PySBD: Pragmatic Sentence Boundary Disambiguation

Nipun Sadvilkar, Mark Neumann
nipun.sadvilkar@episource.com, markn@allenai.org

EPISOURCE

Allen Institute for AI

Introduction

PySBD - a rule-based sentence boundary disambiguation Python package that works out-of-the-box for 22 languages.

Features:

- Domain agnostic rules
- Non Destructive Segmentation
- Multilingual Support
- Robust with 98% test coverage

Processing Stages

"My name is Jonas E. Smith. Please turn to p. 55."

↓ preprocess

"My name is Jonas E\~\~ Smith. Please turn to p\~\~ 55."

↓ segment

["My name is Jonas E\~\~ Smith.", "Please turn to p\~\~ 55."]

↓ postprocess

["My name is Jonas E. Smith.", "Please turn to p. 55."]

Data

1. **English Golden Rules Set (GRS)** -
   48 hand crafted rules by considering various domains.
2. **GENIA Corpus:**
   Linguistically annotated biomedical papers
   Speed Benchmark
3. **OPUS-100 Multilingual Parallel Data**

Results & Comparison to alternatives

Tool	GRS	GENIA
blingfire	75.00	86.95
syntok	68.75	80.90
spaCy	52.08	76.80
spacy dep	54.17	39.20
stanza	72.92	63.40
NLTK	56.25	87.95
PySBD	97.92	97.00

Accuracy (%) of PySBD compared to other open source SBD packages

Tool	Speed(ms)
blingfire	85.24
syntok	1764.11
spaCy	1523.20
spacy dep	26850.69
stanza	48383.46
NLTK	780.49
PySBD	9483.96

Speed benchmark on the entire text of 'The adventures of Sherlock Holmes'

Language	Accuracy (%)
Amharic	80.95%
Arabic	70.40%
Armenian	63.75%
Bulgarian	93.35%
Burmese	48.05%
Chinese	85.35%
Danish	91.40%
Deutsch	80.95%
Dutch	91.40%
French	91.90%
Greek	91.05%
Hindi	88.50%
Italian	90.55%
Japanese	96.45%
Kazakh	63.20%
Marathi	92.60%
Persian	84.95%
Polish	55.48%
Russian	88.55%
Spanish	92.65%
Urdu	77.55%

Accuracy on OPUS 100 multilingual corpus test sets, containing 2000 sentences per language.

Usage

```python
import pysbd
text = "My name is Jonas E. Smith. Please turn to p. 55."
seg = pysbd.Segmenter(language="en", clean=False)
print(seg.segment(text))
# ['My name is Jonas E. Smith.', 'Please turn to p. 55.']
```

Conclusion

- **PySBD** has interpretable rules and are easy to modify
- Highly accurate - 97% English GRS - irrespective of domain
- Robust codebase with 98% test coverage
- Lightweight, easy to integrate with existing NLP pipelines
- Multilingual with 22 language support
- Already being used by 79 projects
- Extensible to handle more edge cases and languages in community driven way

Further information

GitHub Repo:
https://github.com/nipunsadvilkar/pySBD

Research Paper:
https://arxiv.org/abs/2010.09657