

Natural Language Processing: CS 410/510

Introduction

Instructor: Thien Huu Nguyen

Based on slides from: Ralph Grishman



So what are natural languages?

- A primary and natural mode of human communication and reasoning
- representation for most recorded human knowledge
- a very rich and flexible representation (comparing to most formal representations)



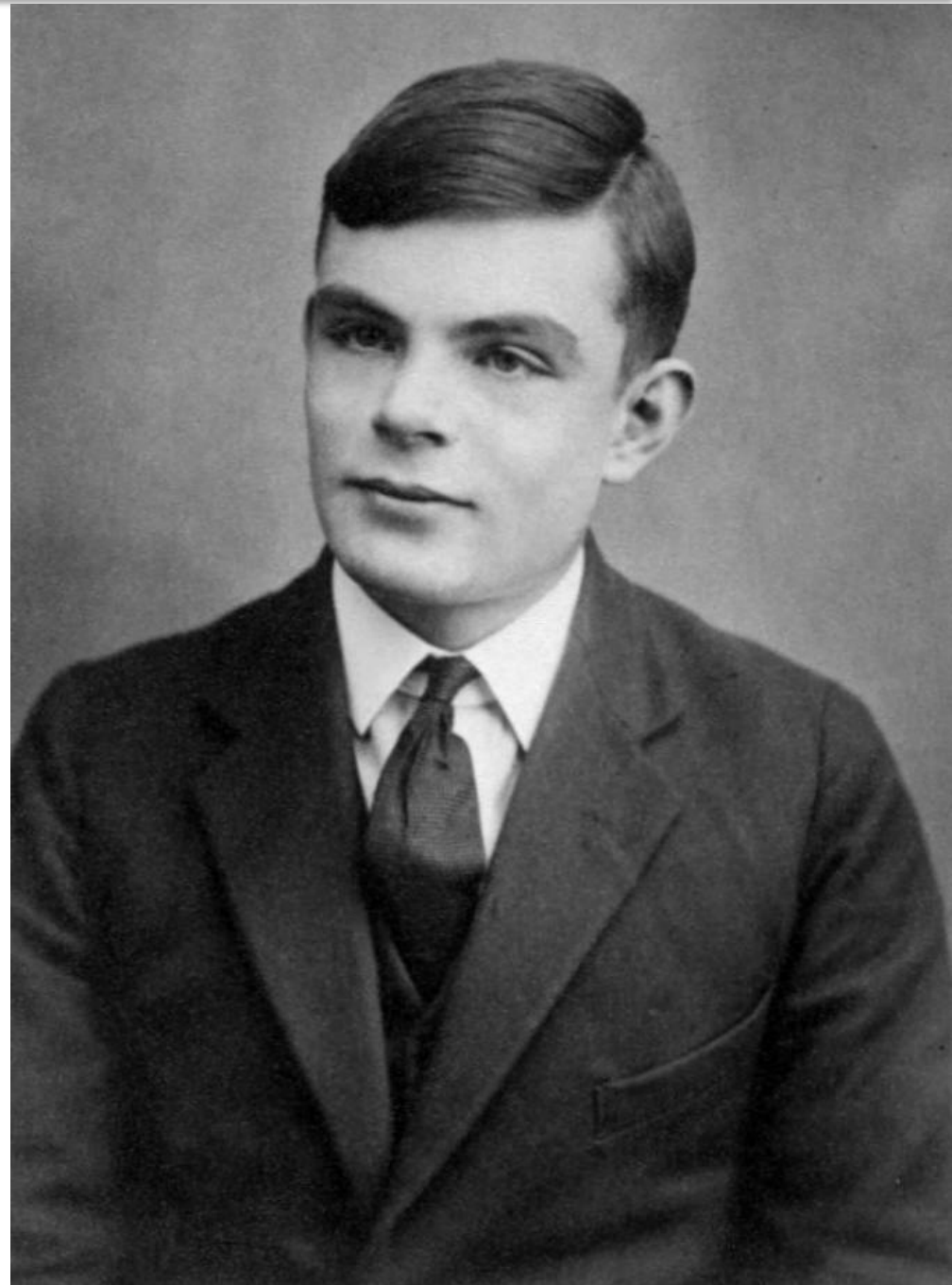
Natural Language Processing (NLP)

- Processing natural languages with computers, enabling computers to understand, generate and analyze natural languages.
- Technically, designing methods and algorithms that take as input and/or produce as output unstructured, natural language data (“Hello World”).
 - More abstract: understanding the underlying principles of human languages to allow computers to mimic human capacity on languages.
 - More abstract: understanding the cognitive processes within human minds based on the human-generated languages
 - More abstract: the computational science for linguistics



The Turing Test

Distinguishing human
vs. computer only
through written
language



Why study NLP?

↑ [-] **CyberByte** 27 points 2 years ago

↓ If you got a billion dollars to spend on a huge research project that you get to lead, what would you like to do?



Michael Jordan, leading machine learning researcher,
UC Berkeley

I'd use the billion dollars to build a NASA-size program focusing on natural language processing (NLP), in all of its glory (semantics, pragmatics, etc).

Intellectually I think that NLP is fascinating, allowing us to focus on highly-structured inference problems, on issues that go to the core of "what is thought" but remain eminently practical, and on a technology that surely would make the world a better place.



Where are we now?

ChatGPT 3.5

- You: What is natural language processing?
- ChatGPT: Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) and linguistics concerned with the interaction between computers and humans through natural language. It focuses on the ability of computers to understand, interpret, and generate human language in a way that is both meaningful and useful.
- NLP involves a variety of tasks and techniques, including:
1. **Tokenization:** Breaking text into smaller units such as words or sentences.
 2. **Part-of-speech tagging:** Assigning grammatical tags (noun, verb, adjective, etc.) to words in a sentence.
 3. **Parsing:** Analyzing the grammatical structure of sentences to understand their syntax.
 4. **Named entity recognition (NER):** Identifying and categorizing entities such as names of people, organizations, locations, etc., within text.
 5. **Sentiment analysis:** Determining the sentiment or emotional tone conveyed by text (positive, negative, neutral).
 6. **Machine translation:** Translating text from one language to another.
 7. **Text generation:** Generating human-like text based on input or context.
 8. **Question answering:** Finding answers to questions posed in natural language.
 9. **Summarization:** Generating a concise summary of a longer text.

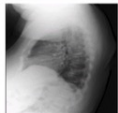


You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. The report needs to contain both the impression and findings sections. You should only generate reports and no other unnecessary information.



Ground Truth:
Impression: No acute intrathoracic process. Findings: The lungs are well expanded and clear. The **cardiomediastinal silhouette**, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. Sternal wires are intact.

GPT-4V:
Impression: Normal chest radiograph with **no acute cardiopulmonary abnormalities**. Findings: Lungs: Clear bilaterally. No focal opacity, consolidation, or mass identified. **Cardiomediastinal silhouette:** Normal size and contour. No mediastinal widening. Bones: Intact without any appreciable fractures or bony lesions. Diaphragm: Intact with no evidence of diaphragmatic hernia. Soft tissues: Unremarkable.



Ground Truth:
Impression: No acute intrathoracic process. Findings: **The lungs are hyperinflated**, accounting for vascular crowding but no evidence of focal opacities. **Cardiomediastinal** and hilar contours are unremarkable. The cardiac size is top normal. There is no pleural effusion or pneumothorax. A right subclavian stent is noted and unchanged from prior examination.

GPT-4V:
Impression: Evidence of increased radiodensity in the mid and lower lung fields, which could suggest possible consolidation or infiltrates. Further clinical correlation is recommended. Findings: **Lungs: Increased radiodensity** in the mid and lower lung fields indicating possible consolidation or infiltrative pathology. **Cardiomediastinal silhouette:** Not entirely visualized, but no obvious abnormalities seen. Bones: Thoracic bony structures are intact with no signs of acute fractures. Diaphragm: Not explicitly mentioned in the report, but we could assume it's intact given no specified abnormalities. Soft tissues: The visible upper abdominal structures are unremarkable.



Ground Truth:
Impression: Increased opacity projecting over the right hemi thorax likely due to increased right pleural effusion with overlying atelectasis, **underlying infectious process not excluded**. Findings: There are relatively **low lung volumes**. There is increased opacity projecting over the right hemi thorax likely due to increased right pleural effusion with overlying atelectasis, **underlying infectious process not excluded**. Possible trace left pleural effusion. The cardiac silhouette is top-normal to mildly enlarged. Mediastinal contours are unremarkable. **No pneumothorax is seen.**

GPT-4V:
Impression: Bilateral interstitial and alveolar opacities, predominantly in the upper and mid lung zones, **suggestive of a possible infectious or inflammatory process**. No overt pleural effusion or pneumothorax. Findings: **The lungs are hyperinflated** with no focal consolidation. Bilateral interstitial and alveolar opacities are observed, more pronounced in the upper and mid lung zones. The cardiac silhouette is within normal limits for size and contour. No visible mediastinal or hilar lymphadenopathy. **No evidence of pneumothorax** or significant pleural effusion. The osseous structures of the chest are intact without visible fractures or lytic lesions.

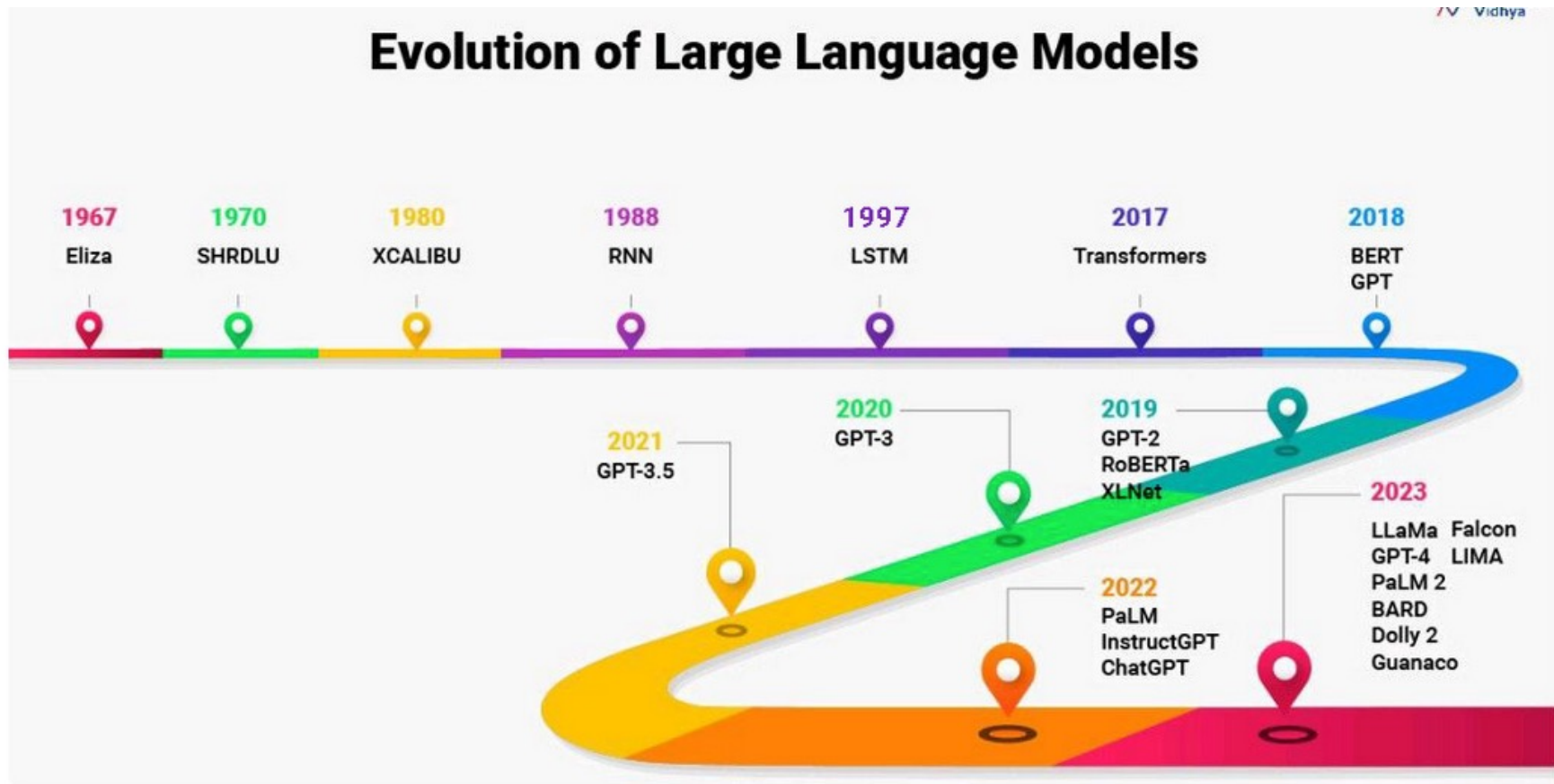


Credit: <https://bito.ai/blog/gemini-1-5-pro-vs-gpt-4-turbo-benchmarks/>
<https://medium.com/@nickyverd/openai-sora-why-you-should-be-excited-and-scared-7bbf0104d732>

Li et al., 2023: A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging



Where are we now?



NLP provides foundations for the most recent breakthroughs in Generative AI.

Credit: <https://www.analyticsvidhya.com/blog/2023/07/beginners-guide-to-build-large-language-models-from-scratch/>



NLP Applications: Machine Translation

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Warren Weaver (in his letter to Prof. Norbert Wiener at MIT in 1947)



NLP Applications: Machine Translation

- Long history of development (since early 1950's)
- Still not perfect for polished translations, but effective for producing rough drafts for post-editing, or for getting the gist of a text
- Quality depends on similarity of language structures (Chinese to English much harder than French to English)



Machine translation (MT)

ChatGPT 3.5 ▾



You

Translate this sentence to Vietnamese: "One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know."



ChatGPT

"Một buổi sáng, tôi bắn một con voi trong bộ đồ ngủ của mình. Làm sao con voi ấy lại vào trong bộ đồ ngủ của tôi, tôi sẽ không bao giờ biết được."



Text

Images

Documents

Websites

Detect language English Vietnamese Spanish ▾

↔ Vietnamese English Arabic ▾

One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know. ✕

Một buổi sáng, tôi bắn một con voi trong bộ đồ ngủ. Làm thế nào anh ấy mặc được bộ đồ ngủ của tôi, tôi sẽ không bao giờ biết. ☆



Machine translation (MT)



NLP Applications: Question Answering

- Originally passage retrieval systems, gradually enriched with NLP
- Example: MIT Start system : <http://start.csail.mit.edu/>
- Major web search tools originally operated as web page retrieval (document retrieval) using word-based strategies but now seek to detect questions they can answer directly (e.g., google.com)
- IBM Watson system to play Jeopardy demonstrated potential of QAS when coupled to a rich variety of knowledge sources; this technology now being applied to medical domain:
https://www.youtube.com/watch?v=WFR3lOm_xhE



NLP Applications: Interactive Systems

- One of the first interactive applications was data base query, but that had limited appeal with written input: people don't like to type a lot; GUIs have been more effective
- Chatbots provide the impression of intelligent conversation (try Eliza: <https://www.cyberpsych.org/eliza/#.WIQlu2PDX8s>
or: <https://www.tolearnenglish.com/free/celebs/alice.php>
or ChatGPT: <https://chat.openai.com/>)
- Conversational agents support simple conversations (using text or speech) for order taking, information
- Smartphones with speech recognition (e.g., SIRI) have greatly increased opportunities for speech input



NLP Applications: Information Extraction

- Conversion of unstructured (or semi-structured) data to structured form
- Resumes: <http://monster.com>
- Name Tagger: GATE: <http://services.gate.ac.uk/annie/>
- More general entity/relation/event taggers
- For search and general news monitoring
 - [Google News](#)
 - [European Media Monitor](#)
- For search and general news monitoring
 - Infectious disease reports (e.g., [the NYU Proteus-Bio system](#))
 - Electronic health records
 - Scientific literature (e.g., [Paleo Deep-Dive](#))



NLP Applications: Text Mining

- Discovering patterns from large text collections
- Direct (“first generation”) approach uses “bag of words” representation of documents
- Second generation: information extraction + data mining = text mining
- Take advantage of social media (e.g., [Dataminr](#))
- Applications:
 - Product monitoring (using sentiment analysis to collect detailed feedback from customers)
 - Situational awareness (rapid detection of emergency situations)
 - Finance (rapid response to financial news)
 - Scientific research (mine large collections of research papers to identify trends and correlations like treatments and adverse reactions)



Challenges in NLP

- Language is discrete, making it more challenging to solve the optimization-related problems.
- Language is compositional (the meaning of a sentence involves understanding the individual words and how to combine/compose the individual meanings to form the overall meaning).
- Language is flexible (there are multiple ways to talk about the same things).
 - I hate spicy food
 - I don't like spicy food
 - If I can choose, I won't take spicy food
- Language is ambiguous (context is often necessary to accurately understand the meanings of a word/sentence).



The ambiguity of language

*“One morning I shot
an elephant in my
pajamas”*

So, I or the elephant is
in my pajamas?



Animal Crackers

The ambiguity of language

*“One morning I shot
an elephant in my
pajamas. How he got
into my pajamas I’ll
never know.”*



Animal Crackers

Relation to other fields

- Linguistics
 - The goal is to describe language
 - Provide simple models which can *predict* language behavior
 - Understand what is *universal* about language
 - Through these formal models, understand how language can be *acquired*
 - Formal linguistic models have been helpful for NLP, but their goals are different from NLP's:
 - A simple counterexample can invalidate a model as a linguistic theory, but would not significantly lessen the value of NLP models
 - NLP must address all phenomena that arise in an application, while linguistics may focus on selected phenomena that are of interest to the language faculty



Relation to other fields

- (symbolic) Artificial Intelligence
 - Classical “symbolic” AI is mostly concerned with generic problem solving strategies and suitable knowledge representation
 - The inherent link between AI and NLP: some NLP problems need the sort of deep reasoning provided by these AI models.
 - But NLP has found increasing success through avoiding such deep reasoning and turning instead to ...



Relation to other fields

- **Statistics and Machine Learning**
 - Early NLP systems (before 1990) were purely symbolic and hand crafted
 - Statistical methods and models have become more widely used in NLP since the mid 1990's
 - Easily trainable and easily computable models have proven much more effective than the more complex hand-crafted models
 - This approach has become more attractive now as lots of training data is available (the “big data” era)



What functionality we need to address NLP applications?

- Consider machine translation
 - at first, people imagined that machine translation is mostly a “data processing” task where a system looks up the words one at a time in a bilingual dictionary, and then maybe has to fix up the translation a bit. However, there is a lot more to do for machine translation:



Machine Translation

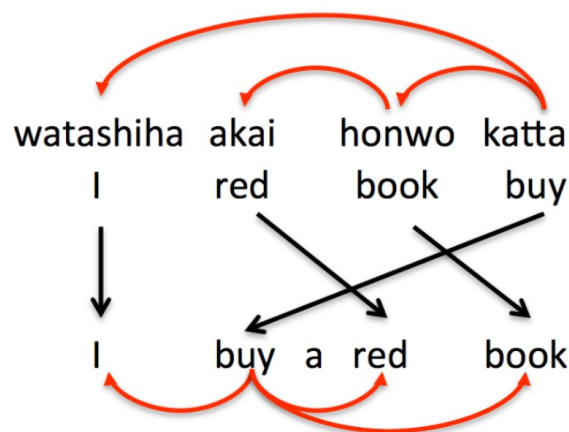
- Word segmentation
 - For some languages (e.g., Chinese, Japanese) there are no spaces between words, so it's not clear what the words are
- Morphology
 - Words appear in different forms, indicating singular vs. plural (for nouns), present tenses vs. past tense (for verbs), nominative vs. accusative case (different words forms for subjects and objects)
 - English has only a few morphological forms, so it's possible to put them all in a dictionary. This doesn't hold for most Western languages; for example, a Spanish verbs could have over 50 forms.



Machine Translation

- Syntax

- Word-for-word translation only works if the word order in the two languages is about the same
- Otherwise, we need to understand the structure of the two languages (i.e., their syntax) to change from one word order to another.
- English has a rather fixed subject-verb-object order (“SVO”), while many more inflected languages have more variable word order and some languages have basically different word orders (e.g., “SOV” for Japanese).



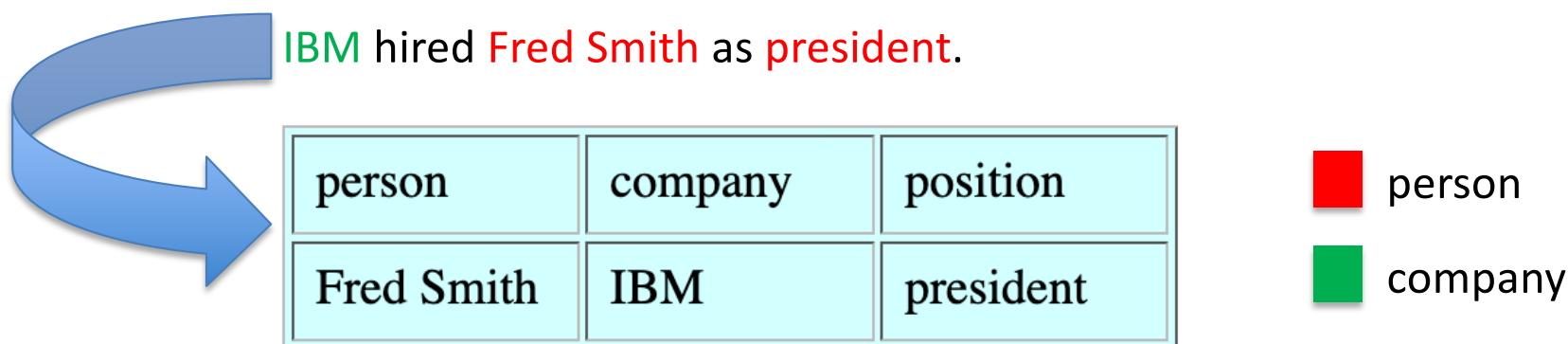
Machine Translation

- Discourse

- Sometime we need to look beyond the individual sentences to obtain proper translation.
- For instance, cross-sentence evidence might be needed to choose the correct word senses.
- Another example arises in *translating into English from languages where subject pronouns can be omitted*, then we need to determine what the subject actually is, so we can pick the right pronoun (i.e., “he”, “she”, or “it”) in English.



Similarly for Information Extraction



- **Name recognition**: knowing “Fred Smith” is a person while “IBM” is a company (these are called named entities).
- **Syntax**: the information may be presented in the passive form (“Fred Smith was hired by IBM”) or in a relative clause (“Fred Smith, who was hired by IBM”); also, there maybe extra modifiers (“IBM yesterday hired Fred Smith as president”).
- **Lexical Semantics**: there might be lots of synonyms for “hired” (i.e., “appointed”, “named”, ...)
- **Discourse – pronoun**: the system might need to determine what the pronoun refers to (“Fred Smith left Compaq last week. IBM hired him yesterday as president.”).

So, for NLP we need:

