For most of us words are how we communicate and, mostly, we don’t give much thought to them beyond that. Sometimes they occur to us spontaneously, in delight or sorrow. Other times we use them in sentences carefully crafted to express nuances of thought. For computer scientists in the field of natural language processing (NLP), however, words are also data, and there’s plenty to go around.

The World Wide Web has become an expanding, limitless repository of text, billions and billions of words, and for Noah Smith and his colleagues it’s a treasure trove — to sift through, ask questions, test better ways to translate languages, and sometimes to make forecasts about collective human behavior. “The general area we work in is natural language processing,” says Smith, associate professor in the Language Technologies Institute at Carnegie Mellon University, one of the world’s leading centers in using computers to solve language-related problems.

“You can imagine anything from more intelligent search engines to answer your questions,” says Smith, “to systems that translate automatically from one language to another.” In late 2010, while Blacklight, PSC’s newest supercomputer (see p. 4), was undergoing shakedown, Smith and his colleagues were experimenting with the new system, work that bore fruit — four papers within six months, in diverse areas of NLP.

“Blacklight has been a very useful resource for us,” says Smith. “We can incorporate deeper ideas about how language works, and we can estimate these more complex models on more data.” Blacklight’s shared memory has been crucial, he observes, because his large-scale models use iterative algorithms that look at the same data over and over again. “Shared memory lets us use many processors in parallel without having to worry about the overhead of passing data over the network or moving model information around.”

A recurring theme of Smith’s modeling, and one of the reasons Blacklight’s shared memory has opened doors in his work, is an “unsupervised” approach to text data, as exemplified by his group’s recent work on word alignment, an important component of automated language translation.
Traditional approaches to NLP have relied in large part on unannotated text, meaning help from humans—in text searches, for instance, a set of keywords to help identify the input of a text, or in automated translation, for instance, links between words in English sentences and their Chinese translations. In general, Smith’s work contrasts with this. “Our approach,” he says, “is to discover the structures of interest from large amounts of unannotated text data.”

“Unsupervised” is a general term for this approach, which allows the model to start from scratch and sift through real-world text, without expensive expert annotations, to build connections in the data for the task it undertakes to accomplish. This has the advantage of not limiting tasks based on whether annotated data is available and, further, offers the potential to uncover linguistic connections less biased by previous thinking.

With translation in particular, observes Smith, unsupervised approaches have become more feasible as huge amounts of text have accumulated on the web, with the United Nations as a prime example.

“Everything that happens at the UN has to be translated by expert translators into the major languages that people speak around the world. The result is what we call “parallel documents”—text in English and corresponding text in Chinese, Arabic and other languages. This data is freely available to everybody.”

For Smith, these parallel documents—and the availability of Blacklight—made it feasible to try an experiment in word alignment, the part of translation in which a model builds statistically-based maps of connections between words in two languages. Smith’s project built alignments between English and Czech, English and Chinese and English and Ordu—languages very different from each other. Czech, for instance, is a language with a complex morphological structure, in which the verb changes depending on whether the subject of the sentence is masculine or feminine, or singular or plural.

“The more data you have,” says Smith, “the better you can do with word alignment, but likewise it becomes more and more expensive computationally.” In some recent work, for this reason, word alignment has relied on human experts to draw links between a subset of words in the two languages as a starting point for the model to train itself. “It gives a nice clean gold standard of what the alignments look like. The problem, of course, is human intervention is costly and you can do it only for a small amount of data.”

With Blacklight’s ability to hold large amounts of data in shared memory, Smith’s unsupervised word-alignment model in all three test cases outperformed other unsupervised alignment models. By a variety of measures, furthermore, including using the model in automated translation programs, his model outperformed supervised approaches, which hadn’t been accomplished in previous natural language modeling.

“Thanks to Blacklight,” says Smith, “we were able to train an unsupervised model that outperforms the supervised approaches. This is because we were using massive amounts of data, along with some sophisticated statistical modeling techniques that had been applied before only in supervised cases. It was an obvious gap, and it was because the amount of computing was prohibitive that people hadn’t tried this before.”

FORECASTING SCHOLARLY IMPACT

How much can the word-data of many texts make it possible to predict how people will respond to other similar texts? For a few years, Smith and his colleague Bryan Routledge from CMU’s Tepper School of Business have explored “text-driven forecasting”—the ability of statistical modeling to discern features in text that can reliably forecast human responses.

With this relatively new application of NLP, they have, for instance, shown that the prevalence of words of identifiable characteristics and flavors in movie reviews can predict with statistical validity whether the movie will make a profit on opening weekend. Another of their projects found that language features of corporate financial reports can predict the volatility of that corporation’s stock price over the next year.

With Blacklight, Smith and Routledge and their collaborators have been testing a more ambitious possibility. Can you forecast the scholarly impact of scientific articles—with “impact” measured as how often an article is cited in other papers or downloaded from the web—from its text content?

For this project, the researchers used two large datasets of scientific papers. The National Bureau of Economic Research (NBER) provides downloadable data on papers, from approximately 1,000 economists, posted in the NBER online archive. A second dataset comprised papers and related citation data from the Association for Computational Linguistics (ACL).

“This gives us paired data,” says Smith. “We have documents, the research papers, and a response that came in later. How much did people download a paper, or cite it? It’s a way to measure a response within the community.”

Unlike translation, this is a supervised problem because the model learns from what happened in the past. For the NBER data, for instance, the model trained on 50 years worth of papers, 1999 to 2009, to learn relationships between text content and the number of downloads. The researchers then tested the model’s ability, based on what it learned from the historical data, to predict download response for papers held out from the training database.

Compared to other data associated with scientific papers—such as an author’s name, the category of topic, what journal, the text—based predictions were significantly more accurate. “The accuracy went up when we used these newer techniques,” says Smith. “Nobody had framed this problem quite this way before.”

Despite its not being an unsupervised model, this project was computationally demanding, says Smith, because of its high number of dimensions. “We’re looking at a very large set of clues from the input text about the future.” And the model is “discriminative”—designed to learn to do a specific task and get it right. “The estimation procedure is computationally expensive, and we’ve run a number of experiments, in which we’ve divided the data in different ways, to see what ideas work.”

Beyond the impact predictions, this model also tracks how scholarly impact of a paper may change over time, which Smith calls “time series changes.” He compares this dimension of the model to related approaches that track “term frequency” in texts, a count of how often a word appears in a given time period. Term frequency by itself, Smith believes, is less revealing than a measurement, such as downloads or citations, that connects term frequency to impact. “While the time series is less computationally expensive than counting words, it gives a fuller understanding of the community response.”

The implications of this kind of text-driven forecasting extend to such possibilities as helping busy people make intelligent choices of what to read. “Can we learn to take this messy and unstructured data and with computers turn it into some deeper meaning representation?” asks Smith. “A human couldn’t read all the papers that come out in a year on the NBER website. They can’t read fast enough. But with this kind of model we can look and do analysis on that large amount of data and come up with trends that tell you what’s going on.”

“These estimation procedures,” he continues, “become expressed in very complicated algorithms, and it can take multiple graduate courses for people to understand how they work, even on one processor. With real data scenarios, this becomes almost frighteningly expensive from a computational standpoint, and people won’t touch it. Having an architecture like Blacklight is what makes this work possible.”

MORE INFO: http://www.pitt.edu/~langlab