SYMBOL KNOWLEDGE EXTRACTION

From a Simple Graphical Language

Jinpeng LI, Harold MOUCHERE, Christian VIARD-GAUDIN

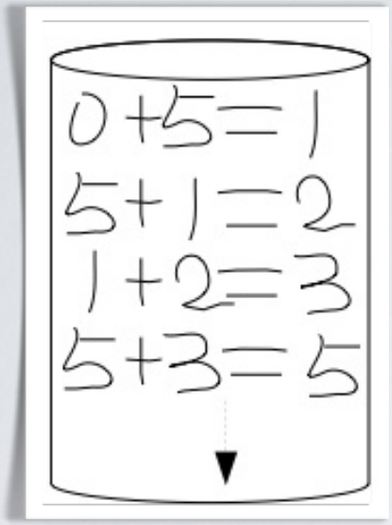
www.projet-depart.org

OUTLINE

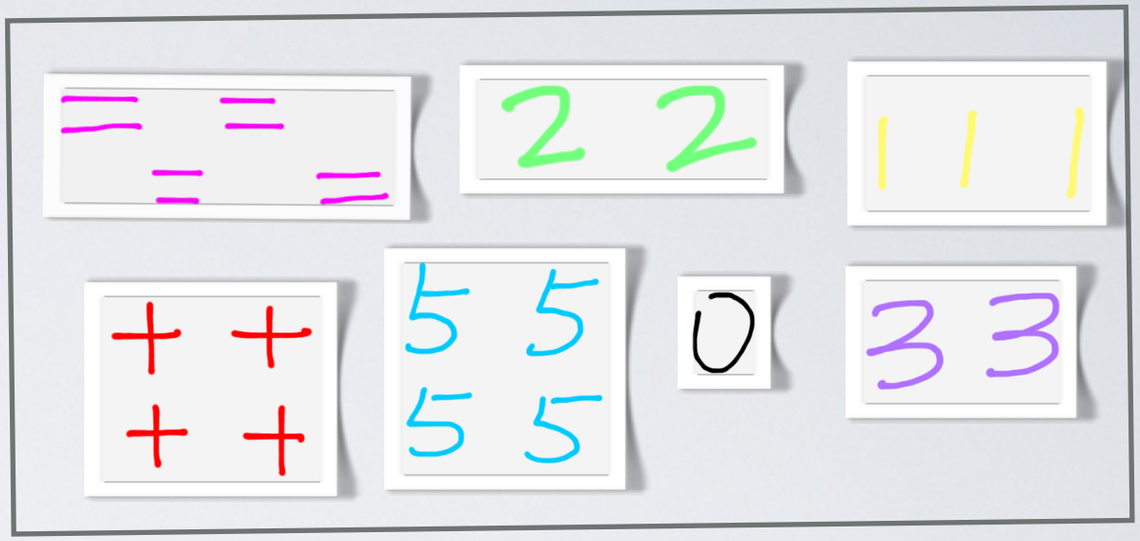
- 1. Background
- 2. Graphical Symbol Knowledge Extraction
  - 2.1.Quantization (Clustering)
  - 2.2.Construction of Relational Graph
  - 2.3.Lexicon Extraction
- 3. Conclusion

BACKGROUND

What is the graphical symbol knowledge extraction?



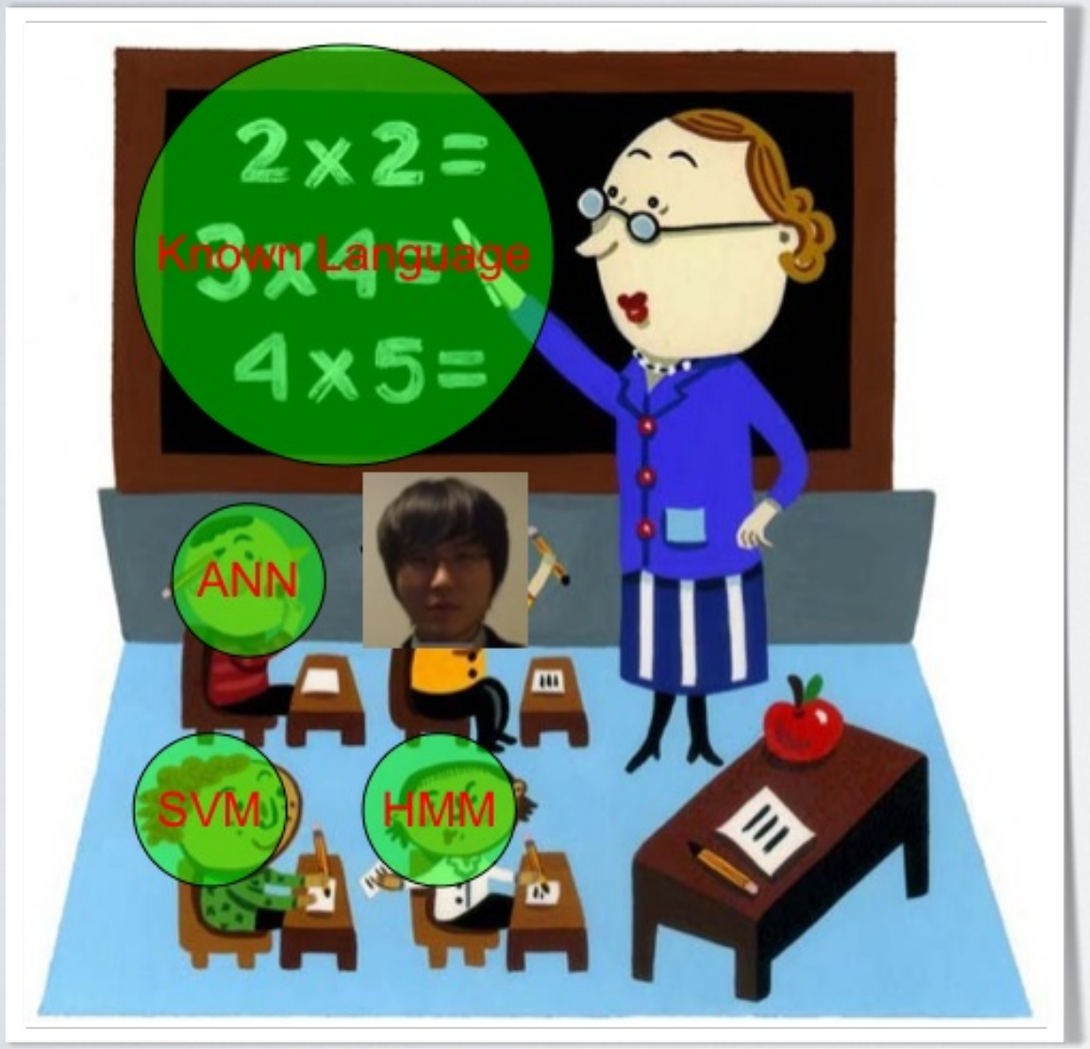
Symbol extraction



Annotation
20 symbols
have to be labelled

Annotation
7 symbols (sets)
have to be labelled

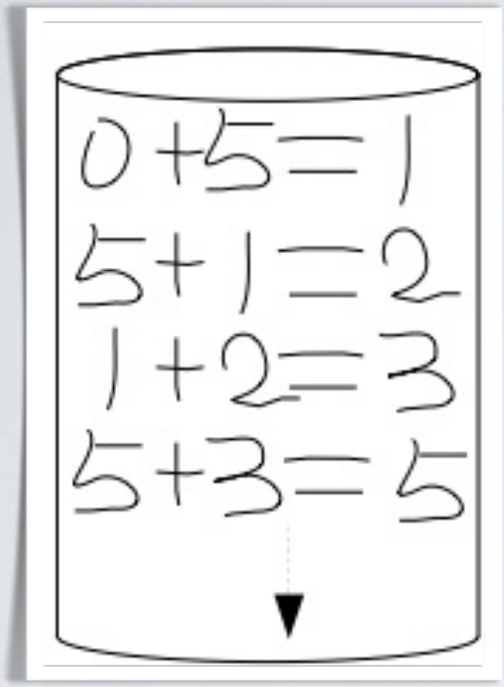
TRADITIONAL GRAPHICAL LANGUAGE RECOGNITION



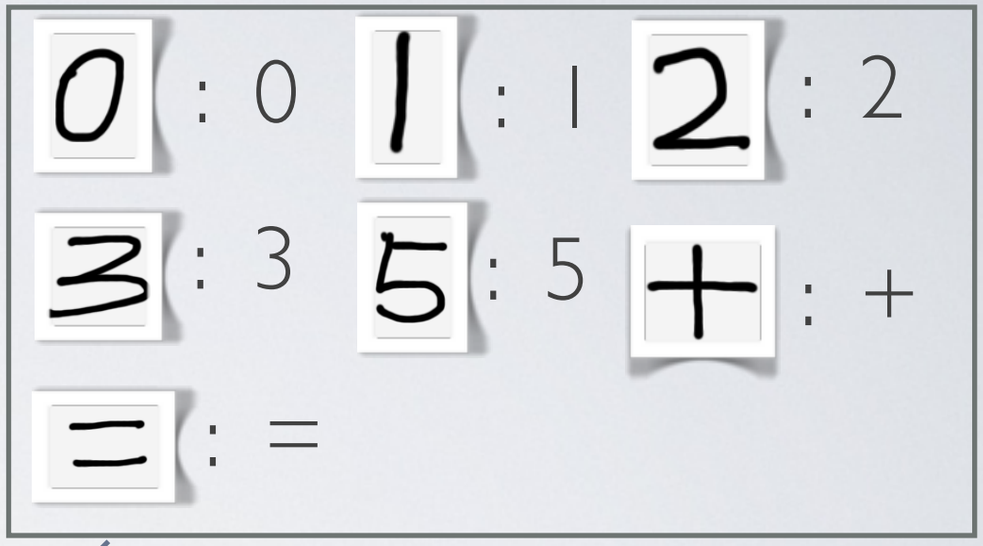
Training

Exams(Tests)

TRADITIONAL GRAPHICAL LANGUAGE RECOGNITION



Known graphical symbols
(defined manually)



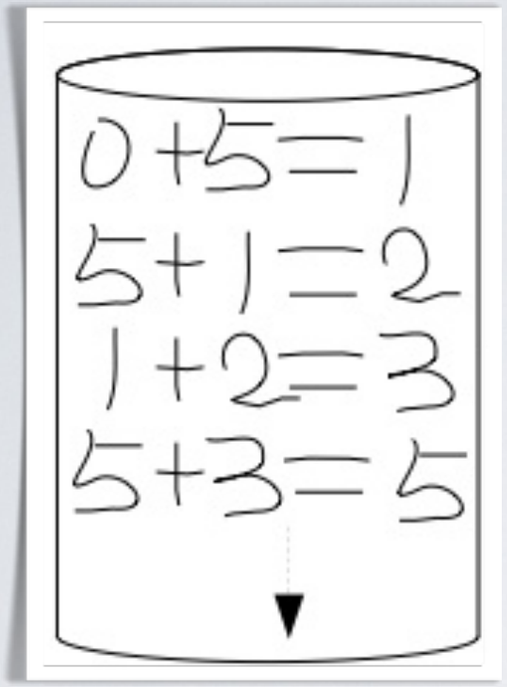
Recognition

Training

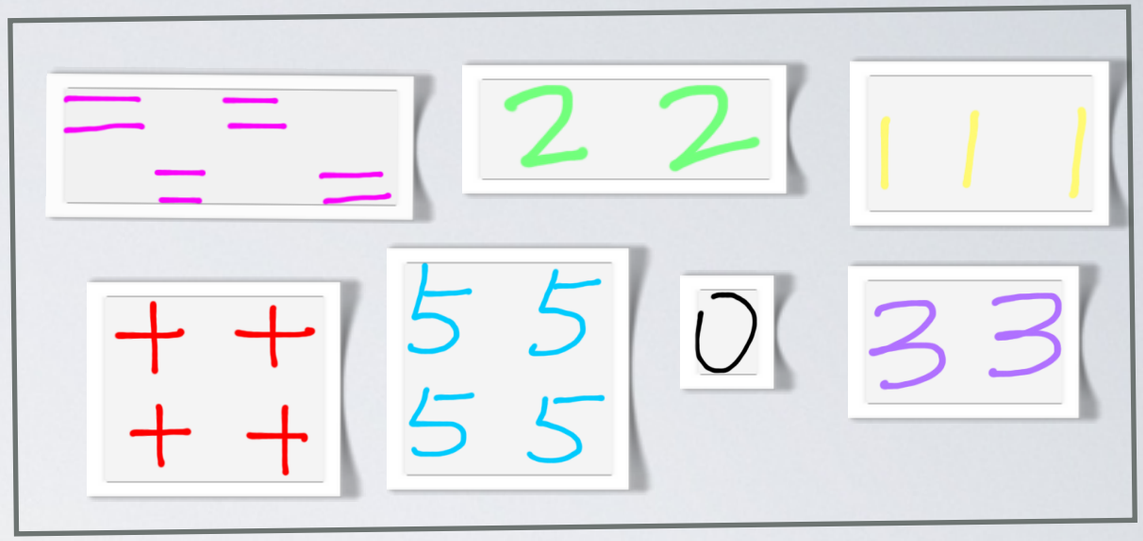
Classifiers
SVM : Support Vector Machine
ANN: Artificial Neural Network
HMM: Hidden Markov Model, etc.

SYMBOL KNOWLEDGE EXTRACTION

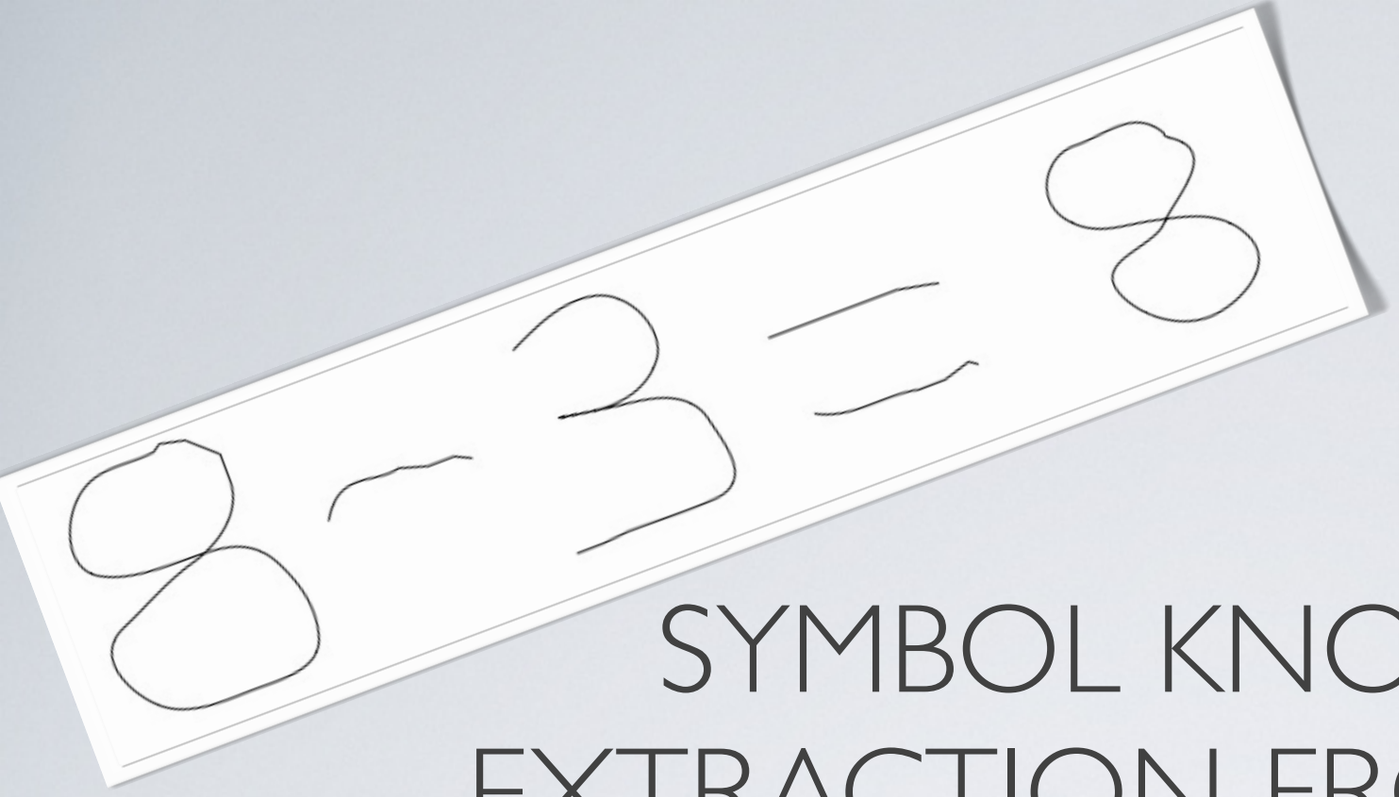
Unknown graphical language



Could we recover or discover these symbols?

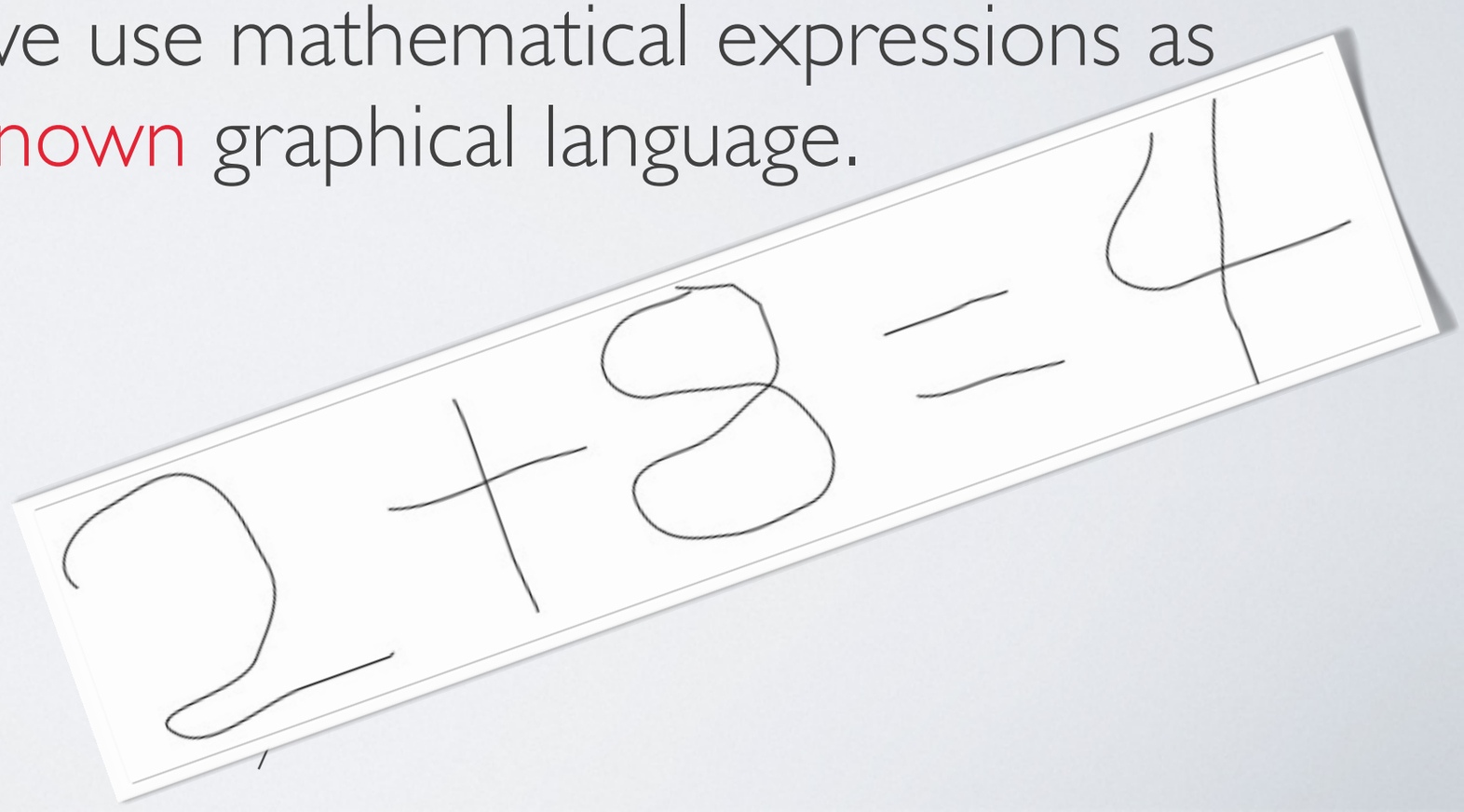


Annotation



SYMBOL KNOWLEDGE EXTRACTION FROM A SIMPLE GRAPHICAL LANGUAGE

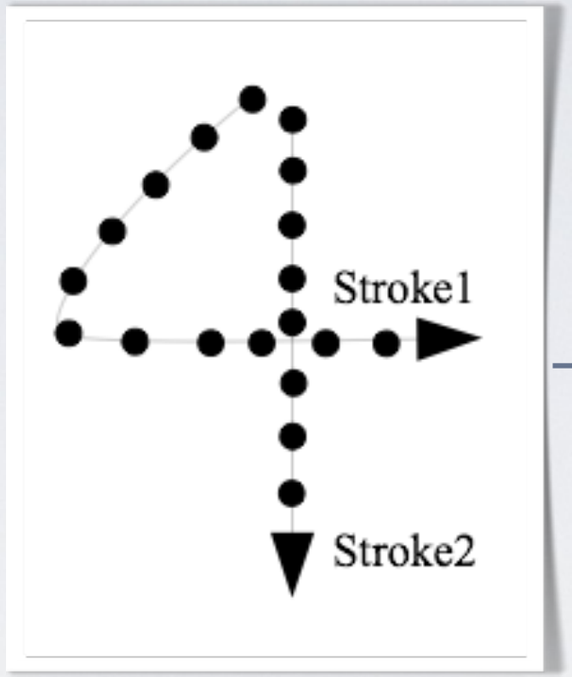
As an example, we use mathematical expressions as an unknown graphical language.



GRAPHICAL LANGUAGE

Online handwritten strokes

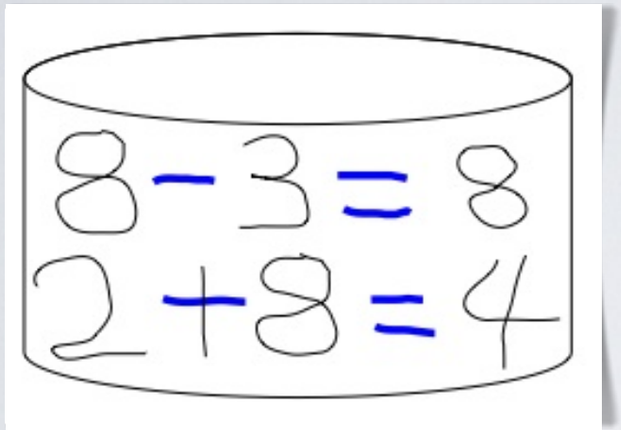
Sampling



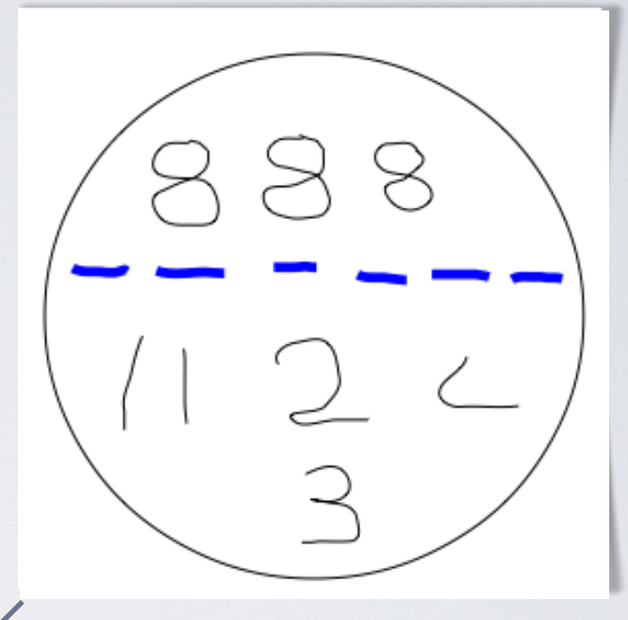
Collected data

Stroke1: $((x_1,y_1),(x_2,y_2),(x_3,y_3),...)$
Stroke2: $((x_1,y_1),(x_2,y_2),(x_3,y_3),...)$

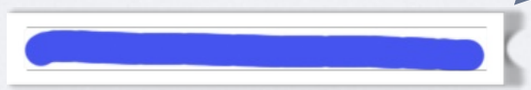
GRAPHICAL SYMBOL KNOWLEDGE EXTRACTION



The base elements are strokes.



One stroke

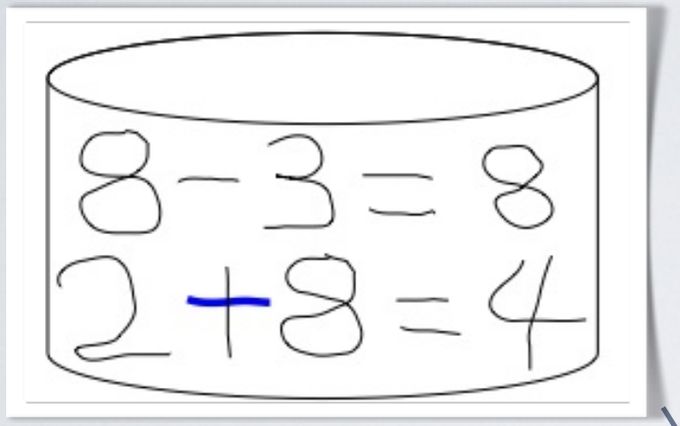


Where is the horizontal stroke from?

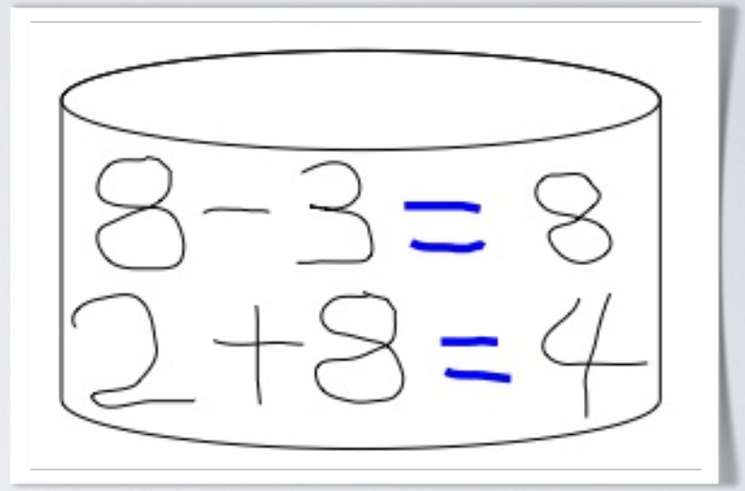
This horizontal stroke repeats six times; it is "frequent".

Grapheme!

GRAPHICAL SYMBOL KNOWLEDGE EXTRACTION



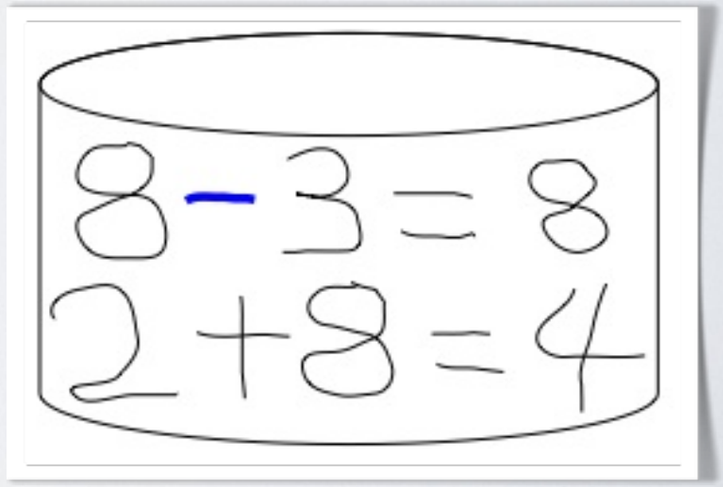
From a part of symbol, "plus"



From two same symbols "equal"

Grapheme!

Where is the horizontal stroke from?

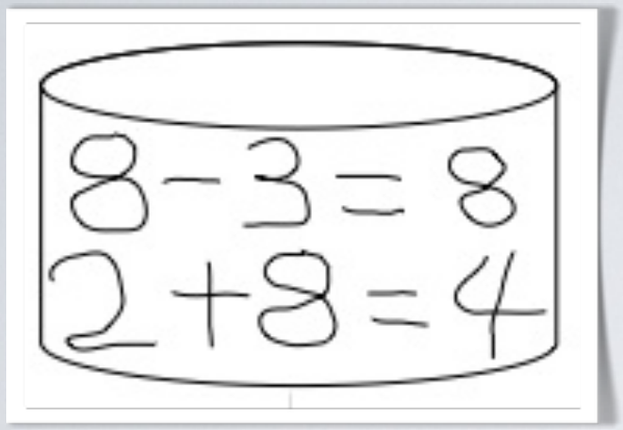


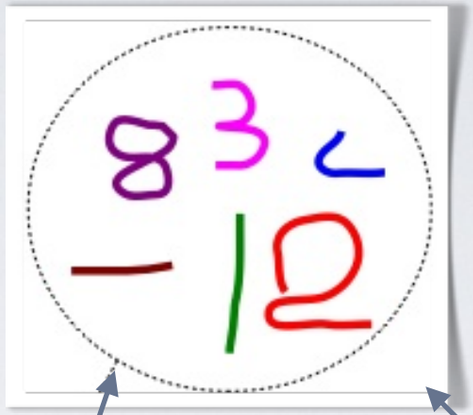
From a symbol, "minus"

OUTLINES

- 1. Background
- 2. Symbol Knowledge Extraction
  - 2.1. Quantization (Clustering)
  - 2.2. Construction of Relational Graph
  - 2.3. Lexicon Extraction
- 3. Conclusion

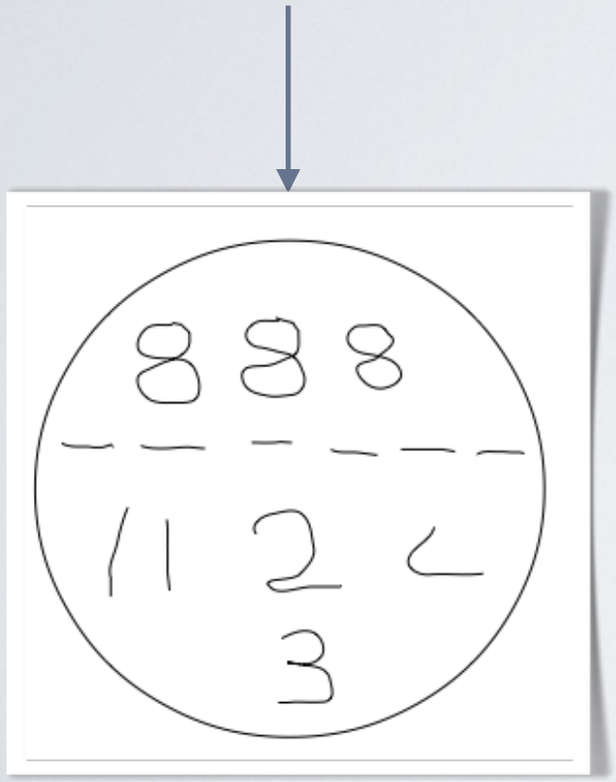
HIERARCHICAL CLUSTERING

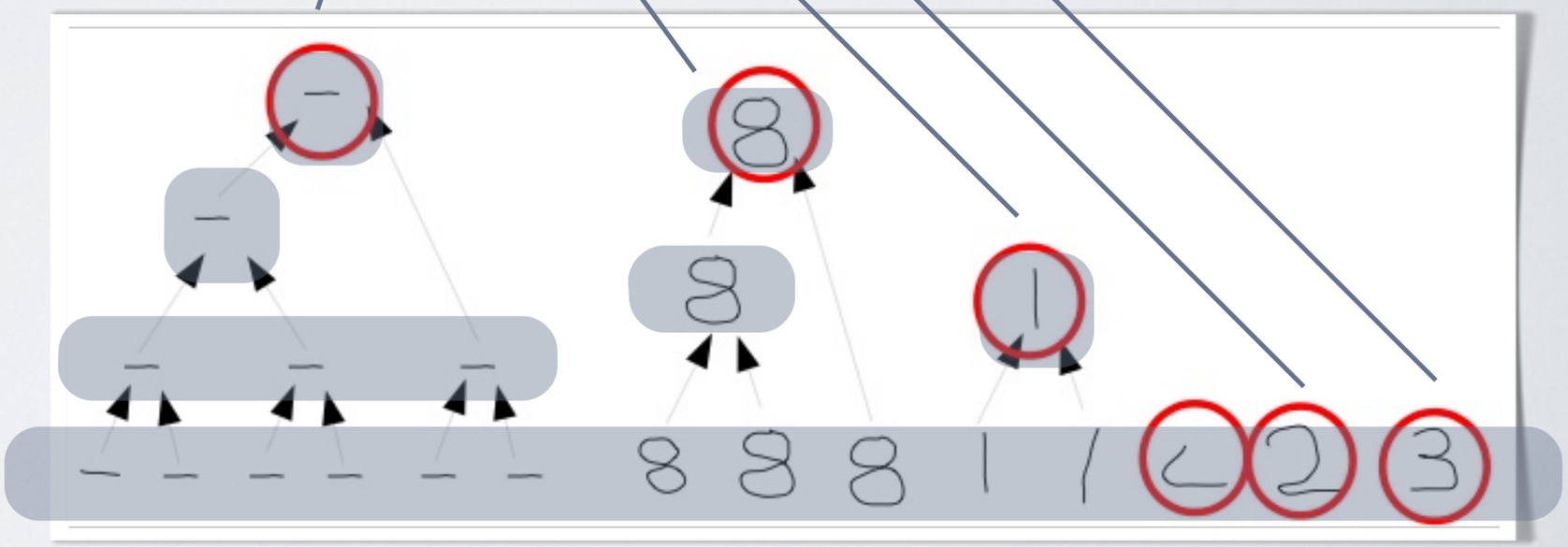




Grapheme

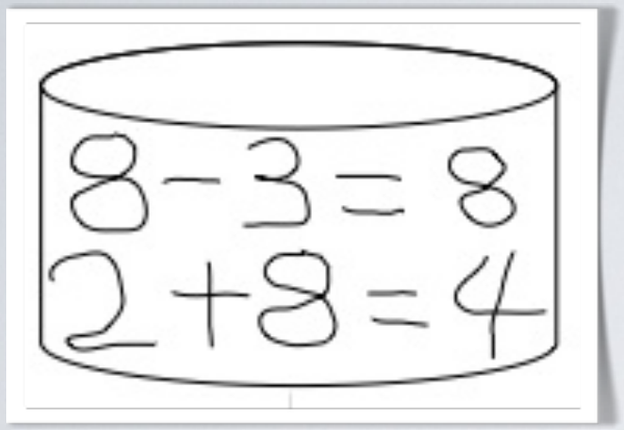
Clustering results:
A set of prototypes

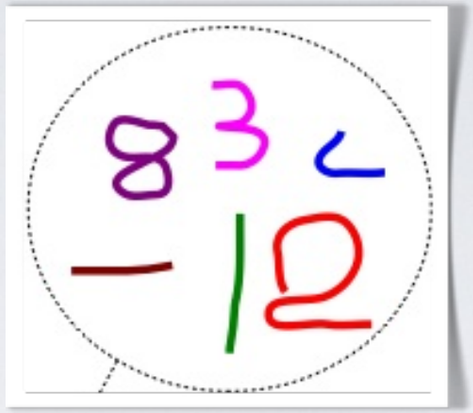




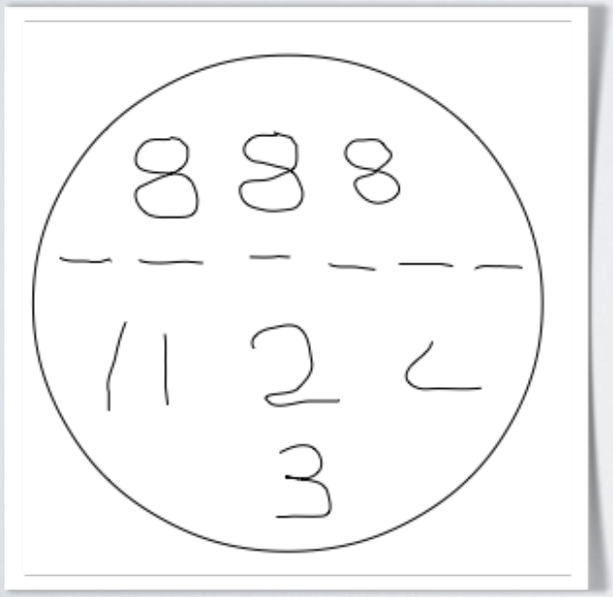
[1] Lance, G. N. & Williams, W.T., A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems, The Computer Journal, 1967, 9, 373-380

QUANTIZATION

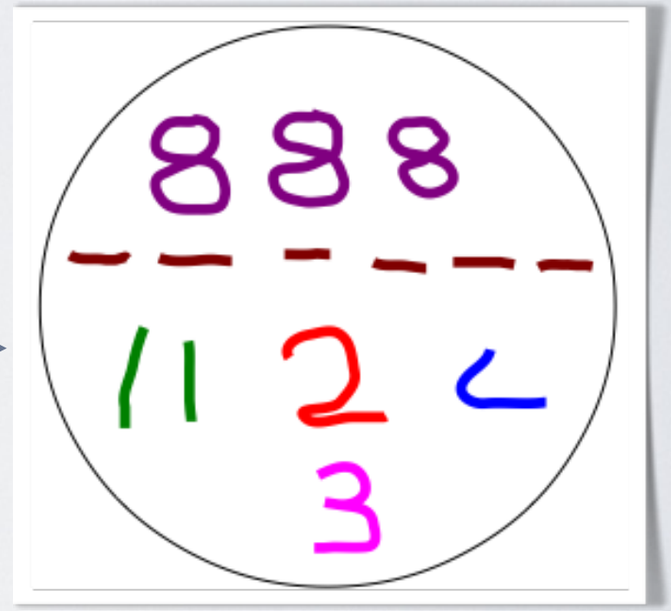




A set of prototypes



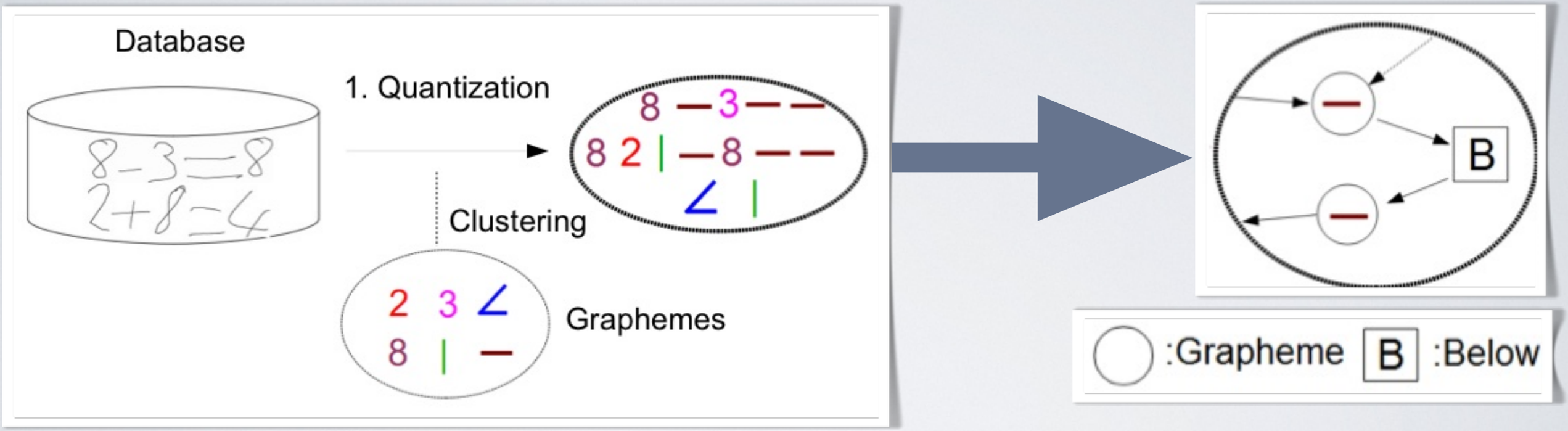
1.Quantization



[1] Lance, G. N. & Williams, W.T., A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems, The Computer Journal, 1967, 9, 373-380

GRAPHICAL SYMBOL DISCOVER

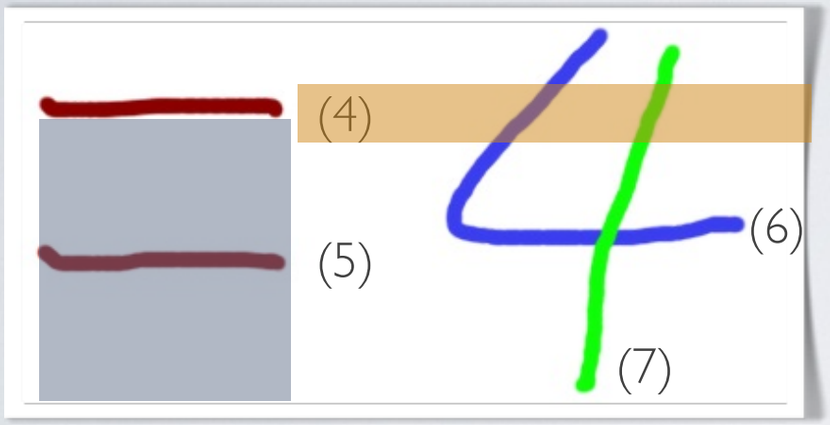
2. Construction of relational graph



SPATIAL RELATIONS

We predefine three spatial relations:

Right , **Below** , and **Intersection**



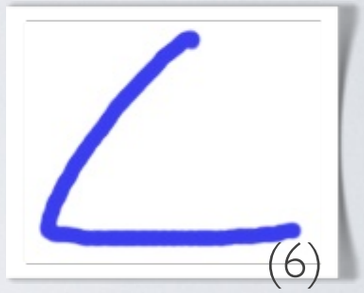
Right:

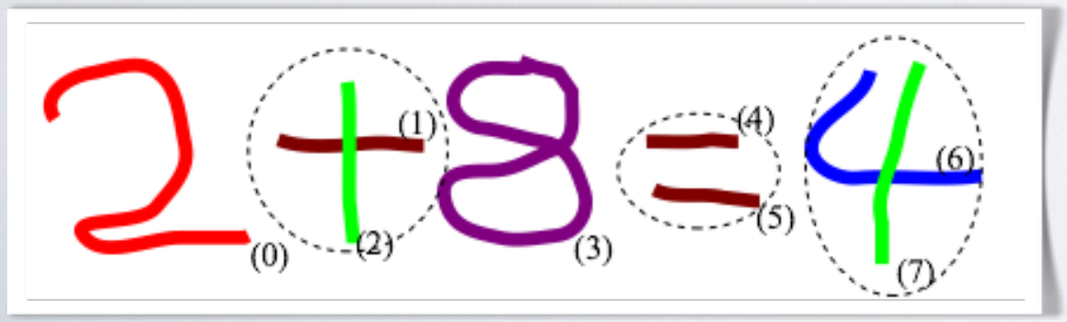
Below:

Reference stroke

R

Closest

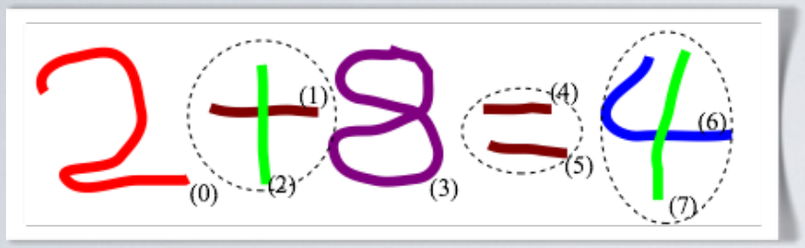




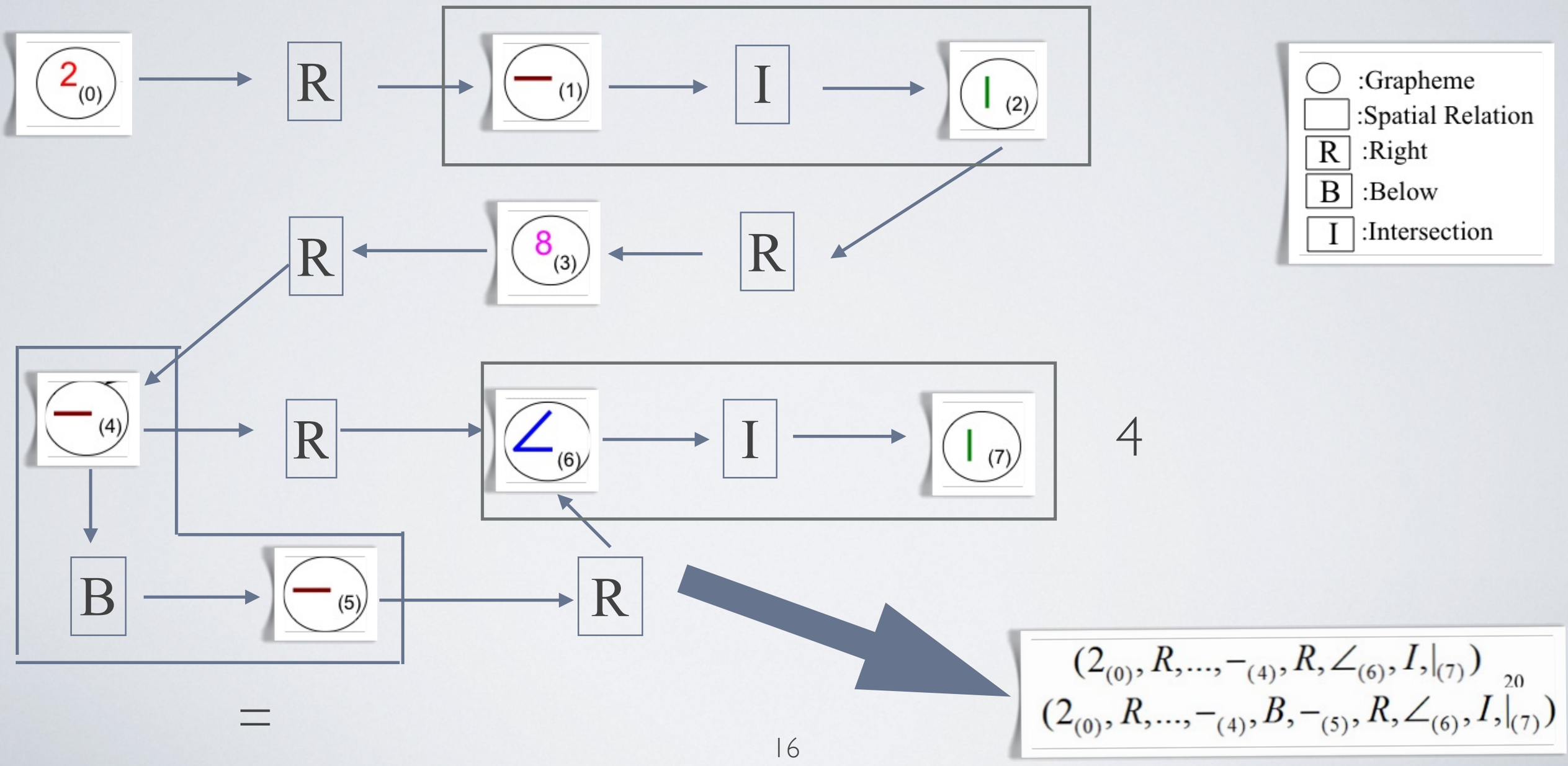
B

RELATIONAL GRAPH

Directed acyclic graph



+

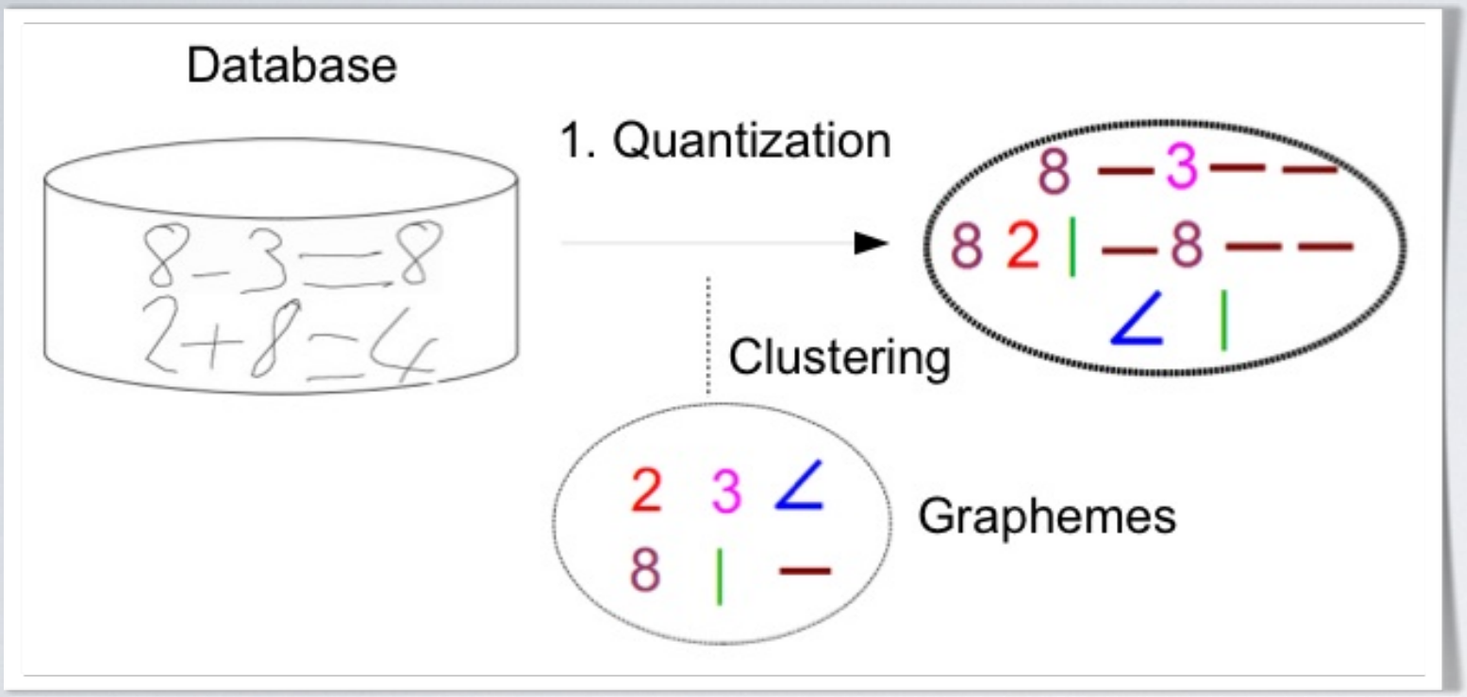


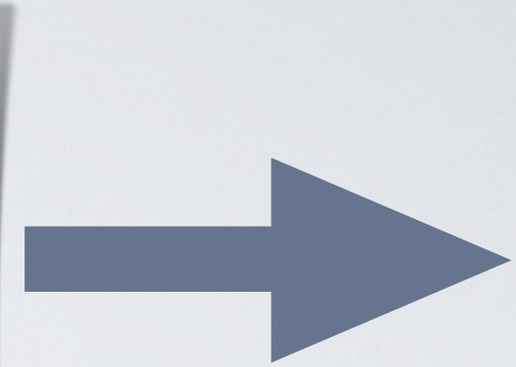
4

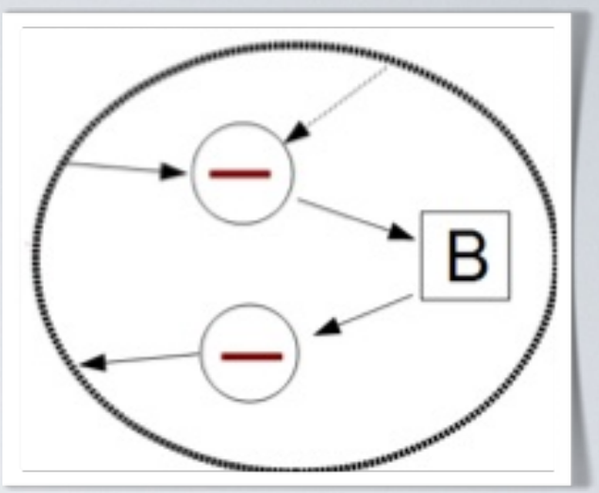
=

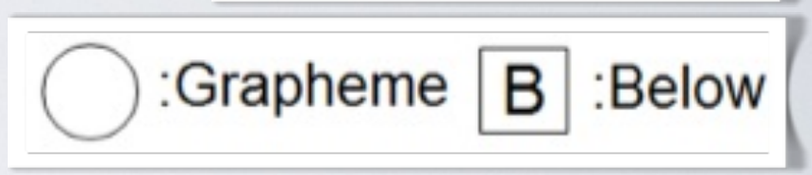
LEXICON EXTRACTION

2. Construction of relational graph



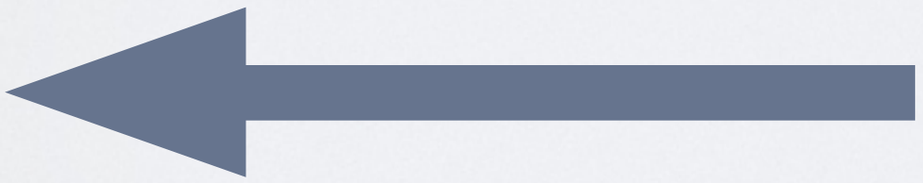




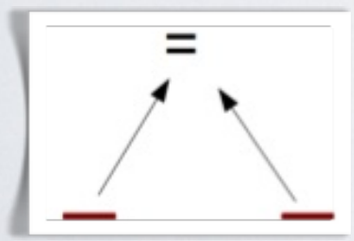


Reduce the description length

$$(2_{(0)}, R, ..., -_{(4)}, R, \angle_{(6)}, I, |_{(7)})_{20}$$
$$(2_{(0)}, R, ..., -_{(4)}, B, -_{(5)}, R, \angle_{(6)}, I, |_{(7)})$$



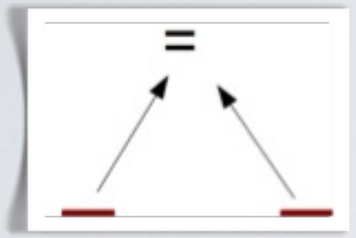
3. Lexicon extraction



MINIMUM DESCRIPTION LENGTH PRINCIPLE

3. Lexicon extraction

Reduce the description length



$$(2_{(0)}, R, \ldots, -_{(4)}, R, \angle_{(6)}, I, |_{(7)})_{20}$$
$$(2_{(0)}, R, \ldots, -_{(4)}, B, -_{(5)}, R, \angle_{(6)}, I, |_{(7)})$$

As a naive example, we try to analyze a sequence, "1234-2/1234".

We define the description length (DL) as the **number of letters**.

DL("1234-2/1234")=11

[2] Marcken, C. D., Linguistic Structure as Composition and Perturbation,
In Meeting of the Association for Computational Linguistics, Morgan Kaufmann Publishers,
1996, 335-341

MINIMUM DESCRIPTION LENGTH PRINCIPLE

As a naive example, we try to analyze a sequence, "1234-2/1234".

We define the description length (DL) as the **number of letters**.

DL("1234-2/1234")=11

If we replace "12" as **S**, DL("**S**34-2/**S**34")+DL("12")=11.

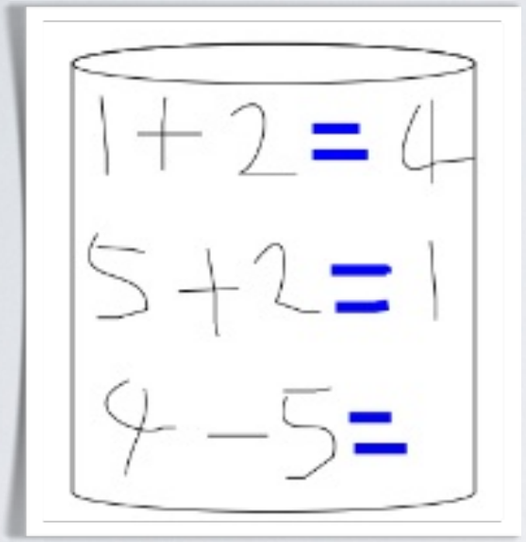
If we replace "123" as **S**, DL("**S**4-2/**S**4")+DL("123")=10.

If we replace "1234" as **S**, DL("**S**-2/**S**")+DL("1234")=9.

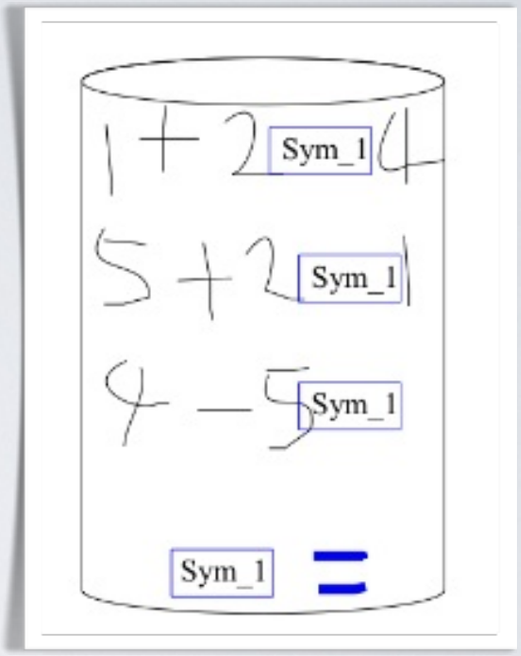
Best lexical unit

[2] Marcken, C. D., Linguistic Structure as Composition and Perturbation, In Meeting of the Association for Computational Linguistics, Morgan Kaufmann Publishers, 1996, 335-341

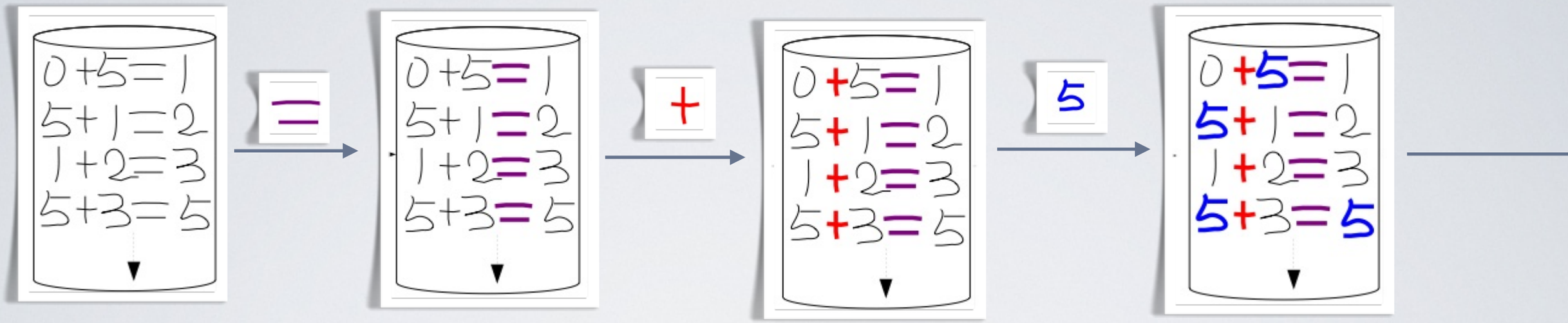
MINIMUM DESCRIPTION LENGTH PRINCIPLE



Replace frequent patterns
in order to compress data



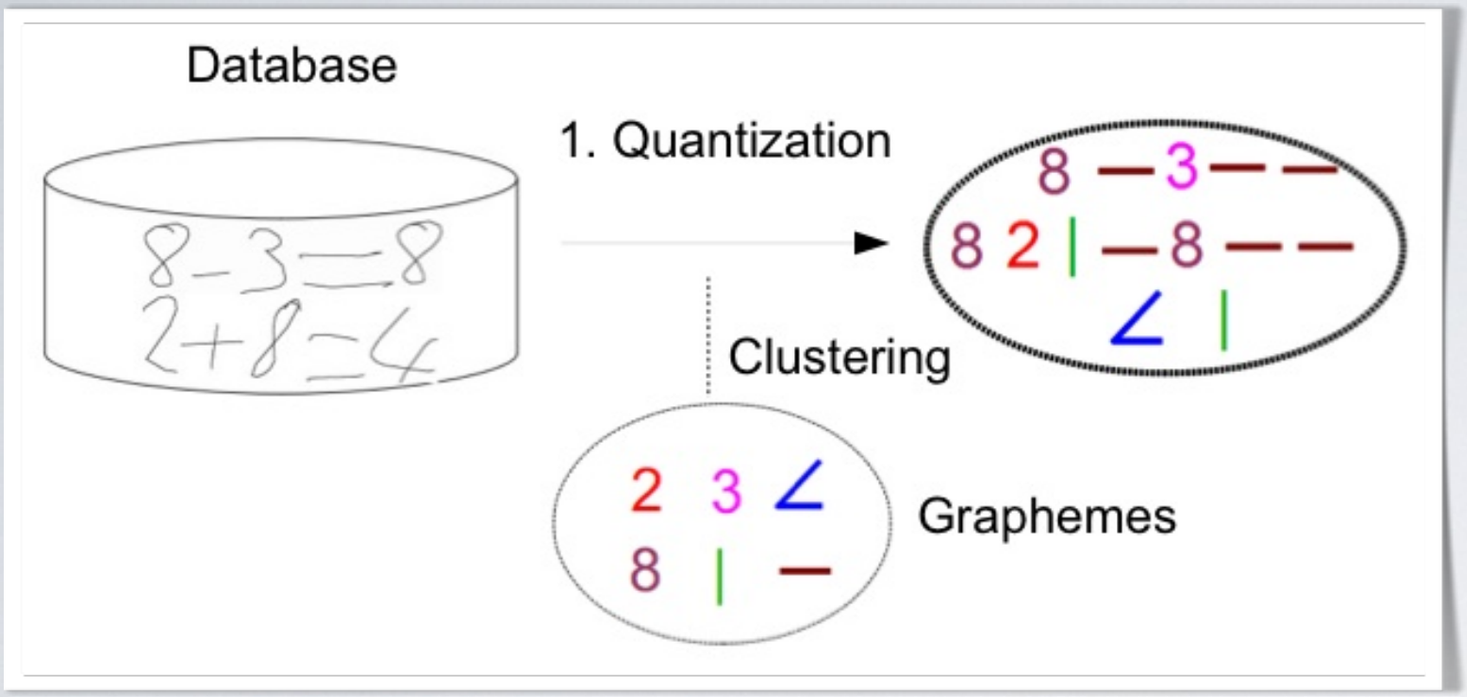
DISCOVER WORDS ITERATIVELY

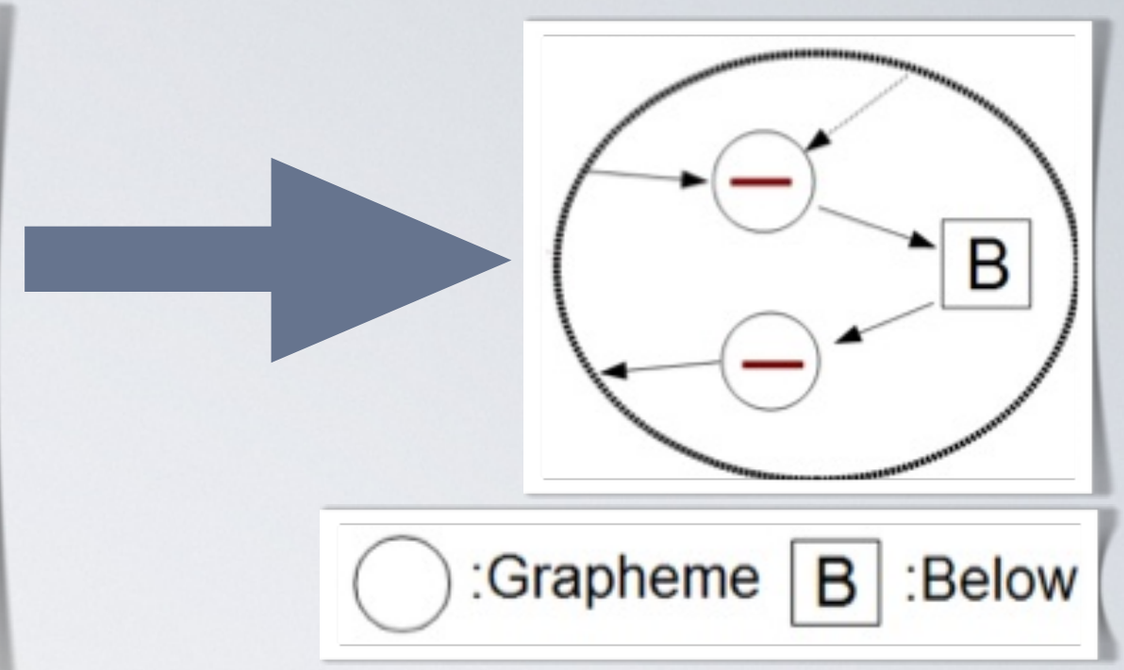


Lexicon: 

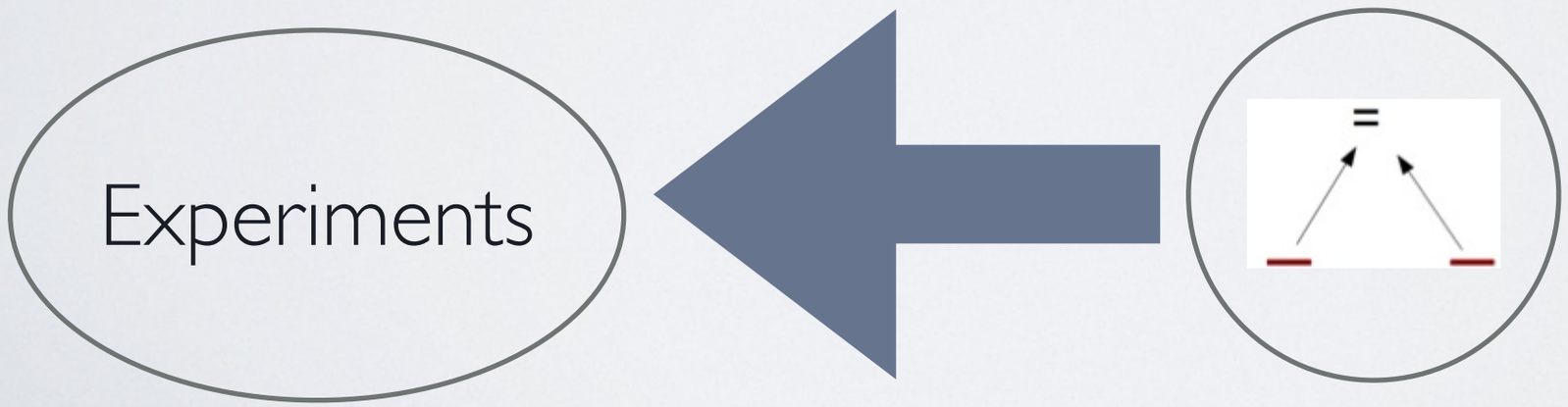
LEXICON EXTRACTION

2. Construction of relational graph





3. Lexicon extraction

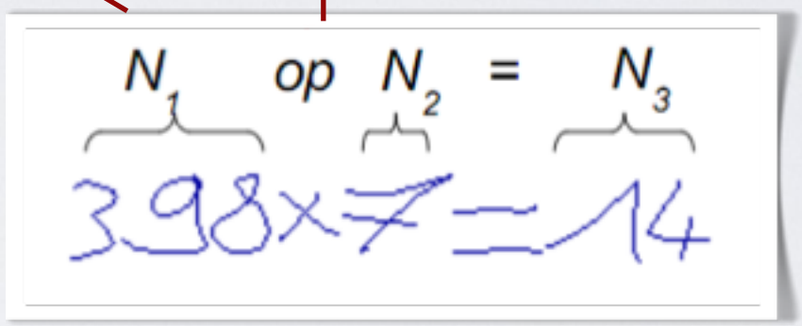


SYNTHETIC DATABASE FROM REAL HANDWRITTEN ISOLATED CHARACTERS

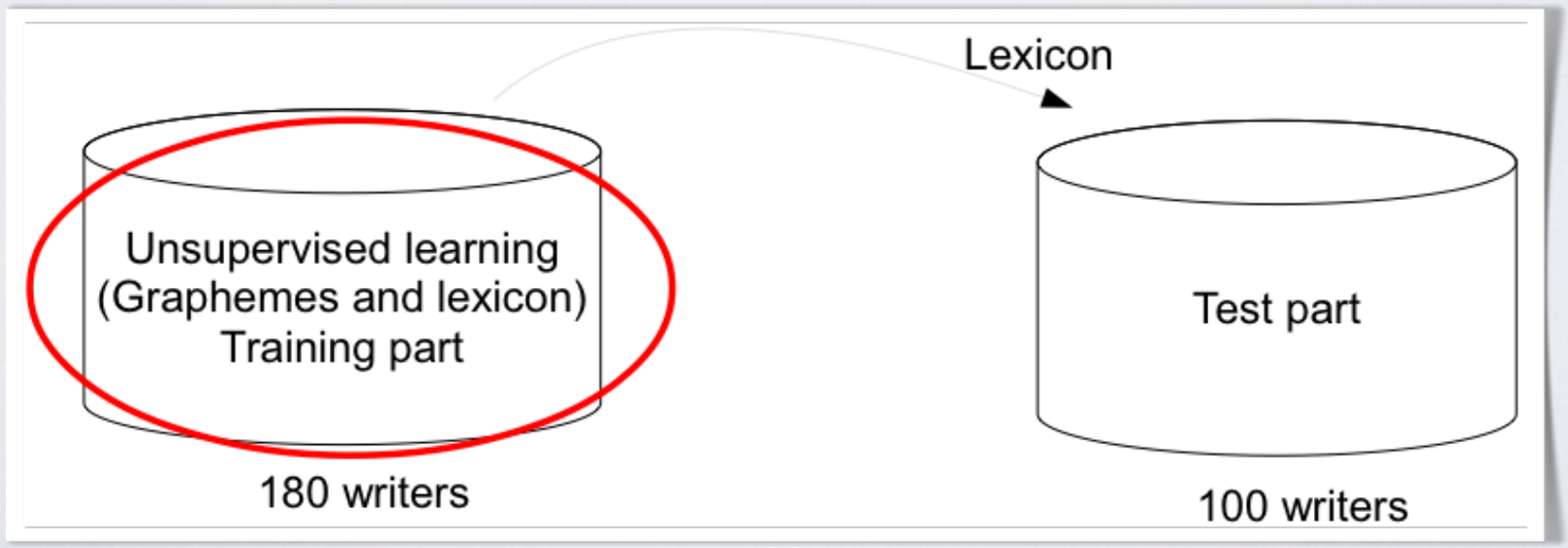
$N_{i=\{1,2,3\}}$ is 70% of 1 digit, 20% of 2 digits and 10% of 3 digits randomly.

$\{0, 1, ..., 9\}$

$\{+, -, \times, \div\}$

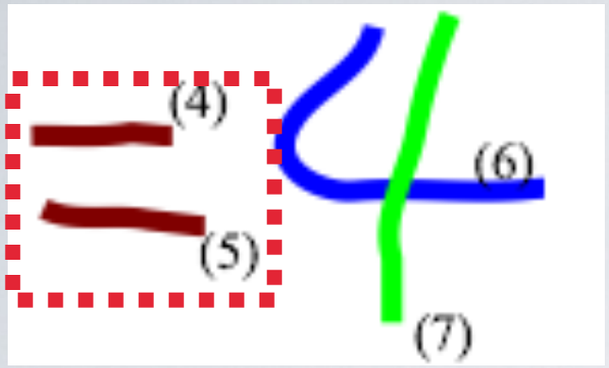


5427 symbols



3035 symbols

RECALL RATE (EXPERIMENTS)



$$R_{\text{Recall}} = \frac{|S(e,G) \cap S(e,L)|}{|S(e,G)|} = 0.5$$

S(e, G):ground-truth for the expression.

$$S(e,G) = \{\{-_{(4)}, -_{(5)}\}, \{\angle_{(6)}, |_{(7)}\}\}$$

S(e, L):hierarchical segmentation using lexicon ***L***.

$$S(e,L) = \{\{-_{(4)}\}, \{-_{(5)}\}, \{-_{(4)}, -_{(5)}\}, \{\angle_{(6)}\}, \{|_{(7)}\}\}$$

We got the recall rate of 74%(2245 symbols) on the test part of our database.

CONCLUSION

- Extraction of graphemes and quantization
- Construction of relational graph
- Lexicon extraction using minimum description length principle
- The recall rate of 74% (2245 symbols) is obtained.

FUTURE WORK

- Reduce the description length on relational graphs instead of sequences [3].
- Unsupervised spatial relation learning for complex spatial relations.

[3]Jinpeng Li, Harold Mouchère and Christian Viard-Gaudin. Unsupervised Handwritten Graphical Symbol Learning Using Minimum Description Length Principle on Relational Graph, International Conference on Knowledge Discovery and Information Retrieval, KDIR 2011, Paris, France.

THANK YOU FOR YOUR ATTENTION

Questions?

Presentation can be downloaded from **LiJinpeng.org**