Training Data Generation

Let's generate training data for the entity linking problem.

We'll cover the following

There are two approaches you can adopt to gather training data for the entity linking problem.

1. Open-source datasets
2. Manual labeling

You can use one or both depending on the particular task for which we have to perform entity linking.

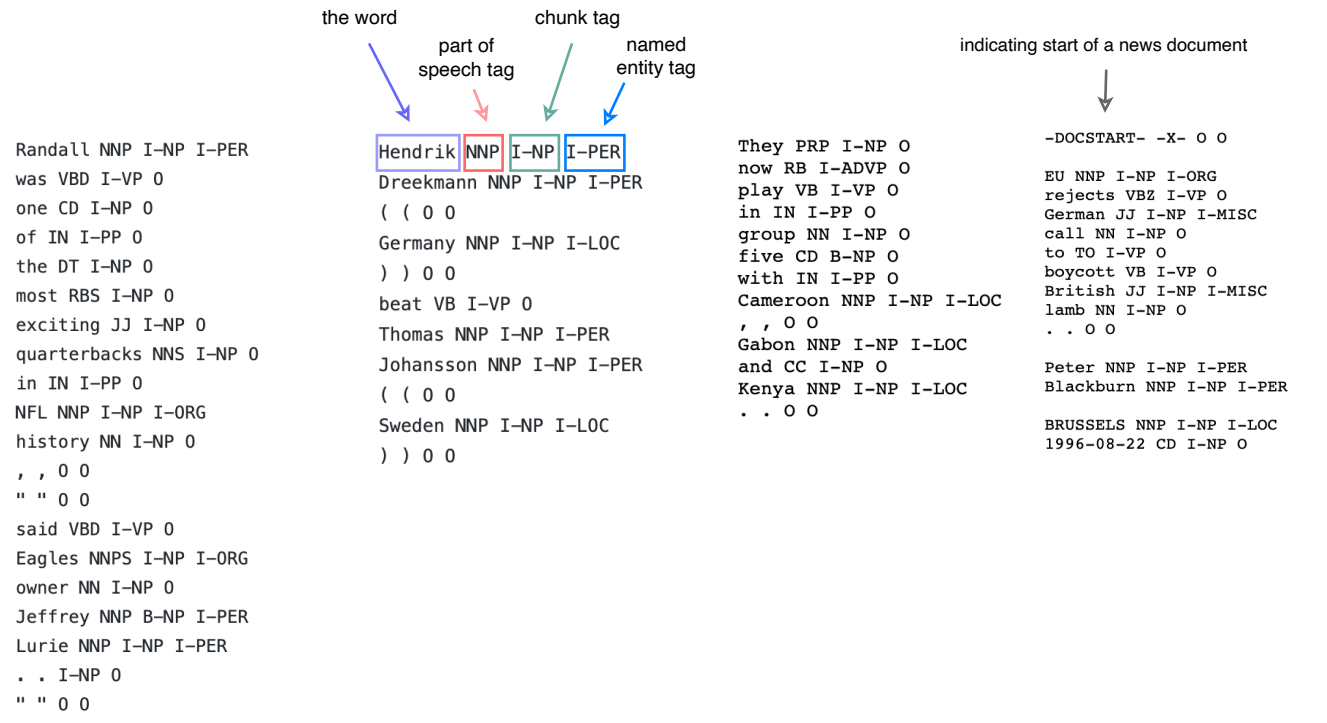
Open-source datasets#

If the task is not extremely domain-specific and does not require very specific tags, you can avail open-source datasets as training data. For example, if you were asked to perform entity linking for a simple chatbot, you could utilize the general-purpose, open-source dataset CoNLL-2003 for *named-entity recognition*.

CoNLL-2003 is built on the Reuters Corpus which contains 10,788 news documents totalling 1.3 million words. It contains train and test files for English and German languages and follows the IOB tagging scheme.

> IOB tagging scheme
> I - An inner token of a multi-token entity
> O - A non-entity token
> B - The first token of a multi-token entity; The B-tag is used only when a tag is followed by a tag of the same type without “O” tokens between them. For example, if for some reason the text has two consecutive locations (type LOC) that are not separated by a non-entity

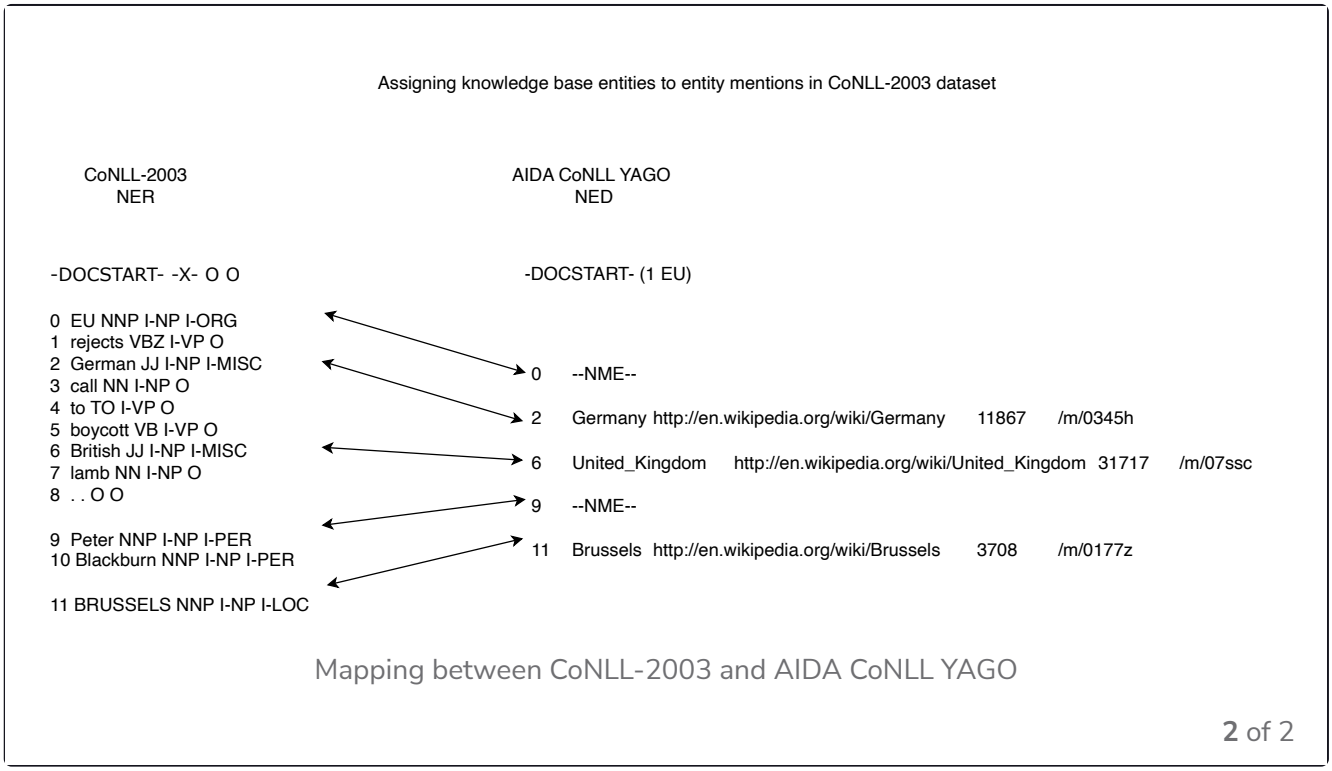
The following are some snippets from the train and test files of CoNLL dataset.



CoNLL-2003 for named entity recognition

For *named-entity disambiguation*, you can utilize the AIDA CoNLL-YAGO Dataset, which contains assignments of entities to the mentions of named entities annotated for the CoNLL-2003 dataset. The entity mentions are

shown in the slide below.



Mapping between CoNLL-2003 and AIDA CoNLL YAGO

2 of 2

Human-labeled data#

Once you have utilized the open-source datasets, we may want to enhance the data and increase its size through manual labelers. The manual labelers will generate training data similar to the open-source datasets, by annotating named entities in text and linking them to corresponding entities in the knowledge base.

Another case where you would generate data through manual labelers is when you require a highly specialized dataset for a specific problem. For example, assume that the problem is related to the medical field; this requires identifying certain domain-specific entities. In such situations, you

requires identifying certain domain-specific entities. In such situations, you need to understand the domain in which you want to perfor it ir

What are the kind of entities you want to recognize and link? When the manual labelers are given hospital data, they will mark doctor names, symptoms, diseases, patient names, types of surgeries, and so on. Hence, you would have tags that are related to the domain of the task.

After labeling the entities the labelers will also link them to the entities in the knowledge base (database) that is being used.

Back

Architectural Components

Next

Modeling

Mark as Completed

Report an Issue