Assignment 2: Lexical semantics (graded part)

General information

- The graded part of Assignment 2 will earn you **30%** of your overall course grade. You will get a grade between 0 and 10.
- **To be done in pairs** – the sames that you chose for Exercise 1 a few weeks ago.
- Please submit your solutions on **Tuesday November 23 at 23:59** at the latest (Barcelona time). Submit it as a single compressed file, containing a single or multiple .py files for the code as well as 2 PDF files: one with your written answers (Exercises 2-4), another one with your slides for Exercise 5.
    - If you used any datasets/resources, please remember to include those files too; but your compressed file cannot be larger than 500MB, so if you use something heavier please put it somewhere on the web and use it in the code via a URL call.

Exercise 1. Deadline: **Wednesday November 3rd at 14:00**.

- Choose a language that you are reasonably proficient in and has a WordNet in NLTK. Choose a word that you like, for which the language's WordNet has at least 3 synsets for a given part of speech of the word. For instance: in Catalan, the verb *estudiar* (*to study*).
- Get 100 random samples of the word in context (around 50 words, or 1 paragraph) from a corpus of choice. Note: the corpus should be lemmatized and part-of-speech tagged. You can select it from NLTK or get it from an external source (for instance, the Universal Dependencies treebank, https://universaldependencies.org).
- Each member of the pair **independently** annotates each sample with a sense (synset) from WordNet.
- Collate the results in a CSV file, with columns *name1*, *name2, sample*. For instance (invented data):

1) choose word 'estudiar' ('to study', in Catalan)
2) define that **s1** (sense 1) is *analyze.v.01;* **s2** *is study.v.03; etc.*
2) annotate data:

    gemma, lucas, sample
    s1, s3, "blah estudiar blah blah..."
    s2, s3, "blih blih estudies blih..."
    s1, s1, "bloh bluh bleh estudiaré..."
    …

- Adapt the code we provide in file 'exercise1_assignment2_updated.py', that prints: the lemma of the word (and its translation into English); the number of synsets; for each

synset, the synset name, its definition, and lemmas; and the contingency table of the annotations by each of you.

- Copy the resulting data to this document: https://docs.google.com/document/d/18HSH0TVJPAZj0qhtmNmy3v8x6TlAA5HbUtzQ4T7dK9I/edit?usp=sharing
- Upload your code and CSV file onto Aula Global, in a compressed file.

Exercise 2.

Analyze your sense-annotated data from Exercise 1: quantitatively, reporting percentage agreement, and qualitatively, discussing the sources of disagreement. Make sure to use **examples** in your discussion. Note: percentage agreement is computed like accuracy (to think: why is it called "agreement" if it's computed identically?). You can report it as a proportion (between 0 and 1), or as a percentage (between 0 and 100).

Exercise 3

State your hypotheses about the following questions, and find out whether the data supports your hypotheses, using Word2Vec and WordNet (use at least 500 data points):

- Which word pairs appear in more similar contexts, synonyms, or antonyms?
- Which word pairs appear in more similar contexts, hypernyms, or hyponyms?

Report quantitative results and a graph visualizing the results. Reflect on your findings (around 50 words).

Exercise 4

Write two mini-essays about how natural language meaning should be captured (100-150 words each): one arguing against the sense enumeration approach (à la WordNet) and in favor of distributional methods, another arguing the opposite, that is, against distributional methods and in favor of sense enumeration. For the second essay, it may be useful to check sections 1-2 of the following article:

> Jose Camacho-Collados and Mohammad Taher Pilehvar (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research*. https://www.jair.org/index.php/jair/article/view/11259/26454

Exercise 5

Carry out an intrinsic semantic evaluation/analysis of word embeddings, as follows.

- Report the results in the form of a **slide presentation** (in PDF format), with 5 slides: 1) title slide, 2) task and dataset (with examples), 3) method, 4) quantitative results (+ visualization, if applicable), 5) results of error analysis (with examples). If you need one more slide for the error analysis, you can add it (no other exception to the 5-slide instruction).
- You will do a brief presentation of the results of this exercise in Lecture 9.

- Unless otherwise stated in the exercises, you can pick embeddings of your choice (provided the ones you choose give good data coverage). A useful resource if FastText, in particular for languages other than English: https://fasttext.cc/.

1. **Word similarity/relatedness**: based on the correlation between cosines for word pairs in semantic space and word pair similarity/relatedness scores given by humans.

- Davide Locatelli (apple) & Ilse Meijer (apple): evaluate textual vs. multimodal embeddings on the MEN dataset (https://staff.fnwi.uva.nl/e.bruni/MEN). Compare the static word embeddings of BERT and VilBERT – which type of embedding correlates the most with human similarity judgments? For the error analysis, Q: For which kind of pairs does VilBERT approximate human judgments better than BERT, and vice versa?
- Eva Richter (kiwi) & Nora Graichen (apple): evaluate one or more distributional model of English of choice using the WordSim353 dataset, in the version that distinguishes between similarity and relatedness (files *wordsim_relatedness_goldstandard.txt* and *wordsim_similarity_goldstandard.txt* from http://alfonseca.org/eng/research/wordsim353.html). In your error analysis, make sure to address the following Qs: Which notion do word embeddings model better, similarity, or relatedness? For which kinds of semantic phenomena are human vs. model similarities the most dissimilar? (Both for similarity and relatedness subsets.)

2. **Other tasks for multi-modal models**: reproduce the experiments in the following paper, using textual and multimodal embeddings: Bruni, E., G. Boleda, M. Baroni, N. K. Tran (2012), Distributional semantics in technicolor. *Proceedings of ACL 2012*, pp. 136-145, Jeju Island, Korea. (**paper**, slides, data, bib). More concretely, compare the static word embeddings of BERT and VilBERT – which type of embedding has better performance, and why? For the error analysis, Q: Which kinds of data points does VilBERT do better than BERT, and vice versa? Hypotheses as to why?

- Audrey Mash (orange) & Aakash Maroti (apple): reproduce Experiment 1 (color association). You can evaluate the models either with accuracy or with the measure proposed in the paper.
- Pol Garriga (apple) & Zixuan Liu (orange): reproduce Experiment 2 (literal vs. non-literal uses of color terms).

3. **Analogies**: solve them with the parallelogram model (see J&M). You will find the Google Analogies dataset developed by Mikolov and colleagues here: http://download.tensorflow.org/data/questions-words.txt, and the Bigger Analogy Test Set (BATS) dataset here: https://vecto.space/data/.

- Florencia Van Rysselberghe (kiwi) & David Arnau (apple): from BATS: i) dataset of things-color analogies (*BATS → Encyclopedic Semantics → things-color*), ii) dataset of country-language analogies. Use English embeddings. Obtain word pairs by taking the word in column 1 and the first word listed in column 2 (ignore the rest). In your error analysis, make sure to address the following Q: Which kind of analogies are solved better? Hypotheses as to why?
- Kosuke Nishio (kiwi) & Manisha Venkat (orange): from BATS, dataset of hyponym analogies and dataset of hypernym analogies (*BATS → Lexicographic Semantics → hypernyms-misc/hyponyms-misc*). Use English embeddings. Obtain word pairs by taking the word in column 1 and the first word listed in column 2 (ignore the rest). In your error analysis, make sure to address the following Q: Which kind of relationship is modeled better?

- Alireza Sayah (orange) & Claudia Avila Cueva (kiwi): from BATS, dataset of animal-sound, animal-young, animal-shelter (*BATS* → *Encyclopedic Semantics* → ...). Use English embeddings. Obtain word pairs by taking the word in column 1 and the first word listed in column 2 (ignore the rest). In your error analysis, make sure to address the following Q: Which kind of relationship is modeled better?
- Ting Yao (orange) & Chenyue Zhou (orange): comparison of Chinese vs. English embeddings in solving the analogy task (use only the 'family' subset). For English, use the Google Analogies dataset. See https://github.com/Leonard-Xu/CWE/blob/master/data/analogy.txt for a dataset of Chinese analogies. Do a language-specific and comparative error analysis: Which embeddings perform better? Hypotheses as to why? What kind of errors do you find in the Chinese embeddings? And in the English embeddings?
- Sophia Harrison (orange) & Marina Bolea (orange): comparison of English vs. Spanish embeddings in solving the analogy task (use only the 'family' part of the datasets). For English, use the Google Analogies dataset. For Spanish, use the dataset by Cardellino (2016), Spanish Billion Words Corpus and Embeddings. Do a language-specific and comparative error analysis: Which embeddings perform better? Hypotheses as to why? What kind of errors do you find in the English embeddings? And in the Spanish embeddings?
- Alba Buendía (orange) & Nicolás Rivera (orange): use English embeddings to solve the analogy task (Google Analogies dataset; use only the 'family' and 'capital-common-countries' subsets). Compare the performance on the two subsets: Which subset is better solved? What kind of errors do you find in each subset?
- Inés Gabanes Anuncibay (orange) & Mar Domínguez Orfila (orange): Use Spanish embeddings to solve the analogy task. Use the dataset of Spanish analogies by Cardellino (2016), Spanish Billion Words Corpus and Embeddings (only 'family' vs. 'capital-common-countries' subsets). Compare the performance on the two subsets: Which subset is better solved? What kind of errors do you find in each subset?
- Matthew Galbraith (orange) & Adelle Hornanska (kiwi): Use the 'city-in-state' subset of the Google Analogies, with English embeddings. What kind of errors do you find? (What is easy/difficult for the model?) Any hypotheses as to why?
- Dalila & Mamen (kiwis): Use Portuguese embeddings to solve the analogy task; more concretely, the 'capital-common-countries dataset' subset of the dataset of analogies in Portuguese available here: https://github.com/nathanshartmann/portuguese_word_embeddings/blob/master/analogies/testset/LX-4WAnalogies.txt. What kind of errors do you find? (What is easy/difficult for the model?) Any hypotheses as to why?