MUC-7 Named Entity Task Definition

Version 3.5
17 September 1997

Nancy Chinchor (chinchor@gso.saic.com)

1. INTRODUCTION

1.1 Scope

The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages).

For many text processing systems, such identifiers are recognized primarily using local pattern-matching techniques. The TEI (Text Encoding Initiative) Guidelines for Electronic Text Encoding and Interchange cover such identifiers (plus abbreviations) together in section 6.4 and explain that the identifiers comprise "textual features which it is often convenient to distinguish from their surrounding text. Names, dates and numbers are likely to be of particular importance to the scholar treating a text as source for a database; distinguishing such items from the surrounding text is however equally important to the scholar primarily interested in lexis."

The task is to identify all instances of the three types of expressions in each text in the test set and to subcategorize the expressions. The original texts contain some SGML tags already; the Named Entity task is to be performed within the text delimited by the SLUG, DATE, NWORDS, PREAMBLE, TEXT, and TRAILER tags.

The system must produce a single, unambiguous output for any relevant string in the text; thus, this evaluation is not based on a view of a pipelined system architecture in which Named Entity recognition would be completely handled as a preprocess to sentence and discourse analysis. The task requires that the system recognize what a string represents, not just its superficial appearance. Sometimes, the right answer is superficially apparent, as in the case of most, if not all, NUMEX expressions, and can be obtained by local pattern-matching techniques. In other cases, the right answer is not superficially apparent, as when a single capitalized word could represent the name of a location, person, or organization, and the answer may have to be obtained using techniques that draw information from a larger context or from reference lists.

The three subtasks correspond to three SGML tag elements: ENAMEX, TIMEX, and NUMEX. The subcategorization is captured by a SGML tag attribute called TYPE, which is defined to have a different set of possible values for each tag element. The markup is described in section 2, below.

1.2 Performance Evaluation

Scoring of this task will be done using the same kinds of metrics that are used for scoring template-filling (information extraction) tasks. For specific information on the scoring, refer to "MUC-7 Scoring System User's Manual," prepared for MUC-7 by SAIC.

Cumulative scores will be generated at several levels of description of the task, e.g.,

- across subtasks,
- for each subtask,

- for the subcategorization aspect of each subtask,
- for each part of the article that is included in the task (<SLUG>, <DATE>, <NWORDS>, <PREAMBLE>, <TEXT>, <TRAILER>).

2. TASK OVERVIEW

2.1 Markup Description

The output of the systems to be evaluated will be in the form of SGML text markup. The only insertions allowed during tagging are tags enclosed in angled brackets. No extra whitespace or carriage returns are to be inserted; otherwise, the offset count would change, which would adversely affect scoring.

The markup will have the following form:

<ELEMENT-NAME ATTR-NAME="ATTR-VALUE" ...>*text-string*</ELEMENT-NAME>

Example:

<ENAMEX TYPE="ORGANIZATION">*Taga Co.*</ENAMEX>

The markup is defined in SGML Document Type Descriptions (DTDs), written for MUC-7 use and maintained by personnel at SAIC. The DTDs enable annotators and system developers to use SGML validation tools to check the correctness of the SGML-tagged texts produced by the annotator or the system. The validation tools are available to MUC-7 participants in the file called muc7-sgml-tools and in the form of the scorer's parser both available via anonymous ftp from ftp.muc.saic.com (or online.muc.saic.com) in the under the MUC subdirectory.

Annotators are using a software tool provided for MUC-7 and MET-2 by SRA Corporation to assist in generating the answer keys to be used for system training and testing.

2.2 Named Entities (ENAMEX tag element)

This subtask is limited to proper names, acronyms, and perhaps miscellaneous other unique identifiers, which are categorized via the TYPE attribute as follows:

ORGANIZATION: named corporate, governmental, or other organizational entity

PERSON: named person or family

LOCATION: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)

2.3 Temporal Expressions (TIMEX tag element)

This subtask is for "absolute" and "relative" temporal expressions only; explanation is provided in appendix B. The tagged tokens are categorized only via the TYPE attribute as follows:

DATE: complete or partial date expression

TIME: complete or partial expression of time of day

The TYPE attribute does not distinguish "absolute" and "relative" temporal expressions from each other.

2.4 Number Expressions (NUMEX tag element)

This subtask is for two useful types of numeric expressions, monetary expressions and percentages. The numbers may be expressed in either numeric or alphabetic form.

The task covers the complete expression, which is categorized via the TYPE attribute as follows:

MONEY: monetary expression

PERCENT: percentage

3. NOTATION RESERVED FOR USE IN THE ANSWER KEYS

3.1 Expressing Alternative Attribute Values

A vertical bar is being used to separate alternative TYPE attribute values in the answer key.

Alternative values will be given when the annotator does not have enough information to make a unique categorization, even considering the context and the annotator's knowledge of the world.

3.2 Expressing Optional Markup (STATUS Attribute)

When it is not certain that a string should be marked up, the annotator will include the STATUS attribute in the markup to indicate that the markup is optional. The only value of the STATUS attribute is "OPT." Examples of its possible use can be found in the appendices, such as in appendix B (holiday names).

3.3 Expressing Alternative or Minimum String Boundaries (ALT or MIN Attribute)

The ALT or MIN attribute will be used when the tagged string contains one or more substrings that should be considered correct for the purposes of scoring the system response. Certain premodifiers ("a," "an," and "the") are automatically ignored by the scoring program (via the "configuration" file) and thus do not need to be specially marked in the key.

ALT was the original term used for this attribute in past NE definitions and scorers, but the markup tool calls it MIN. Either term is appropriate because the alternative string is smaller than the tagged string in all cases.

The ALT or MIN attribute will be used sparingly. A possible TIMEX example is shown below.

"all of 1987"

<TIMEX TYPE="DATE" ALT="1987">*all of 1987*</TIMEX>

4. GUIDELINES FOR MARKUP OF EXCEPTIONAL CONSTRUCTIONS

4.1 Conjunction and Elision in Multi-name, Multi-modifier, and Numeric Range Expressions

Conjoined named entities in general are marked separately except for those in the following categories. All cases in these categories are tagged as *single* expressions.

4.1.1 Multi-name (or multi-number) expressions

A conjoined multi-name expression, in which there is elision of the head of one conjunct, should be marked up as a single expression.

"North and South America"

<ENAMEX TYPE="LOCATION">*North and South America*</ENAMEX>

A similar case occurs with elision in multi-number expressions:

"10- and 20-dollar bills" (i.e. 10-dollar bills and 20-dollar bills)

<NUMEX TYPE="MONEY">*10- and 20-dollar*</NUMEX> *bills*

4.1.2 Multi-modifier expressions

A single-name expression containing conjoined modifiers with no elision also should be marked up as a single expression.

"U.S. Fish and Wildlife Service" (which does NOT mean two entities, i.e. "the U.S. Fish Service and the U.S. Wildlife Service")

<ENAMEX TYPE="ORGANIZATION">*U.S. Fish and Wildlife Service*</ENAMEX>

4.1.3 Numeric range expressions

The subparts of time, date, money, and percentage range expressions should be marked up as parts of a single expression, even if there is no elision of the numeric "units".

"175 to 180 million Canadian dollars"

<NUMEX TYPE="MONEY">*175 to 180 million Canadian dollars*</NUMEX>

"the 1986-87 academic year"

the <TIMEX TYPE="DATE">*1986-87 academic year*</TIMEX>

"from 1990 through 1992”

<TIMEX TYPE="DATE">*from 1990 through 1992*</TIMEX>

4.2 Effects of Tokenization Conventions

The systems must incorporate certain tokenization conventions. These conventions are contained in a separate document titled "Tokenization Rules."

The tokenization conventions for MUC-7 have an impact on the boundaries of the strings to be tagged. For example, the conventions call for treating possessive forms, e.g., "California's," as multiple tokens, unless there is a name such as "McDonald's [burger company]" that is inherently possessive. See the separate documentation titled "Tokenization Rules" for further information and examples.

In various sources there are some special characters used that end up being within the marked string because they are contiguous, but a reader will ignore them. For example, in the Wall Street Journal an @ appears at the beginning of some lines in the headline. In the New York Times News Service articles there are some codes such as "&MD;" which appear and are not always separated by white space from their environment. These will generally be marked up and the scorer will not be able to delete them because of the segmentation problem. Although infrequent, the rule we follow will be to include them if they are string-internal and to exclude them otherwise. It is unlikely that scores will be seriously affected so the scorer will not specially treat these codes.

4.3 Nested Expressions

No nested expressions will be marked. Even in cases where LOCATION (ENAMEX) expressions occur within TIMEX and NUMEX expressions, they are not to be tagged. Also, entity names that appear within ENAMEX tags are *not* to be tagged.

"8:24 a.m. Chicago time"

<TIMEX TYPE="TIME">*8:24 a.m. Chicago time*</TIMEX>

"U.S. $10 million"

<NUMEX TYPE="MONEY">*U.S. $10 million*</NUMEX>

"the U.S. Customs Service"

the <ENAMEX TYPE="ORGANIZATION">*U.S. Customs Service*</ENAMEX>

5. APPENDICES

5.1 Naming Conventions for Section Headlines in the Appendices

1. An "Entity (/Temporal/Numeric)-Expression" identifies something that MUST be tagged;

2. An "Entity-String" identifies that something that MIGHT be tagged, but not in the context described;

3. A "Non-entity" identifies something that is NEVER tagged, according to current MUC/MET conventions.

APPENDIX A. ENAMEX: SPECIFIC GUIDELINES

A.1 Guidelines That Pertain to All Three TYPEs (PERSON, LOCATION, and ORGANIZATION)

A.1.1 Entity-Expressions that Modify Non-entities

Entity names used as modifiers in complex NPs that are not proper names are to be tagged when it is clear to the annotator from context or the annotator's knowledge of the world that the name is that of an organization, person, or location.

"Bridgestone profits"

<ENAMEX TYPE="ORGANIZATION">*Bridgestone*</ENAMEX> *profits*

"the Clinton government"

the <ENAMEX TYPE="PERSON">*Clinton*</ENAMEX> *government*

"Treasury bonds and securities"

<ENAMEX TYPE="ORGANIZATION">*Treasury*</ENAMEX> *bonds and securities*

"U.S. exporters"

<ENAMEX TYPE="LOCATION">*U.S*.</ENAMEX> *exporters*

"Macintosh computers"

<ENAMEX TYPE="ORGANIZATION">*Macintosh*</ENAMEX> *computers*

"West Texas Intermediate crude"

<ENAMEX TYPE="ORGANIZATION">*West Texas Intermediate*</ENAMEX> *crude*

Note that uncapitalized, common-noun designators such as "division" in the phrase "Chrysler division" are *not* considered part of an entity name.

"Chrysler division"

<ENAMEX TYPE="ORGANIZATION">*Chrysler*</ENAMEX> *division*

"the Kennedy family"

the <ENAMEX TYPE="PERSON">*Kennedy*</ENAMEX> *family*

A.1.2 Entity-Expressions that Modify Titles

In addition, entity names modifying person identifiers (where person identifier= title/role and/or name) are to be tagged.

"Mips Vice President John Hime" [Mips is the name of a computer company]

<ENAMEX TYPE="ORGANIZATION">*Mips*</ENAMEX> *Vice President* <ENAMEX TYPE="PERSON">*John Hime*</ENAMEX>

"Treasury Secretary"

<ENAMEX TYPE="ORGANIZATION">*Treasury*</ENAMEX> *Secretary*

"the U.S. Vice President"

the <ENAMEX TYPE="LOCATION">*U.S*.</ENAMEX> *Vice President*

A.1.3 Entity-Strings Embedded in Entity-Expressions

In some cases, multi-word strings that are proper names will contain entity name substrings; such strings are not decomposable; therefore, the substrings are not to be tagged. (See A.1.2 re special cases involving prenominal modifiers of person identifiers.)

"Arthur Anderson Consulting"

<ENAMEX TYPE="ORGANIZATION">*Arthur Anderson Consulting*</ENAMEX>

[no markup for "Arthur Anderson" alone]

"Boston Chicken Corp."

<ENAMEX TYPE="ORGANIZATION">*Boston Chicken Corp*.</ENAMEX>

[no markup for "Boston" alone]

"U.S. Fish and Wildlife Service"

<ENAMEX TYPE="ORGANIZATION">*U.S. Fish and Wildlife Service*</ENAMEX>

[no markup for "U.S." alone]

"Northern California"

<ENAMEX TYPE="LOCATION">*Northern California*</ENAMEX>

[no markup for "California" alone]

"West Texas"

<ENAMEX TYPE="LOCATION">*West Texas*</ENAMEX>

[no markup for "Texas" alone]

A.1.4 Non-Entities Modified by Entity Expressions

(subsumed under A.1.1)

A.1.5 Entity-Expressions that "Possess" Other Entity-Expressions

In a possessive construction, the possessor and possessed ENAMEX substrings should be tagged separately.

"Temple University's Graduate School of Business"

<ENAMEX TYPE="ORGANIZATION">*Temple University*</ENAMEX>*'s* <ENAMEX TYPE="ORGANIZATION">*Graduate School of Business*</ENAMEX>

"Shearson Lehman Hutton's OTC department"

<ENAMEX TYPE="ORGANIZATION">*Shearson Lehman Hutton*</ENAMEX>*'s* <ENAMEX TYPE="ORGANIZATION">OTC</ENAMEX> *department*

"California's Silicon Valley"

<ENAMEX TYPE="LOCATION">*California*</ENAMEX>*'s* <ENAMEX TYPE="LOCATION">*Silicon Valley*</ENAMEX>

"Canada's Parliament"

<ENAMEX TYPE="LOCATION">*Canada*</ENAMEX>*'s* <ENAMEX TYPE="ORGANIZATION">*Parliament*</ENAMEX>

A.1.6 Entity-Expression Aliases

Aliases for entities are to be tagged. Taggable aliases will include the following forms of entity names:

– Acronyms, formed from the initial letter(s) or syllable(s) of successive or major parts of a compound term, for example:

"IBM" [alias for International Business Machines Corp.]

<ENAMEX TYPE="ORGANIZATION">*IBM*</ENAMEX>

"PACTEL" [alias for Pacific Telesys, i.e. Pacific Telephone Systems]

<ENAMEX TYPE="ORGANIZATION">*PACTEL*</ENAMEX>

– Nicknames, examples follow:

"Big Blue" [alias for International Business Machines Corp.]

<ENAMEX TYPE="ORGANIZATION">*Big Blue*</ENAMEX>

"Big Board" [alias for New York Stock Exchange]

<ENAMEX TYPE="ORGANIZATION">*Big Board*</ENAMEX>

"Mr. Fix-It" [nickname for candidate for head of the CIA]

Mr. <ENAMEX TYPE="PERSON">*Fix-It*</ENAMEX>

"the Big Apple" [nickname for New York City]

<ENAMEX TYPE="LOCATION">*the Big Apple*</ENAMEX>

- Truncated Names, provided that the resulting form is clearly a proper noun referring to a specific entity, for example in:

  "Red Sox" [alias for the Boston Red Sox]

  <ENAMEX TYPE="ORGANIZATION">*Red Sox*</ENAMEX>

  "Sears" [alias for Sears Roebuck and Co.]

  <ENAMEX TYPE="ORGANIZATION">*Sears*</ENAMEX>

- Certain metonyms, herein designated "proper" metonyms, which chiefly include references to an organization based on the name of a unique structure or facility in which the organization holds office. The association between the name and the organization should be idiosyncratic enough to justify its inclusion in the dictionary definition of the term (in contrast with "common" metonyms, discussed below), as a kind of nickname for the organization. Some examples follow.

  "The White House announced ..." [alias for the U.S.president's executive organization]

  The <ENAMEX TYPE="ORGANIZATION">*White House*</ENAMEX> *announced ...*

  "The Pentagon announced..."

  The <ENAMEX TYPE="ORGANIZATION">*Pentagon*</ENAMEX> *announced ...*

  Taggable aliases will NOT include the following forms of entity names:

- Common nouns, including pronouns, used in anaphoric reference to taggable entity names, such as

  "IBM announced that the company would lay off ..."

  [no markup for "the company"]

- Aliases that refer to broad industrial sectors, political power centers, etc., rather than to specific organizations. For example, do not tag "Wall Street" as an alias for the U.S. stock market, "Japan Incorporated" as an alias for Japanese Industries, "Uncle Sam" and "Washington" as aliases for the U.S. government, or "Capitol Hill" as an alias for the Congress, since these do not refer to specific organizations. The "Ivy League" refers to a specific set of universities, but does not seem to be a specific organization in its own right. Similarly, the "Axis" (WWII Germany-Japan-Italy) and the "Iron Curtain countries" are aliases for finite sets of entities, but not for specific organizations with corporation-like infrastructures.

- Metonyms, herein designated "common" metonyms, that reference political, military, athletic, and other organizations by the name of a city, country, or other associated location. In these cases, the association between the name's semantic type and the organization is sufficiently predictable and non-idiosyncratic as to preclude a dictionary gloss; hence the name should be tagged as a LOCATION. Some examples of "common" metonyms follow.

  "Germany invaded Poland in 1939."

  <ENAMEX TYPE="LOCATION">*GERMANY*</ENAMEX> *invaded ...*

"Baltimore defeated the Yankees by a score of 4 to 3.

<ENAMEX TYPE="LOCATION">*Baltimore*</ENAMEX> *defeated the* <ENAMEX TYPE="ORGANIZATION">*Yankee*s</ENAMEX> ...

Note that links from LOCATION-tagged names to organizations (e.g. "Baltimore" to the "Baltimore Orioles" baseball team) are left to occur, along with anaphora-resolution, at a processing level higher than Named Entity tagging.

A.1.6.1 Quotation Marks Around an Alias

Quotes are included in the tag if they appear within a person's name.

"Vito "The Godfather" Corleone"

<ENAMEX TYPE="PERSON">*Vito "The Godfather" Corleone*</ENAMEX>

"Corleone, also known as "The Godfather", was the victim of a mob "hit"..."

...also known as <ENAMEX TYPE="PERSON">*"The Godfather"*</ENAMEX>*, was the...*

A.1.6.2 The Definite Article In an Alias

When a definite article is commonly associated with an alias, it also must be tagged.

<ENAMEX TYPE="PERSON">*The Godfather*</ENAMEX>

However, the scoring program ignores a certain list of premodifiers as specified in section 3.3 which may make the scoring in some of these cases more lenient than this rule implies. The scorer does *not* ignore those premodifiers within quotation marks such as inside the tags in A.1.6.1 above.

A.1.7 Miscellaneous Non-Entities

Miscellaneous types of proper names that are *not* to be tagged as ENAMEX include artifacts, other products, and plural names that do not identify a single, unique entity. (For information on the treatment of facilities, see section A.2, below.)

"Wall Street Journal"

[no markup]

"the Campbell Soups of the world"

[no markup]

"Dow Jones Industrial Average"

[no markup, not even for "Dow Jones"]

Note that just as in A.1.1, entity names used as modifiers in complex NPs that are not to be marked as entities are to be tagged when it is clear to the annotator from context or the annotator's knowledge of the world that the name is that of an organization, person, or location. More specifically, cases where the manufacturer and the product are named, the manufacturer will be tagged. The product will not be tagged.

"Ford Taurus"

<ENAMEX TYPE="ORGANIZATION">*Ford*</ENAMEX> *Taurus*

A.2 Guidelines That Pertain Only to ORGANIZATION

A.2.1 Corporate Designators

Corporate designators such as "Co." are considered part of an organization name.

"Bridgestone Sports Co."

<ENAMEX TYPE="ORGANIZATION">*Bridgestone Sports Co*.</ENAMEX>

However, the scorer ignores corporate designators as listed in the configuration file, so the scoring is lenient in this respect. It is possible that at a later date only partial credit will be given if the existing corporate designator is not included in the markup.

A.2.2 Miscellaneous ORG-type Entity-Expressions

Miscellaneous types of proper names that are to be tagged as ORGANIZATION include stock exchanges, multinational organizations, political parties, orchestras, unions, non-generic governmental entity names such as "Congress" or "Chamber of Deputies", sports teams and armies (unless designated only by country names, which are tagged as LOCATION),

"NASDAQ"

<ENAMEX TYPE="ORGANIZATION">*NASDAQ*</ENAMEX>

[a stock exchange]

"European Community"

<ENAMEX TYPE="ORGANIZATION">*European Community*</ENAMEX>

"GOP presidential hopeful"

<ENAMEX TYPE="ORGANIZATION">*GOP*</ENAMEX> *presidential hopeful*

"Machinists union"

<ENAMEX TYPE="ORGANIZATION">*Machinists*</ENAMEX> *union*

"the mayor who built Candlestick Park for the Giants"

the mayor who built Candlestick Park for the<ENAMEX TYPE="ORGANIZATION">

Giants</ENAMEX>

[a sports team]

"In hockey action, Russia defeated France by a score of 7 to 3."

<ENAMEX TYPE="LOCATION">*Russia*</ENAMEX> *defeated* <ENAMEX TYPE="LOCATION">*France*</ENAMEX> ...

A.2.2.1 Articles Appearing with ORGANIZATION-type expressions

Articles appearing with ORGANIZATION-type entity expressions generally do not need to be tagged.

"the University of Chicago"

the <ENAMEX TYPE="ORGANIZATION">*University of Chicago*</ENAMEX>

The scorer does ignore a list of premodifiers that it is given in its configuration file because some of these articles can be included unconsciously during human markup and systems should not be penalized. The answer keys are not consistent with respect to including or excluding articles and the examples in this document reflect the inconsistency of humans. However, the scoring ignores these premodifiers.

A.2.2.2 Generic ORGANIZATION-like Non-entities

Generic entity names such as "the police" and "the government," are not to be tagged.

A.2.3 ORG-type Entity-Expressions Easily Confused With Non-Entities

Miscellaneous types of proper names referring to facilities (e.g., churches, embassies, factories, hospitals, hotels, museums, universities) will be tagged as ORGANIZATION.

"Finger Lakes Area Hospital Corp."

<ENAMEX TYPE="ORGANIZATION">*Finger Lakes Area Hospital Corp*.</ENAMEX>

"Four Seasons Hotels"

<ENAMEX TYPE="ORGANIZATION">*Four Seasons Hotels*</ENAMEX>

"Unification Church"

<ENAMEX TYPE="ORGANIZATION">*Unification Church*</ENAMEX>

"the White House"

the <ENAMEX TYPE="ORGANIZATION">*White House*</ENAMEX>

"Trinity Lutheran Church"

<ENAMEX TYPE="ORGANIZATION">*Trinity Lutheran Church*</ENAMEX>

"General Hospital"

<ENAMEX TYPE="ORGANIZATION">*General Hospital*</ENAMEX>

"The Empire State Building"

[no markup - this is a structure that houses many organizations]

A.2.4 Event-Type Non-Entities

Do not tag event names, even if they refer to events that occur on a regular basis and are associated with institutional structures. However, the institutional structures themselves - steering committees, etc. - should be tagged.

"the Pan-American Games"

[no mark-up]

"the U.S. Olympic Committee"

<ENAMEX TYPE="ORGANIZATION">*U.S. Olympic Committee*</ENAMEX>

A location name that is part of an event name should be tagged if the location name is not rendered in an adjectival form (as in "Pan-American", above).

"China Film Festival"

<ENAMEX TYPE="LOCATION">*China*</ENAMEX> *Film Festival*

Also note that certain designators such as "congress" or "conference" ("Congress of Deputies", "91st Congress", etc.) may refer to events, to organizations, or ambiguously to either one, depending on the context (compare "member of the 91st Congress" and "commencement of the 91st Congress".)

A.3 Guidelines That Pertain Only to PERSON

A.3.1 Titles vs. Generational Designators

Titles such as "Mr." and role names such as "President" are *not* considered part of a person name. However, appositives such as "Jr.", "Sr.", and "III" *are* considered part of a person name.

"Mr. Harry Schearer"

Mr. <ENAMEX TYPE="PERSON">*Harry Schearer*</ENAMEX>

"Secretary Robert Mosbacher"

Secretary <ENAMEX TYPE="PERSON">*Robert Mosbacher*</ENAMEX>

"John Doe, Jr."

<ENAMEX TYPE="PERSON">*John Doe, Jr*.</ENAMEX>

A.3.2 Family Entity-Expressions

Family names are to be tagged.

"the Kennedy family"

the <ENAMEX TYPE="PERSON">*Kennedy*</ENAMEX> *family*

"the Kennedys"

the <ENAMEX TYPE="PERSON">*Kennedys*</ENAMEX>

[alternate form of identification of the Kennedy family entity]

A.3.3 Miscellaneous Personal Non-Entities

Miscellaneous types of proper names that are not to be tagged as PERSON include political groups, laws named after people, diseases/prizes named after people, and saints (because removal of a saint's title leaves a non-unique name).

"The Republicans held a rally."

[no markup]

"the Gramm-Rudman amendment"

[no markup]

"Alzheimer's"

[no markup]

"the Nobel prize"

[no markup]

"St. Michael"

[no markup]

A.4 Guidelines That Pertain Only to LOCATION

Examples of place-related strings that are tagged as LOCATION include named heavenly bodies, continents, countries, provinces, counties, cities, regions, districts, towns, villages, neighborhoods, airports, highways, street names, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, fictional or mythical locations, and monumental structures, such as the Eiffel Tower and Washington Monument, that were built primarily as monuments.

"flew to Plymouth Airport"

flew to <ENAMEX TYPE="LOCATION">*Plymouth Airport*</ENAMEX>

"created a backup at O'Hare International Airport"

created a backup at <ENAMEX TYPE="LOCATION">*O'Hare International Airport*</ENAMEX>

If the name of the airport refers to the organization or business of the airport and not its location or facilities, then it is still marked as a LOCATION..

<ENAMEX TYPE="ORGANIZATION">*Massport*</ENAMEX>*, which owns and operates* <ENAMEX TYPE="LOCATION">*Logan*</ENAMEX>*, defended the attempts ...*

A.4.1 Embedded Locative Entity-Strings and Conjoined Locative Entity-Expressions

The phrase "of <place-name>" following an organization name may or may not be part of the organization name proper. The annotation in the answer key will follow these guidelines: (1) If there is a corporate designator, it marks the end of the organization name; (2) if there is no corporate designator, the "of <place-name>" is part of the organization name.

"Hyundai of Korea, Inc."

<ENAMEX TYPE="ORGANIZATION">*Hyundai of Korea, Inc*.</ENAMEX>

"Hyundai, Inc. of Korea"

<ENAMEX TYPE="ORGANIZATION">*Hyundai, Inc.*</ENAMEX> *of* <ENAMEX TYPE="LOCATION">*Korea*</ENAMEX>

"McDonald's of Japan"

<ENAMEX TYPE="ORGANIZATION">*McDonald's of Japan*</ENAMEX>

"University of California in Los Angeles"

<ENAMEX TYPE="ORGANIZATION">*University of California*</ENAMEX> *in* <ENAMEX TYPE="LOCATION">*Los Angeles*</ENAMEX>

[where the first tagged string functions as organization rather than facility (see A.2, above, re facilities) -- note also that the locative phrase "in Los Angeles" is not included inside the organization tag]

A.4.2 Locative Entity-Expressions Tagged in Succession

Compound expressions in which place names are separated by a comma are to be tagged as separate instances of LOCATION.

"Kaohsiung, Taiwan"

<ENAMEX TYPE = "LOCATION">*Kaohsiung*</ENAMEX>, <ENAMEX TYPE = "LOCATION">*Taiwan*</ENAMEX>

"Washington, D.C."

<ENAMEX TYPE = "LOCATION">*Washington*</ENAMEX>, <ENAMEX TYPE = "LOCATION"> *D.C.*</ENAMEX>

A.4.3 Miscellaneous Locative Non-Entities

Location-related strings that are not to be tagged as LOCATION include the adjectival forms of location names.

"American exporters"

[no markup]

A.4.4 Locative Designators and Specifiers

Designators that are integrally associated with a place name are to be tagged as part of the name. For example, include in the tagged string the word "River" in the name of a river, "Mountain" in the name of a mountain, "City" in the name of a city, etc., if such words are contained in the string.

"Mississippi River"

<ENAMEX TYPE="LOCATION">*Mississippi River*</ENAMEX>

(not: <ENAMEX TYPE="LOCATION">Mississippi</ENAMEX> River)

Determiners also are tagged if they are part of the place name.

"The Hague"

<ENAMEX TYPE="LOCATION">*The Hague*</ENAMEX>

However, the scoring program ignores a certain list of premodifiers as specified in section 3.3 which may make the scoring in some of these cases more lenient than this rule implies.

A.4.4.1 Locative Non-Entities: The Postposed Partitive Specifier

Do not include in the tagged string common noun phrases functioning as partitive-type locative specifiers directly after LOCATION names, such as:

"Mississippi River west bank" (west bank of the Mississippi River)

<ENAMEX TYPE="LOCATION">*Mississippi River*</ENAMEX> *west bank*

A.4.4.2 Exceptional Locative Specifiers Used as Entity Expressions

Note that, due to the political significance of the Jordan River's west bank, the term "West Bank" may, in the context of discussions about the Middle East, assume the status of a named entity expression. A similar example is the term "Left Bank" (of the Seine River) as a name for an area of Paris. Use context and world knowledge to determine whether such a term is being used as a specifying non-entity following a place name, or as an entity expression (a proper noun) representing a particular LOCATION.

A.4.5 Transnational and Subnational Region Names

A.4.5.1 Transnational Locative Entity Expressions

Tag names of continents ("Africa"), multi-country sub-continental regions ("Eastern Europe", "Sub-Saharan Africa"), and multi-country trans-continental regions ("the Middle East", "the Pacific Rim").

A.4.5.2 Subnational Region Names

Do not tag names of sub-national regions when referenced only by compass-point modifiers. Do not tag "the South" or the "mid-West", analogies to "the Middle East" notwithstanding, because, unlike the latter term, their referential value varies from country to country. For example,

"the Southwest region"

[no markup]

Do tag names of sub-national regions when they are associated with specific regions, if they are identifiable even when the name is disassociated from context. Examples include "the Ruhr", "the Auvergne", and "Amazonia". Note that these names generally straddle, or lie within, geo-political jurisdictions such as states or provinces.

A.4.6 Time and Space Modifiers of Locative Entity Expressions

Historic-time modifiers ("former", "present-day") and directional modifiers ("north", "south", "east", "west", "upper", "lower", and combinations thereof) are taggable only when they are intrinsic parts of a location's official name, as in "Upper Volta" or "North Dakota."

Do not include them in tagged expressions when used as ad hoc modifiers that are readily separable from the name.

"former Soviet Union"

former <ENAMEX TYPE="LOCATION">*Soviet Union*</ENAMEX>

"Gaul (present-day France)"

<ENAMEX TYPE="LOCATION">*Gaul*</ENAMEX> *(present-day* <ENAMEX TYPE="LOCATION">*France*</ENAMEX>*)*

"lower Manhattan"

lower <ENAMEX TYPE="LOCATION">*Manhattan*</ENAMEX>

Contrast "Premier of the former Soviet Union" and "formerly Premier of the Soviet Union"; "east Baltimore" and "eastern section of Baltimore"; and "Upper Volta" and "upper section of Volta" to see the separability of these modifiers.

APPENDIX B. TIMEX: SPECIFIC GUIDELINES

B.1 Introduction

Both "absolute" time expressions and certain "relative" time expressions, as specified below (B.1.2), are to be tagged in MUC-7. Note that the tag itself does not differentiate between "absolute" and "relative" types, i.e., all time expressions are labeled with the same type of tag. The salient features of the time expressions that are marked is that whether absolute or relative, they can be anchored on a timeline; unanchored durations, for example, are not marked.

The TIME sub-type is defined as a temporal unit shorter than a full day, such as second, minute, or hour. The DATE sub-type is a temporal unit of a full day or longer. Both DATE and TIME expressions may be either absolute or relative. Both absolute and relative times are tagged as TIME and absolute and relative dates are tagged as DATE.

B.1.1 Absolute Temporal Expressions - Time & Date

To be considered an absolute time expression, the expression must indicate a specific segment of time, as follows:

TIME-tagged expressions

– An expression of minutes must indicate a particular minute and hour, such as "20 minutes after 10" (not "a few minutes after the hour," "20 minutes after the hour").

– An expression of hours must indicate a particular hour, such as "midnight," "twelve o'clock noon," "noon" (not "mid-day," "morning").

DATE-tagged expressions

– An expression of days must indicate a particular day, such as "Monday," "10th of October" (not "first day of the month").

– An expression of seasons must indicate a particular season, such as "autumn" (not "next season").

– An expression of financial quarters or halves of the year must indicate which quarter or half, such as "fourth quarter," "first half." Note that there are no proper names, per se, representing these time periods. Nonetheless, these types of time expressions are important in the business domain and are therefore to be tagged.

– An expression of years must indicate a particular year, such as "1995" (not "the current year").

– An expression of decades must indicate a particular decade, such as "1980s" (not "the last 10 years").

– An expression of centuries must indicate a particular century, such as "the 20th century" (not "this century").

Temporal expressions are to be tagged as a single item. Contiguous subparts (month/day/year) are not to be separately tagged unless they are taggable expressions of two distinct TIMEX sub-types (date followed by time or time followed by date).

"twelve o'clock noon"

<TIMEX TYPE="TIME">*twelve o'clock noon*</TIMEX>

"5 p.m. EST"

<TIMEX TYPE="TIME">*5 p.m. EST*</TIMEX>

"January 1990"

<TIMEX TYPE="DATE">*January 1990*</TIMEX>

"fiscal 1989"

<TIMEX TYPE="DATE">*fiscal 1989*</TIMEX>

"the autumn report"

the <TIMEX TYPE="DATE">*autumn*</TIMEX> *report* "an Indian summer of the soul"

<TIMEX TYPE="DATE" ALT="summer">*an Indian summer*</TIMEX> *of the soul*

"third quarter of 1991"

<TIMEX TYPE="DATE">*third quarter of 1991*</TIMEX>

"the fourth quarter ended Sept. 30"

<TIMEX TYPE="DATE">*the fourth quarter ended Sept. 30*</TIMEX>

"the three months ended Sept. 30" [as paraphrase of "the fourth quarter ended Sept. 30"]

<TIMEX TYPE="DATE">*the three months ended Sept. 30*</TIMEX>

"the first half of fiscal 1990"

<TIMEX TYPE="DATE">*the first half of fiscal 1990*</TIMEX>

"first-half profit"

<TIMEX TYPE="DATE">*first-half*</TIMEX> *profit*

"fiscal 1989's fourth quarter"

<TIMEX TYPE="DATE">*fiscal 1989's fourth quarter*</TIMEX>

"4th period" [of a year]

<TIMEX TYPE="DATE">*4th period*</TIMEX>

"1975 World Series"

<TIMEX TYPE="DATE">*1975*</TIMEX> *World Series*

"February 12, 8 A.M."

<TIMEX TYPE="DATE">*February 12*</TIMEX>,<TIMEX TYPE="TIME">*8 A.M.*</TIMEX>

"by 9 o'clock Monday"

by <TIMEX TYPE="TIME"> *9 o'clock* </TIMEX><TIMEX TYPE="DATE"> *Monday*</TIMEX>

Determiners that introduce the expressions are not to be tagged. Words or phrases modifying the expressions (such as "around" or "about") also will not be tagged. Only the actual temporal expression itself is to be tagged.

"around the 4th of May"

around the <TIMEX TYPE="DATE">*4th of May*</TIMEX>

"shortly after the 4th of May"

shortly after the <TIMEX TYPE="DATE">*4th of May*</TIMEX>

B.1.2 Relative Temporal-Expressions

A relative temporal expression (RTE) indicates a date relative to the date of the document ("yesterday", "today", etc.), or a portion of a temporal unit relative to the given temporal unit ("morning" as the initial part of a specified day). Taggable RTE's include compound temporal expressions containing a deictic marker followed by a time unit, such as "last month" or "next year". If a numeral is included in RTE's of this type, it falls within the scope of the taggable temporal expression ("last two months"). Note that sometimes the deictic marker is postposed, as in "10 years ago" and "four months later". Note also that some RTE's lexicalize deictic markers and time units into a single word, such as "yesterday", which by itself constitutes a taggable expression, and that some RTE's can contain more than one deictic marker, such as "early this year" and "earlier this month." In addition, note that some of the expressions specifically defined as not being absolute temporal expressions are considered markable as relative temporal expressions.

Compound ("marker-plus-unit") temporal expressions, and their lexicalized equivalents, should be tagged as single items. However, if a lexicalized "marker-plus-unit" modifies a contiguous time unit of a different sub-type, they should be tagged as two items. Contrast the following two example markups:

<TIMEX TYPE="TIME">*last night*</TIMEX>

<TIMEX TYPE="DATE">*yesterday* </TIMEX> <TIMEX TYPE="TIME">*evening*</TIMEX>

Sometimes, however, the phrasing is such that the modification and types are non-contiguously arranged as in "8:40 Wednesday night" but marking three items of type TIME-DATE-TIME does not represent the modification accurately. In such cases, mark the entire phrase as a single temporal expression as shown in the following:

"4:15 p.m. Tuesday local time"

<TIMEX TYPE="TIME">*4:15 p.m. Tuesday local time*</TIMEX>

"early Friday evening"

<TIMEX TYPE="TIME">*early Friday evening*</TIMEX>

B.1.3 Miscellaneous Temporal Non-Entities

Indefinite or vague date expressions with non-specific starting or stopping dates will not be tagged. Non-taggable expressions include:

B.1.3.1 Vague Time Adverbials

"now", "recently", etc.

[no markup]

B.1.3.2 Indefinite Duration-of-Time Phrases

"for the past few years"

[no markup]

B.1.3.3 Time-Relative-to-Event Phrases

"since the beginning of arms control negotiations"

[no markup]

"The morning after the July 17 disaster"

The <TIMEX TYPE="TIME">*morning after the* <TIMEX TYPE="DATE">July 17</TIMEX> *disaster*</TIMEX>

B.2 Scope of Temporal Expressions

Absolute time expressions combining numerals and time-unit designators ("A.M., "P.M.", "EST", etc.), or other subparts associated with a single TIMEX sub-type, are to be tagged as a single item. That is, the subparts (such as numbers and time-units) are not to be tagged separately, even in the case of possessive or partitive constructions.

"twelve o'clock noon"

<TIMEX TYPE="TIME">*twelve o'clock noon*</TIMEX>

"5 p.m. EST"

<TIMEX TYPE="TIME">*5 p.m. EST*</TIMEX>

"the first half of fiscal 1990"

the <TIMEX TYPE="DATE">*first half of fiscal 1990*</TIMEX>

B.3 Temporal Expressions Containing Adjacent Absolute and Relative Strings

When a time expression contains both relative and absolute elements, the entire expression is to be tagged. The following examples illustrate some of the ways in which elements of relative and absolute time expressions may combine to form taggable time expressions.

"July last year"

<TIMEX TYPE="DATE">*July last year*</TIMEX>

"the end of 1991"

the <TIMEX TYPE="DATE">*end of 1991*</TIMEX>

"late Tuesday"

<TIMEX TYPE="DATE">*late Tuesday*</TIMEX>

B.4 Holidays

Special days, such as holidays, that are referenced by name, should be tagged.

"because of the observance of All Saints' Day"

because of the observance of <TIMEX TYPE="DATE"> *All Saints' Day* </TIMEX>

B.5 Locative Entity-Strings Embedded in Temporal Expressions

Rarely, multiword strings that are to be tagged as TIMEX will contain LOCATION (ENAMEX) substrings. Include these words within the scope of the tagged expression, but do not apply an embedded LOCATION tag.

"1:30 p.m. Chicago time"

<TIMEX TYPE="TIME">*1:30 p.m. Chicago time*</TIMEX>

Note that the above locative entity-string ("Chicago"), plus the word "time", modifies a contiguous TIMEX expression of the "TIME" sub-type.

Sometimes, however, the phrasing is such that the modification and types are non-contiguously arranged as in "Japan time, 19 February, 8:00 A.M." but marking three items of separately does not represent the modification accurately. In such cases, mark the entire phrase as a single temporal expression as shown in the following:

"Japan time, 19 February, 8:00 A.M."

<TIMEX TYPE="TIME">*Japan time, 19 February, 8:00 A.M.*</TIMEX>

A locative expression should be tagged separately as LOCATION if it is not contiguous to the "TIMEX" type expression, as in:

"In Japan, it would have occurred on 19 February, 8:00 A.M."

In <ENAMEX TYPE="LOCATION">*Japan*</ENAMEX>*, it would have occurred on* <TIMEX TYPE="DATE">*19 February*</TIMEX>, <TIMEX TYPE="TIME">*8:00 A.M.*</TIMEX>

B.6 Temporal Expressions Based on Alternate Calendars

Temporal expressions in terms of alternate calendars, such as fiscal years, the Hebrew calendar, Julian dates and "Star Date," will generally be marked up in accordance with the above guidelines for DATE.

APPENDIX C. NUMEX: SPECIFIC GUIDELINES

C.1 Scope of Numeric Expressions

The entire string expressing the monetary or percentage value is to be tagged.

"20 million New Pesos"

<NUMEX TYPE="MONEY">*20 million New Pesos*</NUMEX>

"$42.1 million"

<NUMEX TYPE="MONEY">*$42.1 million*</NUMEX>

"million-dollar conferences"

<NUMEX TYPE="MONEY">*million-dollar*</NUMEX> *conferences*

"15 pct"

<NUMEX TYPE="PERCENT">*15 pct*</NUMEX>

The word "minus," or the minus sign, should be included in the tagged numeric expression if it is a negative monetary or percentage value.

"minus 15 percent"

<NUMEX TYPE="PERCENT">*minus 15 percent*</NUMEX>

C.2 Numeric Expressions Appearing in Succession

Juxtaposed strings expressing monetary values in two different currencies are to be tagged separately.

"#26 million ($43.6 million)"

<NUMEX TYPE="MONEY">*#26 million*</NUMEX> (<NUMEX TYPE="MONEY">*$43.6 million*</NUMEX>)

C.3 Approximators and Multipliers in the Modification of Numeric Expressions

C.3.1 Approximators

Modifying words that indicate the approximate value of a number or a "relative position" to a number are generally to be excluded from the NUMEX tag if the modifier indicates only some minor imprecision in the known quantity. However, borderline cases such as the quantifier "several" may be optionally included in the tagged string.

"about 5%"

about <NUMEX TYPE="PERCENT">*5%*</NUMEX>

"more than 55%"

more than <NUMEX TYPE="PERCENT">*55%*</NUMEX>

"approximately 20 million New Pesos"

approximately <NUMEX TYPE="MONEY"> *20 million New Pesos* </NUMEX>

"over $90,000"

over <NUMEX TYPE="MONEY">*$90,000*</NUMEX>

C.3.2 Multipliers

Modifiers that indicate the multiplied value of a number unit should be included in the tagged string, if the modifier is a substitute for a specific digit (or the indefinite article or other quantitative determiner) within the monetary or percentage expression.

"several million New Pesos"

<NUMEX TYPE="MONEY">*several million New Pesos*</NUMEX>

"several million dollars"

<NUMEX TYPE="MONEY">*several million dollars*</NUMEX>

In this case, "several" is a substitute for some specific digit such as "3", or "4." Note that the expression remains grammatical if such a digit is substituted for the word "several", but that the expression "about 10 million New Pesos" does NOT remain grammatical if "about" is replaced by a digit. The indefinite article also can be substituted for "several," but not for "about," in the same examples.

C.3.3 Indefinite or Approximate Modifiers Medially Embedded Within Numeric Expressions

MUC/MET conventions do not allow for the tagging of discontinuous structures. If a modifier occurs in the middle of an otherwise taggable numeric expression, the entire expression should be tagged, regardless of whether the modifier seems to be semantically "approximate" or "indefinite".

"30 million plus New Pesos"

<NUMEX TYPE="MONEY">*30 million plus New Pesos*</NUMEX>

"30-plus million New Pesos"

<NUMEX TYPE="MONEY">*30-plus million New Pesos*</NUMEX>

C.4 Locative Entity-Strings Embedded in Numeric Expressions

Rarely, multi-word strings that are to be tagged as NUMEX will contain LOCATION (ENAMEX) substrings. Include these words within the scope of the tagged expression, but do not apply an embedded LOCATION tag.

"U.S. $700 million"

<NUMEX TYPE="MONEY">*U.S. $700 million*</NUMEX>

"the equivalent of less than a U.S. penny"

the equivalent of less than <NUMEX TYPE="MONEY">*a U.S. penny*</NUMEX>

"20 million New Taiwan Dollars"

<NUMEX TYPE="MONEY">*20 million New Taiwan Dollars*</NUMEX>

"US$43.6 million"

<NUMEX TYPE="MONEY">*US$43.6 million*</NUMEX>

C.5 Miscellaneous Numeric Non-Entities

Numeric expressions that do not use currency terms to indicate money values and that do not use percentage terms to indicate percentages are not to be tagged.

"12 points"

[no markup]

"unchanged at 95.05"

[no markup]

"1.5 times"

[no markup]

"about one-third of"

[no markup]

"Fees 1 3/4."

[no markup]

"a fixed 106 7/8"

[no markup]

"priced at 99 1/4"

[no markup]

For more website information contact: Ellen Voorhees
For more evaluation information contact: Nancy Chinchor
Last updated:
Date created: Friday, 12-Jan-01