representations. In Proceedings of IWCS, Potsdam, Germany, 2013.
- 10 Katrin Erk and Sebastian Padó. A structured vector space model for word meaning in context. In Proceedings of EMNLP, Honolulu, HI, 2008.
- 11 Katrin Erk and Sebastian Padó. Exemplar-based models for word meaning in context. In Proceedings of ACL, Uppsala, Sweden, 2010.
- 12 Dan Garrette, Katrin Erk, and Raymond Mooney. Integrating logical representations with probabilistic information using Markov logic. In Proceedings of IWCS, Oxford, UK, 2011.
- 13 Dirk Geeraerts. Vagueness’s puzzles, polysemy’s vagaries. *Cognitive linguistics*, 4:223–272, March 1993.
- 14 Patrick Hanks. Do word meanings exist? *Computers and the Humanities*, 34 (1-2):205–215(11), 2000.
- 15 Adam Kilgarriff. I don’t believe in word senses. *Computers and the Humanities*, 31(2):91–113, 1997.
- 16 Mirella Lapata and Alex Lascarides. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315, 2003.
- 17 Mike Lewis and Mark Steedman. Combined distributional and logical semantics. *TACL*, 1:179–192, 2013.
- 18 Amnon Lotan, Asher Stern, and Ido Dagan. Truthteller: Annotating predicate truth. In Proceedings of NAACL, 2013.
- 19 Bill MacCartney and Christopher D Manning. An extended model of natural logic. In Proceedings of IWCS, 2009.
- 20 Taesun Moon and Katrin Erk. An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology special issue on paraphrasing*, 2013.
- 21 Gregory L. Murphy. *The Big Book of Concepts*. MIT Press, 2002.
- 22 Diarmuid Ó Séaghdha and Anna Korhonen. Probabilistic models of similarity in syntactic context. In Proceedings of EMNLP, Edinburgh, UK, 2011.
- 23 Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.
- 24 Richard Socher, Eric H. Huang, Jeffrey Pennin, Andrew Y. Ng, and Christopher D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In NIPS. 2011.
- 25 Mohammad S Sorower, Janardhan R Doppa, Walker Orr, Prasad Tadepalli, Thomas G Dietterich, and Xiaoli Z Fern. Inverting grice’s maxims to learn rules from natural language extractions. In NIPS, 2011.
- 26 Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. Contextualizing semantic representations using syntactically enriched vector models. In Proceedings of ACL, Uppsala, Sweden, 2010.
- 27 David H. Tuggy. Ambiguity, polysemy and vagueness. *Cognitive linguistics*, 4 (2):273–290, 1993.

- 28 Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. Latent vector weighting for word meaning in context. In Proceedings of EMNLP, Edinburgh, Scotland, UK., 2011.
- 29 Jan van Eijck and Shalom Lappin. Probabilistic semantics for natural language. Draft published on the NASSLLI 2012 webpage, 2011.

5.6 Position statement

Stefan Evert (Universität Erlangen-Nürnberg, DE)

License  Creative Commons BY 3.0 Unported license
© Stefan Evert

At the current time, I am mainly interested in certain mathematical and practical issues of traditional distributional semantic models (DSMs) – which compile representations for single linguistic units (usually words) – and in practical applications that require broad – coverage semantics.

Research questions:

- Impact of DSM type (term–context vs. term–term) and model parameters (span–based vs. dependency–based co–occurrence, frequency weighting, normalization, dimensionality reduction, distance measure,) on the semantic representation.
- To what extent can DSMs be optimized for a particular task? Is there a single representation that captures general word meaning and works well for a broad range of tasks?
- Dimensionality reduction
 - Is it useful?
 - What are the differences between available methods (PCA, SVD, randomized SVD, RI, NMF, LDA,)?
 - How many latent dimensions should be used?
 - Are the reduced representations compatible with simple approaches to compositionality and polysemy?
- Ambiguity and polysemy: DSM vectors represent a weighted average over all senses of the corresponding word.
 - How can different meanings be identified and separated?
 - How can the context–dependent meaning of a word be computed?

If such “traditional” DSMs are applied to larger units (word pairs, phrases or sentences) these are either treated as opaque multiword units or a simplistic approximation to the compositional meaning is used (addition = Schütze’s bag–of–words approach, pointwise multiplication, etc.). I am interested in extensions of these approaches that capture some facets of compositionality while maintaining the simple structure and broad coverage of the simple models. Research questions include:

- Should compositional DSMs aim for a distributional representation (i.e. an approximation of the DSM vector of the phrase as an opaque unit) or merely a distributed representation (i.e. any vector representation; a typical example are convolution–based models such as HRR and BEAGLE).
- What are desirable properties for the DSM distances between compositional units and between such units and individual words?
 - Which of these properties are satisfied by addition, pointwise multiplication and other simple composition operations?

- Can simple compositional representations of large units (e.g. a bag-of-words model for sentences) be seen as a case of ambiguity/polysemy (e.g. a weighted mixture of topics present in a sentence)?

My research interests thus connect to, and overlap with themes 1 (polysemy and vagueness) and 3 (compositionality) of the seminar.

5.7 Statements on Distributional Semantics

Sebastian Löhnner (Heinrich-Heine-Universität Düsseldorf, DE)

License  Creative Commons BY 3.0 Unported license
© Sebastian Löhnner

General judgment. I consider Distributional Semantics a very important methodological achievement for the working semanticist as it offers tools and means for retrieving evidence and data for semantic analysis that were hitherto not available. It appears particularly promising for lexical decomposition, as the co-occurrence of lexical items provides evidence for their combinatory propensities and these, in turn, and to some degree, may provide access to their semantic content. Combinatory propensities are also relevant for a theory of composition, in particular by providing more data that will enable us to broaden the scope of theoretical analysis; but I think the qualitative rule-based approaches to a theory of syntactic, and concomitant semantic, composition are, and will keep being, superior to whatever can be gained by merely statistical methods.

Skepticism. I am very skeptical as to the potential of the DS approach for bringing us closer to an understanding of the cognition of language. Obviously, the brain does not work with this kind of software (the relevance of statistical weights for cognitive learning notwithstanding). As a theoretical semanticist, I am ultimately aiming at an understanding of the cognitive level of language. At present, it appears, there are encouraging developments in the cognitive sciences that open ways for developing cognitive semantic theories of decomposition and composition that take us crucially beyond the first cognitive approaches from the late 20th century (e.g. prototype theory) and the (indirect) insights into semantic cognition that were gained in formal semantics by logical analysis. Trying to model (?) or just do semantic composition with statistical methods might work one day to a certain degree of efficiency (similar to parsing, or machine translation by statistical methods) – but it will not bring us further to an understanding of semantic cognition.

5.8 Putting together the pieces

Louise McNally (UPF – Barcelona, ES)

License  Creative Commons BY 3.0 Unported license
© Louise McNally

I got into linguistics through cognitive science. I was interested in all kinds of big questions about language and the mind, but the more I thought about these questions the more I realized that I understood almost nothing about language, and I didn't think it made much sense to try to answer them without a better idea of how language works. So I decided to

go to grad school in linguistics, hoping that would give me the background I would need to continue in cognitive science. Little did I realize how complex language would turn out to be.

My main goal as a linguist is to understand how lexical meaning (if we can distinguish such a thing) is integrated with general conceptual knowledge, on the one hand, and information coming from reference, on the other, when we interpret phrases and produce them for others. I have pursued this goal by carrying out detailed studies of linguistic phenomena that cannot be understood without some theory of how this integration works. The two most relevant empirical areas I have worked on are modification – the construction of complex descriptive contents – and the many manifestations of the type (conceptual)/token (referential) distinction in language. I have developed most of my work using the tools associated with what I'll loosely refer to as the formal semantics community.

Though it is unquestionable that this community's focus on understanding the connection between words and the world, and on the development of the corresponding tools to do so, has led to huge advances in the theory of meaning, the limitations of this focus have long been known to everyone. One of the things I have found most unsatisfying about formal semantics (though not unsatisfying enough to give up on the whole enterprise) is that these limitations have mostly been quietly ignored. One extremely negative effect of this is that formal semantic research has arguably not had the impact on cognitive science that it could have, and my impression from the last IWCS is that its early contributions to computational linguistics are running a certain risk of being lost. On the bright side, various lines of formal and computational research have addressed several of these limitations. The big pending task is to bring these lines together in a systematic way. This seminar looks like a good opportunity to make some progress. In the rest of this statement, I briefly mention some of the issues that I think should be placed on the table for discussion.

Conceptual vs. referential aspects of meaning: Perhaps the most serious and long-standing problem in study of meaning is the division between approaches and, correspondingly, communities of researchers, according to whether the conceptual or referential dimension of meaning (to say nothing of the social) is the primary focus of interest. Discourse Representation Theory (in a particularly clear way in comparison to other dynamic logics) was perhaps the first systematic attempt to distinguish formally between reference and the descriptive conditions that the referents in our discourse models must satisfy. Though these conditions have generally been modeled as grounded in the world, I do not see any reason in principle why they could not be associated with conceptual contents. I think there is great potential here that is only beginning to be explored (see for example recent work by Erk and colleagues, Kamp, and myself with Baroni and Boleda in this direction).

Though a differentiated treatment of reference vs. descriptive content conditions is not the focus of the richly typed systems that e.g. Pustejovsky and Asher have used to develop more sophisticated analyses of the composition of lexical meanings, these approaches are certainly compatible with such a treatment. In contrast, it is far less obvious how to capture such a distinction in distributional models of meaning, or whether we should even try to do so. The implications of this characteristic of distributional models are profound; I'll come back to them briefly below.

The representation and composition of lexical meanings: The development of rich systems of types and type composition operations has made it possible to express important generalizations concerning the ways lexical meanings are typically modulated in the context of other lexical items – for instance, we can easily capture the role of part-whole relations or the function an entity typically has in accounting for patterns of metonymy. Distributional

semantic models improve in some ways on these systems, but at least in their present state arguably lose ground on others. Since distributional representations reflect not only lexical entailments but also a lot of other contextual information associated with an expression, they are very suggestive as a means of approximating richer conceptual representations, and their behavior under composition offers the hope of analyses of polysemy resolution and phenomena such as metaphor that are more general and finer-grained than those afforded by symbolic systems. However, it is less clear how the composition of distributional representations can be modulated to reflect the salience of the sorts of relations embodied in e.g. qualia structures and that arguably have psychological reality independently of the sheer frequency of their occurrence.

Function words, content words, and the syntax/semantics/discourse interface: We commonly distinguish between so-called content words and function words, the latter serving, for example, to help manage referential relations (e.g. the vs. a) or to guide the integration of new information into the previous discourse (e.g. too). Syntactic and prosodic structures (e.g. left dislocation or a particular pitch accent) also provide crucial, conventionalized information. When one starts using distributional models, this distinction between content words and function words cannot be obviated in the way that it has been relatively easy to obviate in formal semantic theories. This fact raises a number of challenges. If the conventional contributions of function words cannot be represented in distributional models in the same way as those of content words, how should they be represented? Some expressions, such as prepositions, manifest properties both of content words and function words; how do we analyze these? These are questions that the main natural language processing applications using distributional models have been able to ignore so far, but continuing to ignore them will probably impose an upper limit on the quality of NLP applications. We might also aspire to having computational models that help us understand human language processing. For example, it would be interesting to see what our models predict for patterns of semantic change, particularly the well-attested phenomenon of semantic bleaching (the loss of descriptive content associated with an expression over time, often substituted by a strictly referential function). Without an analysis of function words, cognitively realistic language models are not possible.

The analysis of meaning and psychological reality: A model of meaning that is cognitively realistic should be compatible with what we know about how language is acquired, how it is processed in real time, how it connects to the rest of our cognitive systems, what happens in pathological situations, how language changes over time, and how we come to associate new or revised concepts with bits of language. Here are just a few disconnected thoughts about language and cognition that have come to my mind as I have worked with distributional semantic models: 1) One appealing thing about distributional semantic models is that they might allow us to avoid making some difficult decisions about the linguistic meaning/world knowledge boundary. 2) Working with distributional models naturally leads one to think of language as decompositional rather than compositional. This change in perspective has all kinds of interesting implications. 3) Distributional models rely on quantities of data that do not correspond to realistic assumptions about exposure to language during development (a point made in a recent paper by Copestake and Herbelot). If these were to map onto cognitively plausible models, clearly more than just raw statistics would have to be influencing their functioning. But what are these other influences, and how do they work? 4) If I have been critical of formal semantics for its almost exclusive emphasis on referential aspects of meaning, I have developed an entirely new appreciation for these – particularly the special

informativity of the association of words with visual or auditory stimuli – when finding them absent in distributional models.

It should be clear that there'll be no shortage of things to talk about during the week

5.9 Incrementality in Compositional Distributional Semantics

Alessandra Zarcone (Universität Stuttgart, DE), Sebastian Padó (Universität Stuttgart, DE)

License © Creative Commons BY 3.0 Unported license
© Alessandra Zarcone, Sebastian Padó

Overall Interest. Our interest is at the crossroads of computational linguistics and psycholinguistics. We are interested in compositional distributional semantic models (CDSMs) that can both contribute towards NLP as well as account for (aspects of) human sentence comprehension. The following ideas come from a project proposal currently in preparation.

Focus and Desiderata. We feel that a promising direction for CDSMs is provided by tensor-based models [3, 6, 2, 1]: each word is associated with one or more types describing its semantic arity and determining the shape of its distributional semantic representation. For example, nouns can be mapped onto vectors, adjectives onto matrices, and verbs and other function words on higher-order tensors. This enables the formulation of syntax-semantics interfaces that look similar to traditional ones but operate on distributional representations, with the potential to link the benefits of distributional representations with the power of compositionality.

Current models though have some limitations: (a) they are constituency-based rather than based on dependencies (dependency grammar is well-established for many languages, in particular with free word order); (b) they are not incremental, that is, they do not construct semantics in a left-to-right manner (whereas human language processing is to a large degree incremental); (c) they do not incorporate a notion of plausibility for (partial) analyses based on expectations at the level of individual composition operations.

We aim at developing a tensor-based CDSM overcoming such limitations. The steps that we foresee are as follows:

1. **A dependency-based distributional syntax-semantics interface.** This step does not yet take incrementality into account. This simplification allows us to binarize the dependency trees of a large German dependency-parsed corpus into composition trees that express the order of semantic composition (see Figure 1); then we will infer the algebraic type(s) of each lemma (nouns as well as sentences are represented as vectors in \mathbb{R}^n , while other parts of speech will generally be assigned higher-order types); we will finally learn a large lexicon that associates lemmas with distributional representations of appropriate algebraic types, via multi-step regression learning [5]. The free choice of binarization schemes allows us to choose one that leads to well-behaved types both in terms of lexical ambiguity and type complexity.
2. **CDSM-based semantic plausibility scores.** Previous definitions of semantic plausibility for predicate argument combinations [4, 7] were limited to predicate-argument combination. They were based on vector similarity, comparing the expectations about arguments against actual arguments. We assume that these approaches can be generalized to our tensor-based CDSMs, with comes with the potential of generalizing semantic plausibility to a wider range of linguistic constructions. Our central assumption is that sentence plausibility decomposes along the edges of the sentence's composition tree.

3. **Incrementality.** The next step is to adapt the first two models to an incremental setup. We will use an incremental dependency parser and assign a semantic representation to each prefix of the sentence that receives a connected analysis from the parser. The central challenge is that in contrast to step 1, we cannot freely choose the order of compositions; instead, the composition tree must be left-branching. This will introduce a considerably higher degree of lexical ambiguity that has to be managed. Subsequently, we want to define incremental plausibility scores by adapting our plausibility measures to the incremental nature of the analysis, taking advantage of the definition of the plausibility measure in terms of individual edges.
4. **Evaluation.** Given a sentence, the model will be able to return plausibility scores for upcoming words at each time during processing. The psycholinguistic evaluation of these scores will take place through word-by-word prediction of reading times. We will perform a broad-scale prediction of reading times on a corpus of German sentences, hoping to show that our plausibility model can account for a larger amount of variance than other models. The NLP-oriented evaluation will be applied to a state-of-the-art beam search-based dependency parser to re-rank dependency parsing hypotheses, both at the level of complete sentences and during parsing.

References

- 1 Baroni, M., Bernardi, R., Zamparelli, R., 2013. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies* (To appear).
- 2 Baroni, M., Zamparelli, R., 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In: *Proceedings of EMNLP*. Cambridge, MA, pp. 1183–1193.
- 3 Clark, S., Pulman, S., 2007. Combining symbolic and distributional models of meaning. In: *Proceedings of the AAAI Spring Symposium on Quantum Interaction*. Stanford, CA, pp. 52–55.
- 4 Erk, K., Padó, S., Padó, U., 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics* 36 (4), 723–763.
- 5 Grefenstette, E., Dinu, G., Zhang, Y., Sadrzadeh, M., Baroni, M., 2013. Multi-step regression learning for compositional distributional semantics. In: *Proceedings of IWCS*. Potsdam, Germany, pp. 131–142.
- 6 Grefenstette, E., Sadrzadeh, M., Clark, S., Coecke, B., Pulman, S., 2011. Concrete sentence spaces for compositional distributional models of meaning. *Proceedings of IWCS*, 126–134.
- 7 Lenci, A., 2011. Composing and updating verb argument expectations: A distributional semantic model. In: *Proceedings of CMCL*. Portland, Oregon, pp. 58–66.

5.10 Acquiring conceptual knowledge for semantic interpretation

Massimo Poesio (University of Essex, GB)

License  Creative Commons BY 3.0 Unported license
© Massimo Poesio

Summary of my research in the area

Initial Motivations. My first research area was knowledge representation, but I soon stopped working on that to focus on (computational) semantics and pragmatics. I started to look at theories of conceptual knowledge acquisition after a few years working on the use of lexical semantics for anaphora resolution (in particular to interpret bridging references)

and being dissatisfied with the results obtained using WordNet or other lexical resources. Developing theories for acquiring such knowledge automatically seemed both scientifically more interesting and something that was needed to overcome the commonsense knowledge bottleneck found in other areas of AI as well. Over the years however I started getting interested in the acquisition of commonsense knowledge per se and in commonsense knowledge more in general (in particular in cognitive evidence about the way commonsense knowledge is organized).

Acquiring lexical knowledge for resolving bridging references. Our initial efforts were motivated by the work on bridging references carried out with Renata Vieira [18, 17, 24], that had showed that WordNet offered only limited support for this type of interpretation process. Our intuition was that semantic space models like HAL [12] would be quite good at capturing bridging references based on synonymy; our results confirmed this (Poesio et al., 1998). For other types of bridging references we started looking at the unsupervised methods for extracting semantic relations proposed by Hearst [11]. Our work on meronymy indicated that a reasonable precision and recall could be achieved provided that (a) very large corpora were used (Web size), and (b) semantic space models were combined with salience information [20].

Acquisition informed by research on lexical semantics and knowledge representation. As a result of the work on resolving associative references, we started working on theories of commonsense acquisition that incorporated insights from work on lexical semantics (in particular the work by Pustejovsky [23] and formal ontology (in particular the work by Guarino and his lab, [10]). In collaboration with my PhD students Abdulrahman Almuahareb and Eduard Barbu, and then with Marco Baroni, we developed acquisition models that built conceptual relations based on semantic relations extracted from text. With Abdulrahman, we used first unsupervised methods to extract from text attributes, and then supervised methods to build vectors based on qualia theory and Guarino's theory of attributes ([2, 4, 3, 15]). This model also attempted to discriminate between wordsenses ([5]). (A summary of this research can be found in ([16]); a more extensive description in ([1]). With Eduard Barbu, we developed improved models to extract semantic relation-based conceptual descriptions ([21]) and then started using Wikipedia as a corpus ([8]). Finally with Marco Baroni we studied methods using semi-supervised techniques for relation extraction ([9]).

Combining brain evidence with corpus evidence. In recent years, the focus of our research in the area of commonsense knowledge has shifted to using machine learning techniques to study the representation of conceptual knowledge in the brain ([14, 6]) and then using distributional models to predict the activation patterns of concepts ([13, 7]).

Where we stand

At least from a scientific point of view, the only solution to the commonsense bottleneck is to develop models for the acquisition of commonsense knowledge. But the fact remains that although work on using semantic space models for anaphora resolution has continued, the results are still unsatisfactory ([22]). In fact, I would make a more general claim: that so far distributional models have proved successful at tasks that only require collocational or lexical knowledge (checking text coherence, identifying synonymy, etc) but haven't yet been successfully employed in semantic tasks that do require commonsense knowledge. To me the question of why this is the case ought to be one of the central issues for the workshop.

References

- 1 Almuhareb, A. (2006). Attributes in Lexical Acquisition. Ph.D. thesis, University of Essex, Department of Computer Science.
- 2 Almuhareb, A. and Poesio, M. (2004). Attribute- and value-based clustering of concepts. In Proc. of EMNLP , pages 158–165, Barcelona.
- 3 Almuhareb, A. and Poesio, M. (2005a). Concept learning and categorization from the web. In Proc. of Annual Meeting of Cognitive Science Society, pages 103–108, Stresa (Italy).
- 4 Almuhareb, A. and Poesio, M. (2005b). Finding attributes in the web using a parser. In Proc. of the Corpus Linguistics Conference, Birmingham.
- 5 Almuhareb, A. and Poesio, M. (2006). MSDA: A word sense discrimination algorithm. In Proc. of ECAI , Riva del Garda.
- 6 Anderson, A., Murphy, B., and Poesio, M. (2013a). Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience*.
- 7 Anderson, A., Bruni, E., Bordignon, U., Poesio, M., and Baroni, M. (2013b). Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In Proc. of EMNLP .
- 8 Barbu, E. and Poesio, M. (2009). Unsupervised knowledge extraction of taxonomies of concepts from wikipedia. In Proc. RANLP , pages 28–32, Borovets.
- 9 Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel: A distributional semantic model based on property and types. *Cognitive Science*, 34(2), 222–254.
- 10 Guarino, N. (1992). Concepts, attributes and arbitrary relations. *Data and Knowledge Engineering*, 8, 249–261.
- 11 Hearst, M. A. (1998). Automated discovery of wordnet relations. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press.
- 12 Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In Proc. of the 17th Annual Conference of the Cognitive Science Society, pages 660–665.
- 13 Murphy, B., Baroni, M., and Poesio, M. (2009). Eeg responds to conceptual stimuli and corpus semantics. In Proc. of EMNLP , pages 619–627, Singapore.
- 14 Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., and Lakany, H. (2011). Eeg decoding of semantic category reveals distributed representations for single concepts. *Brain and Language*, 117(1), 12–22.
- 15 Poesio, M. and Almuhareb, A. (2005). Identifying concept attributes using a classifier. In T. Baldwin and A. Villavicencio, editors, *Proc. of the ACL Workshop on Deep Lexical Semantics*, pages 18–27, Ann Arbor, Michigan.
- 16 Poesio, M. and Almuhareb, A. (2008). Extracting concept descriptions from the web: The importance of attributes and values. In P. Buitelaar and P. Cimiano, editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 28–44. IOS Press, The Netherlands.
- 17 Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2), 183–216. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.
- 18 Poesio, M., Vieira, R., and Teufel, S. (1997). Resolving bridging references in unrestricted text. In R. Mitkov, editor, *Proc. of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, pages 1–6, Madrid. Also available as HCRC Research Paper HCRC/RP-87, University of Edinburgh.
- 19 Poesio, M., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In Proc. of the AAAI Spring Symposium on Learning for Discourse, pages 82–89, Stanford, CA. AAAI.

- 20 Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to solve bridging references. In Proc. of ACL, pages 143–150, Barcelona.
- 21 Poesio, M., Barbu, E., Giuliano, C., and Romano, L. (2008). Supervised relation extraction for ontology learning from text based on a cognitively plausible model of relations. In Proc. of ECAI Workshop on Ontology Learning and Population, pages 1–5, Patras.
- 22 Ponzetto, S., Versley, Y., and Poesio, M. (2014). Using lexical and commonsense knowledge for anaphora resolution. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 2. Springer.
- 23 Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.
- 24 Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4), 539–593.

5.11 Research overview

Tim Van de Cruys (Paul Sabatier University – Toulouse, FR)

License  Creative Commons BY 3.0 Unported license
© Tim Van de Cruys

My research has explored different algorithms for the modeling of semantic phenomena within the framework of distributional semantics, with a focus on factorization algorithms and tensor algebra. Below is an overview of the research that is most connected to the seminar themes.

Word meaning in context

An important part of my research focuses on factorization, and its application to language. The use of large text collections brings about a large number of contexts in which a word occurs. By using a factorization algorithm, the abundance of individual contexts can be automatically reduced to a limited number of significant dimensions. Characteristic for these dimensions is that they contain latent semantics: the value of a word on a particular dimension indicates the score of the word for a particular semantic field. This is particularly useful for dealing with polysemous words. By determining the latent semantic fingerprint for a particular context, it is possible to weight the word vector accordingly, thus computing the specific meaning of a word in a particular context [3].

Modeling compositionality

Most research in distributional semantics uses matrices as its main mathematical tool, which is useful for the modeling of individual words. If, on the other hand, one wants to model interactions between several words, multi-way co-occurrences need to be taken into account. Multi-way co-occurrences need to be represented within a tensor framework, which is the generalization of a matrix for more than two modes. Tensors may contain any number of n modes. This allows for the treatment of more complex syntactic constructions, such as the combination of a verb and its different complements, or the different modifiers that a verb appears with. Tensors can equally be combined with factorization algorithms, and they can subsequently be used for the modeling of compositional phenomena [4]. The key idea is that compositionality is modeled as a multi-way interaction between latent factors, which are automatically constructed from corpus data. The model can be readily applied to transitive phrases, for which it gives good results.

Position statement

The opposition that exists between distributional approaches and formal approaches to semantics is very much related to the opposition between connectionist and symbolic models within the field of cognitive science; in a way, they provide two different perspectives on the same data. While formal semantics provides a framework for the explicit, symbolic modeling of semantic phenomena, distributional semantics provides a way to deal with those phenomena in a more implicit way, based on simple co-occurrence data. Formal semantics is typically characterized as very successful with respect to the semantic modeling of functional elements and quantification (elements typically not tackled by the distributional approach), while distributional semantics is lauded for its ability to cope with lexical semantics (which is less extensively developed within the formal semantic framework). Yet, nothing seems to prevent the formal or the distributional approach to model the kind of semantic phenomena that are typically more successfully modeled within the other approach. Distributional models are able to get at the generalizations that are typically handled within a formal semantic framework, while nothing prevents the formal semantic approach from explicitly modeling lexical semantics (though the manual modeling of the lexical semantics of individual content words would quickly become a tedious and prohibitively expensive tasks).


Does this mean that one approach should take precedence over the other? Most likely, the best results are obtained by taking a hybrid approach. The ability of the distributional approach to induce generalizations automatically from corpus data is a huge advantage over the manual approach of formal semantics, while the latter provides machinery for inference and entailment which are still problematic within a distributional framework. What exactly should be the role of each framework is a very interesting topic of discussion, that will probably be amply touched upon during the seminar.

References

- 1 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In ICLR 2013, 2013.
- 2 Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1201–1211, Jeju Island, Korea, July 2012. Association for Computational Linguistics.
- 3 Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. Latent vector weighting for word meaning in context. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1012–1022, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics.
- 4 Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. A tensor-based factorization model of semantic compositionality. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1142–1151, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

5.12 Semantics, Communication, and Probability

Jan van Eijck (CWI – Amsterdam, NL)

License  Creative Commons BY 3.0 Unported license
© Jan van Eijck

Logic, Linguistics, and Intelligent Interaction. In logic the distinctions between language, interpretation and communication are quite clear, in natural language understanding less so. But maybe natural language semantics has something to learn from new directions in logic. A first lesson was taught by Richard Montague long ago, but there are some new things to learn now.

Logic, narrowly conceived, is the design and use of formal languages for thought, the study of their strengths and limitations (the trade-off between expressive power and complexity), and the use of these tools in clarifying what goes on in the mind of a mathematician, or in the memory of a computer carrying out a program. Montague’s lesson for NL understanding was that NL can be studied with the methods from logic.

Broadly conceived, logic is the study of intelligent interaction, rational adjustment on the basis of evidence, transformation of our conceptualisations of the world on the basis of received information. See [1] for an overview, and for a logic textbook emphasizing this broader perspective.

Intelligent interaction is also a central topic in natural language understanding, for intelligent interaction is what natural language is for. A desire to explain why human beings are so good at communication using language is one of the reasons for being interested in linguistics.

Formal Models of Communication. In dynamic epistemic logic (see [3] or [2]), a state of affairs is a multi-agent Kripke model, and acts of communication are operations on states of affairs. The Kripke model represents what the agents know (or believe). If an agent a is uncertain about the truth of p , this is represented by an inability of a to distinguish p -worlds from non p -worlds. The act of communication represents how this knowledge (or this belief) gets changed by information exchange. A paradigm example is public announcement. A public announcement of a true fact p has the following effect on a Kripke model. All non p -worlds get removed from the model, and the accessibility relations representing the knowledge or belief of the agents get restricted to the new class of worlds. The result is that p becomes common knowledge among all agents. But many other kinds of communication can be modelled: messages to specific individuals, messages to all agents that happen to pay attention, and so on.

Knowledge, Belief, and Probability. In epistemic/doxastic logic (the logic of knowledge and belief), there is also a new trend, where knowledge and belief are linked to probability theory. Theories of subjective probability [6] agree well with Kripke model representations of knowledge and belief. To turn a Kripke model into a probabilistic model, all one has to do is to add, for each agent, a probability distribution over the set of all worlds to the model [5]. Knowledge of a cannot be linked to certainty: assigning probability 1 to a statement. Belief can be linked to assigning probability $> 1/2$ to a statement. This way, it is possible to explain certain properties of belief that are hard to cope with without bringing in probabilities.

Connection with Natural Language Semantics. Probabilistic semantics for natural language would link language (content words) to the world in a loose way (looser than the traditional truth-functional way), in the perspective of an agent (here is where subjective probabilities of the “knowing subject” come in). Example: vague or uncertain attribution.

“Bonfire is black”. In a probabilistic Kripke model M , in a world w for an agent a , this gets a probability $P_{a,w}$. If the probability is 1, this means that a knows that Bonfire is black, and it follows that it is true that Bonfire is black. In a case where the statement is judged as less than certain by a , we can say that a believes that Bonfire is black. Now it does not follow that it is true that Bonfire is black. Program Work out a probabilistic multi-agent semantics for natural language along these lines. See [4] for a first sketch. Connect up with work on distributional semantics.

References

- 1 Benthem, J. v., van Ditmarsch, H., van Eijck, J., and Jaspars, J. Logic in Action. Internet, 2013. Electronic book, available from <http://www.logicinaction.org>.
- 2 Benthem, J. v., van Eijck, J., and Kooi, B. Logics of communication and change. *Information and Computation* 204, 11 (2006), 1620–1662.
- 3 Ditmarsch, H. v., van der Hoek, W., and Kooi, B. *Dynamic Epistemic Logic*, vol. 337 of Synthese Library. Springer, 2006.
- 4 Eijck, J. v., and Lappin, S. Probabilistic semantics for natural language. To appear in the LIRA Yearbook 2013.
- 5 Eijck, J. v., and Schwarzentruher, F. Epistemic probability logic simplified. Submitted to AAMAS 2014, Paris.
- 6 Jeffrey, R. *Subjective Probability – The Real Thing*. Cambridge University Press, 2004.

5.13 Position statement

Dominic Widdows (Microsoft Bing – Bellevue, US)

License  Creative Commons BY 3.0 Unported license
© Dominic Widdows

My interest in compositional semantics and distributional models began when working on the Stanford Infomap project in the early 2000s, and has continued ever since. After stints at MAYA Design, Google, and Bing, I’m now Director of Language Engineering at Serendipity, a startup with the goal of consumerizing analytics, in which the need for good models for compositional semantics is more pressing than ever!

A differential geometer by training, I had the good fortune to work in an area where tensor products, exterior algebra, linear spans and orthogonal complements are widely used, long before realising that these mathematical models and operations could also be applied to natural language. My early adventures in this space included the use of orthogonal complements for negation and linear sum for disjunction in distributional models built using Latent Semantic Analysis.

We released the software implementation of this work as part of the Infomap NLP package, which after a few years was superseded by the SemanticVectors package, which is freely available at <http://code.google.com/p/semanticvectors>. This in turn led to many collaborations, most notably with Trevor Cohen of the University of Texas Health Sciences Center in Houston. Together, we’ve used the package for literature based discovery, drug repurposing, and most recently, orthographic encoding.

The work on drug repurposing and orthographic encoding highlights two important points for the seminar:

- Distributional models can successful for purposes way beyond the pioneering cases in information retrieval and text classification. In the application to drug repurposing, for

example, they are used much more like a fast, robust, approximate theorem prover.

- These models depend on composition operators that are more varied than the simple vector sum. Their success is partly due to the ready availability of established algebraic methods including orthogonal projection, tensor algebra and matrix multiplication, circular convolution, and permutation.

When applied to vectors with complex or binary numbers as coordinates, these operations, their implementations, and experimental results sometimes differ markedly from those obtained with real numbers as coordinates. This points out a sometimes surprising gap in information retrieval and indeed machine learning: in these rapidly developing empirical fields, we tend to tacitly assume that real numbers are the canonical ground field. This is in marked contrast to physics, where complex numbers are ubiquitous, and logic, where binary numbers are the established starting point. One ongoing personal goal of mine is to encourage theoretical and practical researchers in computational semantics to experiment much more with complex and binary vectors as well as real vectors, in the hope that such investigations may prove as fruitful for information retrieval as they have been for physics and logic.

5.14 Norms and Exploitations in Text Meaning and Word Use

Patrick Hanks (University of Wolverhampton, GB)

License  Creative Commons BY 3.0 Unported license
© Patrick Hanks

It is a truism that meaning depends on context. Corpus evidence shows that normal contexts can be summarized and quantified, revealing the platforms of phraseological norms on the basis of which we communicate with one another (i.e. on the basis of which future meanings may be created). A contrasting but equally important discovery is the fact that the potential for creative exploitations of normal contexts by ordinary language users far exceeds anything that has been dreamed up in speculative linguistic theory. These contrasting aspects of words in use are analysed in [2].

Meanings can be seen as evanescent interpersonal cooperative events that take place between speaker and hearer (or, with displacement in time, between writer and reader). They are created by using and exploiting shared knowledge of conventional patterns of word use. As I said publicly for the first time at a Dagstuhl seminar twenty years ago, words in themselves don't have very much meaning—but they do have meaning potential. Different aspects of this potential are activated when words are put into context and used for some real communicative purpose.

“Many if not most meanings require the presence of more than one word for their normal realization.” – [3]

So we may conclude that human linguistic behaviour is indeed rulegoverned, but there is not just a single monolithic system of rules: instead, language use is governed by two interlinked systems: one set of rules governing normal, idiomatic uses of words and another set of rules governing how we exploit those norms creatively. I call this 'the double helix theory of language in use'. It has a profound effect on the ways in which words are distributed across texts. Thirty years of corpus analysis drives us to the conclusions 1) that human languages are a puzzling mixture of logic and analogy and 2) that the importance of analogy in making meanings has been consistently underrated.

Types of creative exploitation include (among others): using anomalous arguments to make novel meanings ellipsis for verbal economy in discourse metaphors, metonymy, and other figurative uses for stylistic effect and other purposes.

The Pattern Dictionary of English Verbs (PDEV; <http://deb.fi.muni.cz/pdev/>; publicly available work in progress) implements this principle by associating meanings with patterns rather than with words in isolation. In PDEV, a pattern consists of a verb and its valencies (otherwise known as ‘clause roles’ or ‘arguments’). Each argument is populated by an open-ended set of lexical items and phrases, which share, to some extent, a semantic value. This means that different senses of a verb can be distinguished according to the semantic values of its arguments. For example, ‘executing an order’ and ‘executing a plan’ go together; they are distinguished from ‘executing a criminal’. These are two different meanings of the same verb, activated by different collocates, even though, structurally, all three examples have identical syntax.

PDEV’s patterns are analogous to the constructions described in Construction Grammar (e.g. [1]). A difference is that PDEV is corpus-driven. Every English verb (and in due course, every predicator—including predicative adjectives) has been or will be analysed on the basis of corpus evidence. Analogous work is in progress in Spanish and Italian.

Each entry in PDEV has the following components:

- A set of syntagmatically distinct patterns (the phraseological ‘norms’)
- An ‘implicature’ (i.e. the meaning and context) for each pattern
- A set of corpus lines illustrating normal uses of each pattern
- Comparative frequencies of each pattern of use of each verb, showing which patterns are most frequent
- A smaller set of corpus lines illustrating creative exploitations, insofar as these are found in the analysed samples
- A shallow ontology of nouns and noun phrases

The CPA shallow ontology serves as a device for grouping together nouns and noun phrases that distinguish one meaning of a verb from another.

References

- 1 Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.
- 2 Hanks, Patrick. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.
- 3 Sinclair, John. 1998. The lexical item. In Edda Weigand (ed.), *Contrastive Lexical Semantics*. John Benjamins.

5.15 Position statement

Anna Rumshisky (University of Massachusetts – Lowell, US)

License  Creative Commons BY 3.0 Unported license
© Anna Rumshisky

Personal background

My predilection for distributional approaches stems from the early structuralist education I received as a student of theoretical linguistics in Russia. The analysis of minimal pairs of all kinds and of paradigmatic substitutions in syntagmatic contextual patterns in general was one of the first methodological tools taught to a linguist. Therefore as a computational linguist later in life I found myself aligned with corpus-driven distributional approaches to lexical

analysis. My dissertation work on computational lexical semantics for corpus pattern analysis developed quantitative methods for creating contextualized ad-hoc conceptual categories that I felt were needed the proper handling of selectional preferences.

Current position

I briefly summarize below my position on a couple of issues of interest.

Issue 1. Beyond intratextual distributional patterns. At the core of distributional semantics is the notion of concordance, or a set of contexts that the word appears in. However, as we all know, there is only so much you can gain by looking at the “company” which the word “keeps”. This is not how language is used by humans, and not how it is learned – the language is learned in context provided by the circumstance in which linguistic expressions are uttered. What we need to do is to generalize the notion concordance to include this referential intuition, modeling the circumstance as a set of referents and pragmatic factors of the utterance, including the accompanying actions, participants, participants’ intents, etc.

This is a hefty task, for how do you represent all of these different aspects of the “reality” in which something is uttered? There has been some work on linking linguistic expression to (1) visual information (computer vision) and (2) agents and actions (robotics). But so far, from the point of view of language, at least, it’s mostly been limited to “toy” systems ([1], quite a few of the papers in the recent workshops, cf. [2, 3, 4]).

With respect to modeling context of the utterance, or at least representing it well enough to be able to both record it and do something useful with such recordings, we are at the stage where the distributional analysis of text was back in the 50s. I don’t have a solution for how to usefully model and represent the varied aspects of context, but I do think that this is the direction we need to go, and that we desperately need to take it beyond the toy system stage.


Issue 2. Representing compositionality. One of the issues for DS is the scope. What can we actually usefully do with distributional semantics? We know we can do word meanings, more or less. But what else? Personally, I don’t think compositionality through vector addition or multiplication captures any real linguistic intuition for how meanings are built in composition. A composite linguistic expression is built by virtue of its elements successively restricting further and further the meaning potentials for each other, until a meaning for the composite expression is fixed. In a successful communication, a full sentence has a single interpretation. This needs to be reflected in distributional representation of composite expressions.

References

- 1 H. Yu and J. M. Siskind. Grounded Language Learning from Video Described with Sentences. ACL 2013.
- 2 C. Matuszek, S. Tellex, D. Fox, L. Zettlemoyer. Proceedings of 2012 AAAI Grounding Language for Physical Systems.
- 3 T. Darrell, R. J. Mooney, K. Saenko, eds. Proceedings of NIPS 2011 Workshop on Integrating Language and Vision.
- 4 OD. Goldwasser, R. Barzilay, D. Roth, eds. Proceedings of NAACL 2012 Workshop on Semantic Interpretation in an Actionable Context.

5.16 Position statement

Annie Zaenen (*Stanford University, US*)

License  Creative Commons BY 3.0 Unported license
© Annie Zaenen

I am mainly interested in the inferences that can be drawn from texts. My research focuses on the linguistic elements that license inferences about veridicity and existence: what allows us to conclude that a speaker/author is committed to the view that an event has taken place or that an entity exists?

I collaborate with Cleo Condoravdi, Lauri Karttunen and Stanley Peters and occasional Stanford students in a small research unit at CSLI (Stanford), Language and Natural Reasoning (LNR). We assume that there are in natural language constructions and lexical items that signal that a speaker/author presents an event as factual (veridicity) or an entity as existent with certainty, with a high degree of plausibility, or alternatively allow us to conclude that they are impossible or implausible. It is, however, not easy to classify constructions/lexical items according to these inferential properties because we do not have direct access to speakers/authors intentions and, whatever the linguistic elements that signal the allowed inferences are, they interact with other elements in the discourse context that influence the de facto inferences that hearers/readers draw. We think, contra to de Marneffe, Potts and Manning, that it is important to distinguish between the linguistic components and the real world knowledge components that go into drawing conclusions: generalizations that are made over both together will always be constrained by the specifics of the situation in which they were calculated.

We are trying to develop a methodology that allows us to observe how naive language users draw the inferences we are interested in and to translate this understanding into possible annotations of linguistic material for these inferential properties.

At this point we concentrate on the properties of adjectives with clausal complements. Some of those, especially factive constructions, turn out to be more problematic than existing linguistic literature leads one to believe. First the conditions on the factive uses of specific constructions have not been described in enough detail in the existing literature. Whereas 'It was(n't) stupid of John to leave early' is factive, 'It is(n't) stupid to leave early' is not. In certain cases differences in tense can be associated with rather dramatic differences in interpretation 'He was lucky to break even.' is factive or implicative (see below) but 'He will be lucky to break even.' does not mean that the speaker/author thinks that it is likely or sure that the protagonist will fare well.

Lucky in the future seems to be an idiom and the explanation of the differences between the past and the present tense for impersonal evaluatives will most likely be linked to a better understanding of generic interpretations of the present tense, but in other cases, assumed factive expressions are interpreted as implicative and it is not so easy to decide how they should be treated: From a sentence such as: "I was not brave to venture out" one is supposed to conclude that the speaker did venture out. When one looks at the use of such sentences in context, however, one sees that they are often used as implicatives: the speaker did not venture out. But it is premature to simply conclude that for some speakers (non-native speakers?) brave is an implicative adjective. While this may be true for some of them, experimental evidence suggests that for many (native) speakers the interpretation depends on the context: a sentence such as 'He was not stupid to save money.' gets a factive interpretation, while 'He was not stupid to waste money.' gets an implicative interpretation.

A priori it is not clear how such differences such be accounted for: are the adjectives

ambiguous or are the implicative readings ‘performance’ errors? Our investigations suggest that for some speakers some of the adjectives (most clearly lucky, fortunate and stupid) are ambiguous or even only implicative and that, if the rest the variation has to be treated as a performance error, it is one that is very systematically influence by discourse coherence.

With respect to DS this raises the question whether the approach can distinguish between two readings that have rather closely related lexical environments, my suspicion is that it can in principle but it might need to take much more information into account than is done now. With respect to making the distinction between inferences that arise from the pressure of discourse coherence and those that are due to real lexical ambiguity I would like to see a more general discussion.

6 Panel Discussions

During the week, each group carried out intense discussions on the assigned topics, highlighting potential synergies between distributional and formal semantics, pointing out short term as well as long term strategies to implement them.

Each group prepared and presented a summary of their discussions and proposal. These reports were then unified and harmonized by the seminar organizers. The main results of group work are reported in the sections below.

6.1 Polysemy

Polysemy is a central problem for distributional semantics because typically vector representations do not distinguish word senses. Yet, distributional semantics is likely to be able to provide an important contribution to understand and model phenomena such as polysemy and vagueness. Here are some major challenges that distributional semantic model need to address in the near future:

- Can distributional models distinguish types / senses?
- Are there regularities in the model representations and processes corresponding to regularities in the meaning shifts?
- Can they distinguish productivity and conventionality? can we make the implicit information encoded in the vectors explicit (for example in terms of features and meaning components)?
- Can distributional models be augmented?
- Can we use distributional models to evaluate the analyses of semantic theory, for example analyses of meaning shifts?
- Can the distributional models go beyond that and act as a discovery?
- How can distributional semantics better model the notion of meaning potentials?

6.2 Inference

Inference is a stronghold of formal semantics. Conversely, distributional semantics is still not able to address satisfactorily even the most simple cases of natural language inferences. Here are some major issues concerning the treatment of inference with distributional semantics:

- It is necessary to bridge the gap between formal and distributional notions of inference
- Interesting possibilities might arise from the integration with probabilistic inferences
- One major issue is to what extent is DS able to “tap into” contextual information in text

- It is necessary to collect empirical data about examples of inferences, eventually leading to the creation of shareable datasets for model evaluation:
 - simple items exemplifying specific examples of inferences
 - annotated corpus-based examples

6.3 Compositionality

It makes sense to hypothesize that semantic representations include both something distributional and something “structural”/symbolic. We do not have a single agreed-upon hypothesis of what these mixed or parallel representations should be like:

- Overall, the hypothesis space for what sorts of constituents should have distributional representations is:
 1. distributional representations for words only (and/or words and morphemes)
 2. distributional representations for phrases or perhaps clauses
 3. have both word-level and phrase-level distributional representations available
- We see no reason not to exploit both syntax-driven and discourse-driven composition.
- “Flat” semantic representations for the symbolic side (e.g. Hobbs, MRS, other flat underspecified representations) are an alternative approach to compositionality that may address some of the issues raised by Hinrich Schütze as they are not dependent on the availability of a complete syntactic structure
- It would be ideal if the resulting system was psychologically plausible.
- It would also be ideal if the resulting system were useful for NLP applications.
- We should also look for data sets and problems that will get distributional semantic researchers and formal researchers to talk to each other and benefit from what each approach does significantly better than the other. Examples where DS looks promising include:
 1. co-compositionality (e.g., ‘white wine’)
 2. metonymy
 3. explaining highly context-dependent paraphrases that are below (or beyond) the sense level (so not explainable by a lexical resource)
 4. that part of anaphora that depends on lexical content (e.g. cases of quasi-synonymy like ‘his recent appearance at the Carnegie Hall’/ ‘the concert’ / the evening’)
- It would be interesting to tease apart the influence of discourse dynamics on how we identify referents from its influence on how we interpret lexical items.

6.4 Negation

The group decided to focus on negations, because this is a central aspect of natural language semantics, and yet there is no analysis for it in distributional semantics to date:

- Distributional semantics has no treatment for negation, when viewed in the classical definition;
- Distinction between decontextualized and conversational negation;
- Perspectives for Distributional semantics to help identify the comparison sets for the negated item;
- This approach can possibly link to cognitively inspired models of thought.

7 Next Steps

The seminar organizers together with the participants proposed various activities to carry on the discussions started in Dagstuhl:

- Organize a follow-up meeting (3 days) in Pisa, Italy in September 2014
- Provide details about existing datasets (according to a common format) containing interesting linguistic phenomena, to be used as test set for distributional and formal semantic models
- Groups provide a specification over new datasets for challenging, not yet addressed semantic phenomena
- Groups define annotation metadata for the dataset.
- Groups Identify burning topics for next meeting:
 - a. what is the right architecture?
 - b. information structure
 - c. Finding a task/problem where different areas need to be integrated

Participants

- Nicholas Asher
Paul Sabatier University –
Toulouse, FR
- Marco Baroni
University of Trento, IT
- Peter A. Cariani
Harvard Medical School –
Newton, US
- Stephen Clark
University of Cambridge, GB
- Ann Copestake
University of Cambridge, GB
- Ido Dagan
Bar-Ilan University, IL
- Katrin Erk
University of Texas – Austin, US
- Stefan Evert
Univ. Erlangen-Nürnberg, DE
- Patrick W. Hanks
Univ. of Wolverhampton, GB
- Graeme Hirst
University of Toronto, CA
- Jerry R. Hobbs
Univ. of Southern California –
Marina del Rey, US
- Hans Kamp
Universität Stuttgart, DE
- Lauri Karttunen
Stanford University, US
- Alessandro Lenci
University of Pisa, IT
- Sebastian Löhnner
Heinrich-Heine-Universität
Düsseldorf, DE
- Louise McNally
UPF – Barcelona, ES
- Sebastian Padó
Universität Stuttgart, DE
- Massimo Poesio
University of Essex, GB
- James Pustejovsky
Brandeis Univ. – Waltham, US
- Anna Rumshisky
University of Massachusetts –
Lowell, US
- Hinrich Schütze
LMU München, DE
- Mark Steedman
University of Edinburgh, GB
- Suzanne Stevenson
University of Toronto, CA
- Tim van de Cruys
Paul Sabatier University –
Toulouse, FR
- Jan van Eijck
CWI – Amsterdam, NL
- Dominic Widdows
Microsoft Bing – Bellevue, US
- Annie Zaenen
Stanford University, US
- Alessandra Zarcone
Universität Stuttgart, DE

