Natural language is the de-facto standard for storing and transmitting information. Textual data stores information about people and places (e.g., Obama was born in Hawaii), facts about science and engineering (e.g., Ice is frozen water), or simply common-sense facts (e.g., Some mushrooms are poisonous). With the internet, we have unprecedented access to a huge – and growing – amount of text. This presents an immediate practical concern: it becomes infeasible for humans to digest and catalog this influx of information. Attempts at manually extracting knowledge (e.g., Freebase) have led to knowledge bases which are both woefully incomplete and quickly become outdated. In medicine, half of the medical school curriculum becomes obsolete within 5 years of graduation,\(^1\) requiring constant updating. MEDLINE counts 800,000 biomedical articles published in 2013 alone.\(^2\) In academia, studies show that up to 90% of papers are never cited, suggesting that many are never read. In my Ph.D. work, I have both created practical real-world systems to extract information from very large text corpora \([1, 2, 3, 4]\), and made research contributions in theoretically sound inference methods that can leverage unstructured text for knowledge extraction \([5, 6, 7]\).

A key challenge in this research direction is the ability to use a large corpus of plain text to query facts and answer questions which are not verbatim expressed in the text. For example, a statement "the cat ate a mouse" should support even lexically dissimilar queries like "carnivores eat animals" and reject logically contradicted queries (like "no carnivores eat animals"). Or, a long sentence from Wikipedia may include additional information besides the part that supports the query. This contrasts with information retrieval (IR), which simply retrieves lexically similar passages.

A natural formalism for addressing this challenge is **natural logic** – a proof theory over the syntax of natural language. The logic offers computational efficiency and eliminates the need for semantic parsing and domain-specific meaning representations, while still warranting most common language inferences (e.g., negation). Furthermore, the inferences warranted by the logic tend to be the same inferences that are cognitively easy for humans – that is, the inferences humans assume a reader will effortlessly make.

My thesis work explores how to leverage natural logic as a formalism for extracting knowledge not only when it is verbatim written in text, but also when it is implied by some statement in the text. I created a system to (1) extract common-sense knowledge from a large corpus of unannotated text via a search procedure over a soft relaxation of natural logic; (2) simplify complex syntactic structures into maximally informative atomic statements, and (3) incorporate an entailment classifier into this search to serve as an informed backoff. I describe each of the components below, and show that the full system can solve 4\(^{th}\) grade multiple-choice science exam questions.

**Natural Logic for Common Sense Reasoning**

In addition to being practically useful, the ability to extract and make use of knowledge expressed in a large body of text is a core feature of general-purpose intelligent systems which can reason across tasks. A key difference between current NLP systems and the human mind is the latter’s

---

\(^1\) http://uvamagazine.org/articles/adjusting_the_prescription/

\(^2\) https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html
ability to draw upon a huge amount of latent knowledge when solving tasks. As a humorous (real) example, the correct meaning of the headline “Iraqi head seeks arms” relies less on the syntax and local context of the sentence than on our background knowledge about the government of Iraq. Similarly, one reviewer for Pixar’s movie, The Good Dinosaur, wrote: “It’s the second best movie Pixar released [in 2015].” The negative sentiment here comes entirely from the knowledge that Pixar only released two movies in 2015. A central research contribution of my thesis has been the creation of systems that extract or infer from text the sort of background and common-sense knowledge which is useful for this type of general-purpose reasoning.

During my Ph.D. I created a system [5] for inferring the truth or falsehood of common-sense facts from a very large knowledge base of statements about the world. For example, if a premise “the cat ate a mouse” is present in the knowledge base, we should conclude that a hypothesis “no carnivores eat animals” is false. The system constructs a search problem for each queried hypothesis over relaxed natural logic inferences: the surface form of the hypothesis is allowed to mutate until it matches one of the facts in the knowledge base. These mutations correspond to steps in a natural logic proof; a learned cost for each mutation corresponds to the system’s confidence that the mutation is indeed logically valid (e.g., mutating to a hypernym has low cost, whereas nearest neighbors in vector space has high cost). This amounts to high-performance fuzzy theorem prover over an arbitrarily large premise sets. In my experiments, there are 270M premises in the knowledge base and the the system visits 1M candidates per second.

An illustration of a search from the query “no carnivores eat animals” is given below, with the appropriate natural logic relation annotated along the edges:

\[
\text{No carnivores eat animals?} \quad \equiv \quad \text{No animals eat animals} \quad \equiv \quad \text{No animals eat things} \quad \equiv \quad \ldots
\]

This framing of the problem has a number of advantages: unlike most approaches in textual entailment, it can scale to arbitrarily large knowledge bases. Unlike most approaches in relational inference (e.g., Markov Logic Networks), the runtime of the system decreases as the size of the knowledge base grows, since we can run a shallower search in expectation. From the other direction, unlike information retrieval approaches, we remain sensitive to a notion of entailment rather than simply similarity – for example, we can detect false facts in addition to true ones. In an empirical evaluation, we show that we can recover 50% of common sense facts from a subset of ConceptNet at 90% precision – 4x the recall of querying the knowledge base directly.

Simplifying Complex Sentences

A common motif in extracting information from text is the value in converting a complete sentence into a set of atomic propositions. I built a system [6] to extract atomic propositions (e.g., “Obama
was born in Hawaii”) from longer, more syntactically difficult sentences (e.g., “Born in Hawaii, Obama attended Columbia”) by recursively segmenting a dependency tree into a set of self-contained clauses expressing atomic propositions. These clauses are then maximally shortened to yield propositions which are logically entailed by the original sentence, and also maximally concise. For instance, the statement “anchovies were an ideal topping for Italian sailors” yields “anchovies are a topping.”

In addition to being a component in the reasoning engine described in the previous section, we can directly use this method for Open Information Extraction (Open IE) – a flavor of relation extraction where the relation, subject, and object are all allowed to be open domain plain-text strings. On a NIST-run knowledge base population task, we show that our system outperforms UW’s 4th generation Open IE system by 3 F₁.

Solving 4th Grade Science
A key property of natural logic is its ability to interface nicely with statistical models which featurize the surface form of a sentence. In addition to handling complex premises, I extended the inference system to allow for inexact matches against the premises in the knowledge base [7]. This can be thought of as an evaluation function – akin to gameplaying search – which produces a score at each search state for how likely a simple statistical classifier thinks that the state is supported by any premise. This allows us to provide some judgment for every query hypothesis (in contrast to the 50% coverage in the original system over common-sense facts), while still adhering to logically valid inferences where possible and still detecting negation. We evaluate this complete system on 4th grade science exams, and show that we outperform prior work, a strong information retrieval baseline, and a standalone version of the evaluation function. We can achieve a final score of 74% on our practice test, and 67% on unseen test questions.

Future Directions
Modern AI systems successfully solve isolated tasks, but there has been little success in developing systems that elegantly solve multiple varied tasks. My future research will focus on leveraging extracted knowledge to create intelligent systems which generalize across domains, and on creating methods to collect and interpret the type of knowledge which is useful for these general systems. This is, in some ways, a re-thinking of the conventional paradigm where the purpose of a supervised training set is to create systems to solve a single task. Rather, the purpose of supervised data should be to teach a system how to leverage its existing knowledge about the world to a new domain.

More near-term, a natural direction of future work would be to move beyond factoid queries – common-sense or otherwise – towards methods to solve, e.g., word algebra problems or procedural questions (“how do I build a boat?”) without appealing to domain-specific knowledge representations. In some cases, this may take the form of augmentations to natural logic: for example, a propositional extension to natural logic to handle multiple premises and conditional statements. In other cases, this may take the form of vector space models which encode the semantics of a premise set and query in a reusable representation. My Ph.D. work on natural logic for question answering prepares me well for this line of research; my ongoing research on evaluating vector-space models lays the groundwork for research into how these sorts of models can encode semantics in vector-space.

Lastly, I intend to continue developing practical systems, and inferring research insights from these systems. These real-world systems highlight constraints that are likely elided in well-defined core NLP tasks. For example, the importance of operating with limited training data, and the importance of operating under severe class imbalance – usually, an overabundance of negative examples. My experience building systems for the TAC-KBP competition over the last three years suggests natural research directions for solving these challenges. For addressing limited training
data, self-trained and/or bootstrapped methods (like Stanford’s winning system in 2015) show promise for replacing traditional distantly supervised approaches which rely on indirect supervision from hand-curated knowledge bases. The problem of class imbalance in turn suggests research into machine learning algorithms operating under this sort of class imbalance. For instance, creating machine learning algorithms which optimize a [relaxation of] the F-measure directly, rather than training against likelihood.

References


