|  |
| --- |
| **几类不同的词典结构和查找算法对汉语分词系统性能的影响研究** |
|  |
| **Dingzirui Wang** |
| Harbin Institute of Technology |
| 1183710211@stu.hit.edu.cn |
|  |
|  |

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 |

中文词法分析任务是中文信息处理中的一项基本任务，主要包括两大部分，分别是中文分词和中文词性标注。而在中文分词任务中，对效率和正确性影响最大的两个因素就是词典的结构和分词算法的实现。本论文的目的是针对中文分词任务的特点，选取了一些具有代表性的数据结构和算法，尝试对这一问题进行解决，并比较了不同方法之间运行效率和正确性上的差异。

对于词典结构，在最基础的链表的基础上，考虑到单词按照词典序具有有序性这一性质，可以采用二分查找的方法来进行求解。对于查找问题，哈希算法也是一个十分简单而有效的算法。此外，在字符查找问题中，Trie树是一种非常经典的数据结构，本文实现了双数组Trie并进行分析。

然后对于查找算法，本文实现了最基础的前后向最大匹配算法，除此之外，本文还实现了全切分有向图最大匹配算法，并对此进行分析。

最后，本文对以上词典结构和查找算法的两两组合得到的结果进行分析，并研究了各个算法的不足，以及可能的优化和相关工作。

Introduction

|  |
| --- |
| 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 |

在基于中文的自然语言处理中, 由于不像英文中词与词之间有固定的自然分界符, 所以对中文进行分词通常是处理中文信息任务的基础;同时由于中文自身的复杂性, 中文分词也一直是中文信息处理任务的难题。中文分词时进行词性标注、命名实体识别、关键词提取和文本聚类等下游自然语言处理任务的基础，也是语义分析等深层次文本里接任务的基础，在中文搜索引擎、输入法、机器翻译和智能问答系统等应用中起到了支撑作用。中文分词作为中文自然语言处理领域的重要基础研究, 近些年来很多专家学者致力于该领域的研究, 研究方法主要分为三种:a) 基于规则的方法;b) 基于传统机器学习模型的方法;c) 基于深度神经网络模型的方法。

基于规则的方法利用构词原理结合标注的词性等信息, 构建基于句法—语义规则的分析系统, 配合语法信息字典, 并补充了大量消除歧义的信息。Sui Z([2002](#Sui2002))在语法规则的基础上增加了领域特征进行中文分词的处理。基于规则的方法的优点是具有针对性和暂时较高的准确率, 但由于句法构造的领域相关性, 适应性较差, 词典与歧义消解处理难维护。

随着SIGHAN国际中文分词评测Bakeoff的展开, 将中文分词任务视做序列标注问题来解决逐渐成为主流。基于传统机器学习模型的方法主要为基于字标注的机器学习模型方法, 即字在字串的标注问题, 该方法能平等地看待词典词和未登录词的识别。

为了比较几个传统算法性能上的差距，本文主要完成了以下几个工作：

* 实现了多个词典结构和搜索算法。
* 在1998 年人民日报局部语料上取得了优异的结果，一个大规模的中文分词语料库。
* 实验表明传统算法虽然能取得较高的正确率，但在泛化能力和执行效率上有着极大的限制。

Figure 1: 词频和出现次数的统计图，其中横坐标词频，纵坐标为出现次数的自然对数。

Task Definition

本文研究了中文分词任务。这一任务是对一个连续的字序列按照一定的规范重新组合成词序列的过程。特别的，本文将实现的算法应用到1998 年人民日报局部语料上，这是一个可用于中文分词和词性标注的大规模语料库。首先，本文根据已分词的文本来获取词典。在获得词典后，分词系统将会根据词典来对这一语料进行分词。原文中的每一行都会被替换为一系列词，每两个词之间都会用空格来分隔。

Dictionary Analysis

我们首先根据已分词的文本，去除掉词性和所有辅助标注的符号后，将每个单词看作单词表中的一个条目，并记录每个单词出现的频率。此外，出于后续测试中可扩展性的考虑，我们同时提取了一元和二元词袋。

这样得到的词典中，每一个条目为一个三元组，其中为组成词组的单词的数量；为词组出现的频率；为词表中的词组，为全体中文字符的闭包。

整个词典按照的字典序升序排布，便于词典组织结构的处理。

对词典的词频统计如图1所示，其中横坐标为词频，纵坐标为。可以看到绝大多数的词仅出现了一次，占到了总数的74.16%。之后出现次数和词频的对数大致呈负对数的关系，表明二者之间可能可以用一个函数关系来拟合。

最后，词典中出现频率最高的单词为”，”，出现的次数高达74921次。

Tokenize Model

下面介绍本实验中用到的算法和模型结构。

Figure2给出了该模型的一个overview。算法在读取完词典文件后，会根据特定结构将所有单词组织起来。然后，算法（分词器）会逐行对文本进行分词，每次根据特定的规则提取一个可能的词组，并分析该词组是否在词典中存在，若存在，则将其作为一个分词的结果提取出来。

这一节如下组织。首先会介绍程序中各种实现的词典的结构（§4.1），之后会介绍各种分词器的结构（§4.2）。

Dictionary Structure

我们来介绍词典组织结构，其用于在给定了一个词之后，如何在单词表中检索这个词是否存在。

首先我们采用了最原始的链表查询（§4.1.1）。然后基于待查询量为字符串的这一特点，尝试了二分搜索（§4.1.2）和哈希桶（§4.1.3）这两种算法。最后，我们尝试了DAT（§4.1.4），一种预处理时间较长但检索速度极快的结构。

在下文中，为了简化叙述，我们记N为单词表的大小，L为最大的词组长度。

* + 1. Linked List

由于思路简单，实现方便的特性，链表被广泛地应用在各个场景中。

对于该任务，我们每次将新读取的单词放到列表的尾部。对于每次查询，我们遍历整个列表，看待查询的词是否存在。

该结构的插入复杂度为O(1)，查询复杂度为O(N)。因此该结构的预处理性能极佳，但查询的性能非常差，对于大规模的语料库，甚至不能在可接受的时间范围内得到结果。

* + 1. Binary Search

由于单词按照字典序具有有序性这一性质，因此可以使用二分搜索来解决这一问题。

我们按照组成词组的单词数量来进行分类，对于单词表中的每一个词，我们将其放到其长度对应的数组的末尾。由于单词表已经有序，所以按先后顺序放入的词组也是有序的。

对于每次查询，我们按照词组长度由小到大的顺序，在数组中进行二分查找，直到查询到该词组，或发现该词组不存在。

由于单词表的性质，该结构的插入复杂度为O(1)，查询复杂度为O(LlogN)。

* + 1. Hash Bucket

求解一个词组的哈希值分为两步。首先，求解组成词组的每个单字的哈希值；之后，根据每个单子的哈希值，求出整个词组的哈希值。

对于每个单字，由于其采用GBK编码格式，所以我们可以通过其编码，来还原出它的编号，具体来说，有如下公式：

其中和分别表示第i位的低16位和高16位的值。

之后，我们采用BKDR哈希函数来求解词组的哈希值，具体来说，有如下公式：

其中seed为哈希种子，mod一般取一个大质数。在我们的程序中，seed为131，mod为0xFFFFF。

我们会给每一个求得的哈希值创建一个数组，称为桶。产生冲突时，会在每个桶的末尾添加该元素。

对于每次查询，我们首先计算词组的哈希值，然后在哈希值对应的桶中，看该词组是否存在。

该结构的插入复杂度为O(1)，对于查询操作，我们可以认为所有的词几乎均匀地分布在各个桶中，因此查询复杂度为O(N/mod)。

* + 1. Double-Array Trie

在单词检索问题中，Trie树是一个十分经典而有效的结构。而在目前Trie的实现中，Jun-ichi(1992)的双数组实现是较为有效的一种。

我们用上一节中提到的单字哈希来作为分支的依据，该结构的插入复杂度为O(L)，查询复杂度为O(L)。

Tokenize Algorithm

这一节我们介绍分词算法，其用于在给定单词表之后，每次接收一个句子，并按照一定的逻辑来对句子进行分词。

我们主要采用了两种方法。前后向最大匹配（§4.2.1），一种基于枚举的算法；全切分有向图搜索（§4.2.2），一种基于DP的算法。

|  |
| --- |
| 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. Forward/Backward Max Match

As always, the respective call for papers is the authoritative source.

* + 1. Omni-segmentation Algorithm

Experiments

Any preliminary non-archival versions of submitted papers should be listed in the submission form but not in the review version of the paper. Reviewers are generally aware that authors may present preliminary versions of their work in other venues, but will not be provided the list of previous presentations from the submission form.

* + 1. Execution Time

Once a paper has been accepted to the conference, the camera-ready version of the paper should include the author's names and affiliations, and is allowed to use self-references.

* + 1. Tokenize Performance

**MS Word-specific details**: 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).

|  |
| --- |
| 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 |

Related Work

Papers that have been or will be submitted to other meetings or publications must indicate this at submission time in the START submission form, and must be withdrawn from the other venues if accepted by ACL 2020. Authors of papers accepted for presentation at ACL 2020 must notify the program chairs by the camera-ready deadline as to whether the paper will be presented. We will not accept for publication or presentation the papers that overlap significantly in content or results with papers that will be (or have been) published elsewhere.

Authors submitting more than one paper to ACL 2020 must ensure that submissions do not overlap significantly (>25%) with each other in content or results.

Conclusion

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 Section 7.10). **Type single-spaced.** Start all pages directly under the top margin. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section 4. Pages should be numbered in the version submitted for review, but **pages should not be numbered in the camera-ready version**.

File Format

For the production of the electronic manuscript you must use Adobe's Portable Document Format (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. 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 as soon as possible.

|  |  |  |
| --- | --- | --- |
| **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.

**MS Word-specific details**: Lines should be justified, with even spacing between margins (Ctrl+J). 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 LaTeX-formatted documents. Go to Format, Document, Page Setup, and ensure A4 is selected.

* 1. 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
* Top margin: 2.5 cm

|  |
| --- |
| 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 |

* 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.

* 1. 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 or Computer Modern Roman.

Table 1 specifies what font sizes and styles must be used for each type of text in the manuscript.

* 1. Ruler

A printed ruler (line numbers in the left and right margins of the article) should be presented in the version submitted for review, so that reviewers may comment on particular lines in the paper without circumlocution. 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.

**Reviewers:** note that the ruler measurements may 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. In most cases one would expect that the approximate location will be adequate, although you can also use fractional references (e.g., this line ends at mark 291.5).

|  |
| --- |
| 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 |

**MS Word-specific details**: In this Word template, you can add the ruler to each 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.

* 1. Title and Authors

|  |
| --- |
| 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 |

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Place the title centered at the top of the first page, in a 15-point bold font. Long titles should be typed on two lines without a blank line intervening. Put the title 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 “Mitchell” not “MITCHELL”). Do not format title and section headings in all capitals 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.

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.

Start the body of the first page 7.5 cm from the top of the page. **Even in the anonymous version of the paper, you should maintain space for names and addresses so that they will fit in the final (accepted) version.**

* 1. Abstract

Use two-column format when you begin the 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 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 10 point font.

* 1. Text

|  |
| --- |
| 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 |

Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document.

Indent 0.4 cm when starting a new paragraph.

* 1. Sections

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

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)

Graphics

Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns. Color is allowed, but adhere to Section 8's guidelines on accessibility.

**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 10 point text. Captions should be placed below illustrations. Captions that are one line are centered (see Table 1). Captions longer than one line are left-aligned (see Figure 1).

**MS Word-specific details**:

**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 (or ACL Caption Long style for multi-line captions), which places 10 pt below and above the caption.

**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.

|  |
| --- |
| 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 |

* 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.

|  |
| --- |
|  |
| Figure 1: A figure with a caption that runs for more than one line**.** |

This is an example reference to Figure 1.

Hyperlinks

Within-document and external hyperlinks are indicated with Dark Blue text, Color Hex #000099.

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)). 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. Collapse multiple citations as in ([Gusfield, 1997](#Gusfield1997); [Aho and Ullman, 1972](#AhoUllman72)).

Refrain from using full citations as sentence constituents. Instead of

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

write

“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 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](#APA83)). Use full names for authors, not just initials.

|  |
| --- |
| 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 |

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. Authors should not rely on automated citation indices to provide accurate references for prior and related work.

The following text cites various types of articles so that the references section of the present document will include them.

* Example article in a journal: ([Ashok et al., 1981](#Ashok1981))
* Example article in proceedings, with location: ([James et al., 2016](#James2016))
* Example article in proceedings, without location: ([James et al., 2016](#James2016))
* Example arxiv paper: ([Mohammad et al., 2016](#Mohammad2015))

**MS Word-specific details:** 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>.

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.

|  |
| --- |
| 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 |

All camera-ready references are required to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. 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 bibliography.

Appendices:

Appendices, if any, directly follow the text and the

references (but only in the camera-ready; see Appendix A). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

* 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**.

|  |
| --- |
| 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 |

* 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.

1. Accessibility

In an effort to accommodate people who are color-blind (as well as those printing to paper), grayscale readability is strongly 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.

1. MS Word 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

Sui Z, Chen Y. 2002. *The research on the automatic term extraction in the domain of information science and technology*. Proc of the 5th East Asia Forum of the Terminology.

Jun-ichi Aoe, Katsushi Morimoto, and Takashi Sato. 1992. *An efficient implementation of trie structures.* Softw. Pract. Exper. 22, 9 (Sept. 1992), 695–721. DOI:https://doi.org/10.1002/spe.4380220902

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

|  |
| --- |
| 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 |

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.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. *Yara parser: A fast and accurate depen-dency parser*. *Computing Research Repository,* arXiv:1503.06733. Version 2

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.

|  |
| --- |
| 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 |

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)