Linguistics and Computer Science: A Dual Degree Program

Open Day

22.2.2023

Wouldn’t life be easier if

- You could tell the computer what you wanted and it understood you (no programming skills required)?
- You could dictate a letter to the computer, it printed it and then saved it as a file?
- Having no time to read a 1000 page book, you could ask the computer to summarize it for you and it produced a one page summary in a few minutes?
- You could ask the computer to translate for you a text in Japanese which you did not understand?

Explosion of Text Data

- Explosion of linguistic data online
- Social media (twitter, facebook, blogs)
- Linguistic data is not easily amenable to analysis. It requires much processing and **much** insight into the nature and structure of language
- Modifying old computational techniques and tools as well as inventing new ones: for example, no off-the-shelf program can handle Twitter conversations.

A linguist in industry?

Examples:

- ChatGPT
- Information retrieval (esp. search engines that do semantic search)
- Voice recognition/generation – think 'google voice‘ or Siri
- Text classification and text clustering
- Text mining – finding grains of useful info in unstructured text
- Analyzing the language of social media (topic and sentiment extraction from short fragmented and noisy data)

Chat filtering: an example

- A linguistics application for on-line virtual worlds
- Ensuring safety of on-line chat environments
- Filtering chat for inappropriate content
- Filtering is an example of a classification problem: classify text as ‘appropriate’ or ‘inappropriate’

An example: Chat filtering

- The problem of determining whether lines involve inappropriate content is similar to spam detection
- Simple word/phrase matches are not enough:
- It will be too aggressive: “Hag Sameah”
- It will also not be enough: some nefarious lines may pass through
- Most inappropriate talk is made up of completely innocent words!

An example: Chat filtering

- People use innocent words to make inappropriate phrases
- People also find ways to say things with **MixEdCAse** or **b,r,o.k.en w.o.r,ds** that get around the filter
- **We want**: a general way of saying: "*if you see MixED CaSE or br.ok.en word,s it is probably a sign of something bad*"
- **We also want**: to capture inappropriate combinations of innocent words
- What is the general idea?

The idea behind filtering

- Look at the words and other features that make up appropriate and inappropriate chat, and ask *how likely is a word to appear in inappropriate chat?*

- **Example**: If a pair of words such as "stew pit" never appears in regular, appropriate chat, it is probably an indicator of something inappropriate.

- Having done that, we can ask for a new chat segment**, is it likely** to be inappropriate?

That said…

- Filtering is an example of a classification problem
- Computer scientists have developed well known algorithms for classification
- The difficult part is not the algorithm itself, but the fine-tuning of its parameters and manipulating and preprocessing the data
- This requires creative and analytical thinking…
- …and a thorough understanding of how language is used

Another example: Clustering by topic

- Given a set of texts, put together texts that are on the same topic
- Goal: arrive at K clusters, each representing a topic of the text
- Usual method: look for common words
- However….

Clustering Twitter posts

- Clustering super-short twitter posts by topic
- These data are very noisy and fragmented

lets see im on lates on monday so dont start till two and could get down from work .....mail me xxx

Well i'm watching mate running at silly o'clock but i'll be free in the late afternoon

- Resistant to common computational techniques
- Problem: very little to go on!

Clustering Twitter posts

- Possible solutions:
  - Applying spellcheckers to reduce the noise in the data
  - Removing various content-less function words, known as stop-words.
- But still:
- even those posts that share a common topic, do not often share many words in common

Clustering Twitter posts

- Therefore:
- Compare not the words themselves, but the linguistic contexts in which the words in posts appear
- Again: requires thorough understanding of how language is used

Moral

- A combination of linguistics and computer science:
- Efficeint computational algorithms that are guided by insightful understanding of the way language works

The Dual Degree: Linguistics and Computer Science

The Dual Degree: Linguistics and Computer Science

- You will receive a B.Sc. In computer science and a B.A. in linguistics
- You will be able to continue to a Master's degree in either department (some make-up courses may be necessary)

The Dual Degree: Linguistics and Computer Science

- The program is composed of:

1. Required courses on the foundations of linguistics and computer science
2. Electives, which allow you to learn topics that interest you
3. Integrative courses, demonstrating how the combination of linguistics and computer science can solve theoretical and practical problems
4. A final project

The Dual Degree: Linguistics and Computer Science

Coordinated by:

Ariel Cohen, Linguistics

Yuval Pinter, Computer Science

For more information:

arikc@bgu.ac.il

Thank you!