Language Technology

Chapter 1: An Overview of Language Processing

Pierre Nugues

Pierre.Nugues@cs.lth.se

September 2, 2024

Applications of Language Processing

- Spelling and grammatical checkers: *MS Word*, e-mail programs, etc.
- Text indexing and information retrieval on the Internet: *Google*, *Microsoft Bing*, *Yahoo*, or software like *Apache Lucene*
- Translation: *Google Translate*, *DeepL*, *Bing translator*, etc.
- Spoken interaction: Apple Siri, Google Assistant, Amazon Echo
- Speech dictation of letters or reports: *Windows 10*, *macOS*

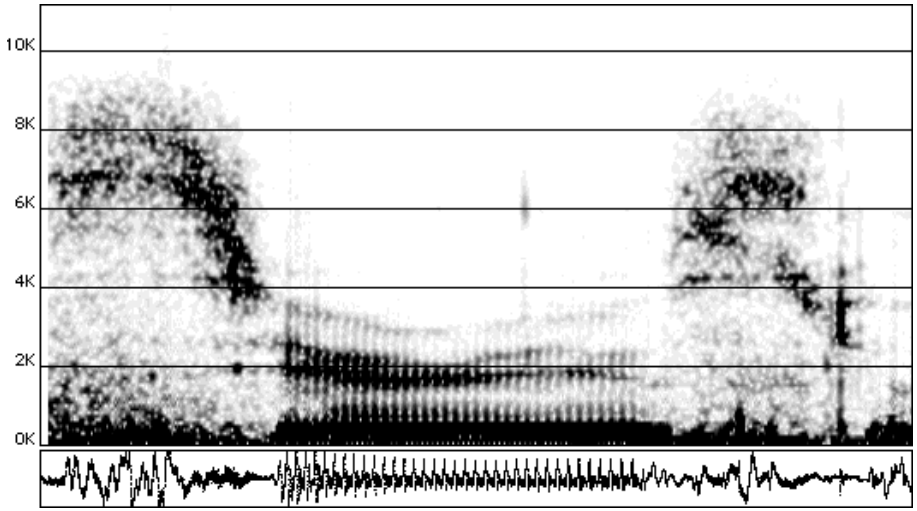
Applications of Language Processing (ctn'd)

- Direct translation from spoken English to spoken Swedish in a restricted domain: *SRI* and *SICS*
- Voice control of domestic devices such as tape recorders: *Philips* or disc changers: *MS Persona*
- Conversational agents able to dialogue and to plan: *TRAINS*
- Spoken navigation in virtual worlds: *Ulysse*, *Higgins*
- Generation of 3D scenes from text: *Carsim*
- Question answering: *IBM Watson* and *Jeopardy!*

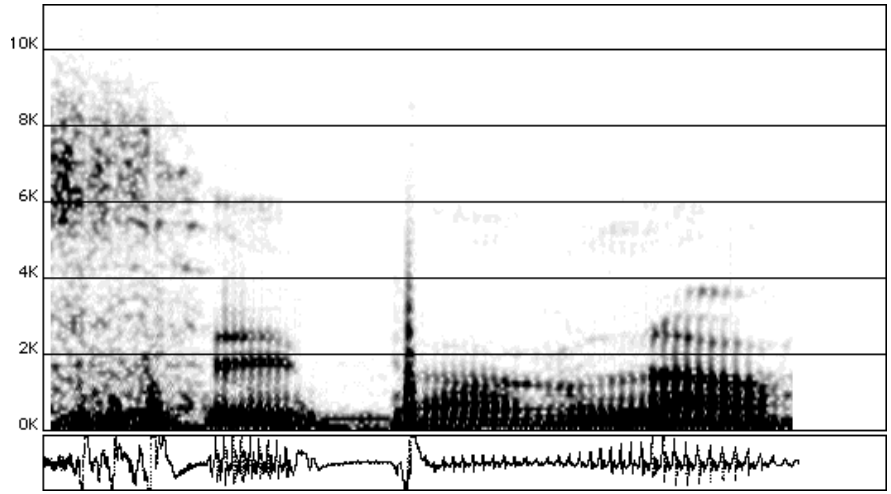
Linguistics Objects and Vocabulary

- Sounds
- Phonemes
- Words and morphology
- Syntax and functions
- Semantics
- Dialogue

Sounds and Phonemes



Serious



C'est par là 'It is that way'

Lexicon and Parts of Speech

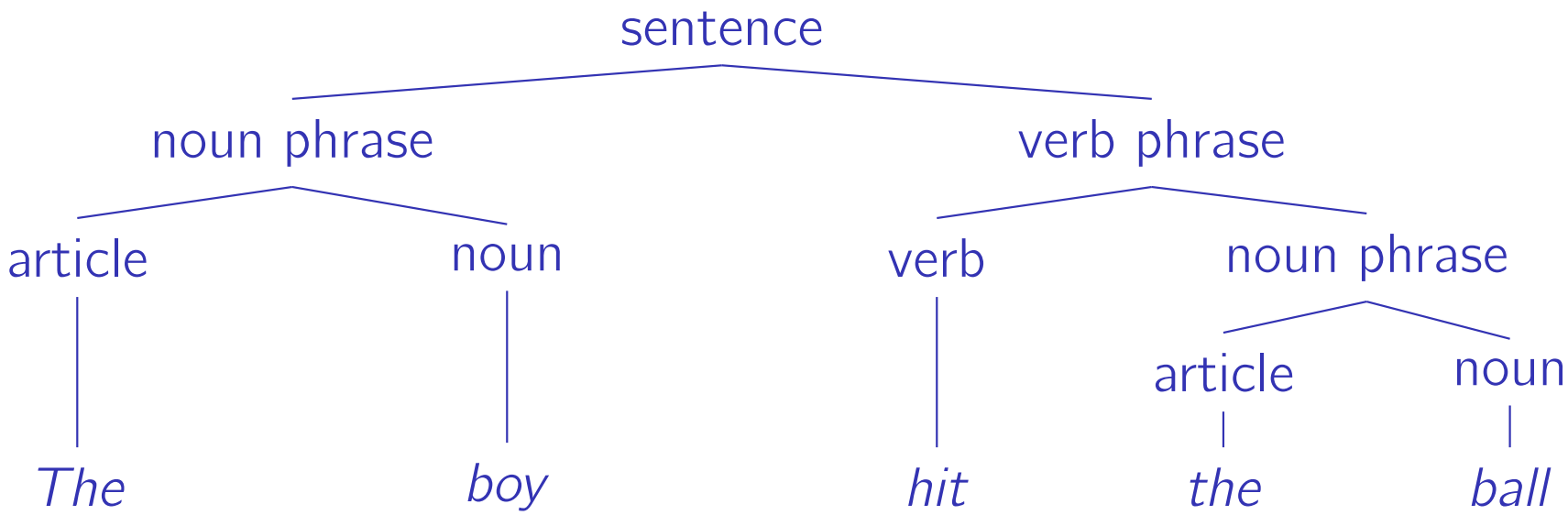
The big cat ate the gray mouse

The/article *big*/adjective *cat*/noun *ate*/verb *the*/article *gray*/adjective *mouse*/noun
Le/article *gros*/adjectif *chat*/nom *mange*/verbe *la*/article *souris*/nom *grise*/adjectif
Die/Artikel *große*/Adjektiv *Katze*/Substantiv *ißt*/Verb *die*/Artikel *graue*/Adjektiv *Maus*/Substantiv

Morphology

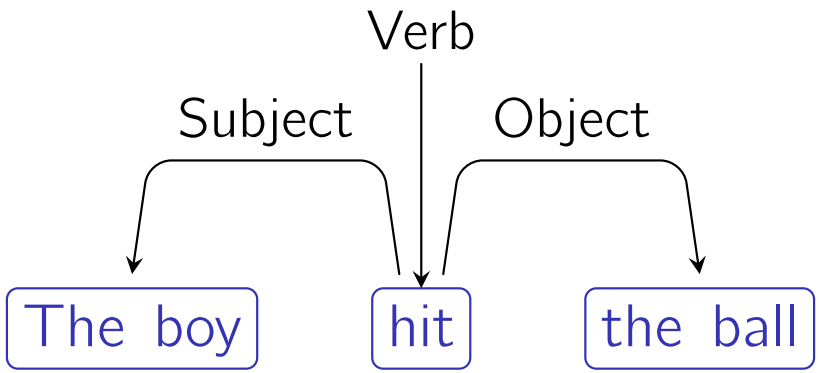
Word	Morphemes	Root form and grammatical features
worked	*work+ed*	*work* + verb + preterit
travaillé	*travaill + é*	*travailler* + verb + past participle
gearbeitet	*ge + arbeit+ et*	*arbeiten* + verb + past participle

Syntactic Tree



Syntax: A Classical View

A graph of dependencies and functions



Semantics

As opposed to syntax:

1. Colorless green ideas sleep furiously.
2. *Furiously sleep ideas green colorless.

Determining the logical form:

Sentence	Logical representation
Frank is writing notes	`writing(Frank, notes).`
François écrit des notes	`écrit(François, notes).`
Franz schreibt Notizen	`schreibt(Franz, Notizen).`

Lexical Semantics

Word senses:

1. **note** (*noun*) short piece of writing;
2. **note** (*noun*) a single sound at a particular level;
3. **note** (*noun*) a piece of paper money;
4. **note** (*verb*) to take notice of;
5. **note** (*noun*) of note: of importance.

Reference

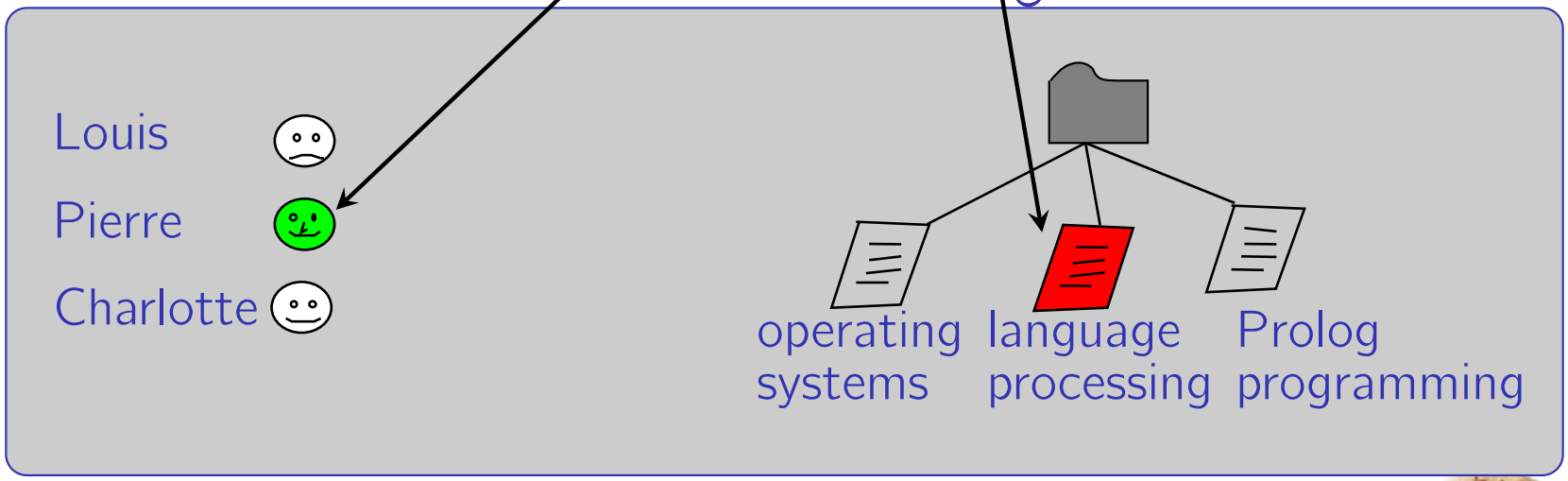
1. Sentence
Pierre wrote notes

2. Logical representation
`wrote(pierre, notes)`

refers to

refers to

3. Real world



Ambiguity

Many analyses are ambiguous. It makes language processing difficult.
Ambiguity occurs in any layer: speech recognition, part-of-speech tagging, parsing, etc.
Example of an ambiguous phonetic transcription:
The boys eat the sandwiches
That may correspond to:
The boy seat the sandwiches; *the boy seat this and which is*; *the buoys eat the sand which is*

Models and Tools

- Linguistics has produced an impressive set of theories and models;
- Inadequate theories in the beginning and lack of data: corpus, dictionaries, or reference (annotated) data;
- Models and tools have matured. Large datasets are now available;
- Tools involve notably finite-state automata, regular expressions, logic, statistics, and machine learning;
- In general, language processing requires significant processing power and new GPU architectures;
- This overall resulted in massive improvements in most areas of NLP.

An Application Example: IBM Watson

- IBM Watson: A system that can answer questions better than any human
- Video: `https://www.youtube.com/watch?v=WFR3lOm_xhE`

- IBM Watson builds on the extraction of knowledge from masses of texts: Wikipedia, archive of the New York Times, etc. and multiple language processing modules

Two Viewpoints

Given a task to solve, natural language processing has seen two approaches:

1. Try to understand the human processing of language and reproduce it;
   - For instance, all the syntactic and semantic phenomena involved in translating a sentence;
2. Try to design a device that replicates the human production/assessment on existing examples.
   - For instance, to produce translations identical to those of human beings, regardless of the process;

The second viewpoint is now hugely dominant.
Validated by benchmarks largely agnostic to means

Datasets

Digitization is a key driver to progress of NLP
Huge datasets are available to many researchers and companies
We can divide them in two categories:

- Annotated text: For instance IMDB movie reviews, positive negative, for categorization. (`https://ai.stanford.edu/~amaas/data/sentiment/`)
- Corpora of raw text. Wikipedia, book archives, Common Crawl (`https://commoncrawl.org/`), Colossal cleaned crawled corpus (156 billion tokens for the English part) (`https://c4-search.apps.allenai.org/`)

Training

Given a mathematical parametrizable model, datasets form the input to adjust the parameters:

1. Annotated corpora: Supervised training for those who have heard of it
2. Raw corpora: We can still train a model with them, for instance to guess missing words. Corpora are repositories of knowledge and language rules. If large enough, they will capture eventually the semantics of words.

Although the second model cannot serve as a final application, we can use it to transform the input in a way that it is more relevant for a subsequent task

This process is called pretraining–fine-tuning

Stanford Question Answering Dataset (SQuAD)

- Consists of 100,000 questions and paragraphs from wikipedia containing the answers
- The answer is a segment in the text (factoid QA)
- Complemented by SQuAD 2.0 with unanswerable questions
- SQuAD started an intense competition. See the impressive leaderboard `https://rajpurkar.github.io/SQuAD-explorer/`

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

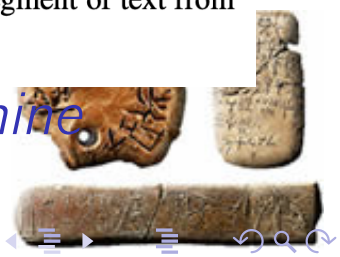
What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

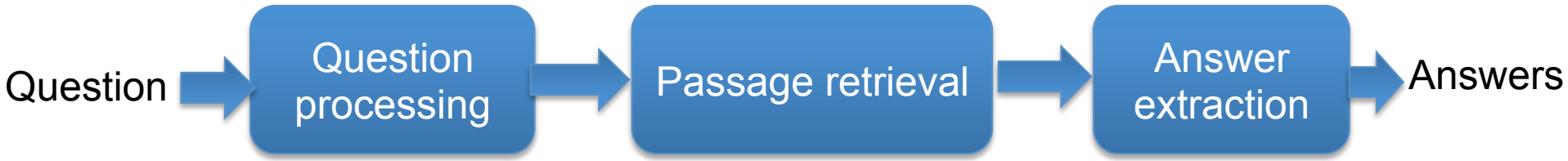
Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

Form Rajpurkar et al., *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, 2016



Architecture Evolution

Questions answering systems started as pipelines of linguistic modules and ended with large language models
Simplified architecture of IBM Watson:

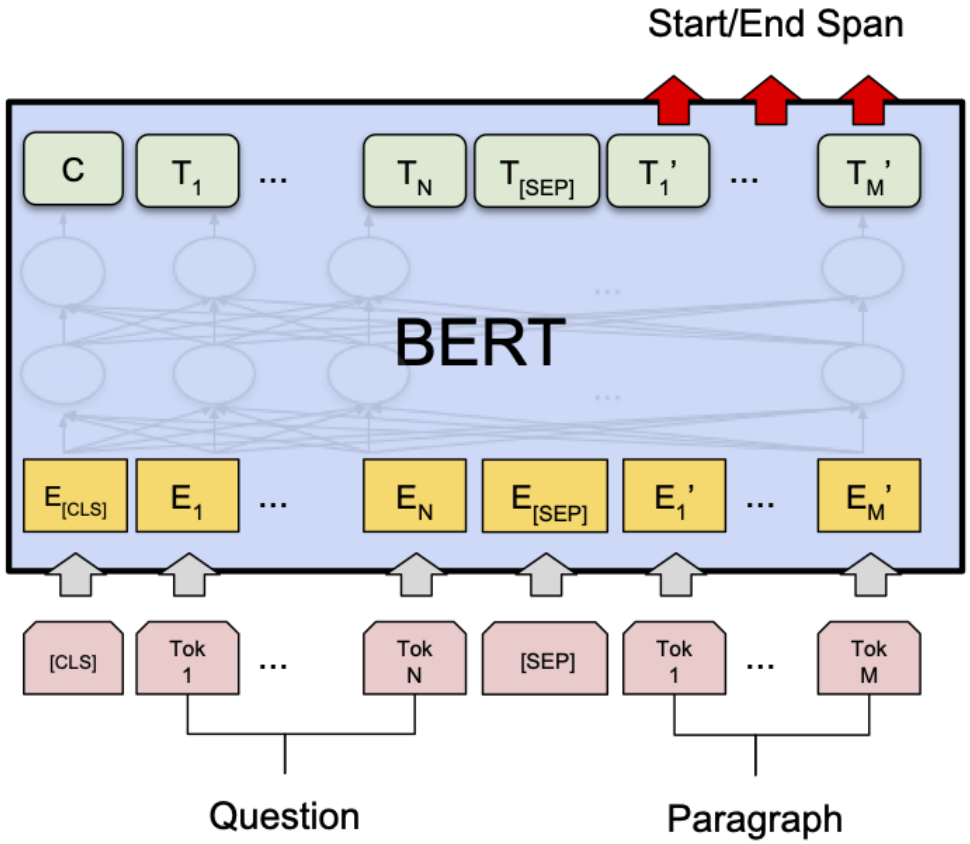


Question parsing and classification: *Syntactic parsing, entity recognition, answer classification*

Document retrieval. Extraction and ranking of passages: *Indexing, vector space model.*

Extraction and ranking of answers: *Answer parsing, entity recognition*

Question Answering with a Transformer



(c) Question Answering Tasks: SQuAD v1.1

from Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019

GLUE Evaluation

To evaluate all the claims, research rely on benchmarks as with SQuAD General Language Understanding Evaluation (GLUE) is another popular benchmark (`https://gluebenchmark.com/`) that three kinds of tasks:

1. Sentence or text classification
2. Classification of a relation in a pair of sentences: similarity, entailment, negation, etc.
3. Question answering

The machine translation tasks are similar:
`https://www.statmt.org/wmt22/results.html`
On large language models: `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`

A Word on Computing Resources

Progress in NLP is not free: Large models now have billions of parameters: 70 for Meta's LLama models.

From Strubell et al (2019)

Consumption	**CO_2e (lbs)**
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

From Touvron et al. (2023) on training LLama

Finally, we estimate that we used 2048 A100-80GB for a period of approximately 5 months to develop our models. This means that developing these models would have cost around 2,638 MWh under our assumptions, and a total emission of 1,015 tCO_2eq. We hope that releasing these models will help to reduce future carbon emission since the training is already done, and some of the models are relatively small and can be run on a single GPU.