

# Improving NER Research Workflows with SeqScore



Brandeis

Constantine Lignos, Maya Kruse,  
and Andrew Rueda

```
pip install seqscore
```

**SeqScore** is a robust, one-stop command line tool for working with NER data, tested on dozens of modern datasets.

## Why? Isn't NER data simple?

**NER data is messy and not standardized.** Many publicly released datasets have errors that can affect evaluation. Most one-off scripts fail when run on multiple datasets.

## Finding and Fixing Data Problems

The **validate** command identifies invalid label transitions (O to I in BIO), and the **repair** command provides options for fixing them. Most datasets have these errors, and they cause trouble!

## Summarizing Data

Want to know what's in an NER dataset? The **summarize** command gives the size of each split (train/dev/test) and counts for all the entity types it contains. The **count** command exports all entity mentions in a tabular format for easy review.

## Training and Evaluating Models

The **process** command maps entity types (fine to coarse) or removes them (who wants MISC anyway?). The **convert** command changes between label encodings (BIO to BIOES, etc.). Finally, **score** provides scores and error analysis tables.

<https://github.com/bltlab/seqscore>