Streaming Fact Extraction for Wikipedia Entities at Web-Scale

Morteza Shahriari Nia, Christan Grant, Yang Peng, Daisy Zhe Wang
Data Science Research Lab
University of Florida
Gainesville, Florida 32611

Milenko Petrovic
Institute for Human and Machine Cognition
15 SE Osceola Ave
Ocala, Florida 34471

Abstract

Wikipedia.org is the largest online resource for free information and is maintained by a small number of volunteer editors. The site contains 4.3 million english articles; these pages can easily be neglected, becoming out of date. Any news-worthy event may require an update of several pages. To address this issue of stale articles we create a system that reads in a stream diverse web documents and recommends facts to be added to specified Wikipedia pages. We developed a three-stage streaming system that creates models of Wikipedia pages, filters out irrelevant documents and extracts facts that are relevant to Wikipedia pages. The systems is evaluated over a 500M page web corpus and 139 Wikipedia pages. Our results show a promising framework for fast fact extraction from arbitrary web pages for Wikipedia.

Wikipedia.org (WP) is the largest and most popular general reference work on the Internet. The website is estimated to have nearly 365 million readers worldwide. An important part of keeping WP usable it to include new and current content. Presently, there is considerable time lag between the publication of an event and its citation in WP. The median time lag for a sample of about 60K web pages cited by WP articles in the *living_people* category is over a year and the distribution has a long and heavy tail (**?**). Such stale entries are the norm in any large reference work because the number of humans maintaining the reference is far fewer than the number of entities.

Reducing latency keeps WP relevant and helpful to its users. Given an entity page, such as *wiki/Boris_Berezovsky_(businessman)*[1], possible citations may come from a variety of sources. Notable news may be derived from newspapers, tweets, blogs and a variety of different sources include Twitter, Facebook, iBlogs, arxiv, etc. However, the actual citable information is a small percentage of the total documents that appear on the web. To help WP editors, a system is needed to parse through terabytes of documents and select facts that can be recommended to particular WP pages.

[1]http://en.wikipedia.org/wiki/Boris_Berezovsky_(businessman)

Previous approaches are able to find relevant documents given a list of WP entities as query nodes (**?**; **?**; **?**; **?**; **?**). Entities of three categories *person*, *organization* and *facility* are considered. This work involves processing large sets of information to determine which facts may contain references to a WP entity. This problem becomes increasingly more difficult when we look to extract relevant facts from each document. Each relevant document must now be parsed and processed to determine if a sentence or paragraph is worth being cited.

Discovering facts across the Internet that are relevant and citable to the WP entities is a non-trivial task. Here we produce an example sentence from a webpage:

"`Boris Berezovsky, who made his fortune in Russia in the 1990s, passed away March 2013.`"

After parsing the sentence, we must first note that there are two entities named *Boris Berezovsky* WP; one a businessman and the other a pianist. Any extraction needs to take this into account and employ a viable distinguishing policy (entity resolution). Then, we match the sentence to find a topic such as *DateOfDeath* valued at *March 2013*. Each of these operations is expensive so an efficient framework is necessary to execute these operations at web scale.

In this paper, we introduce an efficient fact extraction system or given WP entities from a time-ordered document stream. Fact extraction is defined as follows: match each sentence to the generic sentence structure of {*subject — verb — adverbial/complement*} (**?**). The first *subject* represents the entity (WP entity) and *verb* is the relation type (slot) we are interested in (e.g. Table 1). The third component, *adverbial/complement*, represents the value of the associated slot. In our example sentence, the entity of the sentence is *Boris Berezovsky* and the slot we extract is *DateOfDeath* with a slot value of *March 2013*. The resulting extraction containing an entity, slot name and slot value is a *fact*.

Our system contains three main components. First, we pre-process the data and build models representing the WP query entities. Next, we use the models to filter a large stream of documents so they only contain candidate citations. Lastly, we processes sentences from candidate extractions and return slot values. Overall, we contribute the following:

Table 1: The set of possible slot name for each entity type.

Person	Facility	Organization
Affiliate AssociateOf Contact_Meet_PlaceTime AwardsWon DateOfDeath Titles FounderOf EmployeeOf	Affiliate Contact_Meet_Entity	Affiliate TopMembers FoundedBy

- Introduce a method to build models of WP name variations;
- Built a system to filter a large amount of diverse documents using a natural language processing rule-based extraction system;
- Extract, infer and filter entity-slot-value triples of information to be added to KB.

System

In this section, we introduce the main components of the system. Our system is built with a pipeline style architecture giving it the advantage to run each section separately to allow stream processing without blocking the data flow of components (Figure 1). The three logical components are divided into sections entitled *Model* for entity resolution purposes, *Wikipedia Citation* to annotate cite-worthy documents, and *Slot Filling* to generate the actual slot values.

To discover facts for a single WP entity, the first step is to extract aliases of the entity. We extract several name variations from the Wikipedia.org API and from the WP entity page. Also, if the entity type is *person* we can change the order of user names to increase coverage (e.g. 'Boris Berezovsky' $\rightarrow$ 'Berezovsky, Boris'). Next, we iterate over documents in the stream and filter out all documents that do not explicitly contain a string matching the list of entities. To extract relevant facts we perform pattern matching over each sentence that matches the entity based on a dictionary of patterns. If a sentence activates one of the patterns in the dictionary we emit this sentence as a candidate contribution for the WP entity. With the candidate set, we infer new facts from the set and clean up the set by removing the set of values that violate a list of constraints such as duplicates.

Entity Model

We use the Wikipedia.org API to retrieve aliases. The API allows us to requests pages that redirect users to an entity page. For example, if a WP user tries to access the *William Henry Gates* entry they are sent to the page for *Bill Gates* — we treat such redirects as aliases. To extract more aliases we parse the HTML source of a WP entity page. Using regular expressions we extract the bold phrases of the initial paragraph as aliases. This method provides several inline aliases from the wiki page. In WP page for the businesman 'Boris Berezovski' , there is a mention of 'Boris Abramovich Berezovsky' given in bold in the wiki page which obtained by regular expression extraction.

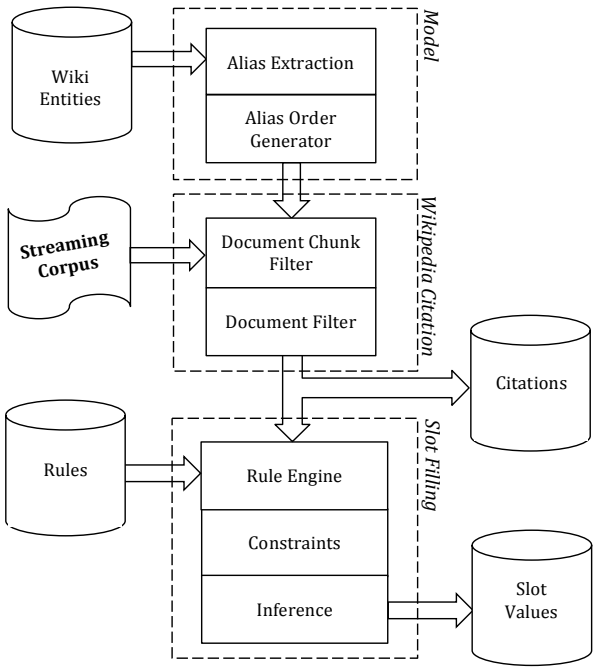

Figure 1: System Architecture. Components are logical groups noted with dotted boxes.

We pass the full set of *person* entities through rules for generating alternate name orders. This module produces various forms of expressing entity names and titles. For example, *Bill Gates* can be written as *Gates, Bill*. This allows the system to capture various notation forms of aliases that appear in text documents.

Wikipedia Citation

The goal of this section is to use the models created to discover a set of documents that are relevant to the WP entity. As a stream of documents come in we first perform a string match between the model aliases and document text. We use this technique as a first filter with confidence because previous work states non-mentioning documents have a low chance of being citable in Wikipedia (**?**). Given our large number of aliases we can be confident that if an alias does not appear in a document it does not need to be cited.

Our system streams in documents in the form of chunk files. Each chunk file contains thousands of documents. This corpus of documents is processed by a two-layer filter system referred to as *Document Chunk Filter* and *Document Filter*. The purpose of these filters is to reduce I/O cost while generating slot values for various entities. Document Chunk Filter removes the chunk files that do not contain a mention of any of the desired entities. Each chunk file may contain thousands of documents — each document is expensive to process. The Document Filter removes documents that do not contain a mention of an entity. This two-level filter allows us to perform detailed slower processing over a smaller set of documents. Not all chunk files contain mention of the entities so filtering out large chunk files early saves on

I/O and processing. Document Chunk Filter discards non-mentioning chunk files and promotes chunk files as soon as an entity mention is found. The document filter additionally notes the sentences that contain entity mentions. This data is passed to the Slot Filling system.

Slot Filling

Streaming Slot Filling (SSF) extracts fact values from sentences according to a list of patterns. Table 1 lists the slot relationships that we looks to extract. In Figure 1 we refer to this task as *Slot Filling*.

SSF reads documents filtered by the Wikipedia Citation step and fetches and tags sentences containing WP entities. All entities are extracted from the document using a natural language processing tool[2]. In the next section, we describe how WP entities are matched against the set of patterns. Following, we discuss our approach to inference over the extracted facts.

Algorithm 1 Slot Value Extraction Pseudocode

List of entities $\mathcal{E} = \{e_0, \ldots, e_{170}\}$
List of patterns $P = \{p_0, \ldots, p_{|P|}\}$
List of documents containing entities $\mathcal{S} = \{s_0, \ldots, s_{|\mathcal{S}|}\}$

```
for si ∈ S do
  for sentence ∈ si do
    for entity ∈ E do
      if Contains(sentence, entity) then
        for pattern ∈ P suitable for entity do
          if Satisfies(sentence, pattern) then
            Emit(sentence, pattern)
```

Rule Engine A pattern is a template of a fact to be extracted and added to a WP entity. Patterns are used to find and extract facts from text. A pattern $\mathcal{P}$ is represented as a five-tuple $\mathcal{P} = \langle p_1, p_2, p_3, p_4, p_5 \rangle$.

The first value, p_1 represents the type of entity. These entity types are in the set $\{\texttt{FAC}, \texttt{ORG}, \texttt{PER}\}$ where `FAC` represents a type of facility, `ORG` represents an organization and `PER` represents a person. p_2 represents a slot name. A list of slot names is present in Table 1. The third element p_3 is the pattern content — a string found in the sentence that identifies a slot name. The extractor looks specifically for pattern content. The pattern evaluator uses a direction (`left` or `right`) found in p_4 to explore sentence. The final element p_5 represents the slot value of a pattern. The type of slot value may be the entity type labeled by the named entity extractor, a noun phrase (`NP`) tagged by a part of speech tagger [3] or a phrase described in the pattern list.

Figure 2 contains the example sentence from the introduction labeled by the rule engine. The matching pattern is ⟨`PER, DateOfDeath,` *`passed away`*`, right, NP`⟩. In the figure, 'Boris Berezovsky' matches as an alias

[2]Lingpipe (available in the dataset) provides entity tags and in-document coreference. http://alias-i.com/lingpipe/

[3]OpenNLP was used for part of speech tagging. http://opennlp.apache.org/

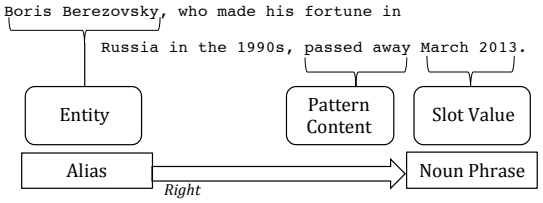

Figure 2: Pattern matching for rule evaluation. The slot value is on the right side of the entity. The pattern context discovered is 'passed away' with a value of 'March 2013'. This conforms to type II of our pattern categories.

for WP entity *wiki/Boris_Berezovsky_(businessman)*, DateOfDeath is the slot name, 'passed away' is the content of the pattern, direction is 'Right' and the actual value of the slot is 'March 2013' which is a Noun Phrase.

There are three types of patterns, each distinguished by different types of slot values (p_5) in the patterns. The matching methods using these three types of patterns are implemented according to the different structures of slot values.

Type I. This pattern type is driven by the entity type. For example, in pattern ⟨`PER, FounderOf,` *`founder`*`, right, ORG`⟩ the `PER` tag means the entity we are finding slot values for is a PER entity; `FounderOf` means this is a pattern for FounderOf slot. *founder* is the word we match in a sentence - the content of pattern occuring in the sentence; `right` means that we are going to the right part of the sentence to match the pattern and find the slot value; ORG means the slot value should be a ORG entity to the right of the entity.

Type II. This pattern type, after finding a matching pattern content, looks for a noun phrase (NP) that is representative of the slot value. For example, pattern ⟨`PER, AwardsWon,` *`awarded`*`, right, NP`⟩ is looking for a noun phrase after the *awarded* that may represent an award. Titles and awards are not named entity tagged hence the use of the part of speech tagger to fetch the noun phrases.

Type III. These types of facts are best discovered by hard coding the slot values. Examples of these include time phrases: ⟨`PER, DateOfDeath,` *`died`*`, right,` *`last night`*⟩. In this pattern, the phrase *last night* is exactly searched in the text last night to the right of the term *died*. The intuition behind this pattern is in news articles that report events in the past relative to the article. For example, an article will mention that a person died 'last night' instead of mentioning precise date-time information. Additionally, part of speech taggers and name entity extractors did not label these terms such as *last night* as a DATE entity.

Constraints and Inference

Our data set contains some duplicate webpages, webpage texts with similar content, and some of the entity tags are incomplete. This causes some duplicates or highly similar content in the extracted list. We implement a filter to remove duplicates or the fact extractions that match patterns that are general and highly susceptible to be noisy. The data contains duplicates and incorrect extractions. We define rules to read

ordered sets of facts to sanitize the output. The input is processed in time order, in a tuple-at-a-time fashion to allow rules to discover noisy slots that appears in close proximity. We define two classes of rules: *deduplication* and *inference* rules.

The output contains many duplicate entries. As we read the list of extracted slots we create rules to define "duplicate". Duplicates can be present in a window of rows; we use a window size of 2 meaning we only be adjacent rows. Two rows are duplicates if they have the same exact extraction or if the rows have the same slot name and a similar slot value or if the extracted sentence for a particular slot types come from the same sentence.

New slots can be deduced from existing slots by defining inference rules. For example, two slots for the task are "FounderOf" and "FoundedBy". A safe assumption is these slot names are biconditional logical connectives with the entities and slot values. Therefore, we can express a rule "X FounderOf Y" $\leftrightarrow$ "Y FoundedBy X" where X and Y are single unique entities. Additionally, we found that the slot names "Contact_Meet_PlaceTime" could be inferred as "Contact_Meet_Entity" if the Entity was a FAC and the extracted sentence contained an additional ORG/FAC tag. We also remove erroneous slots that have extractions that are thousands of characters in length or tool small. Errors of extracting long sentences can typically be attributed to poor sentence parsing of web documents. We have some valid "small" extractions. For example, a comma may separate a name and a title (e.g. "John, Professor at MIT"). But such extraction rules can be particularly noisy, so we check to see if the extracted values have good entity values.

Formatting Requirements in Brief

We need source and PDF files that can be used in a variety of ways and can be output on a variety of devices. AAAI imposes some requirements on your source and PDF files that must be followed. Most of these requirements are based on our efforts to standardize conference manuscript properties and layout. These requirements are as follows, and all papers submitted to AAAI for publication must comply:

- Your .tex file must compile in PDFLaTeX — **no .ps or .eps figure files.**
- All fonts must be embedded in the PDF file — **this includes your figures.**
- Modifications to the style sheet (or your document) in an effort to avoid extra page charges are NOT allowed.
- No type 3 fonts may be used (even in illustrations).
- Your title must follow US capitalization rules.
- LaTeX documents must use the Times or Nimbus font package (do not use Computer Modern for the text of your paper).
- No LaTeX 209 documents may be used or submitted.
- Fonts that require non-English language support (CID and Identity-H) must be converted to outlines or removed from the document (even if they are in a graphics file embedded in the document).
- Two-column format in AAAI style is required for all papers.
- The paper size for final submission must be US letter. No exceptions.
- The source file must exactly match the PDF.
- The document margins must be as specified in the formatting instructions.
- The number of pages and the file size must be as specified for your event.
- No document may be password protected.
- Neither the PDFs nor the source may contain any embedded links or bookmarks.
- Your source and PDF must not have any page numbers, footers, or headers.
- Your PDF must be compatible with Acrobat 5 or higher.
- Your LaTeX source file (excluding references) must consist of a **single** file (use of the "input" command is not allowed.
- Your graphics must be sized appropriately outside of LaTeX (do not use the "clip" command) .

If you do not follow the above requirements, it is likely that we will be unable to publish your paper.

What Files to Submit

You must submit the following items to ensure that your paper is published:

- A fully-compliant PDF file.
- Your LaTeX source file submitted as a **single** .tex file (do not use the "input" command to include sections of your paper — every section must be in the single source file). The only exception is the bibliography, which you may include separately. Your source must compile on our system, which includes the standard LaTeX support files.
- All your graphics files.
- The LaTeX-generated files (e.g. .aux and .bib file, etc.) for your compiled source.
- All the nonstandard style files (ones not commonly found in standard LaTeX installations) used in your document (including, for example, old algorithm style files). If in doubt, include it.

Your LaTeX source will be reviewed and recompiled on our system (if it does not compile, you may incur late fees). **Do not submit your source in multiple text files.** Your single LaTeX source file must include all your text, your bibliography (formatted using aaai.bst), and any custom macros. Accompanying this source file, you must also supply any nonstandard (or older) referenced style files and all your referenced graphics files.

Your files should work without any supporting files (other than the program itself) on any computer with a standard LaTeX distribution. Place your PDF and source files in a single tar, zipped, gzipped, stuffed, or compressed archive. Name your source file with your last (family) name.

Do not send files that are not actually used in the paper. We don't want you to send us any files not needed for compiling your paper, including, for example, this instructions file, unused graphics files, and so forth. A shell script (created by an AAAI member — it might not work without modification on your system) that might help you create the LaTeX source package is included in the Author Kit.

Using LaTeX to Format Your Paper

The latest version of the AAAI style file is available on AAAI's website. Download this file and place it in a file named "aaai.sty" in the TeX search path. Placing it in the same directory as the paper should also work. You must download the latest version of the complete author kit so that you will have the latest instruction set.

Document Preamble

In the LaTeX source for your paper, you **must** place the following lines as shown in the example in this subsection. This command set-up is for three authors. Add or subtract author and address lines as necessary, and uncomment the portions that apply to you. In most instances, this is all you need to do to format your paper in the Times font. The helvet package will cause Helvetica to be used for sans serif, and the courier package will cause Courier to be used for the typewriter font. These files are part of the PSNFSS2e package, which is freely available from many Internet sites (and is often part of a standard installation).

Leave the setcounter for section number depth commented out and set at 0 unless you want to add section numbers to your paper. If you do add section numbers, you must uncomment this line and change the number to 1 (for section numbers), or 2 (for section and subsection numbers). The style file will not work properly with numbering of subsubsections, so do not use a number higher than 2.

```
\documentclass[letterpaper]article
% Required Packages
\usepackage{aaai}
\usepackage{times}
\usepackage{helvet}
\usepackage{courier}
\setlength{\pdfpagewidth}{8.5in}
\setlength{\pdfpageheight}{11in}
%%%%%%%%%%
% PDFINFO for PDFLaTeX
% Uncomment and complete the following for metadata
(your paper must compile with PDFLaTeX)
\pdfinfo{
/Title (Input Your Paper Title Here)
/Author (John Doe, Jane Doe)
/Keywords (Input your paper's keywords in this optional
area)
}
%%%%%%%%%%
% Section Numbers
% Uncomment if you want to use section numbers
% and change the 0 to a 1 or 2
% \setcounter{secnumdepth}{0}
%%%%%%%%%%
% Title, Author, and Address Information
\title{Title}
\author{Author 1 \and Author 2\\
Address line\\
Address line\\
\And
Author 3\\
Address line\\
Address line}
%%%%%%%%%%
% Body of Paper Begins
\begin{document}
\maketitle
...
%%%%%%%%%%
% References and End of Paper
\bibliography{Bibliography-File}
\bibliographystyle{aaai}
\end{document}
```

Inserting Document Metadata with LaTeX

PDF files contain document summary information that enables us to create an Acrobat index (pdx) file, and also allows search engines to locate and present your paper more accurately. **Document Metadata for Author and Title are REQUIRED.**

If your paper includes illustrations that are not compatible with PDFTeX (such as .eps or .ps documents), you will need to convert them. The epstopdf package will usually work for eps files. You will need to convert your ps files to PDF however.

Important: Do not include *any* LaTeX code or nonascii characters (including accented characters) in the metadata. The data in the metadata must be completely plain ascii. It may not include slashes, accents, linebreaks, unicode, or any LaTeX commands. Type the title exactly as it appears on the paper (minus all formatting). Input the author names in the order in which they appear on the paper (minus all accents), separating each author by a comma. You may also include keywords in the Keywords field.

Preparing Your Paper

After the preamble above, you should prepare your paper as follows:

```
\begin{document}
\maketitle
...
\bibliography{Bibliography-File}
\bibliographystyle{aaai}
\end{document}
```

Incompatible Packages

The following packages are incompatible with aaai.sty and/or aaai.bst and must not be used (this list is not exhaustive — there are others as well):

- hyperref

- natbib
- geometry
- titlesec
- layout
- caption
- titlesec
- T1 fontenc package (install the CM super fonts package instead)

Illegal Commands

The following commands may not be used in your paper:

- \input
- \vspace (when used before or after a section or subsection)
- \addtolength
- \columnsep
- \top margin (or text height or addsidemargin or even side margin)

Paper Size, Margins, and Column Width

Papers must be formatted to print in two-column format on 8.5 x 11 inch US letter-sized paper. The margins must be exactly as follows:

- Top margin: .75 inches
- Left margin: .75 inches
- Right margin: .75 inches
- Bottom margin: 1.25 inches

The default paper size in most installations of LaTeX is A4. However, because we require that your electronic paper be formatted in US letter size, you will need to alter the default for this paper to US letter size. Assuming you are using the 2e version of LaTeX, you can do this by including the [letterpaper] option at the beginning of your file: \documentclass[letterpaper]article.

This command is usually sufficient to change the format. Sometimes, however, it may not work. Use PDFLaTeX and include \setlength{\pdfpagewidth}{8.5in} \setlength{\pdfpageheight}{11in} in your preamble.

Do not use the Geometry package to alter the page size. Use of this style file alters aaai.sty and will result in your paper being rejected.

Column Width and Margins. To ensure maximum readability, your paper must include two columns. Each column should be 3.3 inches wide (slightly more than 3.25 inches), with a .375 inch (.952 cm) gutter of white space between the two columns. The aaai.sty file will automatically create these columns for you.

Overlength Papers

If your paper is too long, turn on \frenchspacing, which will reduce the space after periods. Next, shrink the size of your graphics. Use \centering instead of \begin{center} in your figure environment. If these two methods don't work, you may minimally use the following. For floats (tables and figures), you may minimally reduce \floatsep, \textfloatsep, \abovecaptionskip, and \belowcaptionskip. For mathematical environments, you may minimally reduce \abovedisplayskip, \belowdisplayskip, and \arraycolsep. You may also alter the size of your bibliography by inserting \fontsize{9.5pt}{10.5pt} \selectfont right before the bibliography.

Commands that alter page layout are forbidden. These include \columnsep, \topmargin, \topskip, \textheight, \textwidth, \oddsidemargin, and \evensizemargin (this list is not exhaustive). If you alter page layout, you will be required to pay the page fee *plus* a reformatting fee. Other commands that are questionable and may cause your paper to be rejected include \parindent, and \parskip. Commands that alter the space between sections are also questionable. The title sec package is not allowed. Regardless of the above, if your paper is obviously "squeezed" it is not going to to be accepted. Before using every trick you know to make your paper a certain length, try reducing the size of your graphics or cutting text instead or (if allowed) paying the extra page charge. It will be cheaper in the long run.

Figures

Your paper must compile in PDFLaTeX. Consequently, all your figures must be .jpg, .png, or .pdf. You may not use the .gif (the resolution is too low), .ps, or .eps file format for your figures.

When you include your figures, you must crop them **outside** of LaTeX. The command \includegraphics*[clip=true, viewport 0 0 10 10]... might result in a PDF that looks great, but the image is **not really cropped.** The full image can reappear when page numbers are applied or color space is standardized.

Type Font and Size

Your paper must be formatted in Times Roman or Nimbus. We will not accept papers formatted using Computer Modern or Palatino or some other font as the text or heading typeface. Sans serif, when used, should be Courier. Use Symbol or Lucida or Computer Modern for *mathematics only.*

Do not use type 3 fonts for any portion of your paper, including graphics. Type 3 bitmapped fonts are designed for fixed resolution printers. Most print at 300 dpi even if the printer resolution is 1200 dpi or higher. They also often cause high resolution imagesetter devices and our PDF indexing software to crash. Consequently, AAAI will not accept electronic files containing obsolete type 3 fonts. Files containing those fonts (even in graphics) will be rejected.

Fortunately, there are effective workarounds that will prevent your file from embedding type 3 bitmapped fonts. The easiest workaround is to use the required times, helvet, and courier packages with LaTeX2e. (Note that papers formatted

in this way will still use Computer Modern for the mathematics. To make the math look good, you'll either have to use Symbol or Lucida, or you will need to install type 1 Computer Modern fonts — for more on these fonts, see the section "Obtaining Type 1 Computer Modern."

If you are unsure if your paper contains type 3 fonts, view the PDF in Acrobat Reader. The Properties/Fonts window will display the font name, font type, and encoding properties of all the fonts in the document. If you are unsure if your graphics contain type 3 fonts (and they are PostScript or encapsulated PostScript documents), create PDF versions of them, and consult the properties window in Acrobat Reader.

The default size for your type should be ten-point with twelve-point leading (line spacing). Start all pages (except the first) directly under the top margin. (See the next section for instructions on formatting the title page.) Indent ten points when beginning a new paragraph, unless the paragraph begins directly below a heading or subheading.

Obtaining Type 1 Computer Modern for LaTeX. If you use Computer Modern for the mathematics in your paper (you cannot use it for the text) you may need to download type 1 Computer fonts. They are available without charge from the American Mathematical Society: http://www.ams.org/tex/type1-fonts.html.

Title and Authors

Your title must appear in mixed case (nouns, pronouns, and verbs are capitalized) near the top of the first page, centered over both columns in sixteen-point bold type (twenty-four point leading). This style is called "mixed case." Author's names should appear below the title of the paper, centered in twelve-point type (with fifteen point leading), along with affiliation(s) and complete address(es) (including electronic mail address if available) in nine-point roman type (the twelve point leading). (If the title is long, or you have many authors, you may reduce the specified point sizes by up to two points.) You should begin the two-column format when you come to the abstract.

Formatting Author Information Author information can be set in a number of different styles, depending on the number of authors and the number of affiliations you need to display. For several authors from the same institution, use \and:

```
\author{Author 1 \and ... \and Author n\\
Address line \\ ... \\ Address line}
```

If the names do not fit well on one line use:

```
\author{Author 1}\\
{\bf Author 2}\\ ... \\ {\bf Author n}\\
Address line \\ ... \\ Address line}
```

For authors from different institutions, use \And:

```
\author{Author 1\\ Address line \\ ... \\ Address line
\And ... \And Author n\\
Address line\\ ... \\ Address line}
```

To start a separate "row" of authors, use \AND:

```
\author{Author 1\\ Address line \\ ... \\ Address line\\
\AND
Author 2 \\ Address line \\ ... \\ Address line\\
\And
Author 3 \\ Address line \\ ... \\ Address line\\
}
```

If the title and author information does not fit in the area allocated, place \setlength\titlebox{*height*} after the \documentclass line where {*height*} is something like 2.5in.

LaTeX Copyright Notice

The copyright notice automatically appears if you use aaai.sty. If you are creating a technical report, it is not necessary to include this notice. You may disable the copyright line using the \nocopyrightcommand. To change the entire text of the copyright slug, use: \copyrighttext {*text*}. Either of these must appear before \maketitle. Please be advised, however, that *if you disable or change the copyright line and transfer of copyright is required, your paper will not be published.*

Credits

Any credits to a sponsoring agency should appear in the acknowledgments section, unless the agency requires different placement. If it is necessary to include this information on the front page, use \thanks in either the \author or \title commands. For example:

```
\title{Very Important Results in AI\thanks{This work is supported by everybody.}}
```

Multiple \thanks commands can be given. Each will result in a separate footnote indication in the author or title with the corresponding text at the botton of the first column of the document. Note that the \thanks command is fragile. You will need to use \protect.

Please do not include \pubnote commands in your document.

Abstract

The abstract must be placed at the beginning of the first column, indented ten points from the left and right margins. The title Abstract should appear in ten-point bold type, centered above the body of the abstract. The abstract should be set in nine-point type with ten-point leading. This concise, one-paragraph summary should describe the general thesis and conclusion of your paper. A reader should be able to learn the purpose of the paper and the reason for its importance from the abstract. The abstract should be no more than two hundred words in length. (Authors who are submitting short one- or two-page extended extracts should provide a short abstract of only a sentence or so.) **Do not include references in your abstract!**

Page Numbers

Do not **ever** print any page numbers on your paper.

Text

The main body of the paper must be formatted in ten-point with twelve-point leading (line spacing).

Citations

Citations within the text should include the author's last name and year, for example (Newell 1980). Append lowercase letters to the year in cases of ambiguity. Multiple authors should be treated as follows: (Feigenbaum and Engelmore 1988) or (Ford, Hayes, and Glymour 1992). In the case of four or more authors, list only the first author, followed by et al. (Ford et al. 1997).

Extracts

Long quotations and extracts should be indented ten points from the left and right margins.

> This is an example of an extract or quotation. Note the indent on both sides. Quotation marks are not necessary if you offset the text in a block like this, and properly identify and cite the quotation in the text.

Footnotes

Avoid footnotes as much as possible; they interrupt the reading of the text. When essential, they should be consecutively numbered throughout with superscript Arabic numbers. Footnotes should appear at the bottom of the page, separated from the text by a blank line space and a thin, half-point rule.

Headings and Sections

When necessary, headings should be used to separate major sections of your paper. Remember, you are writing a short paper, not a lengthy book! An overabundance of headings will tend to make your paper look more like an outline than a paper.

First-level heads should be twelve-point Times Roman bold type, mixed case (initial capitals followed by lower case on all words except articles, conjunctions, and prepositions, which should appear entirely in lower case), with fifteen-point leading, centered, with one blank line preceding them and three additional points of leading following them. Second-level headings should be eleven-point Times Roman bold type, mixed case, with thirteen-point leading, flush left, with one blank line preceding them and three additional points of leading following them. Do not skip a line between paragraphs. Third-level headings should be run in with the text, ten-point Times Roman bold type, mixed case, with twelve-point leading, flush left, with six points of additional space preceding them and no additional points of leading following them.

Section Numbers The use of section numbers in AAAI Press papers is optional. To use section numbers in LaTeX, uncomment the setcounter line in your document preamble and change the 0 to a 1 or 2. Section numbers should not be used in short poster papers.

Section Headings. Sections should be arranged and headed as follows:

Acknowledgments. The acknowledgments section, if included, appears after the main body of text and is headed "Acknowledgments." This section includes acknowledgments of help from associates and colleagues, credits to sponsoring agencies, financial support, and permission to publish. Please acknowledge other contributors, grant support, and so forth, in this section. Do not put acknowledgments in a footnote on the first page. If your grant agency requires acknowledgment of the grant on page 1, limit the footnote to the required statement, and put the remaining acknowledgments at the back. Please try to limit acknowledgments to no more than three sentences.

Appendices. Any appendices follow the acknowledgments, if included, or after the main body of text if no acknowledgments appear.

References The references section should be labeled "References" and should appear at the very end of the paper (don't end the paper with references, and then put a figure by itself on the last page). A sample list of references is given later on in these instructions. Please use a consistent format for references. Poorly prepared or sloppy references reflect badly on the quality of your paper and your research. Please prepare complete and accurate citations.

Illustrations and Figures

Figures, drawings, tables, and photographs should be placed throughout the paper near the place where they are first discussed. Do not group them together at the end of the paper. If placed at the top or bottom of the paper, illustrations may run across both columns. Figures must not invade the top, bottom, or side margin areas. Figures must be inserted using the \usepackage{graphicx}. Number figures sequentially, for example, figure 1, and so on.

The illustration number and caption should appear under the illustration. Labels, and other text in illustrations must be at least nine-point type.

Low-Resolution Bitmaps. You may not use low-resolution (such as 72 dpi) screen-dumps and GIF files—these files contain so few pixels that they are always blurry, and illegible when printed. If they are color, they will become an indecipherable mess when converted to black and white. This is always the case with gif files, which should never be used. The resolution of screen dumps can be increased by reducing the print size of the original file while retaining the same number of pixels. You can also enlarge files by manipulating them in software such as PhotoShop. Your figures should be a minimum of 266 dpi when incorporated into your document.

LaTeX Overflow. LaTeX users please beware: LaTeX will sometimes put portions of the figure or table or an equation in the margin. If this happens, you need to scale the figure or table down, or reformat the equation. Check your log file! You must fix any overflow into the margin (that means no overfull boxes in LaTeX). If you don't, the overflow text will simply be eliminated. **Nothing is permitted to intrude into the margins.**

Using Color. Your paper will be printed in black and white and grayscale. Consequently, because conversion to grayscale can cause undesirable effects (red changes to black, yellow can disappear, and so forth), we strongly suggest you avoid placing color figures in your document. Of course, any reference to color will be indecipherable to your reader.

Drawings. We suggest you use computer drawing software (such as Adobe Illustrator or, (if unavoidable), the drawing tools in Microsoft Word) to create your illustrations. Do not use Microsoft Publisher. These illustrations will look best if all line widths are uniform (half- to two-point in size), and you do not create labels over shaded areas. Shading should be 133 lines per inch if possible. Use Times Roman or Helvetica for all figure call-outs. **Do not use hairline width lines** — be sure that the stroke width of all lines is at least .5 pt. Zero point lines will print on a laser printer, but will completely disappear on the high-resolution devices used by our printers.

Photographs and Images. Photographs and other images should be in grayscale (color photographs will not reproduce well; for example, red tones will reproduce as black, yellow may turn to white, and so forth) and set to a minimum of 266 dpi. Do not prescreen images.

Resizing Graphics. Resize your graphics **before** you include them with LaTeX. You may **not** use trim or clip options as part of your \includgraphics command. Resize the media box of your PDF using a graphics program instead.

Fonts in Your Illustrations You must embed all fonts in your graphics before including them in your LaTeX document.

References

The aaai.sty file includes a set of definitions for use in formatting references with BibTeX. These definitions make the bibliography style fairly close to the one specified below. To use these definitions, you also need the BibTeX style file "aaai.bst," available in the author kit on the AAAI web site. Then, at the end of your paper but before \enddocument, you need to put the following lines:

\bibliographystyle{aaai} \bibliography{bibfile1,bibfile2,...}

The list of files in the \bibliography command should be the names of your BibTeX source files (that is, the .bib files referenced in your paper).

The following commands are available for your use in citing references:

\cite: Cites the given reference(s) with a full citation. This appears as "(Author Year)" for one reference, or "(Author Year; Author Year)" for multiple references.

\shortcite: Cites the given reference(s) with just the year. This appears as "(Year)" for one reference, or "(Year; Year)" for multiple references.

\citeauthor: Cites the given reference(s) with just the author name(s) and no parentheses.

\citeyear: Cites the given reference(s) with just the date(s) and no parentheses.

Warning: The aaai.sty file is incompatible with the hyperref and natbib packages. If you use either, your references will be garbled.

Formatted bibliographies should look like the following examples.

Book with Multiple Authors
Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Systems.* Reading, Mass.: Addison-Wesley.

Journal Article
Robinson, A. L. 1980a. New Ways to Make Microcircuits Smaller. *Science* 208: 1019–1026.

Magazine Article
Hasling, D. W.; Clancey, W. J.; and Rennels, G. R. 1983. Strategic Explanations in Consultation. *The International Journal of Man-Machine Studies* 20(1): 3–19.

Proceedings Paper Published by a Society
Clancey, W. J. 1983b. Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education. In Proceedings of the Eighth International Joint Conference on Artificial Intelligence, 556–560. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, Inc.

Proceedings Paper Published by a Press or Publisher
Clancey, W. J. 1984. Classification Problem Solving. In *Proceedings of the Fourth National Conference on Artificial Intelligence,* 49–54. Menlo Park, Calif.: AAAI Press.

University Technical Report
Rice, J. 1986. Poligon: A System for Parallel Problem Solving, Technical Report, KSL-86-19, Dept. of Computer Science, Stanford Univ.

Dissertation or Thesis
Clancey, W. J. 1979b. Transfer of Rule-Based Expertise through a Tutorial Dialogue. Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif.

Forthcoming Publication
Clancey, W. J. 1986a. The Engineering of Qualitative Models. Forthcoming.

Producing Reliable PDF Documents with LaTeX

Generally speaking, PDF files are platform independent and accessible to everyone. When creating a paper for a proceedings or publication in which many PDF documents must be merged and then printed on high-resolution PostScript RIPs, several requirements must be met that are not normally of concern. Thus to ensure that your paper will look like it does when printed on your own machine, you must take several precautions:

- Use type 1 fonts (not type 3 fonts)
- Use only standard Times, Nimbus, and CMR font packages (not fonts like F3 or fonts with tildes in the names or fonts—other than Computer Modern—that are created for specific point sizes, like Times˜19) or fonts with strange combinations of numbers and letters
- Embed all fonts when producing the PDF

- Do not use the [T1]fontenc package (install the CM super fonts package instead)

Creating Output Using PDFLaTeX Is Required

By using the PDFTeX program instead of straight LaTeX or TeX, you will probably avoid the type 3 font problem altogether (unless you use a package that calls for metafont). PDFLaTeX enables you to create a PDF document directly from LaTeX source. The one requirement of this software is that all your graphics and images must be available in a format that PDFLaTeX understands (normally PDF).

PDFLaTeX's default is to create documents with type 1 fonts. If you find that it is not doing so in your case, it is likely that one or more fonts are missing from your system or are not in a path that is known to PDFLaTeX.

dvipdf Script Scripts such as dvipdf which ostensibly bypass the Postscript intermediary should not be used since they generally do not instruct dvips to use the config.pdf file.

dvipdfm Do not use this dvi-PDF conversion package if your document contains graphics (and we recommend you avoid it even if your document does not contain graphics).

Ghostscript

LaTeX users should not use GhostScript to create their PDFs.

Graphics

If you are still finding type 3 fonts in your PDF file, look at your graphics! LaTeX users should check all their imported graphics files as well for font problems.

Proofreading Your PDF

Please check all the pages of your PDF file. Is the page size A4? Are there any type 3, Identity-H, or CID fonts? Are all the fonts embedded? Are there any areas where equations or figures run into the margins? Did you include all your figures? Did you follow mixed case capitalization rules for your title? Did you include a copyright notice? Do any of the pages scroll slowly (because the graphics draw slowly on the page)? Are URLs underlined and in color? You will need to fix these common errors before submitting your file.

Improperly Formatted Files

In the past, AAAI has corrected improperly formatted files submitted by the authors. Unfortunately, this has become an increasingly burdensome expense that we can no longer absorb. Consequently, if your file is improperly formatted, it may not be possible to include your paper in the publication. If time allows, however, you will be notified via e-mail (with a copy to the program chair) of the problems with your file and given the option of correcting the file yourself (and paying a late fee) or asking that AAAI have the file corrected for you, for an additional fee. If you opt to correct the file yourself, please note that we cannot provide you with any additional advice beyond that given in your packet. Files that are not corrected after a second attempt will be withdrawn.

LaTeX 209 Warning

If you use LaTeX 209 we will not be able to publish your paper. Convert your paper to LaTeX2e.

Naming Your Electronic File

We request that you name your LaTeX source file with your last name (family name) so that it can easily be differentiated from other submissions. If you name your files with the name of the event or "aaai" or "paper" or "camera-ready" or some other generic or indecipherable name, you bear all risks of loss — it is extremely likely that your file may be overwritten.

Submitting Your Electronic Files to AAAI

Submitting your files to AAAI is a two-step process. It is explained fully in the author registration and submission instructions. Please consult this document for details on how to submit your paper.

Inquiries

If you have any questions about the preparation or submission of your paper as instructed in this document, please contact AAAI Press at the address given below. If you have technical questions about implementation of the aaai style file, please contact an expert at your site. We do not provide technical support for LaTeX or any other software package. To avoid problems, please keep your paper simple, and do not incorporate complicated macros and style files.

AAAI Press
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303
Telephone: (650) 328-3123
E-mail: See the submission instructions for your particular conference or event.

Additional Resources

LaTeX is a difficult program to master. If you've used that software, and this document didn't help or some items were not explained clearly, we recommend you read Michael Shell's excellent document (testflow doc.txt V1.0a 2002/08/13) about obtaining correct PS/PDF output on LaTeX systems. (It was written for another purpose, but it has general application as well). It is available at www.ctan.org in the tex-archive.

Acknowledgments

AAAI is especially grateful to Peter Patel Schneider for his work in implementing the aaai.sty file, liberally using the ideas of other style hackers, including Barbara Beeton. We also acknowledge with thanks the work of George Ferguson for his guide to using the style and BibTeX files — which has been incorporated into this document — and Hans Guesgen, who provided several timely modifications, as well as the many others who have, from time to time, sent in suggestions on improvements to the AAAI style.

The preparation of the LaTeX and BibTeX files that implement these instructions was supported by Schlumberger

Palo Alto Research, AT&T Bell Laboratories, Morgan Kaufmann Publishers, The Live Oak Press, LLC, and AAAI Press. Bibliography style changes were added by Sunil Issar. \pubnote was added by J. Scott Penberthy. George Ferguson added support for printing the AAAI copyright slug. Additional changes to aaai.sty and aaai.bst have been made by the AAAI staff.

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!