Table of Contents

References and Readings

Readings

- Speech and Language Processing (Daniel Jurafsky and James Martin)
  Second Edition, Prentice-Hall, 2008
  ISBN: 0131873210
- Christopher Manning and Hinrich Schütze (1999) Foundations of Statistical Natural Language Processing, Cambridge, Massachusetts, USA. MIT Press.
  https://web.stanford.edu/ jurafsky/slp3/
- Bird et al - NLTK
  http://www.nltk.org

Definition

Artificial Intelligence

Artificial Intelligence is the science and engineering of making intelligent machines.

Definition

Artificial Intelligence

Artificial Intelligence is the science and engineering of making intelligent machines.

Aim of Natural Language Processing

To build intelligent computers that can interact with human being like a human being.

Definition

Natural Language Processing

Natural Language Processing (NLP) is the study of the computational treatment of natural language.

Definition

Natural Language Processing

Natural Language Processing (NLP) is the study of the computational treatment of natural language.

Natural Language Processing

NLP is the application of computational techniques to the analysis and synthesis of natural language and speech.

Definition

What is computational linguistics?

Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "knowledge-based" ("handcrafted") or "data-driven" ("statistical" or "empirical"). Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system. Indeed, the work of computational linguists is incorporated into many working systems today, including speech recognition systems, text-to-speech synthesizers, automated voice response systems, web search engines, text editors, language instruction materials, to name just a few.

Some Info

Prerequisite

Linear algebra: vectors and matrices
Probabilities: random variables, discrete and continuous distributions, Bayes theorem
Programming: Python in a UNIX environment, text manipulation

The Alphabet Soup

NLP (Natural Language Processing)
CL (Computational Linguistics)
IR (Information Retrieval)
SP (Speech Processing)
HLT (Human Language Technology)
NLE (Natural Language Engineering)
ML (Machine Learning)

History

Natural Language Processing had its birth during the world war II. Turing's Enigma decoding machines Bombe was used by UK to decode German messages. Later during the cold war it was developed more sophisticatedly by different countries

1. Machine translation (MT) was the first computer-based application related to natural language and used during World War II. MT took the simplistic view that the only differences between languages resided in their vocabularies and the permitted word orders. Systems developed from this perspective simply used dictionary-lookup for appropriate words for translation and reordered the words after translation to fit the word-order rules of the target language, without taking into account the lexical ambiguity inherent in natural language. This produced poor results

History

1. The apparent failure made researchers realize that the task was a lot harder than anticipated, and they needed a more adequate theory of language. However, it was not until 1957 when Chomsky published Syntactic Structures introducing the idea of generative grammar, did the field gain better insight into whether or how mainstream linguistics could help MT

History

1. During this period, other NLP application areas began to emerge, such as speech recognition. The language processing community and the speech community then was split into two camps with the language processing community dominated by the theoretical perspective of generative grammar and hostile to statistical methods, and the speech community dominated by statistical information theory and hostile to theoretical linguistics.

History

over-enthusiasm

Due to the developments of the syntactic theory of language and parsing algorithms, there was over-enthusiasm in the 1950s that people believed that fully automatic high quality translation systems would be able to produce results indistinguishable from those of human translators, and such systems should be in operation within a few years. It was not only unrealistic given the then-available linguistic knowledge and computer systems, but also impossible in principle.

History

over-enthusiasm

Due to the developments of the syntactic theory of language and parsing algorithms, there was over-enthusiasm in the 1950s that people believed that fully automatic high quality translation systems would be able to produce results indistinguishable from those of human translators, and such systems should be in operation within a few years. It was not only unrealistic given the then-available linguistic knowledge and computer systems, but also impossible in principle.

Big Blow

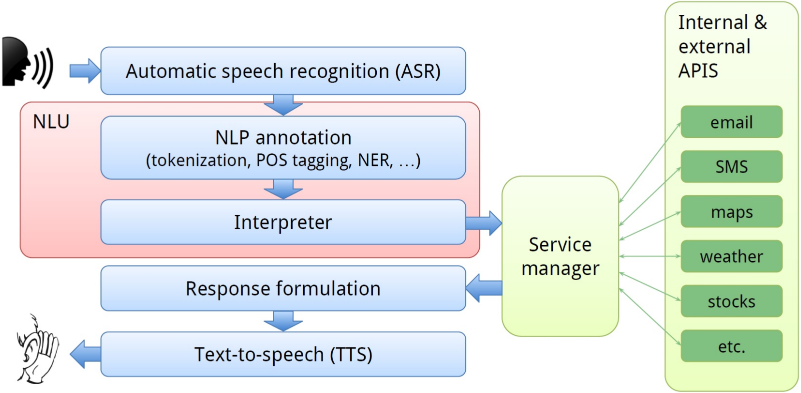
The inadequacies of then-existing systems, led to the ALPAC (Automatic Language Processing Advisory Committee of the National Academy of Science - National Research Council) report of 1966 concluding that MT was not immediately achievable and recommended it not be funded. It resulted in halting NLP research.

History

Ups and Downs

- 1960s: Pattern-matching with small rule-sets
- 1970-80s: Linguistically rich, logic-driven, grounded systems; restricted applications
- 1990s: the statistical revolution in NLP leads to a decrease in NLP work
- 2010s: NLP returns to center stage, mixing techniques from previous decades
- In industry, an explosion in products & services that rely on NLP (Siri, Google Now, Microsoft Cortana, Amazon Echo, ...)
- Systems are impressive, but show their weaknesses quickly. NLP is far from solved and big breakthroughs lie in the future

How do conversational agents work?



Two distinct Focuses of NLP

Natural Language Understanding

This refers to the analysis of language for the purpose of producing a meaningful representation. The task of Natural Language understanding is equivalent to the role of reader/listener. Understanding involves the tasks like mapping the given input in natural language into useful representations and analyzing different aspects of the language.

Two distinct Focuses of NLP

Natural Language Understanding

This refers to the analysis of language for the purpose of producing a meaningful representation. The task of Natural Language understanding is equivalent to the role of reader/listener. Understanding involves the tasks like mapping the given input in natural language into useful representations and analyzing different aspects of the language.

Natural Language Generation

It refers to the process of producing meaningful phrases and sentences in the form of natural language from some internal representation. The task of Natural Language Generation is that of the writer/speaker. It involves various tasks like Text planning (retrieving the relevant content from knowledge base), Sentence planning (choosing required words, forming meaningful phrases), Text Realization (mapping sentence plan into sentence structure)

Stages of Language Processing

- Phonetics and Phonology (interpretation of speech sounds within and across words)
- Morphology (componential nature of words)
- Lexical Analysis (Making dictionaries and storage of words)
- Syntactic Analysis (Parsing of Sentences)
- Semantic Analysis (Meaning representation)
- Pragmatics (purposeful use of language in situations and context)
- Discourse (Process connected sentences)

The Challenge

Ambiguity

The main challenge of NLP is the understanding and modeling of elements within a variable context. In language, words are unique but can have different meanings depending on the context in which they are being evaluated. Ambiguity persists in all the stages of NLP

Phonetics and Phonology

- Homophones (sounds similar): bank (finance), bank (river bank)
- Near Homophones (those with very close sound): accept vs except, affect vs effect, grate vs great, brake vs break, groan vs grown, here vs hear, ball vs bawl, berry vs bury
- word boundary: aajaayenge (will come vs will come today), Carp rice vs car price, I got up late vs I got a plate, fox vs folks
- phrase boundary: Godisnowhere, ground breaking
- Disfluency: ah, um, hmm, etc which gives the user to organise the thought

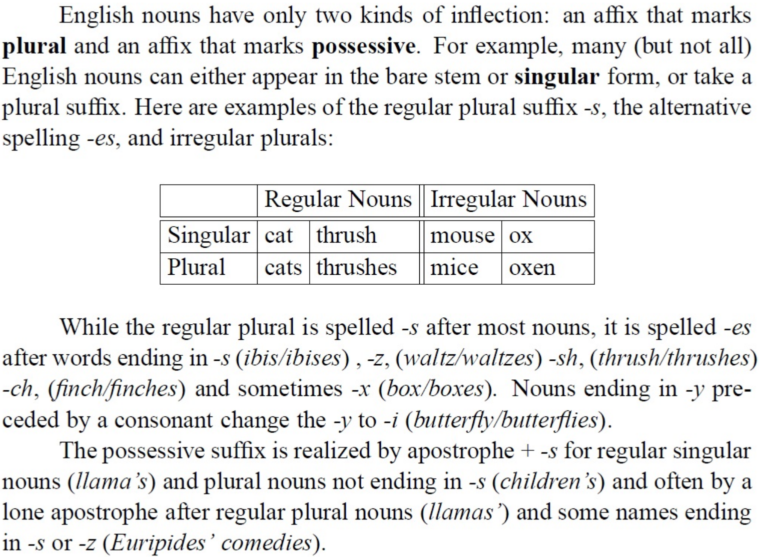
Morphology

- It is the first crucial step in NLU
- It is the study of the way the words are built up from smaller meaning-bearing units (morphemes)
- Two broad classes of morphemes - stems and affixes. The stem is the main morpheme of the word (ex. sing is the stem for singing). Affixes are additional meanings of the morpheme (ex. prefix, suffix, infixes)
- A word can have more than one affix. (ex. unbelievably = un+believe+able+ly)
- Language rich in morphology are Dravidian (Telugu, Tamil, Kannada), Hungarian, Turkish. Languages poor in morphology are English and Chinese. Languages with higher morphology have the advantage of easier processing at higher stages of processing.

Morphology

- Finite State Machines came in handy while handling Morphemes
- There are many ways to combine morphemes to create words. Four methods are Inflectional morphology, derivational morphology, compounding and cliticization

Inflectional morphology



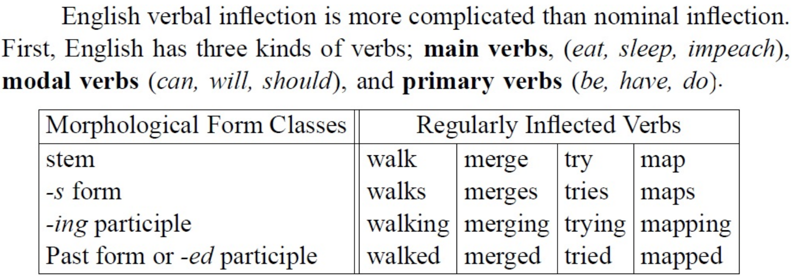
English nouns have only two kinds of inflection: an affix that marks **plural** and an affix that marks **possessive**. For example, many (but not all) English nouns can either appear in the bare stem or **singular** form, or take a plural suffix. Here are examples of the regular plural suffix -*s*, the alternative spelling -*es*, and irregular plurals:

	Regular Nouns		Irregular Nouns	
Singular	cat	thrush	mouse	ox
Plural	cats	thrushes	mice	oxen

While the regular plural is spelled -*s* after most nouns, it is spelled -*es* after words ending in -*s* (*ibis/ibises*) , -*z*, (*waltz/waltzes*) -*sh*, (*thrush/thrushes*) -*ch*, (*finch/finches*) and sometimes -*x* (*box/boxes*). Nouns ending in -*y* preceded by a consonant change the -*y* to -*i* (*butterfly/butterflies*).

The possessive suffix is realized by apostrophe + -*s* for regular singular nouns (*llama’s*) and plural nouns not ending in -*s* (*children’s*) and often by a lone apostrophe after regular plural nouns (*llamas’*) and some names ending in -*s* or -*z* (*Euripides’ comedies*).

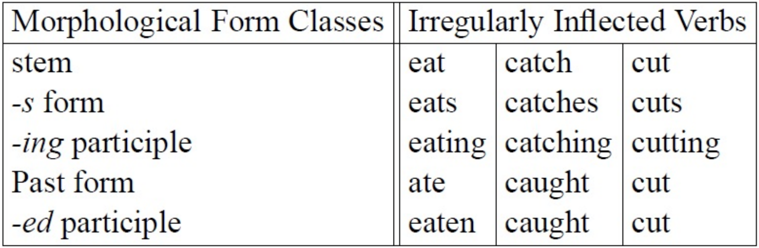
Inflectional morphology



English verbal inflection is more complicated than nominal inflection. First, English has three kinds of verbs; **main verbs**, (*eat, sleep, impeach*), **modal verbs** (*can, will, should*), and **primary verbs** (*be, have, do*).

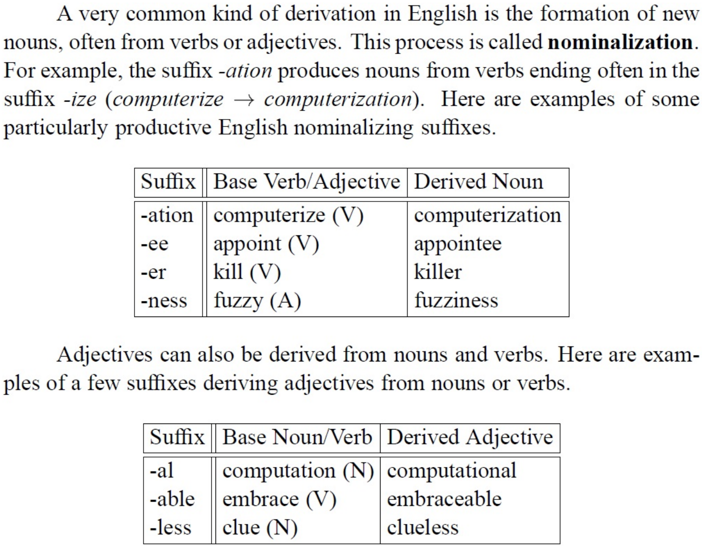
Morphological Form Classes	Regularly Inflected Verbs			
stem	walk	merge	try	map
-*s* form	walks	merges	tries	maps
-*ing* participle	walking	merging	trying	mapping
Past form or -*ed* participle	walked	merged	tried	mapped

Inflectional morphology



Morphological Form Classes	Irregularly Inflected Verbs		
stem	eat	catch	cut
-*s* form	eats	catches	cuts
-*ing* participle	eating	catching	cutting
Past form	ate	caught	cut
-*ed* participle	eaten	caught	cut

Derivational morphology



A very common kind of derivation in English is the formation of new nouns, often from verbs or adjectives. This process is called **nominalization**. For example, the suffix *-ation* produces nouns from verbs ending often in the suffix *-ize* (*computerize* → *computerization*). Here are examples of some particularly productive English nominalizing suffixes.

Suffix	Base Verb/Adjective	Derived Noun
-ation	computerize (V)	computerization
-ee	appoint (V)	appointee
-er	kill (V)	killer
-ness	fuzzy (A)	fuzziness

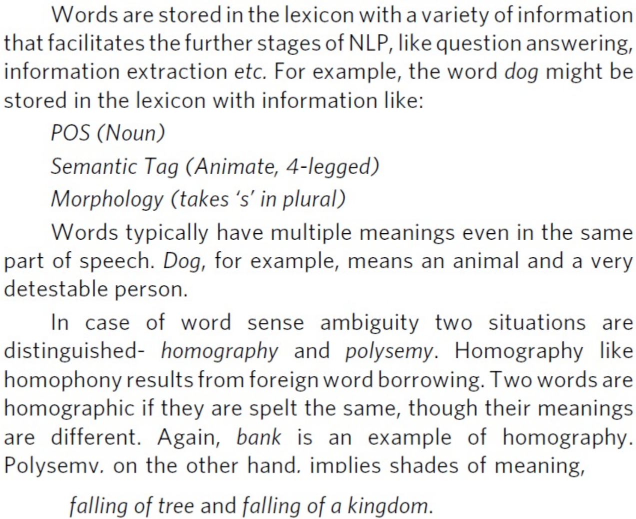
Adjectives can also be derived from nouns and verbs. Here are examples of a few suffixes deriving adjectives from nouns or verbs.

Suffix	Base Noun/Verb	Derived Adjective
-al	computation (N)	computational
-able	embrace (V)	embraceable
-less	clue (N)	clueless

Compounding and Cliticization Morphology

- Compounding is the combination of multiple word stems together. Ex. doghouse, bagpiper, etc.
- Cliticization is the combination of a word stem with a clitic. Ex. am-'m, are-'re, is-'s, will-'ll, have-'ve, had-'d, would-'d, etc.

Lexical Analysis



Words are stored in the lexicon with a variety of information that facilitates the further stages of NLP, like question answering, information extraction *etc*. For example, the word *dog* might be stored in the lexicon with information like:

POS (Noun)

Semantic Tag (Animate, 4-legged)

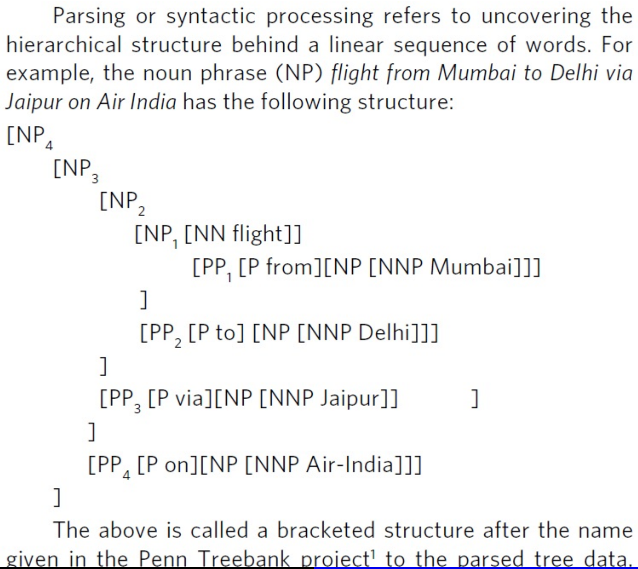
Morphology (takes 's' in plural)

Words typically have multiple meanings even in the same part of speech. *Dog*, for example, means an animal and a very detestable person.

In case of word sense ambiguity two situations are distinguished- *homography* and *polysemy*. Homography like homophony results from foreign word borrowing. Two words are homographic if they are spelt the same, though their meanings are different. Again, *bank* is an example of homography. Polysemy, on the other hand, implies shades of meaning,

falling of tree and *falling of a kingdom*.

Sintax Analysis



Parsing or syntactic processing refers to uncovering the hierarchical structure behind a linear sequence of words. For example, the noun phrase (NP) *flight from Mumbai to Delhi via Jaipur on Air India* has the following structure:

[NP$_4$

[NP$_3$

[NP$_2$

[NP$_1$ [NN flight]]

[PP$_1$ [P from][NP [NNP Mumbai]]]

]

[PP$_2$ [P to] [NP [NNP Delhi]]]

]

[PP$_3$ [P via][NP [NNP Jaipur]]]

]

[PP$_4$ [P on][NP [NNP Air-India]]]

]

The above is called a bracketed structure after the name given in the Penn Treebank project[1] to the parsed tree data.

Sintax Analysis

Now, parsing too faces the challenge of ambiguity called *structural ambiguity*. Structural ambiguity is of two kinds: *scope ambiguity* and *attachment ambiguity*. We give examples of these kinds of ambiguity:

Old men and women were taken to safe locations.

The scope of the adjective (*i.e.*, the amount of text it qualifies) is ambiguous. That is, is the structure *(old men and women)* or *((old men) and women)*?

Another example of scope ambiguity is:

No smoking areas will allow hookahs inside.

Here *no* can qualify the rest of the sentence, meaning thereby *there isn't a smoking area that will allow hookas inside.*

Or

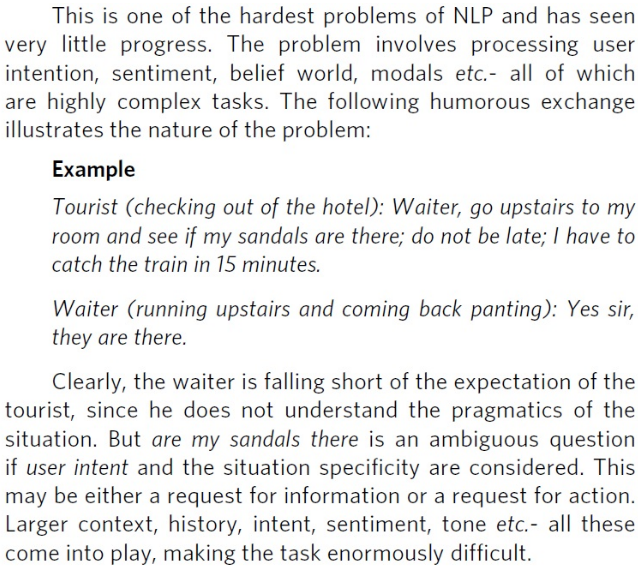
It can qualify only the phrase *smoking areas*, meaning thereby *there are areas designated as no-smoking-areas which, however, allow hookas inside.*

Semantic Analysis

After word forms and structure have been detected, sentence processing devotes itself to meaning extraction. While the meaning of meaning is debatable, there is a general agreement that at the stage of semantic processing, the sentence needs to be represented in one of the unambiguous forms like predicate calculus, semantic net, frame, conceptual dependency, conceptual structure etc.

Now, semantics extraction faces all the challenges arising out of ambiguities of semantic roles or relations. Example:
Visiting aunts can be annoying
aapko mujhe mithaai khilaanii padegii

Pragmatics Processing



This is one of the hardest problems of NLP and has seen very little progress. The problem involves processing user intention, sentiment, belief world, modals *etc.*- all of which are highly complex tasks. The following humorous exchange illustrates the nature of the problem:

Example

Tourist (checking out of the hotel): Waiter, go upstairs to my room and see if my sandals are there; do not be late; I have to catch the train in 15 minutes.

Waiter (running upstairs and coming back panting): Yes sir, they are there.

Clearly, the waiter is falling short of the expectation of the tourist, since he does not understand the pragmatics of the situation. But *are my sandals there* is an ambiguous question if *user intent* and the situation specificity are considered. This may be either a request for information or a request for action. Larger context, history, intent, sentiment, tone *etc.*- all these come into play, making the task enormously difficult.

Discourse Analysis

This is the task of processing connected sentences. All the NLP problems discussed so far, surface when we process connected text. In a speaker-listener scenario, the listener continuously produces hypotheses in his mind and updates them about the world the conversation proposes to create, as the following series of sequence of sentences illustrates:

Sentence-1: John was coming dejected from the school
(who is John: most likely a student?)
Sentence-2: He could not control the class
(who is John now? Most likely the teacher?)
Sentence-3: Teacher should not have made him responsible
(who is John now? Most likely a student again, albeit a special student- the monitor?)
Sentence-4: After all he is just a janitor
(all previous hypotheses are thrown away!)

Research in NLP

Conferences:

ACL/NAACL, EMNLP, SIGIR, AAAI/IJCAI, Coling, HLT, EACL/NAACL, AMTA/MT Summit, ICSLP/Eurospeech

Journals:

Computational Linguistics, TACL, Natural Language Engineering, Information Retrieval, Information Processing and Management, ACM Transactions on Information Systems, ACM TALIP, ACM TSLP

University centers:

Berkeley, Columbia, Stanford, CMU, JHU, Brown, UMass, MIT, UPenn, USC/ISI, Illinois, Michigan, UW, Maryland, etc.
Toronto, Edinburgh, Cambridge, Sheffield, Saarland, Trento, Prague, QCRI, NUS, and many others

Industrial research sites:

Google, MSR, Yahoo!, FB, IBM, SRI, BBN, MITRE, AT&T Labs

The ACL Anthology- http://www.aclweb.org/anthology