Introduction to Statistical NLP (Overview of Empirical NLP)

Instructor: Nick Cercone - 3050 LAS - nick@cse.yorku.ca

A Brief History of Natural Language Research

- Since its inception, one of the primary goals of AI has been the development of computational methods for natural language understanding.
- How to translate the word *pen* appropriately in "The box is in the pen" versus "The pen is in the box" (Bar-Hillel '64).
- Understanding language required not only lexical and grammatical information but semantic, pragmatic, and general world knowledge.
- In the 1970s, AI systems demonstrated interesting aspects of language understanding in restricted domains such as the blocks world (Winograd 1972) or answers to questions about a database of information on moon rocks (Woods 1977) or airplane maintenance (Waltz 1978). During the 1980s,

A Brief History of Natural Language Research

- Developing these systems remained difficult, requiring a great deal of domain-specific knowledge engineering.
- The systems were brittle and could not function adequately outside the restricted tasks for which they were designed.
- In recent years, there has been a paradigm shift in NL research.
- The focus has shifted from *rationalist* methods based on hand-coded rules derived through introspection to *empirical*, or *corpus-based*, methods which is much more data driven and is partially automated by using statistical or machine-learning methods to train systems on large amounts of real language data.

A Brief History of Natural Language Research

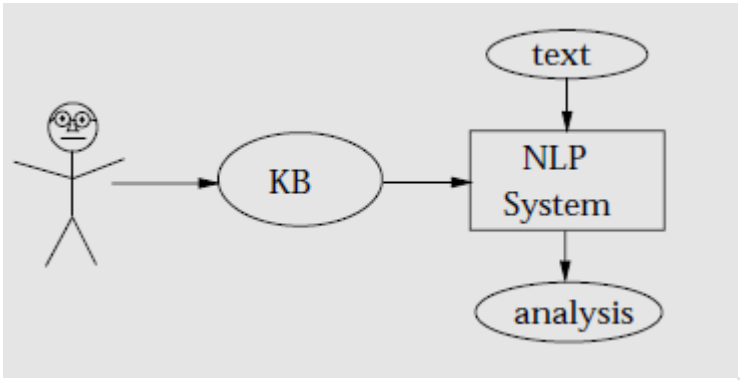


Figure 1. Traditional (Rationalist) Natural Language Processing

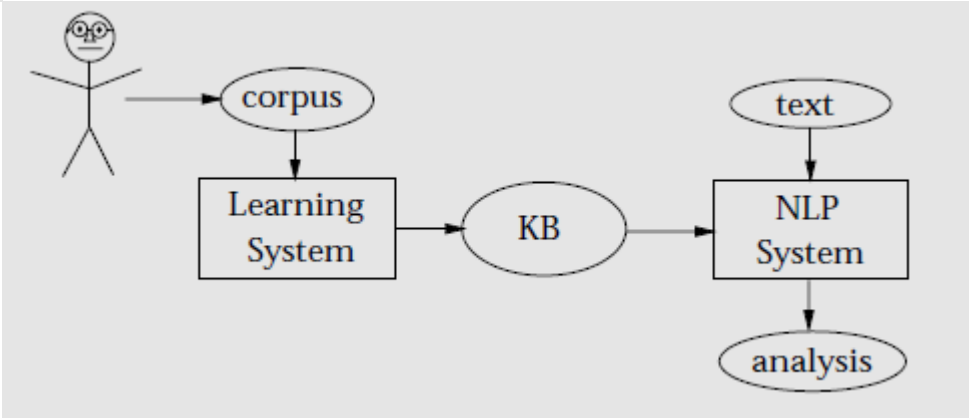


Figure 2. Empirical Natural Language Processing.

A Brief History of Natural Language Research

- Empirical and statistical analyses of NL were popular when behaviorism was thriving in psychology (Skinner 1957), and information theory was newly introduced in electrical engineering (Shannon 1951).
- In linguistics, researchers studied methods to automa-tically learning lexical & syntactic info from corpora.
- The goal was to derive an algorithmic and unbiased methodology for deducing the structure of a language using distributional information, such as the environment in which a word can appear.

A Brief History of Natural Language Research

- This framework parallels that of modern empirical NLP: Given a collection of naturally occurring sentences as input, algorithmically acquire useful linguistic information about the language.
- Chomsky's development of generative linguistics and his critique of existing empirical approaches to language quickly shifted the focus to alternative rationalist methods, with their emphasis on symbolic grammars and innate linguistic knowledge, that is, *universal grammar*.

A Brief History of Natural Language Research

- In the early 1980s, there was some work in automatic induction of lexical and syntactic information from text, based largely on two widely available annotated corpora: the Brown corpus (Marcus, Santorini, and Marcinkiewicz 1993a) and the Lancaster-Oslo–Bergen corpus (Garside, Leech, and Sampson 1987).
- Empiricism spread rapidly throughout the NLP community to a large extent the result of seminal research in speech recognition (Waibel and Lee 1990).

A Brief History of Natural Language Research

- Much of the initial success came from using the noisy-channel model (figure 3), an approach that had proven highly successful in speech recognition.
- Basically, the model assumes that language is generated and then passed through a noisy channel, and the resulting noisy data are received. The goal then is to recover the original data from the noisy data.
- It is fascinating that this simple model has been used successfully in areas of language processing as disparate as spelling correction and machine translation

A Brief History of Natural Language Research

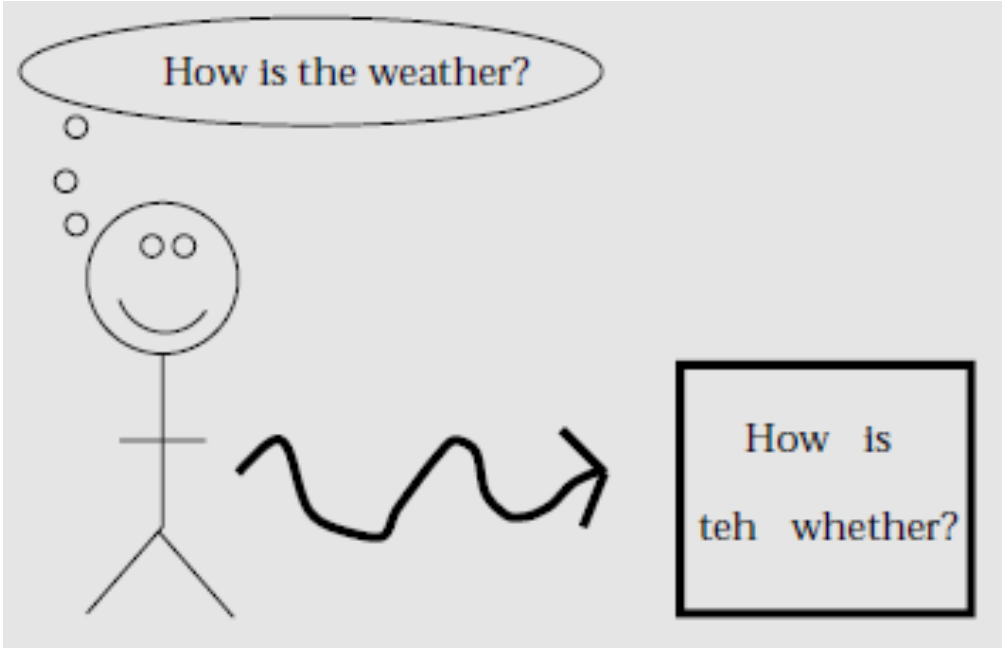


Figure 3. The Noisy-Channel Model.

A Brief History of Natural Language Research

- One of the first successes of corpus-based learning was in *part-of-speech (POS) tagging*.
- A number of techniques can now perform this task at an accuracy close to human performance (>95%); it is a useful preprocessing step in other tasks such as parsing, speech synthesis, and information retrieval.
- Another early influential result was statistical approaches to machine translation trained and tested on bilingual proceedings of the Canadian parliament (Brown et al.' 90).

A Brief History of Natural Language Research

- Another early influential result was statistical approaches to machine translation trained and tested on bilingual proceedings of the Canadian parliament (Brown et al.' 90).
- With the development of *tree banks*, large databases of sentences annotated with syntactic parse trees came an increasing body of research on empirical parsing methods, for example, *probabilistic context-free grammars* (PCFGs) (Charniak 1996; Collins 1996; Pereira and Shabes 1992).

A Brief History of Natural Language Research

- With the development of *tree banks*, large databases of sentences annotated with syntactic parse trees came an increasing body of research on empirical parsing methods, for example, *probabilistic context-free grammars* (PCFGs) (Charniak 1996; Collins 1996).
- Research on empirical methods thrives in other areas as well, such as *word sense disambiguation* (Yarowsky 1992), *prepositional phrase attachment* (Hindle et al 1993), *semantic analysis* (Zelle and Mooney 1996), *anaphora resolution* (Cardie 1992), and *discourse segmentation.*

A Brief History of Natural Language Research

Reasons for the Resurgence of Empiricism

- The recent dramatic increase in empirical research has been attributed to various causes: Empirical methods offer potential solutions to several related, long-standing problems in NLP:
  - (1) *acquisition*, automatically identifying and coding all the necessary knowledge;
  - (2) *coverage*, accounting for all the phenomena in a given domain or application;
  - (3) *robustness*, accommodating real data that contain noise and aspects not accounted for by the underlying model; and
  - (4) *extensibility*, easily extending or porting a system to a new set of data or a new task or domain.

A Brief History of Natural Language Research

Three recent developments have spurred the resurgence in empiricism:

(1) *computing resources*, the availability of relatively inexpensive workstations with sufficient processing and memory resources to analyze large amounts of data;

(2) *data resources*, the development and availability of large corpora of linguistic and lexical data for training and testing systems; and

(3) *emphasis on applications and evaluation*, industrial and government focus on the development of practical systems that are experimentally evaluated on real data.

Categories of Empirical Methods

- Most of the recent work in empirical NLP involves statistical training techniques for probabilistic models such as HMMs and PCFGs (Charniak 1993).
- These methods attach probabilities to the transitions of a finite-state machine or the productions of a formal grammar and estimate these probabilistic parameters based on training data.
- If the training set is preannotated with the structure being learned, learning consists simply of counting various observed events in the training data

Categories of Empirical Methods

- If the corpus is not annotated, an estimation-maximization strategy could be used, for example, the forward-backward algorithm for Markov models and the inside-outside algorithm for PCFGs.
- Novel test examples are then analyzed by determining the most-probable path through the learned automaton or derivation from the learned grammar that generates the given string.
- Other empirical methods gather and use other statistics such as the frequency of each *n*-word sequence (*n-gram*) appearing in the language.

Categories of Language Tasks

- Understanding NL is a complex task and involves many levels of processing and a variety of subtasks. We can divide NLP as follows: speech recognition and spoken-language analysis, syntactic analysis, semantic analysis, discourse analysis and information extraction, and machine translation.
- Figure 4 illustrates the first four as components of an overall language-understanding system

Categories of Language Tasks

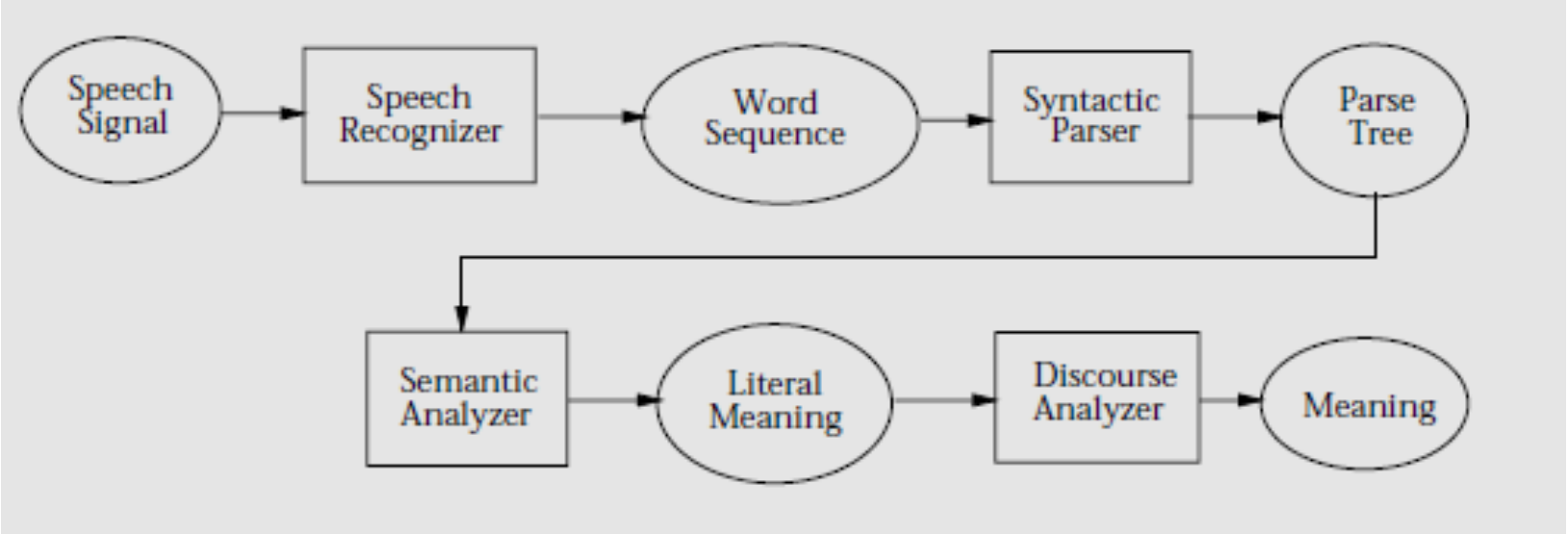


Figure 4. Components of a Natural Language–Processing System.

Categories of Language Tasks

- A different dimension along which empirical methods vary concerns the type of training data required.
- Many systems use *supervised* methods and require *annotated* text in which human supervisors have labeled words with parts of speech or semantic senses or have annotated sentences with syntactic parses or semantic representations.
- Other systems employ *unsupervised* methods and use raw, unannotated text.

Categories of Language Tasks

- Unsupervised learning is generally more difficult and requires some method for acquiring feedback indirectly, such as assuming that all sentences encountered in texts are positive examples of grammatical sentences in the language.
- Finally, it is important to note that traditional rationalist approaches and empirical methods are not incompatible or incommensurate

Categories of Language Tasks

- *Speech recognition* maps a continuous speech signal into a sequence of recognized words. The problem is difficult because of the wide variation in the exact pronunciation of words spoken by different speakers in different contexts. Other problems include homonyms (for example, *pair, pear, pare*), other forms of acoustic ambiguity (e.g., *youth in Asia* and *euthanasia*), and the slurring of words (for example, *didja*) that happens in continuous speech.

Categories of Language Tasks

- *Syntactic analysis* determines the sentence grammatical structure, i.e., how the words are grouped as constituents such as noun phrases and verb phrases.
- A subtask of syntactic analysis is assigning a parts of speech, such as that *saw* acts as a noun in "John bought a saw" and as a verb in "John saw the movie."
- Another ambiguity problem is *attachment*, to determine in "John saw the man on the hill" whether *on the hill* modifies the man or the seeing event.

Categories of Language Tasks

- Great improvements in statistical NLP systems can be attributed to two things: (1) publicly available tree banks and (2) grammar lexicalization.
- The Penn tree bank has released a corpus of 50,000 sentences that have carefully been annotated for syntactic structure by hand.
- PCFGs have been around for a long time. However, they have one big weakness as a tool for modeling language. Given a PCFG, the probability of a particular parse for a sentence is the product of the probability of all rules used to generate the parse.

Categories of Language Tasks

- Given the two sentences along with their parses shown in figure 5, assuming *flew* and *knew* are equally likely as verbs, the PCFG would assign equal probability to these two sentences because they differ in how they were generated only by one rule (*VP flew* vs. *VP knew*).

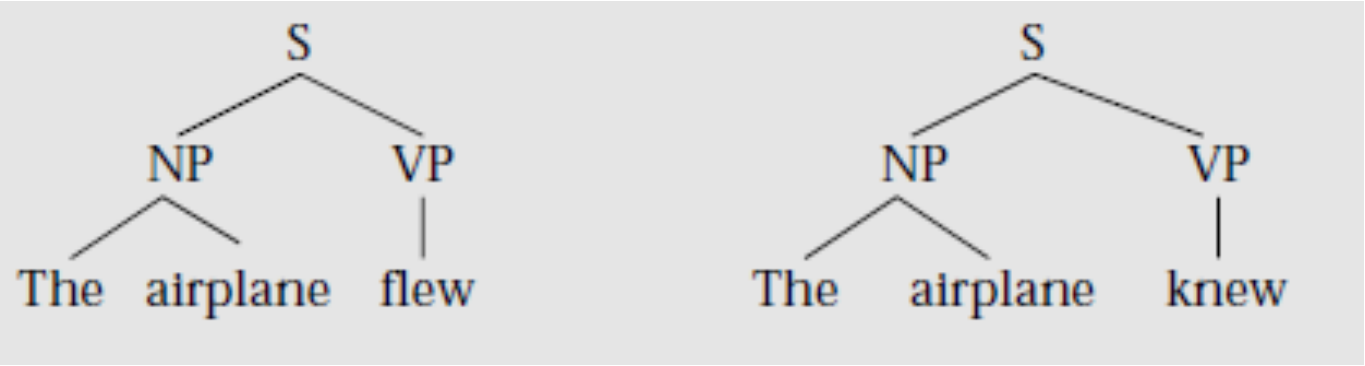


Figure 5. Unlexicalized Parses.

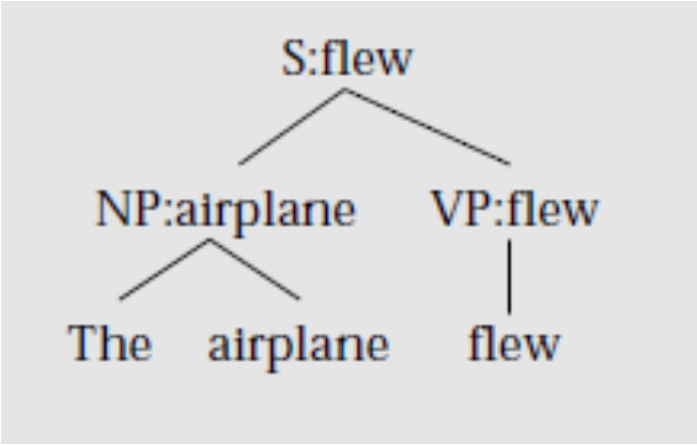
Categories of Language Tasks

- The insight that has led to vast improvements in parsing is that by lexicalizing the grammar, more meaningful statistics can be obtained.
- Eg., a lexicalized parse is shown in figure 6. All nodes in the tree contain information about the words in the sentence, not just the nodes immediately above words.
- In the nonlexicalized parses of figure 5, an *S* node is expanded into an *NP* node and a *VP* node. In a lexicalized parse such as that in figure 6, an *S* node expands into an *airplane NP* node and a *flew VP* node.

Categories of Language Tasks

- Because the probability of *S® NP*: sub*airplane VP*: *flew* will be greater than that of *S® NP*: *airplane VP*: *knew* (airplanes fly, they don't know), the grammar will now give higher probability to the more likely sentence.

Figure 6. A Lexicalized Parse.



Categories of Language Tasks

- *Semantic analysis* maps a sentence to some sort of meaning representation, for example, a logical expression.
- In corpus-based approaches to NLP semantic interpretation in Natural two important subtasks include:

  (1) word-sense disambiguation and

  (2) semantic parsing.

Categories of Language Tasks

- *Word-sense disambiguation* decides which possible meanings for a word is correct in a particular context.
- A classic example is determining whether *pen* refers to a writing instrument or an enclosure in a particular sentence such as "John wrote the letter with a pen" or "John saw the pig in the pen."
- Part of semantic parsing involves producing a *case-role* analysis, in which the semantic roles of the entities referred to in a sentence, such as *agent* and *instrument*, are identified.

Categories of Language Tasks

- *Discourse Analysis and Information Extraction*
- *Discourse analysis* determines how larger intersentential context influences the interpretation of a sentence.
- Tasks in discourse analysis include: (1) coreference resolution and (2) information extraction.
- *Coreference resolution* must determine what phrases in a document refer to the same thing, e.g., pronoun resolution. For example, in "John wanted a copy of Netscape to run on his PC on the desk in his den; fortunately, his ISP included it in their startup package," a pronoun-resolution algorithm would have to determine that *it* refers to a copy of Netscape rather than PC, desk, or den.

Categories of Language Tasks

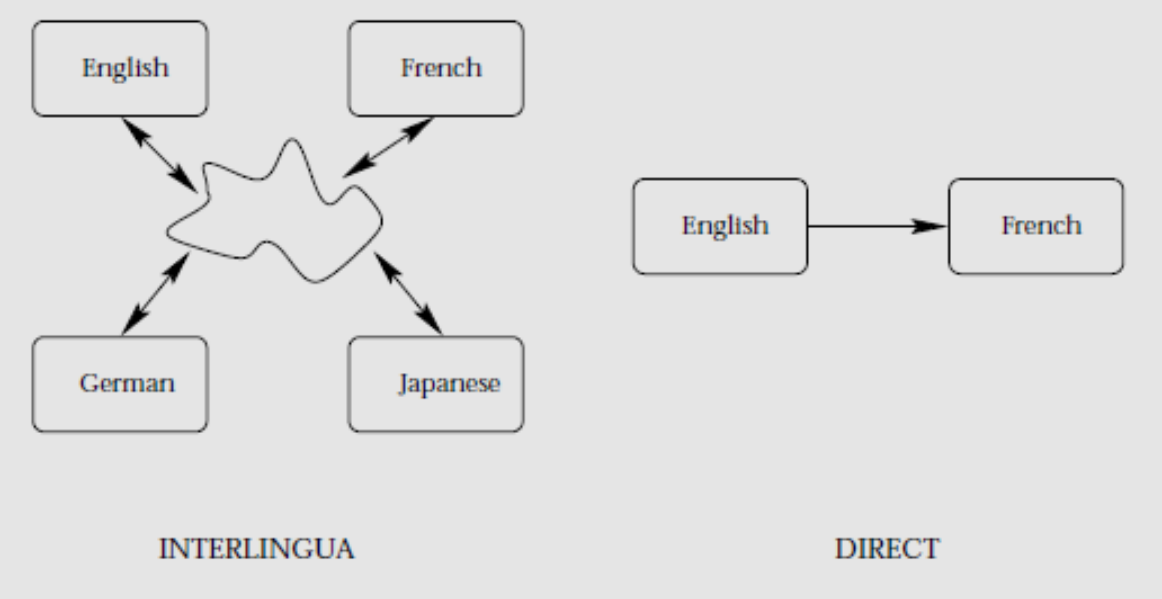
- *Information extraction* is the task of locating specific pieces of data from a NL document.
- Consider analyzing a message from misc.jobs.offered to extract the employer's name, location, type of job, years of experience required, and so on. The information extracted from a collection of messages could then be stored in a database with fields for each of these slots. Text is first linguistically annotated, and then extraction rules are used to map from annotated text to slot filling.

Categories of Language Tasks

- Until recently, these rules were written by hand. By using machine-learning techniques, extraction rules can be learned automatically and achieve performance close to the best manually constructed systems.

Categories of Language Tasks

- *Machine translation* involves translating text from one natural language to another, such as translating English to Japanese, or vice versa. Two approaches:



Other Concluding Remarks

MAKING AN EFFORT

Our so-called limitations, I believe,

apply to faculties we don't apply.

We don't discover what we can't achieve

until we make an effort not to try.