PaCMan : Parallel Corpus Management Workbench

Diptesh Kanojia
CSE, IIT Bombay
diptesh@cse.iitb.ac.in

Manish Shrivastava
CSE, IIIT Hyderabad
mani.shrivastava@gmail.com

Raj Dabre
GSI, Kyoto University
prajdabre@gmail.com

Pushpak Bhattacharyya
CSE, IIT Bombay
pb@cse.iitb.ac.in

Abstract

We present a Parallel Corpora Management tool that aides parallel corpora generation for the task of Machine Translation (MT). It takes source and target text of a corpus for any language pair in text file format, or zip archives containing multiple corresponding text files. Then, it provides with a helpful interface to lexicographers for manual translation / validation, and gives out the corrected text files as output. It provides various dictionary references as help within the interface which increase the productivity and efficiency of a lexicographer. It also provides automatic translation of the source sentence using an integrated MT system. The tool interface includes a corpora management system which facilitates maintenance of parallel corpora by assigning roles such as manager, lexicographer etc. We have designed a novel tool that provides aides like references to various dictionary sources such as Wordnets, Shabdkosh, Wikitionary etc. We also provide manual word alignment correction which is visualized in the tool and can lead to its gamification in the future, thus, providing a valuable source of word / phrase alignments.

1 Introduction

Statistical Machine Translation (SMT) depends primarily on parallel corpora. The quality of parallel corpora available can make or break a SMT system. The process of creating a parallel corpus is neither easy nor cheap. It is essential to produce good quality parallel corpus efficiently to ensure time and cost effectiveness of the process.

Traditional method of parallel corpus creation involves manual translation of every sentence by inputting a monolingual corpus and translating its each sentence. But, strict quality checks and skilled translators need to be employed to ensure correctness and, usually, the process of translation is followed by a validation phase to ensure quality and reliability.

The process of parallel corpora generation can be divided into the following phases: translation, validation and sentence alignment. Furthermore, to help SMT tools like Moses (Koehn et al., 2007), it would be desirable to manually correct word alignments generated by an automatic tool such as GIZA++ (Och and Ney, 2003).

We present a comprehensive workbench to streamline the process of corpora creation for SMT. This common workbench allows for corpora generation, validation, evaluation, alignment and management simultaneously. We aim to simplify the laborious manual task of corpora generation for all language pairs, and provide with aides at each step.

2 Related Work

There are a wide class of document management solutions and products which fall under the category of "corpora and text mining". We find that though a lot of effort has gone into creating tools to aid in corpora generation for lower level NLP tasks such as POS tagging and chunking, but not much work has gone in the direction of corpora generation aid for Machine Translation (MT). The few similar works that we did find are noted below.

PolyPhraZ (Hajlaoui and Boitet, 2004) is one such tool which helps in visualizing, editing and evaluating MT systems on parallel corpora. CasualConc (Imao, 2008) is a parallel concordancer which generates keyword in context concordance lines, word clusters, collocation analysis, and word counts.

MemoQ (Kilgray, 2006) and Trados (SDL, 2007) are also Computer Aided Translation (CAT)

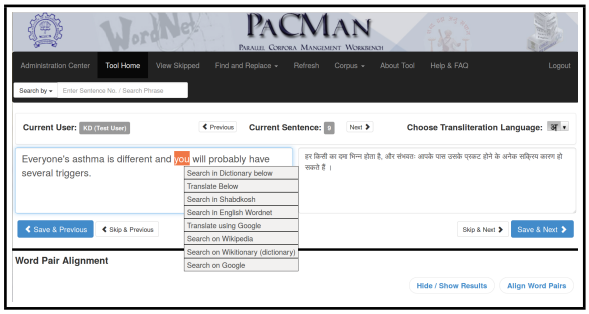

Figure 1: Snapshot of PaCMan on validation / translation screen

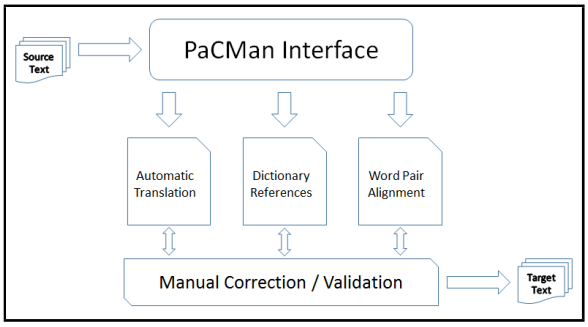

Figure 2: Workflow of the system

systems which are commercially available with features like Translation memory, and Term Extraction.

Wordfast is CAT system having just one free version "WordFast Anywhere". We studied and used the system, but found the interface less intuitive, and hard to use. "WordFast Anywhere" also has an integrated MT system which provides translations via Microsoft Bing and an integrated MT system.

Another system we came across is a web based text corpora development system (Yablonsky, 2003) that focuses on the development of UML-specifications, architecture and implementations of DBMS tools.

None of the above mentioned systems provide a word alignment visualization, which can be corrected manually, and saved to provide perfect phrase tables later.

3 Parallel Corpora Management System

Parallel Corpora Management System (PaCMan) (Figure: 1) is a platform-independent web-based workbench for managing all the processes involved in the generation of good quality parallel corpora. Along with covering the procedural / managerial aspects of the parallel corpora generation process, this tool also provides the means to track and manage the assignment and reporting of tasks in real-time.

As noted earlier, the tasks involved in the generation of a good quality parallel corpora can be broadly classified as follows:

- Translation
- Validation
- Sentence Alignment
- Word Alignment

Apart from these, there is also a need to manage the complete process. For an on-line tool such as the one we present, following activities need to be considered:

- User Registration for Superusers (Manager), translators / validators and aligners
- Task assignment and Tracking
- User session management

We have developed PaCMan to aid at every step of the process while ensuring high quality assistance. In the following sections we cover the various features that the tool provides.

3.1 Translation

Manual translation is the first step towards creating a parallel corpus. It is usually done by experienced translators who rely on their knowledge of source and target language to perform the task. But still, it proves a difficult task. From our experience in creating parallel corpus, we have learned that the task of manual translation can be made much easier if some assistances like dictionary aids, automated translations etc. are provided to the translators. We find that translators could do with some help on the following fronts:

- Translation of rare source language words
- Language input for non-standard scripts (Devanagari in case of Hindi and Marathi)
- Access to automatic translation of source text to speed up the translation process
- Access to previous translations of recurring phrases

In this tool, we have successfully addressed the above mentioned issues.

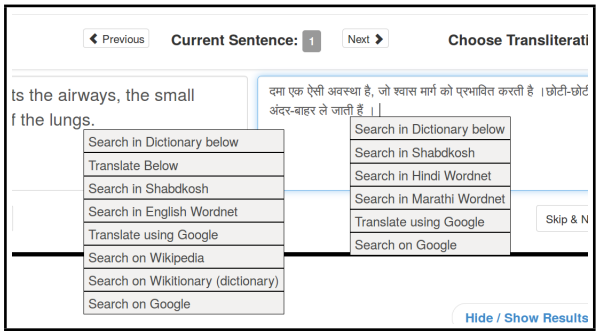

Figure 3: Snapshot of Right Click context menu

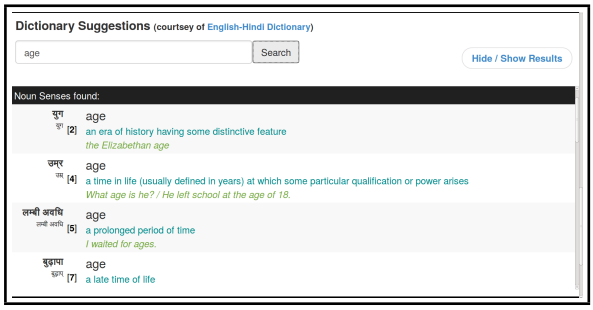

Figure 4: Snapshot of dictionary suggestions

3.1.1 Customized Context Menu Help

It is often the case that the translators, even with their substantial experience, find it hard to recall the meaning of some rare source language words. PaCMan helps overcome this difficulty by integrating various dictionary resources, and references. These dictionaries are provided on the translation screen as right-click menu items. Using these dictionaries the translator can get the meaning of a word simply by highlighting the word and choosing the preferred dictionary from the right-click menu (Figure: 3).

We have also integrated the corresponding UNL dictionary (Bhattacharyya, 2006) in the tool so that a user can select a word and can get its meaning on the same screen simply at the click of a button (Figure: 4). Further, we have integrated language specific Wordnets (Fellbaum, 1998) and (Dipak Narayan and Bhattacharyya, 2002), *ShabdKosh* (a Hindi-English dictionary) (Soni, 2003), Wikipedia and Wiktionary for these languages in the right-click menu. While these dictionaries provide the different meanings of a word, the user can also choose to translate a word or phrase directly by choosing "Google Translate" from the right-click context menu.

We found that the requirement of inputting data in Indian langugages can be easily overcome by integrating a transliteration API, using which the roman script should be converted to the chosed Indian script automatically. Thus, We integrated "Google Transliteration API" in our tool.

Figure 6: Snapshot of Word Alignment Box

Further, if the user wants to see the word's use in contexts other than the given source language sentence, she can choose to fire a Google search query by choosing "Search in Google" option from the right-click menu.

3.1.2 Automatic Translation

It is sometimes easier and faster to correct a translation than to translate from scratch. Also, even if a machine generated translation is syntactically or semantically incorrect, it can give the translator a kernel around which the final correct translation may be built.

Thus, we have integrated Sata-Anuvadak (Anoop Kunchukuttan and Bhattacharyya, 2014), a Moses Decoder based SMT system for multiple Indic languages with our tool which can generate an automatic translation of a given English sentence at the click of a button. In the worst case, the translator would still need to manually translate but she would have gotten a fair idea of the possible translation, given some words get translated correctly.

3.1.3 Previous Translations

Parallel corpora generation very quickly becomes a repetitive task. A translator at times would like to find previously translated sentences. This has been made easy by introducing at ability to search for a word or a phrase and find the parallel sentences with the translations for that particular word or phrase (Figure: 5).

3.2 Validation

Validation or quality control of the translated sentences is the next logical step. Even the best translators can make mistakes, so it is imperative that we re-check the translations. Validation is very similar to translation except that the translations are already available. All the aids available to a

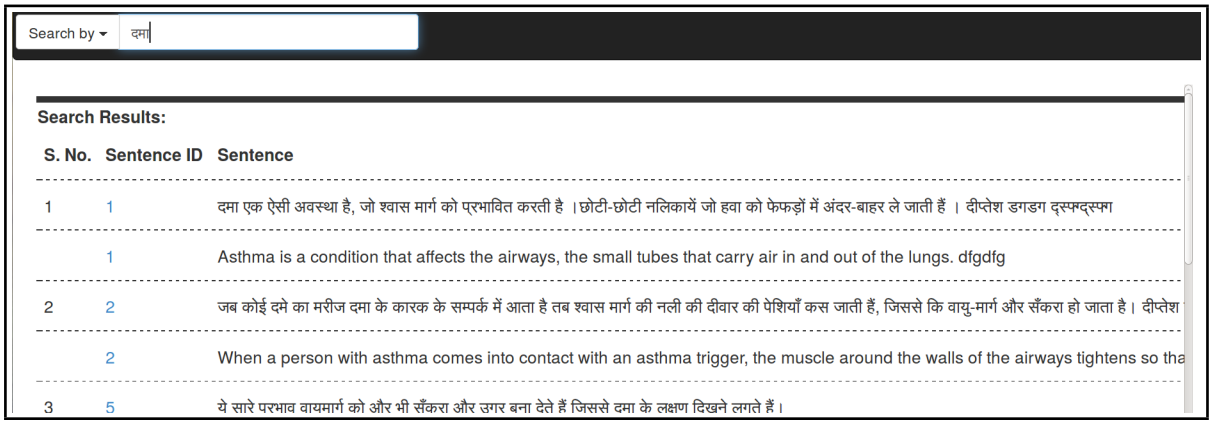

Figure 5: Snapshot of Search Results

translator are also available to a validator. A validator may change a sentence completely or make minor modifications as the need may be.

A validator can also be given fresh parallel corpus which is not generated on this work bench. Such corpus may be uploaded by any superuser responsible for a language pair.

3.3 Word Alignment Visualization and Gamification

The accuracy of a SMT system depends on the quality of word alignment produced by an aligner. For free word ordered languages and languages with rich morphology, aligners can often make many mistakes which eventually have dire effect on decoder accuracy. We aim to provide word level alignment for every sentence in the corpus (Figure: 6). It is expected that this will greatly enhance the performance of an SMT system trained on the alignment data provided by our tool. Currently, we have trained GIZA++ on nearly 48000 sentences.

We provide the means to augment the word alignments to eventually aid in higher accuracy for SMT. We have integrated GIZA++ aligner with our tool to allow a user to correct the alignments provided by it. A user might also choose to do the alignment completely manually using the visual drag and drop interface provided in the tool. Its rich interface lets you align words and phrases with a lot of ease. This can easily be pushed towards a crowd sourced word alignment collector with some gamification. We believe it can lead to a very high amount of perfect word / phrase alignments.

3.4 Administrative Tasks

Corpora Management System provides for various roles a user might need to play in the process of parallel corpus generation. The system is designed to work for multiple language pairs simultaneously. Each corpus can have its own superuser (manager). The superuser then distributes tasks among other users to work.

3.4.1 Corpora Upload and Download

Superusers can upload corpora to be translated, validated or aligned. Corpora upload is in the form of two plain text files in unicode encoding or two archives containing plain text files. For translation, only a single source file needs to be uploaded whereas for any other task two parallel files are needed. Once the assigned task has been performed the registered user can download complete or partial data or can commit the data which the superuser would finally commit to the master database.

3.4.2 Task Assignment

Superusers can assign tasks to different users. A superuser chooses complete or partial corpora and assigns it to a user who can perform the assigned task. At the time of corpora assignment a copy of the data is created specific to the user while the master database remains untouched. The master database can only be changed when the user finishes the task and sends the data back to the superuser. At which point the superuser commits the data back to the master database. A superuser can track the work done by any user who has been assigned a task by her.

3.4.3 Session Management

The system is equipped with both browser based and internal session management. When a user logs in, she is taken directly to the last sentence that was being validated / translated at the time of previous logout.

While navigating through the sentences a user can choose to skip sentences if she so desires. At any point of time she can choose to "View Skipped" sentences and work with them. When she is finished with the skipped sentences, she can go back to the normal view, landing on the last unsaved sentence.

4 Design

The system is developed in PHP and uses MySQL as backend. Various Javascript and jQuery snippets have b een developed along with AJAX to make sure the user stays on the same page. The system has been designed such that user would find every validation aid on the same screen, decreasing the shuffles between screens thus increasing productivity. The advantage of such a design is that the user would not have to change application for different tasks like referring to dictionaries etc.

5 Conclusion and Future Work

We have presented PaCMan, a web-based system that aids parallel corpora generation for the task of machine translation. The system provides an intuitive interface and many valuable features that increase the productivity at every stage of corpora generation process. It provides the user with dictionary references, and automatic translations during translation and validation of parallel corpus. The system also aids word pair alignments. It provides an authentication based corpora management system where multiple parallel corpora can be maintained. The system is open for use to everybody. We expect further extensions of our system will support the development of better parallel corpora and aid machine translation further.

We believe gamification of our word pair alignment feature will help in collecting perfect word / phrase pair alignments which can result in a better performing Moses decoder, and thus a higher accuracy of an SMT system. We aim at providing such an extension of our system soon.

References

Rajen Chatterjee Ritesh Shah Anoop Kunchukuttan, Abhijit Mishra and Pushpak Bhattacharyya. 2014. Shata-anuvadak: Tackling multiway translation of indian languages. In *Proceedings of the 9th Edition of its Language Resources and Evaluation Conference.*, LREC '14.

Pushpak Bhattacharyya. 2006. English - hindi dictionary. `http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/`.

Prabhakar Pande Dipak Narayan, Debasri Chakrabarti and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet - a wordnet for hindi. In *Proceesdings of First International Conference on Global WordNet*, Mysore, India.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Najeh Hajlaoui and Christian Boitet. 2004. Polyphraz: A tool for the management of parallel corpora. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, MLR '04, pages 109–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yasu Imao. 2008. Casualconc. `http://sites.google.com/site/casualconc/`.

Kilgray. 2006. memoq [computer software]. `http://kilgray.com/products/memoq`.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

SDL. 2007. Sdl trados studio [computer software]. `http://www.translationzone.com/products/sdl-trados-studio/`.

Manish Soni. 2003. Shabdkosh. `http://www.shabdkosh.com`.

Serge Yablonsky. 2003. The corpora management system based on java and oracle technologies. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 179–182, Stroudsburg, PA, USA. Association for Computational Linguistics.