About the Corpus

The DEFT corpus contains roughly 7,000 sets of 3-sentence groupings extracted from textbooks of various topics from cnx.org. Each sentence set reflects a context window around which the author of the original text marked a bolded word to indicate a key term. For annotation reasons, the bolded words are not included in the annotated corpus, though you may find them from the textbooks themselves. Each grouping may have multiple term-definition pairs or none at all - it is simply a context window around a likely location for a term.

Annotation Schema

The DEFT annotation schema is comprised of terms and definitions, as well as various auxiliary tags which aid in identifying complex or long-distance relationships between a term-definition pair. With the exception of "implicit" definitions (defined below), all terms are linked in some way to a definition or alias term. The full schema is as follows:

Tag Schema

Term	A primary term.
Alias Term	A secondary or less common name for the primary term. Links to a term tag.
Ordered Term	Multiple terms that have matching sets of definitions which cannot be separated from each other without creating a non-contiguous sequence of tokens. Eg *x and y represent positive and negative versions of definition z, respectively.*
Referential Term	An NP reference to a previously mentioned term tag. Typically *this/that/these + NP* following a sentence boundary.
Definition	A primary definition of a term. May not exist without a matching term.
Secondary Definition	Supplemental information that may qualify as a definition sentence or phrase, but crosses a sentnece boundary.
Ordered Definition	Multiple definitions that have matching sets of terms which cannot be separated from each other. See Ordered Term.
Referential Definition	NP reference to a previously mentioned definition tag. See Referential Term.
Qualifier	A specific date, location, or other condition under which the definition holds true. Typically seen at the clause level.

Relation Schema

Direct-defines	Links definition to term.
Indirect-defines	Links definition to referential term or term to referential definition.
Refers-to	Links referential term to term or referential definition to definition.
AKA	Links alias term to term.
Supplements	Links secondary definition to definition.
Qualifies	Links qualifier to term.

For more information, please refer to our LAW 2019 workshop paper introducing the corpus here.

Format

Train and dev data is provided to you in a CONLL-like tab-deliniated format. Each line represents a token and its features. A single blank line indicates a sentence break; two blank lines indicates a new 3-sentence context window. All context windows begin with a sentence id followed by a period. These are treated as tokens in the data. Each token is represented by the following features:

[TOKEN] [SOURCE] [START_CHAR] [END_CHAR] [TAG] [TAG_ID] [ROOT_ID] [RELATION]

Where:

- SOURCE is the source .txt file of the excerpt
- START_CHAR/END_CHAR are char index boundaries of the token
- TAG is the label of the token (O if not a B-[TAG] or I-[TAG])
- TAG_ID is the ID associated with this TAG (0 if none)
- ROOT_ID is the ID associated with the root of this relation (-1 if no relation/O tag, 0 if root, and TAG_ID of root if not the root)
- RELATION is the relation tag of the token (0 if none).

Test data will be evaluated in the following CONLL-2003-like formats:

- Subtask 1: Sentence Classification
  - [SENTENCE] [BIN_TAG] Where the binary tag is 1 if the sentence contains a definition and 0 if the sentence does not contain a definition.
- Subtask 2: Sequence Labeling
  - [TOKEN] [SOURCE] [START_CHAR] [END_CHAR] [TAG]
- Subtask 3: Relation Extraction
  - [TOKEN] [SOURCE] [START_CHAR] [END_CHAR] [TAG] [TAG_ID] [ROOT_ID] [RELATION] Where ROOT_ID is -1 if there is no relation, 0 if the token is part of the root, and TAG_ID of the root if the token points to the root.

We will provide scripts for converting the original training data format into the appropriate format for sentence classification.

Access and Terms

Data is available on Github. By participating in this competition, you agree to use the train/test/dev splits for their appropriate use (i.e., you will only use train data for training, dev data for development, and so on). Terms and conditions related to the use and distribution of the data is available under Terms and Conditions.