Word Sense Disambiguation

Senses of “Table”

1. The whole family was sitting at the **table$_1$** and enjoying their dinner.

2. This **table$_2$** presents the preliminary results of our investigation.

The word table$_1$ refers to a piece of furniture, while the word table$_2$ means a type of data arrangement in rows and columns.

Overview of the Problem

- ***Problem:*** many words have different meanings or senses ==> there is ambiguity about how they are to be interpreted.
- ***Task:*** to determine which of the senses of an ambiguous word is invoked in a particular use of the word. This is done by looking at the context of the word's use.

Overview of our Discussion

1. ***Introduction***
2. ***Methodology***
3. ***Supervised Disambiguation***: based on a labeled training set.
4. ***Dictionary-Based Disambiguation***: based on lexical resources such as dictionaries and thesauri.
5. ***Unsupervised Disambiguation***: based on unlabeled corpora.
6. **Evaluation** on TWSI and SemEval

Word Sense Disambiguation

1. INTRODUCTION

Introduction

- Word Sense Disambiguation

  Is to determine which of the senses of an **ambiguous word** is invoked in a particular use of the world.

- Ambiguous word(1)

  - He saves half of his salary every month in the bank.
    - bank[1]: The rising ground bordering a lake, river, or sea.
    - bank[2]: An establishment for the custody, loan exchange, …

Word Sense Disambiguation

2. METHODOLOGY

Methodological Preliminaries

- ***Supervised versus Unsupervised Learning***: in supervised learning the sense label of a word occurrence is known. In unsupervised learning, it is not known.
- ***Pseudowords***: used to generate artificial evaluation data for comparison and improvements of text-processing algorithms.
- ***Upper and Lower Bounds on Performance***: used to find out how well an algorithm performs relative to the difficulty of the task.

3. SUPERVISED DISAMBIGUATION

Supervised Disambiguation

- ***Training set***: exemplars where each occurrence of the ambiguous word w is annotated with a semantic label ==> Classification problem.
- ***Approaches***:
  - Bayesian Classification: the context of occurrence is treated as a bag of words without structure, but it integrates information from many words.
  - Information Theory: only looks at informative features in the context. These features may be sensitive to text structure.
  - There are many more approaches using Machine Learning techniques

Supervised Disambiguation: Bayesian Classification I

- ***(Gale et al, 1992)'s Idea:*** to look at the words around an ambiguous word in a large context window. Each content word contributes potentially useful information about which sense of the ambiguous word is likely to be used with it. The classifier does no feature selection. Instead, it combines the evidence from all features.
- ***Bayes decision rule***: Decide s' if $P(s' \mid C) > P(s_k \mid C)$ for $s_k \neq s'$.
- $P(s_k \mid C)$ is computed by Bayes' Rule.

Supervised Disambiguation: Bayesian Classification II

- ***Naïve Bayes assumption***: $P(C|s_k) = P(\{v_j|\ v_j \text{ in } C\}|\ s_k) = \prod_{vj \text{ in } C} P(v_j \mid s_k)$
- The Naïve Bayes assumption is incorrect in the context of text processing, but it is useful.
- ***Decisionrule for Naïve Bayes***: Decide s' if s'=argmax $_{sk}$ $[\log P(s_k)+\sum_{vj \text{ in } C} \log P(v_j \mid s_k)]$
- $P(v_j \mid s_k)$ and $P(s_k)$ are computed via Maximum-Likelihood Estimation, perhaps with appropriate smoothing, from the labeled training corpus.

4. DICTIONARY-BASED DISAMBIGUATION

Dictionary-Based Disambiguation: Overview

- We will be looking at three different methods:
  - Disambiguation based on sense definitions
  - Thesaurus-Based Disambiguation
  - Disambiguation based on translations in a second-language corpus
- Also, we will show how a careful examination of the distributional properties of senses can lead to significant improvements in disambiguation.

Disambiguation based on sense definitions

- ***(Lesk, 1986: Idea)***: a word's dictionary definitions are likely to be good indicators for the sense they define.
- Express the dictionary sub-definitions of the ambiguous word as sets of bag-of-words and the words occurring in the context of the ambiguous word as single bags-of-words emanating from its dictionary definitions (all pooled together).
- Disambiguate the ambiguous word by choosing the sub-definition of the ambiguous word that has the greatest overlap with the words occurring in its . context.

Thesaurus-Based Disambiguation

- ***Idea:*** the semantic categories of the words in a context determine the semantic category of the context as a whole. This category, in turn, determines which word senses are used.
- ***(Walker, 87):*** each word is assigned one or more subject codes which corresponds to its different meanings. For each subject code, we count the number of words (from the context) having the same subject code. We select the subject code corresponding to the highest count.
- ***(Yarowski, 92):*** adapted the algorithm for words that do not occur in the thesaurus but that are very . Informative. E.g., Navratilova --> Sports

One sense per discourse, one sense per collocation

- ***(Yarowsky, 1995)'s Idea***: there are constraints between different occurrences of an ambiguous word within a corpus that can be exploited for disambiguation:
  - ***One sense per discourse***: The sense of a target word is highly consistent within any given document.
  - ***One sense per collocation***: nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship.

5. UNSUPERVISED DISAMBIGUATION

Limit of Supervised Disambiguation

- While knowledge-based and supervised approaches demonstrate top performance in competitions, they suffer from the knowledge acquisition bottleneck.
- Supervised approaches require considerable amount of sense-annotated training data which is expensive to create and often inconsistent.

Unsupervised Disambiguation

- **Idea:** disambiguate word senses without having recourse to supporting tools such as dictionaries and thesauri and in the absence of labeled text. Simply cluster the contexts of an ambiguous word into a number of groups and discriminate between these groups without labeling them.

Two Categories

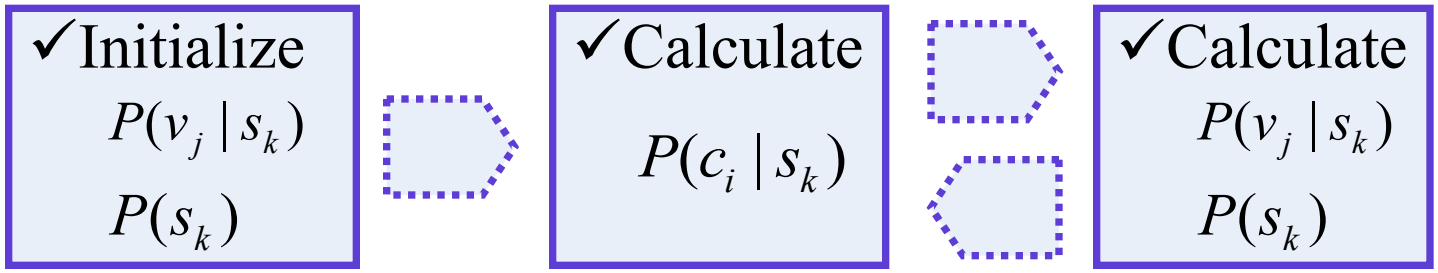
- Unsupervised sense induction methods fall into two categories:
  - Context clustering, such as [Pedersen and Bruce, 1997, Schütze, 1998, Reisinger and Mooney, 2010, Neelakantan et al., 2014, Bartunov et al., 2015]
  - Word similarity graph (ego-network) clustering, such as [Lin, 1998, Pantel and Lin, 2002, Widdows and Dorow, 2002, Biemann,].

Context clustering (1)

- ***(Schutze, 1998):*** The probabilistic model is the same Bayesian model as the one used for supervised classification, but the $P(v_j \mid s_k)$ are estimated using the EM algorithm.

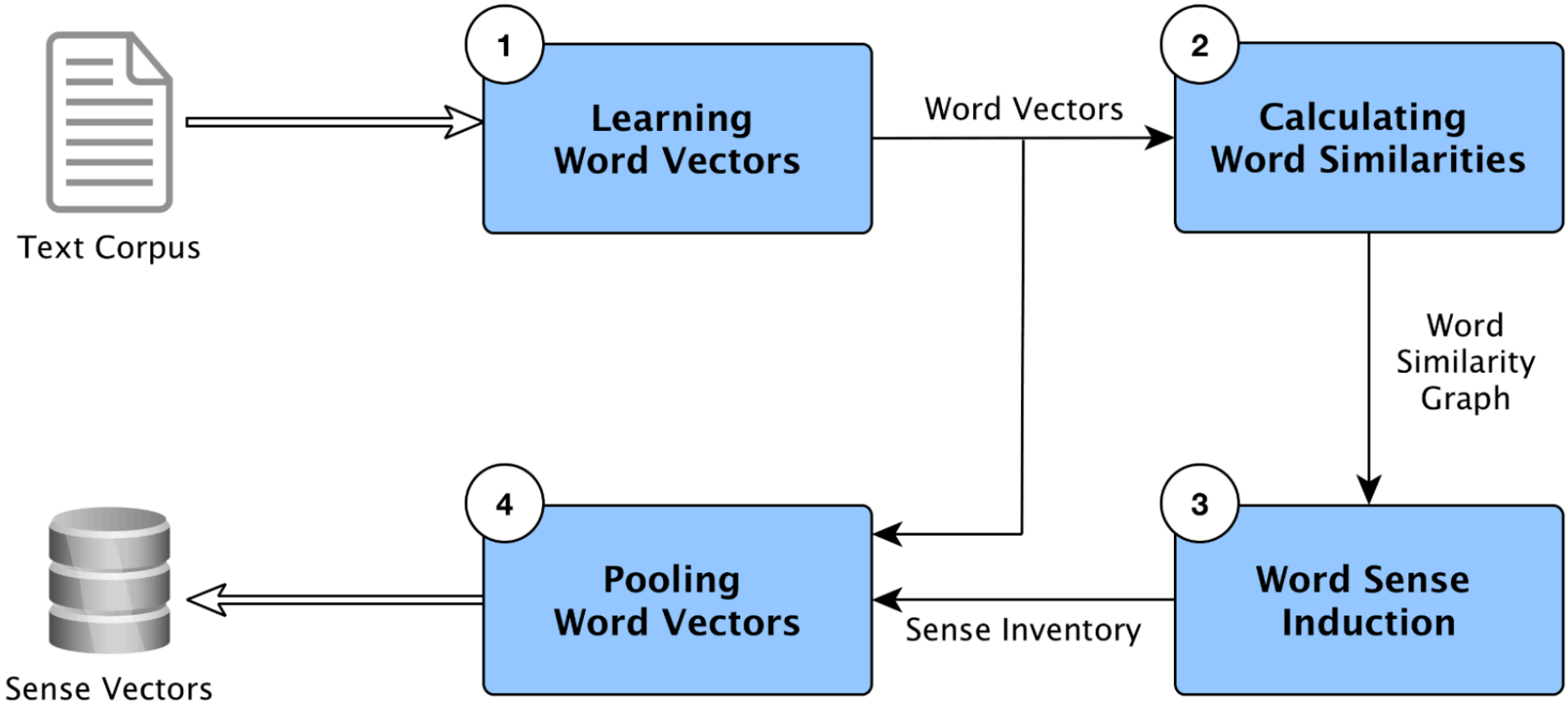
Context clustering (2)

- Context group discrimination
  - A type of sense discrimination.
  - Algorithm – (EM algorithm)

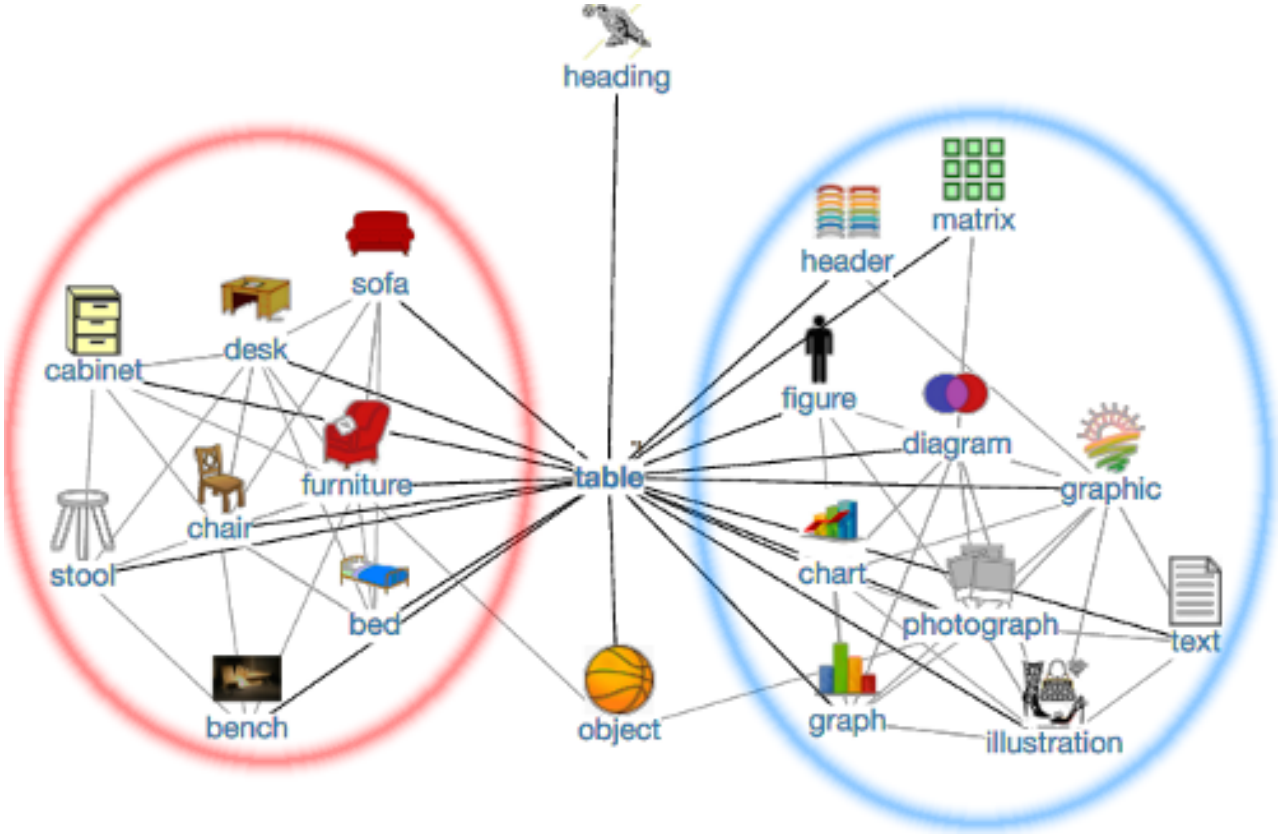


- A range of values of K(sense의 수)
  - is decided by how much the log likelihood increases with each new sense.
  - The more senses there are, the more structure the model has, and therefore it will be able to explain the data better.

word similarity graph (1)

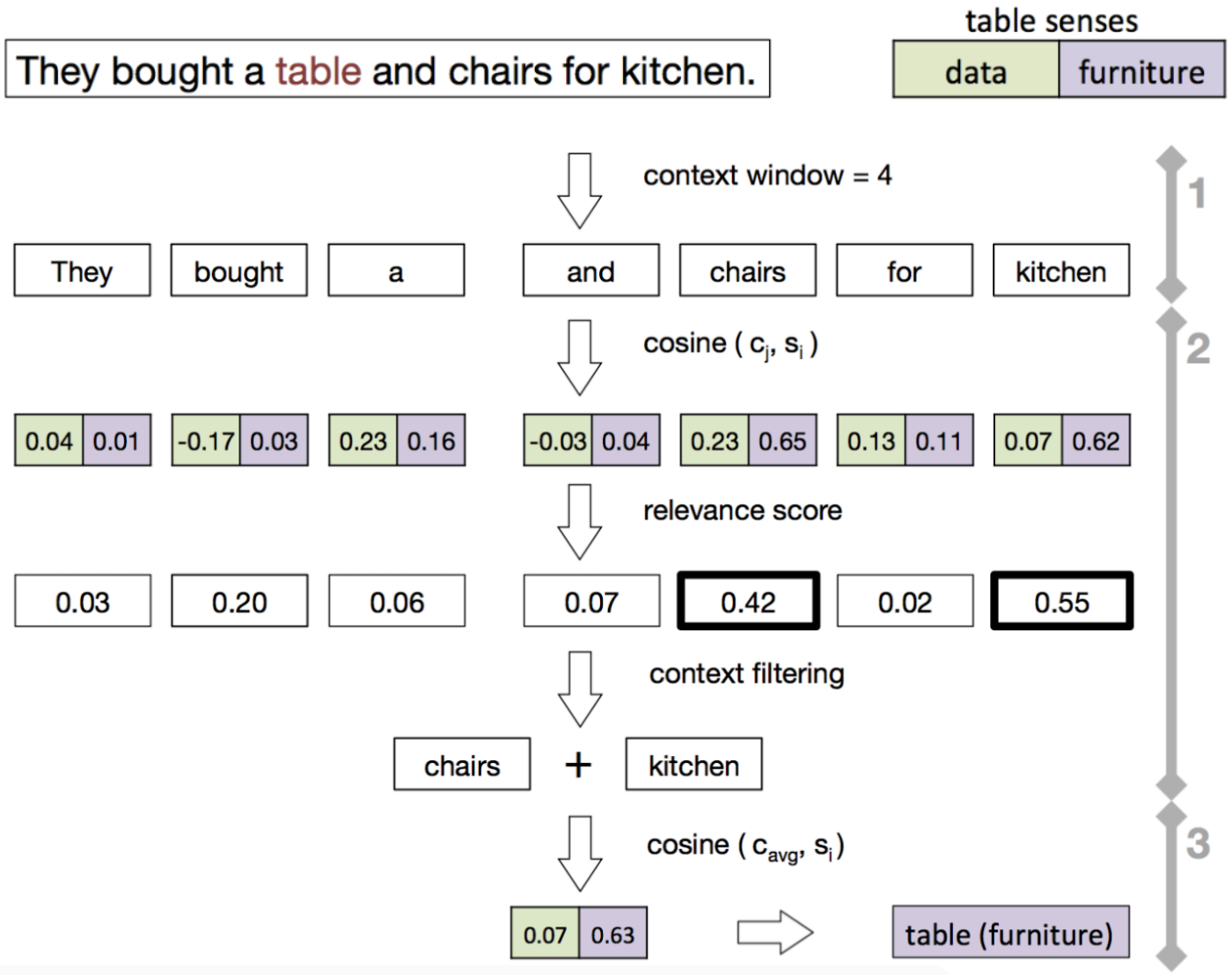


word similarity graph (2)



- The “furniture” and the “data” sense clusters of the word “table”
- Graph clustering using the Chinese Whispers algorithm (Biemann, 2006)

word similarity graph (3)



개별 실습과제 가이드

1. SenseGram 설치
   https://github.com/tudarmstadt-lt/sensegram
2. SenseGram 실행

```
$ python
>>> import sensegram
>>> sv = sensegram.SenseGram.load_word2vec_format(path_to_model/wiki.senses.w2v, binary=True)
```

3. SenseGram 단어 입력 및 결과 확인

```
>>> sv.get_senses("table")
 [('table#0', 0.40206185567), ('table#1', 0.59793814433)]
```

4. 단어 의미 확인

```
>>> sv.most_similar("table#1") [('pile#1', 0.9263191819190979), ('stool#1', 0.918972909450531), ('tray#0',
0.9099194407463074), ('basket#0', 0.9083326458930969), ('bowl#1', 0.905775249004364), ('bucket#0',
0.895959198474884), ('box#0', 0.8930465579032898), ('cage#0', 0.8916786909103394), ('saucer#3',
0.8904291391372681), ('mirror#1', 0.8880348205566406)]
```

5. 문장 입력 및 결과 확인

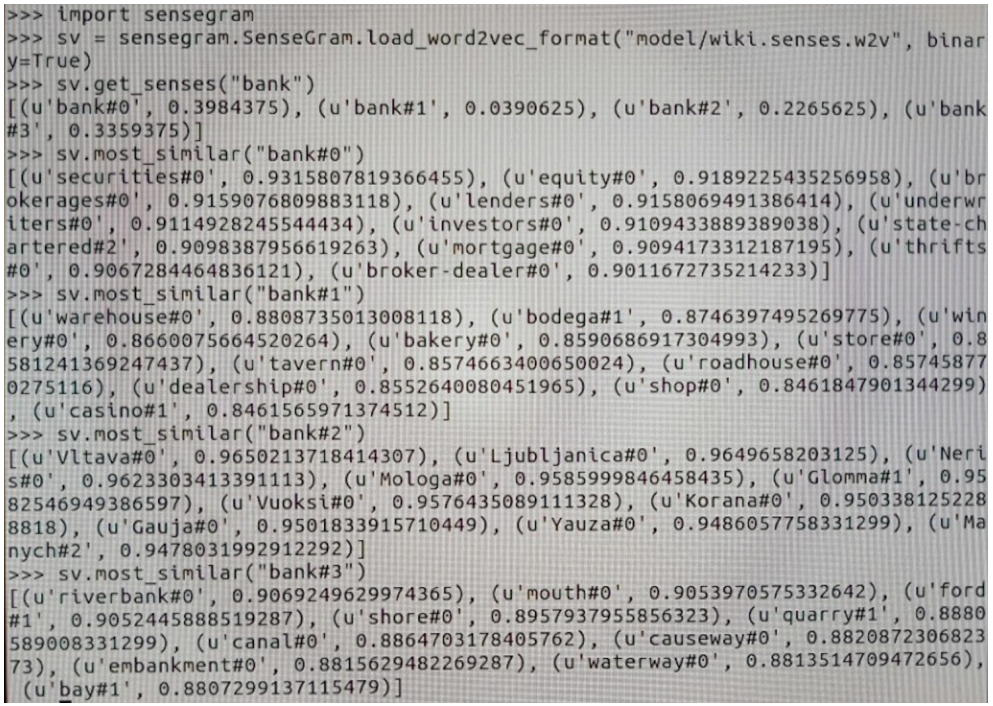
```
>>> wsd_model.dis_text("They bought a table and chairs for kitchen", "table", 14, 19) (u'table#1',
[0.15628162913257754, 0.54676466664654355])
```

개별 실습과제 평가 방법

방법 1) SenseGram 빌드 및 실행 결과 화면 갭쳐

방법 2) 필요 시 설치 과정부터 갭쳐 하면서 설치 과정 간략히 설명, 캡쳐 화면과 실행 과정 간략히 설명, 결과 간략히 설명 하여 레포트로 제출

실행 갭쳐 화면 →



Word Sense Disambiguation

6. EVALUATION

WSD Evaluation on TWSI

- TWSI Turk Bootstrap Word Sense Inventory
- https://www.lt.informatik.tu-darmstadt.de/de/data/twsi-turk-bootstrap-word-sense-inventory/
- TWSI is a large collection of sentences with a single sense-annotated word with relation to an induced sense inventory based on a lexical substitution task.

WSD Evaluation on SemEval

- SemEval (Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems, organized under the umbrella of SIGLEX, the Special Interest Group on the Lexicon of the Association for Computational Linguistics.
- SemEval has evolved from the SensEval word sense disambiguation evaluation series.
- http://alt.qcri.org/semeval2017/