CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies

1 Contacts and URLs

Website: http://universaldependencies.org/conll17

Email contact: ud.conll.shared.task.2017@gmail.com

Mailing list for the organizing committee (if you are not receiving this address, you probably have not confirmed the invitation; ask Dan Zeman for a new invitation): ud-conll-shared-task@googlegroups.com. (Non-members cannot send e-mails to this list; use the above contact instead.)

2 Rules for treebank inclusion

- Test data must contain at least 10,000 syntactic words (possibly after re-splitting). UDv2 annotated test data will not be part of the next UD release and they should not appear in the Github repositories until the shared task is over! (Not even in the dev branch.) Send the data to the shared task organizers by e-mail instead (see the contact above).
- Development data should also contain 10,000 words or more, but this is not so strict. Contact the shared task organizers if you are not able to meet this condition. There is no size requirement for training data: if you have just 20,000 words, split it to 10K dev + 10K test and leave the training data empty.
- The UD validation script will be updated to check conformance with v2 guidelines. A treebank must pass the validation to be included in the shared task. (Test data will be validated only offline.) The validation will probably include a subset of what is now known as “content validation tests” (e.g. check that certain types of relations are left-headed). Lemmas and morphological features are still optional, although treebank owners are strongly encouraged to include them.
- The data must be ready and valid by February 15 (see also the time line below). In exceptional cases we may allow deadline extension for the test data. E.g. the treebank is small, annotation is running but there are only 15,000 words available by the release deadline. The annotation team is confident that they can exceed 20,000 words soon; they will thus ask us to release 10,000 words as development data and wait for the remaining 5,000 words of test data. Obviously we do not want this to become a common practice because we want to announce the set of shared task languages in the beginning of March, and we do not want to withdraw a language later, should the annotators fail to supply the remaining data.
- In order to give us an idea about how many languages we should expect, we ask the teams maintaining individual treebanks to let us know by mid January that they are aiming at meeting the above conditions and have their data in the shared task.
- As mentioned in the shared task specification, there will be a parallel test set of ~1000 sentences in selected shared task languages. DFKI and Google are generously providing translations & annotations in selected languages (English, Spanish, Portuguese, French, Italian, Russian, Japanese, Hindi, Arabic, Indonesian, Chinese, Turkish, German). If you are maintaining a UD treebank in one of these languages, please let us know whether you can check the annotation quality of the parallel data for us. If your language is not listed above but you are willing to translate these sentences from English to your language and annotate it UD v2 style, get in touch too (this is already the case of Swedish, Czech, Finnish and Norwegian).

3 Detailed timeline (for organizers)

Color coding of deadline addressees: participants – “ordinary” data providers – specific data processing – organizing committee.

- December 11 (Sunday): Announcement of the shared task and set up of the shared task website.
Registration for the Shared Task open (by e-mail unless we manage to set up a web registration form).
- Beginning of January: Translations done by DFKI should be available.
- January 10 (Tuesday): Deadline for suggesting additional data by participants (registration necessary).

- January 15 (Sunday): “Commitment deadline” for data contributors – tell the organizers that you intend to have your data included.
- February 15 (Wednesday): The raw data from CommonCrawl, prepared by Turku, must now be available for Milan to start preparing UDPipe (no embeddings needed at this moment) and to train the detokenizer for languages lacking SpaceAfter=No.
- February 15 (Wednesday): Data freeze for UD v2 data that is going to be included in the shared task.
- February 20 (Monday): Trial data publicly available.
- February 28 (Tuesday): Task Registration deadline. Participants have to register to setup their evaluation space and other data, and get access to task data later.
- March 1 (Wednesday): Release of training + development data. That includes versions of training + development data with annotation predicted by UDPipe, and also the raw data with pre-computed word embeddings (Turku) and preprocessing by UDPipe (Milan). CAUTION: UDPipe will probably use models from UD 1.4 and the output must be adjusted to UD 2.0! That involves CONJ-CCONJ but also some features, including language-specific features.
- March 31 (Friday): Google annotation of parallel data from DFKI available.
- April 16 (Sunday): Parallel test data converted to v2 by Dan, language owners should now check them.
- April 23 (Sunday): A preliminary version of the surprise language test data, with manual POS tags, should be available for Milan to start training hyper-parameters.
- April 30 (Sunday): Data-freeze for the parallel test data and for the surprise language test data. All languages that did not take the mainstream path DFKI – Google – Dan – final check, must now be available in UDv2 too. Milan has now 1 week to process the data by UDPipe.
- May 1 (Monday): Descriptions of surprise languages and sample data released. If we want to provide raw data for the surprise languages, it must be ready by now too.
- May 8 – 12 (Monday – Friday): Test phase. We could make the test data available some time between Friday evening and Monday morning (preferably before European midnight Sunday-Monday). The deadline for the system outputs could be Friday 23:59 Samoa Standard Time (UTC-11, i.e. Saturday 12:59 in Central Europe).
- May 15 (Monday): Results announced (someone has to work over the weekend).
- May 26 (Friday): Submission of papers (we will have to negotiate extended deadline because the other ACL workshops will probably have their deadline on Friday May 19).
- June 2 (Friday): Reviews due.
- June 9 (Friday): Final papers due. (While the other workshops probably will have the camera-ready deadline two weeks earlier, Friday May 26.)
- August 3 – 4 (Thursday – Friday): CoNLL conference, Vancouver, Canada