Shared Resources for Multilingual Information Extraction and Challenges in Named Entity Annotation

Shudong Huang
Linguistic Data Consortium
3600 Market St., Ste 810
Philadelphia, PA 19104
shudong@ldc.upenn.edu

Stephanie Strassel
Linguistic Data Consortium
3600 Market St., Ste 810
Philadelphia, PA 19104
strassel@ldc.upenn.edu

Alexis Mitchell
Linguistic Data Consortium
3600 Market St., Ste 810
Philadelphia, PA 19104
amitche0@ldc.upenn.edu

Zhiyi Song
University of Pennsylvania
3600 Market St., Ste Suite 501A
Philadelphia, PA 19104
zhiyi@ling.upenn.edu

Abstract

Progress in natural language processing requires increasing amounts of data and annotation in a growing variety of languages, and research in named entity extraction is no exception. While the value of richly-annotated, large-scale multilingual corpora is undeniable, costs for producing such data are high, underscoring the value of shared resources. As part of the US Government-sponsored Automatic Content Extraction Program (ACE), the University of Pennsylvania's Linguistic Data Consortium has recently created a number of shared resources to support technology evaluations in multilingual information extraction. This paper discusses the challenges of multilingual corpus development, with a particular focus on Chinese named entities. It concludes with a description of the corpora developed to support this research.

1 Introduction

Ongoing research in NLP requires vast amounts of data for system training and development, plus stable benchmark data to measure progress. Researchers require greater and greater volumes of data, representing a growing inventory of human languages and ever more sophisticated annotation. This presents a substantial challenge to the NLP community because human annotation and corpus creation is quite costly, and the availability of high quality language resources remains a central issue for the many communities involved in basic research, technology development and education related to language. The Linguistic Data Consortium (LDC) was founded in 1992 at the University of Pennsylvania, with seed money from DARPA, specifically to address the need for shared language resources. Since then, LDC has created and published more than 283 linguistic databases and has accumulated considerable experience and skill in managing large-scale, multilingual data collection and annotation projects.

Since 1999 LDC has been developing linguistic resources to support information extraction research, including named entity recognition. Recent work in this area falls primarily under the DARPA Program in Translingual Information Detection, Extrac-

tion, and Summarization (TIDES 2002), which combines technologies in detection, extraction, summarization and translation to create systems capable of searching a wide range of streaming multilingual text and speech sources, in real time, to provide effective access for English-speaking users.

Operating under the TIDES umbrella, the Automatic Content Extraction (ACE) program (NIST 2002) builds on the successes of previous extraction research programs like the Message Understanding Conference, or MUC (Chincor 1997). The objective of the ACE Program is to develop extraction technology to support automatic processing of source language data (in the form of natural text, and as text derived from Optical Character Recognition and Automatic Speech Recognition output). This includes classification, filtering, and selection based on the language content of the source data, i.e., the meaning conveyed by the data. Thus the ACE program requires the development of technologies that automatically detect and characterize this meaning. The ACE research objectives are viewed as the detection and characterization of Entities, Relations, and Events.

2 ACE Annotation Tasks

Linguistic Data Consortium develops annotation guidelines, corpora and other linguistic resources to support ACE (LDC 2004). ACE annotators tag broadcast transcripts, newswire and newspaper data in English, Chinese and Arabic, producing both training and test data for common research task evaluations. There are three primary ACE annotation tasks corresponding to the three research objectives: Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), and Event Detection and Characterization (EDC). A fourth annotation task, Entity Linking (LNK), groups all references to a single entity and all its properties together into a Composite Entity.

EDT is the core annotation task, providing the foundation for all remaining tasks. The current ACE task identifies seven types of entities: Person, Organization, Location, Facility, Weapon, Vehicle and Geo-Political Entity (GPEs). Each type is further divided into subtypes (for instance, Organization subtypes include Government, Commercial, Educational, Non-profit, and Other). GPE entities are also assigned roles.

Annotators tag all mentions of each taggable entity within a document, where "mention" is defined as a textual reference to an entity. For every mention, the annotator identifies its maximal textual extent and labels its head if available. Nested mentions are also captured. Each entity is classified according to its type and subtype. Each entity mention is further tagged according to its referential class – specific, generic, attributive, negatively quantified or underspecified. During the LNK annotation task, annotators review the entire document to group mentions of the same entity together; they also label cases of metonymy, where the name of one entity is used to refer to another entity related to it.

During RDC tagging, annotators identify relations that exist between the entities tagged during the EDT task. In EDC, they identify and characterize five types of events in which EDT entities participate.

2.1 Multilingual ACE

In its first two research cycles, the ACE program focused primarily on English language data. Under TIDES, the program has grown to include Arabic and Chinese, as well as exploratory work in Farsi. To move from the basic English tasks into Chinese, Arabic and Farsi, LDC draws on the expertise of fluent bilingual linguists and language scholars. These experts first fully learn the English annotation tasks and complete some training annotation in English. They then apply the English guidelines to texts in the target language, keeping careful note of any constructions that motivate changes or additions to the guidelines. After several rounds of test annotation in the target language, language-specific guidelines are crafted in English, but with examples drawn exclusively from the target language [1] . The new guidelines are then

[1] This means that annotators for non-English ACE tasks must be fluent bilinguals. Customarily, new annotators start by learning the English ACE tasks then move into

extensively tested with pilot annotation by multiple annotators in the target language. Further modifications to the guidelines are made as new patterns in the data are observed. Periodically, ACE annotation tasks have been redefined or modified with an eye to improving annotator consistency.

Each time a new language is targeted, language-specific challenges emerge. The next two sections explore in detail some of the issues of multilingual named entity annotation, contrasting English and Chinese treatments of mention classification, entity identification and entity class assignment in particular.

3 Mention Classification and Entity Identification

ACE annotation begins with detecting mentions of all relevant entities within a document. Mentions of the same entity are coreferenced. This is done in ACE annotation by recording them into the same row in the reference table.

Mentions can be classified into different types according to their morpho-syntactic properties. In earlier ACE specifications, three mention types were defined: nominal (NOM), named (NAM) and pronominal (PRO).

(1) [Al Gore]$_{NAM}$ called it a homecoming, but a comeback was what [he]$_{PRO}$ sought Wednesday as…

(2) …[the Democratic presidential nominee]$_{NOM}$ called out to a rally in [his]$_{PRO}$ grandmother's western Tennessee hometown.

For each mention, ACE annotation records both its head and extent. The extent of a mention is the entirety of the mention phrase, as shown by the brackets above. For a nominal mention, the head is simply the syntactic head of the noun phrase (for example, "nominee" in "the Democratic presidential nominee"). For a pronominal mention, the head is the pronoun itself, and since a pronoun seldom has modifiers, the head and extent usually correspond. Mentions without an overt syntactic head – like "[the old] will be well taken care of" in English or "[该来的]没来, [不该走的]走了" ("*Those who should have come haven't, and those who shouldn't have left have*") in Chinese – were previously tagged as pronominal as well, though the type of "headless" has been newly introduced.

Within ACE a name is atomic, i.e., the head of a name is the entire name string, for example, "Al Gore" in the above example is not tagged into two names, "Al" and "Gore". However, proper names in actual texts may also have modifiers such as non-restrictive relative clauses, prepositional (in English) or pre-modifying locative (in English and Chinese) expressions, etc. In such cases, the extent includes all such modifiers, but the head consists of the name only. There are some other specifics for named mentions. For example, person titles are excluded from the head but are included in the extent.

3.1 Atomicity and Nested Named Mentions

Determining atomicity can itself be a problem for mentions containing strings that are potential mentions of other taggable entities. This is particularly problematic for Chinese, as company names are often prefixed with the name of the country, city or province where the company is based. The annotator must decide whether the prefix is a part of the proper name, or simply a locative modifier of the name. Compare for example, "America Online" and "Sun Microsystems". "America Online" is translated into Chinese as "美国在线". But one frequently sees "美国" ("*America*") prefixed to the Chinese translation of "太阳微系统公司" or "太阳公司" ("*Sun Microsystems*"), particularly at the beginning of a text. The annotator may mistakenly tag the entire string "美国太阳公司" as a single named mention, instead of having two nested mentions, like "[[美国]$_{NAM}$ 太阳公司]$_{NAM}$", where "美国" is a mention for another entity. Annotators are encouraged to consult external resources like entity lists when in doubt; the

their language-specific annotation. This supports a consistent approach to annotation across the multiple languages despite the necessary language-specific modifications.

default rule is to tag two entities so that a relation in the RDC phase can be established between them, achieving more for what ACE is about.

Names of subordinate organizations, especially those labeled with terms that describe their function, frequently consist of common nouns only. Consider, for example, "the Department of State" (or its variant "the State Department). Is it a named mention? A native speaker of English tends to say yes because of the capitalization and because it does uniquely identify an organization. If we treat it as a named mention, the next question is whether "U.S." as in "the U.S. Department of State" should be included in the head of the mention, that is, whether "the U.S. Department of State" is atomic. Even a native speakers disagree on the appropriate treatment.

Since the ACE program is not only concerned with the identification of entities, but also with the relationships between them, there is motivation to tag "U.S" as a separate mention referring to another entity despite the fact that "the U.S. Department of State" is the official name of that department. This way, we capture the subordinate relationship between the organization and government (expressed in the term "U.S."). Another advantage of this approach is flexibility: if future specifications strictly require that the mention of an organization must include the mention of its "parent" entity in the head when the mention string is present in its official name, we can simply merge the two.

This approach has been adopted in Chinese ACE, but it calls into question the distinction between named and nominal mentions. While "State Department" is a fairly unique name, "Ministry of Foreign Affairs" (or its variant "Foreign Ministry") is not. Not only do most countries use this name for that government department, but the term itself is also compositional. In both Chinese and English, the term can be used a nominal mention, as in "两国外交部" ("*the two countries' foreign ministries*"). In addition, English frequently drops capitalization for such expressions. Thus, while it seems reasonable to tag "the foreign ministry" as nominal, it is less clear if "the Chinese foreign ministry" should be treated as nominal or named.

Within Chinese ACE, for mentions like 外交部, if there is a specific entity referenced by the mention in the context, it is tagged as named regardless of whether or not the name of the parent entity is prefixed. If it refers to a group of entities, we treat it as nominal:

(3) [[中国]$_{NAM}$外交部]$_{NAM}$
("*Foreign Ministry of China*")
(4) [[外交部]$_{NAM}$发言人]$_{NOM}$
("*foreign ministry spokesman*")
(5) [[日本]$_{NAM}$外务省]$_{NAM\backslash}$
("*Foreign Ministry of Japan*")
(6) [[美国]$_{NAM}$国务院]$_{NAM\backslash}$
("*the US State Department*")
(7) [[中俄两国]$_{NOM}$外交部]$_{NOM}$
("*the foreign ministries of China & Russia*")

3.2 Complex Mentions: A Solution for Complex Phrasal Structures

Earlier classification of mention types into named, nominal and pronominal is simply based on the syntactic head of the mention phrase: whether it is a common noun, a proper noun, or a pronoun. Since ACE data is not syntactically pre-processed, this led to rules that did not take into account many of the syntactic features that were commonly identified.

Another restriction of the earlier classification scheme was the requirement that only one head was allowed for any given mention and the extent must be a continuous string of words. This creates difficulties for mentions where more than one syntactic head is present. For example, nouns are frequently conjoined and then modified by an adjective or relative clause to form a complex noun phrase like "16 angry men and women". The earlier approach was to simply tag the mentions linearly as follows.

(8) [16 angry men] and [women]

Not only does the approach not correctly reflect the phrasal structure, it also creates problems for other ACE tasks. We cannot, for example, establish a parent-child relationship

between "mother" and "John" for the phrase "John's father and mother" because "John" is not included in the extent headed by "mother".

(9) [John's father] and [mother]

Constructions like these motivated the introduction of complex mentions into the new ACE specifications. A complex mention must have (1) two or more simple or complex mentions directly embedded or (2) another complex mention directly embedded. Because of this embeddedness, a complex mention is never assigned a head – it only has an extent. Since by definition, the inner-most embedded mentions are always simple mentions, they must be assigned a head (except for headless mentions).

For example, appositive constructions, in both English and Chinese, are tagged as complex mentions (APP). The reason is that there is no agreement how an appositive construction should be syntactically analyzed. Rather than being committed to any syntactic theory, we simply treat it as a complex mention:

(10) [[Joe]$_{NAM}$, [the linguist]$_{NOM}$]$_{APP}$, will give a talk on named entities.

(11) [[中国最大的城市]$_{NOM}$[上海]$_{NAM}$]$_{APP}$ 经济发展一直领先。
("*Shanghai, China's largest city, has always been a leader in economic development.*")

In the current ACE task, several types of complex mentions exist, with variation across languages to account for syntactic differences. For instance, while the current English guidelines identify Multi-Mention Constructions (MNH) [2] and Complex Construction with a Relative Clause (ARC), the corresponding Chinese guidelines specify Parallel Multiple Mentions (PMM), Extended Parallel Multiple Mentions (EPM) and Extended Apposition Constructions (EAP).

[2] The latest English version makes a distinction between "multi-mention construction" and "multi-head nominal mention", the latter, a simple mention, referring to the construction where two bare nouns are conjoined. The two types are likely to merge in future revisions.

3.3 PMM and EPM in Chinese

The use of complex mentions can help resolve some problems by revealing underlying syntactic or morphological structures. We will examine PMM and EPM in Chinese as an illustration.

A PMM is defined as a complex mention where two or more mentions or their subparts are conjoined, disjoined, or enumerated. In other words, the multiple mentions inside a PMM are morpho-syntactically parallel to each other, for example, "Apollo [[13]$_{NAM}$ and [14]$_{NAM}$]$_{PMM}$".

If only subparts of multiple mentions are in parallel and the "shared" component(s) is a portion of the head of every mention, we reapply the notion of PMM to this structure. What this means is that a PMM can be embedded within another PMM, but the inner PMM must have at least two parallel words or morphemes, while words or morphemes between the inner PMM and the outer PMM form a complete mention head with each of the words or morphemes inside of the inner PMM. Thus, for the Apollo example, we have "[Apollo [[13]$_{NAM}$ and [14]$_{PMM}$]$_{PMM}$".

If a PMM has head-external expressions, e.g. a relative clause, a demonstrative, a classifier, a nominal/adjectival modifier, etc,. the entire construction is tagged as EPM (extended parallel multiple mentions) with the PMM embedded.

(12) [16 angry [[men]$_{NOM}$ and [women]$_{NOM}$]$_{PMM}$]$_{EPM}$

(13) [[John's]$_{NAM}$ [[father]$_{NOM}$ and [mother]$_{NOM}$]$_{PMM}$]$_{EPM}$

Here are some examples showing how PMM and EPM are applied in Chinese.

(14) [[[黄埔]$_{NAM}$ 和[南浦]$_{NAM}$]$_{PMM}$ 大桥]$_{PMM}$
("*Huangpu Bridge and Nanpu Bridge*")

(15) [报名参展的 [[国家]$_{NOM}$ 和[地区]$_{NOM}$]$_{PMM}$]$_{EPM}$
("*the countries and regions that have enrolled for the exhibition*")

(16) [[俄罗斯]$_{NAM}$ 的[[[明斯克]$_{NAM}$ 和[库尔斯克]$_{NAM}$]$_{PMM}$ 号]$_{PMM}$ 潜艇]$_{NOM}$
("*the Russian submarines, Minsk and Kursk*")

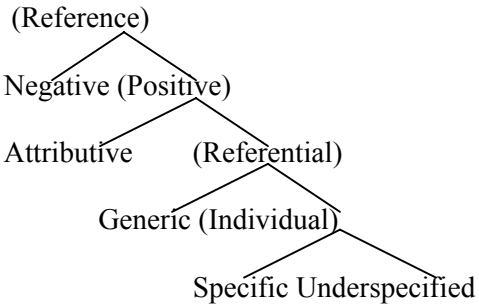
Using PMM and EPM tags makes named entity extraction easier and more precise as each individual mention can be extracted in its entirety even though the surface form is "broken". For example, from "[Apollo [[13]$_{NAM}$ and [14]$_{PMM}$]$_{PMM}$", we can easily extract the names of the two spacecrafts, "Apollo 13" and "Apollo 14".

The introduction of PMM and EPM also helps other tasks of ACE. For example, we can now establish a parent-child relation between "mother" and "John" for (13) above because the underlying extent for the mention headed by "mother" is "John's mother".

4 Entity Class (Referentiality) and Names

The notion of entity class, or referentiality, refers to the relation between a term and the object(s) that the mention is used to talk about. In earlier phases of ACE, we only distinguished between generic and specific references. A *generic* reference was loosely defined as a mention-entity relation where the mention "does not refer to a particular object or particular set of objects in the world". Otherwise the mention was *specific*. Despite many tests designed to help detect generic-hood – primarily for English, annotators frequently had difficulties making the right decision, if any.

This two-way distinction has been revised in the new ACE specifications into a completely new classification of referentiality. The following figure shows the new system of entity classes in the form of a decision tree.



At the top level is the distinction between "negative" and "positive" references. Negative reference means zero reference, that is, the mention refers to an empty set. It is specifically used for mentions with a negative quantifier, such as "no one", "no weapons of mass destruction", "nobody", etc. This class does not apply to Chinese as there are no negative quantifiers in the language.[3]

An attributive mention is also non-referential, but in a different sense. A mention is attributive if it is used to ascribe a property or attribute to another entity. In the follow examples, all mentions in *italics* are attributive.

(17) John is *a linguist*.
(18) John, *the linguist*, will be here.
(19) He is *John*.
(20) He is called *John*.

An attributive mention is always related to another mention in the sentence and only appears in certain syntactic structures, for examples, appositive constructions and predicates led by the copular verb "be". Because attributive mentions are non-referential, we do not record two attributive mentions into the same row in the reference table even if they may be of identical strings.

The notion of generic under the new ACE specifications is very restricted: a generic reference applies only when a mention refers to a class/kind/species of objects or a typical representative of that class/kind/species. So if any property predicates on a generic mention, it means the entire class referred to by the mention has that property, or all/most/any members of that class have the property.

A non-generic referential mention refers to one or more non-representative, individual members of a class/kind/species of entities. This class, also known as individual reference, is further divided into specific and non-specific (or underspecified). A referential mention is specific if the entity or entities referred

[3] The only expression that looks like an English negatively quantified NP is "没有 + N/NP", but it is really a negated existential construction since "没有+ N/NP" cannot freely fill in an NP position in a sentence, for example, in an object position.

to are a specific individual object or a set of specific individual objects related by the speaker regardless of whether they can be named, counted, pointed to, etc. Otherwise, the reference is non-specific, or underspecified.

The notion of underspecificity is essentially a bucket for annotators to throw in any mention that cannot be easily fitted into other classes. The above figure also serves as a decision tree for annotator with "underspecified" being the last label to use. This strategy has boosted consistencies across annotators.

4.1 Revisiting the Named vs. Nominal Distinction

The distinction between generic and individual references helps us better understand the distinction between named and nominal mentions. A concrete common noun, in an abstract sense and without any context, is also a name: it is a class/kind/species name, an idea first put forward by ancient philosophers. If a nominal mention is used as an individual reference in a discourse, the head noun often has to be "individualized" via quantification and/or qualification with determiners, adjectives, relative clauses, etc., although Chinese has fewer morph-syntactic means than English does and context plays a more important role. But the distinction between named and nominal in ACE is not based on the traditional distinction between common and proper noun. Named mentions can only refer to individual entities in ACE.

To see how ACE differentiates named and nominal mentions and how the named and nominal distinction helps the differentiation, consider first the following two examples:

(21) 波音 747 比其它飞机大。
("*Boeing 747 is larger than other airplanes.*")
(22) 中国今年又购买了 12 架波音 747。
("*China purchased another 12 Boeing 747 this year.*")

Intuitively, 波音 747 ("*Boeing 747*") in (21) is more like a name; but it is a name for a *specific kind* of airplanes, or alternatively, a name for an airplane model. However, an airplane model is not a taggable entity in ACE. If we are to tag 波音 747 in (21), we can only treat it as a generic nominal mention referring to a class of airplanes. In (22), 波音 747 is clearly a nominal mention of individual reference.

Consider further the term "Americans" in the following two examples:

(23) Americans are at war with Iraq.
(24) Americans eat more beef than Chinese.

In (23), "Americans" cannot be interpreted as referring to any American. The entity it refers to is actually the country itself, a geopolitical entity (GPE) in ACE. Thus, "Americans" should be tagged as a named mention of specific reference. (24) is ambiguous. Under the interpretation where Americans as a whole consume more beef than Chinese as a whole, "Americans" should be tagged the same way as in (23). But for the other interpretation, where *most* Americans consumes more beef than *most* Chinese, "Americans" refer to person entities and should be tagged as a nominal mention of generic reference.

Just as "Kleenex" is synonymous to "facial tissue" and "Xerox" to "copy machine", "Enron" is now synonymous to "corporation involved in accounting scandals".

(25) 又一家安隆东窗事发。
("*Another Enron just spun off.*")

Here the proper name 安隆 ("*Enron*") is used as a nominal mention, though it is too early to know if "Enron" can make its way into the lexicon.

In short, a named mention under ACE can never be of generic reference whereas nominal mentions can be of any kind of reference. Although most nominal mentions are headed by a common noun, a proper noun can also head a nominal mention.

5. Corpora

As part of the ACE and TIDES information extraction programs, LDC has developed a number of annotated corpora. These corpora all draw on broadcast news, newspaper and newswire data. Sources include data from the

Topic Detection and Tracking corpora, Chinese Treebank, Arabic Treebank and other news data. The table below summarizes data developed thus far for ACE:

Corpus/ Phase	Data Amount (words/language)	Tasks	Languages	Evaluation	Availability
ACE-Pilot	15K training	entities	English	May, Nov 2000	Available 2004
ACE-1	180K training, 45K evaluation	entities	English	Feb 2000	Available 2004
ACE-2	180K training, 45K dev, 45K eval	entities, relations	English, Chinese	Sept 2000	LDC Catalog # LDC2003T11
ACE 2003	100K training, 50K evaluation	entities, relations	English, Chinese, Arabic	Sept 2003	LDC Catalog # LDC2004T09
ACE 2004	300K training, 50K evaluation	entities, relations, events	English, Chinese, Arabic	Fall 2004	Under development

Another resource created to support named entities within information extraction more broadly is the Xinhua Chinese-English Named Entity list, created from Xinhua Newswire's proper name and who's who databases. This corpus contains nearly one million proper names of various kinds, including approximately 500,000 person names, 300,000 place names, 30,000 organization names, and tens of thousands of other name types. The data provides both Chinese to English and English to Chinese name pairs. This corpus, slated for publication in Summer 2004, is currently available to TIDES/ACE participants.

Sponsored common task research programs like TIDES and ACE rely heavily upon such shared resources. In order to allow for expedited delivery of data to a group of researchers participating in a common task evaluation, LDC has developed a new data distribution method by releasing e-corpora. E-corpora provide expedited delivery of training and devtest data in support of formal evaluations. Upon the conclusion of the formal task evaluation, pending negotiations with research sponsors and program coordinators, LDC publishes data more broadly to permit access to these valuable resources to all communities working in linguistic education, research, and technology development.

References

Chen, Ping, 1986, Referent Introducing and Tracking in Chinese Narrative, University of California, Los Angeles, Ph. D. Dissertation.

Chinchor, Nancy, 1997, MUC-7 Named Entity Task Definition Version 3.5 [http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html]

Donnellan, Keith, 1991, Reference and Definite Descriptions. In: Steven Davis (ed), "Pragmatics: a Reader", Oxford, Oxford University Press

LDC, 2004, Automatic Content Extraction [http://www.ldc.upenn.edu/Projects/ACE]

Liberman, Mark and Christopher Cieri, 2002, TIDES Language Resources: A Resource Map for Translingual Information Access, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

NIST, 2002, Automatic Content Extraction [http://www.nist.gov/speech/tests/ace]

Powell, George, 1999, The referential-attributive distinction – a cognitive account, UCL Working Papers in Linguistics 11 (1999)

TIDES, 2002, DARPA Program in Translingual Information Detection Extraction and Summarization [http://www.darpa.mil/iao/TIDES.htm]

Yang, Rong, 2001, Common Nouns, Classifiers, and Quantification in Chinese, Rutgers University, New Brunswick, Ph. D. Dissertation.