
ReadMe File

An entity name normalization method for biomedical articles:
application to diseases and plants

Hyejin Cho, Wonjun Choi, Hyunju Lee

Data Mining & Computational Biology Lab
Gwangju Institute of Science & Technology
Gwangju, South Korea
http://combio.gist.ac.kr/

List of files: “disease_pretrained_word_vectors.zip”

1. **“SSL-modified evidences_5_300_skip.bin”**

2. **“SL-only training data_5_300_skip.bin”**

3. **“SSL-only modified evidences_5_300_skip.bin”**

List of files: “plant_pretrained_word_vectors.zip”

1. **“SSL-modified plant evidences_7_200_cbow.bin”**

2. **“SL-only plant training data_7_200_cbow.bin”**

3. **“SSL-only plant modified evidences_7_200_cbow.bin”**

Description

We are publishing pre-trained word vectors for disease and plant names in the biomedical article, trained on selected PubMed abstracts. These vectors of disease in dimension 300 were obtained using the skip-gram model and vectors of plant in dimension 200 were obtained using the continuous bag of words model. The details are described in the following.

For disease name normalization, we constructed three pre-trained word vector models: (1) semi-supervised learning with unlabeled data of "modified evidence sentences" ("SSL-modified evidences") with 740,353 word vectors, (2) semi-supervised model that used only modified evidence sentences without the training corpus ("SSL-only modified evidences") with 714,575 word vectors, and (3) a supervised learning model with the 592 NCBI disease training corpus ("SL-only training data") with 51,619 word vectors.

To find the optimal hyper-parameters to learn word vectors, we applied different hyper-parameters to the "SSL-modified evidences" model. When the NCBI disease development set was used to select hyper-parameters as window size = 5 and vector dimension = 300, and a skip-gram method was selected.

Next, for plant name normalization, we constructed three plant models: (1) "SL-only plant training data" with 94,338 word vectors, (2) "SSL-only modified plant evidences" with 594,802 word vectors, and (3) "SSL-modified plant evidences" with 649,759 word vectors. For plant evidence sentences, we collected 2,620,684 sentences containing plant names in the NCBI taxonomy database from PubMed abstracts. Note that because "SSL-modified evidences" showed the best performance for disease name normalization, we tested "SSL-modified plant evidences" among the several SSL models.

For selecting proper hyper-parameters, we constructed the "SSL-modified plant evidences" model by applying different hyper-parameters to the plant development set. We selected the hyper-parameters as window size = 7 and vector dimension = 200, and we used the CBOW method.