Advances in natural language processing

Julia Hirschberg1 and Christopher D. Manning2,3

Natural language processing employs computational techniques for the purpose of learning, understanding, and producing human language content. Early computational approaches to language research focused on automating the analysis of the linguistic structure of language and developing basic technologies such as machine translation, speech recognition, and speech synthesis. Today’s researchers refine and make use of such tools in real-world applications, creating spoken dialogue systems and speech-to-speech translation engines, mining social media for information about health or finance, and identifying sentiment and emotion toward products and services. We describe successes and challenges in this rapidly advancing area.

Over the past 20 years, computational linguistics has grown into both an exciting area of scientific research and a practical technology that is increasingly being incorporated into consumer products (for example, in applications such as Apple’s Siri and Skype Translator). Four key factors enabled these developments: (i) a vast increase in computing power, (ii) the availability of very large amounts of linguistic data, (iii) the development of highly successful machine learning (ML) methods, and (iv) a much richer understanding of the structure of human language and its deployment in social contexts. In this Review, we describe some current application areas of interest in language research. These efforts illustrate computational approaches to big data, based on current cutting-edge methodologies that combine statistical analysis and ML with knowledge of language.

Computational linguistics, also known as natural language processing (NLP), is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content. Computational linguistic systems can have multiple purposes: The goal can be aiding human–human communication, such as in machine translation (MT); aiding human–machine communication, such as with conversational agents; or benefitting both humans and machines by analyzing and learning from the enormous quantity of human language content that is now available online.

During the first several decades of work in computational linguistics, scientists attempted to write down for computers the vocabularies and rules of human languages. This proved a difficult task, owing to the variability, ambiguity, and context-dependent interpretation of human languages. For instance, a star can be either an astronomical object or a person, and “star” can be a noun or a verb. In another example, two in-edge methodologies that combine statistical analysis and ML with knowledge of language.

A central finding of this statistical approach to NLP has been that simple methods using words, part-of-speech (POS) sequences (such as whether a word is a noun, verb, or preposition), or simple templates can often achieve notable results when trained on large quantities of data. Many text and sentiment classifiers are still based solely on the different sets of words (“bag of words”) that documents contain, without regard to sentence and discourse structure or meaning. Achieving improvements over these simple baselines can be quite difficult. Nevertheless, the best-performing systems now use sophisticated ML approaches and a rich understanding of linguistic structure. High-performance tools that identify syntactic and semantic information as well as information about discourse context are now available. One example is Stanford CoreNLP (3), which provides a standard NLP preprocessing pipeline that includes POS tagging (with tags such as noun, verb, and preposition); identification of named entities, such as people, places, and organizations; parsing of sentences into their grammatical structures; and identifying co-references between noun phrase mentions (Fig. 1).

Historically, two developments enabled the initial transformation of NLP into a big data field. The first was the early availability to researchers of linguistic data in digital form, particularly through the Linguistic Data Consortium (LDC) (2), established in 1992. Today, large amounts of digital text can easily be downloaded from the Web. Available as linguistically annotated data are large speech and text corpora annotated with POS tags, syntactic parses, semantic labels, annotations of named entities (persons, places, organizations), dialogue acts (statement, question, request), emotions and positive or negative sentiment, and discourse structure (topic or rhetorical structure). Second, performance improvements in NLP were spurred on by shared task competitions. Originally, these competitions were largely funded and organized by the U.S. Department of Defense, but they were later organized by the research community itself, such as the CoNLL Shared Tasks (5). These tasks were a precursor of modern ML predictive modeling and analytics competitions, such as on Kaggle (4), in which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

A major limitation of NLP today is the fact that most NLP resources and systems are available only for high-resource languages (HRls), such as English, French, Spanish, German, and Chinese. In contrast, many low-resource languages (LRLs)—such as Bengali, Indonesian, Punjabi, Cebuano, and Swahili—spoken and written by millions of people have no such resources or systems available. A future challenge for the language community is how to develop resources and tools for hundreds or thousands of languages, not just a few.

Machine translation

Proficiency in languages was traditionally a hallmark of a learned person. Although the social standing of this human skill has declined in the modern age of science and machines, translation between human languages remains crucially important, and MT is perhaps the most substantial way in which computers could aid human–human communication. Moreover, the ability of computers to translate between human languages remains a consummate test of machine intelligence: Correct translation requires not only the ability to analyze and generate sentences in human languages but also a humanlike understanding of world knowledge and context, despite the ambiguities of languages. For example, the French word “bordel” straightforwardly means “brothel”; but if someone says “My room is un bordel,” then a translating machine has to know enough to suspect that this person is probably not running a brothel in his or her room but rather is saying “My room is a complete mess.”

Machine translation was one of the first non-numeric applications of computers and was studied intensively starting in the late 1950s. However, the hand-built grammar-based systems of early decades achieved very limited success. The field was transformed in the early 1990s when researchers at IBM acquired a large quantity of English and French sentences that were translations of each other (known as parallel text), produced as the proceedings of the bilingual Canadian Parliament. These data allowed them to collect statistics of word translations and word sequences and to build a probabilistic model of MT (5).

Following a quiet period in the late 1990s, the new millennium brought the potent combination of ample online text, including considerable quantities of parallel text, much more abundant and inexpensive computing, and a new idea for building statistical phrase-based MT systems.
Rather than translating word by word, the key advance is to notice that small word groups often have distinctive translations. The Japanese "mizu iro" is literally the sequence of two words ("water color"), but this is not the correct meaning (nor does it mean a type of painting); rather, it indicates a light, sky-blue color. Such phrase-based MT was used by Franz Och in the development of Google Translate.

This technology enabled the services we have today, which allow free and instant translation between many language pairs, but it still produces translations that are only just serviceable for determining the gist of a passage. However, very promising work continues to push MT forward. Much subsequent research has aimed to better exploit the structure of human language sentences (i.e., their syntax) in translation systems (7, 8), and researchers are actively building deeper meaning representations of language (9) to enable a new level of semantic MT.

Finally, just in the past year, we have seen the development of an extremely promising approach to MT through the use of deep-learning-based sequence models. The central idea of deep learning is that if we can train a model with several representational levels to optimize a final objective, such as translation quality, then the model can itself learn intermediate representations that are useful for the task at hand. This idea has been explored particularly for neural network models in which information is stored in real-valued vectors, with the mapping between vectors consisting of a matrix multiplication followed by a nonlinearity, such as a sigmoid function that maps the output values of the matrix multiplication onto [−1, 1]. Building large models of this form is much more practical with the massive parallel computation that is now economically available via graphics processing units. For translation, research has focused on a particular version of recurrent neural networks, with enhanced "long short-term memory" computational units that can better maintain contextual information from early until late in a sentence (10, 11) (Fig. 2). The distributed representations of neural networks are often very effective for capturing subtle semantic similarities, and neural MT systems have already produced some state-of-the-art results (12, 13).

A still-underexplored area in MT is getting machines to have more of a sense of discourse, so that a sequence of sentences translates naturally—although work in the area has begun (14). Finally, MT is not necessarily a task for machines to do alone. Rather it can be reconceptualized as an opportunity for computer-supported cooperative work that also exploits human skills (15). In such a system, machine intelligence is aimed at human-computer interface capabilities of giving effective suggestions and reacting productively to human input, rather than wholly replacing the skills and knowledge of a human translator.

**Spoken dialogue systems and conversational agents**

Dialogue has been a popular topic in NLP research since the 1980s. However, early work on text-based dialogue has now expanded to include spoken dialogue on mobile devices (e.g., Apple’s Siri, Amtrak’s Julie, Google Now, and Microsoft’s Cortana) for information access and task-based apps. Spoken dialogue systems (SDSs) also allow robots to help people with simple manual tasks [e.g., Manuela Veloso’s CoBots (16)] or provide therapy for less-abled persons [e.g., Maja Matarić’s socially assistive robots (17)]. They also enable avatars to tutor people in interview or negotiation strategies or to help with health care decisions (18, 19).

The creation of SDSs, whether between humans or between humans and artificial agents, requires tools for automatic speech recognition (ASR), to identify what a human says; dialogue management (DM), to determine what that human wants; actions to obtain the information or perform the activity requested; and text-to-speech (TTS) synthesis, to convey that information back to the human in spoken form (Fig. 3). In addition, SDSs need to be ready to interact with users when an error in speech recognition occurs; to decide what words might be incorrectly recognized; and to determine what the user actually said, either automatically or via dialogue with the user. In speech-to-speech translation systems, MT components are also needed to facilitate dialogue between speakers of different languages and the system, to identify potential mistranslations before they occur; and to clarify these with the speaker.

Practical SDSs have been enabled by breakthroughs in speech recognition accuracy, mainly coming from replacing traditional acoustic feature-modeling pipelines with deep-learning models that map sound signals to sequences of human language sounds and words (20). Although SDSs now work fairly well in limited domains, where the topics of the interaction are known in advance and where the words people are likely to use can be predetermined, they are not yet very successful in open-domain interaction, where users may talk about anything at all. Chatbots following in the tradition of ELIZA (21) handle open-domain interaction by cleverly repeating variations of the human input;
this approach is also being attempted in spoken-chat systems designed to provide a sense of companionship for target audiences such as the elderly or individuals with dementia (22). In spoken dialogue, information about the speaker’s mental state inferred from multimodal information can be used to supplement the system's knowledge of what the user is saying.

There are many challenges in building SDSs, in addition to the primary challenge of improving the accuracy of the basic ASR, DM, and TTS building blocks and extending their use into less-restricted domains. These include basic problems of recognizing and producing normal human conversational behaviors, such as turn-taking and coordination. Humans interpret subtle cues in speakers' voices and facial and body gestures (where available) to determine when the speaker is ready to give up the turn versus simply pausing. These cues, such as a filled pause (e.g., "um" or "uh"), are also used to establish when some feedback from the listener is desirable, to indicate that he or she is listening or working on a request, as well as to provide "grounding" (i.e., information about the current state of the conversation). Non-humanlike latency often makes SDS burdensome, as users must wait seconds to receive a system response. To address this, researchers are exploring incremental processing of ASR, MT, and TTS modules, so that systems can respond more quickly to users by beginning these recognition, translation, and generation processes while the user is still speaking. Humans can also disambiguate words such as "yeah" and "okay," which may have diverse meanings—including agreement, topic shift, and even disagreement—when spoken in different ways. In successful and cooperative conversations, humans also tend to entrain to their conversational partners, becoming more similar to each other in pronunciation, word choice, acoustic and prosodic features, facial expressions, and gestures. This tendency has long been used to subtly induce SDS users to employ terms that the system can more easily recognize. Currently, researchers are beginning to believe that systems (particularly embodied agents) should entrain to their users in these different modalities, and some experimental results have shown that users prefer such systems (23) and even think they are more intelligent (24). Open issues for DM have long been the determination of how to architect the appropriate dialogue flow for particular applications, where existing experimental data may be sparse and some aspects of the dialogue state may not yet have been observed or even be observable from the data. Currently, the most widely used approach is the POMDP (partially observable Markov decision process), which attempts to identify an optimal system policy by maintaining a probability distribution over possible SDS states and updating this distribution as the system observes additional dialogue behavior (25). This approach may make use of the identification of dialogue acts, such as whether the user input represents a question, statement, or indication of agreement, for example.

### Machine reading

The printed word has great power to enlighten. Machine reading is the idea that machines could become intelligent, and could usefully integrate and summarize information for humans, by reading and understanding the vast quantities of text that are available.

In the early decades of artificial intelligence, many researchers focused on the approach of trying to enable intelligent machines by manually building large structured knowledge bases in a formal logical language and developing automated reasoning methods for deriving further facts from this knowledge. However, with the emergence of the modern online world, what we mainly have instead is huge repositories of online information coded in human languages. One place where this is true is in the scientific literature, where findings are still reported almost entirely in human language text (with accompanying tables and diagrams). However, it is equally true for more general knowledge, where we now have huge repositories of information such as Wikipedia (26). The quantity of scientific literature is growing rapidly: For example, the size of the U.S. National Library of Medicine’s Medline index has grown exponentially (27). At such a scale, scientists are unable to keep up with the literature, even in their narrow domains of expertise. Thus, there is an increased need for machine reading for the purposes of comprehending and summarizing the literature, as well as extracting facts and hypotheses from this material.

An initial goal is to extract basic facts, most commonly a relation between two entities, such as “child of” (for instance, Bill Clinton, Chelsea Clinton). This is referred to as relation extraction. For particular domain-specific relations, many such systems have been successfully built. One technique is to use handwritten patterns that match the linguistic expression of relations (e.g., `<PERSON>`’s daughter, `<PERSON>`). Better results can be obtained through the use of...
ML. A structured prediction classifier proposes instances of such relations based on extracted features from the sequence of words and grammatical structure of a sentence (28, 29). Such systems are the mainstay of literature fact-extraction tools in fields such as biomedicine (30, 31).

In many scientific fields, there have been major efforts to build databases of structured information based on the textual scientific record, such as the Gene Ontology database (32) in biomedicine or the PaleoBiology Database for fossil records (33). This has generally been done manually, via concerted work by trained professionals. Using artificial intelligence software to extract these databases, as well as to perform subsequent reasoning and hypothesis generation, has become a major research goal. One subfield where these questions have been actively pursued is pharmacogenomics (34). For example, Percha et al. (35) trained a model of drug-drug interactions based on drug-gene interactions extracted from the literature and were able to use it to predict novel drug-drug interactions.

If a partial knowledge base—for instance, Freebase (36), dbpedia (37), Wikidata (38) (related to Wikipedia), or the Gene Ontology database (32)—has already been extracted from biomedical research articles, then there is an opportunity to automatically align known facts from the knowledge base with putative expressions of those facts in text. The type labels from this mapping can then be used as if they were supervised data for ML information-extraction systems (Fig. 4). This is referred to as distantly supervised relation extraction. Early systems aligned entity mentions and then made the naive assumption that sentences containing a pair of entities expressed every known relation between the two entities in the database (39). More recent systems have used increasingly sophisticated probabilistic inference to discern which textual clauses map to which facts in the knowledge base, or to something else entirely (40, 41). A dramatic recent application of this approach has been the DeepDive system (42), which aims to automate the construction of such systems by providing efficient large-scale learning and inference so a user can simply focus on good features for their domain. PaleoDeepDive, its application to the fossil record, has recently been shown to do a better job at fact extraction from journal articles than the scientist volunteers who maintain the PaleoBiology Database (43).

The relation-extraction task is made general, if less semantically precise, by aiming to extract all relations from any piece of text, a task normally referred to as open information extraction (Open IE) (44). Early work emphasized the development of simple but highly scalable fact-extraction techniques that do not require any kind of hand-labeled data (45). With ever-growing computational power, a second generation of work increasingly emphasized careful use of linguistic structure, which can reliably be extracted with the use of detailed NLP (46).

Currently, a number of avenues are being explored to further extend the ability of computers to build and use knowledge bases starting from textual information. An exciting unification is the proposal for universal schemas (47), which allow simultaneous inference and knowledge-base completion over both the open set of textual relations (such as “born in”) found in Open IE and the more exact schema of databases (such as per:city_of_birth). Even with all of our text-extraction techniques, any knowledge base will only be partial and incomplete; some recent work explores how it can be probabilistically completed to deliver a form of common-sense reasoning (48). Finally, we hope to move beyond simply extracting relations, events, and facts to be able to understand the relations between events (such as causation) and complex multistep procedures and processes. In (49), Berant et al. explore how this can be done for understanding the steps in biological processes, showing that extracting explicit process structures can improve the accuracy of question answering. The flip side of machine reading is to provide question-answering systems, by which humans can get answers from constructed knowledge bases. There has recently been dramatic progress in building such systems by learning semantic parsers (50).

Mining social media

The development of social media has revolutionized the amount and types of information available today to NLP researchers. Data available from sources such as Twitter, Facebook, YouTube, blogs, and discussion forums make it possible to examine relations between demographic information, language use, and social interaction (51). Researchers use Web-scraping techniques, often via application program interfaces provided by websites, to download previously unimaginable amounts and categories of data. Using statistical and ML techniques, they learn to identify demographic information (such as age and gender) from language, track trending topics and popular sentiment, identify opinions and beliefs about products and politicians, predict disease spreading (for instance, with Google Flu Trends: www.google.org/flutrends/) from symptoms mentioned in tweets or food-related illnesses (52), recognize deception in fake reviews (53), and identify social networks of people who interact together online.

In this era of big data, the availability of social media has revolutionized the ways advertisers, journalists, businesses, politicians, and medical experts acquire their data and the ways in which those data can be put to practical use. Product reviews can be mined to predict pricing trends and assess advertising campaigns. Political forums can be searched to predict candidate appeal and performance in elections. Social networks can be examined to find indicators of power and influence among different groups. Medical forums can be studied to discover common questions and misconceptions about sufferers from particular medical conditions so that website information can be improved.

Social media also provide very large and rich sources of conversational data in Web forums that can provide “found” data for the study of language phenomena such as code-switching (mixed
language in bilingual speech), hedging behavior (words and phrases indicating lack of commitment to a proposition such as "sort of"); and hate speech or bullying behavior. Social media exist in a wide variety of languages, including both HRLs and LRLs. These data can be invaluable for enriching ASR language models and developing TTS synthesizers without the need to create costly special-purpose corpora. In turn, these technologies can be useful in producing SDSs in LRL areas. Such systems can provide millions of people with the ability to obtain information over their cell phones which are ubiquitous, even among populations with low literacy rates or whose languages or dialects have no standard written form), similar to the residents of HRL countries. The development of tools for LRLs from found LRL data, by adapting HRL tools, is another important way to use found text data. A particular application of data mining in LRLs is the mining of data collected from Twitter or blogs to provide valuable information for disaster relief organizations, identifying the most serious problems, where they occur, and who is experiencing them.

There are also some drawbacks to social media data mining. There is an increasing concern for privacy issues, particularly for an individual’s control over their own data versus researchers’ desire to mine it. Sites such as Twitter severely limit a researcher’s ability to download data, which impedes speedy corpus collection. There is also a major issue with discovering “ground truth” in online postings, because there is no clear way of validating an individual’s demographic information; the validity of posts concerning events; and most reviews of hotels, restaurants, and products. Aggregating information from multiple sources at similar times can address some validity issues, and sites do attempt to identify spurious reviews, but this issue remains perhaps the most difficult one for those working with social media data.

**Analysis and generation of speaker state**

Speaker states (54), also termed “private states” (55), include opinions, speculations, beliefs, emotions, and any other evaluative views that are personally held by the speaker or writer of a language. Much of the work in NLP has focused on sentiment analysis (identification of positive or negative orientation of textual language) and identification of belief states (committed belief, uncommitted belief, or neutrality of a sentence) on the basis of lexical and syntactic information. Both sentiment and belief constitute attitudes toward events and propositions, although sentiment can also concern attitudes toward objects such as people, organizations, and abstract concepts. Detection of sentiment and emotion in text requires lexical and sentence-level information. Sentiment can be signaled by words conveying positive or negative orientation: For example, “sad,” “worried,” “difficult,” and “weak” are all words with negative orientation, whereas “comfortable,” “important,” “successful,” and “interesting” convey a positive sentiment. Online sentiment dictionaries, such as Whissel’s Dictionary of Affect (56), and systems created from subject-ranked terms, such as Tausczik and Pennebaker’s LIWC (Linguistic Inquiry and Word Count) (57), can be used to assess positive and negative sentiment in a text. More sophisticated approaches to sentiment analysis also seek to identify the holder (source) as well as the object of the sentiment: for instance, who is positive about what person, country, activity, or concept (55).

The speech community has also studied positive and negative attitudes by focusing more generally on the identification of positive and negative emotions, primarily using acoustic and prosodic information. However, more work is currently being done to identify particular emotions, such as Ekman’s classic six basic emotions (anger, disgust, fear, happiness, sadness, surprise), which may be reactions to events, propositions, or objects. There has also been considerable research using features that have proven important in recognizing classic emotions to identify other speaker states (such as deception), medical conditions (such as autism and Parkinson’s disease), speaker characteristics (such as age, gender, likeability, pathology, and personality), and speaker conditions (such as cognitive load, drunkenness, sleepiness, interest, and trust). Corpora collected for such studies have been used in the Interspeech Para-linguistic Challenges, which have been conducted since 2009. Emotion generation has proven a more difficult challenge for TTS synthesis. Although there are some systems (e.g., MARY) that attempt to generate emotions such as depression, aggression, or cheerfulness (58), the best synthesized emotion still comes from corpora recorded for particular emotions by voice talent imitating those emotions.

Sentiment classification is widely used in opinion identification (positive or negative views of people, institutions, or ideas) in many languages and genres. Particular applications abound, such as identifying positive and negative movie or product reviews (59, 60) and predicting votes from congressional records (62) or Supreme Court decisions from court proceedings. Figure 5 illustrates a typical restaurant review, annotated for positive, negative, and neutral sentiment, as well as basic emotions. Mining social media for sentiment or classic emotions has been a particularly popular topic for the purposes of assessing the “public mood” from Twitter, predicting stock market trends, or simply evaluating a community's mental state (62). Social media such as Twitter, blog posts, and forums also provide researchers with very large amounts of data to use in assessing the role of sentiment and emotion in identifying other linguistic or social phenomena (e.g., sarcasm (63), power relationships, and social influence (64)), as well as mental health issues (e.g., depression (65)).

**Conclusion and outlook**

Many times during the past 50 years, enthusiastic researchers have had high hopes that the language-understanding ability of robots in science fiction movies was just around the corner. However, in reality, speech and language understanding did not work well enough at that time to power mainstream applications. The situation has been changing dramatically over the past five years. Huge improvements in speech recognition have made talking to your phone a commonplace activity, especially for young people. Web search engines are increasingly successful in understanding complex queries, and MT can at least yield the gist of material in another language, even if it cannot yet produce human-quality translations. Computer systems trade stocks and futures automatically, based on the sentiment of reports about companies. As a result, there is now great commercial interest in the deployment of human language technology, especially because natural language represents such a natural interface when interacting with mobile phones. In the short term, we feel confident that more data and computation, in addition to recent advances in ML and deep learning, will lead to further substantial progress in NLP. However, the truly difficult problems of semantics, context, and...
knowledge will probably require new discoveries in linguistics and inference. From this perspective, it is worth noting that the development of probabilistic approaches to language is not simply about solving engineering problems: Probabilistic models of language have also been reflected back into linguistic science, where researchers are finding important new applications in describing phonology (66), understanding human language processing (67), and modeling linguistic semantics and pragmatics (68). Many areas of linguistics are therefore becoming more empirical and more quantitative in their approaches.

REFERENCES AND NOTES


ACKNOWLEDGMENTS

C.D.M. holds equity in Google and serves in an advisor role to Idibon, Lift, Lex Machina, and Xseed.
Advances in natural language processing
Julia Hirschberg and Christopher D. Manning

Science 349 (6245), 261-266,
DOI: 10.1126/science.aaa8685

Use of this article is subject to the Terms of Service