Master AIC
Named Entity Recognition assignement
Lab sessions #2 and #3

January 2021

1 Introduction

General instructions and material are available at `https://saharghannay.github.io/courses/cours1/example2/`.

- NER engine: `https://github.com/XuezheMax/NeuroNLP2`
- NER launch script (to be adapted): `https://perso.limsi.fr/pz/upsay/run-ner-q6-20-half.sh`
- Pre-trained word embeddings (from previous lab session): `https://perso.limsi.fr/pz/upsay/QUAERO_FrenchPress-w2v.vec.gz`
- NER config file: `https://perso.limsi.fr/pz/upsay/quaero-100-demi.json`

2 Questions

1. Create descriptive statistics on the corpus: number of words (tokens), sentences, entities of each type.
2. Compare different named entity recognition models that use word embeddings created during Lab session #1.

   The NER system uses word embeddings trained on a corpus (raw text corpus: Quaero broadcast news or Quaero medical) according to a given algorithm (word2vec skip-gram, fastText skip-gram); it is itself trained on an annotated corpus (Quaero broadcast news or Quaero medical) whose texts belong to a given domain (general news, medical). Study the variation in precision, recall, and F-score when the following parameters vary (while keeping the others fixed):

   (a) the size of the training set:
   - use training set subsets of increasing sizes;

   (b) the size of the corpus used to train word embeddings:
   - use raw text samples of increasing sizes to train word embeddings;

   (c) the word embedding model:
   - use word embeddings created with word2vec skip-gram, fastText skip-gram;

   (d) the fit of the domain of the corpus used for named entity recognition and that of the word embeddings:
   - use word embeddings trained on a news corpus vs. a medical corpus of the same size.

   You are welcome to use hyperparameters that reduce training time, such as –num_epochs 10.
3. Please comment on the comparability of the results you obtain on the medical corpus with those of the CLEF eHealth campaign participants.

3 Submission

Please submit a written report addressing these questions as well as the code used to run the experiments to `pz@limsi.fr` and `neveol@limsi.fr` by February 12, 2021.