NTCIR-12 MathIR Task Wikipedia Corpus (v0.2.1)

This is the **revised (v 0.2.1)** version of the 'Wikipedia' corpus for the NTCIR-12 Mathematical Information Retrieval (MathIR) tasks (see http://ntcir-math.nii.ac.jp/introduction/).

v0.1 October 2015, v0.2/0.2.1 February 2016

Richard Zanibbi *(rlaz@cs.rit.edu)*
Document and Pattern Recognition Lab *(https://www.cs.rit.edu/~dprl)*
Department of Computer Science
Rochester Institute of Technology, Rochester, NY, USA

A special thanks to Michal Růžička and the WebMIAS group for generating XHTML files for v0.2.1

Changes from v0.1

See the bottom of this document for additional information on corrections made.

1. Meta data for each article now includes a **docid** attribute (the name of the file, without an extension).
2. Corrected a large number of missing and incorrectly translated articles and formulae.
3. Multiple <body> and <html> tags have been corrected.
4. Handling of '<' and '>' in body text (e.g. for y < z) has been corrected (previously this led to invalid HTML being produced).
5. Translation for some symbol types (e.g. angle brackets) have been corrected.
6. Articles now have a margin for easier reading.
7. (v0.2.1) - files are now in XHTML format, ampersands have been replaced.
8. (v0.2.1) - ampersands have been replaced (including TeX strings) by & to prevent XML parse errors.
9. (v0.2.1) - Formula identifier './' prefixes have been removed.
10. (v0.2.1) - Corpus is now a single compressed .tar.bz2 file.

Overview

All articles provided in the corpus must be included in the search index by for NTCIR-12 MathIR Wikipedia task participants. This is the case for both the keyword + formula query

task, and the formula browsing sub-task. This includes all articles stored in sub-directories named Articles/ and Articles_LaTeXML_Errors/, for example in:

- MathTagArticles/wpmath0000001/Articles
- MathTagArticles/wpmath0000001/Articles_LaTeXML_Errors

For convenience, the collection is broken into two main sections, with 'math' articles that contain explicit <math> tags (*MathTagArticles/* directory, with 16 .tar.bz2 archives), and 'text' articles that do not (*TextArticles/* directory, with 144 .tar.bz2 archives).

The corpus has been split into 160 parts, holding around 2000 articles each, located in sub-directories with the prefix *wpmath* or *wp*. Each part contains articles in .html format, stored in the Article/ and Article_LaTeXML_Errors/ sub-directories. Some additional book-keeping data about the archive contents and conversion process are included in the conversion_data subdirectories.

Across the entire collection, math frequently appears unmarked in .html body text, without an encapsulating tag or mediawiki template (e.g. '*Given this, x <sup> 2 has the property...*'). Formula counts below do not include these unmarked formulas.

Licensing information along with details on data storage, decompression, formula representation, and corpus generation are provided below.

Corpus Contents

319,689 HTML articles

- *MathTagArticles* [~10% of collection]
    - 31,839 articles
- *TextArticles* [~90% of collection]
    - 287,850 articles

592,443 formulae inside <math> tags (approx.)

- *MathTagArticles*
    - 580,068 marked formulae
- *TextArticles*
    - 12,375 marked formulae (many very small, e.g. isolated symbols)

Wikipedia Data and License

The Wikipedia articles included in this collection are being provided under a Creative Commons BY-SA license (http://creativecommons.org/licenses/by-sa/3.0/), as required by the Wikimedia foundation (https://en.wikipedia.org/wiki/Wikipedia:Copyrights#Reusers.27_rights_and_obligations).

The complete original content of all articles in this collection may be obtained from the Aug. 5, 2015 text-only Wikipedia snapshot provided online:

```
https://dumps.wikimedia.org/enwiki/20150805/enwiki-20150805-pages-
articles.xml.bz2
```

Wikipedia articles included in this collection are unchanged aside from alterations resulting from the automated MediaWiki to HTML conversion process.

Our sincerest thanks go to Wikipedia authors and the people at the Wikimedia Foundation for making snapshots publicly available.

Decompressing the Corpus

The corpus can be decompressed using:

```
tar jxf NTCIR12_MathIR_WikiCorpus_v2.1.0.tar.bz2
```

Or **more quickly** using the parallel bzip2 implementation (pbzip2 library):

```
tar xv -I pbzip2 -f NTCIR12_MathIR_WikiCorpus_v2.1.0.tar.bz2
```

If this doesn't work, the decompression be done in two steps, using:

```
pbunzip2 NTCIR12_MathIR_WikiCorpus_v2.1.0.tar.bz2
tar xvf NTCIR12_MathIR_WikiCorpus_v2.1.0.tar
```

MathML Formula Representation

Formulae are translated to MathML, an XML encoding (http://www.w3.org/Math/). Each formula appears as a <math> tag, and is annotated with a unique identifier (the name of the file, followed by the relative offset of the formula in the file, e.g. *id="FileName:0"* for the first formula in *FileName.html*).

LaTeXML (http://dlmf.nist.gov/LaTeXML/) is used to convert each formula from LaTeX to MathML, producing three representations for each formula:

1. **Presentation MathML** (symbol layout (appearance)). This is shown first, directly below a *<semantics>* tag.
2. **Content MathML** (operator tree (mathematical semantics)). Where possible, LaTeXML provides an operator tree representing the mathematical semantics of an expression. This is demarcated by this tag: *<annotation-xml encoding="MathML-Content">.*
3. **LaTeX string** (symbol layout (appearance)), demarcated by: *<annotation-xml encoding="application/x-tex">.*

All articles contain a reference to MathJax (https://www.mathjax.org/), which renders the Presentation MathML as SVG in modern web browsers.

!! v0.2 Addition: Incorrectly Translated Formulas. Formulas may fail to be translated by LaTeXML, in which case the <math> tag will have an attribute *class="LaTeXML:Error"* and contain only the original LaTeX string. Formulae missed during parsing or at other points in the processing chain will appear as a LaTeX string inside a *<span class="LaTeX">* tag. Formulas that end up empty after translation are replaced by the string *[-FormulaError-]*. See the bottom of this document for additional information.

Corpus Creation

1. **Raw Data.** A 'raw' mediawiki dump (i.e. the shorthand markup language used for Wikipedia) was obtained online:

   https://dumps.wikimedia.org/enwiki/20150805/enwiki-20150805-pages-articles.xml.bz2

   This is the Aug. 5, 2015 snapshot of English Wikipedia, omitting non-textual content (e.g. images). The raw dump file was just over 54GB in size.

2. **Article Extraction.** A python program using the 'iterparse' library for iterative parsing was used to:

   - Remove articles that were 'redirect' articles (i.e. entries that only refer to other articles)
   - Remove 'meta' articles about Wikipedia (with titles beginning with "Wikipedia:")
   - Identify articles containing LaTeX formulae, demarcated by <math> tags, as 'math' articles
   - Accept remaining articles as 'text' articles

   After splitting articles into 'math' and 'text' groups:

   - Article entries were converted to title and content fields.
   - Articles were stored in batches of 2000 articles, to avoid overwhelming file systems and make manipulation at the command line feasible.

   The resulting text archive (.dat) files, containing roughly 2000 articles each were stored on disk. In total, there were 3897 archive files extracted from the Wikipedia dump: 16 'math' and 3881 'text.'

3. **Sampling 'Text' Articles.** To keep the size of the corpus relatively small while still insuring that many 'non-math' articles were present in the collection, it was decided to include 144 randomly selected 'text' archives, so that 'math' articles would comprise roughly 10% of the collection. Note that the percentage of math articles in the full English Wikipedia collection is actually *much* smaller (roughly 16 / 3897 = 0.41%).

4. **MediaWiki Text to HTML Conversion.** *pandoc* (http://pandoc.org/) was used to make an initial translation from mediawiki to HTML text. The command used was the following:

   ```
    pandoc --latexmathml -f mediawiki -t html5 -s <mediwikifile.txt> -o
   <outputfile.html>
   ```

   --latexmathml converts <math> tags to <span> tags with a "LaTeX" class attribute. HTML5 was used as the target language for output.

5. **MediaWiki Math Templates to LaTeX Conversion.** A simple recursive-descent parser was implemented in python to convert MediaWiki math 'templates' (e.g. {{frac|1|2}} for '1/2') to LaTeX (these are not demarcated by <math> tags). While it is unlikely that all templates were located, we attempted to translate nearly all templates described online at: **https://en.wikipedia.org/wiki/Category:Mathematical_formatting_templates** (accessed late Sept. 2015; a .pdf snapshot of the list of template names was generated on Oct. 8, 2015, and is included as a .pdf file in this directory).

   Related articles: https://en.wikipedia.org/wiki/Help:Displaying_a_formula, https://en.wikipedia.org/wiki/Template:Math .

6. **MediaWiki Table Templates to HTML Conversion.** Conversion of tables in MediaWiki to HTML syntax was another important task, as these are very frequent in the collection. To avoid conversion failures that were occurring with pandoc when ill-formatted tables were provided, almost all formatting information (e.g. 'align', 'valign', 'width,' etc.) was removed and some other simple normalizations made, but with the goal to preserve all table headers, data and captions (i.e. table content). The MediaWiki table format is described here: https://www.mediawiki.org/wiki/Help:Tables .

7. **Figure Conversion.** To save space, all image tags were removed, but the captions from their associated figures were retained. Figure regions are indented and indicated by a bold **(Figure)** prefix in the final article set.

8. **Math and Table Translation.** LaTeXML (http://dlmf.nist.gov/LaTeXML/) was used to translate LaTeX strings to MathML (see above). Each formula has a unique identifier indicated by an 'id' attribute. Formula identifiers are comprised of the article file name, ':' and the relative offset for the formula in the file. For example, the first formula in file

'Ex_Article.html' is "Ex_Article:0", the second formula "Ex_Article:1," and so on.

9. **Conversion of Archive Files to HTML.** Using a combination of bash scripts, perl one-liners and python programs, the conversion process involved:

    a. Extracting articles from a 'raw' .dat text archive file (see Step 2, above)
    b. Converting MediaWiki to HTML using pandoc, after pre-processing MediaWiki math/table data and modifying figures, followed by some simple post-processing,
    c. Using the Python *BeautifulSoup* library to extract all formula regions, calling LaTeXML to generate MathML content, and then replacing each LaTeX formula with its corresponding LaTeXML output.
    d. Compressing the resulting articles and conversion process data as a .tar.bz2 archive file.

    Note: both 'math' and 'text' .dat archives were converted in the same way. A number of formulae represented by MediaWiki math templates were found (and converted) for the 'text' articles.

10. **(v 0.2.1) XHTML Conversion.** See the file *XHTML_Generation_Notes* in this directory for details from Michal Růžička regarding the conversion of the initially generated HTML files to XHTML.

(v0.2) Notes on Conversion from MediaWiki

1. **Document Names.** Document names were translated as given in the original raw mediawiki file, replacing spaces by underscores (_), and replacing forward slashes (/) by two colons (::).

2. **Newlines.** pandoc is sensitive to spaces in mediawiki text, and in particular will incorrectly translate tables, mediawiki templates (e.g. {{math|...}}) and math regions (<math>...</math>) if there are newlines. This was handled by locating these region types, and replacing all newlines within them by a single space.

3. **MediaWikiTemplates.** pandoc filters any mediawiki templates that it does not recognize, which includes {{math|...}} and {{cite [X]|...}} entries (containing math and citations, respectively). Code was written to parse and preserve these regions. Citations are currently just represented by a list of attribute values for authors, titles, etc. Sometimes citations embedded within outer templates are filtered by pandoc, and so some citations at the bottom of an article are empty.

4. **Incorrectly Translated Formulae.** Formulae that are not converted successfully by LaTeXML have a <math> tag, but with an attribute *class="LaTeXML::Error"*. The LaTeX string is preserved, so that these can still be located. Formula that are empty after translation do not have this attribute, but are instead replaced by *[-FormulaError-]* to

avoid LaTeXML failures. In some cases tags labeled <span class="LaTeX">, representing math regions created by pandoc that were not successfullly translated might also remain (we have tried to minimize these in the new version).

5. **< and >.** pandoc first needs to be run to normalize the input mediawiki, which may contain '<' and '>' in body text (creating invalid HTML). These symbols need to be represented in body text as &lt; and &gt;. Using the 'minimal' formatter in BeautifulSoup (python XML/HTML parsing library) writer preserves this distinction. Previously the 'none' formatter was being used, and losing this distinction between tag brackets and body text.