

TFDS now supports the [Croissant](https://mlcommons.org/croissant)  format (<https://mlcommons.org/croissant>)! Read the [documentation](https://www.tensorflow.org/datasets/format_specific_dataset_builders#croissantbuilder) (https://www.tensorflow.org/datasets/format_specific_dataset_builders#croissantbuilder) to know more.

wiki_auto

- **Description:**

WikiAuto provides a set of aligned sentences from English Wikipedia and Simple English Wikipedia as a resource to train sentence simplification systems. The authors first crowd-sourced a set of manual alignments between sentences in a subset of the Simple English Wikipedia and their corresponding versions in English Wikipedia (this corresponds to the `manual` config), then trained a neural CRF system to predict these alignments. The trained model was then applied to the other articles in Simple English Wikipedia with an English counterpart to create a larger corpus of aligned sentences (corresponding to the `auto`, `auto_acl`, `auto_full_no_split`, and `auto_full_with_split` configs here).

- **Homepage:** <https://github.com/chaojiang06/wiki-auto> (<https://github.com/chaojiang06/wiki-auto>)
- **Source code:** `tfds.text_simplification.wiki_auto.WikiAuto`
(https://github.com/tensorflow/datasets/tree/master/tensorflow_datasets/text_simplification/wiki_auto/wiki_auto.py)
- **Versions:**
 - `1.0.0` (default): Initial release.
- **Supervised keys** (See [as_supervised doc](#) (https://www.tensorflow.org/datasets/api_docs/python/tfds/load#args)): `None`
- **Figure** ([tfds.show_examples](#) (https://www.tensorflow.org/datasets/api_docs/python/tfds/visualization/show_examples)): Not supported.
- **Citation:**

```
@inproceedings{acl/JiangMLZX20,  
  author      = {Chao Jiang and  
                 Mounica Maddela and  
                 Wuwei Lan and
```

```

        Yang Zhong and
        Wei Xu},
    editor      = {Dan Jurafsky and
        Joyce Chai and
        Natalie Schluter and
        Joel R. Tetreault},
    title       = {Neural {CRF} Model for Sentence Alignment in Text Simplification},
    booktitle   = {Proceedings of the 58th Annual Meeting of the Association for Computat
        Linguistics, {ACL} 2020, Online, July 5-10, 2020},
    pages       = {7943--7960},
    publisher   = {Association for Computational Linguistics},
    year        = {2020},
    url         = {https://www.aclweb.org/anthology/2020.acl-main.709/}
}

```

wiki_auto/manual (default config)

- **Config description:** A set of 10K Wikipedia sentence pairs aligned by crowd workers.
- **Download size:** 53.47 MiB
- **Dataset size:** 76.87 MiB
- **Auto-cached** ([documentation](https://www.tensorflow.org/datasets/performances#auto-caching) (https://www.tensorflow.org/datasets/performances#auto-caching)): Yes
- **Splits:**

Split	Examples
'dev'	73,249
'test'	118,074

- **Feature structure:**

```

FeaturesDict({
    'GLEU-score': float64,
    'alignment_label': ClassLabel(shape=(), dtype=int64, num_classes=3),
    'normal_sentence': Text(shape=(), dtype=string),
    'normal_sentence_id': Text(shape=(), dtype=string),
    'simple_sentence': Text(shape=(), dtype=string),

```

```
'simple_sentence_id': Text(shape=(), dtype=string),
})
```

- **Feature documentation:**

Feature	Class	Shape	Dtype	Description
	FeaturesDict			
GLEU-score	Tensor		float64	
alignment_label	ClassLabel		int64	
normal_sentence	Text		string	
normal_sentence_id	Text		string	
simple_sentence	Text		string	
simple_sentence_id	Text		string	

- **Examples** ([tfds.as_dataframe](https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe) (https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe)):

[Display examples...](#)

wiki_auto/auto_acl

- **Config description:** Sentence pairs aligned to train the ACL2020 system.
- **Download size:** 112.60 MiB
- **Dataset size:** 138.83 MiB
- **Auto-cached** ([documentation](https://www.tensorflow.org/datasets/performances#auto-caching) (<https://www.tensorflow.org/datasets/performances#auto-caching>)): Only when `shuffle_files=False` (full)
- **Splits:**

Split	Examples
'full'	488,332

- **Feature structure:**

```
FeaturesDict({
    'normal_sentence': Text(shape=(), dtype=string),
    'simple_sentence': Text(shape=(), dtype=string),
})
```

- **Feature documentation:**

Feature	Class	Shape	Dtype	Description
	FeaturesDict			
normal_sentence	Text		string	
simple_sentence	Text		string	

- **Examples** ([tfds.as_dataframe](https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe) (https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe)):

[Display examples...](#)

wiki_auto/auto_full_no_split

- **Config description:** All automatically aligned sentence pairs without sentence splitting.
- **Download size:** 135.02 MiB
- **Dataset size:** 166.78 MiB
- **Auto-cached** ([documentation](https://www.tensorflow.org/datasets/performances#auto-caching) (<https://www.tensorflow.org/datasets/performances#auto-caching>)): Only when `shuffle_files=False` (full)

- **Splits:**

Split	Examples
'full'	591,994

- **Feature structure:**

```
FeaturesDict({
  'normal_sentence': Text(shape=(), dtype=string),
  'simple_sentence': Text(shape=(), dtype=string),
})
```

- **Feature documentation:**

Feature	Class	Shape	Dtype	Description
	FeaturesDict			
normal_sentence	Text		string	
simple_sentence	Text		string	

- **Examples** ([tfds.as_dataframe](https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe) (https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe)):

[Display examples...](#)

wiki_auto/auto_full_with_split

- **Config description:** All automatically aligned sentence pairs with sentence splitting.
- **Download size:** 115.09 MiB
- **Dataset size:** 141.20 MiB

- **Auto-cached** ([documentation](https://www.tensorflow.org/datasets/performances#auto-caching) (https://www.tensorflow.org/datasets/performances#auto-caching)): Only when `shuffle_files=False` (full)
- **Splits:**

Split	Examples
'full'	483,801

- **Feature structure:**

```
FeaturesDict({
  'normal_sentence': Text(shape=(), dtype=string),
  'simple_sentence': Text(shape=(), dtype=string),
})
```

- **Feature documentation:**

Feature	Class	Shape	Dtype	Description
	FeaturesDict			
normal_sentence	Text		string	
simple_sentence	Text		string	

- **Examples** ([tfds.as_dataframe](https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe) (https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe)):

[Display examples...](#)

wiki_auto/auto

- **Config description:** A large set of automatically aligned sentence pairs.
- **Download size:** 2.01 GiB

- **Dataset size:** 1.76 GiB
- **Auto-cached** ([documentation](https://www.tensorflow.org/datasets/performances#auto-caching) (<https://www.tensorflow.org/datasets/performances#auto-caching>)): No
- **Splits:**

Split	Examples
'part_1'	125,059
'part_2'	13,036

- **Feature structure:**

```
FeaturesDict({
  'example_id': Text(shape=(), dtype=string),
  'normal': FeaturesDict({
    'normal_article_content': Sequence({
      'normal_sentence': Text(shape=(), dtype=string),
      'normal_sentence_id': Text(shape=(), dtype=string),
    }),
    'normal_article_id': int32,
    'normal_article_title': Text(shape=(), dtype=string),
    'normal_article_url': Text(shape=(), dtype=string),
  }),
  'paragraph_alignment': Sequence({
    'normal_paragraph_id': Text(shape=(), dtype=string),
    'simple_paragraph_id': Text(shape=(), dtype=string),
  }),
  'sentence_alignment': Sequence({
    'normal_sentence_id': Text(shape=(), dtype=string),
    'simple_sentence_id': Text(shape=(), dtype=string),
  }),
  'simple': FeaturesDict({
    'simple_article_content': Sequence({
      'simple_sentence': Text(shape=(), dtype=string),
      'simple_sentence_id': Text(shape=(), dtype=string),
    }),
    'simple_article_id': int32,
    'simple_article_title': Text(shape=(), dtype=string),
    'simple_article_url': Text(shape=(), dtype=string),
  }),
})
```

- **Feature documentation:**

Feature	Class	Shape	Dtype	Description
	FeaturesDict			
example_id	Text		string	
normal	FeaturesDict			
normal/normal_article_content	Sequence			
normal/normal_article_content/normal_sentence	Text		string	
normal/normal_article_content/normal_sentence_id	Text		string	
normal/normal_article_id	Tensor		int32	
normal/normal_article_title	Text		string	
normal/normal_article_url	Text		string	
paragraph_alignment	Sequence			
paragraph_alignment/normal_paragraph_id	Text		string	
paragraph_alignment/simple_paragraph_id	Text		string	
sentence_alignment	Sequence			
sentence_alignment/normal_sentence_id	Text		string	
sentence_alignment/simple_sentence_id	Text		string	
simple	FeaturesDict			
simple/simple_article_content	Sequence			
simple/simple_article_content/simple_sentence	Text		string	
simple/simple_article_content/simple_sentence_id	Text		string	
simple/simple_article_id	Tensor		int32	
simple/simple_article_title	Text		string	
simple/simple_article_url	Text		string	

- **Examples** ([tfds.as_dataframe](https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe) (https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe)):

[Display examples...](#)

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (<https://creativecommons.org/licenses/by/4.0/>), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (<https://www.apache.org/licenses/LICENSE-2.0>). For details, see the [Google Developers Site Policies](https://developers.google.com/site-policies) (<https://developers.google.com/site-policies>). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2022-12-06 UTC.