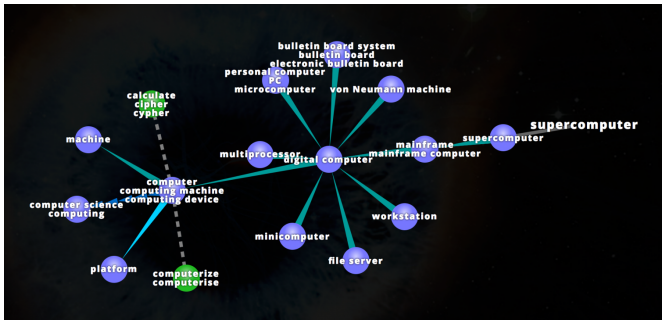
The future of search engines: Researchers combine artificial intelligence, crowdsourcing and supercomputers

3 August 2017



WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. Researchers from The University of Texas at Austin developed a method to incorporate information from WordNet into informational retrieval systems. Credit: Visuwords

How do search engines generate lists of relevant links?

The outcome is the result of two powerful forces in the evolution of information retrieval: artificial intelligence—especially natural language processing—and crowdsourcing.

Computer algorithms interpret the relationship between the words we type and the vast number of possible web pages based on the frequency of linguistic connections in the billions of texts on which the system has been trained.

But that is not the only source of information. The semantic relationships get strengthened by professional annotators who hand-tune results—and the algorithms that generate them—for topics of importance, and by web searchers (us) who, in our clicks, tell the algorithms which connections are the best ones.

Despite the incredible, world-changing success of this model, it has its flaws. Search engine results are often not as "smart" as we'd like them to be, lacking a true understanding of language and human logic. Beyond that, they sometimes replicate and deepen the biases embedded in our searches, rather than bringing us new information or insight.

Matthew Lease, an associate professor in the School of Information at The University of Texas at Austin (UT Austin), believes there may be better ways to harness the dual power of computers and human minds to create more intelligent information retrieval (IR) systems.

Combining AI with the insights of annotators and the information encoded in domain-specific resources, he and his collaborators are developing new approaches to IR that will benefit general search engines, as well as niche ones like those for medical knowledge or non-English texts.

This week, at the Annual Meeting of the Association for Computational Linguistics in Vancouver, Canada, Lease and collaborators from UT Austin and Northeastern University presented two papers describing their novel IR systems. Their research leverages the supercomputing resources at the Texas Advanced Computing Center.

ANNOTATOR CONSENSUS AND ATTRIBUTIONS PROVIDE RATIONALES FOR SEARCH RESULTS

In one paper, led by Ph.D. student An Nguyen, they presented a method that combines input from multiple annotators to determine the best overall annotation for a given text. They applied this

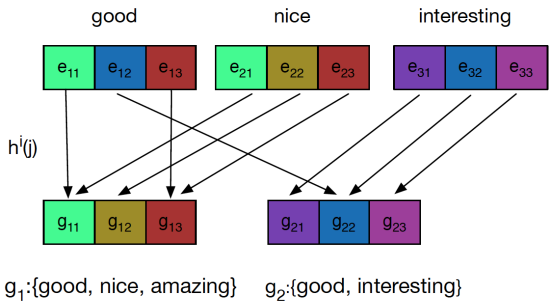
method to two problems: analyzing free-text research articles describing medical studies to extract details of each study (e.g., the condition, patient demographics, treatments, and outcomes), and recognizing named-entities—analyzing breaking news stories to identify the events, people, and places involved.

"An important challenge in natural language processing is accurately finding important information contained in free-text, which lets us extract into databases and combine it with other data in order to make more intelligent decisions and new discoveries," Lease said. "We've been using crowdsourcing to annotate medical and news articles at scale so that our intelligent systems will be able to more accurately find the key information contained in each article."

Such annotation has traditionally been performed by in-house, domain experts. However, more recently crowdsourcing has become a popular means to acquire large labeled datasets at lower cost. Predictably, annotations from laypeople are of lower quality than those from domain experts, so it is necessary to estimate the reliability of crowd annotators and also aggregate individual annotations to come up with a single set of "reference standard" consensus labels.

Lease's team found that their method was able to train a neural network—a form of AI modeled on the human brain—so it could very accurately predict named entities and extract relevant information in unannotated texts. The new method improved upon existing tagging and training methods.

The method also provides an estimate of each worker's label quality, which can be transferred between tasks and is useful for error analysis and intelligently routing tasks—identifying the best person to annotate each particular text.



An example of grouped partial weight sharing, here with two groups. Lease's team stochastically selects embedding weights to be shared between words belonging to the same groups. Weight sharing constrains the number of free parameters that a system must learn, increases the efficiency and accuracy of the neural model, and serves as a flexible way to incorporate prior knowledge, combining the best of human knowledge with machine learning. Credit: Ye Zhang, Matthew Lease, UT Austin; Byron C. Wallace, Northeastern University

LEVERAGING EXISTING KNOWLEDGE TO CREATE BETTER NEURAL MODELS

The group's second paper, led by Ph.D. student Ye Zhang, addressed the fact that neural models for natural language processing (NLP) often ignore existing resources like WordNet—a lexical database for the English language that groups words into sets of synonyms—or domain-specific ontologies, such as the Unified Medical Language System, which encode knowledge about a given field.

They proposed a method for exploiting these existing linguistic resources via weight sharing to improve NLP models for automatic text classification. For example, their model learns to classify whether or not published medical articles describing clinical trials are relevant to a well-specified clinical question.

In weight sharing, words that are similar share some fraction of a weight, or assigned numerical value. Weight sharing constrains the number of free parameters that a system must learn, thereby increasing the efficiency and accuracy of the neural model, and serving as a flexible way to incorporate prior knowledge. In doing so, they combine the best

of human knowledge with machine learning.

"Neural network models have tons of parameters and need lots of data to fit them," said Lease. "We had this idea that if you could somehow reason about some words being related to other words a priori, then instead of having to have a parameter for each one of those word separately, you could tie together the parameters across multiple words and in that way need less data to learn the model. It would realize the benefits of deep learning without large data constraints."

They applied a form of weight sharing to a sentiment analysis of movie reviews and to a biomedical search related to anemia. Their approach consistently yielded improved performance on classification tasks compared to strategies that did not exploit weight sharing.

"This provides a general framework for codifying and exploiting domain knowledge in data-driven neural network models," say Byron Wallace, Lease's collaborator from Northeastern University. (Wallace was formerly also a faculty member at UT Austin and became a frequent user of TACC as well.)

Lease, Wallace and their collaborators used the GPUs (graphics processing units) on the Maverick supercomputer at TACC to enable their analyses and train the machine learning system.

"Training neural computing models for big data takes a lot of compute time," Lease said. "That's where TACC fits in as a wonderful resource, not only because of the great storage that's available, but also the large number of nodes and the high processing speeds available for training neural models."

In addition to GPUs, TACC deploys cutting-edge processing architectures developed by Intel to which the machine learning libraries are playing catch up, according to Lease.

"Though many deep learning libraries have been highly optimized for processing on GPUs, there's reason to think that these other architectures will be faster in the long term once they've been optimized as well," he said.

"With the introduction of Stampede2 and its many core infrastructure, we are glad to see more optimization of CPU-based machine learning frameworks," said Niall Gaffney, Director of Data Intensive Computing at TACC. "Project like Matt's show the power of machine learning in both measured and simulated data analysis."

Gaffney says that in TACC's initial work with Caffe—a deep learning framework developed at the University of California, Berkeley, which has been optimized by Intel for Xeon Phi processors—they are finding that CPUs have roughly the equivalent performance for many AI jobs at GPUs.

"This can be transformative as it allows us to offer more nodes that can satisfy these researchers as well as allowing HPC users to leverage AI in their analysis phases, without having to move to a different GPU enabled system," he said.

By improving core natural language processing technologies for automatic information extraction and the classification of texts, web search engines built on these technologies can continue to improve.

Lease has received grants from the National Science Foundation (NSF), the Institute of Museum and Library Services (IMLS) and the Defense Advanced Research Projects Agency (DARPA) to improve the quality of crowdsourcing across a variety of tasks, scales, and settings. He says that though commercial web search companies invest a lot of resources to develop practical, effective solutions, the demands of industry lead them to focus on problems with commercial application and short-term solutions.

"Industry is great at looking at near-term things, but they don't have the same freedom as academic researchers to pursue research ideas that are higher risk but could be more transformative in the long-term," Lease said. "This is where we benefit from public investment for powering discoveries. Resources like TACC are incredibly empowering for researchers in enabling us to pursue high-risk, potentially transformative research."

Provided by University of Texas at Austin

APA citation: The future of search engines: Researchers combine artificial intelligence, crowdsourcing and supercomputers (2017, August 3) retrieved 19 November 2019 from https://phys.org/news/2017-08-future-combine-artificial-intelligence-crowdsourcing.html