# Introduction to NLP

COMP 473 – Speech and Language Processing

April 8, 2019

# Contents

- What is NLP?
- Some linguistic terminology
- Why is language processing difficult?
- Some NLP applications

# What is NLP?

- Automatic (or semi-automatic) processing of human language.
- Other interchangeably used terms – “Computational Linguistics”, “Language Technology”, “Language Engineering”.
- NLP is essentially multidisciplinary: it is closely related to linguistics. Also has links to research in Cognitive Science, Psychology, Philosophy and Maths (especially Logic).
- Within CS, it relates to formal language theory, compiler techniques, theorem proving, machine learning and human-computer interaction.
- Of course, it is also related to AI, although nowadays, it’s not generally thought of as a part of AI.

# Some linguistic terminology

- Morphology: the structure of words. For example, *unusually* can be thought of as composed of a prefix *un-*, a stem *usual*, and an affix *–ly*.
- Syntax: the way words are used to form phrases, e.g., it is part of English syntax that a determiner such as *the* will come before a noun, and also that determiners are obligatory with certain singular nouns.
- Semantics: Compositional semantics is the construction of meaning (generally expressed as logic) based on syntax. This contrasts to lexical semantics which deals with the meaning of individual words.
- Pragmatics: meaning in context.

# Why is language processing difficult?

- Consider trying to build a system that would answer email sent by customers to a retailer selling laptops and accessories via the Internet.
- This might be expected to handle queries such as the following:
  - Has my order number 4291 been shipped yet?
  - Is FD5 compatible with a 505G?
  - What is the speed of the 505G?

# Why is language processing difficult?

- Assume the query is to be evaluated against a database containing product and order information, with relations such as the following:
- ORDER

| Order number | Date ordered | Date shipped |
|---|---|---|
| 4290 | 2/2/02 | 2/2/02 |
| 4291 | 2/2/02 | 2/2/02 |
| 4292 | 2/2/02 | |

# Why is language processing difficult?

USER: Has my order number 4291 been shipped yet?
DB QUERY: order(number=4291, date_shipped=?)
RESPONSE TO USER: Order number 4291 was shipped on 2/2/02

It might look quite easy to write patterns for these queries, but very similar strings can mean very different things, while very different strings can mean much the same thing.

Below, 1 and 2 look very similar but mean something completely different, while 2 and 3 look very different but mean much the same thing.

1. How fast is the 505G?
2. How fast will my 505G arrive?
3. Please tell me when can I expect the 505G I ordered.

# Why is language processing difficult?

- While some tasks in NLP can be done adequately without having any sort of account of meaning, others require that we construct detailed representations which reflect the underlying meaning rather than the superficial string.
- In fact, in natural languages (as opposed to programming languages), ambiguity is ubiquitous, so exactly the same string might mean different things.
- For instance, in the query:
  - Do you sell Sony laptops and disk drives?

The user may or may not be asking about Sony disk drives.

- Often humans have knowledge of the world which resolves a possible ambiguity, probably without the speaker or hearer even being aware that there is a potential ambiguity.
- But hand-coding such knowledge in NLP applications has turned out to be impossibly hard to do for more than very limited domains.

# Why is language processing difficult?

- The term AI-complete is sometimes used (by analogy to NP-complete), meaning that we'd have to solve the entire problem of representing the world and acquiring world knowledge.
- The guiding principle in current NLP – we're looking for applications which don't require AI-complete solutions: i.e., ones where we can work with very limited domains or approximate full world knowledge by relatively simple techniques.

# Some NLP applications

The following list is not complete, but useful systems have been built for:

- spelling and grammar checking
- optical character recognition
- screen readers for blind and partially sighted users
- augmentative and alternative communication(i.e., systems to aid people who have difficulty communicating because of the disability)
- machine aided translation (i.e., systems which help a human translator, e.g., by storing translation of phrases and providing online dictionaries integrated with word processors, etc.)
- lexicographers' tools
- information retrieval
- document classification (filtering, routing)

# Some NLP applications

- document clustering
- information extraction
- question answering
- Summarization
- text segmentation
- exam marking
- report generation (possibly multilingual)
- machine translation
- natural language interface to databases
- email understanding
- Dialogue systems

# Spelling and grammar checking

- All spelling checkers can flag words which aren't in the dictionary
- If the user can expand the dictionary, or if the language has complex productive morphology, then a simple  list of words isn't enough to do this and some morphological processing is needed.
- More subtle cases involve words which are correct in isolation, but not in context. Syntax could sort some of these cases out.

# Information retrieval, information extraction and question answering

- Information retrieval involves returning a set of documents in response to a user query: Internet search engines are a form of IR.
- However, one change from classical IR is that Internet search now uses techniques that rank documents according to how many links there are to them (e.g., Google's PageRank) as well as the presence of search terms.
- Information extraction involves trying to discover specific information from a set of documents.
- The information required can be described as a template.
- For instance, for company joint ventures, the template might have slots for the companies, the dates, the products, the amount of money involved.
- The slot fillers are generally strings.

# Information retrieval, information extraction and question answering

- Question answering attempts to find a specific answer to a specific question from a set of documents, or at least a short piece of text that contains the answer.
  What is the capital of France?
  Paris has been the French capital for many centuries.
- There are some question-answering systems on the Web, but most use very basic techniques.
- For instance, Ask Jeeves relies on a fairly large staff of people who search the web to find pages which are answers to potential questions.
- The system performs very limited manipulation on the input to map to a known question.
- The same basic technique is used in many online help systems.

# Machine Translation

- MT work started in the US in the early fifties, concentrating on Russian to English.
- A prototype system was publicly demonstrated in 1954.
- MT fund got drastically cut in the US in the mid-60s and ceased to be academically respectable in some places, but Systran was providing useful translations by the late 60s.
- Systran is still going, it now powers AltaVista's BabelFish, http://world.altavista.com and many other translation services on the web.