A Theory of Content

Mark Steedman (with Mike Lewis and Nathan Schneider)

August 2016
Outline

I: Distributional Theories of Content: Collocation vs. Denotation

II: Entailment-based Paraphrase Cluster Semantics (Lewis and Steedman, 2013a, 2014)

III: Multilingual Entailment-based Semantics (Lewis and Steedman, 2013b)

IV: Entailment-based Semantics of Temporality
The Problem of Content

- We have (somewhat) robust wide coverage parsers that work on the scale of Bn of words. They can read the web (and build logical forms) thousands of times faster than we can ourselves.

- So why can't we have them read the web for us, so that we can ask them questions like “What are recordings by Miles Davis without Fender Rhodes piano”, and get a more helpful answer than the following?
what are Miles Davis albums without Fender Rhodes piano – Google Search – Mozilla Firefox

Miles Davis - Get Up With It (CD, Album) at Discogs
https://www.discogs.com/Miles-Davis-Get-Up-With-It/release:1203374

What are some great Fender Rhodes jazz albums/groups (pre-
https://www.quora.com/What-are-some-great-Fender-Rhodes-jazz-album

Fender Rhodes electric piano - Bill Evans
www.billevanswebpages.com/rhodespeche.html

Live-Evil (Miles Davis album) - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Live-Evil_(Miles_Davis_album)

Who was the first artist to record using a Fender Rhodes?
Too Many Ways of Answering The Question

- The central problem of QA is that there are too many ways of asking and answering questions, and we have no idea of the semantics that relates them.
- Your Question: *Has Verizon bought Yahoo?*
- The Text:
  1. Verizon purchased Yahoo. (“Yes”)
  2. Verizon’s purchase of Yahoo (“Yes”)
  3. Verizon owns Yahoo (“Yes”)
  4. Verizon managed to buy Yahoo. (“Yes”)
  5. Verizon acquired every company. (“Yes”)
  6. Yahoo may be sold to Verizon. (“Maybe”)
  7. Verizon will buy Yahoo or Yazoo. (“Maybe not”)
  8. Verizon didn’t take over Yahoo. (“No”)
The Problem

• The hard problem in semantics is not the logical operators, but the content that they apply over.

• How do we define a theory of content that is robust in the sense of generalizing across linguistic form, and compositional in the sense of:
  – being compatible with logical operator semantics and
  – supporting commonsense inference?
Previous Work

- Many have tried to build a form-independent semantics by hand:
  - both in linguistics, as in the “Generative Semantics” of the ’70s and the related conceptual representations of Schank and Langacker;
  - and in computational linguistics, as in WordNet, FrameNet, Generative Lexicon, VerbNet/PropBank, BabelNet, AMR . . .
  - and in knowledge graphs such as FreeBase.
Previous Work

Such hand-built semantic resources are extremely useful, but they are notoriously **incomplete** and **language-specific**.

- So why not let machine learning do the work instead?
- Treat **semantic primitives as hidden**.
- **Mine them** from unlabeled multilingual text, using **Machine Reading**.
One (Somewhat*) New Approach

• Clustering by Collocation
  – Meanings are vectors (etc.)
  – Composition is via Linear Algebraic Operations such as vector addition, matrix multiplication, Frobenius algebra, packed dependency trees, etc.
  – Vectors are good for underspecification and disambiguation (Analogy tasks and Jeopardy questions), and for building RNN embeddings-based “Supertagger” front-ends for CCG parsers, and related transition models for transition-based dependency parsers

* Cf. the MDS “Semantic Differential” (1957), which Wordnet was developed by George Miller partly in reaction to.
For Example: Analogy via Word2Vec

- \text{king} - \text{man} + \text{woman} = ["queen", 0.711819291148071], ["monarch", 0.6189674139022], ["princess", 0.5902431011199951], ["crown prince", 0.5499460697174072], ["prince", 0.5377321243286133]

- \text{picnic} - \text{beer} + \text{wine} = ["wine tasting", 0.5751593112945557], ["picnic lunch", 0.5423362255096436], ["picnics", 0.5164458155632019], ["brunch", 0.509375810623169], ["dinner", 0.5043480396270752]

- \text{right} - \text{good} + \text{bad} = ["wrong", 0.548572838306427], ["fielder Joe Borchard", 0.47464582324028015], ["left", 0.46392881870269775], ["fielder Jeromy Burnitz", 0.45308032631874084], ["fielder Lucas Duda", 0.4393044114112854]

- \text{Bernanke} - \text{USA} + \text{Russia} = ["Ben Bernanke", 0.6536909937858582], ["Kudrin", 0.6301712989807129], ["Chairman Ben Bernanke", 0.6148115396499634], ["Medvedev", 0.6024096608161926], ["Putin", 0.5873086452484131]
Orthogonality in Vector Components

• “A is to B as C is to D” works best when the two components AB and BC are orthogonal i.e. independent, and if B and D are close anyway. Compare:

  - smaller - small + big = ["bigger", 0.7836999297142029], ["larger", 0.586679697036741], ["Bigger", 0.5707237720489502], ["biggest", 0.5240510106086731], ["splashier", 0.5107756853103638]
  - unhappy - happy + fortunate = ["incensed", 0.49339964985847473], ["displeased", 0.4742095172405243], ["unfortunate", 0.46231183409690857], ["frustrated", 0.4529050886631012], ["miffed", 0.4450964927673343]
  - Las Meninas - Velasquez + Picasso = [“Paul Cézanne”, 0.6370980739593506], [“Pablo Picasso”, 0.634435772895813], [“Renoir”, 0.6213735938072205], [“Dubuffet”, 0.619714617729187], [“Degas”, 0.6172788143157959]
  - kill - dead + alive = ["destroy", 0.4605627655982971], ["exterminate", 0.42368459701538086], ["survive", 0.3986499309539795], ["stymie", 0.39753955602645874]
Factorization in Vector Components

- Mitchell and Steedman (2015) show that the orthogality effect holds for a range of morpho-syntactic components, and that in general the cosine of vector differences is a strong predictor of performance on the word analogy task for CBOW, SkipGram, and GloVe.

- But this makes them look rather like old fashioned morpho-syntactic-semantic features male/female, active/inactive, etc.

- It is unclear how to apply logical operators like negation to vectors.

- Beltagy et al. (2013) use vectors to estimate similarity between formulæ in an otherwise standard logical approach.
Another (Somewhat*) New Approach

• Clustering by Denotation:
  – Meanings are automatically-extracted hidden relations, identified by automatic parsing and recognition of Named Entities either in text or in knowledge graphs.
  – Semantic composition is via syntactic derivation and traditional Logical Operators such as ¬, ∧, ∨, etc.
  – Denotations are good for inference of entailment from the text to an answer to your question.
  – They are directly compatible with negation, quantifiers, modality, etc.

II: Entailment-based Paraphrase Cluster Semantics

• Instead of traditional lexical entries like the following:

(1) \begin{align*}
\text{author} & := N/PP[\text{of}] : \lambda x \lambda y. \text{author}'xy \\
\text{write} & := (S \backslash NP)/NP : \lambda x \lambda y. \text{write}'xy
\end{align*}

•—we seek a lexicon capturing entailment via logical forms defined as (conjunctions of) paraphrase clusters like the following:

(2) \begin{align*}
\text{author} & := N/PP_{\text{of}} : \lambda x_{\text{book}} \lambda y_{\text{person}}. \text{relation37}'xy \\
\text{write} & := (S \backslash NP)/NP : \lambda x_{\text{book}} \lambda y_{\text{person}}. \text{relation37}'xy
\end{align*}

• Such a “distributional” lexicon for content words works exactly like the naive lexicon (1) with respect to the semantics of quantification and negation.
Finding Typed Relation Expressions in Text

• We obtain the clusters by parsing (e.g.) Gigaword text with (e.g.) the CCG-based logical-form-building C&C parser, (Bos et al., 2004), using the semantics from Steedman 2012, with a lexicon of the first type (1), to identify expressions relating Named Entities such as Verizon, Yahoo, Scott, Waverley, etc.

• Nominal compounds for the same MUC named entity type are merged.

• Entities are soft-clustered into types according to a suitable method (Topic models, WordNet clusters, FreeBase types, etc.)

• These types are used to distinguish homonyms like the two versions of the born in relation relating PERSONS to DATES versus LOCATIONS
Example

- Obama was born in Hawai‘i.

\[ \text{born} := (S \setminus NP)/PP[\text{in}] : \lambda x \lambda y. \begin{cases} x = \text{LOC} \land y = \text{PER} \Rightarrow \text{rel49} \\ x = \text{DAT} \land y = \text{PER} \Rightarrow \text{rel53} \end{cases} \]

\[ \text{Obama} := \begin{cases} \text{PER} = 0.9 \\ \text{LOC} = 0.1 \end{cases} \]

\[ \text{Hawai‘i} := \begin{cases} \text{LOC} = 0.7 \\ \text{DAT} = 0.1 \end{cases} \]

- The “Packed” Distributional Logical Form

\[ S : \begin{cases} \text{rel49} = 0.63 \\ \text{rel53} = 0.27 \end{cases} \text{hawai‘i'}obama' \]
Directional Entailments

- We now search for potential entailments between such typed relations, where for multiple pairs of entities of type $X$ and $Y$, if we find relation A in the text we often also find relation B stated as well.

Entailment is a directed relation: $X_{\text{person}}$ elected to $Y_{\text{office}}$ does entail $X_{\text{person}}$ ran for $Y_{\text{office}}$ but not vice versa.

- Thus we use an asymmetric similarity measure rather than Cosine.

- Lewis (2015); Lewis and Steedman (2014) apply the entailment graphs of Berant et al. (2012) to generate more articulated entailment structures.
Local Entailment Probabilities

- The typed named-entity technique is applied to (errorfully) estimate local probabilities of entailments:

  a. $p(conquer_{x_{country}}y_{country} \Rightarrow invade_{x_{country}}y_{country}) = 0.9$
  b. $p(invade_{x_{country}}y_{country} \Rightarrow attack_{x_{country}}y_{country}) = 0.8$
  c. $p(invasion\ (of\ x_{country})\ (by\ y_{country}) \Rightarrow attack_{x_{country}}y_{country}) = 0.8$
  d. $p(invade_{x_{country}}y_{country} \Rightarrow invasion\ (of\ x_{country})\ (by\ y_{country})) = 0.7$
  e. $p(invasion\ (of\ x_{country})\ (by\ y_{country}) \Rightarrow invade_{x_{country}}y_{country}) = 0.7$
  f. $p(conquer_{x_{country}}y_{country} \Rightarrow attack_{x_{country}}y_{country}) = 0.4$
  g. $p(conquer_{x_{country}}y_{country} \Rightarrow conqueror\ (of\ x_{country})\ y_{country}) = 0.7$
  h. $p(conqueror\ (of\ x_{country})\ y_{country} \Rightarrow conquer_{x_{country}}y_{country}) = 0.7$
  i. $p(bomb_{x_{country}}y_{country} \Rightarrow attack_{x_{country}}y_{country}) = 0.7$

(etc.)
Global Entailments

- The local entailment probabilities are then used to construct an entailment graph using integer linear programming with a prior $p = 0.25$ with the global constraint that the graph must be closed under transitivity.

- Thus, (f) will be included despite low observed frequency, while other low frequency spurious local entailments will be excluded.

- Cliques within the entailment graphs are collapsed to a single paraphase cluster relation identifier.

- The entailment graph is Boolean, rather than probabilistic.
• A simple entailment graph for relations between countries.
Lexicon

- The lexicon obtained from the entailment graph

  attack := \((S\setminus NP)/NP : \lambda x\lambda y\lambda e. rel_1 x ye\)

  bomb := \((S\setminus NP)/NP : \lambda x\lambda y\lambda e. rel_1 x ye \land rel_4 x ye\)

  invade := \((S\setminus NP)/NP : \lambda x\lambda y\lambda e. rel_1 x ye \land rel_2 x ye\)

  conquer := \((S\setminus NP)/NP : \lambda x\lambda y\lambda e. rel_1 x ye \land rel_2 x ye \land rel_3 x ye\)

  conqueror := \(VP_{pred}/PP_{of} : \lambda x\lambda p\lambda y\lambda e. p y \land rel_1 x ye \land rel_2 x ye \land rel_3 x ye\)

- These logical forms support correct inference under negation, such as that
  conquered entails attacked and didn’t attack entails didn’t conquer
Entailment

• Thus, to answer a question “Did X conquer Y” we look for sentences which subsume the conjunctive logical form $rel_2 \land rel_1$, or satisfy its negation $\neg rel_2 \lor \neg rel_1$.

Note that if we know that invasion-of is a paraphrase of $\text{invade} = rel_2$, we also know invasion-of entails $\text{attack} = rel_1$. 

Steedman, Univ. of Edinburgh RefSemPlus, Bolzano August 2016
Examples from Question-Answer Test Set

- Examples:

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>From Unseen Sentence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>What did Delta merge with?</td>
<td>Northwest</td>
<td>The 747 freighters came with Delta’s acquisition of Northwest</td>
</tr>
<tr>
<td>What spoke with Hu Jintao?</td>
<td>Obama</td>
<td>Obama conveyed his respect for the Dalai Lama to China’s president Hu Jintao during their first meeting</td>
</tr>
<tr>
<td>What arrived in Colorado?</td>
<td>Zazi</td>
<td>Zazi flew back to Colorado...</td>
</tr>
<tr>
<td>What ran for Congress?</td>
<td>Young</td>
<td>. . . Young was elected to Congress in 1972</td>
</tr>
</tbody>
</table>

- Full results in Lewis and Steedman (2013a) and Lewis (2015)
III: Multilingual Entailment Cluster Semantics

- If we can find entailments including paraphrases by observing local entailments between statements in English of relations over typed named entities, there is no reason we shouldn’t consider statements in other languages concerning named entities of the same types as nodes in the same entailment graph.

- Thus from French *Shakespeare est l’auteur de* Mesure pour mesure, and knowledge of how French named entities map to English, we should be able to work out that *être l’auteur de* is a member of the *write* cluster.

- We use cross-linguistic paraphrase clusters to re-rank Moses n-best lists to promote translations that preserve the cluster-based meaning representation from source to target.
**Experiment: Reranking SMT Translations**

- For a source (French) sentence that can be dependency-parsed to deliver a cluster-semantic logical form:

- We Moses-translate (to English) taking the 50-best list and parsing (with C&C) to produce cluster-semantic logical forms.

- If the logical form of the top ranked translation is different from that of the source, we choose whatever translation from the remainder of the n-best list has the logical form that most closely resembles the source cluster semantics.
Reranking SMT

• Example:
  
  **Source:** Le Princess Elizabeth arrive à Dunkerque le 3 août 1999
  
  **SMT 1-best:** The Princess Elizabeth is to manage to Dunkirk on 3 August 1999.
  
  **Reranked 1-best:** The Princess Elizabeth arrives at Dunkirk on 3 August 1999.

• Fluent bilingual human annotators are then asked to choose between the one-best Moses translation and the cluster-based alternative.

<table>
<thead>
<tr>
<th></th>
<th>Percentage of Translations preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-best Moses</td>
<td>5%</td>
</tr>
<tr>
<td>Reranked best</td>
<td>39%</td>
</tr>
<tr>
<td>No preference</td>
<td>56%</td>
</tr>
</tbody>
</table>
Reranking SMT

- Many cases of “no preference” were where Moses and the preferred translation were similar strings differing only in attachment decisions visible only in the logical form.

♫ No parallel text was used in these experiments.

- This is good, because SMT has already used up all of the available parallel text (Och, 2007)!

- Full results in Lewis and Steedman (2013b).
IV: Temporal Semantics

- As in the case of the semantics of content words like nouns and verbs, the semantics of tense, aspect, modality, evidentiality, and intensionality has always seemed to bog down in conflicting and overlapping ontology, and ill-defined or world-knowledge-entangled notions like “inertia worlds”, “relevance”, “extended now”, “perfect time span”, “consequent state”, “preparatory activity”, and the like.
- #Einstein has visited New York (vs. Einstein visited New York).
- #I have forgotten your name but I have remembered it again (vs. I forgot your name but I remembered it again).
- Such relations seem like A Suitable Case for Treatment as hidden relations, letting machine learning find out what the consequent states of people visiting places, forgetting and remembering things, etc. usually are.
A simple entailment graph for relations over events does not yet capture relations of causation and temporal sequence entailment.
Timestamped Data

- We have begun pilot experiments with timestamped news, using the University of Washington NewsSpike corpus of 0.5M newswire articles (Zhang and Weld, 2013).

- In such data, we find that statements that so-and-so is visiting, is in and the perfect has arrived in such and such a place, occur in stories with the same datestamp, whereas is arriving, is on her way to, occur in preceding stories, while has left, is on her way back from, returned, etc. occur in later ones.

- This information provides a basis for inference that visiting entails being in, that the latter is the consequent state of arriving, and that arrival and departure coincide with the beginning and end of the progressive state of visiting.

- We can use it as the input to a neo-Reichenbachian semantics of temporality.
Reconnecting with Logical Operator Semantics

- Some **handbuilt** lexical entries for **auxiliary verbs** (closed-class words):

  has := (S\NP)/VP\textsubscript{en} : \lambda p E \lambda y. consequent-state (p_E y) \text{R} \land \text{R} = \text{NOW}

  will := (S\NP)/VP\textsubscript{b} : \lambda p E \lambda y. priors \Rightarrow imminent-state (p_E y) \text{R})
  \land \text{R} = \text{NOW}

  is := (S\NP)/VP\textsubscript{ing} : \lambda p E \lambda y. progressive-state (p_E y) \text{R} \land \text{R} = \text{NOW}

Reconnecting with Logical Operator Semantics

- Some potentially learnable lexical entries for implicative verbs:

\[
\text{tried} := (S\setminus NP)/VP_{to} : \lambda p_E \lambda y. rel_{\text{try}} p_E y R \land rel_{\text{want}} p_E y R \\
\land \text{preparatory-activity}(p_E y) y R \land R < \text{NOW}
\]

\[
\text{failed} := (S\setminus NP)/VP_{to} : \lambda p_E \lambda y. rel_{\text{try}} p_E y R \land rel_{\text{want}} p_E y R \\
\land \text{preparatory-activity}(p_E y) y R \land \neg p_E y R \land R < \text{NOW}
\]

\[
\text{managed} := (S\setminus NP)/VP_{to} : \lambda p_E \lambda y. rel_{\text{try}} p_E y R \land rel_{\text{want}} p_E y R \\
\land \text{preparatory-activity}(p_E y) y R \land p_E y R \land R < \text{NOW}
\]

\(\checkmark\) Needs negation as failure to find positive entailing text.
Conclusion I: Denotation-based

- Learning over denotations, defined as relations over typed named entities, allows us to build entailment into lexical logical forms for content words via conjunctions of paraphrase clusters.

- The individual conjuncts are potentially language-independent.

- Mining them by machine reading remains a hard task, for which we have no more than proof-of-concept!

- The lexical conjunctions are projected onto sentential logical forms including traditional logical operators by the function words and CCG syntax.

- The sentential logical forms support fast inference of common-sense entailment.
Conclusion II: Collocation-based

- **Learning over Collocations**, represented as a vector space with reduced dimensionality, also represents meanings in terms of hidden components.

- Projection by vector addition remains a hard baseline to beat!

- By **superimposing a number of distinct collocations**, they remain the most powerful mechanism known for resolving ambiguity, as in the use of embeddings and LSTM in parser models.

☞ When Firth (1957/1968):179 made his oft-cited remark about knowing a word by the company it keeps, he was actually talking about disambiguation.
Thanks to:

- Johan Bos (Groningen), Steve Clark (Cambridge), James Curran (Sydney), Brian Harrington (Toronto), Julia Hockenmaier (Illinois), Mirella Lapata, Mike Lewis (Washington), Reggy Long (Stanford), Jeff Mitchell (UCL), Siva Reddy and Nathan Schneider (Georgetown).

- And to http://rare-technologies.com/word2vec-tutorial/#app for running Word2Vec, Congle Zhang and Dan Weld for NewsSpike, and to Google and ERC GramPlus for support.
Conclusions: For Philosophy of Language

- Under more traditional semantic theories employing eliminative definitions these entailments would have been thought of as belonging to the domain of inference rather than semantics, either as meaning postulates relating logical forms or as “encyclopædic” general knowledge.

- Carnap (1952) introduced meaning postulates in support of Inductive Logic, including a model of Probability, basically to keep the model small and consistent.

- Like Katz and Fodor (1963); Katz and Postal (1964); Katz (1971), we are in effect packing meaning postulates into the lexicon.

- This suggests that our semantic representation expresses an a pragmatic empiricist view of analytic meaning of the kind advocated by Quine (1951).
Conclusions: For Psychology

- Do children acquire the meaning of words like “invade” and “conquer” by building entailment graphs?
- I suggest they do, and that this is the mechanism for what Gleitman (1990) called syntactic bootstrapping of the lexicon—that is:
  - Once children have acquired core competence (by semantic bootstrapping of the kind modeled computationally by Kwiatkowski et al. 2012 and Abend et al., 2016), they can detect that “annex” is a transitive verb meaning some kind of attack without knowing exactly what it means.
  - They can then acquire the full meaning by piecemeal observation of its entailments and paraphrases in use.
- This is a major mechanism of cultural inheritance of concepts that would otherwise in many cases take more than an individual lifetime to develop.
Conclusions: For Cognitive Science

- These terms compile into a (still) language-specific Language of Thought (Fodor 1975, *passim*), which is roughly what adult speakers do their thinking in.

- To the extent that the cliques or clusters in the graph are constructed from multilingual text, this meaning representation will approximate the hidden language-independent “private” Language of Mind which the child language learner accesses.

- However, very few terms in any adult logical form correspond directly to the hidden primitives of that Language of Mind. (*red* and maybe *attack* might be exceptions.)

- Even those terms that are cognitively primitive (such as color terms) will not be unambiguously lexicalized in all languages.
Some conceptual primitives, such as that things can only be in one place at a time, probably predate human cognition, and are unlikely to be discoverable at all by machine reading of the kind advocated here.

- These properties are hard-wired into our minds by 600M years of vertebrate evolution.
- These are exactly the properties that Artificial Intelligence planning builds in to the representation via the “Closed World Assumption” and the STRIPS dynamic logic of change.
- Computational Linguistics should learn from AI in defining a Linear Dynamic Logic for distributional clustered entailment semantics.
References


Filip, Hana, 2008. “Events and Maximalization: The Case of Telicity and
Perfectivity.” In Susan Rothstein (ed.), *Theoretical and Crosslinguistic
Approaches to the Semantics of Aspect*, John Benjamins Publishing, volume
110. 217–258.

*Selected Papers of J.R.Firth*, Longmans. 168–205.


Acquisition* 1:1–55.

Hovy, Eduard, Gerber, Laurie, Hermjakob, Ulf, Junk, Michael, and Lin, Chin-Yew,


Steedman, Mark, 1997. “Temporality.” In Johan van Benthem and Alice


Zhang, Congle and Weld, Daniel, 2013. “Harvesting Parallel News Streams to