Research Statement
Mohit Bansal

For artificial intelligent agents to function autonomously in our homes and workplaces, they must be able to effectively understand natural language for instructions and communication. For this, it is necessary to resolve the various deep and subtle semantic (meaning-related) ambiguities in natural language, a challenging goal that involves two fundamental requirements. First, we need diverse, external world knowledge which is simply not present in the standard training datasets used for supervised NLP tasks. Second, we need to develop appropriate machine learning models that can extract the precise disambiguation cues lying latent in such diverse, large-scale data. My research addresses both these requirements by learning novel weakly-labeled and cross-modal semantic representations with accurate, well-formulated disambiguation models, achieving the state-of-the-art on various core NLP tasks and multimodal applications.

I first model world knowledge by automatically extracting pattern statistics from a huge unlabeled corpus of Web text. When augmented with these robust surface statistics, my supervised structured methods do a better job at resolving sentence- and document-level ambiguities, and at learning various types of semantic relations. In addition to pattern-based knowledge, language understanding can also be improved via dense (continuous space), distributed representations of language. These unsupervised embeddings, usually trained on large amounts of unlabeled surrounding context, capture useful, fine-grained meaning relationships but with a generic similarity that mixes concepts such as antonyms and hypernyms. My research learns stronger, relation-discriminative embedding spaces via weak supervision from automatically extracted knowledge bases and improved neural models, achieving strong improvements in various NLP tasks, with human-level performance in some cases.

Grounding cues (i.e., mapping words to physical things in the world) from other modalities such as vision and speech can help resolve tricky pragmatic (contextual) ambiguities which cannot be resolved using any amount of textual data. With colleagues, I show how coreference ambiguities in multi-sentence image captions can be resolved using visual, alignment-based cues in a multimodal structured prediction formulation. I also develop a neural translation style method to successfully understand and execute navigational instructions by using visual cues from an environment map. Moreover, subtle syntactic ambiguities in spoken language can also be resolved using prosodic cues from speech signals.

My research develops diverse learning methods to accurately model the underlying semantics for the task at hand, e.g., from probabilistic graphical models for structured prediction in taxonomy induction and image-text coreference to recursive and recurrent neural networks (with varying loss functions) for embedding learning and sequence-to-sequence tasks, to a combination of structure and neural methods (e.g., coarse-to-fine alignment in encoder-decoder models). A common theme in my model formulations is to employ multiple ‘views,’ e.g., two languages or a paraphrase pair, image and text, navigational instruction and actions in a map image, speech and text, database and language, etc.

1) World Knowledge Semantics via Web Features: World knowledge is required to resolve many of the remaining ambiguity errors in state-of-the-art NLP methods. I show that this deep, subtle knowledge can be robustly captured from large amounts of unlabeled data, if you extract the right statistics and suitably incorporate them into expressive, supervised structured methods. Specifically, I learn disambiguation cues from the vast, all-domain Web via 4 billion n-grams (∼500x larger than Wikipedia) to effectively address various facets of semantics [BK11; BK12; MB13; BBMK14; NBC15].

Taxonomies provide useful semantic category knowledge (e.g., type-of or hypernymy in Fig. 1) to many NLP tasks. However, manual taxonomies like WordNet are incomplete, expensive, and unavailable in most languages. Hence, we need to automatically build such taxonomies from scratch. The first challenge involved is to develop useful domain-independent cues, because a test set from a new domain (e.g., vehicles) may have no word overlap with the taxonomies that we train on (e.g., animals).
To address this, I employ external, Web-attested, unlexicalized features. E.g., “C and other P” could be a useful (unlexicalized) hypernymy feature if “rat” is a type or child (C) of the parent (P) “rodent”, and “rats and other rodents” is the highest-frequency string on the Web between these two terms (and also between various other training child-parent term pairs). We also use siblinghood (coordination) features to mutually help hypernymy by transitivity, again using Web cues, e.g., “rats are similar to squirrels”. Note that we simply collect unlexicalized versions of the 100 most-frequent Web strings between each potential child-parent or sibling pair and let our supervised structured model automatically discover the interesting hypernymy and siblinghood features.

The second challenge is efficient inference over taxonomies with such higher-order (multi-edge) siblinghood features. I address this by formulating taxonomies as directed spanning trees in a probabilistic graphical model with edge, sibling, and tree factors (Fig. 2), and then using approximate message passing (loopy belief propagation) for inference, incorporating standard algorithms for summing and maximizing over directed spanning trees. Overall, on the task of training on and predicting disjoint subhierarchies of WordNet, I achieve substantial improvements over the state-of-the-art [BBMK14] (ACL top-5 paper).

Another domain where I show the power of large-scale unlabeled data in supervised structured learning is syntactic parsing. Parsers are crucial for analyzing sentence structure in several NLP tasks. However, most remaining ambiguity errors in state-of-the-art parsers need external knowledge, e.g., the prepositional phrase (PP) attachment error in Fig. 3, where the PP “from debt” incorrectly modifies the noun phrase “$30 billion” in the Berkeley parser (solid blue edges), whereas the correct attachment (dashed gold edges) is to the verb “raising”. Two example Web-based semantic cues to the correct attachment are lexical co-occurrence (count(“raising from”)) \( \geq \) count(“\$ x billion from”)) and highest-frequency pronominal paraphrases (“raising it from”). I integrate such Web affinity and paraphrase attestations as generalized features into full-sentence, graph-based parsing (i.e., over all potential attachments) via an efficient hash tree method. This novelty allows us to address all error types with a single model, achieving strong improvements in multi-domain dependency and constituent parsing. Discriminative structured learning enables automatically learning useful disambiguating cues, not only reproducing the hand-picked patterns used by previous work, but also discovering several new ones [BK11; NBC15]. I also learn Web features to achieve the state-of-the-art in coreference resolution [BK12], and improve up to human-level performance in intensity ordering of near-synonyms (e.g., “acceptable” < “good” < “great” < “superb”) [MB13].

2) Knowledge-based Representation Learning: Distributed representations, trained on large amounts of external data, also improve language understanding by capturing useful, fine-grained meaning relationships. However, these continuous embeddings are usually learned from unlabeled surrounding context, leading to a generic similarity that conflates antonyms and hypernyms, and does not prove useful for several downstream tasks. My research integrates weak supervision from paraphrases and task-useful context with improved embedding models to learn stronger, discriminative embeddings for words, phrases, and task substructures [BGL14; Ban15; LWBGL15; WBGLR15; WBGL16].

Paraphrastic information can be used to improve unsupervised embedding spaces by learning closer vectors (in an \( n \)-dimensional space) for phrases with the same meaning, while separating the vectors for phrases that differ in meaning. We employ positive and negative paraphrase pairs from an automatically extracted paraphrase database (PPDB) to add such constraints to a compositional paraphrase embedding model, a pair of recursive neural networks that compose parent phrase vectors from children vectors and are tied together by a margin-based loss on the training pairs (Fig. 4) [WBGLR15]. This
model embeds phrases into a low-dimensional space where vector similarity represents paraphrastic strength. Choice of the negative examples is also important: instead of the obvious non-paraphrase pairs with low word overlap, we use pairs that the model incorrectly identified as being very similar. This helps learn to discriminate among the nearest neighbors. The resulting embeddings are the state-of-the-art on various similarity tasks, reaching human-level performance on some. Moreover, our parametric model scores paraphrase pairs more accurately than PPDB’s original confidence scores, while also improving its non-parametric coverage. We extend this to learn universal sentence embeddings via transferable models that outperform competitive baselines on tasks from various domains [WBGL16].

Relatedly, multilingual paraphrases can also add useful constraints to embedding learning, based on the intuition that translational context is invariant across languages. My work uses two-view, deep canonical correlation analysis (Fig. 5) to map monolingual embeddings from two languages to a shared, non-linear, multilingual space [LWBGL15]. The resulting embeddings improve over monolingual and linear CCA embeddings on various similarity tasks and in separating antonyms, hypernyms, and senses. Further, generalizing to three language families and morphological features shows promising initial improvements.

My research also shows how much of the popular unsupervised embeddings (with closer vectors for words with similar surrounding context) do not really help important downstream tasks like syntactic parsing, which intuitively prefers an embedding space in which vectors are closer for words with similar syntactic properties. Hence, I train a neural embedding model using syntactic context (dependency parent, grandparent, and label) and the resulting embeddings give strong gains in both in-domain and out-of-domain dependency parsing. Our syntactic vectors have since been widely used to improve various state-of-the-art parsers and NLP tasks [BGL14]. I also learn continuous representations of higher-order task substructures (e.g., dependency links), which allows a much simpler, smaller, and faster parsing model, with accuracy equal to (and complementary with) state-of-the-art parsers with millions of sparse, template-based features [Ban15]. Currently, I am investigating embedding context for other downstream tasks such as entity recognition, coreference resolution, and sentiment analysis.

3) Cross-modal Knowledge for Language Understanding: Pragmatic ambiguities, with multiple plausible contextual meanings (e.g., “the mug on the table with a crack”), cannot be resolved via any amount of text. Grounding cues (i.e., mapping words to physical things in the world) from other modalities such as vision and speech, when modeled in accurate multimodal frameworks, can help resolve such tricky pragmatic ambiguities.

With vision colleagues, I use visual information to resolve complex coreference ambiguities in textual captions (Fig. 6) [KLBUF14]. Our key, novel intuition here is that if we align nouns and pronouns in multi-sentence captions to their referred objects in the image, then linguistic effects such as coreference emerge naturally. Moreover, our alignments are also seen as an important starting point for other downstream text-image/video tasks such as caption generation and visual question answering. Specifically, we first exploit linguistic information from RGB-D scene descriptions (e.g., spatial prepositional constraints between nouns and pronouns) to improve 3D object detection, while also learning which visual object each noun (pronoun) is referring (aligning) to. This is modeled in a structured prediction framework (a Markov Random Field) with joint text-image potentials. The learned alignments and improved 3D detection, in turn, help resolve grounded and textual coreference cases in the caption (i.e.,
two nouns or pronouns aligning to the same object are coreferent), leading to improvements over the Stanford coreference resolution method. I also use grounding cues from a visual map environment to understand and follow navigational instructions [MBW16]. The key insight is to treat the task as end-to-end translation of navigational instructions to action sequences (via recurrent neural networks), with alignment-based cues between words and the current visual world state. This formulation promotes generalizability by (novelly) avoiding the use of any specialized resources (e.g., semantic parsers and logic-form lexicons); yet, it achieves the best results reported to-date on a benchmark dataset.

Acoustic prosodic cues can help resolve ambiguous spoken queries, e.g., in “toys for kids that are cheap,” an acoustic cue to the correct interpretation (i.e., cheap toys, not cheap kids) will be a subtle prosodic pause after “kids.” With speech colleagues, I get promising parsing improvements on adding such prosodic feature embeddings to the state-of-the-art Stanford neural dependency parser.

Other Work (Language Understanding for Generation and Dialog): Fully-conversational agents, in addition to interpreting ambiguous language, must also be capable of generating statements and answering questions for dialog. My related work includes language generation of a summary of salient events from a large database, e.g., of weather forecasts, again using alignment-based recurrent neural networks to achieve state-of-the-art results [MBW15]. I also work on multiple forms of question answering (Q&A). In [WBGM15], we achieve the state-of-the-art in passage-based Q&A by using syntactic and semantic knowledge. I am also working on Q&A in knowledge bases, where I improve over the state-of-the-art semantic parser via jointly trained question and answer neural embeddings. Addressing multi-step answering to maintain dialog is an important future direction.

Future Directions: My overall, long-term goal is to enhance automatic natural language understanding to human levels for seamless dialog and task execution, which involves addressing various key aspects of language. First, such day-to-day language can be very complex and ambiguous (e.g., contextual or non-literal). For this, I plan to continue modeling newer, deeper facets of semantics such as grounded entailment or implicature (e.g., “run” → move from A to B, increased heart-rate), contextual pragmatics (i.e., interpretation based on the surrounding world and previous context), and non-literal constructions, e.g., metaphors, idioms, sarcasm, and humor. This again entails harnessing rich knowledge sources as well as developing expressive machine learning models, preferably combining structured and neural embedding methods. A second related goal is to develop methods that can generate human-like language with convincing properties such as being polite and interesting (e.g., metaphorical or humorous). I have some initial projects in this direction.

A third requirement is to be able to hold continuous dialog. This is another direction I am pursuing by building upon my Q&A and language generation research. Such a dialog model will learn a regularly-updated knowledge base of the current conversation history and generate relevant answers as well as questions over time. I plan to further enhance these dialog models by incorporating coherence via topic matching, goal-based reinforcement learning, and acoustic and visual cues from the speaker and the environment. Finally, we also need to be able to interpret language from any domain. I plan to continue and expand upon my domain adaptation and transfer learning efforts, e.g., via all-domain Web data, mapping unseen words to task-trained embedding spaces [MBGL15], universal embeddings with transferable models, and zero-shot learning (i.e., having test classes unseen during training).

I am also excited to continue collaborating on the various grounded applications of conversational models. E.g., question answering and dialog for navigation and manipulation (e.g., [WBW16]) in a visual environment will help realize impactful applications like conversational autonomous cars, conversational wheel chairs, and personal healthcare assistants. Here, language will also be useful as an easy, cheap, and robust medium for real-time, correction-based feedback and as discriminative descriptions of similarities and differences when transferring from seen to unseen scenarios.
References


