Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, Thien Huu Nguyen
Department of Computer and Information Science
University of Oregon, Eugene, Oregon, USA
`{minhnv,vietl,apouranb,thien}@cs.uoregon.edu`

Abstract

We introduce **Trankit**, a light-weight **Tran**sformer-based Tool**kit** for multilingual Natural Language Processing (NLP). It provides a trainable pipeline for fundamental NLP tasks over 100 languages, and 90 pretrained pipelines for 56 languages. Built on a state-of-the-art pretrained language model, Trankit significantly outperforms prior multilingual NLP pipelines over sentence segmentation, part-of-speech tagging, morphological feature tagging, and dependency parsing while maintaining competitive performance for tokenization, multi-word token expansion, and lemmatization over 90 Universal Dependencies treebanks. Despite the use of a large pretrained transformer, our toolkit is still efficient in memory usage and speed. This is achieved by our novel plug-and-play mechanism with Adapters where a multilingual pretrained transformer is shared across pipelines for different languages. Our toolkit along with pretrained models and code are publicly available at: `https://github.com/nlp-uoregon/trankit`.
A demo website for our toolkit is also available at: `http://nlp.uoregon.edu/trankit`. Finally, we create a demo video for Trankit at: `https://youtu.be/q0KGP3zGjGc`.

1 Introduction

Many efforts have been devoted to developing multilingual NLP systems to overcome language barriers (Aharoni et al., 2019; Liu et al., 2019a; Taghizadeh and Faili, 2020; Zhu, 2020; Kanayama and Iwamoto, 2020; Nguyen and Nguyen, 2021). A large portion of existing multilingual systems has focused on downstream NLP tasks that critically depend on upstream linguistic features, ranging from basic information such as token and sentence boundaries for raw text to more sophisticated structures such as part-of-speech tags, morphological features, and dependency trees of sentences (called fundamental NLP tasks). As such, building effective multilingual systems/pipelines for fundamental upstream NLP tasks to produce such information has the potentials to transform multilingual downstream systems.

There have been several NLP toolkits that concerns multilingualism for fundamental NLP tasks, featuring spaCy[1], UDify (Kondratyuk and Straka, 2019), Flair (Akbik et al., 2019), CoreNLP (Manning et al., 2014), UDPipe (Straka, 2018), and Stanza (Qi et al., 2020). However, these toolkits have their own limitations. spaCy is designed to focus on speed, thus it needs to sacrifice the performance. UDify and Flair cannot process raw text as they depend on external tokenizers. CoreNLP supports raw text, but it does not offer state-of-the-art performance. UDPipe and Stanza are the recent toolkits that leverage word embeddings, i.e., word2vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017), to deliver current state-of-the-art performance for many languages. However, Stanza and UDPipe's pipelines for different languages are trained separately and do not share any component, especially the embedding layers that account for most of the model size. This makes their memory usage grow aggressively as pipelines for more languages are simultaneously needed and loaded into the memory (e.g., for language learning apps). Most importantly, none of such toolkits have explored contextualized embeddings from pretrained transformer-based language models that have the potentials to significantly improve the performance of the NLP tasks, as demonstrated in many prior works (Devlin et al., 2019; Liu et al., 2019b; Conneau et al., 2020).

In this paper, we introduce **Trankit**, a multilingual Transformer-based NLP Toolkit that over-
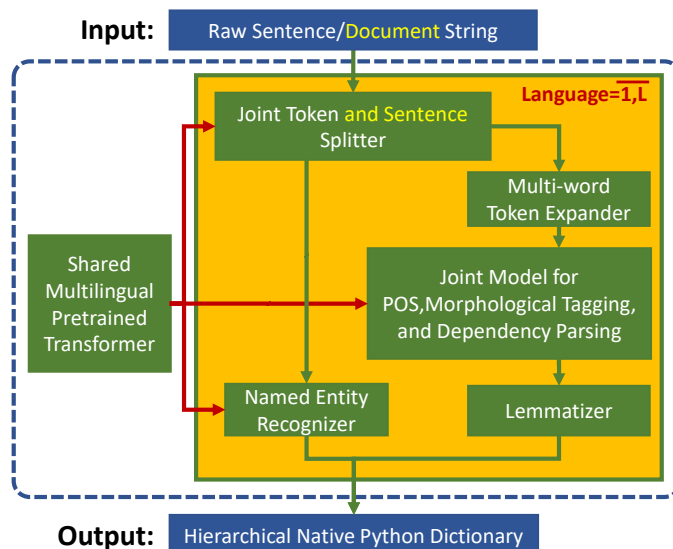
---

[1] `https://spacy.io/`

Figure 1: Overall architecture of Trankit. A single multilingual pretrained transformer is shared across three components (pointed by the red arrows) of the pipeline for different languages.

comes such limitations. Our toolkit can process raw text for fundamental NLP tasks, supporting 56 languages with 90 pre-trained pipelines on 90 treebanks of the Universal Dependency v2.5 (Zeman et al., 2019). By utilizing the state-of-the-art multilingual pretrained transformer *XLM-Roberta* (Conneau et al., 2020), Trankit advances state-of-the-art performance for sentence segmentation, part-of-speech (POS) tagging, morphological feature tagging, and dependency parsing while achieving competitive or better performance for tokenization, multi-word token expansion, and lemmatization over the 90 treebanks. It also obtains competitive or better performance for named entity recognition (NER) on 11 public datasets.

Unlike previous work, *our token and sentence splitter is wordpiece-based* instead of character-based to better exploit contextual information, which are beneficial in many languages. Considering the following sentence:

> *"John Donovan from Argghhh! has put out a excellent slide show on what was actually found and fought for in Fallujah."*

As such, Trankit correctly recognizes this as a single sentence while character-based sentence splitters of Stanza and UDPipe are easily fooled by the exclamation mark *"!"*, treating it as two separate sentences. To our knowledge, this is the first work to successfully build a wordpiece-based token and sentence splitter that works well for 56 languages.

Figure 1 presents the overall architecture of Trankit pipeline that features three novel

transformer-based components for: (i) the joint token and sentence splitter, (ii) the joint model for POS tagging, morphological tagging, dependency parsing, and (iii) the named entity recognizer. One potential concern for our use of a large pretrained transformer model (i.e., *XML-Roberta*) in Trankit involves GPU memory where different transformer-based components in the pipeline for one or multiple languages must be simultaneously loaded into the memory to serve multilingual tasks. This could extensively consume the memory if different versions of the large pre-trained transformer (finetuned for each component) are employed in the pipeline. As such, we introduce a novel plug-and-play mechanism with Adapters to address this memory issue. Adapters are small networks injected inside all layers of the pretrained transformer model that have shown their effectiveness as a lightweight alternative for the traditional finetuning of pretrained transformers (Houlsby et al., 2019; Peters et al., 2019; Pfeiffer et al., 2020a,b). In Trankit, a set of adapters (for transformer layers) and task-specific weights (for final predictions) are created for each transformer-based component for each language while only one single large multilingual pretrained transformer is shared across components and languages. Adapters allow us to learn language-specific features for tasks. During training, the shared pretrained transformer is fixed while only the adapters and task-specific weights are updated. At inference time, depending on the language of the input text and the current active

component, the corresponding trained adapter and task-specific weights are activated and plugged into the pipeline to process the input. This mechanism not only solves the memory problem but also substantially reduces the training time.

2   Related Work

There have been works using pre-trained transformers to build models for character-based word segmentation for Chinese (Yang, 2019; Tian et al., 2020; Che et al., 2020); POS tagging for Dutch, English, Chinese, and Vietnamese (de Vries et al., 2019; Tenney et al., 2019; Tian et al., 2020; Che et al., 2020; Nguyen and Nguyen, 2020); morphological feature tagging for Estonian and Persian (Kittask et al., 2020; Mohseni and Tebbifakhr, 2019); and dependency parsing for English and Chinese (Tenney et al., 2019; Che et al., 2020). However, all of these works are only developed for some specific language, thus potentially unable to support and scale to the multilingual setting.

Some works have designed multilingual transformer-based systems via multilingual training on the combined data of different languages (Tsai et al., 2019; Kondratyuk and Straka, 2019; Üstün et al., 2020). However, multilingual training is suboptimal (see Section 5). Also, these systems still rely on external resources to perform tokenization and sentence segmentation, thus unable to consume raw text. To our knowedge, this is the first work to successfully build a multilingual transformer-based NLP toolkit where different transformer-based models for many languages can be simultaneously loaded into GPU memory and process raw text inputs of different languages.

3   Design and Architecture

Adapters. Adapters play a critical role in making Trankit memory- and time-efficient for training and inference. Figure 2 shows the architecture and the location of an adapter inside a layer of transformer. We use the adapter architecture proposed by (Pfeiffer et al., 2020a,b), which consists of two projection layers Up and Down (feed-forward networks), and a residual connection.

$$c_i = \text{AddNorm}(r_i), h_i = \text{Up}(\text{ReLU}(\text{Down}(c_i))) + r_i \quad (1)$$

where r_i is the input vector from the transformer layer for the adapter and h_i is the output vector for the transformer layer i. During training, all the weights of the pretrained transformer (i.e., gray
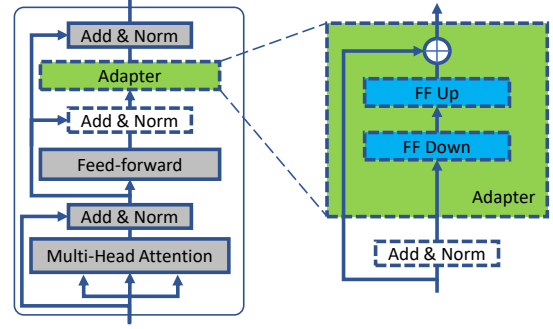

Figure 2: **Left**: location of an adapter (green box) inside a layer of the pretrained transformer. Gray boxes represent the original components of a transformer layer. **Right**: the network architecture of an adapter.

boxes) are fixed and only the adapter weights of two projection layers and the task-specific weights outside the transformer (for final predictions) are updated. As demonstrated in Figure 1, Trankit involves six components described as follows.

Multilingual Encoder with Adapters. This is our core component that is shared across different transformer-based components for different languages of the system. Given an input raw text s, we first split it into substrings by spaces. Afterward, Sentence Piece, a multilingual subword tokenizer (Kudo and Richardson, 2018; Kudo, 2018), is used to further split each substring into wordpieces. By concatenating wordpiece sequences for substrings, we obtain an overall sequence of wordpieces $\mathbf{w} = [w_1, w_2, \ldots, w_K]$ for s. In the next step, $\mathbf{w}$ is fed into the pretrained transformer, which is already integrated with adapters, to obtain the wordpiece representations:

$$x_{1:K}^{l,m} = \text{Transformer}(w_{1:K}; \theta_{AD}^{l,m}) \quad (2)$$

Here, $\theta_{AD}^{l,m}$ represents the adapter weights for language l and component m of the system. As such, we have specific adapters in all transformer layers for each component m and language l. Note that if K is larger than the maximum input length of the pretrained transformer (i.e., 512), we further divide $\mathbf{w}$ into consecutive chunks; each has the length less than or equal to the maximum length. The pretrained transformer is then applied over each chunk to obtain a representation vector for each wordpiece in $\mathbf{w}$. Finally, $x_{1:K}^{l,m}$ will be sent to component m to perform the corresponding task.

Joint Token and Sentence Splitter. Given the wordpiece representations $x_{1:K}^{l,m}$ for this component,

each vector $x_i^{l,m}$ for $w_i \in \mathbf{w}$ will be consumed by a feed-forward network with softmax in the end to predict if w_i is the end of a single-word token, the end of a multi-word token, or the end of a sentence. The predictions for all wordpieces in $\mathbf{w}$ will then be aggregated to determine token, multi-word token, and sentence boundaries for s.

Multi-word Token Expander. This component is responsible for expanding each detected multi-word token (MWT) into multiple syntactic words[2]. We follow Stanza to deploy a character-based seq2seq model for this component. This decision is made based on our observation that the task is done best at character level, and the character-based model (with character embeddings) is very small.

Joint Model for POS Tagging, Morphological Tagging and Dependency Parsing. In Trankit, given the detected sentences and tokens/words, we use a single model to jointly perform POS tagging, morphological feature tagging and dependency parsing at sentence level. Joint modeling mitigates error propagation, saves the memory, and speedups the system. In particular, given a sentence, the representation for each word is computed as the average of its wordpieces' transformer-based representations in $x_{1:K}^{l,m}$. Let $t_{1:N} = [t_1, t_2, \ldots, t_N]$ be the representations of the words in the sentence. We compute the following vectors using feed-forward networks FFN_*:

$$r_{1:N}^{upos} = \text{FFN}_{upos}(t_{1:N}), r_{1:N}^{xpos} = \text{FFN}_{xpos}(t_{1:N})$$
$$r_{1:N}^{ufeats} = \text{FFN}_{ufeats}(t_{1:N}), r_{0:N}^{dep} = [x_{cls}; \text{FFN}_{dep}(t_{1:N})]$$

Vectors for the words in $r_{1:N}^{upos}$, $r_{1:N}^{xpos}$, $r_{1:N}^{ufeats}$ are then passed to a softmax layer to make predictions for UPOS, XPOS, and UFeats tags for each word. For dependency parsing, we use the classification token `<s>` to represent the root node, and apply Deep Biaffine Attention (Dozat and Manning, 2017) and the Chu-Liu/Edmonds algorithm (Chu, 1965; Edmonds, 1967) to assign a syntactic head and the associated dependency relation to each word in the sentence.

Lemmatizer. This component receives sentences and their predicted UPOS tags to produce the canonical form for each word. We also employ a character-based seq2seq model for this component as in Stanza.

---

[2]For languages (e.g., English, Chinese) that do not require MWT expansion, tokens and words are the same concepts.

Named Entity Recognizer. Given a sentence, the named entity recognizer determines spans of entity names by assigning a BIOES tag to each token in the sentence. We deploy a standard sequence labeling architecture using transformer-based representations for tokens, involving a feed-forward network followed by a Conditional Random Field.

4 Usage

Detailed documentation for Trankit can be found at: https://trankit.readthedocs.io.

Trankit Installation. Trankit is written in Python and available on PyPI: https://pypi.org/project/trankit/. Users can install our toolkit via pip using:

```
pip install trankit
```

Initialize a Pipeline. Lines 1-4 in Figure 3 shows how to initialize a pretrained pipeline for English; it is instructed to run on GPU and store downloaded pretrained models to the specified cache directory. Trankit will not download pretrained models if they already exist.

Multilingual Usage. Figure 3 shows how to initialize a multilingual pipeline and process inputs of different languages in Trankit:

```
1  from trankit import Pipeline
2
3  # initialize a multilingual pipeline
4  p = Pipeline(lang='english', gpu=True, cache_dir='./cache')
5  langs = ['arabic', 'chinese', 'dutch']
6  for lang in langs:
7          p.add(lang)
8
9  # tokenize English input
10 p.set_active('english')
11 en = p.tokenize('Rich was here before the scheduled time.')
12
13 # get ner tags for Arabic input
14 p.set_active('arabic')
15 ar = p.ner('وكانت قد ذعاد قبل ذلك رئيس جهاز الامن والاستطلاع للقوات السورية العاملة فى لبنان')
```

Figure 3: Multilingual pipeline initialization.

Basic Functions. Trankit can process inputs which are untokenized (raw) or pretokenized strings, at both sentence and document levels. Figure 4 illustrates a simple code to perform all the supported tasks for an input text. We organize Trankit's outputs into hierarchical native Python dictionaries, which can be easily inspected by users. Figure 5 demonstrates the outputs of the command line 6 in Figure 4.

Training your own Pipelines. Trankit also provides a trainable pipeline for 100 languages via the class `TPipeline`. This ability is inherited from

Treebank	System	Tokens	Sents.	Words	UPOS	XPOS	UFeats	Lemmas	UAS	LAS
Overall (90 treebanks)	**Trankit**	99.23	**91.82**	**99.02**	**95.65**	**94.05**	**93.21**	**94.27**	**87.06**	**83.69**
	Stanza	**99.26**	88.58	98.90	94.21	92.50	91.75	94.15	83.06	78.68
Arabic-PADT	**Trankit**	99.93	**96.59**	**99.22**	**96.31**	**94.08**	**94.28**	**94.65**	**88.39**	**84.68**
	Stanza	**99.98**	80.43	97.88	94.89	91.75	91.86	93.27	83.27	79.33
	UDPipe	99.98	82.09	94.58	90.36	84.00	84.16	88.46	72.67	68.14
Chinese-GSD	**Trankit**	**97.01**	**99.7**	**97.01**	**94.21**	**94.02**	**96.59**	**97.01**	**85.19**	**82.54**
	Stanza	92.83	98.80	92.83	89.12	88.93	92.11	92.83	72.88	69.82
	UDPipe	90.27	99.10	90.27	84.13	84.04	89.05	90.26	61.60	57.81
English-EWT	**Trankit**	98.48	**88.35**	98.48	**95.95**	**95.71**	**96.26**	96.84	**90.14**	**87.96**
	Stanza	**99.01**	81.13	**99.01**	95.40	95.12	96.11	**97.21**	86.22	83.59
	UDPipe	98.90	77.40	98.90	93.26	92.75	94.23	95.45	80.22	77.03
	spaCy	97.44	63.16	97.44	86.99	91.05	-	87.16	55.38	37.03
French-GSD	**Trankit**	**99.7**	**96.63**	**99.66**	**97.85**	-	**97.16**	**97.80**	**94.00**	**92.34**
	Stanza	99.68	94.92	99.48	97.30	-	96.72	97.64	91.38	89.05
	UDPipe	99.68	93.59	98.81	95.85	-	95.55	96.61	87.14	84.26
	spaCy	99.02	89.73	94.81	89.67	-	-	88.55	75.22	66.93
Spanish-Ancora	**Trankit**	99.94	**99.13**	99.93	**99.02**	**98.94**	**98.8**	99.17	**94.11**	**92.41**
	Stanza	**99.98**	99.07	**99.98**	98.78	98.67	98.59	99.19	92.21	90.01
	UDPipe	99.97	98.32	99.95	98.32	98.13	98.13	98.48	88.22	85.10
	spaCy	99.95	97.54	99.43	93.43	-	-	80.02	89.35	83.81

Table 1: Systems' performance on test sets of the Universal Dependencies v2.5 treebanks. Performance for Stanza, UDPipe, and spaCy is obtained using their public pretrained models. The overall performance for Trankit and Stanza is computed as the macro-averaged F1 over 90 treebanks. Detailed performance of Trankit for 90 supported treebanks can be found at our documentation page.

```
1  from trankit import Pipeline
2
3  p = Pipeline(lang='english', gpu=True, cache_dir='./cache')
4
5  doc = '''Hello! This is Trankit.'''
6  # perform all tasks on the input
7  all = p(doc)
```

Figure 4: A function performing all tasks on the input.

the XLM-Roberta encoder which is pretrained on those languages. Figure 6 illustrates how to train a token and sentence splitter with TPipeline.

Demo Website. A demo website for Trankit to support 90 pretrained pipelines is hosted at: http://nlp.uoregon.edu/trankit. Figure 7 shows its interface.

```
// Output
{
  'text': 'Hello! This is Trankit.', // input string
  'sentences': [ // list of sentences
    {
      'id': 1, 'text': 'Hello!', 'dspan': (0, 6), 'tokens': [...]
    },
    {
      'id': 2, // sentence index
      'text': 'This is Trankit.', 'dspan': (7, 23), // sentence span
      'tokens': [ // list of tokens
        {
          'id': 1, // token index
          'text': 'This', 'upos': 'PRON', 'xpos': 'DT',
          'feats': 'Number=Sing|PronType=Dem',
          'head': 3, 'deprel': 'nsubj', 'lemma': 'this', 'ner': 'O',
          'dspan': (7, 11), // document-level span of the token
          'span': (0, 4)     // sentence-level span of the token
        },
        {'id': 2...},
        {'id': 3...},
        {'id': 4...}
      ]
    }
  ]
}
```

Figure 5: Output from Trankit. Some parts are collapsed to improve visualization.

5 System Evaluation

5.1 Datasets & Hyper-parameters

To achieve a fair comparison, we follow Stanza (Qi et al., 2020) to train and evaluate all the models on the same canonical data splits of 90 Universal Dependencies treebanks v2.5 (UD2.5)[3] (Zeman et al., 2019), and 11 public NER datasets provided in the following corpora: AQMAR (Mohit et al., 2012), CoNLL02 (Tjong Kim Sang, 2002), CoNLL03 (Tjong Kim Sang and De Meul-

der, 2003), GermEval14 (Benikova et al., 2014), OntoNotes (Weischedel et al., 2013), and WikiNER (Nothman et al., 2012). Hyper-parameters for all models and datasets are selected based on the development data in this work.

5.2 Universal Dependencies performance

Table 1 compares the performance of Trankit and the latest available versions of other popular toolkits, including Stanza (v1.1.1) with current state-of-the-art performance, UDPipe (v1.2), and spaCy (v2.3) on the UD2.5 test sets. The performance for all systems is obtained using the official scorer

---

[3]We skip 10 treebanks whose languages are not supported by *XLM-Roberta*.

System	Tokens	Sents.	Words	UPOS	XPOS	UFeats	Lemmas	UAS	LAS
Trankit (plug-and-play with adapters)	**99.05**	**95.12**	**98.96**	**95.43**	**89.02**	**92.69**	**93.46**	**86.20**	**82.51**
Multilingual	96.69	88.95	96.35	91.19	84.64	88.10	90.02	72.96	68.66
No-adapters	95.06	89.57	94.08	88.79	82.54	83.76	88.33	66.63	63.11

Table 2: Model performance on 9 different treebanks (macro-averaged F1 score over test sets).

```
1  from trankit import TPipeline
2
3  tp = TPipeline(training_config={
4          'task': 'tokenize',
5          'save_dir': './saved_model',
6          'train_txt_fpath': './train.txt',
7          'train_conllu_fpath': './train.conllu',
8          'dev_txt_fpath': './dev.txt',
9          'dev_conllu_fpath': './dev.conllu'})
10
11 trainer.train()
```

Figure 6: Training a token and sentence splitter using the CONLL-U formatted data (Nivre et al., 2020).

of the CoNLL 2018 Shared Task[4]. On five illustrated languages, Trankit achieves competitive performance on tokenization, MWT expansion, and lemmatization. Importantly, Trankit outperforms other toolkits over all remaining tasks (e.g., POS and morphological tagging) in which the improvement boost is substantial and significant for sentence segmentation and dependency parsing. For example, English enjoys a 7.22% improvement for sentence segmentation, a 3.92% and 4.37% improvement for UAS and LAS in dependency parsing. For Arabic, Trankit has a remarkable improvement of 16.16% for sentence segmentation while Chinese observes 12.31% and 12.72% improvement of UAS and LAS for dependency parsing.

Over all 90 treebanks, Trankit outperforms the previous state-of-the-art framework Stanza in most of the tasks, particularly for sentence segmentation (+3.24%), POS tagging (+1.44% for UPOS and +1.55% for XPOS), morphological tagging (+1.46%), and dependency parsing (+4.0% for UAS and +5.01% for LAS) while maintaining the competitive performance on tokenization, multiword expansion, and lemmatization.

5.3 NER results

Table 3 compares Trankit with Stanza (v1.1.1), Flair (v0.7), and spaCy (v2.3) on the test sets of 11 considered NER datasets. Following Stanza, we report the performance for other toolkits with their pretrained models on the canonical data splits if they are available. Otherwise, their best configurations are used to train the models on the same data splits (inherited from Stanza). Also, for the Dutch

datasets, we retrain the models in Flair as those models (for Dutch) have been updated in version v0.7. As can be seen, Trankit obtains competitive or better performance for most of the languages, clearly demonstrating the benefit of using the pretrained transformer for multilingual NER.

Language	Corpus	**Trankit**	Stanza	Flair	spaCy
Arabic	AQMAR	**74.8**	74.3	74.0	-
Chinese	OntoNotes	**80.0**	79.2	-	69.3
Dutch	CoNLL02	**91.8**	89.2	91.3	73.8
	WikiNER	**94.8**	**94.8**	**94.8**	90.9
English	CoNLL03	92.1	92.1	**92.7**	81.0
	OntoNotes	**89.6**	88.8	89.0	85.4
French	WikiNER	92.3	**92.9**	92.5	88.8
German	CoNLL03	**84.6**	81.9	82.5	63.9
	GermEval14	**86.9**	85.2	85.4	68.4
Russian	WikiNER	92.8	**92.9**	-	-
Spanish	CoNLL02	**88.9**	88.1	87.3	77.5

Table 3: Performance (F1) on NER test sets.

System	GPU		CPU	
	UD	NER	UD	NER
Trankit	4.50×	1.36×	19.8×	31.5×
Stanza	3.22×	1.08×	10.3×	17.7×
UDPipe	-	-	4.30×	-
Flair	-	1.17×	-	51.8×

Table 4: Run time on processing the English EWT treebank and the English Ontonotes NER dataset. Measurements are done on an NVIDIA Titan RTX card.

Model Package	Trankit	Stanza
Multilingual Transformer	1146.9MB	-
Arabic	38.6MB	393.9MB
Chinese	40.6MB	225.2MB
English	47.9MB	383.5MB
French	39.6MB	561.9MB
Spanish	37.3MB	556.1MB
Total size	1350.9MB	2120.6MB

Table 5: Model sizes for five languages.

5.4 Speed and Memory Usage

Table 4 reports the relative processing time for UD and NER of the toolkits compared to spaCy's CPU processing time[5]. For memory usage comparison, we show the model sizes of Trankit and

---

[4] https://universaldependencies.org/conll18/evaluation.html

[5] spaCy can process 8140 tokens and 5912 tokens per second for UD and NER, respectively.
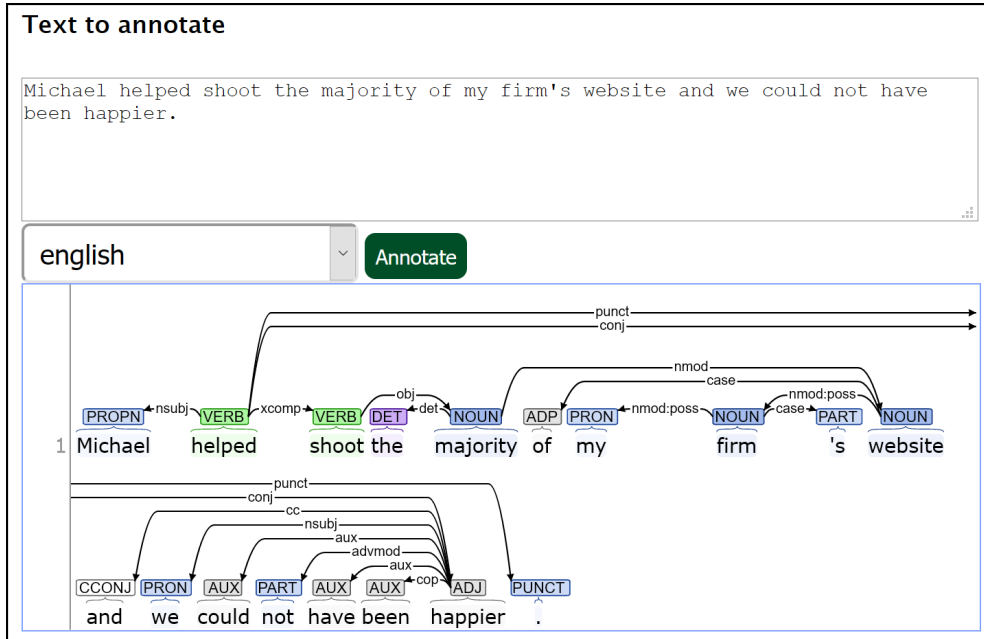
Figure 7: Demo website for Trankit.

Stanza for several languages in Table 5. As can be seen, besides the multilingual transformer, model packages in Trankit only take dozens of megabytes while Stanza consumes hundreds of megabytes for each package. This leads to the Stanza's usage of much more memory when the pipelines for these languages are loaded at the same time. In fact, Trankit only takes 4.9GB to load all the 90 pre-trained pipelines for the 56 supported languages.

5.5 Ablation Study

This section compares Trankit with two other possible strategies to build a multilingual system for fundamental NLP tasks. In the first strategy (called "*Multilingual*"), we train a single pipeline where all the components in the pipeline are trained with the combined training data of all the languages. The second strategy (called "*No-adapters*") involves eliminating adapters from *XLM-Roberta* in Trankit. As such, in "*No-adapters*", pipelines are still trained separately for each language; the pretrained transformer is fixed; and only task-specific weights (for predictions) in components are updated during training.

For evaluation, we select 9 treebanks for 3 different groups, i.e., high-resource, medium-resource, and low-resource, depending on the sizes of the treebanks. In particular, the high-resource group includes Czech, Russian, and Arabic; the medium-resource group includes French, English, and Chinese; and the low-resource group involves Belaru-

sian, Telugu, and Lithuanian. Table 2 compares the average performance of Trankit, "*Multilingual*", and "*No-adapters*". As can be seen, "*Multilingual*" and "*No-adapters*" are significantly worse than the proposed adapter-based Trankit. We attribute this to the fact that multilingual training might suffer from unbalanced sizes of treebanks, causing high-resource languages to dominate others and impairing the overall performance. For "*No-adapters*", fixing pretrained transformer might significantly limit the models' capacity for multiple tasks and languages.

6 Conclusion and Future Work

We introduce Trankit, a transformer-based multilingual toolkit that significantly improves the performance for fundamental NLP tasks, including sentence segmentation, part-of-speech, morphological tagging, and dependency parsing over 90 Universal Dependencies v2.5 treebanks of 56 different languages. Our toolkit is fast on GPUs and efficient in memory use, making it usable for general users. In the future, we plan to improve our toolkit by investigating different pretrained transformers such as mBERT and XLM-Roberta$_{large}$. We also plan to provide Named Entity Recognizers for more languages and add modules to perform more NLP tasks.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-d named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2524–2531, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-ltp: A open-source neural chinese language technology platform with pretrained models. *arXiv preprint arXiv:2009.11616*.

Yoeng-Jin Chu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the International Conference on Learning Representations*.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the national Bureau of Standards B*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning*.

Hiroshi Kanayama and Ran Iwamoto. 2020. How universal are Universal Dependencies? exploiting syntax for multilingual clause-level sentiment detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4063–4073, Marseille, France. European Language Resources Association.

Claudia Kittask, Kirill Milintsevich, and Kairit Sirts. 2020. Evaluating multilingual bert for estonian. *Volume*, 328:19–26.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019a. Neural cross-lingual event detection with minimal parallel resources. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 738–748, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Conference on Neural Information Processing Systems*.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France. Association for Computational Linguistics.

Mahdi Mohseni and Amirhossein Tebbifakhr. 2019. MorphoBERT: a Persian NER system with BERT and morphological analysis. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 23–30, Trento, Italy. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Minh Van Nguyen and Thien Huu Nguyen. 2021. Improving cross-lingual transfer for event argument extraction with language-universal sentence structures. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP) at EACL 2021*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Nasrin Taghizadeh and Heshaam Faili. 2020. Cross-lingual adaptation using universal dependencies. *arXiv preprint arXiv:2003.10816*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical BERT models for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3632–3636, Hong Kong, China. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0. *Linguistic Data Consortium*.

Haiqin Yang. 2019. Bert meets chinese word segmentation. *arXiv preprint arXiv:1909.09292*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot,

Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Xingran Zhu. 2020. Cross-lingual word sense disambiguation using mbert embeddings with syntactic dependencies. *arXiv preprint arXiv:2012.05300*.