ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Values

Version 1.1.2 – 2005.06.10

Linguistic Data Consortium

http://www.ldc.upenn.edu/Projects/ACE/

1. Basic Concepts

1.1 What is a Value?

A value is a string that further characterizes the properties of some entity or event.

We will only be interested in a subset of possible values. Specifically, we will be annotating *NUMERIC*, *CONTACT-INFO*, *TIME*, *JOB-TITLE*, *CRIME* and *SENTENCE* values.

Additionally, we will only be interested in a subset of the possible values for most of these types. For example, only *MONEY* and *PERCENT* will be annotated as NUMERIC values. A complete list of subtypes for each of the value types is provided in the section describing that type.

There are no subtypes provided for the *SENTENCE*, *CRIME* and *JOB-TITLE* value types. There will be no limit to the range of possible subtypes for the value types *SENTENCE*, *CRIME* and *JOB-TITLE* --- all examples of legal punishments, legal offenses and employment positions , respectively, will be taggable values of these types.

It is important to note that there are really two types of things that we are calling values here:

1. Strings that provide potential characterizing information about entities. These strings will always be tagged when they occur in a document. They need not participate in any relation or event.
2. Strings that participate as arguments in Events. These strings are only tagged when they occur within the scope of a taggable Event.

	Entity Characterizing Values	***Event Argument Values***
TYPE ***SUBTYPE***	NUMERIC *PERCENT* *MONEY* CONTACT-INFO *PHONE-NUMBER* *URL* *EMAIL* TIMEX2	JOB-TITLE SENTENCE CRIME
TAGGABLE?	Always tagged when mentioned	Only tagged when used as an argument in an Event

1.2 Extent

The rules for identifying the extent of a value mention will vary from type to type. The specific extent rules for a given type will be provided in the section describing that type. There are, however, some general properties of all value extents that can be mentioned up front.

Many values are mentioned by a noun phrase (NP). The entity task guidelines (Entity_Guidelines_v5.5.doc) introduce a detailed account of the manner in which the full extent of an NP can be identified. That account is repeated here for convenience.

The extent of a mention consists of the entire noun phrase. In case of structures where there is some irresolvable ambiguity as to the attachment of modifiers, the extent annotated should be maximally inclusive. In the case of a discontinuous constituent, the extent goes to the end of the constituent, even if that means including tokens that are not part of the constituent. Thus, in:

The terrorist was charged with conspiracy in the bombing of the USS Cole.

The extent of the mention is the entire noun phrase:

[conspiracy in the bombing of the USS Cole]

The extent includes all the modifiers of a noun phrase, including prepositional phrases and relative clauses.

Generally speaking, tokens are broken at white space, and each item of punctuation is treated as a separate character. As a rule, we do not include punctuation such as commas, periods, and quotation marks in the extent of a mention unless words included within the extent continue on after the punctuation mark. Possessive endings ('s) are treated as separate tokens, and contractions are split (so that "*we're*" becomes the two tokens "*we*" and "*'re*"). Extents must begin at the beginning of a token and end at the end of a token.

These general rules will be most useful in identifying the extent of JOB-TITLE and CRIME values. For the others, the specific rules may prove most useful, since many of these values are expressed with highly formulaic constructions (e.g. a U.S. phone number will almost always be represented as *(XXX) XXX-XXXX* or *1-XXX-XXX-XXXX*)

When annotating values, it will not be necessary to indicate a head. Identification of the extent will suffice.

2. Numeric Values

NUMERIC Values will be limited to the subtypes *MONEY* and *PERCENT*.

For *NUMERIC* values there will be two important parts of the mention:

The first part is the *indicator,* which is used to express the subtype of the *NUMERIC* value. For example, the symbol % would be an indicator that a *NUMERIC* value is of the *PERCENT* subtype. This can be expressed either as a symbol or as a string of words.

The second part is the *number*. This can be either a numeral or a string of words expressing a number.

For *NUMERIC* values, the extent will be the smallest string of words that includes both the *number* and the *indicator* and also any additional quantifiers that might be present such as 'nearly', 'almost' and 'over'. The number and indicator may occur in either order and need not be contiguous. The usual rules about whitespace and punctuation apply (see Section 1.2, above). The usual rules about annotating the full extent of the NP **do not**. We will only annotate the extent relevant to the Numeric value itself. For instance:

［八成］民众

[$400 Million] in stock

2.1 Percent

A *PERCENT* value is mentioned whenever numeric information is presented as a fraction of one-hundred (100).

[百分之二十五]

［四成］

[25%]

[25/100]

[十个百分点]

2.2 Money

A MONEY value is mentioned whenever capital is described in terms of the currency of some country or region (e.g. *US Dollars* or *Euros*).

[$400, 000]

[50 美元]

[20 元]

[十块]

3. Contact-Info

CONTACT-INFO values will be limited to examples of the *PHONE-NUMBER, EMAIL* and *URL* subtypes.

The extent for mentions of *CONTACT-INFO* values will be described independently for each of the subtypes.

3.1 Phone-Number

A *PHONE-NUMBER* value is mentioned whenever there is a string of numerals that can be used to make contact via phone.

The extent of a *PHONE-NUMBER* mention is the smallest sequence of tokens such that all of the numerals in the string are included. This will frequently include internal punctuation such as '-', '.', '(' and ')' and prefixes such as '+'.

[215.555.1111]

[(215) 555-1111]

[+44 23 345-1234]

3.2 Email

An *EMAIL* value is mentioned when someone there is a string of tokens that can be used to make contact electronic mail.

The extent of an *EMAIL* mention is the smallest sequence of tokens such that all of the tokens in the string are included. *EMAIL* value mentions should be contiguous, but will be punctuated by the symbols '@' and '.'.

chwalker@ldc.upenn.edu

president@whitehouse.gov

vice.president@whitehouse.gov

3.3 URL

A *URL* value is mentioned whenever the (virtual) location of a webpage is provided. It is not important that the webpage be the front page (or index page) of some site. A *URL* mention can point to any page directly accessible using a web-browser and the *URL* mentioned.

The extent of an *URL* mention is the smallest sequence of tokens such that all of the tokens in the string are included. *URL* value mentions should be contiguous but will contain tokens of various types (e.g. symbols, numbers and letters).

http://www.ldc.upenn.edu/Projects/ACE/

www.ldc.upenn.edu/Projects /ACE/

google.com

4. Time

For Time value, we are interested in temporal expressions which can reference calendar dates, times of day, dates, ages, seasons, or durations (such as peiods of hours, days or even periods of centuries). For detailed discussion of Time markability and Time extent, please refer to Timex2 Guidelines (Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson G. (2005)."*TIDES 2005 Standard for the Annotation of Temporal Expressions*").

Basically, if a phrase or word falls into one of the following categories, it should be marked as Time value:

- Indexical expressions – you have to know when you are to know what is being referrred to, such as “October 19”, “today”, “that day”, 此时, 此刻 etc.
- Fully referential expression, such as “October 19, 2000”, 10/19/2000 18:45:20
- **Can be oriented on a timeline, or at least be oriented with relation to another time** (past, present, future).
- 前 (as in 前总统), 曾经, 最近, 现在, 目前, 当前, 将来, 以前 (when it means past, not markable when it is used as a localizer, as in 三年以前)

Below are the phrases or words that are not markable:

- Localizers which indicate the time attributes (as-of, within, during, starting, ending, before, after etc), e.g.,在…的时候, 自从…, …以来, …来, 自…始, 为期…, …后, …内.
- Sequencing and ordering expressions, e.g., 接下来, 然后, 其次, 马上, 随后.
- Manner adverbs, e.g., 立刻, 立即, 即刻.
- Non-quantifiable durations, e.g., 暂时, 永远, 长期.
- Bare frequencies, e.g., 常常, 经常, 偶尔.
- Proper names, e.g., 十月革命, 半夜鸡叫, 辛亥革命.

Be careful that some expressions whose markability varies by word sense, e.g., 同时, which is markable when it is used as an anaphor, referring to an explicity time mentioned elsewhere in the text and is used like 当时 (at that time), but is not markable when it is funtioning more like a conjunction.

The extent of a *TIME* mention will be identified as in the TIMEX2 annotation guidelines. Anything modifying the temporal expression is in the extent of Time mention as well. Notes that if a localizer indicates the time attributes such as as-of, within, during, starting, ending, before, after, it should be not included in the extent. For example:

[[三年]前的今天]是[个不平常的日子]。

在[刚刚过去的1999 年]里，中国的国有企业改革力度进一步加大。

从[90 年代初]，广东全省经济发展了。

在[新千年]来临之际，我代表公司管理曾向各位员工表示衷心的感谢。

美国商会中国分会[近日]派出一个25 人组成的代表团，在华盛顿向国会和白宫展开为期[一周]的游说活动。

随着[苏哈托时代]的结束…

在就职仪式上，布什发表了[12 分钟]的就职演说。

张学良[连续三个晚上]观看北京京剧团演出。

[近十年]来，…

[以往十多年]的经验

Normalization of *TIME* values will be done as in the TIMEX2 guidelines. We will not do normalization of Time this year for Chinese.

We will annotate all TIME mentions that occur in either of two places within a given source document:

1. Between the tags <TEXT> and </TEXT>, even when the text is also contained by other tags, such as <POST> and </POST> or <TURN> and </TURN>, as for the other type of annotations applied in ACE.

2. Between the tags <DATETIME> and </DATETIME>, which occurs prior to the <TEXT> tag in source documents.

5. Crime

A *CRIME* value will be mentioned whenever the offense associated with some *JUSTICE* event is explicitly expressed. **Note that there must be a taggable instance of the corresponding *JUSTICE* event for there to be a taggable *CRIME* value.**

Since most *CRIME* value mentions will be expressed in the form of noun phrases, the extent of *CRIME* value mentions will be defined generally, as in Section 1.2 above. In the examples that follow, square brackets are used to indicate the *CRIME* value mention and **bold** font is used to indicate the trigger of the corresponding *JUSTICE* event.

*他因为[谋杀]受到**指控**。*

*他被**控**犯有[抢劫罪]。*

*10 年来，他[贪污 3000 万]，[受贿 500 万]，[两罪]并**罚**，被判处无期徒刑，剥夺政治权利终身。*

6. Sentence

A *SENTENCE* value will be mentioned whenever a punishment for some *JUSTICE* event is explicitly expressed. **Note that there must be a taggable instance of the corresponding *JUSTICE* event for there to be a taggable *SENTENCE* value.**

Since most *SENTENCE* value mentions will be expressed in the form of noun phrases, the extent of *SENTENCE* value mentions will be defined generally, as in Section 1.2 above. In the examples that follow, square brackets are used to indicate the *SENTENCE* value mention and **bold** font is used to indicate the trigger of the corresponding *JUSTICE* event.

> *10 年来，他贪污 3000 万，受贿 500 万，两罪并罚，被**判处**[无期徒刑]，[剥夺政治权利终身]。*
>
> *罪犯被**判**[有期徒刑 15 年]，并**处以**[罚金 5 万元]。*

7. Job-Title

A *JOB-TITLE* value will be mentioned whenever the office associated with some *PERSONNEL* event is explicitly expressed. (For a complete discussion of events and event scopes, please see the Events Guidelines (Events_Guidelines_v5.5)).

Please note that JOB-TITLE value mentions will often be co-extensive with *PERSON* entity mentions. When this happens, both the value and the entity will be annotated. For a complete discussion of the annotation of entities, please see the Entity Guidlelines (Entity_Guidelines_v5.5).

Since most *JOB-TITLE* value mentions will be expressed in the form of noun phrases, the extent of *JOB-TITLE* value mentions will be defined generally, as in Section 1.2 above. In the examples that follow, square brackets are used to indicate the *JOB-TITLE* value mention and **bold** font is used to indicate the trigger of the corresponding *PERSONNEL* event.

> *公司新近**聘**他为[财务总监]。*
>
> *张三**接替**李四成为[公司的新一任高级总工程师]。*