NATURAL LANGUAGE PROCESSING

Peng Lai "Perry" Li[*]

CITE AS: 1 GEO. L. TECH. REV. 98 (2016)
https://perma.cc/S2AU-SKZL

INTRODUCTION

Natural Language Processing ("NLP") is a field of computer science, artificial intelligence, and computational linguistics. Computers operate on the foundation of if/then logic statements, whereas natural human language systems do not; NLP endeavors to bridge this divide by enabling a computer to analyze what a user *said* (input speech recognition) and process what the user *meant*. Because NLP is concerned with the interactions between computer and human (natural) languages, it has important applications in the advancement of artificial intelligence and machine learning, from digital assistant applications, such as Siri and Google Now, to machine translation and sentiment analysis. At the same time, the advancement of NLP requires gathering copious amounts of data from users, thereby raising important legal issues in data ownership, privacy, and security. This article briefly explains the process of NLP and highlights some important legal implications of the technology.

BRIEF HISTORY

NLP research began in the 1950s as the intersection of artificial intelligence and linguistics.[1] In what is known as "the Georgetown-IBM

[*] GLTR Staff Member; Georgetown Law, J.D. expected 2017; Iowa State University, M.Sc. 2006, B.S. 2004. © 2016, Peng Lai "Perry" Li.

[1] Prakash M. Nadkarni, *Natural Language Processing: An Introduction*, J AM MED INFORMATICS ASS'N 2011.

Experiment" of 1954, more than sixty Russian sentences were automatically translated into English by computers (with varying degree of success).[2] Nevertheless, progress in this field was slow for the following decades for two reasons. First, at the time Noam Chomsky, whose theory posits that all natural languages comprise hierarchies of grammars and adhere to a universal set of rules, greatly influenced linguistics; as a result, most NLP systems resembled complex decision trees based on numerous human-devised rules, despite natural language being much more nuanced in reality (*e.g.,* puns, metaphors, and homographs - identically spelled words with multiple meanings).[3] Second, computers of this era had limited processing power and could not undertake the less rule-based approach which was ultimately far more successful.

Beginning in the 1980s, however, advances in both linguistic theory and computational processing power led to NLP based on statistics and probability. This approach replaces deep structural analysis with simple approximation based on probability; more importantly, this method lets the computer learn the natural language on its own (so-called machine learning) by providing the computer with a large body of text (the corpus).[4] As a result, a few simple rules replaced the complex decision tree, and statistical analysis increase the accuracy of NLP. Statistical NLP is now the predominant NLP technology.

TECHNOLOGICAL PROBLEMS IN NLP

A NLP system needs to process the natural language in the following abstractions levels:[5]

- The phonetic or phonological level (*i.e.,* pronunciation);
- The morphological level, which deals with the smallest parts of words that carry meaning, and suffixes and prefixes;
- The lexical level, which deals with lexical meaning of words and parts of speech analyses;
- The syntactic level, which deals with grammar and structure of sentences;

[2] John Hutchins, *The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954* (2006), http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf.

[3] *See generally* Elizabeth D. Liddy, *Natural Language Processing, in* ENCYCLOPEDIA OF LIBRARY AND INFORMATION SCIENCE 2126 (2d ed. NY. Marcel Decker, Inc. 2001).

[4] *Id.*

[5] *Id.*

- The semantic level, which deals with the meaning of words and sentences;
- The discourse level, which deals with the structure of different kinds of text using document structures;
- And the pragmatic level, which deals with the knowledge that comes from the outside world, *i.e.*, from outside the content of the document.

Although detailed discussions for each step deserves its own Technology Explainer and is thus outside the scope of this piece, a common theme among them is that large language corpora and various statistical methods can improve the accuracy of recognition, parsing, and understanding at *each* of the aforementioned steps. This article details how two common statistical methods are used in the phonetic level of NLP (*i.e.*, input speech recognition).

Input Speech Recognition – Can You Hear Me Now?

The goal of input speech recognition is to capture and accurately transform a user's spoken utterance into text. The user's utterance consists of a series of phonetic sounds; for example, the spoken word sequence "cat nip" consists of six phonetic units (phones): "/k/," "/æ/," "/t/," "/n/,""/ɪ/," and "/p/." In terms of probability, the goal of input speech recognition becomes the following: Given the received phonetic sequence (let's call the sequence "A"), find the sequence of words ("W") such that the conditional probability of the word sequence W given the received phonetic sequence A (*i.e.*, conditional probability P(W|A)) is maximized.[6]

The probability maximization problem can be transformed into maximization of the product of two separate probability components, both having real-world significance, based on the statistical theorem known as Bayes Rule.[7] The Bayesian transformation accounts for how likely it is that the processor will accurately observe the phonetic sequence (the acoustic model), and how likely it is for a given word to occur (the language model).

[6] 1 Mark Gales and Steve Young, *The Application of Hidden Markov Models in Speech Recognition, in* FOUNDATIONS AND TRENDS IN SIGNAL PROCESSING NO. 3, 195-304, 204 (2008).
[7] *See id*. at 201.

$$P(W \mid A) = \frac{\overset{\text{acoustic model (HMMs)}}{P(A \mid W)} \bullet \overset{\text{language model}}{P(W)}}{P(A)}$$

Acoustic Model – Bridging Sound and Words

The first probability component in the numerator of the Bayes transformation is P (A|W); that is, given that the word sequence ***is*** “cat nip,” what the probability of receiving an observed phonetic sequence is. This is called the Acoustic Model.[8] The following table provides examples of what, given that the word sequence is “cat nip,” the respective probabilities of overserved phonetic sequences are:[9]

Phonetic Sequence	Likelihood (qualitative)	Probability (as example)
“/k/,” “/æ/,” “/t/,” “/n/,”“/I/,” “/p/.”	Complete certainty	100%
“/k/,” “/æ/,” “/t/,” “/n/,”“/I/,” **“/t/.”**	Very likely, except for an error while capturing the last phone	90%
“/d/,” “/o/,” “/n/,” “/k/,”“/I/,” “/I/.”	Very unlikely (the phonetic sequence is more likely for the word “Donkey”)	2%

Because speeches (phonetic sequences) are time-sequential, a statistical model known as the Hidden Markov Model is particularly well-suited for modeling speech.[10] Acoustic models in most state-of-the-art NLP systems today are variations of Hidden Markov Models. In practice, a voice recognition system can learn a user’s voice by analyzing the particularities in her voice, such as inflections and the pronunciation of certain consonants, when she speaks a predetermined training phrase.

[8] *Id.*

[9] The numeric probability values in the table below are merely for illustration purposes.

[10] Liddy, *supra* note 3.

Language Model – God Save the Gerbil

The second probability component in the numerator of the Bayes transformation is P(W) (*i.e.,* how likely the word sequence W occurs). This is called the Language Model. In natural speech of a given language, certain word sequences appear much more frequently than other sequences. For example, in the English language, the phrase "God save the king" occurs much more frequently than the phrase "God save the gerbil." Therefore, different probabilities of occurrence can be assigned to the two word sequences, where the first probability greatly exceeds the second. When the voice recognition system detects that the first three words in a user's speech is "God save the," it can consider the respective frequencies of occurrence of the two example word sequences above in determining what the user said. Such word group in a language system is called **n-gram** (*i.e.,* the phrase "God save the King" has four words and is therefore a 4-gram).[11] Armed with a large corpus of text in a certain language and a super computer, a comprehensive statistical analysis can be performed on any n-gram (although in practice n is usually limited to five or less); that is what Google did for several languages based on corpora of texts between the year 1500 A.D. and 2008 A.D. The results and their subsequent publication have significantly advanced computational linguistic research.[12]

Context-Based Language Model – Know Thyself

More refinements to the language model can further improve the language model. For example, in the United States, the phrase "God save the King" is much less common than the phrase "God save the country." If a user is determined to be an American English speaker (such determination can be readily made today by, *e.g.,* using the user's GPS location on her smartphone, detecting her American accent, verifying her U.S. phone number or time zone information), predicating the 4-gram frequency of occurrence on a corpus of American English instead of English spoken everywhere can thus improve the accuracy of prediction.

This type of refinement, called contextual language modeling, can be extended to a variety of contexts: for example, in smartphone digital assistant

[11] DAN JURAFSKY AND JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING: AN INTRODUCTION TO NATURAL LANGUAGE PROCESSING, COMPUTATIONAL LINGUISTICS, AND SPEECH RECOGNITION 85 (2d ed. 2008).

[12] Alex Franz and Thorsten Brants, *All Our N-grams Belong to You*, GOOGLE RESEARCH BLOG (Aug. 3, 2006), http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html.

applications, contextual language models can be based on the app being used, the recipient(s) of certain communication (if a communication/social-network app is being used), the user's GPS location and direction of travel, web browser history, calendar entries, and biometric information (such as temperature or heart rate), and much more.[13]

In the age of cloud computing, it is even possible to use contextual information from groups of users to create a community-based context.[14] For example, if a large number of users that either have a MIT email address or are near the MIT campus speak or write texts linking the term "the Sponge" with the proper name Simmons Hall (a building resembling a sponge), the NLP system can automatically create a community-based contextual language model for everyone who fits a "MIT student or affiliate" criteria and make "Simmons Hall" and "the Sponge" synonymous in the community language model. Ultimately, improving the accuracy becomes a process of narrowly defining both the speaker (what *I*, the user, have previously said or written) and the context of the speech (*i.e.*, whether I am sending an email to my boss in my office or texting my spouse on a commuter train heading home), much like how humans improve accuracy of understanding using context.

Legal Implications

The continued advancement of NLP has important legal implications on data privacy and security. As discussed, NLP thrives on voluminous data in the forms of text corpora and contextual information; the more a NLP system knows about a user's individual's manner of speech, frequently used words, habits, social connections, physiological state, and other information, the better it can process the user's natural language utterances. NLP applications, such as digital assistants (*e.g.*, Siri, Google Now, and Amazon Alexa) have greatly improved capability and accuracy as a result of having a vast amount of corpus and contextual data as training material. However, the user is providing important personal data by using these NLP applications, and the current technology trend is moving toward unconscious sharing of data, frequently without user consent.[15] For example, NLP systems such as Amazon's Alexa can be "always on" in the background to constantly monitor and process the user's

[13] *See, e.g.*, Training an at least partial voice command system, U.S. Patent Application No. 20140278413 at [41–2] (filed Mar. 15, 2013).

[14] *Id*. at 113–4.

[15] *See generally* Stacey Gray, *Always On: Privacy Implications of Microphone-Enabled Devices*, FUTURE OF PRIVACY F. (Apr. 2016), https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf.

speech as she goes about her daily routine, regardless of whether she is aware of such constant monitoring.[16] Furthermore, current NLP systems transmit collected data to remote servers, which alone has the enormous computation power required to achieve the near-instantaneous processing. Such data transfer raises not only information privacy but also security issues, as the data is prone to hacking.

Finally, community-based contextual language models also pose novel legal issues. For example, if a prominent Silicon Valley technology company has gathered enough corpus and contextual data to create an Arabic language model specific to Queens, New York, can law enforcement request (or even compel) access to the language model in order to use it for law enforcement purpose? The complicated issue is that the language model predicates on, yet is distinct from, the corpus and contextual data collected from the user; therefore, who owns the model and who can access it are destined to be contentious legal issues.

CONCLUSION

A profoundly important bridge between human and computer languages, NLP technology will continue to advance at dizzying speed, propelling the progress of artificial intelligence and machine learning along the way. At the same time, the advancement of NLP requires gathering copious amounts of data from users, thereby raising important legal issues in data ownership, privacy, and security. Such novel legal issues will likely come into focus as NLP technology becomes more advanced and mainstream.

[16] Kim Komando, *3 Gadgets That Are Always Listening and How to Stop Them*, USA TODAY (Oct. 6 2015), http://www.usatoday.com/story/tech/columnist/komando/2015/10/02/3-gadgets-always-listening-and-how-stop-them/73191644/.