Multi-Doc Summarization COVID-19

Sharon Wang

Week 5

CatE

Performance of CatE depends on:

1. **corpus context statistics**; if a category name only appears a few times in the corpus, there may not be good contexts for the category name
2. **input category names;** if a category name doesn't have other category names to be distinguished from, the results may be too general
    a. for category names China, Italy, United_States, and New_Zealand, the output for Italy is: ->
    b. solution: add other countries (e.g. Spain, France, Germany, ...) to the list of category names
    c. alternative solution: discard the other country names from the output

Italy
Spain
Lombardy
Europe
Italys
France
Germany
Milan
Venice
Veneto
Italy

Corpus Context Statistics

1. Count the number of **declarative sentences** containing
   a. the category name
   b. a **subject, predicate,** and **object** (using SRL)
2. Only run CatE on the category name if the number of declarative sentences > ~50.

Input Category Names

- To get more specific results for category names, use SpaCy and WordNet to find other words that are in the same classification as the category name.
- For example, for the category name “Italy,” use SpaCy to classify it as a country.

- Then use WordNet to generate a certain number of other words that are both countries and appear in the corpus.
- **Alternatively, can go through the corpus and add any words classified as a country to the category name list.**

- Then add these words as other category names and run CatE for a better result.

Results

- Applying these new rules, the new category names will be close to: **China, Italy, United_States, Spain, France, Germany, Iran, Indonesia, Switzerland, etc.**

- Notes:
- New_Zealand was removed because there were only 9 declarative sentences containing “New_Zealand”, and a subject, predicate, and object in the corpus.
- Spain, France, and Germany, etc were added because they are other countries and they appear in the corpus in quality sentences more than 50 times.

CatE Category Names (Countries)

```
Category (China):
Wuhan mainland Chinese Hubei Chinas Li human-to-human China  Shanghai Beijing
Category (Italy):
Tuscany happen in ITALYThe Christie Korea northern Italys Lombardy Stettler towns
Category (United_States):
recent_days 41000 53000 confirmed_coronavirus_cases surpassed topped surged exceeds total_infections neared
Category (Spain):
hard-hit devastation hardest-hit morgues Tamlin overwhelmed overwhelming hallways Maisonneuve-Rosemont hospitals
Category (France):
Bishops French Paris QESBA parents lessons English Kingdom courses  exams
Category (Germany):
Daniele Rome Italian Alberto Spain Milan Madrid Spains wheeled Ho
Category (Iran):
Europe Egypt  Italy Milan Korea Spain Asia Italys Lombardy Italian
Category (Indonesia):
Willy KurniawanReuters worlds Smithfield_Foods biggest giant halted second-largest tankers company said
Category (Switzerland):
Anutin Bordeaux Charnvirakul returned from reporters Speaking Dominic Networks helpful  spoke
Category (Ukraine):
Op Croizer recorders Unifier Defence authorization recorder Navys Capt helmet
Category (Greece):
military NATO troops peacekeeping mission soldiers Forces helicopter reservists fighter
Category (Turkey):
returned 27-year-old who travelled woman who returning travelled 14-day internationally flew recently
Category (Algeria):
confirmed in cruise  to travel BC — Two Three    March Princess cruise 50s who three
Category (Lithuania):
Oriental Huntington Palm Orpheum Finals Boutique Mapelli Mozzi Alpine Ski
Category (Slovakia):
shopping_malls showrooms shops close  restaurants retail movie theatres venues spas
Category (Denmark):
gradual_reopening public_transport phased-in restrictions to approach workplaces loosening gradual — such province — if
Category (Australia):
Jacinda Bahrain hardline Zealand tighten bother Medicare List Ardern favouring
Category (Egypt):
Iran  COVID-19 in whereabouts coronavirus in asked self-isolate Iran case  Ayatollah Shahab
Category (Sweden):
Asian_countries wallop squandered watchdogs critics stoked likelier stereotypes mid-January inclined
Category (South_Africa):
strategy  expressed_concern highlighting diversified economy creating wrought prop fast-growing broadest
Category (Slovenia):
Cyprus border controls barred traffic restrict crossing Canada-US crossings entry
Category (Austria):
bakeries supermarkets stores pharmacies grocery soups foods hardware sauce meats
Category (Argentina):
Tucumán Miroli Schaffner Américo immunologist pathology Network  associate_professor Alfredo Susy
Category (Ghana):
727 three new 405 its total also Tuesday On Wednesday Thursday afternoon
```

- The countries we are testing is **China, Italy,** and **United_States**. The other category names are included so they do not show up in the first 3 countries' generated phrases. (There are even more countries, but they couldn't fit onto the slide).
- This improved the output, but it's still not perfect.
- For example, "Korea" was generated for Italy, even though it was included as a category name.

- However, the error is much lower now so as a solution, results that are in the same classification as the category names could be deleted? I.e., delete "Korea" from Italy's phrases since Korea is a country.