Potential paper:

Entity identification of scientific papers/content

A machine-learning approach:

Goal: Given a long scientific text (representing paper content), identify important entities/features as facts

Option1: n-grams. Drawback: irrelevant phrases e.g. data not shown

Option2: machine-learning approach

an entity is represented by a word sequence say s1 s2 s3 where si is one word

Features:

1. POSTag Pattern: e.g. JJ NN, NN/ JJ JJ NN

2. freq(s1)

3. freq(s2)

4. freq(s3)

5. freq(s1 s2 s3)

6. isAcronym