Annotate 'Translation' Pairs

The following describes an annotation scheme for so-called 'translation pairs' – pairs of words, both from vocabularies of English at different points in time.

Where do the word pairs come from?

The Royal Society Corpus (RSC) contains English scientific texts from the 1660s to the 1920s. These texts were divided decade-wise and for each of the 27 decades a Word2Vec model was trained. These models – or spaces – can be connected with a method called "Gromov-Wasserstein Optimal Transport" which takes two spaces and matches up points which have similar layouts within their space.

The result is a *coupling*: a translation table which lists for each combination of two words (one from one space, one from the other) the likelihood that these two words are translations of each other (i.e., that they express the same concept).

Why is this needed?

As language changes over time, we can't be sure that all of these spaces use the same words to express the same concepts (that's why we see the coupling of spaces as a bilingual task and call these pairs 'translation' pairs).

Of course, a solid coupling connects many pairs of concepts which are expressed with the same word; this is the easiest case of a true positive because it can be detected by simple string matching. However, there are always borderline cases in which a translation pair doesn't string-match. Some of these cases are pure noise (e.g., pairs like "x" and ";"), others are actually true positives, like "connexion" and "connection".

We need information about the types of mismatches in these pairs in order to evaluate how well the coupling really works.

The Task

Files

We have coupled several pairs of spaces and obtained multiple lists of translation pairs, two of which are up for annotation: `annotations_1.txt` and `annotations_2.txt`. They each contain 1500 pairs, one pair per line, preceeded by brackets and followed by a number which is irrelevant for this task. Here is an example (see below for more):

```
[]   word1     word2   0.123
```

Labels

Each label is one character long and case insensitive.

1. `O` — orthographic or morphological difference: these mismatches are purely based on how a word is written, or because the differing part carries grammatical meaning
2. `S` — semantically similar: the two words could (loosely spoken) be synonyms of each other; it's easy to think of one sentence where both make sense (or can even be used interchangeably)
3. `R` — semantically related: the two words come from the same word field, but they have clearly different meanings
4. `A` — antonymy: the two words are opposites of each other
5. `N` — noise: these pairs don't make sense in any way
6. `X` — wildcard: no straightforward decision possible; maybe the writers of the annotation scheme missed something

How to annotate

- into the brackets, insert the label that best describes the relationship between the two words of a pair
- only use one label per line
- when interpreting the words, take into account that we're dealing with language from scientific texts.
- try to use external lexical resources (Leo, Linguee, dict.cc)
- annotating one pair should take about 30 seconds *on average*; don't spend much more than 1 minute on the difficult cases!
- be conservative: when in doubt, use `R` (related) rather than `S` (similar) or `A` (antonymy)
- don't be afraid to label something with `N` (noise) or `X` (wildcard)

Examples

`O` — orthographic and morphological differences

```
 [O]   knowlege            knowledge           0.333
[O]   connexion           connection          0.319
[O]   fallen              falls               0.223
[O]   signs               sign                0.221
```

`S` — semantically similar

```
[S]   producing          forming          0.296
[S]   procure            obtain           0.252
[S]   assured            sure             0.354
[S]   article            essay            0.247
[S]   retained           absorbed         0.233
[S]   phaenician         punic            0.191
```

`R` — semantically related

```
[R]   artery             vein             0.347
[R]   daily              annually         0.315
[R]   confirmed          convinced        0.264
[R]   steam              moisture         0.254
[R]   medicines          hemlock          0.260
[R]   stony              calcareous       0.307
[R]   mars               venus            0.238
[R]   eat                food             0.237
[R]   cells              pores            0.206
[R]   tide               noon             0.307
[R]   lime-water         vinegar          0.297
[R]   judgment           determination    0.248
[R]   argument           demonstration    0.210
[R]   chapter            memoir           0.192
[R]   reasoning          rules            0.189
[R]   lake               bay              0.175
[R]   polypes            suckers          0.180
[R]   fruit              seed             0.184
[R]   earthquakes        eclipses         0.183
[R]   carlsbad           florence         0.183
[R]   flow               changes          0.179
```

`A` — antonymy

```
[A]   inferior           superior         0.351
[A]   floor              roof             0.218
```

`N` — noise

```
[N]   w                  58               0.307
[N]   whereof            em               0.309
[N]   minus              rectangle        0.302
[N]   1753               1769             0.250
[N]   moves              6~               0.255
[N]   hist.              par.             0.330
[N]   ancients           l'abbe           0.425
[N]   pliny              sig.             0.383
[N]   allow              24~              0.227
[N]   henry              charles          0.226
[N]   american           paul             0.220
[N]   bb                 hh               0.216
[N]   painting           jupiter          0.214
[N]   xx                 xi               0.208
[N]   a2                 ~2               0.174
[N]   21/                03               0.174
[N]   modern             astronomers      0.174
```

`X` — wildcard

```
[X]   palmyrene          phaenician       0.238
```