Parallel Aligned Treebanks at LDC: New Challenges Interfacing Existing Infrastructures

Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Mohamed Maamouri, Ann Bies, Nianwen Xue

Linguistic Data Consortium, University of Pennsylvania
Philadelphia, PA 19104 USA
Computer Science Department, Brandeis University
Waltham, MA 02453 USA
Email: {xuansong,strassel,sgrimes,safa,maamouri,bies}@ldc.upenn.edu xuen@brandeis.edu

Abstract

Parallel aligned treebanks (PAT) are linguistic corpora annotated with morphological and syntactic structures that are aligned at sentence as well as sub-sentence levels. They are valuable resources for improving machine translation (MT) quality. Recently, there has been an increasing demand for such data, especially for divergent language pairs. The Linguistic Data Consortium (LDC) and its academic partners have been developing Arabic-English and Chinese-English PATs for several years. This paper describes the PAT corpus creation effort for the program GALE (Global Autonomous Language Exploitation) and introduces the potential issues of scaling up this PAT effort for the program BOLT (Broad Operational Language Translation). Based on existing infrastructures and in the light of current annotation process, challenges and approaches, we are exploring new methodologies to address emerging challenges in constructing PATs, including data volume bottlenecks, dialect issues of Arabic languages, and new genre features related to rapidly changing social media. Preliminary experimental results are presented to show the feasibility of the approaches proposed.

Keywords: machine translation; word alignment; parallel aligned treebank

1. Introduction

PATs are parallel treebanks annotated with morphological and syntactic structures that are aligned at sentence as well as sub-sentence levels. They are valuable resources for natural language processing and other research fields, such as automatic word alignment system training and evaluation, transfer-rule extraction, word sense disambiguation, translation lexicon extraction, and cultural heritage and cross-linguistic studies. With machine translation particularly, PAT data can help to improve system performance with enhanced syntactic parsers (Marecek, 2011), with better learned empirical rules for flexibly capturing factual and linguistic knowledge of language pairs (Simov et al, 2011), with effective syntactic re-ordering of languages that are far-apart (DeNero & Uszkoreit, 2011), and with reduced automatic word error rate (Ittercheriah & Roukos, 2005). However, such corpora are hard to find. Prominent PAT efforts are the Japanese-English-Chinese PAT (Uchimoto et al. 2004), the English-German-Swedish PAT (Volk et al., 2006), and the SMULTRON corpus built in English, Swedish, and German (Gustafson-Capkova et al., 2007). A recent contribution is the PAT data at LDC, notable for its large volume (Table 1). It is a part of the GALE Program funded by DARPA (Defense Advanced Research Projects Agency), targeting Arabic, Chinese and English languages.

The GALE program significantly improved MT technologies while exposing impediments for further MT breakthroughs. Arabic-English translation systems don't port well to dialect texts. Models for both Arabic and Chinese have to handle unfamiliar genres due to rapidly changing social media. The BOLT (Broad Operational Language Translation) Program, another five-year DARPA program in the wake of GALE, aims to tackle these issues with the ultimate goal of enhancing MT accuracy. To support this objective, the paper envisions several avenues of future PAT creation at LDC, emphasizing alignment strategies. The paper is organized as follows: Section 2 focuses on existing PAT infrastructures and methodologies; Section 3 points out potential challenges and discusses corresponding approaches to meet these challenges; and Section 4 is the conclusion.

	Arabic-English PAT			
Genre	*Arabic Words*	*ATB Tokens*	*English Words*	*Segments*
NW	198558	290064	261303	8322
BN	201421	259047	266601	12109
BC	n/a	n/a	n/a	n/a
WB	19296	28138	26382	853
	Chinese-English PAT			
Genre	*Chinese Characters*	*English Words*	*CTB Words*	*Segments*
NW	240920	164161	145925	5322
BN	n/a	n/a	n/a	n/a
BC	176448	91650	122714	7156
WB	129594	89866	82585	3920

Table 1: PAT Corpora at LDC

2. Existing Infrastructures and Methodologies

2.1 Arabic-English PAT

The "Arabic language", as a general term, includes Modern Standard Arabic (MSA) and a collection of spoken dialects. "Arabic" herein refers to MSA.

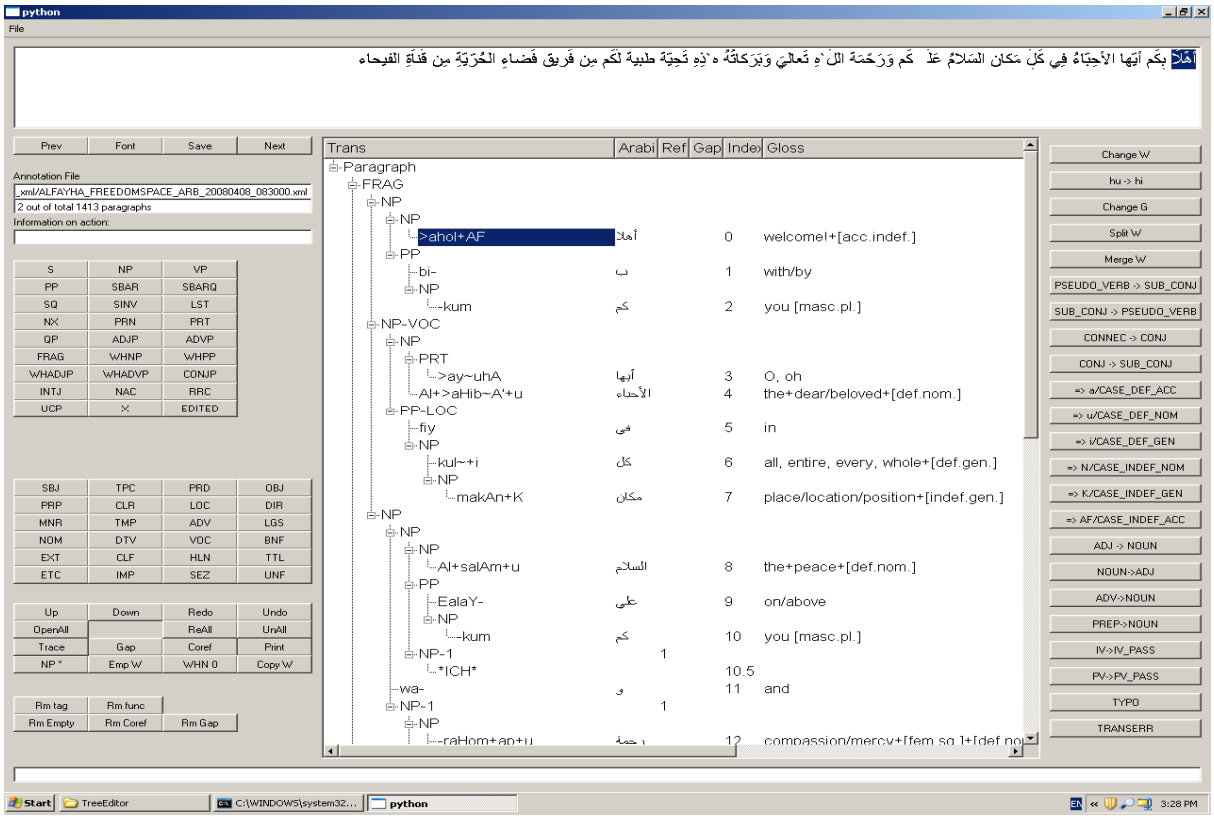


Figure 1: Arabic Treebank Annotation Tool

The MSA-English PAT data has four genres: newswire, broadcast news, broadcast conversation and weblogs. Source data are harvested from TV/broadcast programs, news agencies or online resources and subsequently sentence-segmented and translated into English. Treebank annotations are further conducted on source texts and their translations. The final PAT corpora are created with an infrastructure that integrates existing parallel treebank annotations through two levels of alignment with corpus-wide data mapping and indexing.

Treebank annotations are taken from the Penn Arabic Treebank (ATB) and its corresponding Penn English Treebank (EATB). The ATB and EATB are composed of two annotation levels: morphological/part-of-speech (POS) and syntactic/tree annotation. The former provides part-of-speech, morphological, and gloss information. The latter focuses on the constituent structures of sentences and word sequences, analyzing functional categories for each non-terminal node, and identifying null elements, co-reference, traces, etc. (Maamouri et al., 2008; Maamouri & Bies, 2004). English Treebanks were also created at LDC (Bies et al., 2012). LDC developed an annotation tool to assist treebank annotation (Figure 1).

The sentence alignment is performed during the translation stage and automatically re-aligned during the PAT process in the case of misalignment. The word-level alignment is performed on leaf tokens, producing ground-level alignment. Tokens are directly extracted from the Penn English and Arabic treebanks. The English texts are tokenized in the following ways: words separated by white spaces, contractions split, punctuations separated from surrounding words, and the apostrophe (',s) treated as a separate token. Hyphens are separate tokens, but some of them are treated as part of words. Because of rich morphological features, tokenization or segmentation of Arabic texts is more complicated. We adopt Arabic treebank tokenization which splits clitics (except "determiner") into separate tokens to allow for finer alignment. Annotation markups, such as "empty category" markers from treebank annotation, are tokenized into separate tokens. Punctuation marks are also separate tokens.

Tokenized tokens are treated as minimum units for alignment because they are atomic semantic units during translation. We adopted two methods for aligning words (for both Arabic-English and Chinese-English alignments, see Chinese examples in Figure 2 and 3): minimum-match

and attachment. The minimum match approach aims to identify complete and minimal semantic translation units, generating alignments of minimal semantic units which may be one-to-one, many-to-one or many-to-many links (Figure 2). The attachment approach is proposed to deal with unaligned words. The unaligned words, also known as "spurious words", are usually contextually or functionally required for semantic equivalence. However, they do not have surface structure translation equivalence. The attachment method attaches these unaligned words to constituent head words to indicate phrasal constituent dependency or collocation dependency (Figure 3). The unaligned words at the sentence or discourse level are not attached to any words. The words attached are also labeled with the "GLUE" tag in Arabic PATs.

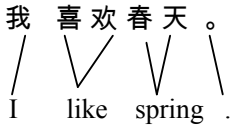
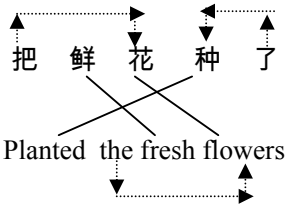


Figure 2: Minimum Match Alignment



(Note: unaligned 把, 了 and "the" attached to head words)
Figure 3: Attachment Approach

Two types of alignment links ("translated-correct" and "translated-incorrect") and two types of word markups ("not-translated correct" and "not-translated incorrect") are designed to capture general linguistic information and language specific features. The "translated-correct" links are the most common alignment links, indicating valid translation pairs. The "translated-incorrect" link type covers instances of erroneous translations lexically, grammatically or both. "Not-translated incorrect" is applied to cases with a loss of semantic meaning and an absence of surface structure representation. For unaligned words, such as omissions or insertions of words, we use the "not-translated correct" markup for capturing cross-lingual features (Figure 4).

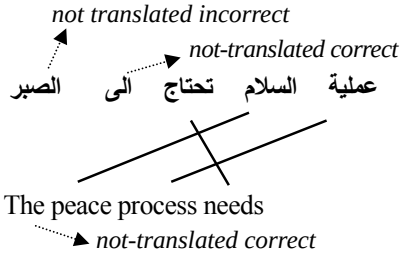


Figure 4: Word Markups

The alignment and treebank annotations are presented in four types of files: raw, tokenized, word aligned, and treebanked documents. Files with an identical base filename have the same number of lines, and the annotations of a specific line share the same line and token numbers. In the treebank and alignment files, there are no physical token strings. Instead, there are only token IDs corresponding to token strings in tokenized files. Trees are structured in labeled brackets of the Penn treebank format, where the tree leaves contain POS tags, with token IDs corresponding to the numbers in the tokenized file. In most cases, there is one tree per line. In the cases of multiple trees on one line, they are separated by whitespace. In a word alignment file, each line contains a set of alignments for a given sentence. The alignments are separated by spaces. Each alignment is represented in the format of "s-t(linktype)", *s* and *t* being a list of comma delimited source and translation token IDs respectively.

2.2 Chinese-English PAT

The Chinese-English PAT is built in a way similar to the Arabic PAT except for a slight difference in the alignment infrastructure. An additional level of alignment was introduced to form a three-level alignment which we further enriched with linguistic tags. Treebank annotations for the Chinese-English PAT are taken from the Penn Chinese treebank (CTB) and its corresponding English treebank (ECTB). Both treebanks are segmented, POS tagged, and syntactically-annotated.

A particular feature of CTB data is that, before the treebank process, source Chinese data are segmented into leaf tokens according to the word segmentation scheme proposed by the Penn Chinese treebank team (Xue et al., 2005). Word segmentation is a challenging data pre-processing step required for many Chinese NLP applications because of lack of word boundaries. The simplest kind of segmentation is character tokens. More sophisticated segmentation schemes group one or more characters into a word. CTB tokens are of the latter kind. As the CTB tokens sometimes are not the minimum translation units, we cannot use them as the minimum atomic units for ground or base level alignment for PATs. We further tokenize CTB tokens into individual characters for the manual character-level alignment. The CTB-word level alignment is then automatically generated from this ground-level character alignment. Therefore, we have a three-level alignment infrastructure for Chinese-English PATs. Like the Arabic-English alignment, we also applied the same two alignment approaches for the Chinese PAT (Figure 2 and 3).

To improve automatic word alignment and ultimately MT quality, researchers are exploring the possibility of incorporating additional linguistic information to word alignment. This was also a research focus for the GALE program. We manually tag special linguistic features on top of character-level alignment and automatically propagate them to the alignment at the CTB word level.

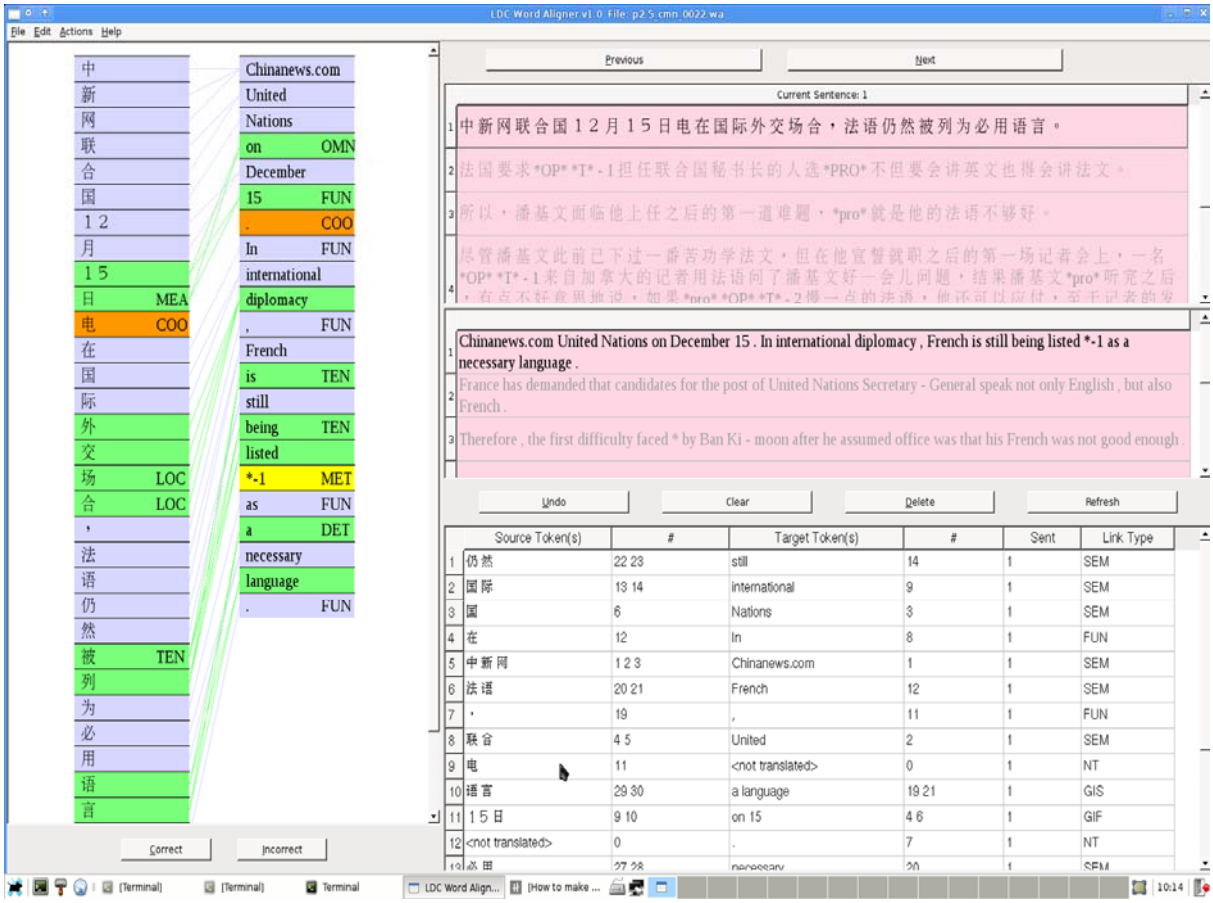


Figure 5: Alignment and Tagging Annotation Tool

To more comprehensively address linguistic idiosyncrasies of both Chinese and English, including the Chinese particle 的 (DE), we further classified "translated correct links" into seven link types, and further classified the "not-translated correct" markup tag into fourteen word tags. These link and word types are illustrated in Table 2 and 3 (Li et al., 2010). LDC developed an interface to support the manual alignment and tagging annotation (Figure 5).

Chinese-English PAT corpora are structured and formatted similarly to Arabic-English PAT data, except that in the alignment file, tags are attached to an alignment pair. A word can have multiple tags, each corresponding to a character within the word. The tags are formatted for the convenience of easy removal since some less sophisticated translation systems may not need all such tagging information. As all the unaligned and attached words are categorized and tagged, users can choose to retain and investigate a particular type or several types of linguistic features. Therefore, the tagging data structure is designed to facilitate use for all models and by all users.

Link Tags	*Examples*
Semantic	这所(this)大学(university) [this ***university***]
Function	***在***(in)这个(this)森林(forest)[***in*** this forest]
Grammatically-inferred	***把项目*** (project)提交(submit) [finish this ***project***]
Contextually-Inferred	欢迎***收听***BBC***新闻***[Welcome to BBC ***news***]
DE-clause	出版(publish)***的***书(book) [book ***that*** was published]
DE-modifier	春天(spring)***的***(of)天空(sky) [the sky ***of*** the spring]
DE-possessive	大学(university) ***的***(from)教授(professors) [professors ***from*** the university]

Table 2: Link Types

Word Tags	*Examples*
Omni-function-preposition	***将*** 项 目 (project) 完 成 (finish) [finish the project]
Tense/passive	提 出 (raise) ***的*** 问 题 (issue) [the issue raised]
Measure word	五 (five) ***家*** 商 店 (shop) [five shops]
Clause marker	他(he)犯(made)错(mistake) [the mistake ***which*** he made]
Determiner	老 师 (teacher) 没 (not) 来 (come)[***The*** teacher didn't come]
TO-infinitive	让(ask)他(him)发 言(talk) [ask him ***to*** talk]
Co-reference	校 长 (principal) 说 (said) 将 要 (would)…[The principal said ***he*** would….]
Possessive	学 校 (school) 师 (teacher) 生 (students) [the teachers and students ***of*** this school]
DE-modifier	跑 (run) ***地*** 慢 (slowly) [run slowly]
Local context	欢 迎 (welcome) ***收 看*** (CCTV) [Welcome to CCTV]
Rhetorical	美国(U.S.)学者(scholar)和(and) 中 国 (China) ***学 者*** (scholar) [scholars from the U.S. and China]
Sentence marker	冬天(winter)很(very)干燥(dry) ***的*** [Winter is very dry]
Context-obligatory	下雪 (rains)了[***It*** snows]
Non-context-Obligatory	不久(soon)我(I)***就***(already)出发 (left)了[Soon I left]

Table 3: Word Tags

3. New Challenges and Approaches

3.1 Scaling up PAT Production

Supervised MT methods rely on a considerable amount of PAT data to learn coherent language phenomena. High throughput of PAT creation is beyond our reach. LDC is currently exploring ways to break such bottlenecks.

3.1.1 Pre-automatic Alignment

For faster alignment, we plan to adopt a two-step semi-automatic annotation process, i.e., automatic alignment followed by human correction. Two aligners were compared and used to automatically produce the alignments: GIZA++ aligner (Och and Ney, 2003) and the Berkeley Aligner (Liang et al., 2006). GIZA++ produces one-to-one or one-to-many alignments while the Berkeley aligner has the potential to produce many-to-many links. We use an automatic alignment method that is high precision and low recall. We did an experiment with Arabic-English parallel texts and generated alignments using GIZA++. GIZA was run twice, taking turns with Arabic or English as the source language in the source-translation pair, as the alignment output is not symmetric. From these two alignment files we took the intersection -- the set of alignments appearing in both alignment direction files from GIZA (Table 4). We compared the GIZA-generated alignment with a gold-standard two-pass manual alignment and obtained the accuracies in Table 5. The result shows automatic alignments can expect to have about 72% precision. Annotators must correct the 28% of the alignments which are incorrect.

Size of Intersection (proposed alignments)	141677
Correct Alignments	102001
Incorrect Alignments	39676
Missed Alignments	116615

Table 4: Automatic Alignment Results

Precision	102001/(102001+39676) =0.71995
Recall	102001/(102001+116615) =0.46657
*F-score(2*precision*recall/precision+recall)*	0.56621146

Table 5: Automatic Alignment Accuracy

For Chinese-English alignment, we performed supervised and unsupervised automatic alignment. The unsupervised training was performed using approximately 200,000 tokens of Chinese newswire parallel text while supervised training would include manual alignments on the parallel texts. The precision of the supervised training was high, ranging between 90-93%. We also did experiments to measure annotator agreement and to compare the time used for pure human annotation with the time for correcting and annotating automatic alignments. Results show that when using GIZA++ alignments the agreement is at .91, and it is somewhat lower when not using GIZA++ alignments. The annotation speed on GIZA results is about 20% faster than annotation on blank files (Grimes et al., 2012). It is assumed that GIZA++ alignment could be further improved by deleting links from union alignments for higher precision, or adding links to intersection alignments for higher recall. A 20% increase in speed is significant but we hope for more.

3.1.2 Automatic Tagging

Automatic tagging was specifically designed for the Chinese-English alignment. The automation was performed on certain word types and on 6 types of alignment links out of the total of 7 types. The mechanism

for tagging alignment links was embedded into the alignment tool, labeling an alignment link when it is created based on word features. This automation accuracy is 100%, greatly saving tagging annotation time. For instance, if an annotator tags an unaligned word as "contextual marker", then the alignment containing this tagged word would be a "contextually inferred" link. A tagged function word can be used as an indicator to automatically tag a "grammatically-inferred" link. The automation is realized during the human annotation process, with the automatic mechanism embedded into the alignment tool. In contrast, the automatic tagging of words is a pre-processing step prior to human annotation. Two types of word tags are pre-tagged: MRK and MET. MRK is used to label annotation markup words inherited from upstream annotation such as "traces" markups in treebank and various markups from transcription annotation such as "speaker id". MET is used for format-related symbols or words. The automation of the two labels shows a high accuracy. Semi-automatic tagging is also run on certain types of words, such as "determiners", but the accuracy is low, needing further human correction.

3.1.3 Automatic Alignment of More Language Pairs

It's likely that future PAT corpora at LDC could be extended to more language pairs or dialects besides Chinese and Arabic (MSA), such as MSA-Iraqi or Iraqi-Levantine alignment. Automatic generation of a third language pair alignment (L1-L3) based on alignments of two existing manual language pairs (L1-L2 and L2-L3) with a pivot language (L2) shows the possibility of automatic (Arabic) dialect-English alignment based on existing manual MSA-English and dialect-dialect alignments. This automation allows us to reuse existing alignment resources and technologies. We successfully realized automatic generation of MSA-English alignment based on Iraqi-MSA alignment and Iraqi-English alignment.

3.2 Diversifying Annotation Resources

3.2.1 Arabic PAT

The GALE researchers significantly leveraged the state-of-the-art Arabic-English machine translation technology in translating MSA source sentences. However, dialects in predominantly Arabic MSA text have hindered progress in further raising Arabic-English MT performance. The BOLT project initiated a devoted effort towards this dialect issue. LDC's PAT is a part of this mission.

To address the dialect problem, a systematic study of dialect features is highly desired. The lexical and structural features of a particular dialect can be learned from a dialect-English PAT. To test whether the existing guidelines are applicable for aligning a dialect and English, we did annotation on two Iraqi and English files (10,000 words). The guidelines fit well with the new task and the result is encouraging. In annotation, we also detected some interesting and special dialect features which can be captured using our tagging mechanism. For instance, some words suggest regional varieties of the Iraqi dialect, such as شنهي ("what is it"), الوادم ("people"), and دحجت ("looked"), which are typically used in Southern Iraq. These mixed features of dialects could also be captured using our tagging mechanism.

Another potential dialect resource is the dialect-MSA PAT. This type of data is used to train systems to learn divergences and similarities between MSA and that dialect. Such knowledge enables a reuse of existing MSA resources and translation technologies. We investigated to see what features we can find by aligning and tagging a pair of dialect and MSA files. The conversation file we took in our example is in Iraqi, consisting of 1446 words. It was translated into MSA. We designed 5 tags for their similar or divergent features: 2 for lexical content words and 3 for function words. Results show that out of the 232 tagged Iraqi words, 98 words indicate an orthographical/lexical difference (there is no standard orthography for dialects), 63 a semantic match with context disambiguation, 65 a function word match, and 3 a mismatch in structure. The findings reveal great similarity between Iraqi and MSA in semantic meaning and conspicuous difference in orthography.

3.2.2 Chinese PAT

Compared to the Arabic-English translation quality, the Chinese-English translation accuracy was low due to wide syntactic differences between the two languages. With the Chinese-English PAT, there are two potential directions to expand the current infrastructure. One is to add more refined linguistic tags. Through a deeper classification of tags, machines can learn to disambiguate subtle word sense. However, incorrect use or overuse of tag-rich data may confuse less sophisticated systems. Therefore, we design and format tag information for convenient removal or selective use, supporting advanced as well as baseline model training. The other possible expansion is the alignment of units larger than characters or CTB words. Aligning structural units is a worthwhile future track for superior Chinese-English MT quality.

3.2.3 New Genres

Translation performance in both Arabic and Chinese language pairs is greatly degraded when tested on unfamiliar genres. The unknown word issue affecting all languages has been intensified by an influx of internet information and diversified social media. In genres such as text messages, emails, or conversations, the data domain has also deviated from the "news" focus. We'll extend genre types to include text messages, emails and free-style conversations for the next stage of PAT production at LDC. This type of data will be a new resource to study genre-idiosyncratic features and solve related issues.

We performed an initial study of the features of these

genres, including conversations in AOL, MSN, Yahoo!, and QQ, text messages via ATT, T-mobile, Verizon, Sprint, CMCC, Unicom, and a newly collected email dataset. Features observed can be briefly classified into, but not limited to: 1) inappropriate content and language, such as obscene, sexual and threatening language, or words or phrases associated with fraudulent schemes, chain letters and other common types of unsolicited email or messages; 2) non-standard use of punctuation or signs such as multiple exclamation marks "!!!", question marks "???" or special "%&#$@*" characters, or use of all capital letters; 3) frequent use of emoticons; 4) informal abbreviations, slang, acronyms, etc. in text messages, such as "lol" and "IMO"; 5) numbers representing words such as numbers 520 in Chinese for "I love you" (我爱你) and 748 for (去死吧) "go to hell"; 6) old words taking new meaning such as "杯具", originally only indicating "cups", now having the new meaning "tragedy"; 7）new coined words like "宅男" (referring to "youth staying at home addicted to internet, computer games, etc."). These new language features are posing challenges to data processing and annotation for PATs. The inability to handle and describe these features will eventually affect translation quality. Some of these features can be handled during the data collection stage, such as numbers or emoticons, while other features, such as genre idiosyncratic features, can be described utilizing our alignment and tagging structure.

4. Conclusion

This paper describes methodologies and the existing infrastructure for creating PAT corpora at LDC for the GALE program, including the annotation process, challenges and approaches. While interfacing the existing infrastructure, we are exploring new methodologies to address emerging challenges in constructing PATs for the BOLT program, such as the data volume bottleneck, dialect issues of Arabic languages, and new genres related to rapidly changing social media. Several feasible approaches have been proposed and experimented with to expand the existing framework, covering topics from automatic alignment and tagging for efficiency to the enriched multi-lingual, multi-layer, and multi-genre alignment for scaling up MT performance. The linguistic resources described in this paper will be made available to the broader research community via publication in LDC's catalog.

5. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, BOLT Program Grant No. HR0011-11-C-0145. 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

Bies A., Mott J., Warner C., Kulick, S. (2012). English Translation Treebank: An-Nahar Newswire. LDC Catalog Number: LDC2012T02.

DeNero, J., Uszkoreit, J. (2011). Inducing Sentence Structure from Parallel Corpora for Reordering. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pp. 193-203.

Grimes, S., Peterson, K., Li, X. (2012). Automatic Word Alignment Tools to Scale Production of Manually Aligned Parallel Text. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation.* Istanbul, Turkey.

Gustafson-Capkova, S., Samuelsson, Y., and Volk, M. (2007). SMULTRON (version 1.0) - The Stockholm MULtilingual Parallel Treebank. Department of Linguistics, Stockholm University, Sweden.

Ittycheriah, A. & Roukos, S. (2005). A Maximum Entropy World Aligner for Arabic-English Machine Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 89-96.

Li, X., Ge, N., Grimes, S., Strassel, S. M. and Maeda, K. (2010). Enriching word alignment with linguistic tags. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation.* Valletta, Malta.

Li, X., Ge, N., Strassel, S.M., (2009). Guidelines for Chinese-English Word Alignment. Linguistic Data Consortium, University of Pennsylvania. http://projects.ldc.upenn.edu/gale/task_specifications/GALE_Chinese_alignment_guidelines_v4.0.pdf

Li, X., Ge, N., Strassel, S.M., (2009). Tagging Guidelines for Chinese-English Word Alignment. Linguistic Data Consortium, University of Pennsylvania. http://projects.ldc.upenn.edu/gale/task_specifications/GALE_Chinese_WA_TaggingGuidelines_V1.0.pdf

Liang, P., Taskar, B., Klein D. (2006). Alignment by Agreement. In *Proceedings of NAACL 2009*.

Maamouri, M., & Bies, A. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In Proceedings of COLING 2004. Geneva, Switzerland.

Maamouri, M., Bies, A., Kulick, S. (2008). Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation.* Marrakech, Morocco.

Marecek, D. (2011). Combining Diverse Word-Alignment Symmetrizations Improves Dependency Tree Projection. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*. Tokyo, Japan. pp.144-154

Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19-51.

Simov, K., Osenova, P., Laskova, L., Savkov, A., Kancheva, S. (2011). Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment. *In Proceedings of the Second Workshop on Annotation and Exploitation of Parallel Corpora*. Hissar, Bulgaria.

Uchimoto, K., Zhang, Y., Sudo, K., Murata, M., Sekine, S. and Hitoshi, I. (2004). Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. *In Proceedings of the Workshop on Multilingual Linguistic Resources*. Geneva,

Switzerland. pp. 63-70.
Volk, M., Gustafson-Capková, S., Lundborg, J., Marek, T., Samuelsson, Y. and Tidström, F. (2006). XML-based phrase alignment in parallel treebanks. *In Proceedings of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*. Trento. pp. 93–96.
Xue, N., Xia, F., Chiou, F. and Palmer, M. (2005). The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2):207-238.