There is no simple right answer to this question, however we can discuss a little bit about it.

In conlleval.pl the evaluation was based on correctly marked entities, not tokens and in python script it's basically grounded on tokens(macro, micro and etc.). in this method(conlleval) an entity is correctly marked with both starting and ending type in document. But on the other method, we just care about individual tokens and not entities so it seems that conlleval is more closer to our goal for NER. However there is some problem in this method as well like we tagged a word which is belong to PERSON, wrongly to LOC(location) we will get a false negative for PERSON and a false positive for LOC, but this problem is still on the another evaluation metric as well so they are tie in this.

As conlleval also return F1 and accuracy to us so it will be useful to see those results as well beside other measurements.

Seems that F1 is better way for measurement rather than accuracy because it can gives us a high accuracy even when there is no document returned. But F1 still has some problem about handling labeling and boundary error. Our system might notice about an entity but give it a wrong label like below:

(I/O/O *live*/O/O *in*/O/O *Palo*/LOC/ORG *Alto*/LOC/ORG ./O/O)

Or it might give it wrong boundaries like below:

Unless/O/PERS *Karl*/PERS/PERS *Smith*/PERS/PERS *resigns*/O/O)

Or doing both mistake at the same time which those errors will be labeling and boundary errors. when it comes to counting true positive and true negatives will be okay. But for precision and recall since we have a mistake like those above would cost 2 times for us. So F1 somehow is punishing mistake more severely other than the way it should be, while conlleval by considering and removing those will cost less to us. So all in all I think conlleval would be better choice for this task.