

Chapter 1

Introduction

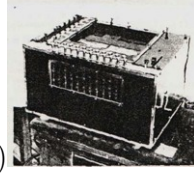
1.1 What is NLP?

- Analysis of a natural language (wait, what's a natural language)
- Generation of a natural language
- Sometimes both analysis and generation
- Representation of a natural language (but usually in the service of generation and/or analysis)

1.1.1 Super Brief History of NLP

- NLP was indistinguishable from MT for a long time for some weird reason
- Athanasius Kircher – ‘polygraphy’ (1630) for translation without understanding
 - side note: kircher is really weird guy! he
 - incorrectly deciphered hieroglyphics
 - Went into Vesuvius’ erupting crater for his geology research
 - invented a cat piano
 - check him out at the Museum of Jurassic Technology just outside Culver City
- Becher mechanical meta-language for language-to-meaning (in the service of NLP):





- Georges Artsruni mechanical brain (1930)
- Computers proposed for translation (1949...3 years after computer invented)

- ACL founded (as the “Association for Machine Translation”) in 1962



1.1.2 Applications

- Chatbots
- customer service agents
- Question Answering (search)
- Translation
- Grammar Correction
- Summarization
- Article Writing
- Opinion/sentiment analysis
- essay evaluation/plagiarism detection
- Knowledge base construction
- author identification
- ...

1.1.3 A Dialogue of the near-future

HUMAN: Can you get me a table at N/Naka tomorrow at 7?

AI: How many?

H: Four of us.

A: Sorry, there's nothing at 7 PM. They have an opening at 10 PM.

H: Ugh, that's sort of late. What about Sakura?

A: Sakura has an opening at 9:30 PM.

H: Can you email Brooke, Mike, and Liz, and ask if 9:30 is too late?

A: OK, email sent.

A: Liz says it's fine.

H: OK, let's book the table at Sakura.
 A: OK, all set for 4 people at 9:30 PM.
 [some time later...]
 H: You know what, just take the N/Naka one at 10 PM.
 A: OK, N/Naka is booked at 10 PM for 4 people and I cancelled the reservation at Sakura.

What parts of this conversation are possible with NLP tech today (cf. Google Duplex demo)? Which aren't? Which seem likely to be coming soon, based on your intuition?

1.1.4 Find the following NL 'tasks' in the above dialogue, or see where they would be useful

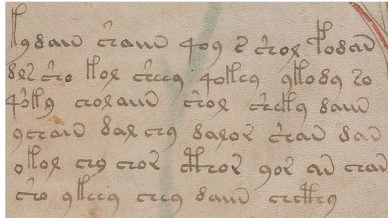
- Spelling Correction
- Named Entity Information Extraction
- Question Answering
- Coreference Resolution
- Grounding
- Ambiguity Resolution
- Summarization
- Information Retrieval
- Generation
- Translation

1.1.5 Connections to Other Areas

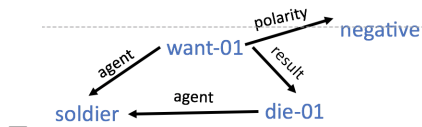
- Machine Learning – the biggie, now. NLP: ML is a tool we use. ML: NLP is a data set we use.
- Linguistics – for ML folks think of this as 'guided search' both within a problem and when considering what the problems are. But also consider we are trying to understand language and can use NLP/CL techniques to do so
- Cognitive Science / Psychology – see also Linguistics, but a level up. We are fascinated by humans' ability to learn language that is far better than computers' and we're not quite sure why this is. We probably won't get into real cog sci/psych theories, though
- information theory – language is a means by which humans communicate and the communication capacity/compression/confusability of communications is baked in to our studies, particularly when discussing (cross-)entropy and mutual information

- theory of computation – very important for search over complex spaces (e.g. for a syntactic tree or semantic graph) and for recognizing what transformations are and are not possible. Strongly connected via chomsky hierarchy, of which we'll only talk a little
- data science, political science, etc. – NLP: these are good subdomains to try our stuff. X science: NLP is a good tool to demonstrate my theories
- Other areas you're interested in not covered here?

1.2 Linguistic Stack – from low to high ambiguity (from shallow to deep)

- pre-text (speech channel)
 - phonetics – what mouth sound has been produced?
 - * [l] = alveolar lateral approximant (**l**ace)
 - * [ɾ] = alveolar tap (**r**ace)
 - * [r] = alveolar trill (**r**ey (Spanish))
 - phonology – what are the meaningfully distinct sounds (governed by each language)?
 - * English: [ɹ] vs. [r] conflated
 - * Japanese: [ɾ] vs. [r] vs. [l] conflated
 - * Hindi: ɖ (dental) vs. ɖʱ (dental, glottal) distinct, etc.
 - * cf https://en.wikipedia.org/wiki/Hindustani_phonology examples (retroflex are good ones to confuse my american ears)
- pre-text (vision channel)
 - orthography – what makes a character? Particularly difficult when dealing with unknown writing set, especially handwritten.
 
 - morphology – what are the minimal meaning-laden parts of a word that are useful to distinguish? (Why distinguish? For handling novelty (wug test), collapsing statistics...)
 - English is pretty weak here:
 - inflection: ‘talks’ = ‘talk (verb) + s (present 3rd singular)’ or ‘talk (noun) + s (plural).

- Turkish agglutination: *uygarlaştıramadıklarımızdanmışsınızcasına* = “(behaving) as if you are among those whom we could not civilize”
- words (lexemes)
 - Ok, not really a level but it’s important to recognize what we consider a word, especially when data processing
 - Is a word ‘a single unit of meaning?’ ‘text separated by whitespace?’
 - What about Chinese? Or Thai? Or long compounds/agglutinations in Turkish, German, Finnish?
 - What about whitespace-separated units that function noncompositionally (‘New York’, ‘take out’)?
 - What about hyphenated and punctuated text? (Tricky example: ‘New York-New Haven Railroad’)
- syntax – how to properly put words together to form a sentence
 - part-of-speech tags:
 - the/DT (determiner) blue/JJ (adjective) boat/NN (noun)
 - *boat/NN blue/JJ the/DT
 - Constituencies:
 - S = NP[the blue boat] VP[sailed home]
 - * VP[sailed home] NP[the blue boat]
- semantics – what does a word in a sentence mean, and how do the words meaningfully relate to each other?
 - Consider the sentence ‘The soldier did not want to die.’
 - What is meant by ‘want’ – desire? lack?
 - Who is doing the wanting? Who is doing the dying? What is (not) wanted?



- pragmatics – what does the *speaker* (as opposed to the sentence) mean in conversational context?

HUMAN: Can you get me a table at N/Naka tomorrow at 7?

*AI: Yes, I have that ability.

AI: OK, your reservation is made.

- discourse – what information is conveyed subtextually, as a result of context (interpretation of sentence in context to other sentence or sequence, overall intent of text or dialogue)?

WAITER: What would you like for dinner?

DINER: I had a heavy lunch.

WAITER: Let me tell you about our salads.

1.3 Ambiguity is the enemy of NLP

Humans are seemingly able to integrate lots of context, world knowledge, tone, etc. clues to clearly disambiguate ‘bank’ and ‘mean’ and ‘latex’, find no problem in deciding to not pick apart individual meaning in ‘make a decision’ or ‘take out’ or ‘make up’, and can easily conclude that if you can *gronfle* sixty *milchanks* in one hour, then after one minute you will have *gronfled* one *milchank*. Errors in NLP are chiefly due to not having sufficient context.

1.3.1 Funny (English) Examples made less funny by linguistic analysis – pick out the misinterpreted phenomenon!

- Enraged Cow Injures Farmer With Axe
- Ban on Nude Dancing on Governor’s Desk
- Teacher Strikes Idle Kids
- Hospitals are Sued by 7 Foot Doctors
- Iraqi Head Seeks Arms
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- Dinosaurs didn’t read. Now they are extinct.

1.3.2 Issues in ambiguity, richness

“We saw the woman with the telescope wrapped in paper” – who has the telescope? What is the paper wrapping? Is this perception or assault? Humans have two major readings of this (why?) but it’s hard to keep computers from considering the unlikely ones unlikely.

“Every fifteen minutes a woman in this country gives birth. Our job is to find that woman, and stop her!” Groucho Marx – ambiguity of semantics (‘a woman’)

‘The soldier was not afraid to die’ vs ‘The soldier did not fear death.’ – There are lots of ways to express the same thing and to us this is not an issue, but without proper intervention these are completely distinct sentences to a computer model.

Humans often produce intentionally obfuscated language, possibly to only target a subgroup. These obfuscations can change very quickly; it's tough to keep up! There are good reasons to keep up! Examples:

- Call me at six niner i three triple 0 dos
- u ship them? smh. ikr, lolol, yolo.
- 14 words now and then 88 later if u want

1.4 Structure Of the class

- The overall structure of learning is:
 - a tiny bit of linguistics
 - a refresher on probability you should know
 - mini-course on important aspects of machine learning (linear models, nonlinear models)
 - discussion of data
 - discussion of evaluation
 - some core techniques oriented toward parts of the linguistic stack
 - various end-goal subfields (MT, IE, Dialogue, maybe QA)
- I also want you to get experienced digesting the latest papers and to break the class up a bit so I'm not lecturing the whole time. Everybody will choose one paper from NAACL 2021 <https://aclanthology.org/events/naacl-2021/> and will make a 15-minute presentation on it in class. (You can't present your own work, BTW.) We'll all read the paper ahead of time and engage in active discussion on the paper using piazza and in-class.
- Based on class interest we can add or subtract topics, especially near the end of the class.
- There are three HW assignments and a project. As shown below, the majority of your effort should be in writing clearly and communicating what you have done.

1.5 Evaluation

- no punishment curve (but a reward curve if needed);
- be an active, engaged participant and do all the work and an A is easy to get.

1.5.1 Homeworks – 3x10% = 30% total

- Expect substantial programming in most (if not all) of them
- Coding should be strictly in Python – in some cases we may provide useful templates for parts of the assignment.
- Writeups should be strictly in L^AT_EX; learn how to use it now (if you don't already)!
- Communicate well; write clearly and simply, use appropriate figures and graphs.
- There will be an expectation of self-exploration, reading papers to get good ideas to reimplement or build upon.
- **ALL CODE MUST BE YOUR OWN AND MAY NOT BE COPIED** which includes solutions you find on the web and code from others in the class. We run code-checking software; it is smart enough to defeat attempts to obfuscate, and I take cheating very seriously (see below).
- Grading (approximate and totally subjective) presuming that you actually did what was asked:
 - about 50% – did you clearly communicate your description of what you implemented, how you implemented it, what your experiments were, and what conclusions you drew from them? This includes appropriate use of graphics and tables where warranted that clearly explain your point. This also includes well written explanations that tell a compelling story. Grammar and syntax are a small part of this (maybe 5%) but much more important is the narrative you tell. Also a part of this is that you clearly acknowledged your sources and influences with appropriate bibliography and, where relevant, cited influencing prior work.
 - about 20% – is your code correct? Did you implement what was asked for, and did you do it correctly?
 - about 20% – is your code well-written, documented, and robust? Will it run from a different directory than the one you ran it in? Does it rely on hard-codes? Is it commented and structured such that we can read it and understand what you are doing?
 - about 10% – did you go the extra mile? Did you push beyond what was asked for in the assignment, trying new models, features, or approaches? Did you use motivation (and document appropriately) from another researcher trying the same problem or from an unrelated but transferrable other paper? **THIS IS NOT EXTRA CREDIT! YOU CANNOT RECEIVE 100% WITHOUT COMPLETING THIS PART!**. There is no extra credit on homeworks.

1.5.2 In-class paper presentation = 10%

- The schedule contains paper listings under “presentation”; everyone will sign up for (at least one) NAACL 2021 paper to present to the class and lead discussion (approx 15 minutes presentation, 5-10 minutes discussion, but this is flexible)

- Everyone will read the papers (see below) and prepare questions ahead of time to facilitate discussion
- You will explain the paper, getting into key details and insights as well as the context of the paper (you may have to look at key papers that cited this paper as well as key influential works that the paper cites)
- Slides or a handout may be helpful and are a good idea but are not mandatory
- Submit your top 3 papers as a response to a poll we will send; we'll do our best to give everyone their top choice but can't promise. If you don't pick one we'll pick for you!

1.5.3 Project—5% (proposal) + 5% (version 1 of report) + 10% (final presentation) + 20%(final report) = 40%

- Two people per project
- Reproduce results in an ACL 2021 paper
- Write up your results clearly
- Proposal is due in two weeks from the beginning of class!

1.5.4 Pre-written paper presentation questions for others' presentations— $10/(N-1)\% \times N-1 = 10\%$

- Before the class in which a presentation is going to be given (i.e. in 2020, by 9:59 am Pacific time on the day of the class), post at least one question regarding each presentation (usually 1 but occasionally 2) to the appropriate location on piazza.
- During the presentation, if the question isn't answered or isn't answered sufficiently, bring it up and engage in discussion with the class.
- There probably won't be enough time for everyone; if the question is unresolved it should be discussed on piazza (after or before the presentation)
- The main goal of this is to ensure that you've read the papers and are engaging in discussion.

1.5.5 Other class participation—10%

- Ask general questions in class, engage in discussion
- Ask questions and engage in discussion on piazza – answer each others' questions before instructors weigh in
- Propose topics to cover

- Ask questions of fellow students during project presentations
- Answer questions when I call on you and be in class (occasional absences understandable).

1.5.6 Late Days

- Work is done at 11:59:59 anywhere on earth (=4 AM the next day, for PST) on the announced due date (except for the final report).
- You get four *cumulative* late days for homeworks and project proposals (no late days for final project report or missed presentations). Thereafter, 20% off per day.
- Late group project proposals will deduct from both team members' late day accounts.

1.5.7 Office Hours

- 2021: Mine are 2pm–3pm before class, in SAL 311 (or on zoom).
- Elan's are TBD
- Come bounce ideas off of us (particularly related to project proposal/project)

1.6 Don't Cheat

- Read the USC honor code; it applies, and I will abide by it.
- All work you turn in should be your own.
- This includes anything written and all code.
- All work must be originally done **for this class** by you. Self-plagiarizing is still plagiarizing!
- That means that if you have done any of these assignments before in a previous class, you should either not look at your previous work, or (better) come talk to me and I will give you an alternate assignment
- Similarly, for the project, you may not re-present a research project or paper you have previously worked on or are currently working on; this should be entirely new work.
- If we have determined you have violated the honor code we will invoke punishments as deemed necessary; this can mean a zero on an assignment, a reduced letter grade in the class, or even a failing grade. Punishments can occur at any time after violations (usually at the worst possible time). I hate doing this but I will if necessary (ask around).

This is intended to be a fairly comprehensive list of policies and provisions but something may have been missed; other policies or changes to existing policy may be announced and will supersede any conflicting statements made here.