NLP Report

Guchuan Qiu

Simon Business School

This report will represents 4 potential approaches to NLP (reference by TOPBOTS) and particular method I was using.

1. Distributional Approach

These methods typically turn content into word vectors for mathematical analysis and perform quite well at tasks such as part-of-speech tagging, dependency parsing, and semantic relatedness. They can be applied widely to different types of text without the need for hand-engineered features or expert-encoded domain knowledge.

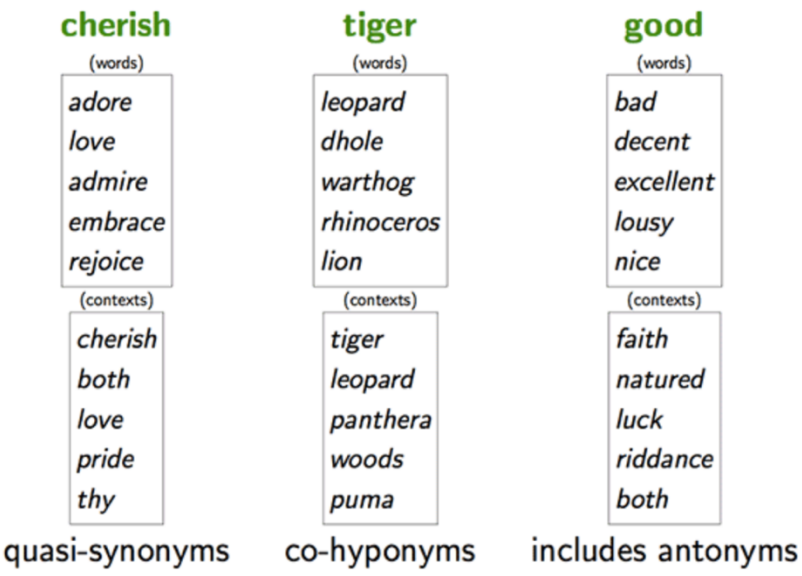

Figure 1: Distributional Approach

2. Frame-based Approach

Professor provides the example of a commercial transaction as a frame. In such situations, you typically have a seller, a buyers, goods being exchanged, and an exchange price. Sentences that are syntactically different but semantically identical – such as "Cynthia sold Bob the bike for $200" and "Bob bought the bike for $200 from Cynthia" – can be fit into the same frame. Parsing then entails first identifying the frame being used, then populating the specific frame parameters – i.e. Cynthia, $200.

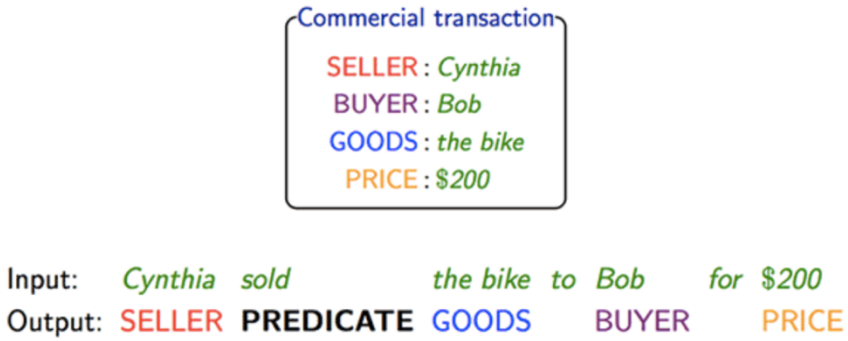

Figure 2: Frame-based Approach

3. Model-theoretical Approach

The approach is the category of semantic analysis falls under it. The advantages of model-based methods include full-world representation, rich semantics, and end-to-end processing, which enable such approaches to answer difficult and nuanced search queries. The major con is that the applications are heavily limited in scope due to the need for hand-engineered features.

4. Interactive learning

It is a viable approach to tackling both breadth and depth in language learning is to employ dynamic, interactive environments where humans teach computers gradually. In such approaches, the pragmatic needs of language inform the development.

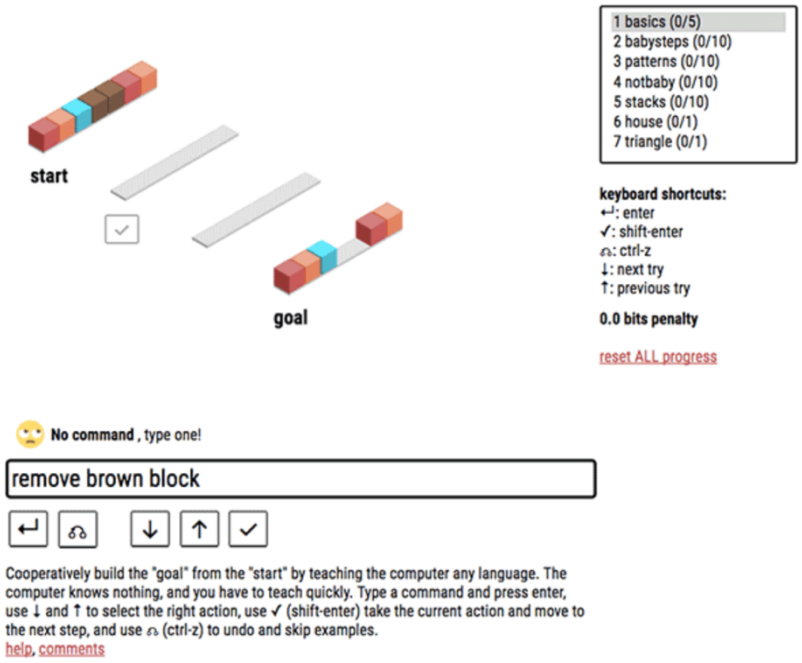

Figure 3: Interactive Learning

My particular method

Now, I'll introduce my approach to this problem. First I integrated the data file of Facebook from 2011 to 2015. I removed sparse items with a sparsity of 0.996, using the transformations and *dtm* buildings which discussed in class. Afterwards, I would get the frequency of the cauliflower corresponding to the date(year and month). Then I used for loop function to generate a list containing frequency of 'cauliflower' in each every month, by strsplitting the row names of "-". One problem is that the date is not ordered by sequence at first. That's why I added '0' before units if the length of month is less than 10.

	date	freq_cau
fpost-2011-1.csv	2011.01	32
fpost-2011-2.csv	2011.02	32
fpost-2011-3.csv	2011.03	19
fpost-2011-4.csv	2011.04	14
fpost-2011-5.csv	2011.05	30
fpost-2011-6.csv	2011.06	28
fpost-2011-7.csv	2011.07	30
fpost-2011-8.csv	2011.08	36
fpost-2011-9.csv	2011.09	59
fpost-2011-10.csv	2011.10	79
fpost-2011-11.csv	2011.11	63
fpost-2011-12.csv	2011.12	51
fpost-2012-1.csv	2012.01	91
fpost-2012-2.csv	2012.02	64
fpost-2012-3.csv	2012.03	76

Showing 1 to 16 of 60 entries, 2 total columns

Figure 4: Data Frame of Cauliflower

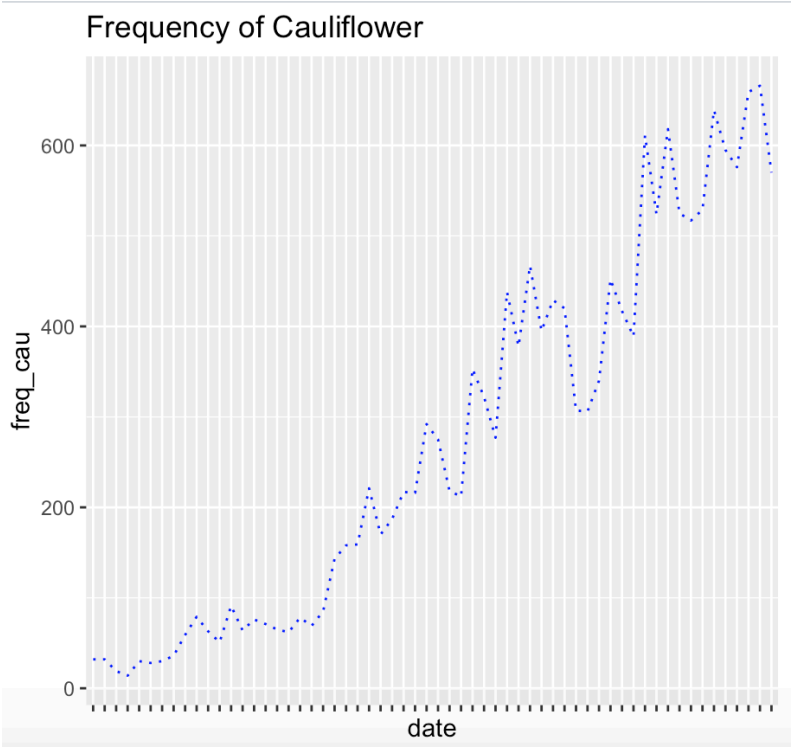

Figure 5: Frequency of Cauliflower

My main findings:

• 'cauliflower' generally has an upward trend from 2011 to 2015, which means that 'cauliflower' is more frequently being mentioned on social media.

• Frequency of the term 'cauliflower' is the most on 2015-11 as 666 times, and the least on 2011-4 as 14 times.

• Based on these observation, it concludes that people are gradually more paying attention on healthy food and sharing them on social media.

• Deeper analysis may be needed for further explorations.