The TERN 2004 Evaluation Plan *(DRAFT)*:
Time Expression Recognition and Normalization

1 INTRODUCTION

The objective of the Automatic Content Extraction (ACE) program is to develop natural language processing technology to support automatic understanding of textual data. This includes classification, filtering, and selection based on the meaning conveyed by the data. Thus, the ACE program requires the development of technologies that automatically detect and characterize this meaning.

The Time Expression Recognition and Normalization (TERN) evaluation is based on work that began in 1999 to establish a set of useful guidelines for text annotation and data interchange[1]. The guidelines define a tag called TIMEX2, including attributes for expressing the normalized, intended meaning or value of a broad range of temporal expressions. The work extends the Message Understanding Conferences' definition of the TIMEX category of named entity to include a broader variety of expressions and to offer a normalization scheme.

TIMEX2 is influencing the definition of ACE tasks, in which temporal expressions covered by TIMEX2 will contribute to filling temporal attributes for extracted relations and events. Thus, the production of TIMEX2 annotations is viewed as an ACE component technology. The TERN evaluation is open to sites that want to develop this type of component technology. The evaluation will be offered in both English and Chinese.

The year 2004 is the first year that TERN has been offered. Therefore, the task and the evaluation procedures are expected to be subject to substantial revision in the course of the effort. In particular, the Chinese task guidelines will be improved in light of annotation experience and feedback from evaluation participants.

2 TASK DEFINITION

The TIMEX2 task requires that temporal expressions mentioned in the source data be detected and normalized according to the *2003 Standard for the Annotation of Temporal Expressions* by Ferro et al. Guidelines that are particular to Chinese are documented (with extensive examples) in a draft supplement that has been prepared as a separate document. Participants are expected to follow the guidelines as updated and posted on the TIMEX2 website.[2]

Temporal expressions to be marked include both absolute expressions ("July 17, 1999", "12:00", "the summer of '69") and relative expressions ("yesterday," "last week," "the next millennium"). Also markable are durations ("one-hour", "two weeks"), event-anchored expressions ("two days before departure"), and sets of times ("every week"). The degree to which these expressions can be normalized given the current TIMEX2 guidelines varies according to the type and specificity of the expression.

The relevance of phrases as TIMEX2 expressions is based on the notion of lexical triggers. Table 1 below is excerpted from section 3.1 of the English guidelines. Rules for determining the extent of a relevant expression are defined in section 5 of the guidelines. Table 2 below, which is from section 4.1, outlines the normalization attributes that are currently defined for TIMEX2. The normalization aspect of the TERN evaluation will include all the defined attributes except NON_SPECIFIC (and COMMENT, which is for use only by human annotators).

Table 1: Sample Lexical Triggers and Non-Triggers

Part of Speech	Lexical Triggers	Non-Triggers
Noun	minute, afternoon, midnight, day, night, weekend, month, summer, season, quarter, year, decade, century, millennium, era, semester, [the] future, [the] past, time, period, point	instant, jiffy, episode, occasion, tenure, timetable, reign
Proper name	(unique identifier for temporally-defined events:) Monday, January, New Year's Eve, Washington's Birthday, Solstice	
Specialized time patterns	8:00, 12/2/00, 1994, 1960s	
Adjective	recent, former, current, future, past, daily, monthly, biannual, semiannual, daytime, daylong, onetime, ago, preseason	early, ahead, next, subsequent, frequent, perpetual, later, contemporary, simultaneous, preceding, previous, existing
Adverb	currently, lately, hourly, daily, monthly, ago	earlier, immediately, instantly, forthwith, meanwhile, heretofore, previously, next, beforehand, following, later soon, sooner, shortly, eventually, occasionally, once, still, again

Table 2: TIMEX2 Tag Attributes

Attribute	Function	Example
VAL	Contains a form of the date/time.	VAL="1964-10-16"
MOD	Captures temporal modifiers.	MOD="APPROX"
ANCHOR_VAL	Contains a normalized form of an anchoring date/time.	ANCHOR_VAL="1964-10-16"
ANCHOR_DIR	Captures the relative direction/ orientation between VAL and ANCHOR_VAL.	ANCHOR_DIR="BEFORE"
SET	Identifies expressions denoting sets of times.	SET="YES"
NON_SPECIFIC	Identifies non-specific expressions.	NON_SPECIFIC="YES"
COMMENT	Contains any comments the annotator wants to add.	COMMENT="context garbled"

[1] This work was funded by the DARPA TIDES (Translingual Information Detection, Extraction and Summarization) program and was carried out collaboratively by L. Ferro, I. Mani and G. Wilson (MITRE Corp.), L. Gerber (Northrop Grumman Information Technology), and B. Sundheim (SPAWAR Systems Center).

[2] http://timex2.mitre.org

3 EVALUATION

3.1 EVALUATION CONDITIONS

Two evaluation conditions will be supported for the 2004 evaluation. When participants submit their system output to NIST for scoring[3], they must identify which of these two conditions they wish to be evaluated on.

- **Recognition + Normalization.** For this condition, which is the complete processing condition, systems are evaluated on their performance on detection, bracketing, and normalization.
- **Recognition only.** For this condition, systems are evaluated only on their performance on detection and bracketing.

Both of the above conditions are also subject to the condition that each document is processed independently of other documents. Systems are expected to take advantage of the publication/broadcast date found in the document (header, body and/or trailer) to use as a reference date for normalizing TIMEX2s in the document, and to take advantage of other temporal clues in the document to further aid in normalization, as required.

3.1.1 DETECTION (RECOGNITION)

The basic prerequisite for carrying out the full TERN task is to detect relevant time expressions that are mentioned in the source language data. The scoring of detection is very generous for the 2004 evaluation; minimal overlap in the extent of the reference and the system output tags is required for detection. As long as there is at least one overlapping character, the tags will be aligned, unless there is a better alignment possible in cases involving embedded tags. Any alignment of a system output tag is scored as a correct detection.

All types of TIMEX2s are weighted equally in the scoring in the 2004 evaluation. In the future, it may prove advisable to give greater weight to some expressions based on their relative frequency of occurrence, their amenability to full normalization, or some other criterion.

3.1.2 BRACKETING (EXTENT RECOGNITION)

Extent recognition measures the ability of the system to correctly determine the extent of the TIMEX2s, *for all correctly detected expressions*. In the 2004 evaluation, the extent of the reference and system output tags must match exactly for the extent of a system output tag to be scored as correct.

3.1.3 ATTRIBUTE VALUE ASSIGNMENT (NORMALIZATION)

Attribute value assignment measures the ability of the system to correctly assign the normalization attribute values of the TIMEX2s, *for all correctly detected expressions*. These attributes include VAL, MOD, SET, ANCHOR_VAL and ANCHOR_DIR. Scores will be reported across all these attributes and separately for each attribute.

3.2 SCORING CATEGORIES AND METRICS

The result of aligning one reference item with one system output item (or of leaving one reference item or system output item unaligned) yields four kinds of tallies:

- **CORR (Correct):** The two items under comparison are considered identical.
- **INCO (Incorrect):** The two items under comparison are not identical.
- **MISS (Missing):** A reference has no system output aligned with it.
- **SPUR (Spurious):** A system output has no reference aligned with it.

Given a set of tallies, there are several values calculated in the alignment and final scoring.

- **POSS (Possible):** The number of items in the reference that contribute to the final score:

$$POSS = CORR + INCO + MISS$$

- **ACT (Actual):** The number of items in the system output:

$$ACT = CORR + INCO + SPUR$$

- **REC (Recall):** The quantity of reference items that is present in the system output:

$$REC = CORR/POSS$$

- **PREC (Precision):** The quantity of system output items that is actually in the reference:

$$PREC = CORR / ACT$$

Van Rijsbergen's **F-measure** is used to combine recall and precision measures into one measure:

$$F = 2((\beta)+1.0)*P*R / 2((\beta)*P)+R$$

, where beta is the relative weight of precision and recall. In the TERN 2004 evaluation, they are weighted equally.

Measures of **Undergeneration**, **Overgeneration**, **Substitution**, and **Error** (per system output) are also calculated from the tallies:

$$UND = MISS / POSS$$

$$OVG = SPUR / ACT$$

$$SUB = INCO / CORR + INCO$$

$$ERR = INCO + SPUR + MISS / CORR + INCO + SPUR + MISS$$

The above *error-based metrics* provide more specific insight into the sources of error than do the *recall and precision-based (RP) metrics*, and experience from past named entity tasks suggests that they offer a more reliable basis on which to distinguish systems that are performing a task at a high level.

4 CORPUS SUPPORT

Corpus support consists of providing data to train, develop and evaluate TERN systems. Corpora are assembled from a variety of sources selected from broadcast news programs, newspapers, and newswire reports.

The sources for the overall ACE corpus for English for 2004 include broadcast news and newswire from the Topic Detection and Tracking program (TDT4), translated documents from Chinese Treebank and the Arabic Treebank, and transcripts of telephone speech (Fisher Conversational Telephone speech collection). The overall Chinese corpus is from TDT4 and the Chinese Treebank. The TERN corpora are drawn from the same pool of ACE documents, but there may be little overlap between the TERN and ACE evaluations in terms of the specific documents selected for annotation.

4.1 THE 2004 DEVELOPMENT CORPUS

Two versions of each document are provided:

- DOCID.sgm: the original document without TIMEX2 annotation. The document contains a series of SGML tags. All the text between the begin doc tag <DOC> and end doc tag </DOC> is to be evaluated.

[3] The TIMEX2 evaluation software that is being used for TERN is available for download from the TIMEX2 web site.

- DOCID.tmx.sgml: this version of the document is the answer key for TERN. It is an in-line SGML document. This is also the required form for TERN system output.

Corpus statistics are listed in Table 3:

Table 3: 2004 TERN system development corpus statistics

	2004 corpora	Prior-years ACE corpora
English Resources	~150K words ?? docs	2002 ACE evalset 67K words 180 docs
		2003 ACE development ~100K words ~250 docs
Chinese Resources	~70K words[4] ~?? docs	2003 ACE development ~30K words ~90 docs

It is anticipated that all the English documents will be doubly annotated; time permitting, more documents will be singly annotated as additional training data. For Chinese, it is anticipated that 50K words will be doubly annotated, and the rest may be only singly annotated.

4.2 The 2004 Evaluation Corpus

The 2004 evaluation corpus will be 50K words for English and 50K words for Chinese. All documents to be included in the 2004 evaluation set will be doubly annotated.

5 Evaluation Schedule

Date (2004)	Event
12 April - 1 July	Increments of training data released by LDC to sites that respond to the CFP
30 June	Last day to register as evaluation participant
2-13 August	Evaluation text corpus (English/Chinese) available to participants. Participants must return results (system output) to NIST within 24 hours of receipt of the evaluation corpus.
13 August	Last day for participants to submit official results to NIST
20 September	Evaluation scores released by NIST to participants
23 September	One-day meeting in conjunction with ACE workshop

6 Submitting Results To NIST

6.1 Rules

No changes to the system are allowed once the evaluation data are released. Adaptive systems may of course change themselves in response to the source data that they process, but the test data is not to be examined by humans and ***no human intervention is allowed*** prior to the submission of the site's test results to NIST.[5]

6.2 Packaging Results

Due to the short period of time between the date system output files are due at NIST and the beginning of the workshop, it will expedite scoring and releasing of results if participants follow the outlined procedure for submitting results. This will enable quick unpacking and scoring of several site submission files with minimum human intervention.

STEP1: Create a top-level directory for each of the ***languages*** attempted:

Example: $> mkdir chinese english

STEP2: Create a subdirectory that identifies whether you elected to do the complete TERN task or just the recognition portion:

Example: $> mkdir english/tern-full english/tern-recog

STEP3: In each of these subdirectories make one directory for each system submitted (choose a name that identifies your site, BBN, SHEF, SRI…):

Example: $> mkdir english/tern-full/NIST1_primary

Example: $> mkdir english/tern-full/NIST2_contrastive1

Example: $> mkdir chinese/tern-recog/NIST1_primary

Example: $> mkdir chinese/tern-recog/NIST1_contrastive1

STEP4: Deposit all system output files in the appropriate system directory. Include a system description in this same directory (see section 6.3 for details).

STEP5: Create a compressed tar file of your results and transfer them to NIST by FTP (ftp://ijaguar.ncsl.nist.gov/incoming). After successful transmission send e-mail to mark.przybocki@nist.gov identifying the name of the file submitted. Alternatively you may send the compressed tar file directly to mark.przybocki@nist.gov.

6.3 System Descriptions

A valuable tool in discovering strengths and weakness of different algorithmic approaches is the use of system descriptions. Each participant should prepare a brief description of each system submitted and include the description with their submission of results. These system descriptions will be distributed to each participant at the time NIST releases of results.

A typical system description should include:

- A description of the algorithmic approach taken in the development of each system that was submitted for evaluation
- System resources required to process the test set
- A sense of how long it took to run the test (as if running on a single computer)
- A clear description of the primary system and all contrastive systems.

[4] Two Chinese characters count as one word.

[5] It sometimes happens that a system bug is discovered during the course of processing the test data. In such a case, repairs may be possible that allow a more accurate assessment of the underlying performance of a system. If this happens, modified results may be accepted, provided that an explanation of the modification is provided and provided that the original results are also be submitted.