Enhanced Infrastructure for Creation and Collection of Translation Resources

Zhiyi Song, Gary Krug, Stephanie Strassel, Kazuaki Maeda
Linguistic Data Consortium, University of Pennsylvania

1 Introduction

Statistical Machine Translation (MT) systems have achieved impressive results in recent years, in part by using large quantities of parallel text for system training and development. Linguistic Data Consortium at the University of Pennsylvania supports the development of language technology like MT, as well as language-related research and education, by creating and distributing linguistic resources on a large scale. Increasingly these resources include substantial volumes of parallel text. Ma & Cieri (2006) described 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. The current paper describes recent adaptations that have significantly expanded the scope, variety, quality, efficiency and cost-effectiveness of LDC's development of translation resources.

2 Context and Scope of Translation 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, 2006, 2008, 2009), the ACE 07 Entity Translation program (Song and Strassel, 2008) and the REFLEX Less Commonly Taught Languages (LCTL) program (Simpson et al, 2009).

These programs and others have required LDC to expand the scope and complexity of its translation efforts, branching out into new genres (including broadcast news, broadcast conversation, weblogs, newsgroups and others), new language pairs including some less commonly taught or previously under-resourced languages, and new methodologies. Table 1 summarizes recent translation efforts.

Language Pair	Approximate Volume (Words)	Research Programs	Genres[1]	Methodology[2]
Arabic > English	100M +	TIDES, GALE, NIST MT, ACE	BN, BC, NW, WB	create, harvest, acquire
Chinese > English	100M +	TIDES, GALE, NIST MT, ACE	BN, BC, NW, WB	create, harvest, acquire
English > Chinese	200K +	NIST MT, ACE, GALE	BN, NW, WB	create
English>Arabic	100K +	ACE, GALE	BN, NW, WB	create
Bengali>English *Pashto>English* *Punjabi>English* *Tagalog>English* *Tamil>English* *Thai>English* *Urdu>English* *Uzbek>English*	250-500K + per language pair	NIST MT, REFLEX-LCTL	NW, WB	create, harvest

Table 1. Recent LDC Translation Efforts

3 Enhanced Infrastructure for Parallel Text Creation

3.1 Data Pipeline

Since manually-created parallel text is in high demand to tackle genre- and language-specific issues for MT system training and evaluation, LDC has focused a lot of effort toward improving the infrastructure and pipeline so that a large amount of parallel text of various genres and languages can be created in a short timeline, with relatively low cost. Generally, all manually translated parallel text now follows the standardized pipeline shown in Figure 1. This includes data selection, sentence segmentation, translation format conversion, manual translation, and data validation. The pipeline improves workflow management functionality and ensures parallel text aligned perfectly at the sentence level.

[1] Genre description: BN is broadcast news, BC is broadcast conversation, WB includes weblog, newsgroup and BBS posts, NW is newswire.

[2] Methodology description: create is manual translation, harvest is scraping parallel text resources from the web, acquire is translation resources acquiring from known repositories.

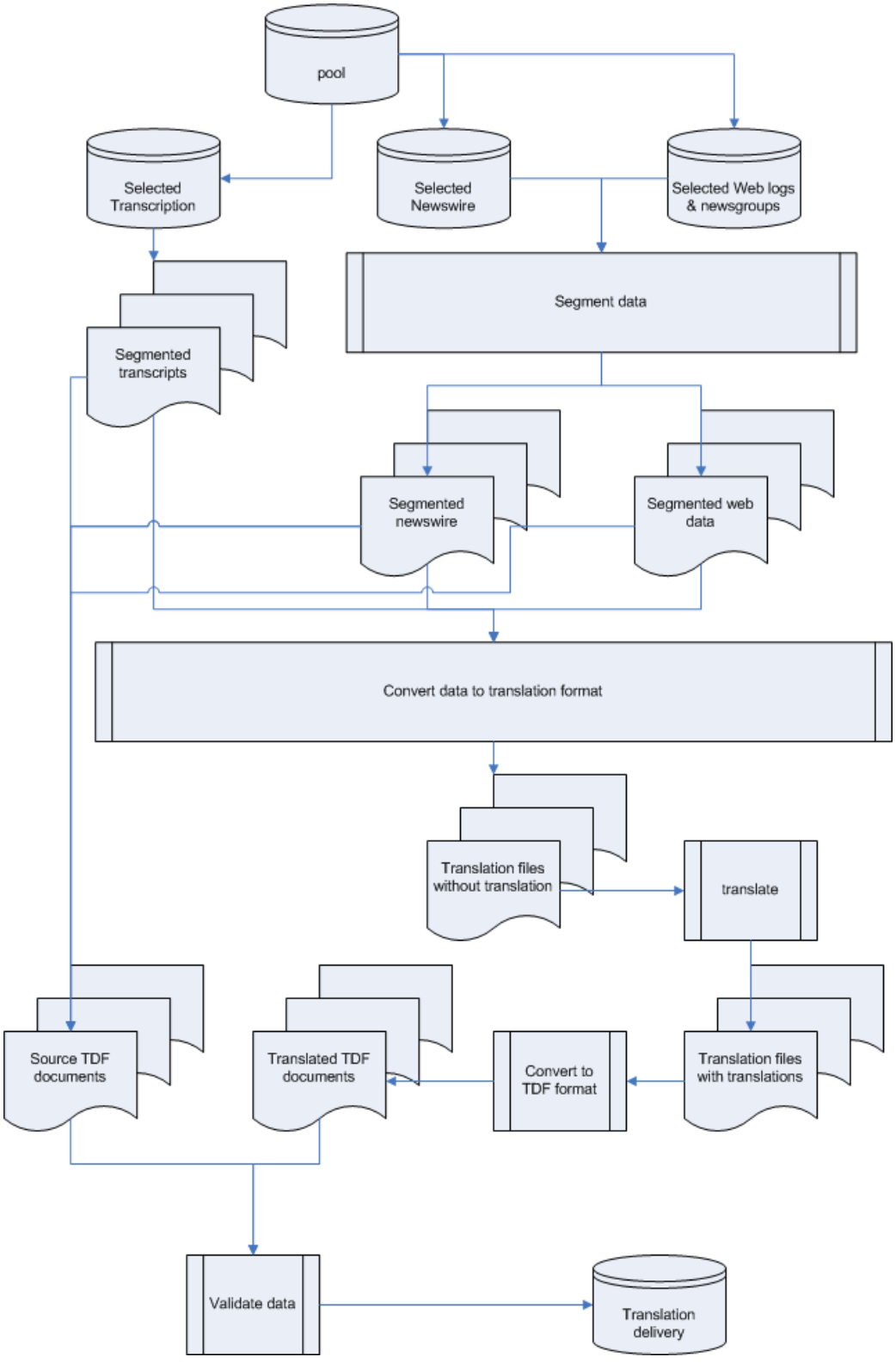

Figure 1. Standardized Translation Pipeline

3.2 Tracking and Management Database

In addition to designing and standardizing the manual translation pipeline, LDC has created a translation database that tracks all manually-translated parallel texts as they progress through the translation pipeline, as shown in Figure 2. The database tracks the following information for every file: translation agency and team, duration of translation process, payment information, and status of quality control stages. It also allows us to provide quick and exhaustive information of the parallel texts to other downstream annotation tasks, such as Treebank and Word Alignment annotation.

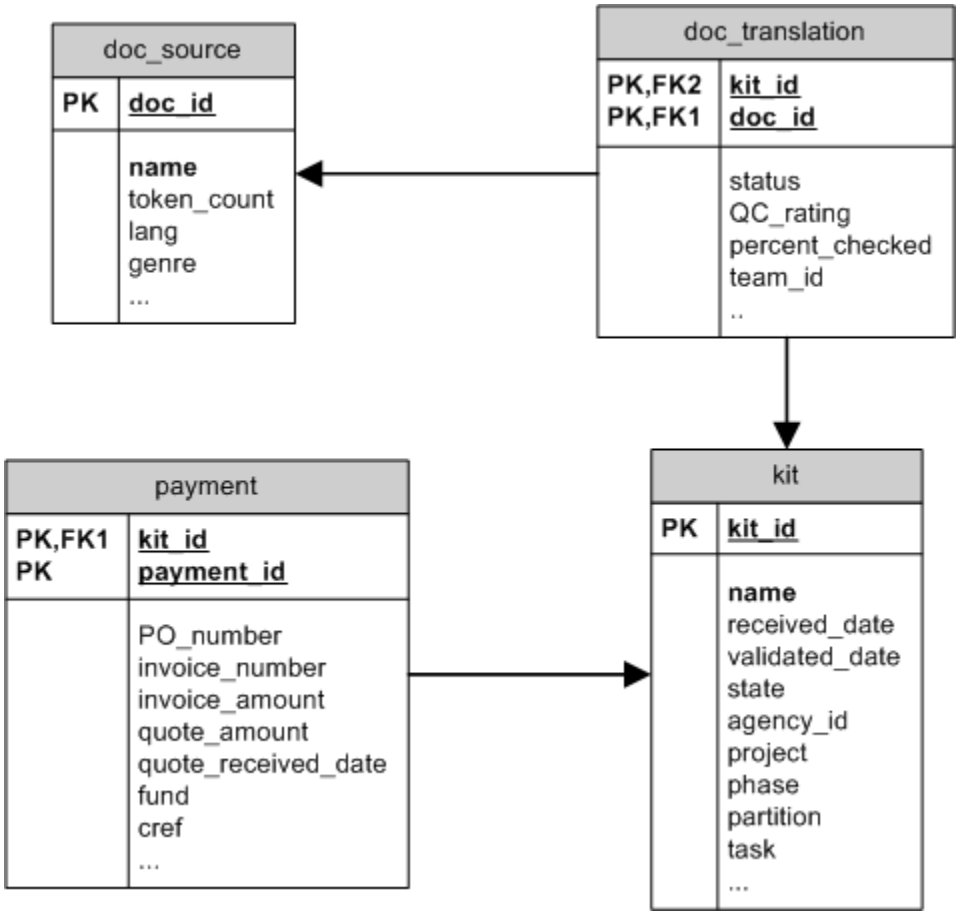

Figure 2. Translation Database

3.3 Selection of Data for Translation

Generally, all manually-created translation resources undergo a standard process of manual selection. Lists of candidate files are presented to annotators for manual judgment. Factors like sources, topics, difficulty, and epoch are balanced to suite specific needs of users and sponsors. After selection, LDC conducts checks over the data pool to detect and remove files with undesired

features, such as traditional Chinese characters in the Chinese data, or duplicated content within and across datasets and packages.

As statistical MT systems are trained with more and more parallel text, they require "novel" training data that uses resources effectively. To respond to this need, LDC has begun creating parallel text with sentence-based selection by calculating the frequency of n-grams and perplexity of all sentences in the pool of existing MT training data and generating list of candidate sentences of the data pool[3]. Sentences with the highest N-grams are reviewed and selected for manual translation. This process raises some challenges in data processing and handling, and the benefits of this approach have yet to be reported by system developers.

3.4 Guidelines for Translators

LDC has improved translation guidelines (LDC, 2009) to address challenges and issues regarding specific genres and languages. The general structure of the guidelines is stable, which describes translation data format, data delivery methods, and translation QC. LDC has specific guidelines that address translation from Arabic (A), Chinese(C) and Urdu(U) to English(E). These guidelines can be easily adapted to other languages. For each language pair (A>E, C>E, U>E), the guidelines demonstrate good and bad translation examples and include instructions for handling language-specific issues, such as pro-dropping, serial verbs, proper names, numbers, punctuations or idiomatic phrases.

The guidelines also provide instructions for managing specific challenges in audio transcripts and web files. For transcription data, the guidelines include rich examples of disfluencies like filled pauses, partial words, restarts, and speaker noises. An effort has also been made in normalizing transcription and translation mark-up for system development and downstream annotation tasks.

Translation guidelines and related documentation is available at http://projects.ldc.upenn.edu/gale/Translation/

3.5 Quality Control

For training data, LDC provides regular, targeted feedback to translation teams by reviewing 10% of each delivery. The feedback is provided in 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. These QC reports also provide LDC with detailed records to consult before assigning new translation work.

[3] The sentence selection scripts were kindly provided by IBM (Arabic selection) and SRI (Chinese selection), which are both GALE participants.

For evaluation data, LDC reviews 100% of the eval pool to correct and improve the manual translation. “Alternatives” are added to address source ambiguity, following existing translation alternative guidelines. LDC currently has alternative guidelines for both Arabic and Chinese languages. To support evaluation translation QC, LDC built a new translation QC toolkit that uses components of the XTrans speech transcription and annotation tool and enables more extensive QC and improved real time rates (Friedman, 2008).

3.6 Parallel Text Harvesting

LDC has also developed a set of software tools for identifying potential parallel text resources in a large pool of online multilingual documents. LDC regularly runs 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 set of SGML or XML markups. The document mapping module of the BITS system is then run to identify pairs of possible parallel documents. Once pairs are identified, we segment each document into sentences using a sentence-segmenter and then run the Champollion sentence aligner (Ma, 2006) to create sentence mapping tables. For the GALE program, LDC has created and distributed over 82,000 document pairs of potential parallel text in Arabic and English, and 67,000 document pairs of potential parallel text in Mandarin Chinese and English.

4 Conclusion

For the above mentioned programs, LDC has created and distributed dozens of translation corpora (both parallel text and multiple translations) to the MT program participants. LDC will wherever possible distribute them more broadly, for example, to its members and licensees, through the usual mechanisms including publication in LDC's catalog or on its website, and several GALE program translation resources have already appeared in LDC's catalog:

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

With an improved infrastructure, pipeline, tracking database and QC procedures, as well as stable translation guidelines which are easy to adapt to address new challenges and program requests, LDC has created a large amount of parallel text with high efficiency and controlled costs for multiple programs, including DARPA TIDES and GALE, the NIST Open Machine Translation Evaluation series, the REFLEX-LCTL program, and ACE. The infrastructure and guidelines described here are easily tailored to programs with specific requirements for addressing language- and genre-specific issues. The improved techniques for harvesting existing online translation resources enable LDC to collect significant amounts of parallel text at minimal cost.

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.

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,* Marakesh, Morocco

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,* Marakesh, Morocco

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

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

NIST (National Institute of Standards and Technology) (2006). *NIST 2008 Open MT Evaluation*, http://www.nist.gov/speech/tests/mt/2006

NIST (National Institute of Standards and Technology) (2008). *NIST 2008 Open MT Evaluation*, http://www.nist.gov/speech/tests/mt/2008

NIST (National Institute of Standards and Technology) (2009). *NIST 2008 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). Enitty Translation And Alignment in the ACE-07 ET Task. In *Proceedings of LREC-2008,* Marakesh, 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