Subspace Transform Thought for Bilingual Word Translation

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Continuous space's representations of words has showed good performance in many tasks of Natural Language Processing (NLP). Bilingual distributed representations of words can benefit Statistical Machine Translation (SMT). And some motheds for learning bilingual representations are to build the bridge between the source words and target words which have the same meaning by a linear transform matrix. Howover we found that most transform learning of bilingual represetations is complished between the entire source language space and the entire target language space, which make the transform matrix capture the local word-pair relationships badly.Instead of the entire spaces, we try to devide the subspaces of the source language appropriately, then we compute the transform matrix in the subspaces respectively. Experiments on bilingual word translation show that the proposed method can achieve better performance on a word similarity task and the several bilingual word translation task.

1 Introduction

Distributed representations has been studied for a long time which were raised with the language model before (). The CBOW and Skip-gram models can capture a large number of recise syntactic and semantic word relationships (**?**). In the representation space, similar-semantic words usually have close cosine distence such as vec("man") - vec("women") + vec("king") is close to vec("queen"). For the adventange of semanticse, word representations have been applied in many NLP tasks successfully(). Especially, neural network models in NLP prefer to build input matrix extracted from word representations ().

Statictial Machine Translation (SMT) is a sequence-to-sequence prediction task (**?**)), which receives the source sentence and predicts the target sentence. Phrase-base machine translation (MT) often performs better than word-base MT (). Bilingual word representations which include can help to improve MT results (**?**). Word repersentations of two languages can build a relationship by a linear transformation matrix (**?**). The mapping through the matrix shows in Figure 1.

However, we found that most methods to compute the linear transformation matrix are based on the whole source space and the whole target space. These methods utilize the word pairs which have the same or very similar meanings in source space and target space to compute a transformation matrix. The transformation matrix's dimension is fixed by the word embedding dimsension: that means whatever the word pairs' quantity is, matrix's shape will not change. If the source space is very large, the capacity of transformation can not be improved. Besides that, the transformation matrix in the entire space can not capture the local word-pair relationships well. Besides, as shown in Figure2, when two kinds of local word pair relation appear in the spaces: one is straight linear transformation and the other is rotational transformation, the only matrix cannot learn the relationships very well.

We propose a novel method for bilingual word representations. Firstly, we devide the subspaces of the source language appropriately by unsupervised learning method and record the subspace center. The quantity of the subspaces can be adjusted according to the size of the entire continuous space. And the transformation matrixs are computed in these subspaces respectively. This method can improve the capacity of linear mapping matrix in bilingual space. The thought of subspace can be used in the methods which utilize the linear relationship in bilingual space, such as ().

2 General Instructions

Manuscripts must be in single-column format. The title, authors' names and complete addresses must be centred at the top of the first page, and any full-width figures or tables (see the guidelines in Subsection 2.5). **Type single-spaced.** Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section 4. Do not number the pages.

2.1 Electronically-available resources

We strongly prefer that you prepare your PDF files using LaTeX with the official COLING 2016 style file (coling2016.sty) and bibliography style (acl.bst). These files are available in coling2016.zip at "Instructions for authors" section of `http://coling2016.anlp.jp`. You will also find the document you are currently reading (coling2016.pdf) and its LaTeX source code (coling2016.tex) in coling2016.zip.

You can alternatively use Microsoft Word to produce your PDF file. In this case, we strongly recommend the use of the Word template file (coling2016.dot) in coling2016.zip. If you have an option, we recommend that you use the LaTeX2e version. If you will be using the Microsoft Word template, we suggest that you anonymise your source file so that the pdf produced does not retain your identity. This can be done by removing any personal information from your source document properties.

2.2 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). PDF files are usually produced from LaTeX using the *pdflatex* command. If your version of LaTeX produces Postscript files, you can convert these into PDF using *ps2pdf* or *dvipdf*. On Windows, you can also use Adobe Distiller to generate PDF.

Please make sure that your PDF file includes all the necessary fonts (especially tree diagrams, symbols, and fonts with Asian characters). 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 PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the 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.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the paper. When working with `dvips`, for instance, one should specify `-t a4`.

If you cannot meet the above requirements for the production of your electronic submission, please contact the publication chairs as soon as possible.

2.3 Layout

Format manuscripts with a single column to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper are:

- Left and right margins: 2.5 cm
- Top margin: 2.5 cm
- Bottom margin: 2.5 cm
- Width: 16.0 cm
- Height: 24.7 cm

Papers should not be submitted on any other paper size. If you cannot meet the above requirements for the production of your electronic submission, please contact the publication chairs above as soon as possible.

2.4 Fonts

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. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

The **Times New Roman** font, which is configured for us in the Microsoft Word template (coling2016.dot) and which some Linux distributions offer for installation, can be used as well.

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	
captions	11 pt	
sub-captions	9 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

2.5 The First Page

Centre the title, author's name(s) and affiliation(s) across the page. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process.

Title: Place the title centred at the top of the first page, in a 15 pt bold font. (For a complete guide to font sizes and styles, see Table 1) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author's names(s), and the affiliation 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., use "Schlangen" not "SCHLANGEN"). Do not format title and section headings in all capitals as well except for proper names (such as "BLEU") that are conventionally in all capitals. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronical paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

Abstract: Type the abstract between addresses and main body. The width of the abstract text should be smaller than main body by about 0.6 cm on each side. Centre the word **Abstract** in a 12 pt 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 pt font.

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

Indent when starting a new paragraph. Use 11 pt for text and subsection headings, 12 pt for section headings and 15 pt for the title.

Licence: Include a licence statement as an unmarked (unnumbered) footnote on the first page of the final, camera-ready paper. See Section 2.9 below for details and motivation.

2.6 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. Do not number subsubsections.

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). Append lowercase letters to the year in cases of ambiguity. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972). Also refrain from using full citations as sentence constituents. We suggest that instead of

> "(Gusfield, 1997) showed that ..."

you use

> "Gusfield (1997) showed that ..."

If you are using the provided LaTeX and BibTeX style files, you can use the command `\newcite` to get "author (year)" citations.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g.,

> "We previously showed (Gusfield, 1997) ..."

should be avoided. Instead, use citations such as

> "Gusfield (1997) previously showed ... "

Please do not use anonymous citations and do not include any of the following when submitting your paper for review: acknowledgements, project names, grant numbers, and names or URLs of resources or tools that have only been made publicly available in the last 3 weeks or are about to be made public. Papers that do not conform to these requirements may be rejected without review. These details can, however, be included in the camera-ready, final paper.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation 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). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (Association for Computing Machinery, 1983).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

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**.

2.7 Footnotes

Footnotes: Put footnotes at the bottom of the page and use 9 pt text. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2]

[1] This is how a footnote should appear.

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

2.8 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Colour 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. Caption of the Figure." "Table 1. Caption of the Table." Type the captions of the figures and tables below the body, using 11 pt text.

Narrow graphics together with the single-column format may lead to large empty spaces, see for example the wide margins on both sides of Table 1. If you have multiple graphics with related content, it may be preferable to combine them in one graphic. You can identify the sub-graphics with sub-captions below the sub-graphics numbered (a), (b), (c) etc. and using 9 pt text. The LaTeX packages wrapfig, subfig, subtable and/or subcaption may be useful.

2.9 Licence Statement

As in COLING-2014, we require that authors license their camera-ready papers under a Creative Commons Attribution 4.0 International Licence (CC-BY). This means that authors (copyright holders) retain copyright but grant everybody the right to adapt and re-distribute their paper as long as the authors are credited and modifications listed. In other words, this license lets researchers use research papers for their research without liegal issues. Please refer to `http://creativecommons.org/licenses/by/4.0/` for the licence terms.

Depending on whether you use American or British English in your paper, please include one of the following as an unmarked (unnumbered) footnote on page 1 of your paper. The LaTeX style file (coling2016.sty) adds a command `blfootnote` for this purpose, and usage of the command is prepared in the LaTeX source code (coling2016.tex) at the start of Section 1 "Introduction".

- This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: `http://creativecommons.org/licenses/by/4.0/`

- This work is licenced under a Creative Commons Attribution 4.0 International License. License details: `http://creativecommons.org/licenses/by/4.0/`

We strongly prefer that you licence your paper as the CC license above. However, if it is impossible for you to use that license, please contact the publication chairs, Hitoshi Isahara and Masao Utiyama (`isahara@tut.jp, mutiyama@nict.go.jp`), before you submit your final version of accepted papers. (Please note that this license statement is only related to the final versions of accepted papers. It is not related to papers submitted for review.)

3 Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration "translation".

4 Length of Submission

The maximum submission length is 8 pages (A4), plus two extra pages for references. Authors of accepted papers will be given additional space in the camera-ready version to reflect space needed for changes stemming from reviewers comments.

Papers that do not conform to the specified length and formatting requirements may be rejected without review.

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.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

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.