|  |
| --- |
| **Final project literature review** |
| **Self-supervised pipeline for automated product replacement based on neural language models** |
| **Andrei Ionut Damian**  **XCS224U Spring 2020** |
|  |
|  |
|  |
|  |

Abstract

|  |
| --- |
| 000  001  002  003  004  005  006  007  008  009  010  011  012  013  014  015  016  017  018  019  020  021  022  023  024  025  026  027  028  029  030  031  032  033  034  035  036  037  038  039  040  041  042  043  044  045  046  047  048  049 |

In past years we have seen a lot of cross-domain application from various deep learning areas to particular real-life cases with apparently little connection. One of those areas is that of applying deep representation learning based on neural language processing to business analytics system. This document presents the summarization and synthesis of several approaches in the area of Natural Language Understanding as well as several papers relevant to the current state of the art in the area of recommender systems based on neural language processing and representation learning in general. This cross-domain analysis is required to further define and test the proposed hypothesis of inferring product replacement retail and distribution systems. We will try to review the proposed natural language understanding tools as well as define a potential outline of the proposed representation learning pipeline.

Introduction and problem definition

# Deep representation learning is arguably one of the most important areas of machine learning as it allows us to automatically – and potentially in un-supervised fashion – discover rich features of the raw data. Since the introduction of the well-known word-embeddings generation algorithms in early 2015, the area on natural language processing and understanding has seen a tremendous improvement. Although it might seem that the actual state-of-the-art resides strongly on contextual embeddings and/or complex multi-head self-attention architectures, it is all based on the initial “basic” steps in the area of semantic vector space models.

# Based on this rich area of research and experimentation a lot of development has been done in the area of applying NLP/NLU approaches to commercial transaction systems including recommender systems. One the most common approaches, further described later in section 2, is to view a retail transaction (a list of purchased items) as a natural language object – a context window, the whole retail transactional database as the actual text corpus and the item SKU database as the “word vocabulary”. This approach of creating a hypothesis similarity between the natural language representation learning and business analytics systems has proved successful in various cases and it is still a very active area of research.

# The intuition behind our proposed deep learning pipeline is that we could, at least in theory, generate, through multiple modeling iterations, powerful-enough semantic vector space embeddings for each individual item (product) so that we can infer replacements (synonym) items and propose them in the case on original item shortages – all of these in a self-supervised setting.

|  |
| --- |
| 050  051  052  053  054  055  056  057  058  059  060  061  062  063  064  065  066  067  068  069  070  071  072  073  074  075  076  077  078  079  080  081  082  083  084  085  086  087  088  089  090  091  092  093  094  095  096  097  098  099 |

# Papers summaries

In the following sub-sections, a deeper look will be given at various approaches and “tools” we plan to employ in our proposed experiment. We will analyze a mix of fundaments of deep representation learning including but not limited to semantic vector space model generation as well as representation learning based recommender systems. As our target is to clearly define our related concepts and approaches we will present each approach in a general perspective and also pinpoint our the take-aways.

## *GloVe*

One of the most important papers relevant to our subject is *“GloVe: Global Vectors for Word Representation”* (Pennington, Socher, & Manning, 2014) as it describes one the core representation learning methods required to generate sound semantic word embeddings beside the similarly well-known *word2vect* (Mikolov, Chen, Corrado, & Dean, 2013). The main intuition behind *GloVe* is based on the fact that statistics of word co-occurrence in a corpus is the primary source of information of the proposed unsupervised method for learning word representations. Thus, the proposed word vector generation pipeline first constructs the matrix of co-occurrence then applies a weighted least squares regression using as target the natural logarithm of the word counts as defined by the below objective function.

|  |
| --- |
| 150  151  152  153  154  155  156  157  158  159  160  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180  181  182  183  184  185  186  187  188  189  190  191  192  193  194  195  196  197  198  199 |

|  |
| --- |
| 100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144  145  146  147  148  149 |

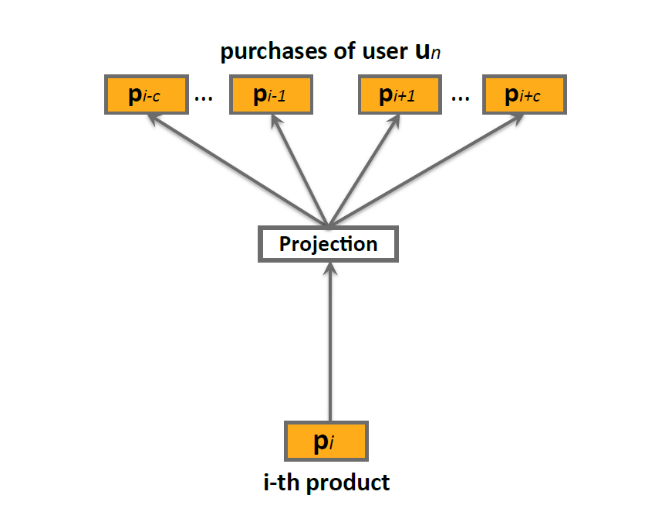
 (1)

Figure 1 - The skip-gram architecture similar with original word2vec paper. Image from original paper by Grbovic et al

Finally, as the model generates through the optimization process to sets of word vectors (*W* and ), the *GloVe* algorithm summarizes the two matrices in order to obtain the final semantic vector space models.

Several other aspects of interest can be found in the paper for the construction of our hypothesis and one of the most important is the correlation between word vector dimension (and thus model capacity) and the task-related accuracy (see Figure 1).

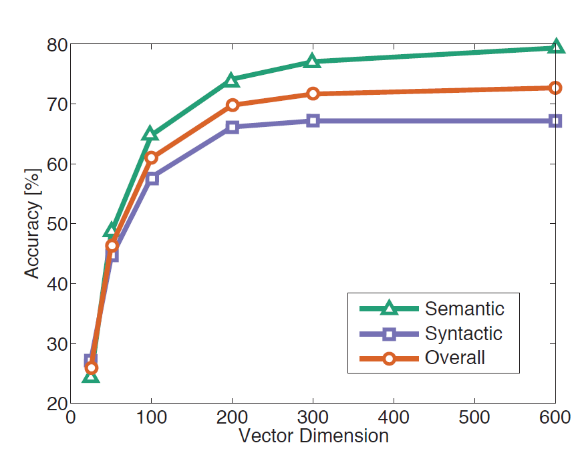
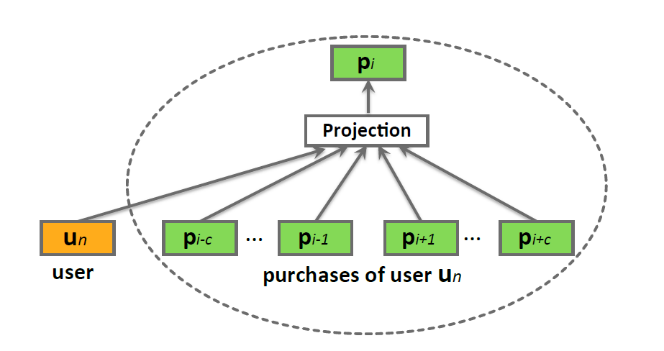
As the intuition suggests increasing model capacity (bigger vector space dimension) allows capturing richer semantic knowledge that in turns leads to more useful higher-level features and better accuracy down the stream. This is another of the main take-aways for our proposed pipeline.

Figure 2 - Accuracy on a specific task as a function of vector size. Image from original paper by Pennington et al

## Product recommendations

The logical next step in our related work review would be to analyze one of the early research papers that proposed the encoding of item SKUs in a similar manner to that of words. Grbovic et al proposed in their paper “*E-commerce in your inbox: Product recommendations at scale*” (Grbovic, et al., 2015) a direct approach of constructing product embeddings based on word2vec skip-gram model.

The authors propose a direct analogy between a certain product (item) in a basket and the focal word in the skip-gram context as presented in Figure 1.

|  |
| --- |
| 250  251  252  253  254  255  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285  286  287  288  289  290  291  292  293  294  295  296  297  298  299 |

|  |
| --- |
| 200  201  202  203  204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219  220  221  222  223  224  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240  241  242  243  244  245  246  247  248  249 |

Figure 3 - Bag-of-words like approach to enrich the vector space model with user meta information. Picture based on the original paper by Grbovic et al

Another proposed aspect was that of enriching the vector space model with meta information such as user-identification in a similar method with *doc2vec* approach – a follow-up of the original word2vec – by Le and Mikolov (Le & Mikolov, 2014). In Figure 3 is presented this particular approach, this time being based on the BOW algorithm (predicting a focal item based on the context) instead of the skip-gram (predicting the context starting from a focal item).

In terms of actual vector space model applications, the authors propose two different hypotheses:

* straight correlation between items cosine distances and the actual real-life similarities of the products (namely the **prod2vec-topK** approach)
* a second more elaborate heuristic approach of finding viable “complementarity” baskets of products that might be sold well together. This second proposed predictive model, called **prod2vec-cluster** by the authors, leverages the hypothesis that similar items can be grouped together using distance metrics with K-Means algorithm. Furthermore, close clusters generated in this manner yield potentially well correlated products and picking items from different close clusters will generate a well-defined complementarity basket.

As a final note, this particular work is one of the earliest and most well known in the area of applying representation learning and neural language modelling to the problem of retail recommender system, however for our proposed pipeline item embeddings we plan to employ *GloVe* approach to vector space generation as well as use the discounted positive pointwise mutual information taken from the matrix of item co-occurrence.

## Retrofitting

The next logical step in our related work review is to revisit research work in the area of enriching vector space models based on retrofitting. For this subject we decided to analyze two important papers: “*Retrofitting Word Vectors to Semantic Lexicons*” by Faruqui et al (Faruqui, et al., 2014) and the more recent work of Benjamin J. Lengerich, Andrew L. Maas and Christopher Potts “*Retrofitting Distributional Embeddings to Knowledge Graphs with Functional Relations*” (Lengerich, Maas, & Potts, 2017).

In order to better understand the motivation for the review of vector space models retrofitting approaches (and in a later section the review of the counter-fitting approaches) we have to revisit the work done by *Grbovic at al* as well as other teams that pursued the goal of obtaining products semantic vector space models. In all above approaches the authors base their work on the shallow hypothesis that similar items (*words*) measured by cosine distance might have a synonym-like or an entailing-like relationship. However, this as possible for product vector space models as it is with natural language vector space models counterparts. For example, a *GloVe-300* vector space model, pretrained on *Wikipedia* and *Gigaword* might tell us that *'adolescent'*, *'teenager'*, *'puberty'* are very similar to each other based on their pairwise cosine distance and actually ‘*adolescent’* is more similar to ‘*puberty’* than it is to ‘*teenager’*. Exactly the same stands with product semantic vector space models – just using a raw embedding generated with *GloVe* or *word2vect* is not enough to get rich and reliable properties of the items.

This is the point where adding meta-information similar to the *synonymity graphs* in natural language might greatly aid us in further refining the semantic power of our product vector space model.

General Instructions

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, authors’ names and complete addresses, which must be centered at the top of the first page, and any full-width figures or tables (see the guidelines in Subsection 3.6). Lines should be justified, with even spacing between margins (Ctrl+J). Single-spaced lines are permitted, but **authors are encouraged to use Paragraph spacing at Multiple, 1.05pt**, with **Font character spacing condensed with kerning of 0.1pt**,and **Margins at 0.98in**, for consistency with A4 paper and documents formatted with LaTeX. Go to Format, Document, Page Setup, and make sure A4 is selected. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section 5.

Pages have a ruler, header and footer for initial submission, with header at 0.3 in from top and footer at 0.4 in from bottom. When you first create your submission on softconf, please fill in your submitted paper ID where \*\*\* appears in the \*\*\* at the header at the top. If the paper is accepted, remove the header, footer (page numbers), and the ruler for the final version (camera-ready). **Do not number the pages in the camera-ready version**.

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 2019 MS Word style contains a 2.5 in table beneath the title for this purpose.

The Ruler

The ACL 2019 style defines a printed ruler that should be present 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.

In this Word template, you can add the ruler to each individual page by copying it from another page. The Ruler consists of separate left-hand rulers and right-hand rulers. Each ruler is a text box, sized 10 in x 0.45 in, flushed with the sides of the page, with a table containing numbers. The table has no borders, and the edges are pushed to the edges of the textbox. The left part of the ruler is justified left, left indent 0.1 in/right indent 0 in, and saved in the “ACL Ruler Left” style; and the right-hand ruler, justified right had left indent 0 in/right indent 0.1 in, saved in the “ACL Ruler Right” style. Number text is in Arial, and spacing between each line is Multiple, 1.05 in, with 4.5 pt below each line. After pasting a new ruler onto a page, highlight the ruler text and press **F9** to update the numbers. **Align** the text box **Middle**, and **Center**, and then **Arrange** Behind Text.

**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.*, this paragraph ends at mark 141.5).

Electronically-available Resources

ACL provides this description in LATEX2e (acl2019.tex) and PDF format (acl2019.pdf), along with the LATEX2e style file used to format it (acl2019.sty) and an ACL bibliography style (acl2019.bst) and example bibliography (acl2019.bib). These files are all available at [http://naacl2019.org/downloads/ acl2019-latex.zip](http://naacl2019.org/downloads/%20naaclhlt2019-latex.zip). A Microsoft Word template file (acl2019-word.docx) and example submission pdf (acl2019-word.pdf) is available online. We strongly recommend the use of these style files, which have been appropriately tailored for the ACL 2019 proceedings.

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; on Windows, you can use Adobe's Distiller, an online service such as <http://go4convert.com/>, or if you have cygwin installed, you can use dvipdf or ps2pdf.

Please make sure that your PDF file includes all the necessary formatting, hyperlinks, and 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. 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.

Layout

Format manuscripts two columns 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

|  |  |  |
| --- | --- | --- |
| **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 | 10 pt |  |
| abstract text | 10 pt |  |
| bibliography | 10 pt |  |
| footnotes | 9 pt |  |

Table 1: Font guide.

* Top margin: 2.5 cm
* Bottom margin: 2.5 cm
* Column width: 7.7 cm
* Column height: 24.7 cm
* Gap between columns: 0.6 cm

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

|  |
| --- |
| 300  301  302  303  304  305  306  307  308  309  310  311  312  313  314  315  316  317  318  319  320  321  322  323  324  325  326  327  328  329  330  331  332  333  334  335  336  337  338  339  340  341  342  343  344  345  346  347  348  349 |

|  |
| --- |
| 350  351  352  353  354  355  356  357  358  359  360  361  362  363  364  365  366  367  368  369  370  371  372  373  374  375  376  377  378  379  380  381  382  383  384  385  386  387  388  389  390  391  392  393  394  395  396  397  398  399 |

Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. If Times Roman is not available, you may use the Times New Roman font, which is often provided by default and only slightly different.

The First Page

Center the title, author name(s) and affiliation(s) across both columns (or, in the case of initial sub- mission, 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 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 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.*, use “Mitchell,” not “MITCHELL”). Do not use 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 electronic 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 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 0.6cm 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. Use 11 points for text. Indent when starting a new paragraph, about 0.16 in. This is accomplished with Right click, Styles, “ACL First Line”.

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. In bold, use 11 points for subsection headings, 12 points for section headings. Do not number subsubsections.

**Citations:** Citations within the text appear in parentheses as ([Gusfield, 1997](#Gusfield1997)) or, if the author's name appears in the text itself, as Gusfield ([1997](#Gusfield1997)). Collapse multiple citations as in ([Gusfield, 1997](#Gusfield1997); [Aho and Ullman, 1972](#AhoUllman72)). Append lowercase letters to the year in cases of ambiguities. Treat double authors as in ([Aho and Ullman, 1972](#AhoUllman72)), but write as in ([Chandra et al., 1981](#ChandraEtAl1981)) when more than two authors are involved. Also refrain from using full citations as sentence constituents.

We suggest that instead of

“([Gusfield, 1997](#Gusfield1997)) showed that ...”

you use

“Gusfield ([1997](#Gusfield1997)) showed that ...”

**Hyperlinks:** Within-document and external hyperlinks are indicated with Dark Blue text, Color Hex #000099.In order to create hyperlinks between citations and references, as you insert each full reference in the References section, highlight it and then select Insert, Bookmark. Link back to the reference from its citations in the text by highlight the citation, right clicking, and selecting Insert, Cross-Reference, then selecting the Bookmark you’ve saved. Highlight the citation again to give make it dark blue (included in this theme), if it is not automatically applied. If there are problems saving the hyperlinks when you convert the document to PDF, use an online converter such as <http://go4convert.com>.

|  |
| --- |
| 450  451  452  453  454  455  456  457  458  459  460  461  462  463  464  465  466  467  468  469  470  471  472  473  474  475  476  477  478  479  480  481  482  483  484  485  486  487  488  489  490  491  492  493  494  495  496  497  498  499 |

|  |
| --- |
| 400  401  402  403  404  405  406  407  408  409  410  411  412  413  414  415  416  417  418  419  420  421  422  423  424  425  426  427  428  429  430  431  432  433  434  435  436  437  438  439  440  441  442  443  444  445  446  447  448  449 |

**Digital Object Identifiers:** As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. As of 2017, we are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus, please ensure that you use References that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at <http://aclanthology.info/>.

As examples, we cite ([Goodman et al., 2016](#GoodmanEtAl2016)) to show you how papers with a DOI will appear in the bibliography. We cite ([Harper, 2014](#Harper2014)) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the references.

As reviewing will be double-blind, the submit- ted 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](#Gusfield1997)) ...”

should be avoided. Instead, use citations such as

“Gusfield ([1997](#Gusfield1997)) previously showed ... ”

**Please do not use anonymous citations** and donot include acknowledgements when submitting your papers. Papers that do not conform to theserequirements may be rejected without review.

**References:** Gather the full set of references together under the heading **References**; place the section before any Appendices. 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](#APA83)). 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 Association for Computing Machinery ([1983](#ACM83)) *Computing Reviews*.

The bibliography style described here roughly fits the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Submissions should accurately reference prior and related work, including code and data. If a piece of prior work appeared in multiple venues, the version that appeared in a refereed, archival venue should be referenced. If multiple versions of a piece of prior work exist, the one used by the authors should be referenced.

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

Footnotes

**Footnotes:** Put footnotes at the bottom of the page and use 9 point font. They may be numbered or referred to by asterisks or other symbols.[[1]](#footnote-1) Footnotes should be separated from the text by a line.[[2]](#footnote-2)

Figures and Tables

**Creating:** To create a new Figure or Table, insert a Text Box where you want it to appear (generally, centered at the top of a column close to where it is referred to) and then fill it in with the Figure (or Table). Highlight and right click to add Caption, with the ACL Caption style, which places 10 pt below and above the caption.

**Placement:** Place figures and tables in the paper near where they are first discussed, as close as possible to the top of their respective column. Wide figures and tables may run across both columns and should be placed at the top of a page.

|  |
| --- |
|  |
| Figure 4: Figure caption**.** |

In MSWord, authors can place a Figure (*e.g.*, a graphic and its caption) inside the rows of a 2 x 1 table (2 rows and 1 column) with invisible borders. Specify table positioning by right-clicking its handle in the upper left corner. Place the image in the center of the first row, and the caption in the center of the second row.

**Captions:** Provide a caption for every table and figure; 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 10 point text. Captions should be placed below illustrations. Captions that are one line are centered. Captions longer than one line are left-aligned.

**Numbering:** To update numbering, highlight all the relevant text (*e.g.*, **Ctrl-A + F9**). This will update all the numbering applicable to tables, figures, equations, and headings.

**Cross-referencing:** To add a cross reference to a figure or table:

* Place the mouse pointer at the location where you wish to add the cross-reference.
* Click on the **Insert** menu, (then click **Reference**), and then **Cross-reference** in the **Links** panel.
* In the **Cross-reference** dialog box, click the caption to which you are building the text reference.
* For a figure, under **Reference Type**, click **Figure**.
* Under Insert Reference To, click Only Label and Number, then click OK.
* Once the reference is in place, apply the ‘Normal’ font style (size 11, no bold face).

|  |
| --- |
| 550  551  552  553  554  555  556  557  558  559  560  561  562  563  564  565  566  567  568  569  570  571  572  573  574  575  576  577  578  579  580  581  582  583  584  585  586  587  588  589  590  591  592  593  594  595  596  597  598  599 |

|  |
| --- |
| 500  501  502  503  504  505  506  507  508  509  510  511  512  513  514  515  516  517  518  519  520  521  522  523  524  525  526  527  528  529  530  531  532  533  534  535  536  537  538  539  540  541  542  543  544  545  546  547  548  549 |

* This is an example reference to Figure 1.

Equations

An example equation is shown below:

(1)

To add new equations, authors are encouraged to copy this existing equation line, and then replace with the new equation. The numbering and alignment of equation line elements is automatic. To update equation numbering, press **Ctrl-A + F9**. Note: this will only update the number to the right of the equation; to update numbering within the text you must create a cross-reference.

**Cross-referencing:** To create a cross-reference for an equation:

* Create a bookmark for it.
* Select the number to the right of the equation. Go to **Insert**, **Bookmark** (in the **Links** panel),andthen create a name for your equation. Press **Add** to create the bookmark.
* To refer back, place the mouse pointer at the location where you wish to add the cross reference.
* Go to **Insert, Cross-reference** (in the **Links** panel).In the dialogue box, select **Bookmark** and **Bookmark Text** from each dropdown list. Uncheck **Insert as Hyperlink**, then click **OK**.
* This will make it such that whenever a new equation is added, the references to the equation will be updated when **Ctrl-A + F9** is pressed.
* This an example cross-reference to Equation (1).

Accessibility

In an effort to accommodate people who are color-blind (as well as those printing to paper), grayscale readability for all accepted papers is encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. A simple criterion: All curves and points in your figures should be clearly distinguishable without color.

From Submitted to Camera-Ready

|  |
| --- |
| 650  651  652  653  654  655  656  657  658  659  660  661  662  663  664  665  666  667  668  669  670  671  672  673  674  675  676  677  678  679  680  681  682  683  684  685  686  687  688  689  690  691  692  693  694  695  696  697  698  699 |

|  |
| --- |
| 600  601  602  603  604  605  606  607  608  609  610  611  612  613  614  615  616  617  618  619  620  621  622  623  624  625  626  627  628  629  630  631  632  633  634  635  636  637  638  639  640  641  642  643  644  645  646  647  648  649 |

To remove submission formatting for the camera-ready document, delete the header and footer on the first page (double clicking in the area, then delete the text). This will remove them from the rest of the document. To remove the ruler, highlight the lists of numbers and delete.

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”.

Length of Submission

The ACL 2019 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 nine (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 figures and tables that are part of the main text 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.

ACL 2019 does encourage the submission of additional material that is relevant to the reviewers but not an integral part of the paper. There are two such types of material: appendices, which can be read, and non-readable supplementary materials, often data or code. Do not include this additional material in the same document as your main paper. Additional material must be submitted as one or more separate files, and must adhere to the same anonymity guidelines as the main paper. The paper must be self-contained: it is optional for reviewers to look at the supplementary material. Papers should not refer, for further detail, to documents, code or data resources that are not available to the reviewers. Refer to Appendix A and Appendix B and for further information.

Workshop chairs may have different rules for allowed length and whether supplemental material is welcome. As always, the corresponding call for papers is the authoritative source.

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 2019 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 NAACL 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.

STREAM Tools

This Microsoft Word file was updated in 2016 with STREAM Tools, designed for creating well-formatted reports and papers with Microsoft Word (Mamishev, 2010; Mamishev, 2013).

Acknowledgments

The acknowledgements should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for 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.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](http://dl.acm.org/citation.cfm?doid=322234.322243). *Journal of the Association for Computing Machinery*, 28(1):114-133. https://doi.org/10.1145/322234.32224.

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

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. [Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing](http://aclweb.org/anthology/P16-1001). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. <https://doi.org/10.18653/v1/P16-1001>.

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

Mary Harper. 2014. [Learning from 26 languages: Pro- gram management and science in the babel program](http://aclweb.org/anthology/C14-1001). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. <http://aclweb.org/anthology/C14-1001>.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

1. Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here.

1. Supplementary Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.

1. This is how a footnote should appear.

   Note the line separating the footnotes from the text. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)