WIKIPEDIA

Natural language processing

Natural language processing (**NLP**) is a subfield of underlined linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural-language generation.

An automated online assistant providing customer service on a web page, an example of an application where natural language processing is a major component.[1]

Contents

History

Natural language processing has its roots in the 1950s. Already in 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence, a task that involves the automated interpretation and generation of natural language, but at the time not articulated as a problem separate from artificial intelligence.

Symbolic NLP (1950s - early 1990s)

The premise of symbolic NLP is well-summarized by John Searle's Chinese room experiment: Given a collection of rules (e.g., a Chinese phrasebook, with questions and matching answers), the computer emulates natural language understanding (or other NLP tasks) by applying those rules to the data it is confronted with.

- **1950s**: The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem.[2] However, real progress was much slower, and after the ALPAC report in 1966, which found that ten-year-long research had failed to fulfill the expectations, funding for machine translation was dramatically reduced. Little further research in machine translation was conducted until the late 1980s when the first statistical machine translation systems were developed.

- **1960s**: Some notably successful natural language processing systems developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between 1964 and 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction. When the "patient" exceeded the very small knowledge base, ELIZA might provide a generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?".

- **1970s**: During the 1970s, many programmers began to write "conceptual ontologies", which structured real-world information into computer-understandable data. Examples are MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot Units (Lehnert 1981). During this time, the first many chatterbots were written (e.g., PARRY).

- **1980s**: The 1980s and early 1990s mark the hey-day of symbolic methods in NLP. Focus areas of the time included research on rule-based parsing (e.g., the development of HPSG as a computational operationalization of generative grammar), morphology (e.g., two-level morphology[3]), semantics (e.g., Lesk algorithm), reference (e.g., within Centering Theory[4]) and other areas of natural language understanding (e.g., in the Rhetorical Structure Theory). Other lines of research were continued, e.g., the development of chatterbots with Racter and Jabberwacky. An important development (that eventually led to the statistical turn in the 1990s) was the rising importance of quantitative evaluation in this period.[5]

Statistical NLP (1990s - 2010s)

Up to the 1980s, most natural language processing systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of machine learning algorithms for language processing. This was due to both the steady increase in computational power (see Moore's law) and the gradual lessening of the dominance of Chomskyan theories of linguistics (e.g. transformational grammar), whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing.[6]

- **1990s**: Many of the notable early successes on statistical methods in NLP occurred in the field of machine translation, due especially to work at IBM Research. These systems were able to take advantage of existing multilingual textual corpora that had been produced by the Parliament of Canada and the European Union as a result of laws calling for the translation of all governmental proceedings into all official languages of the corresponding systems of government. However, most other systems depended on corpora specifically developed for the tasks implemented by these systems, which was (and often continues to be) a major limitation in the success of these systems. As a result, a great deal of research has gone into methods of more effectively learning from limited amounts of data.

- **2000s**: With the growth of the web, increasing amounts of raw (unannotated) language data has become available since the mid-1990s. Research has thus increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms can learn from data that has not been hand-annotated with the desired answers or using a combination of annotated and non-annotated data. Generally, this task is much more difficult than supervised learning, and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available (including, among other things, the entire content of the World Wide Web), which can often make up for the inferior results if the algorithm used has a low enough time complexity to be practical.

Neural NLP (present)

In the 2010s, representation learning and deep neural network-style machine learning methods became widespread in natural language processing, due in part to a flurry of results showing that such techniques[7][8] can achieve state-of-the-art results in many natural language tasks, for example in language modeling,[9] parsing,[10][11] and many others. This is increasingly important in medicine and healthcare, where NLP is being used to analyze notes and text in electronic health records that would otherwise be inaccessible for study when seeking to improve care.[12]

Methods: Rules, statistics, neural networks

In the early days, many language-processing systems were designed by symbolic methods, i.e., the hand-coding of a set of rules, coupled with a dictionary lookup:[13][14] such as by writing grammars or devising heuristic rules for stemming.

More recent systems based on machine-learning algorithms have many advantages over hand-produced rules:

- The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not at all obvious where the effort should be directed.
- Automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar input (e.g. containing words or structures that have not been seen before) and to erroneous input (e.g. with misspelled words or words accidentally omitted). Generally, handling such input gracefully with handwritten rules, or, more generally, creating systems of handwritten rules that make soft decisions, is extremely difficult, error-prone and time-consuming.
- Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on handwritten rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task. In particular, there is a limit to the complexity of systems based on handwritten rules, beyond which the systems become more and more unmanageable. However, creating more data to input to machine-learning systems simply requires a corresponding increase in the number of man-hours worked, generally without significant increases in the complexity of the annotation process.

Despite the popularity of machine learning in NLP research, symbolic methods are still (2020) commonly used

- when the amount of training data is insufficient to successfully apply machine learning methods, e.g., for the machine translation of low-resource languages such as provided by the Apertium system,
- for preprocessing in NLP pipelines, e.g., tokenization, or
- for postprocessing and transforming the output of NLP pipelines, e.g., for knowledge extraction from syntactic parses.

Statistical methods

Since the so-called "statistical revolution"[15][16] in the late 1980s and mid-1990s, much natural language processing research has relied heavily on machine learning. The machine-learning paradigm calls instead for using statistical inference to automatically learn such rules through the analysis of large *corpora* (the plural form of *corpus*, is a set of documents, possibly with human or computer annotations) of typical real-world examples.

Many different classes of machine-learning algorithms have been applied to natural-language-processing tasks. These algorithms take as input a large set of "features" that are generated from the input data. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

Some of the earliest-used machine learning algorithms, such as decision trees, produced systems of hard if-then rules similar to existing hand-written rules. However, part-of-speech tagging introduced the use of hidden Markov models to natural language processing, and increasingly, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to the features making up the input data. The cache language models upon which many speech recognition systems now rely are examples of such statistical models. Such models are generally more robust when given unfamiliar input, especially input that contains errors (as is very common for real-world data), and produce more reliable results when integrated into a larger system comprising multiple subtasks.

Since the neural turn, statistical methods in NLP research have been largely replaced by neural networks. However, they continue to be relevant for contexts in which statistical interpretability and transparency is required.

Neural networks

A major drawback of statistical methods is that they require elaborate feature engineering. Since 2015,[17] the field has thus largely abandoned statistical methods and shifted to neural networks for machine learning. Popular techniques include the use of word embeddings to capture semantic properties of words, and an increase in end-to-end learning of a higher-level task (e.g., question answering) instead of relying on a pipeline of separate intermediate tasks (e.g., part-of-speech tagging and dependency parsing). In some areas, this shift has entailed substantial changes in how NLP systems are designed, such that deep neural network-based approaches may be viewed as a new paradigm distinct from statistical natural language processing. For instance, the term *neural machine translation* (NMT) emphasizes the fact that deep learning-based approaches to machine translation directly learn sequence-to-sequence transformations, obviating the need for intermediate steps such as word alignment and language modeling that was used in statistical machine translation (SMT). Latest works tend to use non-technical structure of a given task to build proper neural network.[18]

Common NLP tasks

The following is a list of some of the most commonly researched tasks in natural language processing. Some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks.

Though natural language processing tasks are closely intertwined, they can be subdivided into categories for convenience. A coarse division is given below.

Text and speech processing

Optical character recognition (OCR)
> Given an image representing printed text, determine the corresponding text.

Speech recognition
> Given a sound clip of a person or people speaking, determine the textual representation of the speech. This is the opposite of text to speech and is one of the extremely difficult problems colloquially termed "AI-complete" (see above). In natural speech there are hardly any pauses between successive words, and thus speech segmentation is a necessary subtask of speech recognition (see below). In most spoken languages, the sounds representing successive letters blend into each other in a process termed coarticulation, so the conversion of the analog signal to discrete characters can be a very difficult process. Also, given that words in the same language are spoken by people with different accents, the speech recognition software must be able to recognize the wide variety of input as being identical to each other in terms of its textual equivalent.

Speech segmentation
> Given a sound clip of a person or people speaking, separate it into words. A subtask of speech recognition and typically grouped with it.

Text-to-speech
> Given a text, transform those units and produce a spoken representation. Text-to-speech can be used to aid the visually impaired.[19]

Word segmentation (Tokenization)
> Separate a chunk of continuous text into separate words. For a language like English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language. Sometimes this process is also used in cases like bag of words (BOW) creation in data mining.

Morphological analysis

Lemmatization
> The task of removing inflectional endings only and to return the base dictionary form of a word which is also known as a lemma. Lemmatization is another technique for reducing words to their normalized form. But in this case, the transformation actually uses a dictionary to map words to their actual form.[20]

Morphological segmentation
> Separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology (*i.e.*, the structure of words) of the language being considered. English has fairly simple morphology, especially inflectional morphology, and thus it is

often possible to ignore this task entirely and simply model all possible forms of a word (*e.g.*, "open, opens, opened, opening") as separate words. In languages such as Turkish or Meitei,[21] a highly agglutinated Indian language, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.

Part-of-speech tagging

Given a sentence, determine the part of speech (POS) for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech.

Stemming

The process of reducing inflected (or sometimes derived) words to a base form (*e.g.*, "close" will be the root for "closed", "closing", "close", "closer" etc.). Stemming yields similar results as lemmatization, but does so on grounds of rules, not a dictionary.

Syntactic analysis

Grammar induction[22]

Generate a formal grammar that describes a language's syntax.

Sentence breaking (also known as "sentence boundary disambiguation")

Given a chunk of text, find the sentence boundaries. Sentence boundaries are often marked by periods or other punctuation marks, but these same characters can serve other purposes (*e.g.*, marking abbreviations).

Parsing

Determine the parse tree (grammatical analysis) of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses: perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human). There are two primary types of parsing: *dependency parsing* and *constituency parsing*. Dependency parsing focuses on the relationships between words in a sentence (marking things like primary objects and predicates), whereas constituency parsing focuses on building out the parse tree using a probabilistic context-free grammar (PCFG) (see also *stochastic grammar*).

Lexical semantics (of individual words in context)

Lexical semantics

What is the computational meaning of individual words in context?

Distributional semantics

How can we learn semantic representations from data?

Named entity recognition (NER)

Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization). Although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case, is often inaccurate or insufficient. For example, the first letter of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized. Furthermore, many other languages in non-Western scripts (e.g. Chinese or Arabic) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names. For example, German capitalizes all nouns, regardless of whether they are names, and French and Spanish do not capitalize names that serve as adjectives.

Sentiment analysis (see also Multimodal sentiment analysis)

Extract subjective information usually from a set of documents, often using online reviews to determine "polarity" about specific objects. It is especially useful for identifying trends of public opinion in social media, for marketing.

Terminology extraction

The goal of terminology extraction is to automatically extract relevant terms from a given corpus.

Word sense disambiguation

Many words have more than one meaning; we have to select the meaning which makes the most sense in context. For this problem, we are typically given a list of words and associated word senses, e.g. from a dictionary or an online resource such as WordNet.

Relational semantics (semantics of individual sentences)

Relationship extraction

Given a chunk of text, identify the relationships among named entities (e.g. who is married to whom).

Semantic parsing

Given a piece of text (typically a sentence), produce a formal representation of its semantics, either as a graph (e.g., in AMR parsing) or in accordance with a logical formalism (e.g., in DRT parsing). This challenge typically includes aspects of several more elementary NLP tasks from semantics (e.g., semantic role labelling, word sense disambiguation) and can be extended to include full-fledged discourse analysis (e.g., discourse analysis, coreference; see Natural language understanding below).

Semantic role labelling (see also implicit semantic role labelling below)

Given a single sentence, identify and disambiguate semantic predicates (e.g., verbal frames), then identify and classify the frame elements (semantic roles).

Discourse (semantics beyond individual sentences)

Coreference resolution

Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities"). Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names to which they refer. The more general task of coreference resolution also includes identifying so-called "bridging relationships" involving referring expressions. For example, in a sentence such as "He entered John's house through the front door", "the front door" is a referring expression and the bridging relationship to be identified is the fact that the door being referred to is the front door of John's house (rather than of some other structure that might also be referred to).

Discourse analysis

This rubric includes several related tasks. One task is discourse parsing, i.e., identifying the discourse structure of a connected text, i.e. the nature of the discourse relationships between sentences (e.g. elaboration, explanation, contrast). Another possible task is recognizing and classifying the speech acts in a chunk of text (e.g. yes-no question, content question, statement, assertion, etc.).

Implicit semantic role labelling

Given a single sentence, identify and disambiguate semantic predicates (e.g., verbal frames) and their explicit semantic roles in the current sentence (see Semantic role labelling above). Then, identify semantic roles that are not explicitly realized in the current sentence, classify them into arguments that are explicitly realized elsewhere in the text and those that are not specified, and resolve the former against the local text. A closely related task is zero anaphora resolution, i.e., the extension of coreference resolution to pro-drop languages.

Recognizing textual entailment

Given two text fragments, determine if one being true entails the other, entails the other's negation, or allows the other to be either true or false.[23]

Topic segmentation and recognition

Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.

Argument mining

The goal of argument mining is the automatic extraction and identification of argumentative structures from natural language text with the aid of computer programs. [24] Such argumentative structures include the premise, conclusions, the argument scheme and the relationship between the main and subsidiary argument, or the main and counter-argument within discourse. [25][26]

Higher-level NLP applications

Automatic summarization (text summarization)

Produce a readable summary of a chunk of text. Often used to provide summaries of the text of a known type, such as research papers, articles in the financial section of a newspaper.

Book generation

Not an NLP task proper but an extension of natural language generation and other NLP tasks is the creation of full-fledged books. The first machine-generated book was created by a rule-based system in

1984 (Racter, *The policeman's beard is half-constructed*).[27] The first published work by a neural network was published in 2018, *1 the Road*, marketed as a novel, contains sixty million words. Both these systems are basically elaborate but non-sensical (semantics-free) language models. The first machine-generated science book was published in 2019 (Beta Writer, *Lithium-Ion Batteries*, Springer, Cham).[28] Unlike *Racter* and *1 the Road*, this is grounded on factual knowledge and based on text summarization.

Dialogue management
Computer systems intended to converse with a human.

Document AI
A Document AI platform sits on top of the NLP technology enabling users with no prior experience of artificial intelligence, machine learning or NLP to quickly train a computer to extract the specific data they need from different document types. NLP-powered Document AI enables non-technical teams to quickly access information hidden in documents, for example, lawyers, business analysts and accountants.[29]

Grammatical error correction
Grammatical error detection and correction involves a great band-width of problems on all levels of linguistic analysis (phonology/orthography, morphology, syntax, semantics, pragmatics). Grammatical error correction is impactful since it affects hundreds of millions of people that use or acquire English as a second language. It has thus been subject to a number of shared tasks since 2011.[30][31][32] As far as orthography, morphology, syntax and certain aspects of semantics are concerned, and due to the development of powerful neural language models such as GPT-2, this can now (2019) be considered a largely solved problem and is being marketed in various commercial applications.[33]

Machine translation
Automatically translate text from one human language to another. This is one of the most difficult problems, and is a member of a class of problems colloquially termed "AI-complete", i.e. requiring all of the different types of knowledge that humans possess (grammar, semantics, facts about the real world, etc.) to solve properly.

Natural language generation (NLG):
Convert information from computer databases or semantic intents into readable human language.

Natural language understanding (NLU)
Convert chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate. Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural language concepts. Introduction and creation of language metamodel and ontology are efficient however empirical solutions. An explicit formalization of natural language semantics without confusions with implicit assumptions such as closed-world assumption (CWA) vs. open-world assumption, or subjective Yes/No vs. objective True/False is expected for the construction of a basis of semantics formalization.[34]

Question answering
Given a human-language question, determine its answer. Typical questions have a specific right answer (such as "What is the capital of Canada?"), but sometimes open-ended questions are also considered (such as "What is the meaning of life?"). Recent works have looked at even more complex questions.[35]

General tendencies and (possible) future directions

Based on long-standing trends in the field, it is possible to extrapolate future directions of NLP. As of 2020, three trends among the topics of the long-standing series of CoNLL Shared Tasks can be observed:[36]

- Interest on increasingly abstract, "cognitive" aspects of natural language (1999-2001: shallow parsing, 2002-03: named entity recognition, 2006-09/2017-18: dependency syntax, 2004-05/2008-09 semantic role labelling, 2011-12 coreference, 2015-16: discourse parsing, 2019: semantic parsing).
- Increasing interest in multilinguality, and, potentially, multimodality (English since 1999; Spanish, Dutch since 2002; German since 2003; Bulgarian, Danish, Japanese, Portuguese, Slovenian, Swedish, Turkish since 2006; Basque, Catalan, Chinese, Greek, Hungarian, Italian, Turkish since 2007; Czech since 2009; Arabic since 2012; 2017: 40+ languages; 2018: 60+/100+ languages)
- Elimination of symbolic representations (rule-based over supervised towards weakly supervised methods, representation learning and end-to-end systems)

Cognition and NLP

Most more higher-level NLP applications involve aspects that emulate intelligent behaviour and apparent comprehension of natural language. More broadly speaking, the technical operationalization of increasingly advanced aspects of cognitive behaviour represents one of the developmental trajectories of NLP (see trends among CoNLL shared tasks above).

Cognition refers to "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses."[37] Cognitive science is the interdisciplinary, scientific study of the mind and its processes.[38] Cognitive linguistics is an interdisciplinary branch of linguistics, combining knowledge and research from both psychology and linguistics.[39] Especially during the age of symbolic NLP, the area of computational linguistics maintained strong ties with cognitive studies.

As an example, George Lakoff offers a methodology to build natural language processing (NLP) algorithms through the perspective of cognitive science, along with the findings of cognitive linguistics,[40] with two defining aspects:

1. Apply the theory of conceptual metaphor, explained by Lakoff as "the understanding of one idea, in terms of another" which provides an idea of the intent of the author.[41] For example, consider the English word *"big"*. When used in a comparison (*"That is a big tree"*), the author's intent is to imply that the tree is *"physically large"* relative to other trees or the authors experience. When used metaphorically (*"Tomorrow is a big day"*), the author's intent to imply *"importance"*. The intent behind other usages, like in *"She is a big person"* will remain somewhat ambiguous to a person and a cognitive NLP algorithm alike without additional information.

2. Assign relative measures of meaning to a word, phrase, sentence or piece of text based on the information presented before and after the piece of text being analyzed, e.g., by means of a probabilistic context-free grammar (PCFG). The mathematical equation for such algorithms is presented in US patent 9269353 (https://worldwide.espacenet.com/textdoc?DB=EPODOC&IDX=US9269353):

$$RMM(token_N) = PMM(token_N) \times \frac{1}{2d} \left(\sum_{i=-d}^{d} ((PMM(token_{N-1}) \times PF(token_N, token_{N-1}))_i \right)$$

   Where,

   RMM, is the Relative Measure of Meaning
   token, is any block of text, sentence, phrase or word
   N, is the number of tokens being analyzed
   PMM, is the Probable Measure of Meaning based on a corpora
   d, is the location of the token along the sequence of **N-1** tokens
   PF, is the Probability Function specific to a language

Ties with cognitive linguistics are part of the historical heritage of NLP, but they have been less frequently addressed since the statistical turn during the 1990s. Nevertheless, approaches to develop cognitive models towards technically operationalizable frameworks have been pursued in the context of various frameworks, e.g., of cognitive grammar,[42] functional grammar,[43] construction grammar,[44] computational psycholinguistics and cognitive neuroscience (e.g., ACT-R), however, with limited uptake in mainstream NLP (as measured by presence on major conferences[45] of the ACL). More recently, ideas of cognitive NLP have been revived as an approach to achieve explainability, e.g., under the notion of "cognitive AI".[46] Likewise, ideas of cognitive NLP are inherent to neural models multimodal NLP (although rarely made explicit).[47]

See also

- *1 the Road*
- Automated essay scoring
- Biomedical text mining
- Compound term processing
- Computational linguistics
- Computer-assisted reviewing
- Controlled natural language
- Deep learning
- Deep linguistic processing
- Distributional semantics

- Foreign language reading aid
- Foreign language writing aid
- Information extraction
- Information retrieval
- Language and Communication Technologies
- Language technology
- Latent semantic indexing
- Native-language identification
- Natural language programming
- Natural language search

References

1. Kongthon, Alisa; Sangkeettrakarn, Chatchawal; Kongyoung, Sarawoot; Haruechaiyasak, Choochart (October 27–30, 2009). *Implementing an online help desk system based on conversational agent*. MEDES '09: The International Conference on Management of Emergent Digital EcoSystems. France: ACM. doi:10.1145/1643823.1643908 (https://doi.org/10.1145%2F1643823.1643908).

2. Hutchins, J. (2005). "The history of machine translation in a nutshell" (http://www.hutchinsweb.me.uk/Nutshell-2005.pdf) (PDF).

3. Koskenniemi, Kimmo (1983), *Two-level morphology: A general computational model of word-form recognition and production* (http://www.ling.helsinki.fi/~koskenni/doc/Two-LevelMorphology.pdf) (PDF), Department of General Linguistics, University of Helsinki

4. Joshi, A. K., & Weinstein, S. (1981, August). Control of Inference: Role of Some Aspects of Discourse Structure-Centering (https://www.ijcai.org/Proceedings/81-1/Papers/071.pdf). In *IJCAI* (pp. 385-387).

5. Guida, G.; Mauri, G. (July 1986). "Evaluation of natural language processing systems: Issues and approaches". *Proceedings of the IEEE*. **74** (7): 1026–1035. doi:10.1109/PROC.1986.13580 (https://doi.org/10.1109%2FPROC.1986.13580). ISSN 1558-2256 (https://www.worldcat.org/issn/1558-2256). S2CID 30688575 (https://api.semanticscholar.org/CorpusID:30688575).

6. Chomskyan linguistics encourages the investigation of "corner cases" that stress the limits of its theoretical models (comparable to pathological phenomena in mathematics), typically created using thought experiments, rather than the systematic investigation of typical phenomena that occur in real-world data, as is the case in corpus linguistics. The creation and use of such corpora of real-world data is a fundamental part of machine-learning algorithms for natural language processing. In addition, theoretical underpinnings of Chomskyan linguistics such as the so-called "poverty of the stimulus" argument entail that general learning algorithms, as are typically used in machine learning, cannot be successful in language processing. As a result, the Chomskyan paradigm discouraged the application of such models to language processing.

7. Goldberg, Yoav (2016). "A Primer on Neural Network Models for Natural Language Processing". *Journal of Artificial Intelligence Research*. **57**: 345–420. arXiv:1807.10854 (https://arxiv.org/abs/1807.10854). doi:10.1613/jair.4992 (https://doi.org/10.1613%2Fjair.4992). S2CID 8273530 (https://api.semanticscholar.org/CorpusID:8273530).

8. Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron (2016). *Deep Learning* (http://www.deeplearningbook.org/). MIT Press.

9. Jozefowicz, Rafal; Vinyals, Oriol; Schuster, Mike; Shazeer, Noam; Wu, Yonghui (2016). *Exploring the Limits of Language Modeling*. arXiv:1602.02410 (https://arxiv.org/abs/1602.02410). Bibcode:2016arXiv160202410J (https://ui.adsabs.harvard.edu/abs/2016arXiv160202410J).

10. Choe, Do Kook; Charniak, Eugene. "Parsing as Language Modeling" (https://aclanthology.coli.uni-saarland.de/papers/D16-1257/d16-1257). *Emnlp 2016*.

11. Vinyals, Oriol; et al. (2014). "Grammar as a Foreign Language" (https://papers.nips.cc/paper/5635-grammar-as-a-foreign-language.pdf) (PDF). *Nips2015*. arXiv:1412.7449 (https://arxiv.org/abs/1412.7449). Bibcode:2014arXiv1412.7449V (https://ui.adsabs.harvard.edu/abs/2014arXiv1412.7449V).

12. Turchin, Alexander; Florez Builes, Luisa F. (2021-03-19). "Using Natural Language Processing to Measure and Improve Quality of Diabetes Care: A Systematic Review" (http://journals.sagepub.com/doi/10.1177/19322968211000831). *Journal of Diabetes Science and Technology*: 193229682110008. doi:10.1177/19322968211000831 (https://doi.org/10.1177%2F19322968211000831). ISSN 1932-2968 (https://www.worldcat.org/issn/1932-2968). PMID 33736486 (https://pubmed.ncbi.nlm.nih.gov/33736486).

13. Winograd, Terry (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language* (http://hci.stanford.edu/winograd/shrdlu/) (Thesis).

14. Schank, Roger C.; Abelson, Robert P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Hillsdale: Erlbaum. ISBN 0-470-99033-3.

15. Mark Johnson. How the statistical revolution changes (computational) linguistics. (http://www.aclweb.org/anthology/W09-0103) Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics.

16. Philip Resnik. Four revolutions. (http://languagelog.ldc.upenn.edu/nll/?p=2946) Language Log, February 5, 2011.

17. Socher, Richard. "Deep Learning For NLP-ACL 2012 Tutorial" (https://www.socher.org/index.php/Main/DeepLearningForNLP-ACL2012Tutorial). www.socher.org. Retrieved 2020-08-17. This was an early Deep Learning tutorial at the ACL 2012 and met with both interest and (at the time) skepticism by most participants. Until then, neural learning was basically rejected because of its lack of statistical interpretability. Until 2015, deep learning had evolved into the major framework of NLP.

18. Annamoradnejad, I. (2020). Colbert: Using bert sentence embedding for humor detection (https://arxiv.org/abs/2004.12765). arXiv preprint arXiv:2004.12765.

19. Yi, Chucai; Tian, Yingli (2012), "Assistive Text Reading from Complex Background for Blind Persons", Camera-Based Document Analysis and Recognition, Springer Berlin Heidelberg, pp. 15–28, CiteSeerX 10.1.1.668.869 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.668.869), doi:10.1007/978-3-642-29364-1_2 (https://doi.org/10.1007%2F978-3-642-29364-1_2), ISBN 9783642293634

20. "What is Natural Language Processing? Intro to NLP in Machine Learning" (https://www.gyansetu.in/what-is-natural-language-processing/). GyanSetu!. 2020-12-06. Retrieved 2021-01-09.

21. Kishorjit, N.; Vidya, Raj RK.; Nirmal, Y.; Sivaji, B. (2012). "Manipuri Morpheme Identification" (http://aclweb.org/anthology//W/W12/W12-5008.pdf) (PDF). Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP). COLING 2012, Mumbai, December 2012: 95–108.

22. Klein, Dan; Manning, Christopher D. (2002). "Natural language grammar induction using a constituent-context model" (http://papers.nips.cc/paper/1945-natural-language-grammar-induction-using-a-constituent-context-model.pdf) (PDF). Advances in Neural Information Processing Systems.

23. PASCAL Recognizing Textual Entailment Challenge (RTE-7) https://tac.nist.gov//2011/RTE/

24. Lippi, Marco; Torroni, Paolo (2016-04-20). "Argumentation Mining: State of the Art and Emerging Trends" (https://dl.acm.org/doi/10.1145/2850417). ACM Transactions on Internet Technology. 16 (2): 1–25. doi:10.1145/2850417 (https://doi.org/10.1145%2F2850417). ISSN 1533-5399 (https://www.worldcat.org/issn/1533-5399). S2CID 9561587 (https://api.semanticscholar.org/CorpusID:9561587).

25. "Argument Mining - IJCAI2016 Tutorial" (https://www.i3s.unice.fr/~villata/tutorialIJCAI2016.html). www.i3s.unice.fr. Retrieved 2021-03-09.

26. "NLP Approaches to Computational Argumentation – ACL 2016, Berlin" (http://acl2016tutorial.arg.tech/). Retrieved 2021-03-09.

27. "U B U W E B :: Racter" (http://www.ubu.com/historical/racter/index.html). www.ubu.com. Retrieved 2020-08-17.

28. Writer, Beta (2019). Lithium-Ion Batteries. doi:10.1007/978-3-030-16800-1 (https://doi.org/10.1007%2F978-3-030-16800-1). ISBN 978-3-030-16799-8.

29. "Document Understanding AI on Google Cloud (Cloud Next '19) - YouTube" (https://www.youtube.com/watch?v=7dtl650D0y0). www.youtube.com. Retrieved 2021-01-11.

30. Administration. "Centre for Language Technology (CLT)" (https://www.mq.edu.au/research/research-centres-groups-and-facilities/innovative-technologies/centres/centre-for-language-technology-clt). Macquarie University. Retrieved 2021-01-11.

31. "Shared Task: Grammatical Error Correction" (https://www.comp.nus.edu.sg/~nlp/conll13st.html). www.comp.nus.edu.sg. Retrieved 2021-01-11.

32. "Shared Task: Grammatical Error Correction" (https://www.comp.nus.edu.sg/~nlp/conll14st.html). www.comp.nus.edu.sg. Retrieved 2021-01-11.

33. "About Us | Grammarly" (https://www.grammarly.com/about). www.grammarly.com. Retrieved 2021-01-11.

34. Duan, Yucong; Cruz, Christophe (2011). "Formalizing Semantic of Natural Language through Conceptualization from Existence" (https://web.archive.org/web/20111009135952/http://www.ijimt.org/abstract/100-E00187.htm). International Journal of Innovation, Management and Technology. 2 (1): 37–42. Archived from the original (http://www.ijimt.org/abstract/100-E00187.htm) on 2011-10-09.

35. Mittal (2011). "Versatile question answering systems: seeing in synthesis" (https://hal.archives-ouvertes.fr/hal-01104648/file/Mittal_VersatileQA_IJIIDS.pdf) (PDF). International Journal of Intelligent Information and Database Systems. 5 (2): 119–142. doi:10.1504/IJIIDS.2011.038968 (https://doi.org/10.1504%2FIJIIDS.2011.038968).

36. "Previous shared tasks | CoNLL" (https://www.conll.org/previous-tasks). www.conll.org. Retrieved 2021-01-11.

37. "Cognition" (https://www.lexico.com/definition/cognition). *Lexico*. Oxford University Press and Dictionary.com. Retrieved 6 May 2020.

38. "Ask the Cognitive Scientist" (http://www.aft.org/newspubs/periodicals/ae/summer2002/willingham.cfm). *American Federation of Teachers*. 8 August 2014. "Cognitive science is an interdisciplinary field of researchers from Linguistics, psychology, neuroscience, philosophy, computer science, and anthropology that seek to understand the mind."

39. Robinson, Peter (2008). *Handbook of Cognitive Linguistics and Second Language Acquisition*. Routledge. pp. 3–8. ISBN 978-0-805-85352-0.

40. Lakoff, George (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Philosophy; Appendix: The Neural Theory of Language Paradigm*. New York Basic Books. pp. 569–583. ISBN 978-0-465-05674-3.

41. Strauss, Claudia (1999). *A Cognitive Theory of Cultural Meaning*. Cambridge University Press. pp. 156–164. ISBN 978-0-521-59541-4.

42. "Universal Conceptual Cognitive Annotation (UCCA)" (https://universalconceptualcognitiveannotation.github.io/). *Universal Conceptual Cognitive Annotation (UCCA)*. Retrieved 2021-01-11.

43. Rodríguez, F. C., & Mairal-Usón, R. (2016). Building an RRG computational grammar (https://www.redalyc.org/pdf/1345/134549291020.pdf). *Onomazein*, (34), 86-117.

44. "Fluid Construction Grammar – A fully operational processing system for construction grammars" (https://www.fcg-net.org/). Retrieved 2021-01-11.

45. "ACL Member Portal | The Association for Computational Linguistics Member Portal" (https://www.aclweb.org/portal/). *www.aclweb.org*. Retrieved 2021-01-11.

46. "Chunks and Rules" (https://www.w3.org/Data/demos/chunks/chunks.html). *www.w3.org*. Retrieved 2021-01-11.

47. Socher, Richard; Karpathy, Andrej; Le, Quoc V.; Manning, Christopher D.; Ng, Andrew Y. (2014). "Grounded Compositional Semantics for Finding and Describing Images with Sentences" (https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00177). *Transactions of the Association for Computational Linguistics*. **2**: 207–218. doi:10.1162/tacl_a_00177 (https://doi.org/10.1162%2Ftacl_a_00177). S2CID 2317858 (https://api.semanticscholar.org/CorpusID:2317858).

Further reading

- Bates, M (1995). "Models of natural language understanding" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC40721). *Proceedings of the National Academy of Sciences of the United States of America*. **92** (22): 9977–9982. Bibcode:1995PNAS...92.9977B (https://ui.adsabs.harvard.edu/abs/1995PNAS...92.9977B). doi:10.1073/pnas.92.22.9977 (https://doi.org/10.1073%2Fpnas.92.22.9977). PMC 40721 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC40721). PMID 7479812 (https://pubmed.ncbi.nlm.nih.gov/7479812).

- Steven Bird, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media. ISBN 978-0-596-51649-9.

- Daniel Jurafsky and James H. Martin (2008). *Speech and Language Processing*, 2nd edition. Pearson Prentice Hall. ISBN 978-0-13-187321-6.

- Mohamed Zakaria Kurdi (2016). *Natural Language Processing and Computational Linguistics: speech, morphology, and syntax*, Volume 1. ISTE-Wiley. ISBN 978-1848218482.

- Mohamed Zakaria Kurdi (2017). *Natural Language Processing and Computational Linguistics: semantics, discourse, and applications*, Volume 2. ISTE-Wiley. ISBN 978-1848219212.

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press. ISBN 978-0-521-86571-5. Official html and pdf versions available without charge. (http://nlp.stanford.edu/IR-book/)

- Christopher D. Manning and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press. ISBN 978-0-262-13360-9.

- David M. W. Powers and Christopher C. R. Turk (1989). *Machine Learning of Natural Language*. Springer-Verlag. ISBN 978-0-387-19557-5.