AIDA 2018 Resources for Constrained Training Condition V1.3

September 10, 2018

Resources allowed under the constrained evaluation condition must:

- Be publicly available for free Web download or via the Linguistic Data Consortium
- Not violate the Time Machine Principle

The AIDA 2018 Russia-Ukraine scenario begins November 26, 2013. Resources that predate the scenario all meet the requirements of the Time Machine Principle. Some resources that do not predate the scenario may also be allowable if they follow the spirit of the Time Machine Principle and are unlikely to overlap with the scenario.

The following resources are ALLOWED under the constrained evaluation condition:

Data distributed for the SM-KBP 2018 pilot evaluation

- All LDC packages released under the TAC 2018 Evaluation License Agreement
- Ukraine-Russia-Relations-Scenario-Document_2018-07-20_v1.2.pdf
    - Note: This violates the Time Machine Principle, but is OK to use for the pilot; for future evaluation, information in the scenario document will likely predate the evaluation topics
- The seedling training corpus and seedling training annotation:
    - LDC2018E01, LDC2018E45, LDC2018E52, LDC2018E53, LDC2018E63
    - Note that additional annotation done on the seedling corpus (beyond annotations specific to the training topics) would count as unconstrained data. After M9, the training topics will hopefully be earlier than all of the eval topics, so that participants would be allowed to do unlimited additional annotation on the training topic documents under the constrained condition.

All LDC data packages released before the scenario, including but not limited to:

- ACE 2005 English dataset: https://catalog.ldc.upenn.edu/ldc2006t06

Web documents that predate the scenario:

- Voice of America articles (list compiled by RPI): https://tac.nist.gov/2018/SM-KBP/guidelines/en_voa_pre_20131126.txt

Generic lexical resources:

- WordNet
- VerbNet
- FrameNet
- PropBank

Common NLP datasets:

- AMR (LDC2017T10)
- MPQA: http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/
- GFBF: http://mpqa.cs.pitt.edu/corpora/gfbf/
- universal dependencies (recently annotated training corpora): http://universaldependencies.org/
- CoNLL 2003 English NER dataset: https://www.clips.uantwerpen.be/conll2003/ner/
- OpenSubtitles 2018 Russian-English. This is parallel data extracted from movie subtitles (P. Lison and J. Tiedemann, 2016, OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles, LREC 2016): http://opus.nlpl.eu/download.php?f=OpenSubtitles2018/en-ru.txt.zip
- OpenSubtitles 2018 Ukrainian-English. This is parallel data extracted from movie subtitles (P. Lison and J. Tiedemann, 2016, OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles, LREC 2016): http://opus.nlpl.eu/download.php?f=OpenSubtitles2018/en-uk.txt.zip
- IARPA BABEL corpora
  - https://docs.google.com/spreadsheets/d/1Cv7_VNlGYYf7HDSaXkZeE5Bv0nvPl1t_qWTPcR37eVg/edit?usp=sharing
  - Notes: It seems the last files were recorded mid-2015, but the content (both informational and the languages) does not really interfere with the example incidents/scenarios. OK to use for speech AM training, but DO NOT extract any additional knowledge or information.
- Russian translation (Kutuzov & Kunilovskaya) of Google Analogies (Mikolov et al.)
  - http://rusvectores.org/static/testsets/ru_analogy.txt
  - Date: approx 12/21/2017 (English original: 1/162013)
- ConceptNet: http://conceptnet.io/
- Pretrained components that are available as open source toolkits:
  - Spacy
  - DRAGNN (formerly SyntaxNet)

Common image and video public datasets for training generic visual classes of objects, people, locations, actions, events, etc.:

- TRECVID video collections such as (Semantic Indexing IACC , Sound&Vision, and MED)
- ActivityNet
- OpenImage V4
- "Labeled Faces in the Wild"
  - http://vis-www.cs.umass.edu/lfw/
    Direct Link: http://vis-www.cs.umass.edu/lfw/lfw.tgz
  - Notes: Collected in 2007
- The learned Deep Network parameters for https://github.com/davidsandberg/facenet which embed a face image as a vector come from: "CASIA WebFace Database"

    - http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html
    - Notes: Collected by 2014
- "MS-Celeb-1M"
    - https://www.microsoft.com/en-us/research/project/ms-celeb-1m-challenge-recognizing-one-million-celebrities-real-world/
    - Notes: Collected by 2017
- "Flicker 30k"
    - http://web.engr.illinois.edu/~bplumme2/Flickr30kEntities/
      Direct Link: http://web.engr.illinois.edu/~bplumme2/Flickr30kEntities/
    - Notes: 2014-2015, Contains captions for images
- "MSR-VTT"
    - http://ms-multimedia-challenge.com/2017/dataset
    - Notes: 2017, Contains captions for videos
- "COCO dataset"
    - http://cocodataset.org/#home
    - Notes: Collected by 2014
- "ImageNet"
    - https://www.kaggle.com/c/imagenet-object-detection-challenge/data
    - http://image-net.org/download-imageurls
- "FaceNet"
    - https://github.com/davidsandberg/facenet
- VGGFace2
    - https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/
    - Notes: Collected in 2017. Creative Commons Attribution 4.0 International license
- WIDER Face
    - http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/
    - Notes: Collected in 2016.
- CelebA
    - http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
    - Notes: Collected in 2015

general-purpose KBs that are fairly static over time:

- GeoNames

Broader KBs that have versions that do not violate the time machine principle:

- Wikipedia: OK to use a Wikipedia dump from before November 2013. There are monthly xml data dumps, going back to 2014 and before. Wikipedia historical data: https://lists.wikimedia.org/pipermail/xmldatadumps-l/
- DBPedia (version 3.9 or earlier): The datasets were extracted from Wikipedia dumps generated in late March / early April 2013. A Wikidata cross-language link dump from June 2013 was used to interconnect concepts between languages.

- YAGO 2.5.3 (wikipedia dump 2012-12-01). More information is available here: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/archive/