CoNLL#: Fine-grained Error Analysis and a Corrected Test Set for CoNLL-03 English

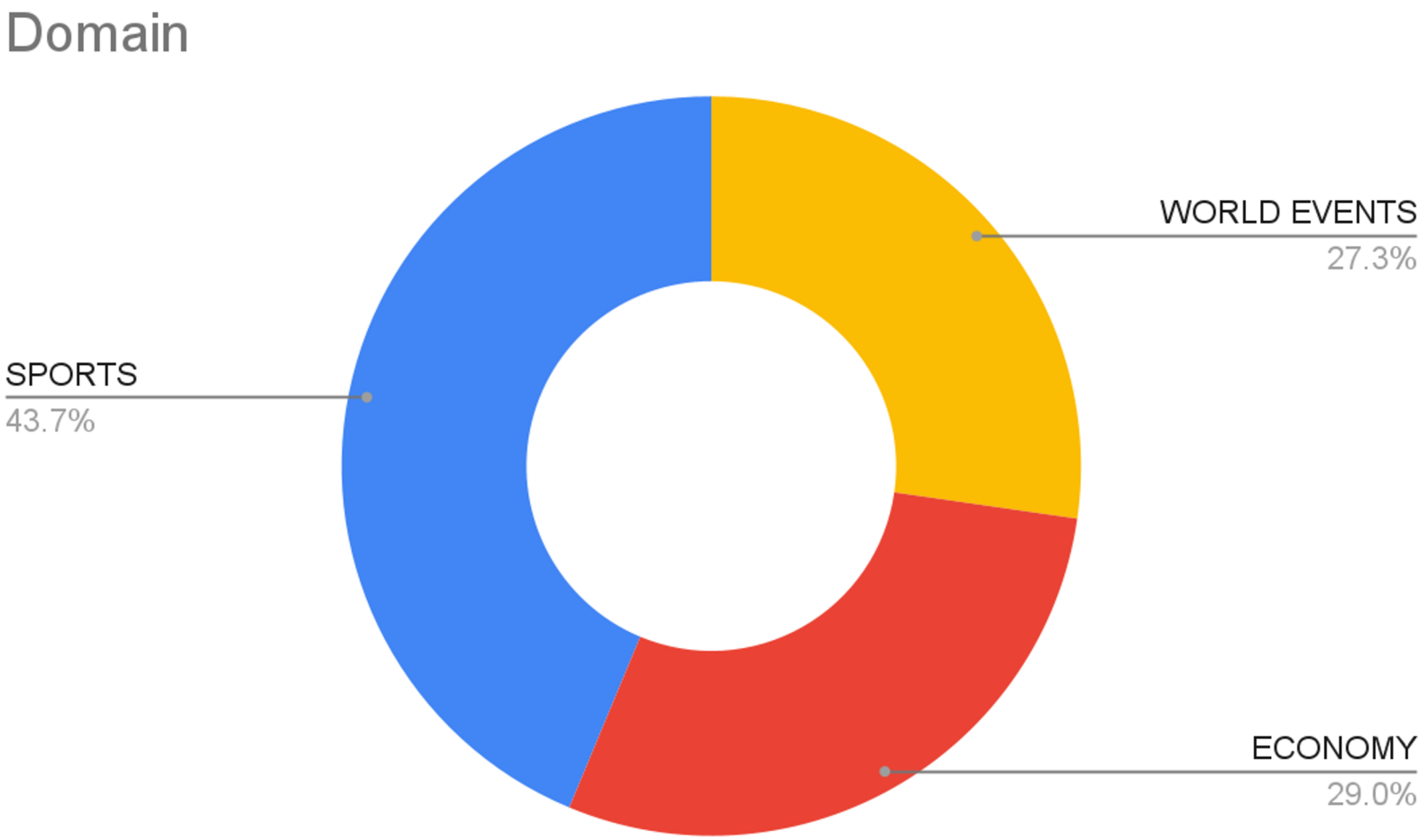
Andrew Rueda, Elena Álvarez Mellado, **Constantine Lignos**

Brandeis
UNIVERSITY

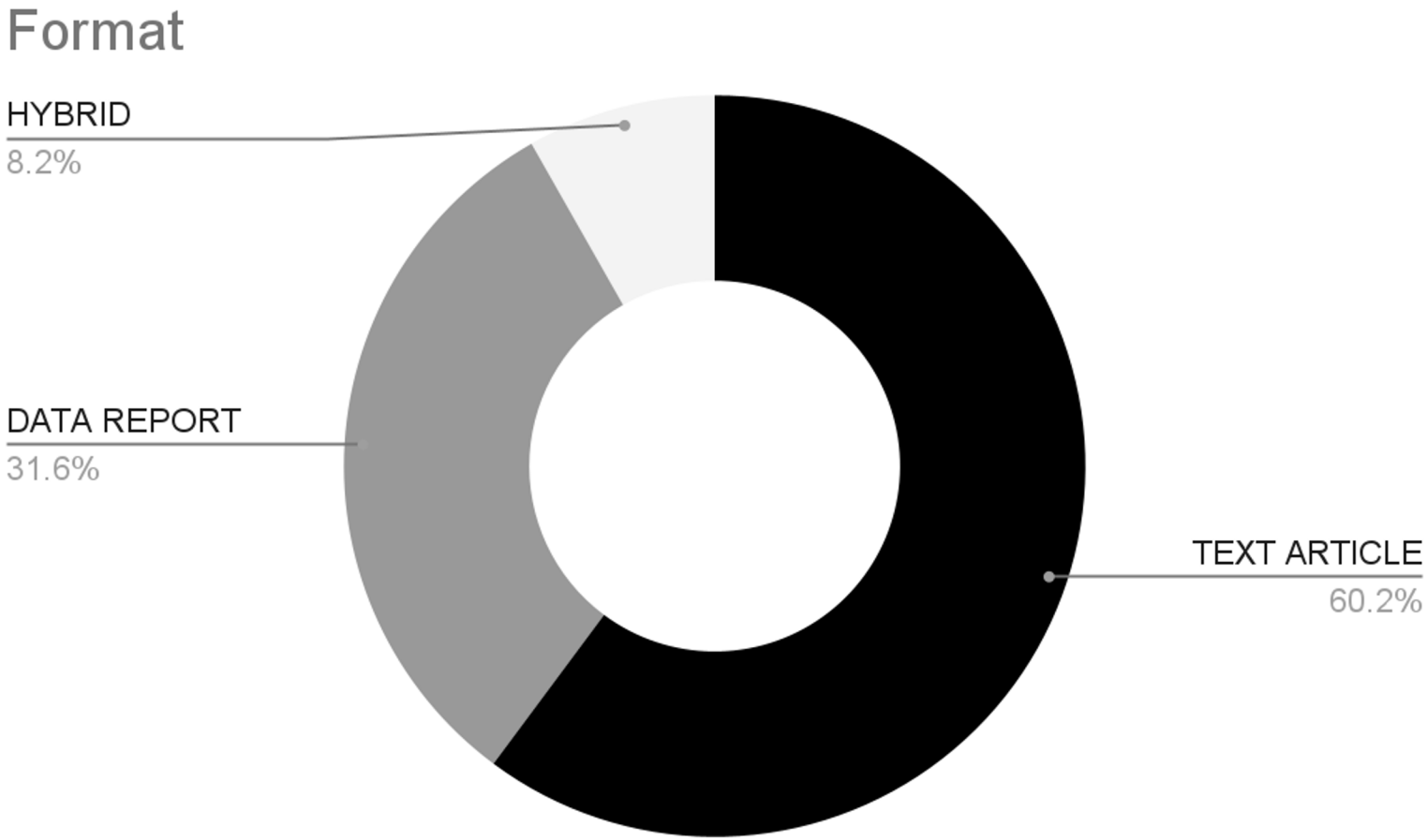
One-slide summary

- Performance on the CoNLL-03 English dataset seems to have plateaued
- We investigated what the state-of-the-art models were getting wrong
- Like many others, concluded we need to make corrections to the test set, which we release as CoNLL#
- We also annotated all documents in the test set for the domain (sports, etc.) and text format (text, tables, etc.)
- Key findings:
  - Original test set and other corrected versions contain many sentence/token boundary errors that affect performance
  - Economy-related documents (not sports) show the lowest performance
  - Our corrections result in an increase of about 2-3 F1 points on test
- Paper, data release, and more information here: https://github.com/bltlab/conll-sharp

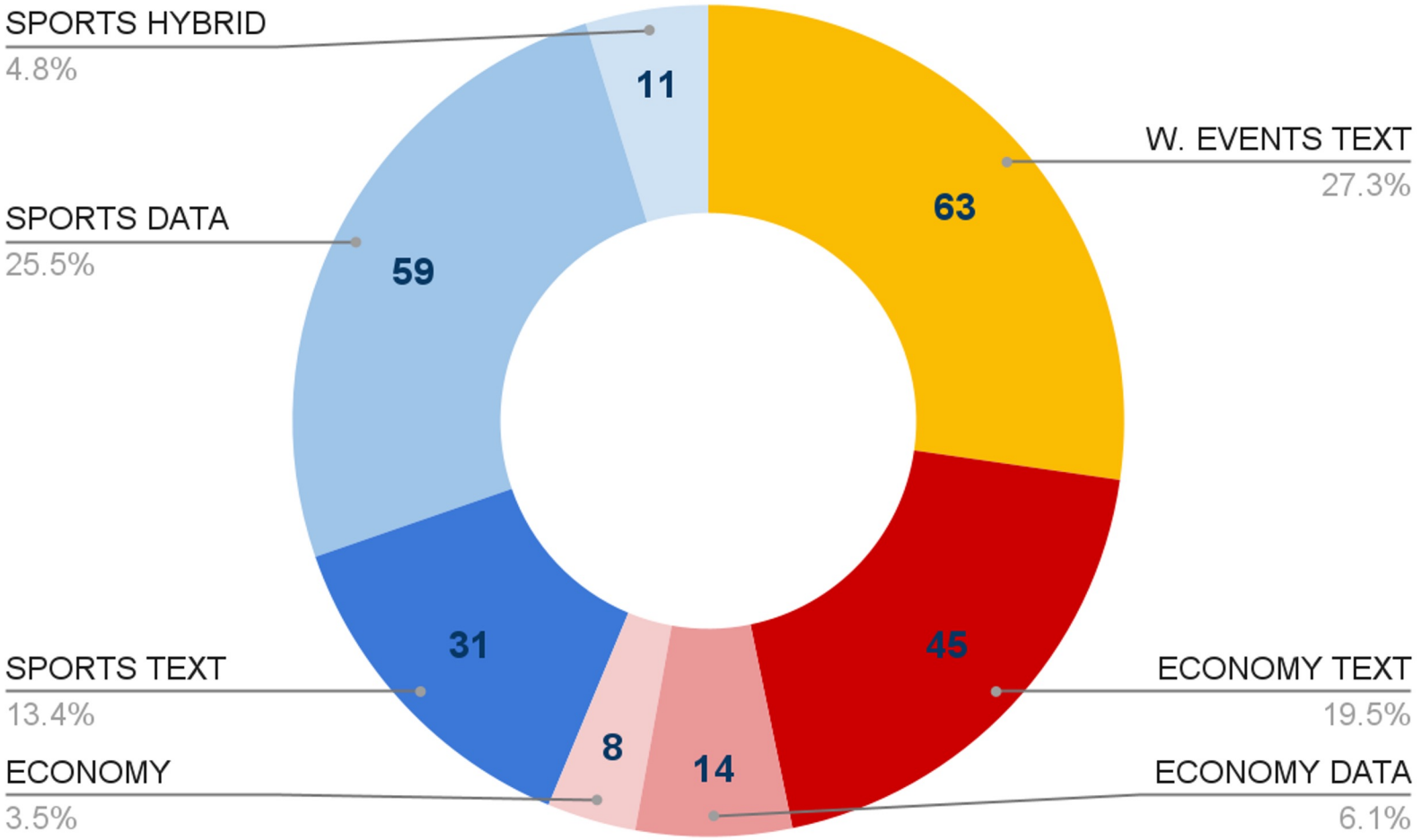
What's in the test set?



What's in the test set?



What's in the test set?



What did we change in the test set?

Error fix	Count	Example
Token splits	5	*JosepGuardiola* → *Josep Guardiola*
Bad hyphen fixes	27	*SKIING-WORLD CUP* → *SKIING - WORLD CUP*
Sentence boundary fixes	63	[*Results of National Basketball*] [*Association games on Friday*]
Label fixes	457	*Tasmania* `LOC` → *Tasmania* `ORG`

What didn't we change?

- Left training data untouched
- We did not change the annotation guidelines or simplify the task
- We left the intentional design choices of CoNLL-03 English, but corrected the test set to be more consistent with the annotation guidelines and the intended tokenization/sentence splits

Test set performance

Model	CoNLL-03	CoNLL#
XLM-R FLERT	93.64	95.98
LUKE	**94.44**	**97.10**
ASP-T0-3B	93.88	96.50

2-3 points of F1 increase due to test set corrections!

Conclusion

- We provide a new corrected CoNLL-03 English test set and analysis of performance based on document domains and types
- Our corrections lead to 2-3 points of F1 increase **without simplifying the annotation task**
- Paper, data release, and more information here: https://github.com/bltlab/conll-sharp