All About Natural Language Processing

Natural Language Processing is the machine handling of written and spoken human communications. Methods draw on linguistics and statistics, coupled with machine learning, to model language in the service of automation.

What Good Is NLP for Business?

There are myriad applications. Every business process (or personal need) that involves speech or text — with volume, velocity, or complexity sufficient to push you to seek automated assistance — can benefit from Natural Language Processing. Let's review, systematically, what NLP can do for you. Here are 22 facets, with examples that illustrate both implementations and R&D initiatives. Let's start with computing's second-oldest application, search, and then explore NLP uses from everyday to analytical to unusual.

Information Extraction and Search

If all the world's information were neatly binned in database fields, we wouldn't need search. Information retrieval would be nothing more than queries. But instead, notionally, 80 percent of business-relevant information originates in unstructured form, primarily text. The vast majority of that text is "natural language" (as opposed to formal language, found for instance in a computer programming or algebraic equation). Google and Bing and other search systems use NLP to extract terms from text (#1) to populate their indexes and to parse search queries (#2). Those terms may include "named entities" such as people, companies, brands, ticker symbols, and places. Other features of interest may include dates, addresses, URLs, and the like; NLP will automate extraction of pattern-identified information (#3) and extraction of attributes associated with terms (#4) whether factual or subjective: expensive watch, black car, 4.6 kg fish.

The more advanced engines apply NLP to identify relationships (#5) ("this is a that") in order to build their knowledge graphs. NLP feeds the computational knowledge engines behind Apple Siri , Wolfram Alpha, and Google Now as well as resources for your own lexical analyses such as Lexaltyics' Concept Matrix, built via NLP application to the Wikipedia dataset to identify "concept topics" and "facets" as well as associated sentiment. According to Lexalytics CEO Jeff Catlin, "these features allow users to easily build classifiers for very broad topics as well as roll-up opinions into buckets of similarity." Pingar's Taxonomy Generator is another take on the same idea: Use NLP methods to build a knowledge structure for later application to search, classification, and other business information-management needs.

Concepts, Topics, Sentiment, and Similarity, Plus Notes on Methods

"Buckets of similarity": Those would be categories determined by an analyst or via statistical clustering. Classification is the act of placing cases into categories based on attributes or into clusters based on best fit. Classification (#6) is part of the NLP task,

whether it involves grouping terms or documents. One variety of term grouping involves creating conceptual classes, for instance "vehicle manufacturers" from Fiat, Ford, General Motors, Nissan, Toyota, etc. Another variety involves coreference — multiple ways of referring to a given thing; to illustrate, "Barack H. Obama is the 44th President of the United States. His story is the American story… President Obama was born in Hawaii" refers to a given person in four underlined ways, one of them via a pronoun ("his") that refers to that person only in context.

Want to see real-world entity extraction and coreference? Try Language Computer Corporation's Cicero system demo. Process the Web page where I found the above lines, http://www.whitehouse.gov/administration/president-obama. Click on one of the "he" or "his" occurrences in the marked-up text and you'll see that these pronouns have been correctly resolved to "President Obama."

I suppose I'll grant numbers to concept extraction (#7) and to topic extraction (#8), that is, to information extraction (per the previous subsection) that involves abstraction. Sentiment is also abstract, although sentiment analysis (#9) can be characterized (in a very simplistic way) as simply another classification problem, whether involving the usual positive/negative/neutral categories, more nuanced emotion categories (e.g., angry, happy, sad), or intent signals (e.g., to buy, sell, renew, cancel). Visit the Web site of text-analytics mavens Daedalus for an online sentiment classification demo. The Nerily online demo will extract a variety of other text features.

Sentiment analysis and opinion mining are central topics for me. I've written a lot about them, and I organize a twice-yearly conference, the Sentiment Analysis Symposium, next up May 8, 2013 in New York, preceded on May 7 by an optional, half-day Research & Innovation session and an optional, half-day Practical Sentiment Analysis tutorial. A disclosure: Lexalytics, cited above and later in this article, is a sponsor.

All of this information extraction is what makes NLP a key asset for text analytics, which models and structures the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation. (That's a definition I wrote back in 2007, in a TechWeb article, that made its way to Wikipedia.)

I'll digress to explain that you can automate human handling of many Natural Language Processing tasks, via a crowdsourcing using CrowdFlower for "human-powered sentiment analysis" and other systems built on platforms such as Amazon Mechanical Turk. Also, you can also extract sentiment and other information by analyzing non-textual sources that range from transaction records to images and speech.

We'll get back to speech bit later. For now, I'll cite one last function related to classification and similarity, and then let's change tacks. That last for-now function is plagiarism detection (#10), essentially passage-similarity evaluation across retrieved text, as explained on the PAN-13 conference site, with a bit of data and source code to get the Python programmers among you started. (PAN = Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection. I guess PAAINDD is kind of awkward as a an acronym.)

Spelling, Grammar, and Style

Want to right gud? Lucky for you: NLP is built into your favorite word-processing software. Spell check (#11) is NLP at its most basic. Spell check will flag a word that's not in the dictionary and maybe suggest corrections. If you have ever written a document with Microsoft Word (or OpenOffice, Google Docs or any of countless other authoring environments), you've seen a spelling checker. But spell check won't identify the two errors in "I went their at tree o'clock." Try that sentence in JSpell or at SpellCheck.net as proof. For syntactic errors, you need a grammar checker. And how does a machine check grammar?

A linguistic approach to grammar checking might involve resolving parts of speech, via sentence diagramming (#12), illustrated here, part-of-speech tagging (#13), as seen in a Univ. of Illinois demo system, or via study of syntactic relations (#14), à la this Connexor demo. (I'm a bit behind myself, actually. Syntactic parsing is one method of discerning relationships among entities, my #5, above.)

What do some of the tools out there think of my writing? I pasted the three-sentence paragraph above into one. It found "3 critical writing issues" — two alleged spelling errors and an accusation of wordiness — and said my writing is "weak, needs revision." (Free access doesn't provide detail so I'll withhold the tool's name.) Try some others: LanguageTool open source proofreading software (which I didn't find particularly useful, but you might) and Stilus from my friends at Daedalus.

Two more varieties of stylistic analysis to mention: Lymbix analyzes e-mail sentiment via the ToneCheck tool, and automated social-comment moderation is another interesting application, although I've been unable to identify an independent provider comparable to Adaptive Semantics, which the Huffington Post bought back in 2010.

Summarization and Translation

Text summarization (#15) is the first of several NLP functions I'll cite that involve both natural-language understanding (covered in #1 through #14) and natural-language generation. A summarizer has to understand the source text sufficiently to generate a shortened version that is faithful to the content and purpose of the original. Abstracting is a related function

Visionary researcher Hans Peter Luhn described an approach to automatic text abstracting in has April, 1958 IBM Journal paper, The Automatic Creation of Literature Abstracts: "Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the 'auto-abstract'."

Developer Andreas Gohr, at his SplitBrains.org site, provides a Web interface to Nadav Rotem's open source Open Text Summarizer code. Try it!

Machine translation (#16) is a wonderful NLP application. It doesn't require explanation; I'll just point you to Google translate where you can try it yourself. Note the automatic language identification (#17) feature.

Translation involves more than just rendering words from one language to another. Each language has its own syntax and idioms. A translator, whether a human or a machine, needs to make sense of the text provided and to make sense in the destination language. That is, like summarization, machine translation involves natural-language generation. So does the next example.

Question Answering

IBM Watson is the most prominent example of question answering (#17) at work: Information retrieval that produces usable guidance — situationally-relevant facts, in a form that reflects question context — to respond to a query. When Watson played Jeopardy, it formulated responses as questions; very different from how it will respond to medical-diagnostic challenges. I'll point you to an academic illustration, START from Boris Katz and associates at MIT, and also refer you to the EasyAsk and Inbenta Web sites for explanations how Q-A can work in general business contexts.

Speech Recognition

Let's recognize that speech is natural language too, and cite speech recognition (#18) and speech generation or synthesis (#19) as two more NLP functions.

Speech is more than just spoken text. It conveys genre, sentiment, mood, and emotion, detectable from word and sentence inflection (an interrogatory sentence — a question — is inflected up at the end) and from changes in speech volume and rapidity and other indicators. Check out a recent IEEE Spectrum podcast, Teaching Computers to Hear Emotions, an interview with University of Rochester Professor Wendi Heinzelman, and you'll hear what I mean.

You don't have to render the spoken word as text in order to make analytical use of it, also speech transcription (#20) certainly counts as an NLP function. Plenty of academic and industrial work has been done on phonological analysis, which examines sounds and sound patterns, and there are industrial systems that perform voice search (#21) on phonemes and patterns. If you'd like to see phonetic transcription in action, check out Daedalus's online demo.

On the flip side, text-to-speech (#22) — having the machine read to you with properly accented pronunciation, inflection, pacing, etc. — is another bit of NLP. Ivona, recently acquired by Amazon — the software is already used on the Kindle Fire — has a cool online demo that will read for you in a wide variety of languages and accents.

Building Blocks

Finally, a note on tools, on the bits and pieces of code you can apply to hobble together your own solution, and on learning more. A disclaimer however: I didn't intend, in this article, to systematically catalog available software and services, open source or other.

Cognitive linguist Christopher Phipps observes, in his Lousy Linguist blog, "luckily, the NLP field has matured into an open access friendly crowd, so there are lots of resources freely available." Phipps focuses on text understanding, and for that, there's no better catalog than Stanford University NLP's page, "Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources," although as Phipps cautions, it's not for newbies. I list a number of open source tools in a last-year blog article, What are the most powerful open-source sentiment-analysis tools? Two I didn't include there, because they're not optimized for sentiment (that article's topic), are Apache OpenNLP. and the Mallet machine-learning toolkit.

You also have at your disposal a plethora of service offerings that implement NLP, invokable via online APIs, most with free for either trials or limited use. Off the top of my head, there are: AlchemyAPI, Apicultur, Bitext, Clarabridge, ConveyAPI, OpenAmplify, Pingar, Saplo, Semantria (backed by the Lexalytics Salience engine), and Viralheat. Mashape lists many more. Capabilities, quality, and cost vary widely. Some do only entity or sentiment tagging while others do more-elemental text analysis. The Apicultur service and Jacob Perkins' Web API for Python NLTK, at text-processing.com, are examples of the latter. I'll withhold detail and judgments but maybe write them out in an article at some point, and I'm also not going to write now about install-yourself software options, which aren't as easy to simply try as a Web API.

As for learning more, other than via do-it-yourself, what better way than through an online course? Coursera has one going, taught by Michael Collins of Columbia University, and the videos and lecture materials from Christopher Manning's popular Stanford University course are available online. A third option is the Statistics.com course taught by Dr. Nitin Indurkhya, planned for a July 19 start.

Now That You Know

In business contexts, you're most often going to apply NLP in conjunction with collection, integration, and analysis of disparate forms of online, social, and enterprise data. All that text-and speech-extractable goodness I've been discussing: In today's world of heterogeneous big data, it doesn't stand on its own. This statement is true for business analytics – you get lift by applying and integrating an appropriate variety of methods and data — and it's also true for activities that seemingly don't involve non-textual or non-speech sources, for activities such as Web search. Even in those latter cases, smart, sense-making engines take into account your profile, location, past online/on-social activities, and social connections, in conjunction with NLP, to provide the best situationally-relevant results.

Natural-language processing can do many things for you. It's an essential tool for leading-edge analytics. Understanding is just the start.

Source text: http://breakthroughanalysis.com/2013/03/04/all-about-natural-language-processing/