An Introduction To Natural Language Processing

Chapter 1: Introduction and Roadmap

Jacob Eisenstein

This course

Natural Language Processing is the set of methods for making language accessible to computers.

- This course is about learning what methods are available,
- ...how and why they work,
- ...and how they can best be applied to real problems.

Day 1

- Natural language processing and its neighbors
- Three themes in natural language processing

Natural language processing and its neighbors

Natural language processing draws on a diverse array of intellectual traditions.

- **Linguistics**
- **Machine learning**
- **Artificial intelligence**
- **Computer science**

NLP also raises questions about **human-computer interaction** and **ethics, fairness,** and **accountability**.

Linguistics

The goal of **linguistics** is understand how language works — possibly using computational techniques. For example:

- What are the major language families and how are they related to each other?
- What are the principles that determine whether a sentence is grammatical? Can we identify shared principles that explain grammaticality across many different kinds of languages?
- How and why do languages change?
- How do people learn their first language? What, if anything, is different when they learner their second language?

Natural language processing leverages insights from linguistics to build language technology.

Machine learning

Machine learning makes it possible to build complex computer programs from examples.[1]

- Due to the complexity of natural language, virtually all successful NLP applications today involve some amount of machine learning.
- For example, machine translation systems are built from examples of translations, not from rules or dictionaries.
- For this reason, this course begins by building a foundation in machine learning.
- *If you're already familiar with machine learning, think about what makes language a unique application domain.*

[1] Kevin P. Murphy (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Artificial intelligence

The goal of artificial intelligence (AI) is to automate human mental capabilities.[2]

- Language is a fundamental aspect of human intelligence.
- Language can help solve the **knowledge bottleneck**, giving AI systems the knowledge they need to make useful inferences.
- Reasoning is sometimes essential for language understanding:

(1) a. The trophy doesn't fit in the suitcase because **it** is too **big**.[3]
b. The trophy doesn't fit in the suitcase because **it** is too **small**.

[2] Stuart J Russell and Peter Norvig (2009). *Artificial intelligence: a modern approach*. 3rd. Prentice Hall.

[3] This notation introduces a linguistic example.

Computer science

Natural language processing draws on several aspects of "core" computer science:

- Natural language can be modeled using **formal language theory**, building on similar theoretical tools that are used to analyze programming languages.
- Natural language data requires efficient algorithms, which can be analyzed in terms of **time and space complexity**.
- These algorithms must be implemented on diverse architectures, including distributed systems, GPUs, and mobile devices.

This course will draw on basic tools from complexity theory,[4] and will highlight connections to other areas of computer science.

[4] Michael Sipser (2012). *Introduction to the Theory of Computation*. Cengage Learning.

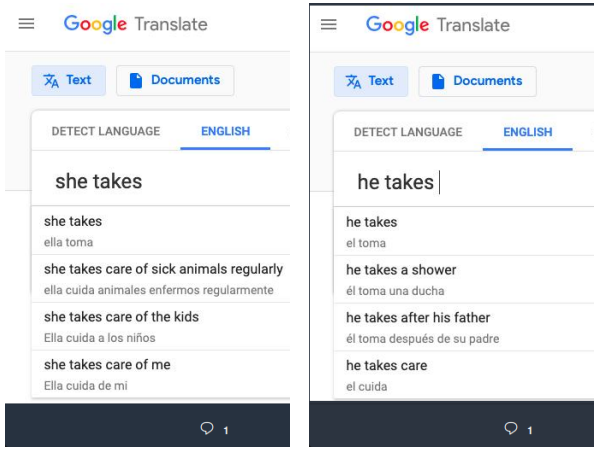
Ethical and social dimensions

Natural language processing raises many broader social questions. Here are just a few:[5]

Access. Who are the users? For example, whose language is translated *from*, and whose is translated *to*?

Bias. Does NLP replicate and reinforce social biases that are present in textual data?

Privacy. What is the role of NLP with respect to freedom expression and surveillance?



[5] Screenshots retrieved by Ian Stewart on February 13, 2019

Three themes in natural language processing

1. Learning and knowledge
2. Search and learning
3. Relational, compositional, and distributional perspectives

Learning and knowledge

Given the dominance of machine learning, what role is left for linguistic theory? Some possibilities:

- The **NLP stack**: A series of systems transforms text from raw strings into progressively higher level linguistic representations.
- **Preprocessing**: The base representation for machine learning is a set of linguistically meaningful features.
- **Model design**: The architecture of the learning algorithm is designed to reflect linguistic principles.
- **Nothing**: Language is just another kind of data, and language processing is just another learning problem.

Learning and knowledge

- The **poverty of the stimulus** hypothesis: children learn language so quickly that our brains must be configured for language in advance.[6]
- Similarly, machine learning theory argues that generalization requires an **inductive bias** toward the desired model.
- In theory, adding a bias toward theoretically-motivated linguistic structures should make machine learning more effective.
- But in practice, recent progress has mostly gone in the opposite direction, with language-neutral learning techniques playing an increasingly important role.

[6] Steven Pinker (2003). *The language instinct: How the mind creates language*. Penguin UK.

Search and learning

Many natural language processing problems can be written mathematically in the form of optimization,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})}{\operatorname{argmax}} \, \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}), \tag{1}$$

where,

- $\boldsymbol{x}$ is the input, which is an element of a set $\mathcal{X}$;
- $\boldsymbol{y}$ is the output, which is an element of a set $\mathcal{Y}(\boldsymbol{x})$;
- Ψ is a scoring function (also called the **model**), which maps from the set $\mathcal{X} \times \mathcal{Y}$ to the real numbers;
- $\boldsymbol{\theta}$ is a vector of parameters for Ψ;
- $\hat{\boldsymbol{y}}$ is the predicted output, which is chosen to maximize the scoring function.

Search and learning

Many natural language processing problems can be written mathematically in the form of optimization,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})}{\operatorname{argmax}} \, \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}), \tag{1}$$

- The **search** problem is to find $\hat{\boldsymbol{y}}$. When $\mathcal{Y}(\boldsymbol{x})$ is too large to enumerate, search is often done by combinatorial optimization.
- The **learning** problem is to select the parameters $\boldsymbol{\theta}$, usually by minimizing some function of a labeled dataset, $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}$. This problem is usually solved by numerical optimization.
- The use of generic optimization algorithms for search and learning can simplify the engineering of NLP software, but it can also limit flexibility.

Relational, compositional, and distributional perspectives

Any element of language can be described from at least three perspectives:

Relational: How does it relate to other elements of the language? For example: equivalence, opposition, implication.

Compositional: How is its meaning a function of the meanings of its component parts?

Distributional: In what contexts does it tend to appear?

The relational perspective

Who works for the college newspaper?

(2) Umashanthi interviewed Ana.
She works for the college newspaper.

The relational perspective

Who works for the college newspaper?

(2) Umashanthi interviewed Ana.
She works for the college newspaper.

- A **journalist** works for a **newspaper**, and also performs **interviews**.
- This creates a relational link between the interviewer (Ana) and the newspaper worker (She).
- An **ontology** is a network of relations between fundamental semantic units. WordNet is one example.

The relational perspective

Borges' *Celestial Emporium of Benevolent Knowledge*[7] divides animals into:

> *(a) belonging to the emperor; (b) embalmed; (c) tame; (d) suckling pigs; (e) sirens; (f) fabulous; (g) stray dogs; (h) included in the present classification; (i) frenzied; (j) innumerable; (k) drawn with a very fine camelhair brush; (l) et cetera; (m) having just broken the water pitcher; (n) that from a long way off resemble flies.*

It's hard to design an ontology that satisfies everyone!

[7] Borges 1993.

The compositional perspective

- Many elements in language can be analyzed as functions of smaller constituent parts:
  - $journalists = journalist{+}s = (journal{+}ist){+}s$

The **principle of compositionality** is often attributed to German philosopher Gottlob Frege.

The compositional perspective

- Many elements in language can be analyzed as functions of smaller constituent parts:
  - $journalists = journalist{+}s = (journal{+}ist){+}s$
  - $journal = jour{+}nal$ (in French)

The **principle of compositionality** is often attributed to German philosopher Gottlob Frege.

The compositional perspective

- Many elements in language can be analyzed as functions of smaller constituent parts:
  - $journalists = journalist{+}s = (journal{+}ist){+}s$
  - $journal = jour{+}nal$ (in French)
  - These parts (**morphemes**) play the same role in other words, like *soloists* and *optimists*.
- The same approach can be applied to phrases, sentences, and beyond.
- But some phrases must be analyzed **non-compositionally**, e.g. *kick the bucket*, *shoot the breeze*.

The **principle of compositionality** is often attributed to German philosopher Gottlob Frege.

The distributional perspective

"You shall know a word by the company it keeps"[8]

(3) a. The blubber served them as fuel.
b. ...extracting it from the blubber of the large fish ...
c. Amongst oily substances, blubber has been employed as a manure.

- These examples link *blubber* to other words that appear in similar contexts, like *fat*, *pelts*, and *barnacles*.
- The distributional perspective is implemented in techniques like **word2vec**.

[8] Firth 1957.

Relational, compositional, and distributional perspectives

Any element of language can be described from at least three perspectives:

Relational: How does it relate to other elements of the language? For example: equivalence, opposition, implication.

Compositional: How is its meaning a function of the meanings of its component parts?

Distributional: In what contexts does it tend to appear?

Relational, compositional, and distributional perspectives

Any element of language can be described from at least three perspectives:

Relational: How does it relate to other elements of the language? For example: equivalence, opposition, implication.

Compositional: How is its meaning a function of the meanings of its component parts?

Distributional: In what contexts does it tend to appear?

Natural language processing research is constantly engaged in finding new syntheses of these views of language.

References I

- Borges, Jorge Luis (1993). *Other Inquisitions 1937–1952*. Translated by Ruth L. C. Simms. University of Texas Press.
- Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press.
- Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- Pinker, Steven (2003). *The language instinct: How the mind creates language*. Penguin UK.
- Russell, Stuart J and Peter Norvig (2009). *Artificial intelligence: a modern approach*. 3rd. Prentice Hall.
- Sipser, Michael (2012). *Introduction to the Theory of Computation*. Cengage Learning.