The TimeBank Corpus

Version: 1.0

Release Date: July 21, 2002

Authors: Beth Sundheim, Dragomir Radev

TERQAS Corpus Working Group Members: Patrick Hanks, Lisa Ferro, Dragomir Radev, Roser Saurí, Beth Sundheim, James Pustejovsky.

Introduction

This is general information about the document corpora being used for TERQAS. Information on the question corpora will be found separately.

Corpora to be Annotated

Annotation corpora are divided into training, devtest and evaltest. Training sets are on NRRC host computer. Devtest and Evaltest data are in password- protected zip files on the NRRC host computer. Contact Beth to receive password if you need the files in order to do annotation or testing.

Two subsets of training sets have been identified for early annotation:

1. A set of 6 articles, 3 from ACE and 3 from DUC, were annotated by multiple TERQAS participants, with an aim to revealing significant issues with the annotation guidelines. The annotation of those 6 articles is being perfected through an adjudication process to serve as an initial TimeML gold standard.
2. A set of 44 relatively short articles are being annotated by "serious" TERQAS annotators, led by Roser Sauri and Andy See at Brandeis. This set includes 22 ACE articles (newswire and broadcast news) and 22 Propbank articles. The goal is to complete single annotation on as many of these documents as possible by the end of the workshop, preferably with some adjudication and correction done by Roser and Andy.

In addition to the total of 50 articles described above, another 250 articles have been identified for annotation by the end of September. These articles include ones from ACE, DUC and Propbank. A subset of 24 relatively short articles will be annotated by multiple TERQAS annotators, and an assessment made of interannotator agreement, using comparison and scoring software from Mitre. This set includes 5 ACE newswire, 5 ACE broadcast news, 2 DUC "Sununu", 3 DUC "Iraq", and 9 Propbank.

1. DUC (TIPSTER)

a. 3 Training clusters (all docs from each cluster)

d03 - biography (Sununu) -- total of 11 docs
d09 - single event (Iraq) -- total of 16 docs
d16 - sequence of events (earthquakes) -- total of 8 docs

b. 5 Devtest clusters (5 docs from each cluster)

d25 - biography
d40 - biography

d21 - opinion
d20 - sequence of events
d33 - single event

c. 5 Evaltest clusters (5 docs from each cluster)

d42 - biography
d47 - biography
d26 - opinion
d36 - sequence of events
d60 - single event

2. ACE (TDT2) : Jan-Jun 1998

These sets are available both with and without manually generated TIMEX2 tags.

a. Broadcast news (ABC, CNN, PRI, VOA)

100 articles (50 training, 25 devtest, 25 evaltest)

Method of selection: Removed sports and pop culture articles from ACE data set, plus most of the shortest (1K) files, leaving 100 docs. Identified general topic of each doc. Each of the three TERQAS data sets includes proportional number of docs from each source (ABC, CNN, PRI, VOA). The first *n* docs (in alpha order by filename) were selected as training. The rest were arbitrarily assigned to devtest or evaltest, but with some conscious balancing by topic (U.S. news vs. foreign news, for example).

b. Newswire (AP & NYT)

99 articles (49 training, 25 devtest, 25 evaltest)

Method of selection: Removed all sports stories, leaving approximately 100 docs. Every second one was assigned to training; the remaining were alternately assigned to devtest and evaltest.

3. Propbank (from Treebank2)

216 WSJ articles (166 training, 25 devtest, 25 evaltest)

The Propbank/Treebank2 documents lack header information, including any reference time for the document. We tracked down the original documents from a Tipster CD, and assigned them file names that match the Propbank/Treebank2 file names, but with ".orig" (original) appended. These original files will be used for TERQAS annotation. Due to time constraints, a rigorous procedure was not used to select the articles. Most of the first ones to be annotated are taken from the shortest of the first 100 that were matched. The rest also come from the top of the list of files that were matched, but with less regard for length.

Reference Corpora (Corpora that are good to have around)

1. TIPSTER : These are the DUC clusters that are not included in the corpora to be annotated.

2. TDT2 : This is the whole corpus, including the small portions that are in the corpora to be annotated.

3. TREEBANK : Right now, all docs are included, including the Propbank articles that are in the corpora to be annotated.

4. PROPBANK : This excludes the portions that are in the corpora to be annotated.

5. REUTERS-21578

6. AP (HANKS)

7. NAMTC (LDC) -- The North American News Text Corpus is a collection of journalistic text in English from newswire and newspaper sources in the United States. The sources and time periods covered by this collection are as follows: LA Times & Washington Post (May 1994 - August 1997), NY Times News Syndicate (July 1994 - December 1996), Reuters News Service, general and financial (April 1994 - December 1996), WSJ (July 1994 : December 1996)

8. ProMed

9. ENTHUSIAST dialogue corpus (appointment scheduling dialogues), with TIMEX2 tags

10. BNC (British National Corpus)

Basic Paths to the Data

All corpora are on the NRRC host at Mitre. The paths originate with /workshops/terqas/data/doc-corpora/.

1. The first 6 training docs:

.../doc-corpora-training/target-100-docs/first-6-docs/

2. The rest of the first set of 50 training docs:

.../doc-corpora-training/ACE-broadcast-news-training/for-group-A-roser-22/
.../doc-corpora-training/ACE-broadcast-news-training/for-group-B-andys-22/

.../doc-corpora-training/ACE-newswire-training/for-group-A-roser-22/
.../doc-corpora-training/ACE-newswire-training/for-group-B-andys-22/

3. The training set of 24 docs for use in interannotator study:

.../doc-corpora-training/target-100-docs/interannotator-study-docs/

4. The rest of the training docs:

.../doc-corpora-training/new-target-for-sept/ace-bnews/
.../doc-corpora-training/new-target-for-sept/ace-nwire/
.../doc-corpora-training/new-target-for-sept/duc03a/
.../doc-corpora-training/new-target-for-sept/duc09b/
.../doc-corpora-training/new-target-for-sept/duc16c/
.../doc-corpora-training/new-target-for-sept/propbank/

5. The devtest docs: .../doc-corpora-devtest/

6. The evaltest docs: .../doc-corpora-evaltest/

7. The reference corpora: .../doc-corpora-reference/ (all subdirectories)