Major Tasks within NLP

The following is a list of some of the most commonly researched tasks in NLP. Note that 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. What distinguishes these tasks from other potential and actual NLP tasks is not only the volume of research devoted to them but the fact that for each one there is typically a well-defined problem setting, a standard metric for evaluating the task, standard [corpora](http://en.wikipedia.org/wiki/Corpora) on which the task can be evaluated, and competitions devoted to the specific task.

[Automatic summarization](http://en.wikipedia.org/wiki/Automatic_summarization): Produce a readable summary of a chunk of text. Often used to provide summaries of text of a known type, such as articles in the financial section of a newspaper.

[Coreference resolution](http://en.wikipedia.org/w/index.php?title=Coreference_resolution&action=edit&redlink=1): Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities"). [Anaphora resolution](http://en.wikipedia.org/wiki/Anaphora_resolution) is a specific example of this task, and is specifically concerned with matching up [pronouns](http://en.wikipedia.org/wiki/Pronoun) with the nouns or names that they refer to. The more general task of coreference resolution also includes identify so-called "bridging relationships" involving [referring expressions](http://en.wikipedia.org/wiki/Referring_expression). 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](http://en.wikipedia.org/wiki/Discourse_analysis): This rubric includes a number of related tasks. One task is identifying the [discourse structure](http://en.wikipedia.org/w/index.php?title=Discourse_structure&action=edit&redlink=1) of 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](http://en.wikipedia.org/wiki/Speech_act) in a chunk of text (e.g. yes-no question, content question, statement, assertion, etc.).

[Machine translation](http://en.wikipedia.org/wiki/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](http://en.wikipedia.org/wiki/AI-complete)", i.e. requiring all of the different types of knowledge that humans possess (grammar, semantics, facts about the real world, etc.) in order to solve properly.

[Morphological segmentation](http://en.wikipedia.org/w/index.php?title=Morphological_segmentation&action=edit&redlink=1): Separate words into individual [morphemes](http://en.wikipedia.org/wiki/Morphemes) and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the [morphology](http://en.wikipedia.org/wiki/Morphology_(linguistics)) (i.e. the structure of words) of the language being considered. [English](http://en.wikipedia.org/wiki/English_language) has fairly simple morphology, especially [inflectional morphology](http://en.wikipedia.org/wiki/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](http://en.wikipedia.org/wiki/Turkish_language), however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.

[Named entity recognition](http://en.wikipedia.org/wiki/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). Note that, although [capitalization](http://en.wikipedia.org/wiki/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 word 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](http://en.wikipedia.org/wiki/Chinese_language) or [Arabic](http://en.wikipedia.org/wiki/Arabic_language)) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names. For example, [German](http://en.wikipedia.org/wiki/German_language) capitalizes all [nouns](http://en.wikipedia.org/wiki/Noun), regardless of whether they refer to names, and [French](http://en.wikipedia.org/wiki/French_language) and [Spanish](http://en.wikipedia.org/wiki/Spanish_language) do not capitalize names that serve as [adjectives](http://en.wikipedia.org/wiki/Adjective).

[Natural language generation](http://en.wikipedia.org/wiki/Natural_language_generation): Convert information from computer databases into readable human language.

[Natural language understanding](http://en.wikipedia.org/wiki/Natural_language_understanding): Convert chunks of text into more formal representations such as [first-order logic](http://en.wikipedia.org/wiki/First-order_logic) structures that are easier for [computer](http://en.wikipedia.org/wiki/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 languages concepts. Introduction and creation of language metamodel and ontology are efficient however empirical solutions. An explicit formalization of natural languages 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.[[5]](http://en.wikipedia.org/wiki/Natural_language_processing#cite_note-4)

[Optical character recognition](http://en.wikipedia.org/wiki/Optical_character_recognition) (OCR): Given an image representing printed text, determine the corresponding text.

[Part-of-speech tagging](http://en.wikipedia.org/wiki/Part-of-speech_tagging): Given a sentence, determine the [part of speech](http://en.wikipedia.org/wiki/Part_of_speech) for each word. Many words, especially common ones, can serve as multiple [parts of speech](http://en.wikipedia.org/wiki/Parts_of_speech). For example, "book" can be a [noun](http://en.wikipedia.org/wiki/Noun) ("the book on the table") or [verb](http://en.wikipedia.org/wiki/Verb) ("to book a flight"); "set" can be a [noun](http://en.wikipedia.org/wiki/Noun), [verb](http://en.wikipedia.org/wiki/Verb) or [adjective](http://en.wikipedia.org/wiki/Adjective); and "out" can be any of at least five different parts of speech. Note that some languages have more such ambiguity than others. Languages with little [inflectional morphology](http://en.wikipedia.org/wiki/Inflectional_morphology), such as [English](http://en.wikipedia.org/wiki/English_language) are particularly prone to such ambiguity. [Chinese](http://en.wikipedia.org/wiki/Chinese_language) is prone to such ambiguity because it is a [tonal language](http://en.wikipedia.org/wiki/Tonal_language) during verbalization. Such inflection is not readily conveyed via the entities employed within the orthography to convey intended meaning.

[Parsing](http://en.wikipedia.org/wiki/Parsing): Determine the [parse tree](http://en.wikipedia.org/wiki/Parse_tree) (grammatical analysis) of a given sentence. The [grammar](http://en.wikipedia.org/wiki/Grammar) for [natural languages](http://en.wikipedia.org/wiki/Natural_language) is [ambiguous](http://en.wikipedia.org/wiki/Ambiguous) and typical sentences have multiple possible analyses. In fact, perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human).

[Question answering](http://en.wikipedia.org/wiki/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?").

[Relationship extraction](http://en.wikipedia.org/wiki/Relationship_extraction): Given a chunk of text, identify the relationships among named entities (e.g. who is the wife of whom).

[Sentence breaking](http://en.wikipedia.org/wiki/Sentence_breaking) (also known as [sentence boundary disambiguation](http://en.wikipedia.org/wiki/Sentence_boundary_disambiguation)): Given a chunk of text, find the sentence boundaries. Sentence boundaries are often marked by [periods](http://en.wikipedia.org/wiki/Full_stop) or other [punctuation marks](http://en.wikipedia.org/wiki/Punctuation_mark), but these same characters can serve other purposes (e.g. marking [abbreviations](http://en.wikipedia.org/wiki/Abbreviation)).

[Sentiment analysis](http://en.wikipedia.org/wiki/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 the social media, for the purpose of marketing.

[Speech recognition](http://en.wikipedia.org/wiki/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](http://en.wikipedia.org/wiki/Text_to_speech) and is one of the extremely difficult problems colloquially termed "[AI-complete](http://en.wikipedia.org/wiki/AI-complete)" (see above). In [natural speech](http://en.wikipedia.org/wiki/Natural_speech) there are hardly any pauses between successive words, and thus [speech segmentation](http://en.wikipedia.org/wiki/Speech_segmentation) is a necessary subtask of speech recognition (see below). Note also that in most spoken languages, the sounds representing successive letters blend into each other in a process termed [coarticulation](http://en.wikipedia.org/wiki/Coarticulation" \o "Coarticulation), so the conversion of the analog signal to discrete characters can be a very difficult process.

[Speech segmentation](http://en.wikipedia.org/wiki/Speech_segmentation): Given a sound clip of a person or people speaking, separate it into words. A subtask of [speech recognition](http://en.wikipedia.org/wiki/Speech_recognition) and typically grouped with it.

[Topic segmentation](http://en.wikipedia.org/wiki/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.

[Word segmentation](http://en.wikipedia.org/wiki/Word_segmentation): Separate a chunk of continuous text into separate words. For a language like [English](http://en.wikipedia.org/wiki/English_language), this is fairly trivial, since words are usually separated by spaces. However, some written languages like [Chinese](http://en.wikipedia.org/wiki/Chinese_language), [Japanese](http://en.wikipedia.org/wiki/Japanese_language) and [Thai](http://en.wikipedia.org/wiki/Thai_language) do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the [vocabulary](http://en.wikipedia.org/wiki/Vocabulary) and[morphology](http://en.wikipedia.org/wiki/Morphology_(linguistics)) of words in the language.

[Word sense disambiguation](http://en.wikipedia.org/wiki/Word_sense_disambiguation): Many words have more than one [meaning](http://en.wikipedia.org/wiki/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 from an online resource such as [WordNet](http://en.wikipedia.org/wiki/WordNet" \o "WordNet).

In some cases, sets of related tasks are grouped into subfields of NLP that are often considered separately from NLP as a whole. Examples include:

[Information retrieval](http://en.wikipedia.org/wiki/Information_retrieval) (IR): This is concerned with storing, searching and retrieving information. It is a separate field within computer science (closer to databases), but IR relies on some NLP methods (for example, stemming). Some current research and applications seek to bridge the gap between IR and NLP.

[Information extraction](http://en.wikipedia.org/wiki/Information_extraction) (IE): This is concerned in general with the extraction of semantic information from text. This covers tasks such as [named entity recognition](http://en.wikipedia.org/wiki/Named_entity_recognition), [coreference resolution](http://en.wikipedia.org/w/index.php?title=Coreference_resolution&action=edit&redlink=1" \o "Coreference resolution (page does not exist)), [relationship extraction](http://en.wikipedia.org/wiki/Relationship_extraction), etc.

[Speech processing](http://en.wikipedia.org/wiki/Speech_processing): This covers [speech recognition](http://en.wikipedia.org/wiki/Speech_recognition), [text-to-speech](http://en.wikipedia.org/wiki/Text-to-speech) and related tasks.

Other tasks include:

[Stemming](http://en.wikipedia.org/wiki/Stemming) is the process for reducing inflected (or sometimes derived) words to their [stem](http://en.wikipedia.org/wiki/Word_stem), base or [root](http://en.wikipedia.org/wiki/Root_(linguistics)) form—generally a written word form. The stem need not be identical to the [morphological root](http://en.wikipedia.org/wiki/Morphological_root) of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. [Algorithms](http://en.wikipedia.org/wiki/Algorithm) for stemming have been studied in [computer science](http://en.wikipedia.org/wiki/Computer_science) since 1968. Many [search engines](http://en.wikipedia.org/wiki/Search_engine) treat words with the same stem as [synonyms](http://en.wikipedia.org/wiki/Synonym) as a kind of [query broadening](http://en.wikipedia.org/w/index.php?title=Query_broadening&action=edit&redlink=1), a process called conflation. Stemming programs are commonly referred to as stemming algorithms or stemmers.

[Text simplification](http://en.wikipedia.org/wiki/Text_simplification) is an operation used in [natural language processing](http://en.wikipedia.org/wiki/Natural_language_processing) to modify, enhance, classify or otherwise process an existing corpus of human-readable text in such a way that the grammar and structure of the prose is greatly simplified, while the underlying [meaning](http://en.wikipedia.org/wiki/Meaning_(linguistic)) and [information](http://en.wikipedia.org/wiki/Information) remains the same. Text simplification is an important area of research, because natural human languages ordinarily contain complex compound constructions that are not easily processed through [automation](http://en.wikipedia.org/wiki/Automation).

[Text-to-speech](http://en.wikipedia.org/wiki/Text-to-speech) is the artificial production of human [speech](http://en.wikipedia.org/wiki/Speech). A computer system used for this purpose is called a speech synthesizer, and can be implemented in [software](http://en.wikipedia.org/wiki/Software) or [hardware](http://en.wikipedia.org/wiki/Computer_hardware).

[Text-proofing](http://en.wikipedia.org/wiki/Text-proofing) digital proof reading.

[Natural language search](http://en.wikipedia.org/wiki/Natural_language_user_interface) are a type of [computer human interface](http://en.wikipedia.org/wiki/User_interface) where linguistic phenomena such as verbs, phrases and clauses act as UI controls for creating, selecting and modifying data in software applications.

In [interface design](http://en.wikipedia.org/wiki/Interface_design) natural language interfaces are sought after for their speed and ease of use, but most suffer the challenges to understanding wide varieties of ambiguous input.[[1]](http://en.wikipedia.org/wiki/Natural_language_user_interface#cite_note-0) Natural language interfaces are an active area of study in the field of [natural language processing](http://en.wikipedia.org/wiki/Natural_language_processing) and [Computational linguistics](http://en.wikipedia.org/wiki/Computational_linguistics). An intuitive general Natural language interface is one of the active goals of the [Semantic Web](http://en.wikipedia.org/wiki/Semantic_Web).

[Query expansion](http://en.wikipedia.org/wiki/Query_expansion)  is the process of reformulating a seed query to improve retrieval performance in [information retrieval](http://en.wikipedia.org/wiki/Information_retrieval) operations.[[1]](http://en.wikipedia.org/wiki/Query_expansion#cite_note-0) In the context of web [search engines](http://en.wikipedia.org/wiki/Search_engine), query expansion involves evaluating a user's input (what words were typed into the search query area, and sometimes other types of [data](http://en.wikipedia.org/wiki/Data)) and expanding the search query to match additional documents. Query expansion involves techniques such as:

Finding [synonyms](http://en.wikipedia.org/wiki/Synonym) of words, and searching for the synonyms as well

Finding all the various [morphological](http://en.wikipedia.org/wiki/Morphology_(linguistics)) forms of words by [stemming](http://en.wikipedia.org/wiki/Stemming) each word in the [search query](http://en.wikipedia.org/wiki/Search_query)

Fixing [spelling errors](http://en.wikipedia.org/wiki/Typographic_error) and automatically searching for the corrected form or suggesting it in the results

Re-weighting the terms in the original query

[Truecasing](http://en.wikipedia.org/wiki/Truecasing)  is the problem in [natural language processing](http://en.wikipedia.org/wiki/Natural_language_processing) (NLP) of determining the proper [capitalization](http://en.wikipedia.org/wiki/Capitalization) of words where such information is unavailable. This commonly comes up due to the standard practice (in [English](http://en.wikipedia.org/wiki/English_language) and many other languages) of automatically capitalizing the first word of a sentence. It can also arise in badly-cased or noncased text (for example, all-lowercase or all-uppercase[text messages](http://en.wikipedia.org/wiki/Text_messages)). Truecasing aids in many other NLP tasks, such as [named entity recognition](http://en.wikipedia.org/wiki/Named_entity_recognition), [machine translation](http://en.wikipedia.org/wiki/Machine_translation) and [automatic content extraction](http://en.wikipedia.org/w/index.php?title=Automatic_content_extraction&action=edit&redlink=1)[[1]](http://en.wikipedia.org/wiki/Truecasing#cite_note-0).

Reading List

<http://www.inf.ed.ac.uk/teaching/courses/anlp/reading.html>

University of Edinburgh’s School of Informatics

ADVANCED NATURAL LANGUAGE PROCESSING

READING LIST

The course will use the following text book:

Daniel Jurafsky and James H. Martin (2009). Speech and Language Processing (2nd Edition). Prentice Hall.

The following papers serve as background reading:

Srinivas Bangalore and Aravind K. Joshi (1999). [Supertagging: An Approach to Almost Parsing](http://www.aclweb.org/anthology-new/J/J99/J99_2004.pdf). Computational Linguistics 25(2), 237-265.

Frank Keller (2010). [Cognitively Plausible Models of Human Language Processing.](http://www.aclweb.org/anthology/P/P10/P10-2012.pdf) Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Short Papers, 60-67. Uppsala, Sweden.

Philipp Koehn (2010). [Statistical Machine Translation](http://www.statmt.org/book/). Cambridge University Press.

Vincent Ng (2010). [Supervised Noun Phrase Coreference Research: The First Fifteen Years.](http://www.aclweb.org/anthology-new/P/P10/P10_1142.pdf) Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 1396-1411. Uppsala, Sweden.

Joakim Nivre (2010). [Dependency Parsing.](http://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2010.00187.x/pdf) Language and Linguistics Compass 4/3, 138-152. [Accessible only from within the university domain.]

Michael Strube (2007). [Corpus-based and machine learning approaches to anaphora resolution: A critical assessment.](http://www.inf.ed.ac.uk/teaching/courses/anlp/readings/strube.inbook07.pdf) In M. Schwarz-Friesel, M. Consten and M. Knees (eds.), Anaphors in Text, 207-222. John Benjamins. [Accessible only from within the university domain.]

Bonnie Webber, Markus Egg and Valia Kordoni (2010). [Discourse Structure for Language Technology](http://www.inf.ed.ac.uk/teaching/courses/anlp/readings/wek10_4.pdf). Submitted to Journal of Natural Language Engineering. [Accessible only from within the university domain.]

<http://www.nltk.org/book>

Natural Language Toolk Kit. (Uses Python)

<http://www.gelbukh.com/clbook/>

Computational Linguistics book Online

<http://en.wikipedia.org/wiki/Blocks_world>

Blocks Word

The blocks world is one of the most famous planning domains in [artificial intelligence](http://en.wikipedia.org/wiki/Artificial_intelligence). The program was created by [Terry Winograd](http://en.wikipedia.org/wiki/Terry_Winograd) and is a limited-domain [natural-language](http://en.wikipedia.org/wiki/Natural_language) system that can understand typed commands and move blocks around on a surface.