# Designing a Semantic Ground Truth for Mathematical Formulas

Alan P. Sexton, Volker Sorge

School of Computer Science
University of Birmingham

Masakazu Suzuki

Department of Mathematics
Kyushu University

# Motivation: Ground Truth Sets

- Ground Truth Sets are manually annotated or validated sets of training data
- Important tools for many recognition tasks
- In Document Analysis, ground truth data is crucial for the design, training and testing of algorithms
- Ground truth for OCR consists of images of single characters together with their correct syntactic interpretation
- Bespoke ground truth sets are developed for particular domains
- Laborious manual task as there is only limited automation available

# Motivation: GTS for Maths

- Currently only one (collection of) ground truth sets for mathematics
- Designed in the context of the Infty project
- Over 680,000 characters from 30 different articles
- Statistical information about the relative occurrence of and relationships between neighbouring characters
- Characters appear many times in database, there is a large amount of information that can be mined from the database

# Motivation: Semantic Ground Truth

- To fully recognise a mathematical expression, capturing its correct semantics is crucial
- Proper semantic markup is essential for processing, translation or correct pronunciation by a screen reader

# Motivation: Semantic Ground Truth

- To fully recognise a mathematical expression, capturing its correct semantics is crucial
- Proper semantic markup is essential for processing, translation or correct pronunciation by a screen reader

Example: Is $\binom{n}{k}$ a binomial coefficient ($n$ choose $k$) or a vector?

- Impossible to answer without context.
- Semantic ground truth could base a decision on well-founded scientific data

# Constructing a Semantic Ground Truth Set

- Combine syntactic information like font information spacing and relative baseline positions with semantic information

  Example:

  1 is the character “one”, in 11 point, CMM font, and represents the integer “1”.
- Use a syntactic ground truth set as basis

  Gives the syntactic meaning to each character occurring in a collection of documents
- Add information to every mathematical expression and character or symbol occurring in it

# Semantic Ground Truth

We use a two step approach:

1. semantic ground truth for mathematical characters and symbols depending on their context
   Example: $f(x\,y)$
   $f$ can be a simple variable or represent a function.

# Semantic Ground Truth

We use a two step approach:

1. semantic ground truth for mathematical characters and symbols depending on their context
   Example: $f(x\,y)$
   $f$ can be a simple variable or represent a function.
2. semantic ground truth for entire mathematical expressions and some sub-expressions
   Example:
   $$v^T = (1\;2\;3) \qquad\qquad \pi = (1\;2\;3)$$
   1 is always an integer, but is once contained in a vector $v$ and once in a permutation $\pi$.

# Semantic Annotation of Symbols

- Annotate mathematical symbols occurring in a syntactic ground truth set
- Annotations based on three levels:
  1. Subject area
  2. Usage of a symbol
  3. Definition within a given context.
- Enables description of different levels of granularity

# Symbol Annotation: Subject Area

- One annotation attribute for a symbol's origin in some mathematical field
- Refer to the general mathematical field the document belongs to from which a symbols was extracted
- Use two first digits of the AMS Mathematics Subject Classification of 2000
- Entered globally for all characters in a document

# Symbol Annotation: Symbol Usage

- Symbols often have different meaning depending on
  - mathematical subject area
  - the local context in the document
- Record the exact mathematical usage of each symbol in the formula from which it was extracted
- Is it a function symbol, an operator, a relation etc.?

# Symbol Annotation: Symbol Usage

- Symbols often have different meaning depending on
  - mathematical subject area
  - the local context in the document
- Record the exact mathematical usage of each symbol in the formula from which it was extracted
- Is it a function symbol, an operator, a relation etc.?

Example: Consider the symbol $g$:

$$\begin{aligned} g &\in G \\ g &\in B^A \end{aligned}$$

Two distinct meanings: element a group $G$ or function $A \longrightarrow B$

# Symbol Annotation: Definition

- Most fine-grained semantic annotation
- Associate every symbol as far as possible with a mathematical definition
- Take the context of the document into account!
- Use an existing system, e.g. OpenMath as reference
- Possible problems:
  - Content dictionaries can not be mapped onto the semantics in a paper
  - No content dictionaries available for some subjects

# Expression Ground Truth

- Semantics of symbols, relationships to neighbours not enough
- Semantics of expressions and sub-expressions is important
- Build abstract syntax trees for expressions
- Leafs would be fully annotated characters and symbols
- Inner nodes inherit subject area but need definitions assigned

# Expression Ground Truth

- Semantics of symbols, relationships to neighbours not enough
- Semantics of expressions and sub-expressions is important
- Build abstract syntax trees for expressions
- Leafs would be fully annotated characters and symbols
- Inner nodes inherit subject area but need definitions assigned

Example: (1 2 3) in group theory

- Symbol annotations: open fence, three ordinaries, closed fence
- Three ordinaries in turn have definition annotations as integers
- Entire AST has definition annotation of permutation

# Automation: Basics

- Much of the semantics will have to be assigned manually
- Goal is to automate as much as possible and have a user correct the result
- Automate some annotation via grammars and parsing techniques
- Annotate definitions using word spotting
  - Hangman style approach
  - Assign one definition to a symbol/expression
  - Similar expressions in the rest of the document get the same annotation automatically
  - Check and correct

# Automation: Constructing ASTs

- Semantic Ground Truth is built on top of a syntactic ground truth set
- Syntactic Ground Truth gives us “perfect” knowledge: character, baseline, bounding box, position, etc.
- We have tool that can
  - extract precise data from PDF documents
  - use a grammar approach to reconstruct math expressions
- Use that tool with the syntactic ground truth data to extract ASTs

# Automation: Symbol Usage

- Automatic recognition of symbol usage by symbol and spatial analysis
- Exploit relative distance between symbols in a formula
- Re-engineer basic layout rules from traditional mathematical typesetting
- Similar to the implementation in LaTeX

# Automation: Symbol Usage

- Automatic recognition of symbol usage by symbol and spatial analysis
- Exploit relative distance between symbols in a formula
- Re-engineer basic layout rules from traditional mathematical typesetting
- Similar to the implementation in LaTeX

Example:

$$\begin{array}{rcl} x\mathcal{R}y & \rightarrow & y\mathcal{R}x \\ x \mathrel{\mathcal{R}} y & \rightarrow & y \mathrel{\mathcal{R}} x \end{array}$$

In the second case $\mathcal{R}$ would be considered as a relation symbol

# Automation: Symbol Usage

| | Ord | Op | Bin | Rel | Open | Close | Punct | Inner |
|---|---|---|---|---|---|---|---|---|
| Ord | 0 | 1 | (2) | (3) | 0 | 0 | 0 | (1) |
| Op | 1 | 1 | $*$ | (3) | 0 | 0 | 0 | (1) |
| Bin | (2) | (2) | $*$ | $*$ | (2) | $*$ | $*$ | (2) |
| Rel | (3) | (3) | $*$ | 0 | (3) | 0 | 0 | (3) |
| Open | 0 | 0 | $*$ | 0 | 0 | 0 | 0 | 0 |
| Close | 0 | 1 | (2) | (3) | 0 | 0 | 0 | (1) |
| Punct | (1) | (1) | $*$ | (1) | (1) | (1) | (1) | (1) |
| Inner | (1) | 1 | (2) | (3) | (1) | 0 | (1) | (1) |

# Discussion

Some potential problems:

- Ambiguities in the meaning of mathematical notation can not be resolved by considering a single article of the ground truth set, but will need a background knowledge of the mathematical literature in the field
- Current semantic formalisation in the OpenMath content dictionaries are not sufficient for annotating given data
- OpenMath formalisations are not at the right level to give semantic meaning to "human oriented" mathematics

# Conclusions

- Currently we have the design of a semantic ground truth for mathematics
- It could have major impact on correct recognition and content markup of mathematical formulae
- We even have the funding to build it on top of a syntactic ground truth set
- Main Problem: The Infty ground truth set can not be used due to copyright restrictions