Fast, Scalable Phrase-Based SMT Decoding

Anonymous ACL submission

Abstract

The utilization of statistical machine translation (SMT) has grown enormously over the last decade, many using open-source software developed by the NLP community. As commercial utilization has increased, there has been a pressing need that is optimized for their requirements. Specifically, faster phrase-based decoding, and more efficient utilization of modern multicore servers.

We present in this paper a re-assessment of the major components of phrase-based decoding and decoder implementation with particular emphasis on speed and scalability to multicore machines. The result is a drop-in replacement for the Moses decoder which is up to fifteen times faster and scales almost linearly with the number of cores. Furthermore, the decoder makes less search errors than the current Moses decoder.

1 Introduction

SMT has been one of the outstanding success story from the NLP community in the last decade. It has transition from a mostly research discipline to services such as Google Translate, Microsoft Translator Hub, as well as services and products built around offline products such as the open-source Moses toolkit. The latter has spawned a cottage industry encompassing a range of organizations and services from small language service providers that use SMT to reduce translation cost to large inter-govermental organizations such as the EU and the UN that provides high volume translation.

For high volume users, decoding is a largest and most critical part of the translation process which needs to be fast and efficient. However, it has been noticed that the Moses decoder, amongst others, is unable to efficiently use multiple CPU cores that are now common on modern servers (reviewed paper, github discussion). That is, the time taken to decode a test set does not substantial decrease when more cores are used, in fact, decoding time may increase when more cores are added. The issue will only become more noticeable as the commercial use of SMT grows and the number of cores in servers increases.

There have be speculation on the causes of the inefficiency as well as remedies. This paper is the first we know of that seeks to tackle this problem head on. We present an phrase-based decoder that is not only significantly faster than the Moses baseline for single-threaded operation, but is able to scale run multiple threads on multicore machines with only a slightly loss in linear speed. Model scores and functionality are compatible with Moses to aid comparison and ease of transition for users. All source code will be made available under an open-source license.

1.1 Prior Work

There are a number of open-source SMT projects, most includes a decoder. The most well known is Moses, which supports phrase-based models, hierarchical phrase-based as well as various syntax-based models. Joshua also supports hiearchical and syntax models and has recently supported phrase-based models. Phrasal supports a number of variants of the phrase-based model. CDEC supports hierarchical and syntactic models.

A number of the decoders support multithreading whilst others use alternative methods such as Hadoop or external scripts to parallelize decoding. We shall investigate the efficiency of using parallelizing decoding using the multi-processor approach. None of the decoder focus on multithreads decoding.

(Recently reviewed) describes running multiple processes of the Moses decoder for increased speed.

Other prior work look to optimizing specific components of decoding. (Liang and Chiang) describes the cube-pruning and cube-growing algorithm for decoding which allows the tradeoff between speed and translation quality to the adjusted with a simple parameter. (KenLM) and (DALM) describes fast, efficient datastructures for language models. (Zen) describes an implementation of a phrase-table for an SMT decoder that is loaded on demand, reducing the initial loading time and memory requirements. (CompactPT) extends this by compressing the on-disk phrase table and lexicalized re-ordering model resulting in impressive speed gains over previous work.

(mtplz) is perhaps closest in intent to this work. This takes a wholistic approach to decoding, describing a novel decoding algorithm which is fis focused on better decoding speed. It also describes a number of implementation details for faster decoding. However, the decoding algorithm is only able to incorporate one stateful feature function which precludes some of the useful decoding configurations which contains multiple stateful feature functions. It does not include a load-on-demand phrase table, therefore, cannot be used in a commercial environment where phrase-table has not be filtered with a know test set for any realistic size phrase-table. Neither did this paper analyze the scalability of their work to multicore servers.

The rest of the paper will be broken up into the following sections. Next, we will describe the phrase-based model and the major implementation components, with particular emphasis on decoding time shortcomings. We will then describe modifications to improve decoding speed and present results. We conclude in the last section discuss suggested improvements and future work.

2 Phrase-Based Model

The objective of decoding is to find the target translation with the maximum probability, given a source sentence. That is, for a source sentence s, the objective is to find a target translation $\hat{t}$ which has the highest conditional probability $p(t|s)$. Mathematically, this is written as:

$$\hat{t} = \arg\max_t p(t|s) \qquad (1)$$

where the *arg max* function is the search. The log-linear model generalizes Equation 1 to include more component models and weighting each model according to the contribution of each model to the total probability.

$$p(t|s) = \frac{1}{Z} \exp(\sum_m \lambda_m h_m(t,s)) \qquad (2)$$

where λ_m is the weight, and h_m is the feature function, or 'score', for model m. Z is the partition function which can be ignored for optimization.

2.1 Beam Search

A translation of a source sentence is created by applying a series of translation rules which together translate each source word once, and only once. Each partial translation is called a *hypothesis*, which is created by applying a rule to an existing hypothesis. This process is called *hypothesis expansion* and starts with a hypothesis that has translated no source word and ends with a completed hypothesis that has translated all source words. The highest-scoring completed hypothesis, according to the model score, is returned as most probable translation, $\hat{t}$. Incomplete hypotheses are referred to as partial hypotheses.

Each rule translates a contiguous sequence of source words but successive translation options do not have to be adjacent on the source side, depending on the distortion limit. However, the target output is constructed strictly left-to-right from the target string of successive translation options. Therefore, successive translation options which are not adjacent and monotonic in the source causes translation reordering.

A beam search algorithm is used to create the completed hypothesis set efficiently. Partial hypotheses are organized into stacks where each stack holds a number of comparable hypotheses. Hypotheses in the same stack have the same coverage cardinality $|C|$, where C is the coverage set, $C \subseteq \{1, 2, ...|s|\}$ of the number of source words translated. Therefore, $|s| + 1$ number of stacks are created for the decoding of a sentence s.

There are two main optimization to the search that we shall investigate. Firstly, the search creates and destroy a large number of hypothesis objects in memory which puts a heavy burden on the operating system. Also, the speed of memory access is dependent on whether the data is in the CPU cache. We shall optimize the search algorithm to use memory pools and object pools, replacing the operating system's general purpose memory management with our own application-aware management. We shall also seek to re-use recently accessed information to increase likelihood of the data being in the CPU cache.

Secondly, we shall investigate different stack configurations other than coverage cardinality to see whether they can improve the speed / model score ratio.

2.2 Feature Functions

Features functions are the h_m in Equation 2, calculating a score for each hypothesis which is then weighted and summed to give an overall score.

Standard feature functions in the phrase-based model include:

1. log transforms translation model probabilities, $p_{TM}(t|s)$ and $p_{TM}(s|t)$, and word-based translation probabilities $p_w(t|s)$ and $p_w(s|t)$,
2. log transforms of the lexicalized re-ordering probabilities,
3. log transforms of the target language model probability $p(t)$,
4. a distortion penalty
5. a phrase-penalty,
6. a word penalty,
7. an unknown word penalty.

The first three feature functions frequently trained on data and require the feature to read the model from files. The other feature functions do not require model files. We shall investigate the first two feature functions for optimization.

2.3 Translation Model

Load-on-demand 'binary' phrase-tables are often used for MT deployment due to the advantages of fast loading and fast querying speed, and because they can be used with large phrase-tables. We therefore focus on optimizing decoding speed with these phrase-tables, specifically the Probing PT.

We shall look at the caching strategies to reduce the number of phrase-table lookups. We shall also investigate the datastructures used by the phrase-table and their impact on decoding speed.

2.4 Lexicalized Reordering Model

3 Experimental Setup

We trained two phrase-based systems using the Moses toolkit, with standard settings. The first system was trained on most of the publicly available Arabic-English data from Opus (Jrg Tiedemann, 2012,) consisting of over 69 million parallel sentences, and tuned on a held out set. The second system was trained on the French-English Europarl corpus. The phrase-tables were then pruned, keeping only the top 100 entries per source phrase, according to $p(t|s)$. All models files were then binarized; thelanguage models were binaized using KenLM (???), the phrase table using Probing PT (???), lexicalized reordering model using the compact datastructure described in ???. These binary formats were choosen for their best-of-class multithreaded performance. Table 1 gives details of the resultant sizes of the model files.

	ar-en	fr-en
Phrase table	17	5.8
Language model	3.1	1.8
Lex-re model	2.3	637MB

Table 1: Model sizes in GB

For testing decoding speed, we used a subset of the training data, Table 2. The two test set have differing characterics that we are interested in analyzing, ar-en have short sentences while fr-en have overly long sentences.

	ar-en	fr-en
# sentences	800k	200k
# words	5.8m	5.9m
Avg words/sent	7.3	29.7

Table 2: Model sizes in GB

Where we need to compare the model score of the algorithms, we used a held out set; ??? for ar-en and ??? for fr-en.

Standard Moses phrase-based configurations are used, except that we use the cube-pruning algorithm (???) with a pop-limit of 400, rather than the basic phrase-based algorithm. The cube-pruning algorithm is often employ by users who require fast decoding as it gives them the ability to trade speed with translation quality with a simple pop-limit parameter.

4 Results

5 BLAH BLAH

The following instructions are directed to authors of papers submitted to and accepted for publication in the ACL 2016 proceedings. All authors are required to adhere to these specifications. Authors are required to provide a Portable Document Format (PDF) version of their papers. The proceedings will be printed on A4 paper. Authors from countries where access to word-processing systems is limited should contact the publication chairs as soon as possible. Grayscale readability of all figures and graphics will be encouraged for all accepted papers (Section 6.8).

Submitted and camera-ready formatting is similar, however, the submitted paper should have:

1. Author-identifying information removed
2. A 'ruler' on the left and right margins
3. Page numbers
4. A confidentiality header.

In contrast, the camera-ready **should not have** a ruler, page numbers, nor a confidentiality header. By uncommenting `\aclfinalcopy` at the top of this document, it will compile to produce an example of the camera-ready formatting; by leaving it commented out, the document will be anonymized for initial submission. Authors should place this command after the `\usepackage` declarations when preparing their camera-ready manuscript with the ACL 2016 style.

6 General Instructions

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, as well as the authors' names and complete addresses (only in the final version, not in the version submitted for review), which must be centered at the top of the first page (see the guidelines in Subsection 6.4), and any full-width figures or tables. Type single-spaced. Do not number the pages in the camera-ready version. Start all pages directly under the top margin. See the guidelines later regarding formatting the first page.

The maximum length of a manuscript is eight (8) pages for the main conference, printed single-sided, plus two (2) pages for references (see Section 7 for additional information on the maximum number of pages).

By uncommenting `\aclfinalcopy` at the top of this document, it will compile to produce an example of the camera-ready formatting; by leaving it commented out, the document will be anonymized for initial submission. When you first create your submission on softconf, please fill in your submitted paper ID where `***` appears in the `\def\aclpaperid{***}` definition at the top.

The review process is double-blind, so do not include any author information (names, addresses) when submitting a paper for review. However, you should maintain space for names and addresses so that they will fit in the final (accepted) version. The ACL 2016 LaTeX style will create a titlebox space of 2.5in for you when `\aclfinalcopy` is commented out.

6.1 The Ruler

The ACL 2016 style defines a printed ruler which should be presented in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document without the provided style files, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\aclfinalcopy` command in the document preamble.)

Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g., the first line on this page is at mark 096.5), although in most cases one would expect that the approximate location will be adequate.

6.2 Electronically-available resources

ACL provides this description in LaTeX2e (`acl2016.tex`) and PDF format (`acl2016.pdf`), along with the LaTeX2e style file used to format it (`acl2016.sty`) and an ACL bibliography style (`acl2016.bst`) and example bibliography (`acl2016.bib`). These files are all available at `acl2016.org/index.php?article_id=9`. We strongly recommend the use of these style files, which have been appropriately tailored for the ACL 2016 proceedings.

6.3 Format of Electronic Manuscript

For the production of the electronic manuscript, you must use Adobe's Portable Document Format (PDF). This format can be generated from postscript files: on Unix systems, you can use `ps2pdf` for this purpose; under Microsoft Windows, you can use Adobe's Distiller, or if you have cygwin installed, you can use `dvipdf` or `ps2pdf`. Note that some word processing programs generate PDF that may not include all the necessary fonts (esp. tree diagrams, symbols). When you print or create the PDF file, there is usually an option in your printer setup to include none, all, or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. *Before sending it, test your* PDF *by printing it from a computer different from the one where it was created*. Moreover, some word processors may generate very large postscript/PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the postscript and/or PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying "Output to a file", then convert the file to PDF.

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble.

Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

6.4 The First Page

Center the title, author name(s) and affiliation(s) across both columns (or, in the case of initial submission, space for the names). Do not use footnotes for affiliations. Use the two-column format only when you begin the abstract.

Title: Place the title centered at the top of the first page, in a 15 point bold font. (For a complete guide to font sizes and styles, see Table 4.) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 1in from the top of the page, followed by a blank line, then the author name(s), and the affiliation(s) on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (e.g., "Mitchell," not "MITCHELL"). The affiliation should contain the author's complete address, and if possible, an electronic mail address. Leave about 0.75in between the affiliation and the body of the first page.

Command	Output
`{\"a}`	ä
`{\^e}`	ê
`{\`i}`	ì
`{\.I}`	İ
`{\o}`	ø
`{\'u}`	ú
`{\aa}`	å

Command	Output
`{\c c}`	ç
`{\u g}`	ğ
`{\l}`	ł
`{\~n}`	ñ
`{\H o}`	ő
`{\v r}`	ř
`{\ss}`	ß

Table 3: Example commands for accented characters, to be used in, e.g., BIBTeX names.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.25in on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The abstract text should be in 10 point font.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers in the camera-ready manuscript.

Indent when starting a new paragraph. For reasons of uniformity, use Adobe's **Times Roman** fonts, with 11 points for text and subsection headings, 12 points for section headings and 15 points for the title. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default; see section 6.3 above). Note that the latter is about 10% less dense than Adobe's Times Roman font.

6.5 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals.

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Using the provided LaTeX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972); this is accomplished with the provided style using commas within the `\cite` command, e.g., `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved.

References: We recommend including references in a separate `.bib` file, and include an example file in this release (`naalhlt2016.bib`). Some commands for names with accents are provided for convenience in Table 3. References stored in the separate `.bib` file are inserted into the document using the following commands:

```
\bibliography{acl2016}
\bibliographystyle{acl2016}
```

References should appear under the heading **References** at the end of the document, but before any Appendices, unless the appendices contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a reference as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Authors' full names rather than initials are preferred. You may use **standard** abbreviations for conferences[1] and journals[2].

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

Acknowledgment sections should go as a last (unnumbered) section immediately before the references.

6.6 Footnotes

Footnotes: Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[3] Footnotes should be separated from the text by a line.[4] Footnotes should be in 9 point font.

[1] https://en.wikipedia.org/wiki/List_of_computer_science_conference_acronyms

[2] http://www.abbreviations.com/jas.php

[3] This is how a footnote should appear.

[4] Note the line separating the footnotes from the text.

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
abstract text	10 pt	
captions	9 pt	
caption label	9 pt	bold
bibliography	10 pt	
footnotes	9 pt	

Table 4: Font guide.

6.7 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns and should be placed at the top of a page. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "**Figure 1:** Figure caption.", "**Table 1:** Table caption." Type the captions of the figures and tables below the body, using 9 point text. Table and Figure labels should be bold-faced.

6.8 Accessibility

In an effort to accommodate the color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. Here we give a simple criterion on your colored figures, if your paper has to be printed in black and white, then you must assure that every curves or points in your figures can be still clearly distinguished.

7 Length of Submission

The ACL 2016 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content, plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page (up to 9 pages with unlimited pages for references) so that reviewers' comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references. For both long and short papers, all illustrations and appendices must be accommodated within these page limits, observing the formatting instructions given in the present document. Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

8 Double-blind review process

As the reviewing will be blind, the paper must not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g., "We previously showed (Smith, 1991) ..." must be avoided. Instead, use citations such as "Smith previously showed (Smith, 1991) ..." Papers that do not conform to these requirements will be rejected without review. In addition, please do not post your submissions on the web until after the review process is complete (in special cases this is permitted: see the multiple submission policy below).

We will reject without review any papers that do not follow the official style guidelines, anonymity conditions and page limits.

9 Multiple Submission Policy

Papers that have been or will be submitted to other meetings or publications must indicate this at submission time. Authors of papers accepted for presentation at ACL 2016 must notify the program chairs by the camera-ready deadline as to whether the paper will be presented. All accepted papers must be presented at the conference to appear in the proceedings. We will not accept for publication or presentation papers that overlap significantly in content or results with papers that will be (or have been) published elsewhere.

Preprint servers such as arXiv.org and ACL-related workshops that do not have published proceedings in the ACL Anthology are not considered archival for purposes of submission. Authors must state in the online submission form the name of the workshop or preprint server and title of the non-archival version. The submitted version should be suitably anonymized and not contain references to the prior non-archival version. Reviewers will be told: "The author(s) have notified us that there exists a non-archival previous version of this paper with significantly overlapping text. We have approved submission under these circumstances, but to preserve the spirit of blind review, the current submission does not reference the non-archival version." Reviewers are free to do what they like with this information.

Authors submitting more than one paper to ACL must ensure that submissions do not overlap significantly ($> 25\%$) with each other in content or results. Authors should not submit short and long versions of papers with substantial overlap in their original contributions.

Acknowledgments

Do not number the acknowledgment section. This section should not be presented for the submission version.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.