Enhanced Infrastructure for Creation and Collection of Translation Resources

Zhiyi Song, Stephanie Strassel, Gary Krug, Kazuaki Maeda
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810 Philadelphia PA 19104 USA
E-mail:{zhiyi, strassel, gkrug, maeda}@ldc.upenn.edu

Abstract

Statistical Machine Translation (MT) systems have achieved impressive results in recent years, due in large part to the increasing availability of parallel text for system training and development. This paper describes recent efforts at Linguistic Data Consortium to create linguistic resources for MT, including corpora, specifications and resource infrastructure. We review LDC's three-pronged approach to parallel text corpus development (acquisition of existing parallel text from known repositories, harvesting and aligning of potential parallel documents from the web, and manual creation of parallel text by professional translators), and describe recent adaptations that have enabled significant expansions in the scope, variety, quality, efficiency and cost-effectiveness of translation resource creation at LDC.

1. Introduction

Statistical Machine Translation (MT) systems have achieved impressive results in recent years, due in large part to the increasing availability of parallel text for system training and development. Linguistic Data Consortium at the University of Pennsylvania has undertaken a number of recent efforts to develop linguistic resources for MT on a large scale. We utilize a three-pronged approach, including acquisition of existing parallel text from known repositories, harvesting and aligning of potential parallel documents from the web, and manual creation of parallel text by professional translators. This paper reviews each approach in detail, and describes recent adaptations that have significantly expanded the scope, variety, quality, efficiency and cost-effectiveness of our translation resource development efforts.

2. Context for Resource Creation

Driving these adaptations is LDC's role in resource creation and distribution for a number of sponsored technology evaluation programs, including DARPA GALE (Strassel, 2006), the NIST Open MT evaluation series (NIST, 2009), the REFLEX Less Commonly Taught Languages (LCTL) program (Simpson et al, 2009), the ACE Program Entity Translation Pilot Task (Song & Strassel 2008), and MADCAT (Strassel 2009).

These programs and others have required LDC to expand the scope and complexity of its translation efforts, first by branching out into new genres. While many existing parallel text corpora focus on newswire (NW), the demands of programs like GALE have required an extension into spoken and unstructured domains. Broadcast news (BN) transcripts in multiple languages is one such genre; while spoken (and thus subject to occasional speech errors, disfluencies, filled pauses and the like) this genre primarily consists of read speech, and is only a mild extension from the challenges of newswire translation. In contrast, broadcast conversation (BC) -- consisting of transcripts from talk shows, roundtable discussions, call-ins and the like -- is significantly more challenging, with multiple speakers engaging in rapid, spontaneous, and frequently overlapping speech. For Arabic, broadcast conversation transcripts are also marked by heavy use of colloquial varieties (as opposed to Modern Standard Arabic) which

Language Pair	Approximate Volume (Words)	Genres	Methodology
Arabic > English	100M +	BN, BC, NW, WB, VAR	manual translation, parallel text harvesting, acquisition of existing manual translations
Chinese > English	100M +	BN, BC, NW, WB	manual translation, parallel text harvesting, acquisition of existing manual translations
English > Chinese	250K +	BN, BC, NW, WB	manual translation
English>Arabic	250K +	BN, BC, NW, WB	manual translation
Bengali>English	250-500K + per language pair	NW, WB	manual translation, parallel text harvesting
Pashto>English			
Punjabi>English			
Tagalog>English			
Tamil>English			
Thai>English			
Urdu>English			
Uzbek>English			

Table 1: Summary of Recent LDC Translation Efforts

presents additional challenges for transcription as well as translation. Unstructured text genres like weblogs and newsgroups (WB) present additional challenges. These genres often include threaded discussions with posts by multiple authors, who employ non-standard punctuation, spelling, grammar combined with heavy use of abbreviation and emoticons. Non-standard lexical items and novel coinages are also typical, as are other colloquial usages including (as with BC) use of colloquial Arabic varieties instead of MSA. Perhaps the most challenging recent domain expansion has been in support of the DARPA MADCAT Program, which has required translation of Arabic handwritten texts in a wide variety of document types (VAR) including forms, letters, memos and ledgers. These materials are characterized by all the challenges of web text, plus the added complications of poor legibility, handwriting errors and ambiguous reading order. The topical focus of these documents also proves difficult, with a prevalence of military jargon and named entities (particularly place and organization names) that are heavily region-specific. This data necessitates specialized knowledge on the part of the translator, and additional quality control measures prior to data release.

Beyond expansion into new genres, LDC's recent efforts have also required an extension in to new linguistic varieties. Most notably, the LCTL program required translation efforts in eight language pairs for which existing parallel text resources are scarce. Supporting translation in these languages demanded extensions to LDC's existing infrastructure including guidelines and quality control practices. Similar enhancements were also demanded to account for the prevalence of colloquial Arabic varieties for some projects.

Finally, the translation task itself has evolved in response to program demands. In addition to producing large volumes of commercial-quality basic translations for use as training data, LDC is increasingly required to adjust the quality, translation style and approach to meet particular evaluation requirements; for instance by creating multiple human translations and adjudicating them into a single gold standard; by providing alternative translations for ambiguous phrases; or by performing translation post-editing -- taking the output of an MT system and manually adjusting it to produce an accurate and fluent reference.

Table 1 above summarizes recent translation efforts across programs and languages.

3. Enhanced Infrastructure for Parallel Text Creation

3.1 Data Pipeline

Given increased demand for large volumes of manually created parallel text in multiple languages and genres, across projects often with overlapping (and aggressive) timelines, investment in standardized infrastructure and tools is essential to maintaining consistency, efficiency and maximum through-put.

To this end, LDC has developed a standardized translation data pipeline that remains stable across projects and tasks. This stability permits translators, managers and programmers alike to focus primarily on the translation task and data itself, rather than on the procedural approach.

Data Selection

The pipeline begins with data selection. A pool of candidate documents is assembled consisting of LDC collected data. This pool may be automatically assembled based on genre, language, epoch or other characteristics, or it may be manually constructed based on other features to suit specific needs of users and sponsors. For instance, in the case of GALE evaluation data, candidate files are carefully selected through a multi-stage process involving a combination of manual review (for language/dialect, genre, content, topic, and other features) and automatic sub-sambling (e.g. to match desired estimated translation error rate or n-gram and perplexity distribution).

As statistical MT systems are trained with more and more parallel text, they require "novel" training data that uses resources effectively. Whereas traditional selection methods produce full-document translations (a whole newswire article, blog post or broadcast story), the new method targets individual high-yield sentences -- where high-yield is defined having features that are novel compared to existing stores of training data. In collaboration with GALE research teams including IBM and SRI, we have developed a process to utilize information about n-grams and perplexity in existing translation models, then automatically select a set of novel candidate sentences from a pool of previously-unseen data. Humans review the ranked list of candidates, discarding anything that is not suitable for translation. Initial feedback from GALE teams is positive, but it remains to be determined whether this method provides the anticipated boost to system performance.

Processing

After selection additional scans may be conducted, for instance to flag identical content or to verify that no overlap exists between training and test data partitions. Finally, formatting validation steps ensure consistent use of markup, character encoding and the like.

SU Segmentation

Once data has been selected for translation, documents are automatically or manually segmented into Sentence Units (SU), following guidelines developed by LDC for this task. Automatic SU segmentation is created by using language-specific segmentation tools developed by LDC and trained on existing manually-segmented data. Manual SU segmentation is generally performed as part of the transcription process in the case of spoken genres; or as a standalone annotation task in the case of text genres. LDC's XTrans Toolkit (Glenn et. al., 2009) contains a module for performing SU segmentation. The manual SU segmentation task can be done quite efficiently, and is an essential component of LDC's translation pipeline. Seg-

mentation into sentences prior to manual translation results in parallel text that is perfectly aligned at the sentence level, which significantly enhances the value of the resulting resource and enables downstream manual and automatic tasks (including word alignment).

Format Conversion and Translation Assignment

After SU segmentation, data is converted into a translator-friendly format, consisting of UTF-8 encoded plain text documents, where each numbered segment of source data is paired with a corresponding blank numbered line, as in the following example for Arabic:

<ar=1> Arabic text
<en=1> [blank line]
<ar=2> Arabic text
<en=2> [blank line]

This format supports the goal of sentence-aligned translations, and its simplicity reduces translator-introduced formatting errors. The formatted files are then assembled into "kits" for outsourcing to one or more of LDC's vetted translation vendors. A script for kit creation permits translation managers at LDC to quickly generate customized kits for assignment to multiple agencies from a large pool of selected, formatted data. Kit customization might consider data volume, genre, translator expertise, file length, and level of difficulty. In general, an effort is made to evenly distribute sources and genres across different translation teams.

Post-Procssing and Distribution

After translation, kits are checked in by translation teams then validated with a suite of scripts developed to transform the translated data into a deliverable corpus. Processing scripts extract translated text lines from the incoming translation file, then verify a number of features including document formatting and text encoding, presence of translation for all segments, and consistent use of expected translation markup (for instance, to indicate translator uncertainty or translation alternatives). Automatic patches are applied wherever possible, and in some cases translation teams are contacted for clarification or correction of significant problems. Validated translations are then converted into the designated translation distribution format, which can be a tab-delimited file (.tdf), SGML, XML, or some other standard required by the program or end user.

Figure 1 presents a graphical representation of the standard LDC translation pipeline.

3.2 Tracking and Management Database

LDC's core translation infrastructure is grounded in a custom MySQL database that tracks every file at every stage of the translation pipeline, enabling consistent and efficient process management. Indexed by document, the database contains document metadata like language, genre

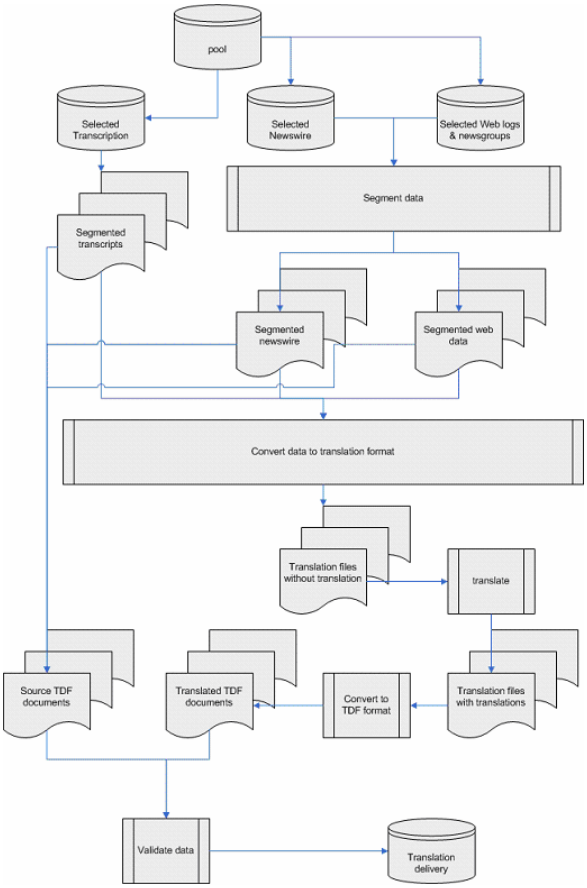

Figure 1: Standardized Translation Pipeline

and token count; it also indicates what data set(s) - by project, phase and train/dev/eval partition - the document appears in. Administrative functions and translation vendor maintenance are also provided: the database tracks kit assignments, assignments, due dates, and financial information for each data set. Finally, the database records the location of source and translation data on LDC's file servers; this information is then exploited by downstream processing scripts. Fields can be populated via batch imports, and files are added at the planning stages so that each step in the pipeline can be tracked and recorded. The database is also the backend for LDC's Translation Extranet, currently under development. This resource will allow translation teams to access pending assignments, check files in and out, validate completed kits, view quality control reports and other feedback, generate payment requests, and otherwise manage translation assignments in a convenient, one-stop approach. Database query support also permits managers to easily retrieve relevant information and respond to site or sponsor inquiries (e.g., which kit it belongs to, when it was delivered, what the QC score was, which agency translated it, whether it has been processed, etc.). Managers can also query kit and file status, for instance to determine what deliveries are pending, or what files for a given epoch, source and genre have never appeared in an evaluation set for any program; this is an essential benefit to selecting suitable material for other projects and downstream tasks like word alignment or

treebanking. The translation database is illustrated in Figure 2.

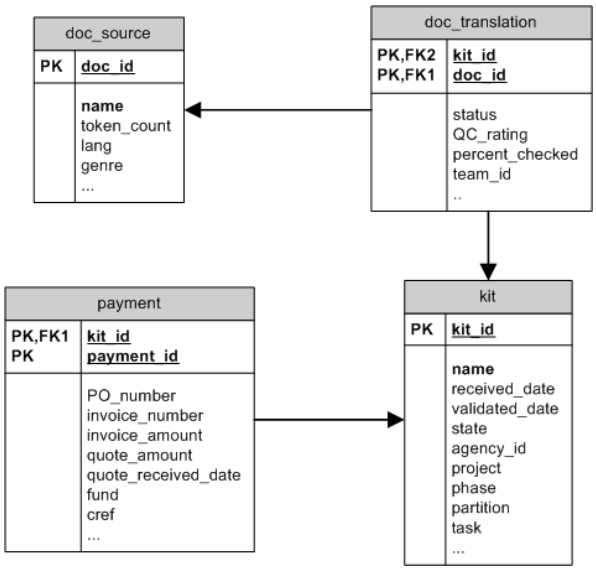

Figure 2: Translation Database Configuration

3.3 Translation Guidelines

In addition to the technical infrastructure described above, LDC has also produced extensive documentation to systematize and standardize human translation approaches for the range of programs and tasks we support. The core resource here is a set of language- and task-specific translation guidelines (LDC, 2009), which are the procedural manuals translation teams must follow when creating parallel text. Given the algorithms employed by statistical MT systems and the goals of the various evaluation programs we support, LDC's translation guidelines typically emphasize accuracy and fidelity to the source text over any other considerations, including fluency or stylistic tone.

The overall structure of the guidelines is stable across tasks and languages, and includes information about expected data formats, delivery methods and LDC quality control procedures. The guidelines specify the required makup of the translation team and state that teams may use an automatic machine translation system and/or a translation memory system to assist them during translation. Translation teams are instructed to provide substantial documentation along with delivered data, including

• Translator and proofreader profiles consisting of name or pseudonym, native language, second languages, age and years of translation experience. When multiple translation teams are used, also indicate team membership for each person.

• Work assignment information consisting of the team number or the unique identifier for the translator and proofreader for each file in the data set.

• The name and version number of any translation system or translation memory used.

• A description of any additional quality control procedures or other relevant parameters that affect the translation

A clause also states that if for any reason translators are uncomfortable working with any particular document included in their assignment, they may contact LDC to request a replacement. This statement is necessary given the fact that LDC translation teams are located around the world, sometimes in politically sensitive climates, and may be uneasy working with material from particular American media or other sources.

The bulk of the documentation, which may be refined for each project/language pair, consists of a series of general rules for handling various (language-specific and general) linguistic constructs, genre features or other considerations, along with numerous examples of preferred and dispreferred translations. Translators are required to follow the guidelines' specifications for translating factual errors, proper names, idiomatic expressions and more. Additional rules address language-specific constructs like pro-dropping, serial verbs, and the presence of colloquial varieties (e.g. for Arabic). Instructions for handling genre-specific challenges – like typos, neologisms, emoticons, and other features of web data, and disfluencies, filled pauses, partial words, restarts, and speaker noises for speech data – are also included. Addressing these special cases in the translation guidelines assures consistency where there would otherwise be variability if individual translators relied on their own best judgment. Furthermore, clear markup that indicates typos, translator uncertainty, and made-up words allows sites and evaluators to treat these instances differently when necessary.

Additional guidelines exist for specialized tasks. These include post-editing, where translators modify machine translation output to generate a gold standard translation, and translation alternative generation, where translators produce multiple translation options for cases where the source text is ambiguous as to its meaning[1] -- for instance, in the case of Arabic proper names where context does not disambiguate between a choice of literal translation or transliteration of the proper name.

Special guidelines also exist for translating novel (single) sentences automatically selected via the methods described in Section 3.1. The single-sentence translation task is particularly challenging because of the lack of full document context. To overcome this burden LDC provides translation teams with an html version of the full document with the targeted sentence highlighted, along with the standard translator-friendly format. Translators are instructed to avoid relying on extra-sentential information for specific translation choices; for instance, named

[1] This task does not target modelling of predictable variation, e.g. synonyms or standard syntactic alternatives.

entities that do not appear in the assigned sentence cannot be used to resolve pronoun co-reference or gender. Task-specific guidelines also exist for the translation of handwritten documents. In this task translators are provided with a digital transcript of the text to be translated, along with the image file of the original handwritten document. Reading order is also indicated; this is key, given that many images consist of tables and forms, where it may not be immediately apparent which sentence or chunk of text should be read/translated first. Given the especially challenging nature of these two translation tasks, translation teams generally require a longer timeline for translation, as well as a 5-10% premium on per-token translation fees.

The various translation guidelines are living documents, regularly updated to include more examples and address new translation issues, or to incorporate feedback provided by translators. Translation agencies are always required to use the most up-to-date version of the guidelines.

3.4 Quality Control

All translations produced by professional translation vendors undergo some level of quality control (QC) at LDC, prior to distribution and publication. Although all of our translation teams have been thoroughly vetted and tested by us prior to any assignments being issued, our translation model -- guidelines, genres, timeline and volume -- is challenging even for seasoned professional.

We adopt two distinct QC models, one generally applied to training data and another for evaluation data. In the light QC scenario typically used for training data, a subset of each translation delivery (typically 10-20% of the total token count) is randomly selected and checked by fluent bilingual annotators trained in the appropriate procedure. Annotators apply specific scoring mechanisms according to a rubric included in the translation guidelines provided to agencies. Translation errors are categorized as syntactic, lexical, poor usage, or typographic (significant spelling or punctuation mistakes), with specific points deducted for each type of error. Translation submissions whose total points deducted exceeds a particular threshold are returned to the agency; payment is withheld until corrections are completed on the entire translation set (not just the files that were reviewed) and the revised translation delivery meets QC standards. Even for submissions that pass the quality threshold, feedback is provided to translators via a standardized written report, with examples of poor translations and scores for each dataset. Translation teams have found this approach to be beneficial for both training and evaluating their translators. QC results are also encoded in LDC's translation database, which can be reviewed in determining future tasking assignments.

The full QC scenario typically employed for evaluation and development data is significantly more extensive and time-consuming. Full QC involves multiple stages, each targeting a different facet of the translation problem:

- A source-language dominant bilingual annotator checks submitted translation for errors and omissions;
- A source-language dominant bilingual senior annotator checks for remaining errors, improves fluency, corrects and standardizes named entities;
- A target-language dominant bilingual annotator improves fluency and adds translation variants where required;
- A target-language *monolingual* annotator reviews for fluency and consistency, and flags questionable regions for re-assignment to Stage 2.

In most cases and for most projects, 100% of evaluation data is subject to this careful and time-consuming review. The process is made significantly more efficient by the existence of a customized translation QC user interface, developed by LDC as an extension to the XTrans Transcription Toolkit (Glenn et. al., 2009, Friedman et. al., 2008).

3.5 Parallel Text Harvesting

Beyond the creation and acquisition of manual translations of the type described above, LDC has also developed a set of software tools for identifying potential parallel text resources in a large pool of online multilingual documents. We regularly run these tools on resources where parallel text might be found (Maeda et. al., 2008); such resources include newswire articles from multilingual news agencies, such as AFP (Agence France Presse) and Xinhua News Agency.

These documents come in a variety of formats. All source files are converted into a text format with a predefined SGML or XML markup standard. The document mapping module of the Bilingual Internet Text Search (BITS) system (Ma, 1999) is then run to identify pairs of possible parallel documents. Once pairs are identified, we automatically segment each document into and then run the Champollion sentence aligner (Ma, 2006) to create sentence mapping tables. This process can result in high yields; LDC has used this method to harvest over 82,000 Arabic-English document pairs and 67,000 Chinese-English document pairs for distribution to GALE program participants.

4. Conclusion

In response to a constellation of recent demands for new kinds of translation resources, LDC has developed a robust and flexible translation pipeline that combines enabling technical infrastructure, detailed task specifications and fully documented best practices. These efforts have allowed resource creation to become more efficient and adaptive, with increased emphasis on automation and utilization of emergent technology to improve and aug-

ment the data pipeline. This stable and adaptive infrastructure has permitted LDC to meet and often exceed requirements for training, development and evaluation data sets in multiple languages and genres, in support of concurrent projects with demanding and frequently overlapping timelines.

Many of the resources described here are already available to the research community at large. Translation guidelines and task specifications are freely available on LDC's website, while annotation toolkits like QCTrans are targeted for free, open-source distribution. Many corpora produced using the methods described here are already published in LDC's catalog, with several more slated for publication in the coming months. Recent corpora from the GALE program are listed in Table 2.

.

Catalog Number	Title
LDC2007T23	GALE Phase 1 Chinese Broadcast News Parallel Text - Part 1
LDC2008T08	GALE Phase 1 Chinese Broadcast News Parallel Text - Part 2
LDC2008T18	GALE Phase 1 Chinese Broadcast News Parallel Text - Part 3
LDC2007T24	GALE Phase 1 Arabic Broadcast News Parallel Text - Part 1
LDC2008T09	GALE Phase 1 Arabic Broadcast News Parallel Text - Part 2
LDC2009T02	GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 1
LDC2009T06	GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 2
LDC2008T02	GALE Phase 1 Arabic Blog Parallel Text
LDC2008T06	GALE Phase 1 Chinese Blog Parallel Text
LDC2009T03	GALE Phase 1 Arabic Newsgroup Parallel Text - Part 1
LDC2009T09	GALE Phase 1 Arabic Newsgroup Parallel Text - Part 2
LDC2009T15	GALE Phase 1 Chinese Newsgroup Parallel Text - Part 1
LDC2010T03	GALE Phase 1 Chinese Newsgroup Parallel Text - Part 2

Table 2: Recent Parallel Text Corpora from LDC

5. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

6. References

Friedman, L., Lee, H., Strassel, S. (2008). A Quality Control Framework for Gold Standard Reference Translations: The Process and Toolkit Developed for GALE. In *Proceedings of LREC-2008,* Marrakech, Morocco.

Glenn, M., Strassel, S. and Lee, H. (2009). XTrans: a speech annotation and transcription tool. In *Proceedings of Interspeech-2009,* Brighton, England.

LDC (Linguistic Data Consortium) (2009). LDC Human Translation Guidelines, *http://projects.ldc.upenn.edu/gale/Translation*.

Maeda, K., Ma, X., and Strassel, S. (2008). Creating Sentence-Aligned Parallel Text Corpora from a Large Archive of Potential Parallel Text using BITS and Champollion. In *Proceedings of LREC-2008,* Marrakech, Morocco.

Ma, X. (2006). Champollion: A Robust Parallel Text Sentence Aligner. In *Proceedings of LREC-2006*, Genoa, Italy.

Ma, X., Cieri, C. (2006). Corpus Support for Machine Translation at LDC. In *Proceedings of LREC-2006*, Genoa, Italy

Ma, X., Liberman, M. (1999). BITS: A Method for Bilingual Text Search over the Web. Machine Translation Summit VII, September 13th, 1999, Kent Ridge Digital Labs, National University of Singapore.

NIST (National Institute of Standards and Technology) (2009). *NIST 2009 Open MT Evaluation*, http://www.nist.gov/speech/tests/mt/2009

Simpson, H, Maeda, K, Cieri, C. (2009). Basic Language Resources for Diverse Asian Languages: A Streamlined Approach for Resource Creation. In *Proceedings of the 7th Workshop on Asian Language Resources*, ACL-IJCNLP 2009, Suntec, Singapore.

Song, Z, Strassel S. (2008). Entity Translation and Alignment in the ACE-07 ET Task. In *Proceedings of LREC-2008,* Marrakech, Morocco.

Strassel, S., Cieri, C. Cole, A., Dipersio, D., Liberman, M., Ma, X., Maamouri, M., Maeda, K. (2006). Integrated Linguistic Resources for Language Exploitation Technologies. In *Proceedings of LREC-2006*, Genoa, Italy.

Strassel, S. (2009). Linguistic Resources for Arabic Handwriting Recognition. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.