Feedback on Project One

HG2051

Francis Bond

NTU

2021-10-28

Data

	Synsets No.	Score	Lemmas No.	Score
All	5,282	-0.025	6,017	-0.025
Non-Zero	2,189	-0.059	2,416	-0.063
Positive	1,000	+0.341	1,099	+0.354
Negative	1,189	-0.396	1,317	-0.411

Table: Sentiment Data

- Show how much data we have, as well the scores
- Only show a reasonable number of significant figures
- Weird that there were fewer lemmas than synsets
  as synsets can have multiple lemmas!
  turns out there is a bug in nltk.wordnet

!!! I will give numbers ignoring the bug
then fix it and give you new numbers

Synonyms

Synsets	Total	Score
All	1077	0.175
Non-zero	349	0.2626

- Higher than expected
- Some clearly weird values:
  Synset('paroxysm.n.01') paroxysm = -0.5, fit = 0.34, convulsion, -0.34
  Synset('free_will.n.01') free will = 0.34, discretion = -0.34
- Need to look at the corpus and/or wordnet

Synonyms I

- loop through each lemma, look for the synset, check if we have already done it

```
ss = l.synset()
    if ss.name() in known:
        continue
    else:
        known.add(ss.name())
```

- Look for the lemmas that have sentiment

```
### lemmas with sentiment
if nonzero:
    lems = [ll for ll in ss.lemmas() \
            if (ll in lsnt) and lsnt[ll]  !=0]
else:
    lems = [ll for ll in ss.lemmas() if ll in lsnt]
```

Synonyms II

- Find the difference for each pair

```
if len(lems) > 1:
   for l1 in lems:
       for l2 in lems:
           ## this makes sure we only compare once
           if l1 > l2:
               sdiff.append(abs(lsnt[l1]-lsnt[l2]))
   diffs.append(np.mean(sdiff))
   ### print pairs with a big difference
   if np.mean(sdiff) > .5:
       print(ss, np.mean(sdiff),
             [(x,  lsnt[x]) for x in lems])
```

Synset based relations

	All	Score	Non-Zero	Score
similar	331	+0.179	131	+0.219
hyponym	371	+0.145	72	+0.259
holo location	0	+nan	0	+nan
holo member	8	+0.003	0	+nan
holo part	87	+0.058	2	+0.350
holo portion	0	+nan	0	+nan
holo substance	2	+0.000	0	+nan
holonym	0	+nan	0	+nan
entails	23	+0.107	1	+0.283
causes	16	+0.184	4	+0.254

- None of them are very close
- Similar is most close, still some possible issues

Synset based relations discussion

- **white** “being of the achromatic color of maximum lightness; having little or no hue owing to reflection of almost all incident light” was given a negative score even though it appears to be neutral in meaning. It is possible that the sense **white** “anemic looking from illness or emotion” should have been the correct tag.[1]
- Similarly for **serious** it is likely that “of great consequence” and “causing fear or anxiety by threatening great harm” were confused.
- Sometimes the error may be in the structure of wordnet itself. For example **proud** only has a single meaning, even though many lexicons distinguish more: e.g. from wiktionary “Feeling honoured (by something); feeling happy or satisfied about an event or fact; gratified” vs “Having too high an opinion of oneself; arrogant, supercilious ”.

[1] It was *His dark eyes, glaring out of the white mask of his face, were full of horror and astonishment as he gazed from Sir Henry to me.*

Lemma based relations

	All	Score	Non-Zero	Score
antonym	125	+0.288	40	+0.655
antonym opposite	125	+0.132	40	+0.167
derivation	530	+0.140	185	+0.184
also	16	+0.117	2	+0.258
pertainym	3	+0.069	0	+nan

- Derivation is pretty close
- Antonym is large, as we can expect
  calculate the sum (difference when one is reversed)
  then the difference is very small

```
 diff = (abs(lsnt[s1] + lsnt[s2]))
```

Analysis

- Checking these gives good feedback on both the corpus annotation and the structure of wordnet
  it should be done regularly
- To extend to un-annotated synsets
  - Automatically do if there are compatible scores from multiple relations (e.g. antonym and derivation)
  - Automatically do if there are compatible scores from other resources
  - Otherwise suggest for tagging only, ...