Extending Levenshtein edit distance with phonological features for cognate identification task

|  |  |
| --- | --- |
| 000  00  00  00  00  00  00  00  00  00  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |

|  |
| --- |
|  |
| Anonymous EMNLP submission |
|  |
|  |
|  |
|  |
|  |

Abstract

The paper presents an automated task-based evaluation framework for an extension to the Levenshtein edit distance metric. This extension explicitly represents linguistic phonological features of compared characters so the metric can use information about characters’ internal structure rather than treat them as elementary atomic units of comparison. The proposed framework allows us to test alternative configurations of phonological features, and automatically select feature arrangements, which on a range of evaluation parameters systematically show greater improvements over the traditional character-based Levenshtein edit distance. Resources for computing Levenshtein edit distance with phonological features for a number of alphabets are made available as open software.

Introduction

This paper present a methodology for the development and automated evaluation of a linguistic features set that extends the traditional Levenshtein edit distance metric used for the task of cognate identification.

Cognate identification is important for a range of different applications; in our experimental settings we use it for assisting MT developers in creating cognate lexicon for hybrid MT between closely related languages, some of these languages are under-resourced (e.g., Ukrainian and Russian, Portuguese and Spanish, Dutch and German). For many of such language pairs there are no electronic dictionaries, and there are only small parallel corpora available with limited lexical coverage. Typically these parallel corpora can provide translations for frequently used general words, but miss the ‘long tail’ of less frequent, often topic-specific or terminological words. However, in closely related languages these words are often cognates, which creates a possibility to rapidly extend bilingual lexicons in semi-automated way using non-parallel, comparable corpora and automated cognate identification techniques.

In this task, cognate candidates are generated from word lists created from large monolingual comparable corpora in both languages The assumption is that the developers have good linguistic intuition of both languages and work through lists of cognate candidates, checking which pairs can be added to the bilingual dictionary. Their productivity depends on whether cognates are presented high up in the list of candidates, ideally at the top, or at least in the top N items, where N lines should at most fit on a single screen.

Levenshtein edit distance (Levenshtein, 1966) is typically used to compare word pairs from different languages and determine if they are cognate candidates. It is computed for every pair of words in the two word lists (their Cartesian product), the search space may be restricted to matching part-of-speech codes, if this annotation is available for the corpus. However, there are several problems with the traditional Levenshtein metric, one of which is that all characters are treated equally. As a result, words that are intuitively close may receive a large distance score, e.g.,

Ukrainian (Uk) “*жовтий*” (zhovtyi)=‘yellow’

Russian (Ru) “*жёлтый*” (zheltyi) = ‘yellow’

(Lev distance = 3),

where, for historical reasons, similar sounds are represented by different characters: [o] by Uk ‘о’ and Ru ‘ё’, [y] by Uk ‘и’ and Ru ‘ы’. On the other hand, words that are not cognates and are intuitively far apart, still receive the same distance scores, such as:

Uk “*жовтий*” (zhovtyi) = ‘yellow’ and

Ru “*жуткий*” (zhutkiy) = ‘dismal’ (Lev = 3).

Some existing modifications and extensions of the Levenshtein metric introduce weightings for different character mapping, but these weights need to be set or empirically determined for each specific mapping: compared characters do not have internal structure, so there is no way to predict the weights in advance for any possible pair in a principled way.

In this paper we present an automated task-based evaluation framework for an extension of the Levenshtein edit distance metric, which explicitly represents linguistic phonological features of compared characters, so the metric can use information about characters’ internal structure rather than treat them as elementary atomic units of comparison. These distinctive features have been proposed for modeling dialectological variation and historical changes in languages (Nerbonne & Heeringa, 1997), and may be also applied to a range of computational linguistic tasks, such as transliteration, morphological alternations and cognate identification. However, there are multiple ways of identifying, representing and structurally arranging these features, so there is a need for a methodology for evaluating alternative feature configurations. In our previously published pilot experiment, a smaller-scale manual evaluation indicated the need to use hierarchical phonological feature structures for consonants rather than flat feature vectors used in dialectology (Anonymous, 2016). An automated evaluation methodology proposed in this paper allows us to design and calibrate feature structures in a systematic way.

In this paper we present a framework for evaluating different arrangements of phonological features using the task of automated cognate identification. Apart from practical applications mentioned above, this task also has standardized evaluation metrics, e.g., cognate in top-N candidates, which can be applied automatically and used in the feature-engineering task for selection of optimal feature arrangements. Our evaluation results show greater improvement for hierarchical feature structures, on a number of automatically computed parameters, compared to our previous pilot experiment, and also rule out unproductive modifications of the metric.

Phonological distinctive features and their application for cognate identification

The theory of phonological distinctive features, first proposed by Roman Jacobson (Anderson, 1995: 116), associates each phoneme (an elementary segmental unit of speech which can distinguish meanings) with a unique set of values for categories, which may apply to larger classes of sounds. For example, phoneme [t] has the following values for the categories:

*‘type’: consonant*

*‘voice’: unvoiced*

*‘maner of articulation’: plosive;*

*‘active articulation organ’: front of the tongue*

*‘passive articulation organ’: alveolar*

Phoneme [d] has the same articulation, but is pronounced with the use of vocal cords, so it differs only in the value of one distinctive feature,

*‘voice’: voiced*

with all other categories and values remaining the same.

In historical development of languages and in morphological variation within a language the sound changes more often apply only to values of certain distinctive features within characters, but not to the whole category-value system, e.g.: Ge “Tag” = Nl “dag” (‘day); Ge: “machen” = Nl “maken” (‘make’). Therefore, in languages where the writing system is at least partially motivated by pronunciation, it would be useful to represent the phonological distinctive features, in order to differentiate between varying degrees of closeness for different classes of characters, e.g., vowels, sonants and consonants, or sounds with identical or similar articulation. Greater closeness between characters in terms of their phonological features has important linguistic and technical applications, such as modeling dialectal variation, historical change, morphological and derivational changes in words, such as stem alternations in inflected forms.

Phonological distinctive features have been integrated into the Levenshtein distance metric in the following way (c.f. Anonymous, 2016: 123):

(1) Substitution cost for two characters is calculated as 1 [minus] F-measure between Precision and Recall of their feature overlap; this allows calculating the cost for character with different numbers of features, and the metric remains symmetric. Intuitively, to substitute [t] with [d] in “Tag” → “dag” we need to re-write only one feature out of 5, so the cost is 0.2 rather than 1.

(2) The order of matching the distinctive features was found to be important. Sections 3 and 4 describe an experiment on comparing two different arrangements of features: as flat feature vectors and as feature hierarchies, where matching lower level features is a pre-condition for attempting to match lower level features. Hierarchical organization achieved better performance compared to traditional flat feature vectors. Intuitively this means that not all feature categories should be treated equally, some of them have higher priority and license comparison of lower level features

(3) Insertion and deletion costs have been calibrated for the range between 0.2 and 1 using the proposed evaluation framework, described in this paper in Section 3. Optimal performance on cognate identification was achieved for cost of insertion = deletion = 1.

For the task of cognate identification, the introduction of these features distinguishes different types of character substitutions and gives more accurate prediction of the degree of closeness between compared characters and words, e.g., for the word pairs discussed above, where Levenshtein distance =3 in both:

Ukrainian (Uk) “жовтий” (zhovtyi) = ‘yellow’

ж 'type:consonant', 'voice:ff-voiced',

'maner:ff-fricative', 'active:ff-fronttongue',

'passive:ff-palatal'

о 'type:vowel', 'backness:back',

'height:mid', 'roundedness:rounded',

'palate:nonpalatalizing'

в 'type:consonant', 'voice:fl-voiced',

'maner:fl-fricative', 'active:fl-labial',

'passive:fl-bilabial'

т 'type:consonant', 'voice:pf-unvoiced',

'maner:pf-plosive', 'active:pf-fronttongue',

'passive:pf-alveolar']

и 'type:vowel', 'backness:front',

'height:closemid', 'roundedness:unrounded',

'palate:nonpalatalizing']

й 'type:consonant', 'voice:am-sonorant',

'maner:am-approximant','active:am-midtongue',

'passive:am-palatal'

Russian (Ru) “жёлтый” (zheltyi) = ‘yellow’ (non-matching characters)

ё // highlighted matches with Uk “о”

**'type:vowel', 'backness:back',**

**'height:mid',** **'roundedness:rounded'**,

'palate:palatalizing'

л // highlighted matches with Uk “в”

**'type:consonant',** 'voice:lf-sonorant', '

maner:lf-lateral', 'active:lf-fronttongue',

'passive:lf-alveolar'

ы // highlighted matches with

**'type:vowel',** 'backness:central',

**'height:closemid', 'roundedness:unrounded',**

**'palate:nonpalatalizing'**

Calculation of the GLev metric for (Ru) “жёлтый” (zheltyi) = ‘yellow’:

0.0 1.0 2.0 3.0 4.0 5.0 6.0

1.0 **0.0** 1.0 2.0 3.0 4.0 5.0

2.0 1.0 **0.2** 1.2 2.2 3.2 4.2

3.0 2.0 1.2 **1.0** 2.0 3.0 4.0

4.0 3.0 2.2 2.0 **1.0** 2.0 3.0

5.0 4.0 3.2 3.0 2.0 **1.2** 2.2

6.0 5.0 4.2 4.0 3.0 2.2 **1.2**

cf.: Metric calculated with Ru “жуткий” (zhutkiy) ‘dismal’, while Lev=2 for both, GLev = 2.0 > 1.2

0.0 1.0 2.0 3.0 4.0 5.0 6.0

1.0 **0.0** 1.0 2.0 3.0 4.0 5.0

2.0 1.0 **0.2** 1.2 2.2 3.2 4.2

3.0 2.0 1.2 **1.0** 1.2 2.2 3.2

4.0 3.0 2.2 2.0 **1.8** 2.2 3.0

5.0 4.0 3.2 3.0 2.8 **2.0** 3.0

6.0 5.0 4.2 4.0 3.8 3.0 **2.0**

Set-up of the experiment and results

The proposed methodology is an automated performance-based evaluation for testing different settings of phonological categories and values for the extended Levenshtein metric. The experiment is set up in the following way:

(1) A small Ukrainian – Russian and Russian-Ukrainian dictionary was used to compile a gold-standard translation glossary of 11k Ukrainian words, each having one or more Russian translation equivalents.

(2) Translation equivalents in the glossary have been crosschecked with monolingual Russian and Ukrainian word lists compiled from PoS-tagged corpora for these languages. Only those entries in the gold standard were retained which exist in both words lists, so the cognate identification programme can in principle find them.

(3) An additional requirement has been introduced that in both word lists the cognates should be tagged with the same part-of-speech. This reduced the search space for cognates and reduced computing time needed to calculate distances for the entries in the Russian word list. Frequency information was not used, since the monolingual word lists from both languages came from diverse corpora of different sizes. Still calculation time is a limiting factor in evaluation, so 809 entries from the gold standard were randomly chosen for computting different types of Levenshtein edit distance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Experiment** | **Med topN** | **Top 1** | **Top 5** | **Top 10** | **Top 25** |
| BaseL Lev | 50 | 206 | 328 | **360** | 382 |
| Phon Lev H | **47.5** | **240** | **334** | 359 | **385** |
| **Improv BaseL** | +5% | +16.5% | +2% | 0% | +1% |
|  |  |  |  |  |  |
| Phon Lev V | 87.5 | **215** | 289 | 319 | 349 |
| *DiffBase L* | -75% | +4.4% | -10% | -11% | -9% |
|  |  |  |  |  |  |
| PhonLevi=0.2 | 125.5 | **216** | **291** | 315 | 342 |
| PhonLevi=0.4 | 54.5 | **230** | 307 | 334 | 367 |
| PhonLevi=0.6 | **48** | **235** | 328 | 354 | **385** |
| PhonLevi=0.8 | **40** | **240** | **337** | 359 | **391** |
| **Improv BaseL** | +20% | +16.5% | +3% | 0% | 2% |
| Improv PhLevH | +15% | 0% | +1% | 0% | 2% |

Table : Automated evaluation of metric settings.

(4) For each Ukrainian entry in the gold standard, the top-N list has been calculated from the Russian word list for each of available translation equivalents.

(5) Importantly, as the evolution methodology is automatic, all translation equivalents available in the gold standard are treated equally: in this stage no distinction is made between cognates and non-cognate equivalents. This removes the need for the manual filtering of the gold standard and also naturally covers ‘near-cognates’ or words with cognate morphemes where only parts of words match. Since the baseline and the modified Levenshtein metric are evaluated on the same gold standard, performance figures are relative and show the difference in finding translation equivalents for any degree of ‘cognateness’.

(6) Different methods are compared by the following parameters: Median top-N number for the metric; In top-1, top-5, top-10 and top-25.

(7) The following settings were compared:

(a) Baseline Levenshtein edit distance;

(b) Levenshtein distance extended with phonological features with flat feature vectors

(c) Levenshtein distance extended with hierarchical phonological features (where manner and active place of articulation are treated as top-level features which need to be matched in order for other features to match

(d) Variants of the (b) and (c) metric with different insertion / deletion values – between 0.2 and 0.8.

The results of the evaluation experiment are presented in Table 1, where:

BaseL Lev = baseline Levenshtein metric

Phon Lev H = Phonological extension to Levenshtein metric with feature hierarchy

Phon Lev V = Phonological extension to Levenshtein metric with flat feature vectors

PhonLevi=0.X = Phonological extension to Levenshtein metric with modified insertion / deletion cost: i0.2 = the cost of insertion deletion is set to 0.2, i0.8 = is set to 0.8 (it is set to 1 in the Phon Lev metrics.

Discussion of the results, conclusion

It can be seen from Table 1 that:

(1) Hierarchical phonological Levenshtein metric outperforms the baseline on the Top 1 and Top 2 measures, the median rank improvements is +5%

(2) Flat phonological feature vector metric on all measures performs worse than the baseline. This can be interpreted as the need to take into account the order of matching higher-level features. Match of low-level features is not meaningful if higher-level features are not matched.

(3) The metric with insertion / deletion cost set to 0.8 outperforms both the baseline and the Hierarchical phonological Levenshtein metric, especially on the Median Top N, Top 1 and Top5 measures. This can be interpreted as the need to scale down insertion cost moderately, since the average substitution cost is down.

The results show that phonological extension to the Levenshtein edit distance metric on the task of cognate identification outperforms the character-based baseline. The proposed framework also allows accurate calibration of the feature arrangement and other parameter settings to the metric. Future work will include systematic evaluation of different possible feature hierarchies and costs, and metrics application to other tasks, such as transliteration.

Jji

sjijijs

**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 as they will fit in the final (accepted) version. The EMNLP 2017 MS Word style contains a 2.5 in table beneath the title for this purpose.

The Ruler

The EMNLP 2017 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. Each Ruler is a textbox, sized 10 in x 8.3 in and aligned middle and center, with a table inside of it. The table has two columns: The left-hand ruler and the right hand ruler. The table has no borders, and its 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 “EMNLP Ruler Left” style; and the right-hand ruler, justified right had left indent 0 in/right indent 0.1 in, saved in the “EMNLP 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 starts at mark 138.5).

Electronically-available Resources

EMNLP provides this description in LATEX2e (emnlp2017.tex) and PDF format (emnlp2017.pdf), along with the LATEX2e style file used to format it (emnlp2017.sty) and an EMNLP bibliography-style (emnlp2017.bst) and

and example bibliography (emnlp2017.bib). These files are all available at [http://emnlp2017.org/downloads/emnlp17-latex.zip](http://emnlp2017.org/downloads/acl17-latex.zip). A Microsoft Word template file (emnlp17-word.docx) and example submission pdf (emnlp17-word.pdf) is available at [http://emnlp2017.org/downloads/emnlp17-word.zip](http://emnlp2017.org/downloads/acl17-word.zip). We strongly recommend the use of these style files, which have been appropriately tailored for the EMNLP 2017 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.

|  |  |
| --- | --- |
|  |  |

* Left and right margins: 2.5 cm
* Top margin: 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 | 11 pt |  |
| abstract text | 11 pt |  |
| bibliography | 10 pt |  |
| footnotes | 9 pt |  |

Table : Font guide.

* 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.

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 0.6 cm 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 11 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, “EMNLP First Line”. 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.

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 12 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 ...”

**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.

**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**. However, in a submission for review the appendices should be filed as a separate PDF.

**Acknowledgments:** A section for acknowledgments to funding agencies, colleagues, collaborators, etc. should go as a last (unnumbered) section immediately before the references. Keep in mind that, during review, anonymization guidelines apply to the contents of this section too.

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-2) Footnotes should be separated from the text by a line.[[2]](#footnote-3)

Graphics

**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 EMNLP 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 1: 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 11 point text.

**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).
* 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 the 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

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.

Length of Submission

The EMNLP 2017 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. Supplementary material in the form of appendices does not count towards the page limit.

However, note that supplementary material should be supplementary (rather than central) to the paper, and that reviewers may ignore supplementary material when reviewing the paper. Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

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.

Supplemental Materials

EMNLP 2017 encourages submitting software and data that is described in the paper as supplementary material. EMNLP 2017 also encourages reporting preprocessing decisions, model parameters, and other details necessary for the exact replication of the experiments described in the paper. Papers may be accompanied by supplementary material, consisting of software, data, pseudo-code, detailed proofs or derivations that do not fit into the paper, lists of features or feature templates, parameter specifications, and sample inputs and outputs for a system. Appendices are considered to be supplementary materials, and should be submitted as such.

|  |  |
| --- | --- |
|  |  |

The paper should be self-contained and not rely on the supplementary material. Reviewers are not asked to review or even download the supplemental material. If the pseudo-code or derivations or model specifications are an important part of the contribution, or if they are important for the reviewers to assess the technical correctness of the work, they should be a part of the main paper, not as appendices.

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 authors' 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.

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 EMNLP 2017 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 EMNLP-related workshops that do not have published proceedings in the EMNLP 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 EMNLP 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 adapted from the ACL 2017 Word template and was updated 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. This is how a footnote should appear. [↑](#footnote-ref-2)
2. Note the line separating the footnotes from the text. [↑](#footnote-ref-3)