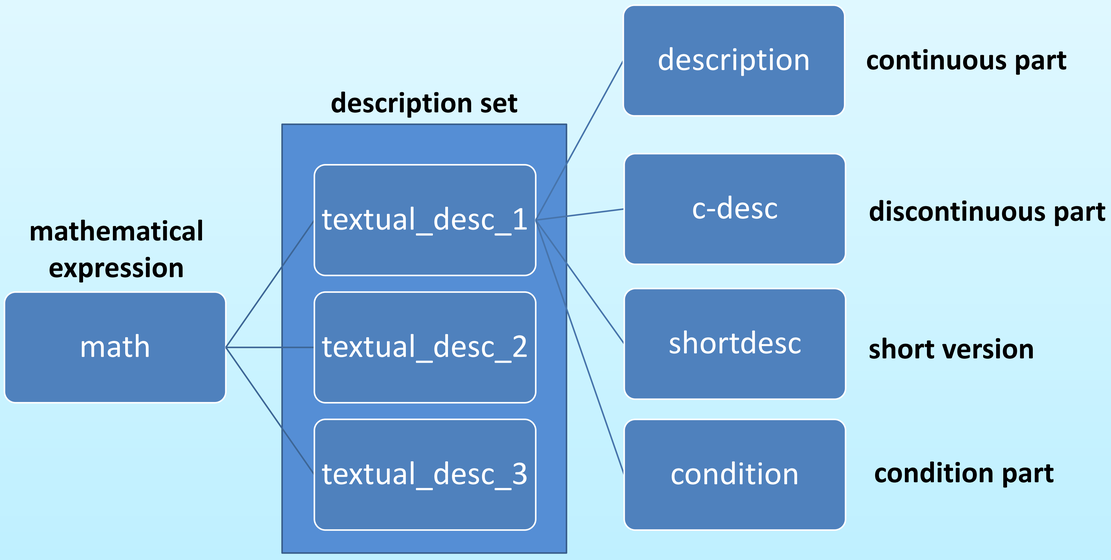
Annotating Scientific Papers for Mathematical Formula Search

Giovanni Yoko Kristianto[1], Goran Topic[2], Minh-Quoc Nghiem[3], Akiko Aizawa[1,2]
[1]Department of Computer Science, The University of Tokyo, Tokyo, Japan
[2] National Institute of Informatics, Tokyo, Japan
[3] Department of Informatics, The Graduate University for Advanced Studies, Tokyo, Japan

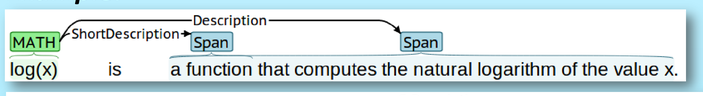
Goal

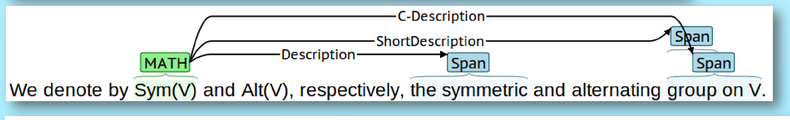
Development of a knowledge base providing relationships between mathematical formulas and corresponding descriptions.

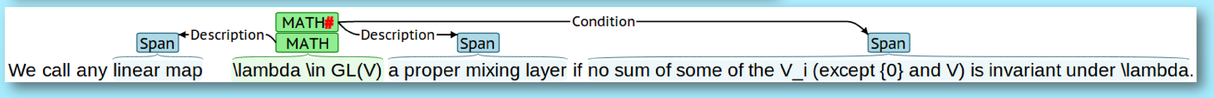
Annotation Design



Example:







annotation tool: *http://brat.nlplab.org/*
cite: *arXiv:0806.4135v1 [math.GR]*

Evaluation

- **strict matching: extracted descriptions must be exactly the same as annotation result**
- **soft matching: accepting extracted descriptions that contain, are contained, or overlap with annotation result**

Automatic Extraction

Methods:

- **pattern matching: consists of seven predefined sentence patterns**
- **CRF:**
  - **noun phrase as description candidate**
  - **syntactic features: sentence patterns, relative position of descriptions toward expressions, POS tags**
  - **lexical features: word unigrams, bigrams, and trigrams**

Method	Strict Matching			Soft Matching		
	Precision	Recall	F1-score	Precision	Recall	F1-score
pattern	25.53	20.84	22.91	55.41	44.80	49.44
CRF	73.60	30.09	42.46	80.08	40.30	53.29

Using the Annotation Result

Semantic Search

- **Input: natural language description**
- **Output: related mathematical formula**

Input	Output	
	Formulas	Descriptions
“”entropy”	$H(T\|L) = -\sum_{t \epsilon T} P(t\|L)\log(t\|L)$	the entropy of tags at a leaf L of the tree $\mathcal{T}$
	$\overline{H}_{\mathcal{T}(T)} = \sum_{L \epsilon \mathcal{T}} P(L)H(T\|L)$	the average entropy of tags in the tree

Semantic Browsing

- **Input: mathematical formula**
- **Output: descriptions of the formula, including explanation of variables and subexpressions**

Conclusion

- **Our annotation design supports the annotation of continuous, discontinuous, and complex descriptions**
- **Annotation results can be used as a training data in automatic description extraction. The subsequently extracted data can be used for semantic searching and semantic browsing of mathematical expressions**