0 - what-is-nlp

November 29, 2021

These notebooks are found in the repo: https://github.com/fastai/course-nlp

1 1. What is NLP?

1.1 What can you do with NLP?

NLP is a broad field, encompassing a variety of tasks, including:

- Part-of-speech tagging: identify if each word is a noun, verb, adjective, etc.)
- Named entity recognition NER): identify person names, organizations, locations, medical codes, time expressions, quantities, monetary values, etc)
- Question answering
- Speech recognition
- Text-to-speech and Speech-to-text
- Topic modeling
- Sentiment classification
- Language modeling
- Translation

Many techniques from NLP are useful in a variety of places, for instance, you may have text within your tabular data.

There are also interesting techniques that let you go between text and images:

from Lesson 9 of Practical Deep Learning for Coders 2018.

1.2 This course

In this course, we will cover these applications: - Topic modeling - Sentiment classification - Language modeling - Translation

1.2.1 Top-down teaching approach

I'll be using a *top-down* teaching method, which is different from how most CS/math courses operate. Typically, in a *bottom-up* approach, you first learn all the separate components you will be using, and then you gradually build them up into more complex structures. The problems with this are that students often lose motivation, don't have a sense of the "big picture", and don't know what they'll need.

If you took the fast.ai deep learning course, that is what we used. You can hear more about my teaching philosophy in this blog post or in this talk.

Harvard Professor David Perkins has a book, Making Learning Whole in which he uses baseball as an analogy. We don't require kids to memorize all the rules of baseball and understand all the technical details before we let them play the game. Rather, they start playing with a just general sense of it, and then gradually learn more rules/details as time goes on.

All that to say, don't worry if you don't understand everything at first! You're not supposed to. We will start using some "black boxes" that haven't yet been explained, and then we'll dig into the lower level details later. The goal is to get experience working with interesting applications, which will motivate you to learn more about the underlying structures as time goes on.

To start, focus on what things DO, not what they ARE.

1.3 A changing field

Historically, NLP originally relied on hard-coded rules about a language. In the 1990s, there was a change towards using statistical & machine learning approaches, but the complexity of natural language meant that simple statistical approaches were often not state-of-the-art. We are now currently in the midst of a major change in the move towards neural networks. Because deep learning allows for much greater complexity, it is now achieving state-of-the-art for many things.

This doesn't have to be binary: there is room to combine deep learning with rules-based approaches.

1.3.1 Case study: spell checkers

This example comes from Peter Norvig:

Historically, spell checkers required thousands of lines of hard-coded rules:

A version that uses historical data and probabilities can be written in far fewer lines:

Read more here.

1.3.2 A field in flux

The field is still very much in a state of flux, with best practices changing.

1.3.3 Norvig vs. Chomsky

This "debate" captures the tension between two approaches:

- modeling the underlying mechanism of a phenomena
- using machine learning to predict outputs (without necessarily understanding the mechanisms that create them)

This tension is still very much present in NLP (and in many fields in which machine learning is being adopted, as well as in approachs to "artificial intelligence" in general).

Background: Noam Chomsky is an MIT emeritus professor, the father of modern linguistics, one of the founders of cognitive science, has written >100 books. Peter Norvig is Director of Research at Google.

From MIT Tech Review coverage of a panel at MIT in 2011:

"Chomsky derided researchers in machine learning who use purely statistical methods to produce behavior that mimics something in the world, but who don't try to understand the meaning of that behavior. Chomsky compared such researchers to scientists who might study the dance made by a bee returning to the hive, and who could produce a statistically based simulation of such a dance without attempting to understand why the bee behaved that way. “That’s a notion of scientific success that’s very novel. I don’t know of anything like it in the history of science,” said Chomsky.”

From Norvig’s response On Chomsky and the Two Cultures of Statistical Learning:

“Breiman is inviting us to give up on the idea that we can uniquely model the true underlying form of nature’s function from inputs to outputs. Instead he asks us to be satisfied with a function that accounts for the observed data well, and generalizes to new, previously unseen data well, but may be expressed in a complex mathematical form that may bear no relation to the”true” function’s form (if such a true function even exists). Chomsky takes the opposite approach: he prefers to keep a simple, elegant model, and give up on the idea that the model will represent the data well.”

- Noam Chomsky on Where Artificial Intelligence Went Wrong: An extended conversation with the legendary linguist
- Norvig vs. Chomsky and the Fight for the Future of AI

1.3.4 Yann LeCun vs. Chris Manning

Another interesting discussion along the topic of how much linguistic structure to incorporate into NLP models is between Yann LeCun and Chris Manning:

Deep Learning, Structure and Innate Priors: A Discussion between Yann LeCun and Christopher Manning:

On one side, Manning is a prominent advocate for incorporating more linguistic structure into deep learning systems. On the other, LeCun is a leading proponent for the ability of simple but powerful neural architectures to perform sophisticated tasks without extensive task-specific feature engineering. For this reason, anticipation for disagreement between the two was high, with one Twitter commentator describing the event as “the AI equivalent of Batman vs Superman”.

…

Manning described structure as a “necessary good” (9:14), arguing that we should have a positive attitude towards structure as a good design decision. In particular, structure allows us to design systems that can learn more from less data, and at a higher level of abstraction, compared to those without structure.

Conversely, LeCun described structure as a “necessary evil” (2:44), and warned that imposing structure requires us to make certain assumptions, which are invariably wrong for at least some portion of the data, and may become obsolete within the near future. As an example, he hypothesized that ConvNets may be obsolete in 10 years (29:57).

1.4 Resources

Books

We won’t be using a text book, although here are a few helpful references:

- **Speech and Language Processing**, by Dan Jurafsky and James H. Martin (free PDF)
- **Introduction to Information Retrieval** by By Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (free online)

- **Natural Language Processing with PyTorch** by Brian McMahan and Delip Rao (need to purchase or have O'Reilly Safari account)

Blogs

Good NLP-related blogs: - Sebastian Ruder - Joyce Xu - Jay Alammar - Stephen Merity - Rachael Tatman

1.5 NLP Tools

- Regex (example: find all phone numbers: 123-456-7890, (123) 456-7890, etc.)
- Tokenization: splitting your text into meaningful units (has a different meaning in security)
- Word embeddings
- Linear algebra/matrix decomposition
- Neural nets
- Hidden Markov Models
- Parse trees

Example from http://damir.cavar.me/charty/python/: "She killed the man with the tie."

Was the man wearing the tie?

Or was the tie the murder weapon?

1.6 Python Libraries

- nltk: first released in 2001, very broad NLP library
- spaCy: creates parse trees, excellent tokenizer, opinionated
- gensim: topic modeling and similarity detection

specialized tools: - PyText - fastText has library of embeddings

general ML/DL libraries with text features: - sklearn: general purpose Python ML library - fastai: fast & accurate neural nets using modern best practices, on top of PyTorch

1.7 A few NLP applications from fast.ai students

Some things you can do with NLP:

- How Quid uses deep learning with small data: Quid has a database of company descriptions, and needed to identify company descriptions that are low quality (too much generic marketing language)

- Legal Text Classifier with Universal Language Model Fine-Tuning: A law student in Singapore classified legal documents by category (civil, criminal, contract, family, tort,...)

- Democrats 'went low' on Twitter leading up to 2018: Journalism grad students analyzed twitter sentiment of politicians

- Introducing Metadata Enhanced ULMFiT, classifying quotes from articles. Uses metadata (such as publication, country, and source) together with the text of the quote to improve accuracy of the classifier.

1.8 Ethics issues

1.8.1 Bias

- How Vector Space Mathematics Reveals the Hidden Sexism in Language
- Semantics derived automatically from language corpora contain human-like biases
- Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them
- Word Embeddings, Bias in ML, Why You Don't Like Math, & Why AI Needs You

1.8.2 Fakery

OpenAI's New Multitalented AI writes, translates, and slanders

He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse. (interview with Avi Ovadya)

- Generate an audio or video clip of a world leader declaring war. "It doesn't have to be perfect — just good enough to make the enemy think something happened that it provokes a knee-jerk and reckless response of retaliation."

- A combination of political botnets and astroturfing, where political movements are manipulated by fake grassroots campaigns to effectively compete with real humans for legislator and regulator attention because it will be too difficult to tell the difference.

- Imagine if every bit of spam you receive looked identical to emails from real people you knew (with appropriate context, tone, etc).

How Will We Prevent AI-Based Forgery?: "We need to promulgate the norm that any item that isn't signed is potentially forged."