LEARNING OPEN DOMAIN KNOWLEDGE FROM TEXT

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Gábor György Angeli

May 2016

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Christopher D. Manning) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Percy Liang)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Dan Jurafsky)

Approved for the Stanford University Committee on Graduate Studies

Preface

This thesis describes a new technique for learning open-domain knowledge from unstructured web-scale text corpora, making use of a probabilistic relaxation of natural logic – a logic which uses the syntax of natural language as its logical formalism. We begin by reviewing the theory behind natural logic, and propose a novel extension of the logic to handle propositional formulae.

We then show how to capture common sense facts: given a candidate statement about the world and a large corpus of known facts, is the statement likely to be true? This is treated as a search problem from the query statement to its appropriate support in the knowledge base over valid (or approximately valid) natural logical inference steps. This approach achieves a 4x improvement at retrieval recall compared to lemmatized lookup, maintaining above 90% precision.

We then extend the approach to handle longer, more complex premises by segmenting these utterance into a set of atomic statements entailed through natural logic. We evaluate this system in isolation by using it as the main component in an Open Information Extraction system, and show that it achieves a 3% absolute improvement in F1 compared to prior work on a competitive knowledge base population task.

Finally, we address how to elegantly handle situations where we could not find a supporting premise for our query. To address this, we create an analogue of an evaluation function in gameplaying search: a shallow lexical classifier is folded into the search program to serve as a heuristic function to assess how likely we would have been to find a premise. Results on answering 4th grade science questions show that this method improves over both the classifier in isolation, a strong IR baseline, and prior work.

Acknowledgements

Contents

List of Tables

List of Figures

Chapter 1

Conclusions

In this dissertation, I have explored methods to leverage natural logic for extracting open domain knowledge from large-scale text corpora. Unlike fixed-schema knowledge bases, this approach allows querying arbitrary facts. Unlike open-domain knowledge bases – such as Open Information Extraction approaches – this approach (1) does not limit the representation of facts to subject/relation/object triples; and (2) allows for rich inferences to be made so that we can find facts which are not only in the knowledge base, but also *inferred* by some known fact. From the other direction, unlike shallow information retrieval approaches, which also operate over large text corpora, the approach in this dissertation is robust to logical subtleties like negation and monotonicity. We have applied this method to three areas: we have shown that we can predict the truth of common-sense facts with high precision and substantially higher recall than using a fixed knowledge base. We have shown that we can segment complex sentences into short atomic propositions, and that this is effective for a practical downstream task of knowledge base population. Lastly, we have shown that we can incorporate an *evaluation function* encoding a simple entailment classifier, and that the hybrid of this evaluation function and our natural logic search is effective for question answering.

In Chapter **??** I reviewed the current theory behind natural logic as a logical formalism. We reviewed a theory of denotations over lexical items, a notion of monotonicity over arguments to quantifiers and other functional denotations, and then introduced Monotonicity Calculus as a logic to reason over these monotonic functions. We then introduced *exclusion* to deal with antonomy and negation, and showed how we can extend Monotonicity Calculus to incorporate this additional expressive power. Lastly, I introduced a brief sketch of a propositional natural logic, which would allow for jointly reasoning about multiple natural language utterances (for instance, the disjunctive syllogism). I encourage future research into this propositional natural logic, and further research into the use of natural logic in place of conventional (e.g., first-order) logics for language tasks.

In Chapter **??**, I introduce NaturalLI – a large-scale natural logic reasoning engine for common sense facts. I show that natural logic inference can be cast as a search problem, and that the *join table* of MacCartney and Manning [1] can be more elegantly represented as a finite state machine we transition through during search. I show that we can not only perform strictly warranted searches, but also learn a confidence for likely valid mutations; this allows the system to improve its recall by matching not only strictly valid premises, but also likely valid premises that it finds through the search. I show that our system improves recall by 4× over lemmatized knowledge base lookup when assessing whether commonsense facts are true given a source corpus of 270 million unique short propositions.

In Chapter **??**, I move from short propositions to longer sentences, and introduce a method for segmenting and trimming a complex sentence into the types of short utterances that NaturalLI can operate over. This is done in two steps: First, complex sentences are broken into clauses, where each clause expresses one of the main propositions of the sentence (alongside potentially many additional qualifiers). This is done by casting the problem as a search task: as we search down the dependency tree, each edge either corresponds to a split clause (possibly interpreting the subject / object of the governor as the subject of the dependent), or the search is told to stop, or the search is told to continue down that branch of the tree but not to split off that clause. These clauses are then maximally shortened according to valid natural logic mutations to yield maximally informative atomic propositions. These propositions can then either be used as propositions for a system like NaturalLI, or segmented further into OpenIE relation triples. Using segmented triples from this system outperforms prior OpenIE systems on a downstream relation extraction task by 3 F_1.

In Chapter **??**, we extend NaturalLI to operate over dependency trees and incorporate the method for creating atomic propositions from Chapter **??** to allow NaturalLI to operate over a more complex premise set. In addition, we introduce a method for combining a shallow entailment classifier with the more formal NaturalLI search. At each step of the search, this classifier is run against a set of candidate premises; if any of these search states get close enough to a candidate premise according to the classifier, the fact is taken to be possibly true. This behaves as a sort of evaluation function – akin to evaluation functions in gameplaying algorithms – and allows for both (1) improving the recall of NaturalLI, and (2) creating a reasonable confidence value for likely entailment or contradiction even when the query cannot be formally proven or disproven. I show that this method outperforms both strong IR baselines and prior work on answering multiple choice 4$^{\text{th}}$ grade science exam questions.

There are a number of interesting and natural directions for future work in this area, which I will briefly discuss below:

Propositional Natural Logic Section **??** sketches a simplistic propositional natural logic, based around a simple proof theory. However, this is presented both without a formal proof of consistency or completeness, and without an associated model theory. Furthermore, I have skirted issues of proper quantification and other first-order phenomena. It would clearly be benificial to the NLP community to have a natural logic which can operate over multiple premises, and more work in this area is I believe both useful and exciting.

Downstream Applications This dissertation has presented a means of inferring the truth or falsehood of common-sense facts, but has only scratched the surface of downstream applications which can make use of this information. There is I believe an interesting avenue of research which attempts to improve core NLP algorithms beyond what can be obtained with statistical methods by leveraging the common-sense knowledge acquired from large unsupervised text corpora. For example, perhaps a parser which is more aware of facts about the world could correctly disambiguate prepositional attachments (e.g., *I ate the cake with a fork / cherry*).

Natural Logic for Cross-Domain Question Answering The applications in this dissertation have focused on factoid-style true/false queries. However, much of question answering is either (1) non-factoid (e.g., procedural) questions, or (2) requires finding a textual answer to the question (e.g., *Who is the president of the US?)*. Extending NaturalLI to handle these questions is a potential means of creating a truly cross-domain question-answering system. If all premises are encoded in text, and all questions are given in text, then there is no notion of a schema or domain-specific model / named entity tag set / etc. which would limit the scope of questions that could be asked of the system. For the first time, the same system could be asked both what color the sky is, and where Barack Obama was born.

I hope that this dissertation can inspire research in the direction of open-domain, broad coverage knowledge extraction, and encourage researchers to consider natural logic and its extensions as the foundation for storing and reasoning about this sort of knowledge. As humans, we have chosen language as the means we store and represent knowledge, and I believe intelligent computers should do the same.

Bibliography

[1] Bill MacCartney and Christopher D Manning. Modeling semantic containment and exclusion in natural language inference. In *Coling*, 2008.