Multilingual Resources for Entity Extraction

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

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

Abstract

Progress in human language technology requires increasing amounts of data and annotation in a growing variety of languages. Research in Named Entity extraction is no exception. Linguistic Data Consortium is creating annotated corpora to support information extraction in English, Chinese, Arabic, and other languages for a variety of US Government-sponsored programs. This paper covers the scope of annotation and research tasks within these programs, describes some of the challenges of multilingual corpus development for entity extraction, and concludes with a description of the corpora developed to support this research.

1 Introduction

Ongoing research in human language technology (HLT) requires vast amounts of data for system training and development, plus stable benchmark data to measure ongoing progress. Researchers require greater and greater volumes of data, representing a broadening inventory of human languages and ever more sophisticated annotation. This presents a substantial challenge to the HLT community because human annotation and corpus creation is quite costly. New approaches to research require not tens but hundreds and thousands of hours of speech data, and millions of words of text. The availability of high quality language resources remains a central issue for the many communities involved in basic research, technology development and education technology development and education related to language. The role of international data centers continues to evolve to accommodate emerging needs in the speech and language technology community (Liberman and Cieri 2002).

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 241 linguistic databases and has accumulated considerable experience and skill in managing large-scale, multilingual data collection and annotation projects. LDC has established itself as a center for research into standards and best practices in linguistic resource development, while participating actively in ongoing HLT research.

LDC has had a major role in creating annotated corpora and other resources to support named entity extraction, as well as larger information extraction activities, for a number of years. Current work in this area falls under a handful of research programs. The DARPA Program in Translingual Information Detection, Extraction, and Summarization (TIDES 2002) 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. TIDES core languages are English, Mandarin and Arabic; second tier languages are Korean, Spanish, and Japanese. The primary medium is text though this includes speech recognition output. The TIDES research tasks require broadcast transcripts and news texts to be annotated for entities, relations, and events; categorized by topic; translated; summarized; and processed in a variety of other ways.

Another of the TIDES Program goals is to produce technology that can be easily ported to handle new natural languages. To this end, the TIDES Surprise Language Exercise (LDC 2003b) challenges researchers to produce working systems for a previously untargeted language within a constrained time span (for instance, a single calendar month).

Currently operating under the TIDES umbrella, the Automatic Content Extraction program (NIST 2002) builds on the successes of previous extraction research programs. 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. LDC provides data and annotations to support these program goals.

Another DARPA program, Evidence Extraction and Link Detection (EELD 2002), draws on linguistic resources created by LDC to promote its research goals. The EELD program aims for development of technologies and tools for automated discovery, extraction and linking of sparse evidence contained in large amounts of classified and unclassified data sources. EELD is developing detection capabilities to extract relevant data and relationships about people, organizations, and activities from message traffic and open source data. LDC has provided domain-specific entity-tagged corpora in support of the EELD technology evaluation.

2 From MUC to ACE

While LDC's current resource development efforts support ACE and related programs in particular, ACE is hardly the first program to tackle named entities and the larger information extraction problem. The Message Understanding Conference Program (MUC) (NIST 1999a) focused on named entity extraction, coreference relations among noun phrases, the identification of selected relations, and events. High system performance within the English newswire domain motivated an expansion of the named entity task after MUC-7. In 1999, the DARPA Hub-4 NE Project (Chinchor et. al. 1999) expanded the domain of source data to include broadcast news transcripts.

LDC joined the group of those developing corpora to support named entity research in that same year, providing annotations for the TIDES Information Extraction-Entity Recognition (IE-ER) task (NIST 1999b). In the following year, LDC began to develop corpora and other resources to support the ACE program.

ACE is substantially similar in scope to these earlier extraction programs, though slightly different in focus. ACE adds new varieties of annotated data to the information extraction domain. Annotators tag newswire, broadcast news transcripts and newspaper data. Additionally, research sites are evaluated on their performance on degraded ASR and OCR output. Under the TIDES umbrella, the ACE program supports the multilingual resource and system development, focusing currently on Chinese and Arabic.

ACE also modifies the inventory of entity types targeted by the MUC tasks (Chinchor et. al. 1997). While MUC considered three entity types (person, organization, location), ACE further divides locations into geo-political entities and facilities, while the newest phase also adds weapons, substances, and vehicles (similar to the MUC artifact category). Coreference is preserved in ACE, while generic entities and metonymy are tackled explicitly.

ACE brings together many of the separate tasks evaluated under different components of the MUC program. All ACE tasks -- entities, relations and events -- evaluate not only recognition but also characterization of these phenomena. While the MUC Template Relation and Scenario Template tasks targeted relations and events plus their attributes, the focus of these tasks was domain specific. ACE tasks, on the other hand, are defined to be more general and domain-independent.

3 Annotation

Named entity annotation is a core component of ACE, but the scope of the annotation required by the program builds substantially on this.

3.1 The Task

There are three main ACE tasks: Entity Detection and Tracking, Relation Detection and Characterization, and Event Detection and Characterization.

Entity Detection and Tracking is the most fundamental of the ACE tasks and was the sole focus of the ACE Pilot effort as well as ACE Phase One. The entity task provides a foundation for the remaining annotation and research tasks. ACE annotators identify five types of entities (Mitchell et. al. 2002). The first two types, Person and Organization, remain substantially similar to their definitions under MUC. Locations within ACE are limited to geographical entities such as landmasses, bodies of water, and geological formations. Two new entity types are tagged under ACE: Facilities, which include buildings and other permanent man-made structures and real estate improvements; and GPEs, which are geographical regions defined by political and/or social groups. GPEs are composite in nature, typically having a government, a populace, and a geographic location, as well as some more abstract notion of statehood.

A GPE subsumes and does not distinguish between a nation, its region, its government and its people. However, annotators also assign a role to each textual reference (mention) of a GPE, indicating which of these aspects is most prominent for that mention. In the example below, the two entity mentions refer to the governmental (rather than people or location) aspect of the entities, so in both cases the mentions would be tagged as GPEs with Organization roles:

> **{Russia}** recently held discussions with **{the US}** regarding the ongoing crisis.

ACE annotators tag all mentions of each entity within a document, whether named, nominal or pronominal. For every mention, the annotator identifies the maximal extent of the string that represents the entity. Nested mentions are also captured. Each entity is classified according to its type, and co-reference among mentions is recorded. While the ACE Pilot Annotation effort did not explicitly deal with metonymic entities and generics, Phase One of ACE added these elements to the entity annotation and research tasks.

Metonymy occurs when a single string of text makes reference to multiple entities. Generally, these distinct entities are related to each other in some way. For example, in this sentence,

> **{Beijing}** will not continue sales of anti-ship missiles to {Iran}.

Beijing, though literally referring to the name of the capital of China, is being used as a reference to the government of China. The relationship between "Beijing the city" and "Beijing the seat of China's government" triggers this metonymic reference. When metonymic references occur, ACE annotators create two separate entities, one for each reference.

An entity is generic when it does not refer to a particular object or set of objects in the world. Generic entities include references to general types of objects, hypothetical objects and generalizations across sets of objects. Annotators apply the rules of mention coreference to generic entities, and specifically classify each entity as specific or generic.

In future ACE efforts the set of targeted entities will expand to include vehicles, weapons, and substances. Entity subtypes will also be added.

The second phase of the ACE program added relation detection and characterization to the suite of annotation and research tasks. This task targets five relation types (Mitchell et. al. 2002b): Role, Part, Located, Near, and Social. For example, Role relations link people to organizations and GPEs in employment, affiliate, and citizenship relationships. The Social relation links people in personal, familial or professional relationships. Each relation type is further classified according to its subtype. For instance, the Role relation includes Management, General Staff, Member, Owner, Founder, and Citizen-Of subtypes.

For every relation, annotators identify two primary arguments (namely, the two ACE entities that are linked) as well as the relation's temporal attributes (Sundheim 2001). Temporal information is drawn from pre-existing TIMEX2 [1] annotation (Ferro et. al. 2001) wherever those values are explicitly linked to a relation. LDC annotators also create more general, relative time attributes de-

[1] TIMEX2 annotation, supported by the ACE program, provides a framework for the normalized representation of temporal expressions.

rived from the tense of the verb that heads the predication of the relation. Relations that are supported by explicit textual evidence are distinguished from those that depend on contextual inference on the part of the reader.

The following is an example of an explicit relation of type Located:

> {**President Bush**} was in {**New York**} Thursday.

The textual evidence supports the relation between the entities *President Bush* and *New York*, with the temporal attributes *was* and *Thursday*.

Future phases of ACE will refine the relation task to highlight new relations that are of particular interest to the program, and to allow finer categorization of some existing types.

ACE Phase Three adds a new challenge: recognition and characterization of events. Definition of a set of general event types and subtypes is currently underway. Targeted types include Interaction, Movement, Transfer, Creation and Destruction events. Annotators label event arguments (agent, patient and the like) and attributes (temporal, locative) according to a type-specific template. They further tag the textual mention or anchor for each event and categorize it by type and subtype. For example, the sentence below contains reference to an Interaction event.

> {**Colin Powell**} and {**Jiang Zemin**} *held high-level talks* in {**Beijing**} last week.

Annotators extract the specific text reference to the event (*held high-level talks*); identify the meeting participants (*Colin Powell, Jiang Zemin*) as arguments of the event; tag the locative (*Beijing*) and temporal (*held, last week*) attributes.

The event task will expand in future phases of ACE to include additional event types and subtypes, as well characterization of relations between events.

3.2 The Process

The large amounts of data, multilingual focus and the number and range of annotation tasks required by the ACE program lends itself to a team-based approach to annotation. A single project manager provides oversight for all LDC ACE activities. Language-specific lead annotators work directly with teams of part-time (typically student) annotators, providing training, monitoring progress and generally supervising the annotation staff. The project manager works with lead annotators to develop and maintain the formal ACE annotation task definitions and guidelines (LDC 2003a).

The complexity of ACE annotation requires annotators with a solid background in linguistics, particularly syntax and semantics. New annotators first become familiar with the basic concepts and terminology and study the annotation guidelines before annotating several sets of training files. Throughout the training process, supervisors provide periodic feedback, comparing the trainee's annotation to a gold standard, identifying discrepancies and refining the annotator's approach to the data and understanding of guidelines and rules. Not until an annotator has achieved a certain level of accuracy and speed is he permitted to tag actual data.

The annotation work environment is designed to encourage regular discussion and "groupthink" among the annotation team. Problems and questions are logged for future reference, and teams meet regularly to discuss outstanding issues. A web-based annotation manual contributes to the team approach. This reference complements the formal task definition, documenting decisions about how to handle problematic constructions and outlier examples. Because its content is developed solely by ACE annotators, the web guidelines also function as a training tool. New annotators regularly add to the guidelines, focusing on the aspects of the ACE tasks that are most difficult for them.

During production annotation, separate annotators conduct at minimum two complete passes over the data. First pass annotation creates the initial markup, and a second pass reviews the existing annotation for consistency and accuracy. Second passing is typically conducted by more experienced senior annotators. A targeted third pass is performed to further enhance annotation quality. During the third pass, lead annotators review the annotated data to catch common errors and ensure consistent treatment of difficult constructions.

In addition to multiple passes over all ACE data, an additional 5% to 10% of the data is completely re-annotated from scratch by separate annotators. Results of this dual annotation are compared and discrepancies adjudicated in order to establish inter-annotator agreement scores and identify areas

of lingering confusion or inconsistency. While rates of inter-annotator agreement for ACE named entities are comparable to MUC consistency levels, the results for the more complex annotation tasks are considerably lower. Particular challenges include the coreference of generic entities and the use of metonymy, GPE roles, and implicit vs. explicit relations.

The first two phases of ACE annotation utilized MITRE's Alembic Workbench (Day 1997), which was customized for the ACE tasks. With the expansion into new languages and the addition of events, LDC began development of a locally designed, locally supported ACE toolkit. Utilizing the Annotation Graphs model (Bird and Liberman 2001), the toolkit provides for customized, platform-independent, multilingual ACE annotation. At present the toolkit supports entity tagging only; focused relation and event tagging modules are under development. The toolkit will also support customized functions for second passing, comparison and adjudication of dually-annotated files, and additional quality control features including queries of the annotation database.

3.3 Multilingual ACE

In its first two phases the ACE program has focused primarily on English language data. Under TIDES, the program has grown to include new languages. LDC is supporting this expansion with production annotation in Arabic and Chinese, as well as exploratory work in Farsi.

LDC has completed development of entity annotation guidelines in Chinese and Arabic. Full-scale Chinese annotation is well underway, while Arabic annotation is just beginning. 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, new guidelines are crafted in English, but with examples drawn exclusively from the target language[2]. The new guidelines are then 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.

Each time a new language is targeted, language-specific challenges emerge. For Chinese, one of the most difficult problems is the lack of agreed-upon rules for word segmentation. While English is written with white space around each new word, "word" is not a fundamental concept in Chinese, and characters are written without white space. Because entity annotation requires annotators to select both the maximal extent of a mention as well as the mention's head, it becomes difficult for annotators to agree on the exact series of characters that constitute the head of a mention. Annotation guidelines for Chinese must include rules for dealing with this issue.

Chinese also presents difficulties for tagging generic entities. The rules for identifying generics in English rely in part on tests surrounding the existence of determiners. However, determiners do not exist in Chinese, and this required the creation of new annotation guidelines for generics that rely solely on context. Similarly, Arabic often uses determiners in a way that is different from English. For instance, in Arabic it is common to use a construction with a determiner when referring to a class of entities:

> The horse is a wonderful animal.

rather than a bare plural, more common in English:

> Horses are wonderful animals.

The complexity of Arabic morphology presents a very different set of problems. Unlike Chinese and English, Arabic commonly uses pronoun affixes. For ACE, this means that any annotation tool must allow partial words to be tagged as mentions of entities in Arabic, while disallowing this for other languages.

In addition to these linguistic differences, some distinctive stylistic qualities of Chinese and Arabic news reporting present challenges for annotators and are worthy of note.

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

Many of these challenges are based in cultural differences. For example, many industries in China are government owned and operated. Consequently, names of organizations are often quite different than their English counterparts, and guidelines written with English naming conventions in mind are inadequate for handling common Chinese name constructions like "Beijing School Number 4".

Further, organizations located outside of China are often referred to with their country's name preceding the company name. This presents a challenge for annotator consistency, since it is often unclear whether to include the country as part of the extent of the company name.

Arabic news sources regularly use very long sentences with multiple clauses. This presents the annotator with different kinds of mention extent and coreference decisions than found in English news data. Mention extents are typically longer and contain more nested mentions, and pronominal references to entities are more easily confused.

Another set of problems extends beyond any language-specific considerations; these have to do with the infrastructure needed to support a large-scale multilingual data creation effort. Finding qualified native speaker annotators with adequate training in linguistics and eligibility to work in the United States is a serious challenge. Further, expanding ACE into new languages is not simply a matter of addressing the linguistic questions, but also tackling the technical ones. Maintaining data formats and annotation tools that can accommodate not only multiple annotation tasks, but also multiple languages and multiple character sets and encodings presents a significant problem.

Despite the range of issues described above, porting the ACE annotation task into new languages is relatively straightforward. The fundamental work of moving into a new language for ACE involves identifying the syntactic and morphological (i.e., surface) constructions that are used to refer to the entities, relations and events of interest. This is not an insubstantial task, and requires both the insights of trained linguists and many rounds of pilot annotation and exploration of the data. However, the fundamental concepts targeted by ACE, and the underlying semantic content discussed in the annotated texts, remain substantially similar from one language to the next.

4 Corpora

As part of the ACE program, and to further support both the DARPA TIDES and DARPA EELD 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 materials.

Corpus development for the ACE program began in 1999. Initially, the Pilot Phase was designed to develop a basic task definition for entity detection and tracking. Multiple research sites including MITRE, BBN, NYU, and LDC annotated the same set of 15,000 words of English data to establish a shared understanding of the annotation guidelines and resolve any inter-annotator discrepancies. This data supported technology evaluations in May and November 2000.

In ACE Phase 1, the research and annotation tasks were expanded to address metonymy and generic entities. Multiple research sites joined LDC in annotating 180,000 words of training data to support a February 2002 evaluation. LDC was solely responsible for annotating an additional 45,000 words of evaluation data.

ACE Phase 2 required research sites to additionally detect and characterize relations between entities. During this phase of ACE, LDC acted as sole annotation site and also took on responsibility for developing and maintaining annotation guidelines. Phase 2 used the entire ACE Phase 1 corpus as training data, and added an additional 45,000 words of new evaluation data. Both training and evaluation data were annotated for entities plus relations. In support of the EELD Program, LDC annotators tagged another 30,000 words of domain-specific training data plus 20,000 words of test data for entities and relations. A September 2002 evaluation tested system performance for both Entities and Relations.

LDC is currently producing English test data to augment the existing corpora in support of a Fall 2003 TIDES extraction evaluation; in addition, LDC is creating data and annotations for multilingual extraction research in Chinese and Arabic. 100,000 words of Chinese Treebank and 10,000 words of Arabic Treebank have already been annotated for entities.

Alongside corpus development, LDC is working in parallel to expand and refine the existing set of ACE tasks. These modifications are being made with input from both the TIDES Extraction and ACE communities. For ACE Phase 3, LDC will annotate 300,000 words of data in each of three languages: English, Chinese and Arabic; pilot annotation in Farsi is also targeted. Ultimately, all three annotation tasks -- entities, relations and events -- will be represented in the data. The corpora developed by LDC to support ACE, EELD, and TIDES Extraction are currently available to program participants only (LDC 2003c). General publication of the ACE Pilot and ACE Phase 1 Corpora is slated for Summer 2003; upon publication, the data will be available to LDC members as well as non-members. The remaining ACE and related corpora will be published after the conclusion of these programs' evaluation cycles.

Outside of the ACE program, LDC has developed a handful of additional resources for multilingual extraction research. As part of the TIDES Surprise Language Exercise, LDC collects and creates linguistic resources in a previously untargeted language in an extremely compressed time span. During a two-week dry run in March 2003, the target was Cebuano, a language of the Philippines. Within the span of a few days, LDC created 250,000 words of monolingual text, built a 20,000 word lexicon, created 25,000 words of parallel text, built a morphological parser, and completed named entity tagging of 32,000 words of text.

Given the severe time constraints of the exercise, named entity annotators used a trimmed-down version of the MUC Named Entity Guidelines rather than the more complex full MUC or ACE guidelines. Despite the time constraints, inter-annotator consistency remained high when LDC-tagged data was compared with data tagged by annotators at BBN. A similar set of resources for a new surprise language will be developed during the Surprise Language evaluation in June 2003. All of the data developed for Surprise Language is currently available to TIDES participants, and will be released as a general publication at the conclusion of the Exercise.

A final 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 2003, is currently available to TIDES participants.

Much of the material described above is based upon large volumes of text and speech best collected from commercial providers. Commercial sources may require the negotiation of agreements that permit the distribution of data to researchers while constraining the use of the material to linguistic education, research, and technology development. LDC coordinates all necessary intellectual property arrangements for data developed under multiple research programs including TIDES, ACE, and EELD to make resources gathered in this way available to the broader research communities.

Sponsored common task research programs like TIDES and ACE rely heavily upon such shared resources. LDC was in fact created specifically to facilitate research sharing. 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 known as ECorpora. ECorpora target expedited delivery of training and devtest data to 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

Bird, Stephen and Mark Liberman, 2001, A Formal Framework for Linguistic Annotation. [http://agtk.sourceforge.net/]

Chinchor, Nancy, et al., 1999, Named Entity Recognition Task Definition v1.4. [ftp://jaguar.ncsl.nist.gov/ace/phase1/ne99_taskdef_v1_4.pdf]

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]

Day, David. 1997, Alembic Workbench User's Guide. [http://www.mitre.org/tech/alembic-workbench/manual/]

EELD, 2002, DARPA Program in Evidence Extraction and Link Detection [http://www.darpa.mil/iao/EELD.htm]

Ferro, Lisa, et al., 2001, TIDES Temporal Annotation Guidelines Version 1.0.2.

LDC, 2003a, Automatic Content Extraction [http://www.ldc.upenn.edu/Projects/ACE/]

LDC, 2003b, Surprise Language Project [http://www.ldc.upenn.edu/Projects/SurpriseLanguage]

LDC, 2003c, TIDES Project [http://www.ldc.upenn.edu/Projects/TIDES/]

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.

Mitchell, A., et al. 2002a. Annotation Guidelines for Entity Detection and Tracking (EDT) Version 2.5. [http://www.ldc.upenn.edu/Projects/ACE]

Mitchell, A., et al. 2002b. Annotation Guidelines for Relation Detection and Characterization (RDC) Version 3.6. [http://www.ldc.upenn.edu/Projects/ACE]

NIST, 1999a, Message Understanding Conference [http://www.itl.nist.gov/iaui/894.02/related_projects/muc/]

NIST, 1999b, TIDES Information Extraction-Entity [http://www.nist.gov/speech/tests/ie-er/er_99/er_99.htm]

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

Sundheim, Beth, 2001, Preliminary RDC Guidelines for Time Attributes Version 1.0.

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