Artificial Intelligence

Project 2 :
Language Learning Using Hypernetworks

2009. 5. 4.

Ko, Younggil (ykko@bi.snu.ac.kr)
Kim, Seungyeon (sykim@bi.snu.ac.kr)

Biointelligence laboratory

Contents

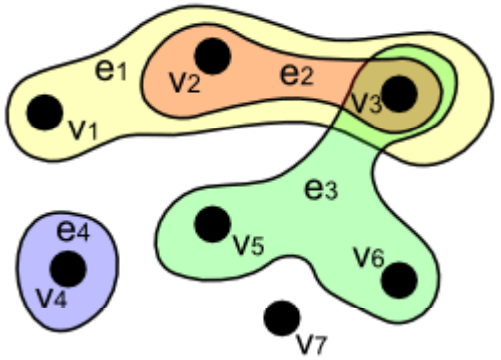
- Outline
- Language Learning using Hypernetworks
- Data set for Language Learning
- Tool for Hypernetworks
- Guide to Writing Reports
  - Style, mandatory contents, optional contents
- Submission guide / Marking scheme
- Demo on the Tool

Outline

- Goal
  - Understand Hypernetworks & machine learning deeper
  - Practice research and technical writing
- Language Learning (sentence completion)
  - The problem is to predict each word in a sentence based on surrounding words
  - Consider each word as an attribute and train a Hypernetwork with sentences
- Data Set
  - Sentences from ‘Friends’ and BBC documentary

Brief Introduction to Hypernetworks

- Hypernetworks
  - Representation and learning method based on weighted hypergraph
    - http://en.wikipedia.org/wiki/Hypergraph
  - Generate hyperedges with sampling and manage a library of weighted hyperedges
  - Learning strategy
    - (explained in pp. 5~8)



- Reference: [IEEE-CIM, 2008] B.-T. Zhang, Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3):49-63, 2008.

Language Learning with Hypernetworks (1/8)

- Sentence Completion
  - We want to complete a sentence which has some missing words.

- How to complete the sentence?
  - Motivation: predict the blank based on the pattern of word co-occurrences in some specific corpus of sentences

Language Learning with Hypernetworks (2/8)

- Goal : Sentence Completion
  - Train a hypernetwork to be able to recall any sentence in the given corpus
  - To complete sentences which contain missing words with the trained hypernetwork

  - Target function to be maximized for the sentence completion problem
    - It is the accuracy in basic
    - f(x) := (number of correct answers) / (number of whole tests)

    - **A test** : try to guess a word in a sentence with a trained hypernetwork, and compare the result with the original sentence in dataset.
      - **How? (see the following page)**

Language Learning with Hypernetworks (3/8)

- Learning steps
  - Sampling step
    - Randomly choose (n) words in each sentence while preserving the order of words.
    - Repeat (m) times to get samples per sentence
    - {(1,2,3,4,5,6),(n, m=3)} :: (1,3,4), (2,3,5), (4,5,6)
  - Weight update step
    - Guess each word with the current Hypernetwork.

Language Learning with Hypernetworks (4/8)

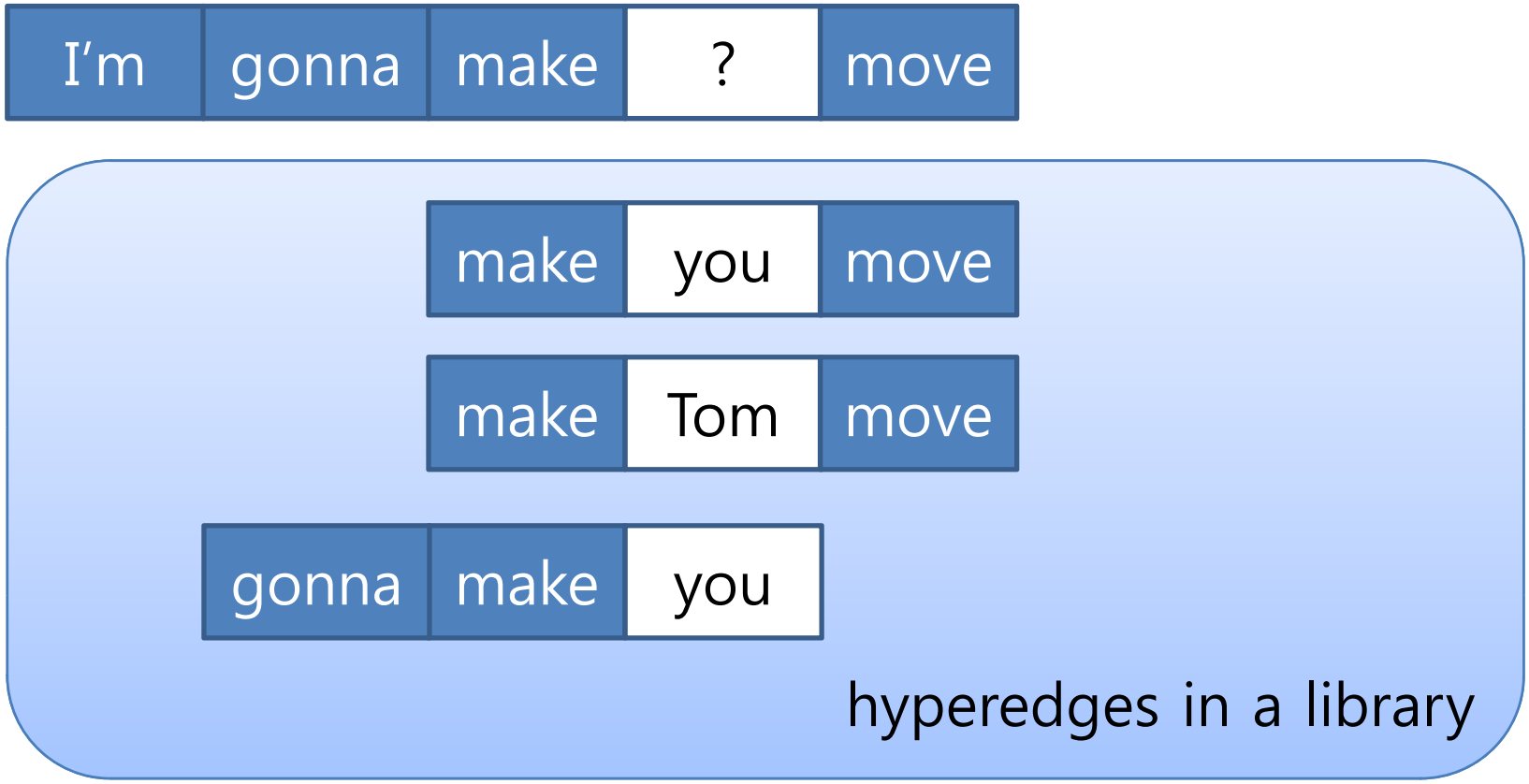
- Learning steps
  - Weight update step(con't)
    - Weight update
      - Test every word with current Hypernet
      - If correct : do nothing
      - If incorrect
        - » Add a constant score (weight update rate) for every **related** hyperedges.
        - » If some hyperedges are not exist in the HN, regard their score as 0

Language Learning with Hypernetworks (5/8)

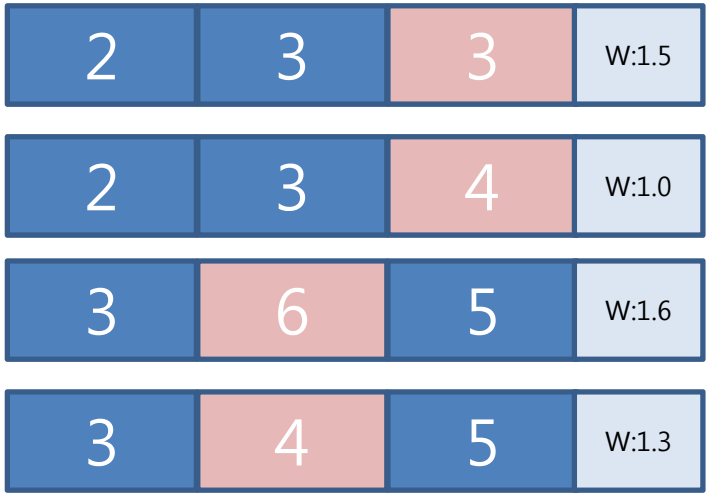
- How to complete the sentence?
  - Hypernetworks : set of hyperedges with weights
  - Assume HN={ (1,2,3 : 3), (2,3,4 : 1), (1,2,4 : 5) }
  - Guess (1,2,?,4) : see following pages
    - (1,2,3 : 3), (2,3,4 : 1) : score(3) = 4
    - (1,2,4 : 5) : score(4) = 5
    - We assume unknown word as '4' with highest score, 5

Language Learning with Hypernetworks (6/8)

- Sentence Completion



Language Learning with Hypernetworks (7/8)



3 : 1.5
4 : 1.0 + 1.3
6 : 1.6

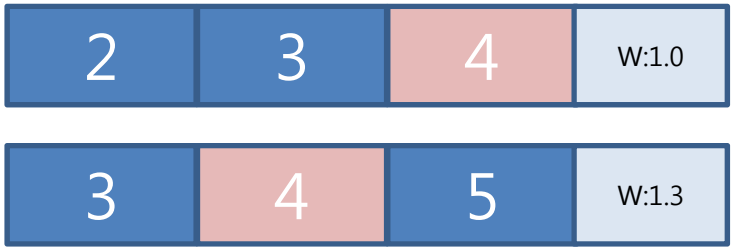
Answer → 4!!

Language Learning with Hypernetworks (8/8)

- We can also analysis word associations.

- We can enumerate the associativity of words based on following features of hypernetworks
  - the co-occurrence of words in a hyperedge
  - the weight of the hyperedge

- For example
  - 4 is associated with
    - *2* with weight 1.0
    - *3* with weight 2.3(1.0+1.3)
    - *5* with weight 1.3



Data Set (Friends & BBC)

- English sentences from movie subtitles
- Friends
  - Well known TV situation comedy
  - Captions from Season 1~10
  - 5,000 sentences
- BBC Documentary
  - Captions from three series on space, bird, and wild
  - 5,000 sentences

Data Set (cont'd)

- Each sentence is translated to integer form based on dictionary file.
  - "This is not even a date"
  - → "33,34,35,36,27,37"
- Experiment with
  - friends_training.cvs, bbc_training.csv
- Original sentence file
  - friends_original.txt, bbc_original.txt

Tool For Hypernetworks

- Language Game (for this project)
  - Sentence Completion
  - Language Classification
  - Word Association
- Multimodal Game
  - Language to Image Generation
  - Image search using language query

Report Contents – Mandatory (1/2)

- System description
  - Used software and running environments
- Result graphs and tables
  - Do several experiments and calculate average & standard deviations
- Analysis & discussion
  - Very Important

Report Contents – Mandatory (2/2)

- Basic experiments
  - Draw learning curves for each training set
    - Graph type 1: accuracy vs. epoch, with *orders* of hyperedges ex) 3,4,5
    - Graph type 2: accuracy vs. epoch, with different weight update parameters ex) 0.01, 0.1, 0.5, 1, 2, 5, 10
      - Note : initial weight is assigned as 1.0 for each hyperedge
  - Comparison between two training sets
    - Graph type 3: learning curves for Friends and BBC sentences in one graph

Report Contents – Optional

- Various experiments and analyses
  - Comparing learning curves
    - w/ various setting(varying order & update parameters & training data)
  - Comparing ***word associations***
    - w/ different training set
  - Comparing sentence completion results
    - w/ different training set
    - Test (1,2,?,4) **(find out suitable queries)**
      - For training set A: ? = 3
      - For training set B: ? = 5, why?

Reports Style

- **English only**, Scientific journal-style
  - How to Write A Paper in Scientific Journal Style and Format
    http://abacus.bates.edu/~ganderso/biology/resources/writing/HTWsections.html

Experimental process	Section of Paper
What did I do in a nutshell?	Abstract
What is the problem?	Introduction
How did I solve the problem?	Materials and Methods
What did I find out?	Results
What does it mean?	Discussion
Who helped me out?	Acknowledgments (optional)
Whose work did I refer to?	Literature Cited
Extra Information	Appendices (optional)

Submission Guide

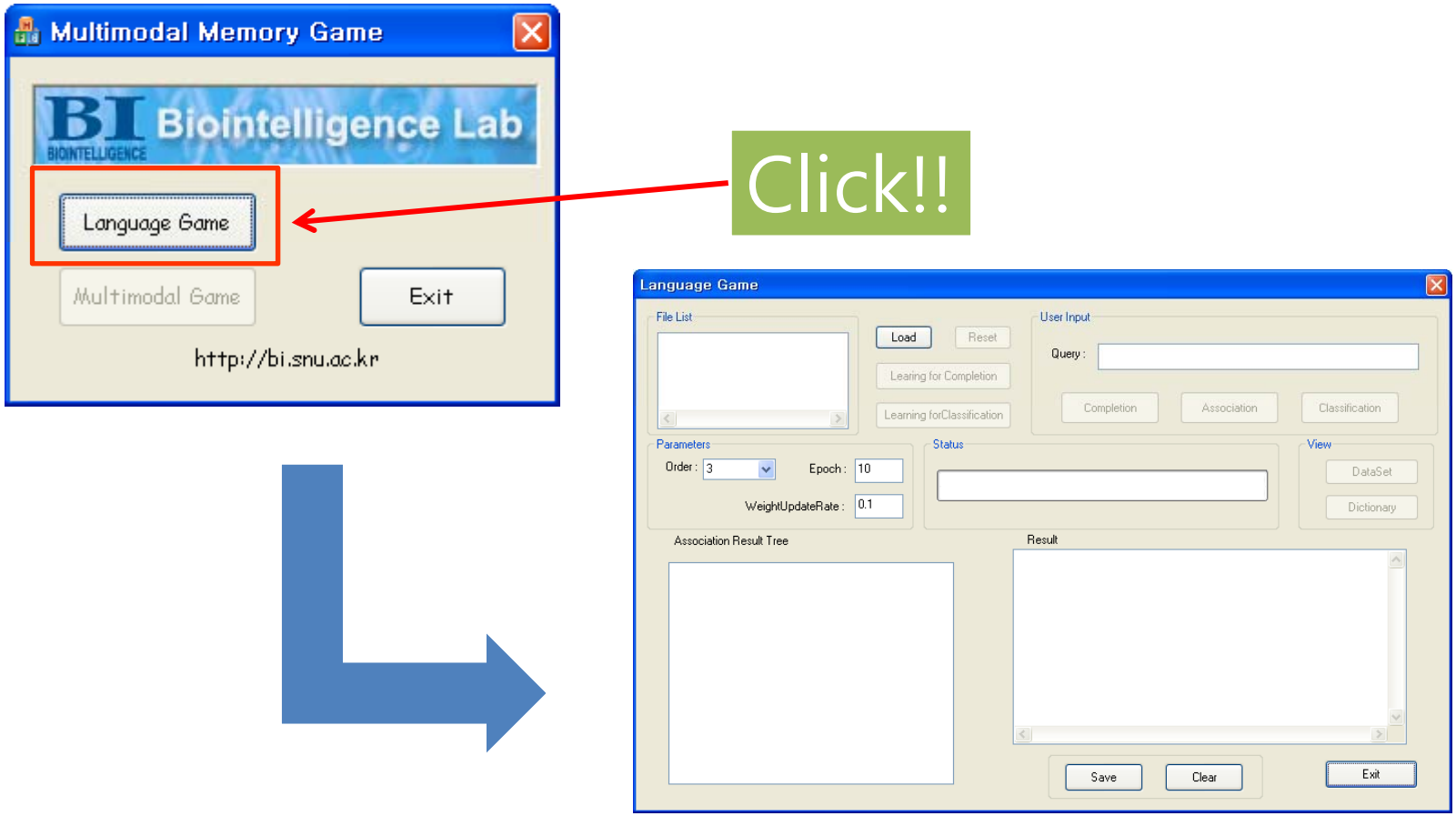
- Due date: May 27, 13:00
- Submit both ‘hardcopy’ and ‘email’
  - Hardcopy submission to the office (301-417)
  - E-mail submission to ykko@bi.snu.ac.kr
    - Subject : [AI Project2 Report] Student number, Name
  - Length: report should be summarized within 12 pages.
  - If you build a program by yourself, submit the source code with comments
- Objective: NOT the accuracy and your programming skill, but your creativity and research ability.
- **Individual project! You have to do it by yourself.**

Marking Scheme

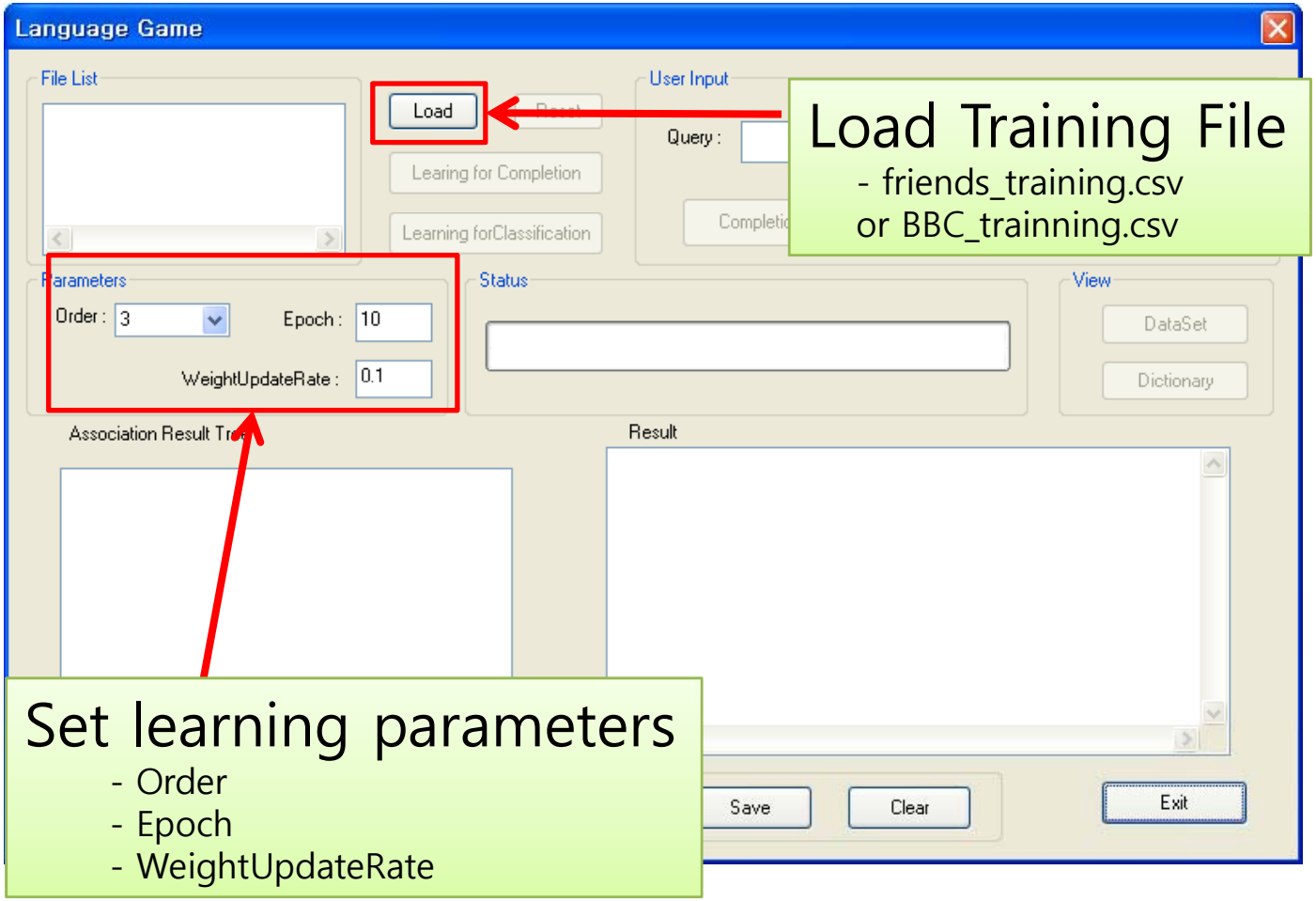
- 40 points for experiment & analysis
  - Extra 3 points per additional experiment
- 20 points for the report
- 6 points for overall organization
- Late work
  - (- 10%) per one day (8 points)
  - Maximum 7 days

Demo – How to Start

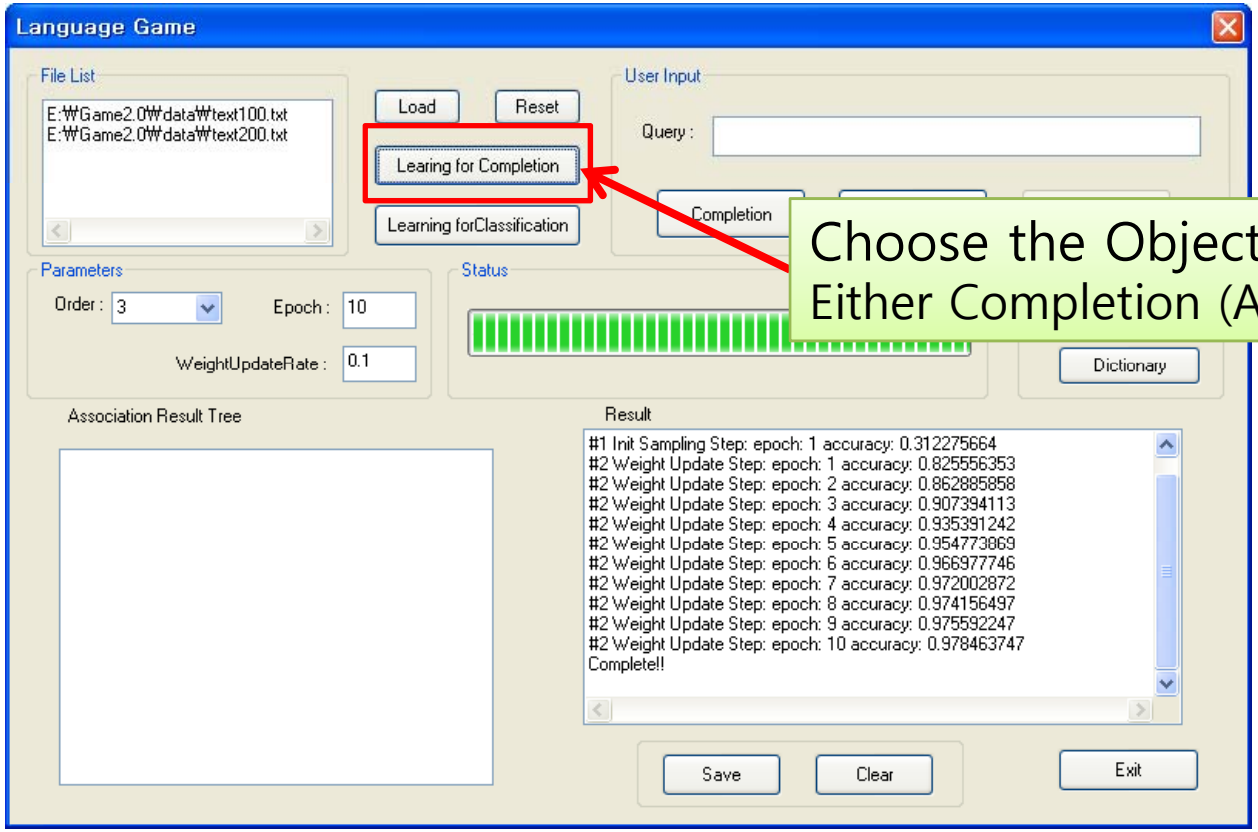
- Execute: MemoryGame_2.0.exe



Setting data and parameters



Do Learning - Completion

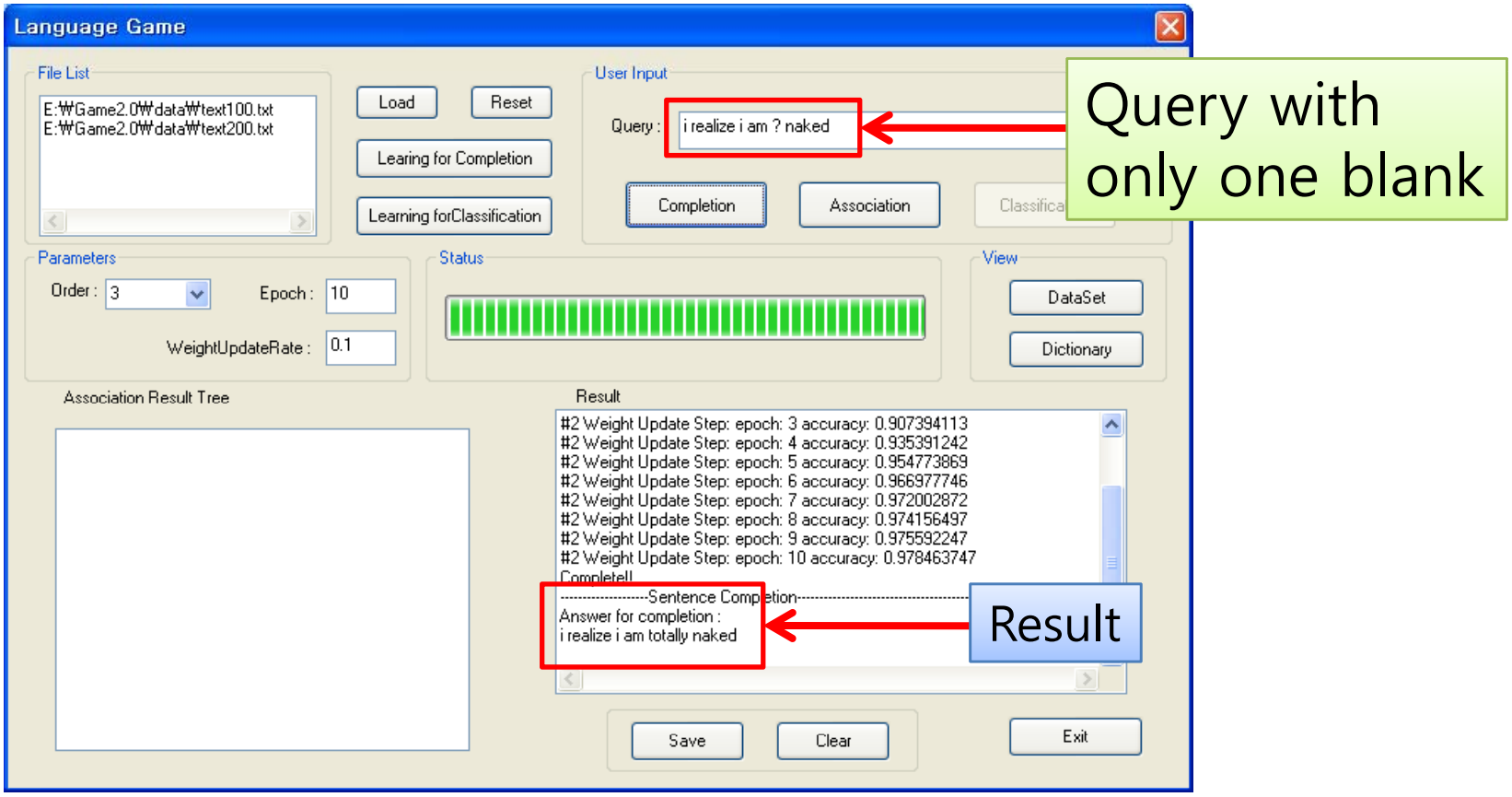


Choose the Objective of Learning:
Either Completion (Association)

Warning: it takes much time

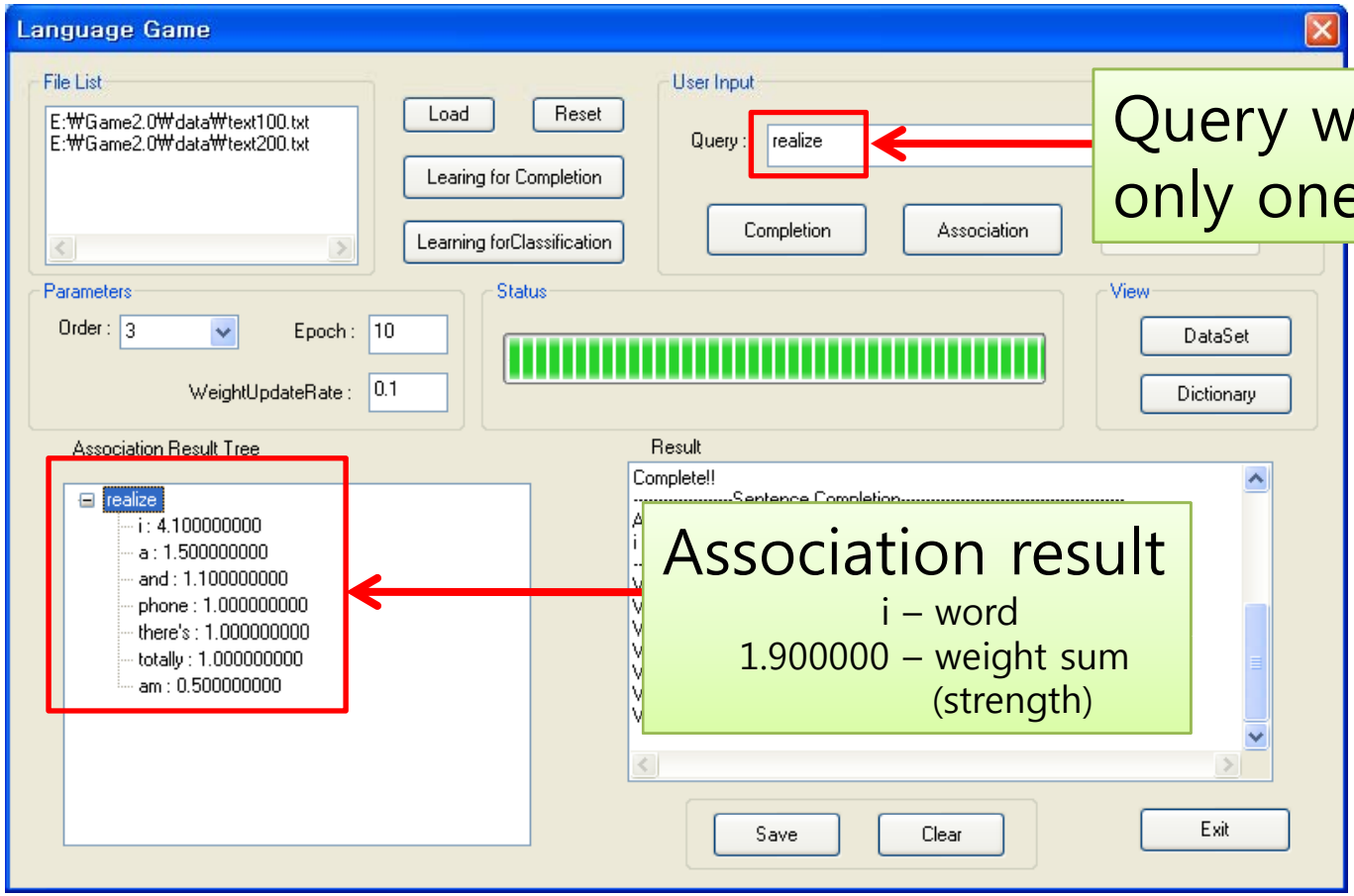
Test (sentence completion)

Sentence Completion

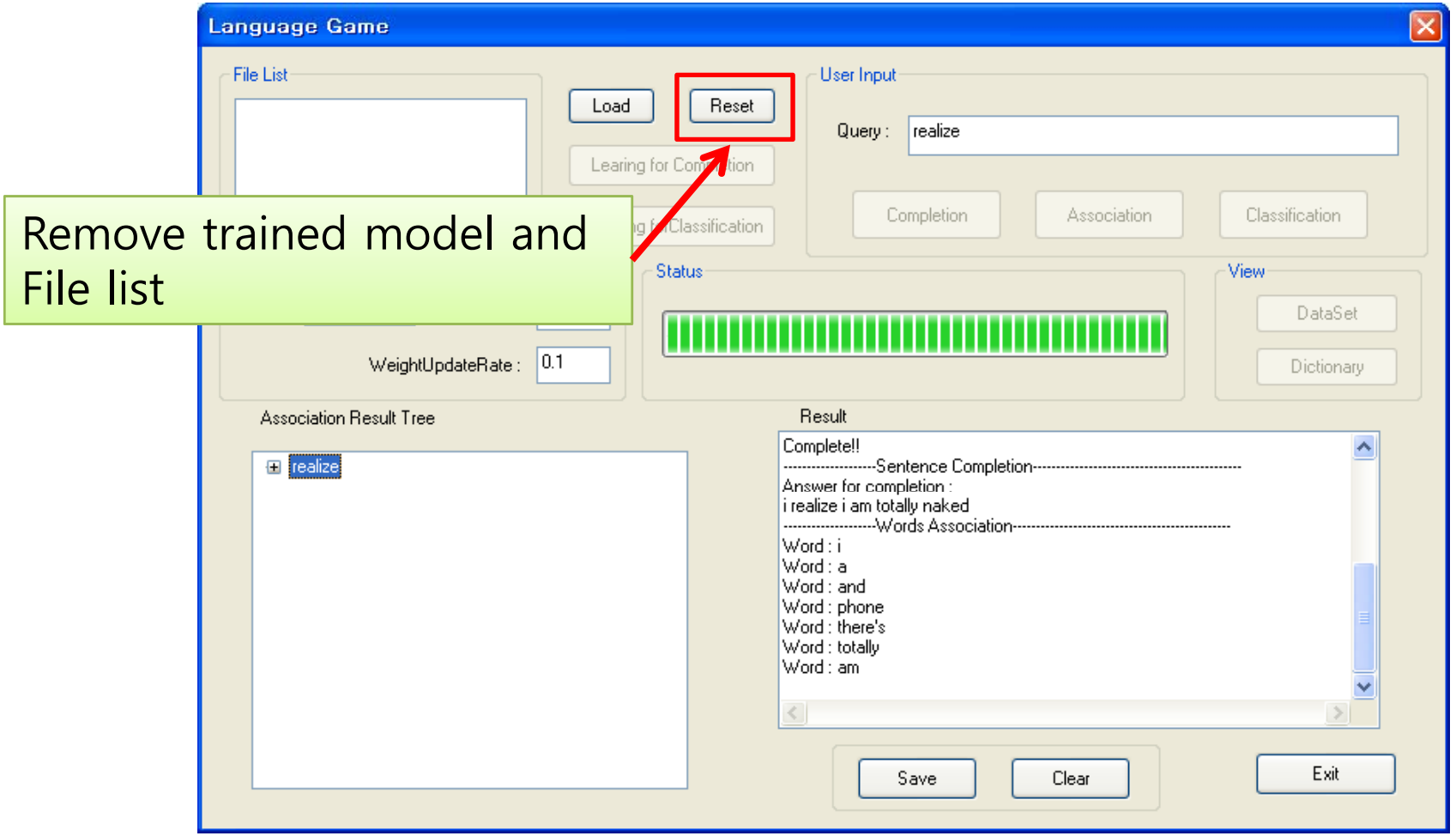


Learning for Completion

Word Association

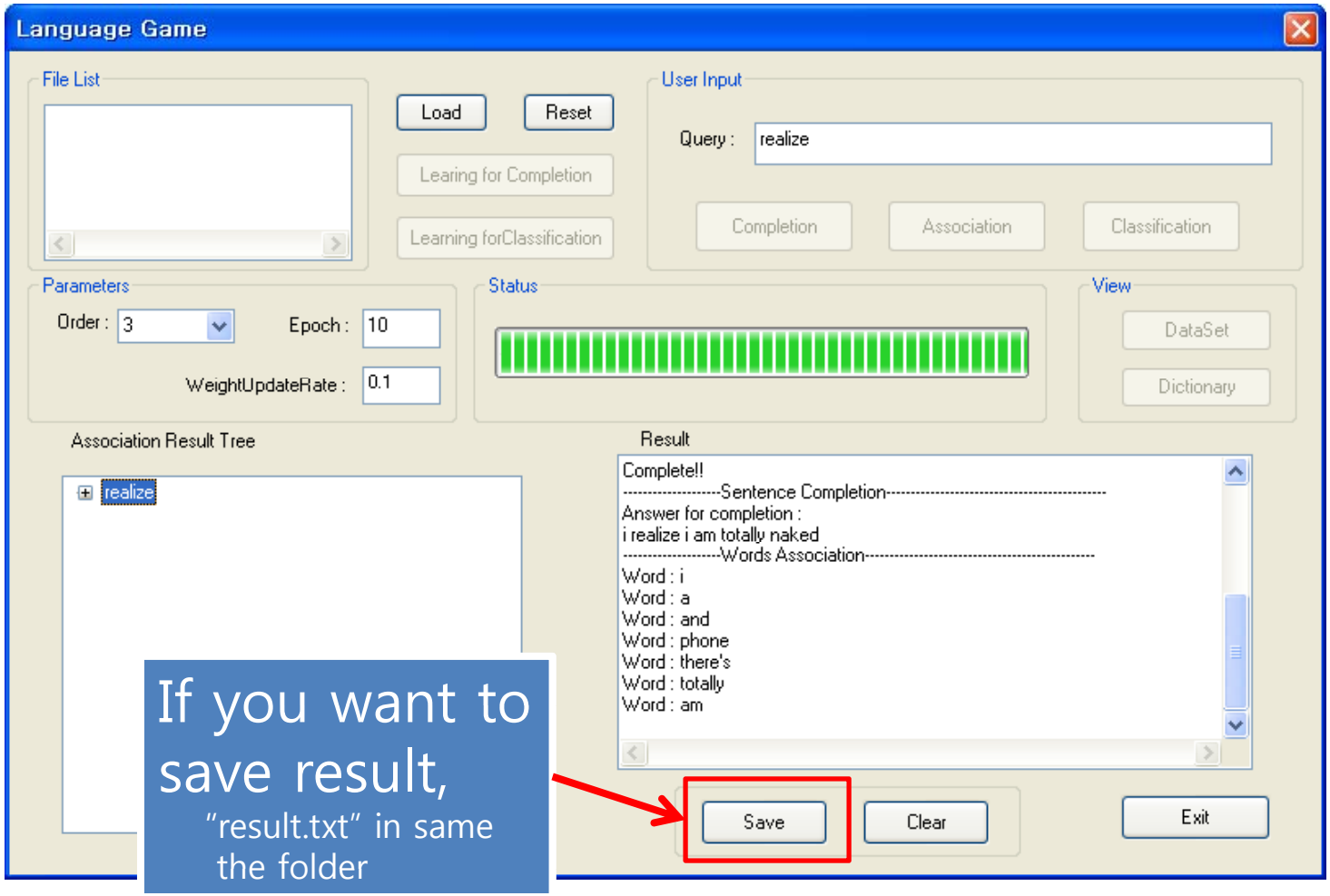


Reset



Remove trained model and File list

Save result


Language Game
File List
Load
Reset
User Input
Query :
realize
Learing for Completion
Learning forClassification
Completion
Association
Classification
Parameters
Order :
3
Epoch :
10
WeightUpdateRate :
0.1
Status
View
DataSet
Dictionary
Association Result Tree
realize
Result
Complete!!
----------Sentence Completion----------
Answer for completion :
i realize i am totally naked
----------Words Association----------
Word : i
Word : a
Word : and
Word : phone
Word : there's
Word : totally
Word : am
Save
Clear
Exit
If you want to save result,
"result.txt" in same the folder

Warning!!

- Program path can not have any Korean or other language except English
  - "C:₩Documents and Settings₩바탕 화면₩MemoryGame2.0"
    - Not accepted.
  - "C:₩Documents and Settings₩MemoryGame2.0"
    - It's OK.
- Current program does not allow making new training files
  - Dictionary file is fixed.
  - If you want to, make dictionary file too.
- If you have any question about the program, visit the office 301-417 (Tel. 880-1835)
  - Youngkil, Ko (ykko@bi.snu.ac.kr)