REVIEW

# Advances in natural language processing

Julia Hirschberg[1]* and Christopher D. Manning[2,3]

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 benefiting 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 interpretations are possible for the headline "Teacher strikes idle kids," depending on the noun, verb, and adjective assignments of the words in the sentence, as well as grammatical structure. Beginning in the 1980s, but more widely in the 1990s, NLP was transformed by researchers starting to build models over large quantities of empirical language data. Statistical or corpus ("body of words")–based NLP was one of the first notable successes of the use of big data, long before the power of ML was more generally recognized or the term "big data" even introduced.

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 (*1*), 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 (*3*). 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 nonnumeric 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

[1]Department of Computer Science, Columbia University, New York, NY 10027, USA. [2]Department of Linguistics, Stanford University, Stanford, CA 94305-2150, USA. [3]Department of Computer Science, Stanford University, Stanford, CA 94305-9020, USA.
*Corresponding author. E-mail: julia@cs.columbia.edu

(*6*). Rather than translating word by word, the key advance is to notice that small word groups often have distinctive translations. The Japanese 水色 "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 Mataric'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;

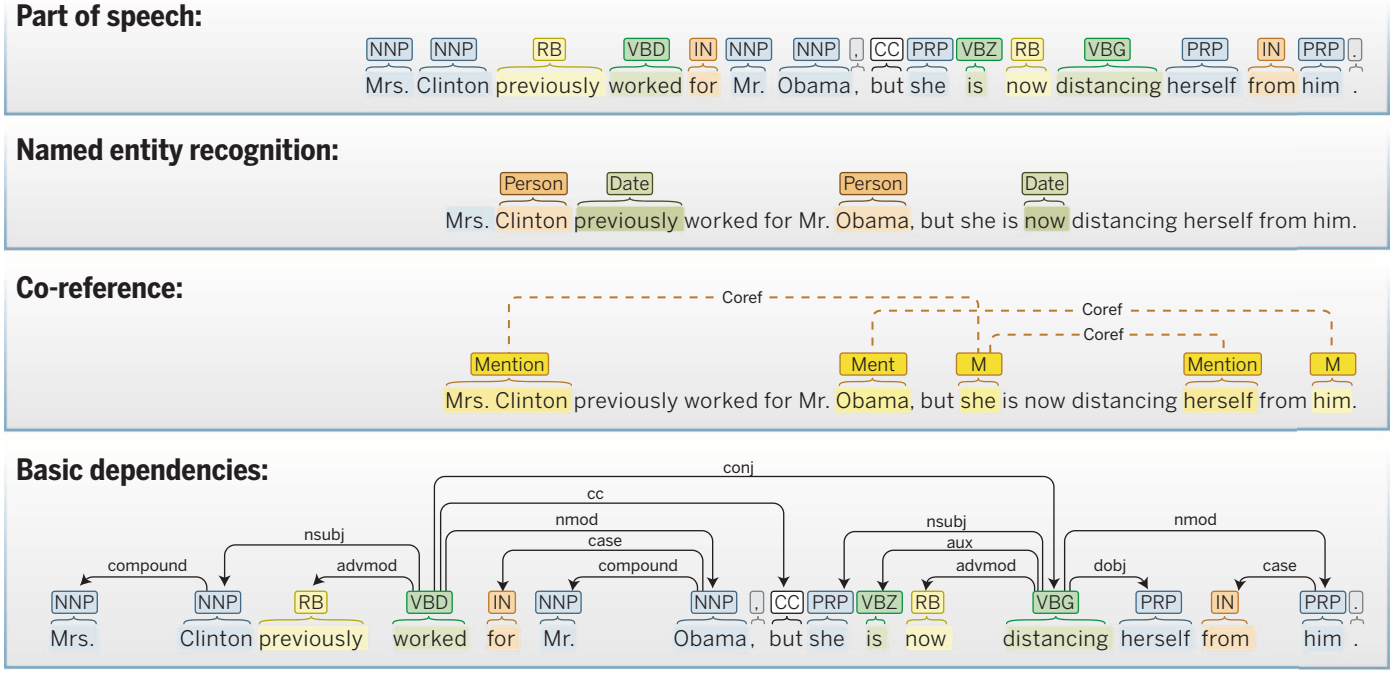


**Fig. 1. Many language technology tools start by doing linguistic structure analysis.** Here we show output from Stanford CoreNLP. As shown from top to bottom, this tool determines the parts of speech of each word, tags various words or phrases as semantic named entities of various sorts, determines which entity mentions co-refer to the same person or organization, and then works out the syntactic structure of each sentence, using a dependency grammar analysis.

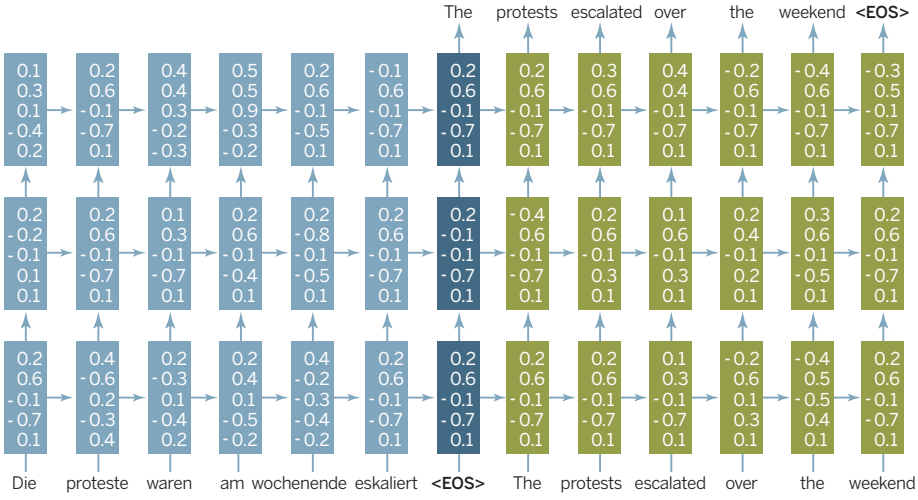


**Fig. 2. A deep, recurrent neural MT system (*10*).** Initially trained on parallel sentences that translate each other, the model learns a representation of each word as a real-valued vector and internal parameter matrices so as to optimize translation quality. The trained network can then translate new sentences. Each arrow represents a computation unit of a matrix multiplication followed by a nonlinear transformation; the small vectors shown in the illustration might really be 1000-dimensional. The recurrent network first encodes the meaning of the source sentence (left side, blue). It maintains an internal state representing a partial sentence, which is updated after each new word is read (horizontal arrows). Having the upper network layers [mapped to by additional (upper vertical) computation arrows] makes this a deep recurrent network. Adding depth improves the ability of the model to learn, generalize, and remember. Once the end of the sentence (denoted by <EOS>) is reached (middle, dark blue), the network additionally starts to produce a word of translated output at each step from its internal state (using a multiclass logistic regression–style model). During translation generation (right side, green), the last generated word is fed in as the input at each step. From the stored hidden state and this input, the model calculates the next word of the translation. This process repeats until <EOS> is generated.

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

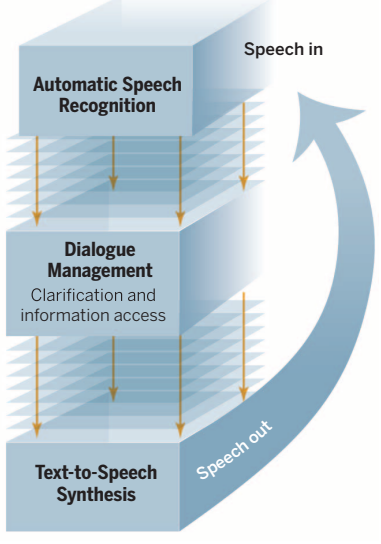


**Fig. 3. A spoken dialogue system.** The three main components are represented by rectangles; arrows denote the flow of information.

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 naïve 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

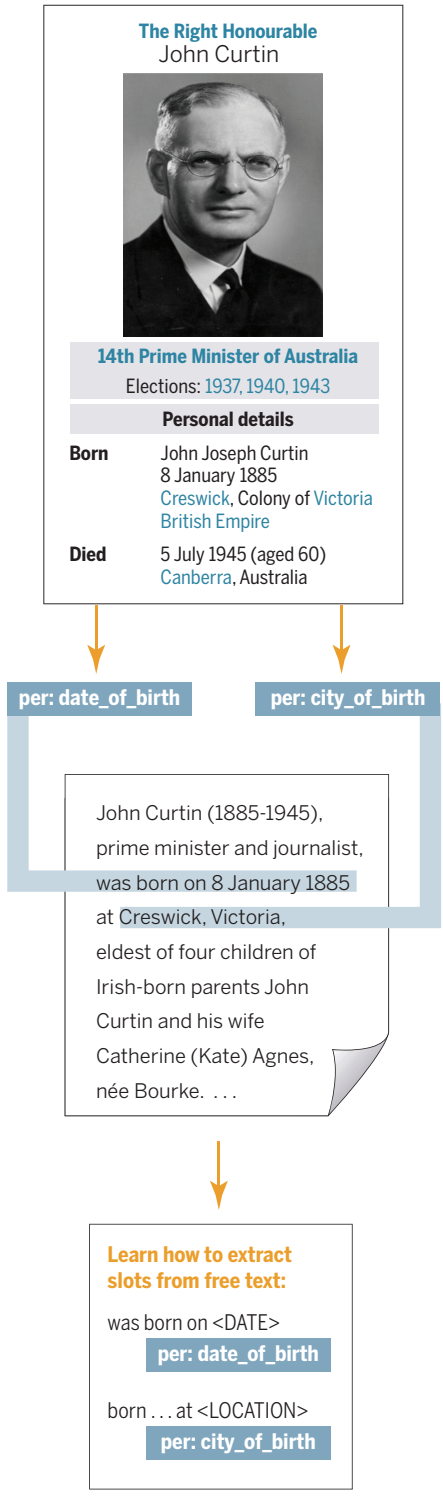


**Fig. 4. Distantly supervised learning.** In this approach, facts in a structured knowledge representation are projected onto pieces of text that mention the people, places, dates, etc., that appear in knowledge-base entries. This projection is noisy, but when done over large quantities of text, it provides enough signal to successfully learn good classifiers for extracting relations from text. [Photo source: National Library of Australia, http://nla.gov.au/nla.pic-an12267621]

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 Paralinguistic 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 (*61*) 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

Breakfast on Broadway is a new place focusing on, you guessed it, breakfast/brunch. Went there last Sunday around 1. The food was not bad but the service was pretty terrible. We had to wait 15 minutes just to get menus and another 30 to get something to eat. And there were only a few tables occupied! If you don't mind the wait though, the price is right. I'll probably give it another try. Maybe they need time to get their act together.

Breakfast on Broadway is a new place focusing on, you guessed it, breakfast/brunch. Went there last Sunday around 1. The food was not bad but **[Anger:** the service was pretty terrible**]**. **[Disgust:** We had to wait 15 minutes just to get menus and another 30 to get something to eat. And there were only a few tables occupied!**]** If you don't mind the wait though, the price is right. I'll probably give it another try. **[Uncertainty:** Maybe they need time to get their act together.**]**

**Fig. 5. Manually annotated text analysis on a sample restaurant review.** Sentiment analysis is shown on the left (blue, positive sentiments; red, negative; gray, neutral). In the emotion analysis on the right, emotions are shown in bold type and delineated by square brackets. Note in particular the importance of going beyond simple keyword analysis; for example, "not" has scope over "bad," which might mislead simple systems. Also, the presence of "hedge" words and phrases, which muddle the intended meaning (e.g., "pretty," which has a positive connotation, modifying the negative word "terrible"), somewhat decreases the negative score of the next clause.

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 themselves becoming more empirical and more quantitative in their approaches.

## REFERENCES AND NOTES


1. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations* (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 55–60.
2. Linguistic Data Consortium, www.ldc.upenn.edu/.
3. CoNLL Shared Tasks, http://ifarm.nl/signll/conll/.
4. Kaggle, www.kaggle.com.
5. P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer, *Comput. Linguist.* **19**, 263–311 (1993).
6. P. Koehn, F. J. Och, D. Marcu, "Statistical phrase-based translation," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, 2003), pp. 48–54.
7. D. Chiang, "A hierarchical phrase-based model for statistical machine translation," *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, 2005), pp. 263–270.
8. M. Galley, M. Hopkins, K. Knight, D. Marcu, "What's in a translation rule?" in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2004)* (Association for Computational Linguistics, Stroudsburg, PA, 2004).
9. B. Jones, J. Andreas, D. Bauer, K. M. Hermann, K. Knight, "Semantics-based machine translation with hyperedge replacement grammars," in *Proceedings of COLING 2012* (Technical Papers, The COLING 2012 Organizing Committee, Mumbai, India, 2012), pp. 1359–1376.
10. I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, Eds. (Curran Associates, Red Hook, NY, 2014), pp. 3104–3112.
11. D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," http://arxiv.org/abs/1409.0473 (2015).
12. M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, W. Zaremba, "Addressing the rare word problem in neural machine translation," http://arxiv.org/abs/1410.8206 (2015).
13. S. Jean, K. Cho, R. Memisevic, Y. Bengio, "On using very large target vocabulary for neural machine translation," http://arxiv.org/abs/1412.2007 (2015).
14. S. Stymne, C. Hardmeier, J. Tiedemann, J. Nivre, "Feature weight optimization for discourse-level SMT," in *Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT)* (Association for Computational Linguistics, Stroudsburg, PA, 2013), pp. 60–69.
15. S. Green, J. Chuang, J. Heer, C. D. Manning, "Predictive translation memory: A mixed-initiative system for human language translation," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, Honolulu, HI, 5 to 8 October 2014 (Association for Computing Machinery, New York, 2014), pp. 177–187.
16. S. Rosenthal, J. Biswas, M. Veloso, "An effective personal mobile robot agent through symbiotic human-robot interaction," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, Toronto, Canada, 10 to 14 May 2010 (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2010), pp. 915–922.
17. J. Fasola, M. J. Matarić, *J. Human-Robot Interact.* **2**, 3–32 (2013).
18. M. Core, H. C. Lane, D. Traum, "Intelligent tutoring support for learners interacting with virtual humans," in *Design Recommendations for Intelligent Tutoring Systems* (U.S. Army Research Laboratory, Orlando, FL, 2014), vol. 2, pp. 249–257.
19. D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, L.-P. Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, Paris, France, 5 to 9 May 2014 (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2014), pp. 1061–1068; http://aamas2014.lip6.fr/proceedings/aamas/p1061.pdf.
20. G. Hinton *et al*., *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
21. J. Weizenbaum, *Commun. ACM* **9**, 36–45 (1966).
22. Y. Nonaka, Y. Sakai, K. Yasuda, Y. Nakano, "Towards assessing the communication responsiveness of people with dementia," in *12th International Conference on Intelligent Virtual Agents (IVA'12)* (Springer, Berlin, 2012), pp. 496–498.
23. C. Nass, Y. Moon, B. J. Fogg, B. Reeves, D. C. Dryer, *Int. J. Hum. Comput. Stud.* **43**, 223–239 (1995). .
24. H. Giles, A. Mulac, J. J. Bradac, P. Johnson, "Speech accommodation theory: The next decade and beyond," in *Communication Yearbook* (Sage, Newbury Park, CA, 1987), vol. 10, pp. 13–48.
25. S. Young, M. Gasic, B. Thomson, J. Williams, *Proc. IEEE* **101**, 1160–1179 (2013).
26. Wikipedia, www.wikipedia.org/.
27. L. Hunter, K. B. Cohen, *Mol. Cell* **21**, 589–594 (2006).
28. A. Culotta, J. Sorensen, "Dependency tree kernels for relation extraction," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, 2004), pp. 423–429.
29. K. Fundel, R. Küffner, R. Zimmer, *Bioinformatics* **23**, 365–371 (2007).
30. J. Björne *et al*., *Comput. Intell.* **27**, 541–557 (2011).
31. S. Van Landeghem *et al*., *PLOS ONE* **8**, e55814 (2013).
32. M. Ashburner *et al*. The Gene Ontology Consortium, *Nat. Genet.* **25**, 25–29 (2000).
33. PaleoBiology Database, https://paleobiodb.org/.
34. A. Coulet, K. B. Cohen, R. B. Altman, *J. Biomed. Inform.* **45**, 825–826 (2012).
35. B. Percha, Y. Garten, R. B. Altman, *Pac. Symp. Biocomput.* **2012**, 410–421 (2012).
36. Freebase, www.freebase.com/.
37. dbpedia, http://dbpedia.org/.
38. Wikidata, www.wikidata.org/.
39. M. Mintz, S. Bills, R. Snow, D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Association for Computational Linguistics, Stroudsburg, PA, 2009), vol. 2, pp. 1003–1011.
40. M. Surdeanu, J. Tibshirani, R. Nallapati, C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju Island, South Korea, 12 to 14 July 2012 (Association for Computational Linguistics, Stroudsburg, PA, 2012), pp. 455–465.
41. B. Min, R. Grishman, L. Wan, C. Wang, D. Gondek, "Distant supervision for relation extraction with an incomplete knowledge base," in *Proceedings of NAACL-HLT 2013*, Atlanta, GA, 9 to 14 June 2013 (Association for Computational Linguistics, Stroudsburg, PA, 2013), pp. 777–782.
42. DeepDive, http://deepdive.stanford.edu/.
43. S. E. Peters, C. Zhang, M. Livny, C. Ré, *PLOS ONE* **9**, e113523 (2014).
44. E. Etzioni, M. Banko, M. J. Cafarella, "Machine reading," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, Boston, MA, 16 to 20 July 2006 (AAAI Press, Menlo Park, CA, 2006), vol. 2, pp. 1517–1519.
45. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, "Open information extraction from the web," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI 2007)* (Morgan Kaufmann, San Francisco, 2007), pp. 2670–2676.
46. O. Etzioni, A. Fader, J. Christensen, S. Soderland, Mausam, "Open information extraction: The second generation," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 16 to 22 July 2011 (AAAI Press, Menlo Park, CA, 2011), pp. 3–10.
47. S. Riedel, L. Yao, A. McCallum, B. M. Marlin, "Relation extraction with matrix factorization and universal schemas," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (HLT NAACL 2013)* (Stroudsburg, PA, 2013), pp. 74–84.
48. G. Angeli, C. D. Manning, "NaturalLI: Natural logic inference for common sense reasoning," in *Proceedings of the 2014 Conference on Emprical Methods in Natural Language Processing*, Doha, Qatar, 25 to 29 October 2014 (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 534–545.
49. J. Berant, V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, C. D. Manning, "Modeling biological processes for reading comprehension," in *Proceedings of the 2014 Conference on Emprical Methods in Natural Language Processing*, Doha, Qatar, 25 to 29 October 2014 (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 1499–1510.
50. A. Fader, L. Zettlemoyer, O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *Proceedings of the Conference on Knowledge Discovery and Data Mining* (KDD) (Association for Computing Machinery, New York, 2014), pp. 1156–1165.
51. M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More* (O'Reilly Media, Sebastopol, CA, ed. 2, 2013).
52. N. Elhadad, L. Gravano, D. Hsu, S. Balter, V. Reddy, H. Waechter, "Information extraction from social media for public health," in *KDD at Bloomberg Workshop, Data Frameworks Track (KDD 2014)* (Association for Computing Machinery, New York, 2014).
53. M. Ott, C. Cardie, J. T. Hancock, "Estimating the prevalence of deception in online review communities." in *Proceedings of the 21st International Conference on World Wide Web Conference*, Lyon, France, 16 to 20 April 2012 (Association for Computing Machinery, New York, 2012), pp. 201–210.
54. J. Liscombe, thesis, Columbia University (2007).
55. J. Wiebe, T. Wilson, C. Cardie, *Lang. Resour. Eval.* **39**, 165–210 (2005).
56. C. Whissell, "The dictionary of affect in language," in *Emotion: Theory, Research and Experience*, R. Plutchik, H. Kellerman, Eds. (Academic Press, London, 1989).
57. Y. R. Tausczik, J. W. Pennebaker, *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
58. O. Türk, M. Schröder, *IEEE Trans. Audio Speech Lang. Proc.* **18**, 965–973 (2010).
59. B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, July 2002 (Association for Computational Linguistics, Stroudsburg, PA, 2002), vol. 10, pp. 79–86.
60. H. Wang, M. Ester, "A sentiment-aligned topic model for product aspect rating prediction," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 25 to 29 October 2014 (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 1192–1202.
61. M. Thomas, Bo Pang, L. Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," in *Proceedings of the 2006 Conference on Emprical Methods in Natural Language Processing*, Sydney, Australia, 22 to 23 July 2006 (Association for Computational Linguistics, Stroudsburg, PA, 2006), pp. 327–335.
62. J. Bollen, H. Mao, A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 17 to 21 July 2011 (AAAI Press, Menlo Park, 2011), pp. 450–453.
63. R. Gonzalez-Ibanez, S. Muresan, N. Wacholder, "Identifying sarcasm in Twitter: A closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, 19 to 24 June 2011 (Association for Computational Linguistics, Stroudsburg, PA, 2011), pp. 581–586.
64. O. Biran, S. Rosenthal, J. Andreas, K. McKeown, O. Rambow, "Detecting influencers in written online conversations," in *Proceedings of the 2012 Workshop on Language in Social Media*, Montreal, Canada, 7 June 2012 (Association for Computational Linguistics, Stroudsburg, PA, 2012), pp. 37–45.
65. L.-C. Yu, C.-Y. Ho, "Identifying emotion labels from psychiatric social texts using independent component analysis," in *Proceedings of COLING 2014* (Technical Papers, Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 837–847.
66. B. Hayes, Z. Londe, *Phonology* **23**, 59–104 (2006).
67. R. Levy, *Cognition* **106**, 1126–1177 (2008).
68. N. D. Goodman, D. Lassiter, "Probabilistic semantics and pragmatics: Uncertainty in language and thought," in *Handbook of Contemporary Semantics*, C. Fox, S. Lappin, Eds. (Blackwell, Hoboken, NJ, ed. 2, 2015).


## ACKNOWLEDGMENTS

C.D.M. holds equity in Google and serves in an advising role to Idibon, Lilt, Lex Machina, and Xseed.

10.1126/science.aaa8685

**Advances in natural language processing**
Julia Hirschberg and Christopher D. Manning (July 16, 2015)
*Science* **349** (6245), 261-266. [doi: 10.1126/science.aaa8685]

Editor's Summary



*Science* (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 