Semantic Extraction and Enrichment of Natural Language and Mathematical Discourse for Mathematical Search

Ștefan Anca

KWARC - Knowledge Adaptation and Reasoning for Content
Jacobs University, Bremen, Germany

September 9, 2009
Context: Scientific Documents with Math Formulas

Corpora of interest:

- **arXMLiv** translated corpus from the Cornell University **arXiv** Library
  - Presentation **MathML** or intermediary XML math (**XMath**) → **non-semantic** formats (**not useful** for search)
  - Over 400000 converted documents
  - **Very high incidence** of mathematical formulas and standard formulations (theorems, definitions, proofs, etc.)
- **Connexions** online platform for publishing user content
  - Content **MathML** → **semantic** format (**useful** for search)
  - Over 12000 documents, 3400 with semantic mathematical formulas
  - **Low incidence** of standard mathematical formulations

Tools employed:

- **LaTeXML** - converts **arXiv** (**arxiv.org**) into **arXMLiv** (**arxmliv.kwarc.info**) (**\LaTeX** math ⇒ Presentation MathML)
- The **LaMaPuN Architecture** - semantic enrichment of **LaTeXML** output (**XMath** + context + user ⇒* **Content MathML**)
**MathWebSearch** - Search for Mathematical Structure

- **http://search.mathweb.org**
- Can only index formulas in semantic representation (Content MathML or OpenMath)
- Stores all the mathematical terms by their structure in a substitution-tree
- Provides instantiation, generalization and unification search
- Sentido editor for math query input
- Indexes the Connexions repository (∼ 85000 terms)
**Applicable Theorem Search - Search for Theorems**

Theorem: If we take \( n \geq 0 \), then we know that
\[
\sum_{i=0}^{n} i = \frac{n \cdot (n+1)}{2}.
\]

Query: \( \sum_{k=0}^{25} k \)

Match: \( \sum_{i=0}^{n} i = \frac{n \cdot (n+1)}{2}, \ i \rightarrow k, \ n \rightarrow 25. \)

Conclusion: we know that \( \sum_{i=0}^{n} i = \frac{n \cdot (n+1)}{2} \).

Hypothesis: we take \( n \geq 0 \)

http://betasearch.mathweb.org

- Search for fixed-structure natural language patterns (idioms) which express theorem relations: if \( H \) then \( C \), let \( H \) then \( C \), etc.
- Use idioms as natural language patterns for semantic information extraction
- Find and extract (index) theorems with mathematical universals
- Use MathWebSearch generalization search on queries with constants to retrieve applicable theorems
- Use natural language context to infer semantics about mathematical content:
  For all \( x \), there exists a \( y \), such that \( 4^x = 2^y \)
Knowledge Adaptation and Reasoning for Content

Group Details

- Homepage: [http://kwarc.info](http://kwarc.info)
- Based at Jacobs University, Bremen, Germany
- Led by Prof. Dr. Michael Kohlhase
- Main Research Focus: knowledge representation with a view towards applications in knowledge management, especially for documents with mathematical content

Projects involving:

- Representing documents with mathematical content through semantic mark-up (OMDoc [http://omdoc.org](http://omdoc.org), sTeX), browsing and annotating (SWiM, panta rhei, CPoint)
- Management of change for structured documents (locutor, TNTBase, CCWord)
- Semantic extraction from XML documents (Krextor, Idiom Spotter)
- Semantic enrichment of mathematical terms in XML documents (LaMaPUn)
- Processing, validating and rendering OMDoc documents (JOMDoc, MMT)
- Integrating web services into interactive mathematical documents (JOBAD)
- Converting the arXiv database of LaTeX documents to an XML format (arXMLiv)
- Semantic search on XML documents with mathematical content (MathWebSearch, MaTeSearch, Applicable Theorem Search)
### Idiom Spotter - Semantic Extraction from Informal Math

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Connexions</th>
<th>Saarbrücken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total files</td>
<td>11712</td>
<td>10239</td>
</tr>
<tr>
<td>Files with idioms</td>
<td>451</td>
<td>9947</td>
</tr>
<tr>
<td>Idioms found</td>
<td>1794</td>
<td>215044</td>
</tr>
<tr>
<td>Avg idioms per file</td>
<td>0.15</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Freq. Cnx</th>
<th>Freq. Saarb</th>
</tr>
</thead>
<tbody>
<tr>
<td>assume H1 then C1</td>
<td>29</td>
<td>1755</td>
</tr>
<tr>
<td>conclude D1 is D2</td>
<td>22</td>
<td>3176</td>
</tr>
<tr>
<td>define D1 to be D2</td>
<td>58</td>
<td>4911</td>
</tr>
<tr>
<td>given H1 then C1</td>
<td>43</td>
<td>1809</td>
</tr>
<tr>
<td>H1 if and only if C1</td>
<td>56</td>
<td>25979</td>
</tr>
<tr>
<td>H1 implies C1</td>
<td>170</td>
<td>30593</td>
</tr>
<tr>
<td>C1 only if H1</td>
<td>102</td>
<td>27964</td>
</tr>
<tr>
<td>C1 only when H1</td>
<td>35</td>
<td>1553</td>
</tr>
<tr>
<td>if H1 then C1</td>
<td>1195</td>
<td>108633</td>
</tr>
<tr>
<td>let H1 then C1</td>
<td>61</td>
<td>6915</td>
</tr>
<tr>
<td>suppose H1 then C1</td>
<td>23</td>
<td>1756</td>
</tr>
<tr>
<td><strong>Theorem patterns</strong></td>
<td><strong>1714</strong></td>
<td><strong>206957</strong></td>
</tr>
</tbody>
</table>

- **Idiom**: natural language formulation which follows a certain fixed word and syntax pattern
- Extract semantic relations from scientific texts → **structured knowledge**
- **Connexions corpus** - user-authored online content
- **Saarbrücken corpus** - selection of math publications from ARXMLIV
- All idioms extracted contain at least a mathematical term
- **Saarbrücken corpus** obviously better for Theorem extraction
- No correct content representation of math in Saarbrücken corpus!
The **LaMaPUn Architecture**: A project pursuing semantic enrichment, ambiguity resolution of mathematics in the **arXiv** corpus.

- Semantically enrich the XML math outputted by an initial stage of **LaTeXML**, to reach content-level semantics $\Rightarrow$ **MathML**, **OpenMath**
- **Preprocessing**: correct the math-related human encoding mistakes (e.g. “$1^{\text{st}}$”, “$\{\textbf{x} - \{\textbf{y}\}$”, $\text{last}(x) \rightarrow l \cdot a \cdot s \cdot t(x)$)
- **Semantic Blackboard**: represent the XML documents in an RDF Database
- **Semantic Analysis Modules**: plug into the Blackboard and perform semantic processing, results stored as stand-off annotations
  - Mathematical Formula Disambiguation
  - Content-Based Formula Disambiguation
- **Semantic Result** and **Output Generation**: merge annotations with original documents to obtain the semantically enriched result, outputted as XHTML or **OMDoc** with Content/Presentation **MathML**.
- **Visualization and Feedback**: allow users/authors to review/correct inferred semantics
LaMaPUn Architecture

Preprocessing
- arXMLiv
  - .noparse.xml
  - Semantic Purification (D.Ginev)

Semantic Blackboard
- RDF Representation
  - MFD (C. Jucovschi)
  - CEFU (M. Grigore)
  - Result: Parallel RDF Annotations
  - RDF DB

Semantic Result
- Stand-off annotations
  - .tex.xml
  - LaTeXML semantic defaulting of XMath formula structure

Visualization and Feedback
- Stand-off annotations
  - JavaScript Annotation Framework
  - .xhtml
  - JavaScript Annotation Framework (C. David)
  - Applicable Theorem Search (S.Anca)

Generate OMDoc/XHTML
- Stand-off annotations
  - .pre.omdoc
  - .omdoc
  - .xhtml
  - Semantic Aggregator (D.Ginev)
  - XML\textsc{-}post\_pmml\_omml -keep\_math
  - LaTeXML-omdoc.xsl
  - XSLT
  - LaTeXML\_xhtml\_xsl