

**SNLP07**  
**International Symposium**  
**on Natural Language Processing**

# **Effective Proximity Distance** **For Word-Based Context**

---

Graduate School of Information Science,  
Nagoya University, Japan

Masato HAGIWARA, Yasuhiro OGAWA, Katsuhiko TOYAMA

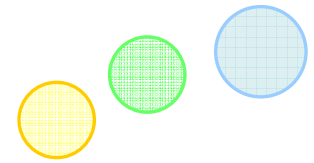
# Introduction

---



- Distributional similarity
  - Captures lexical semantic relatedness
  - Applications
    - Synonym acquisition, word sense disambiguation, etc.
  - Based on the *distributional hypothesis* [Harris 95]
    - “Semantically similar words share similar contexts”
    - The more similar the contexts are, the more related

# Example of distributional similarity



Guess what *tezgüino* is: [Lin 98]

- A bottle of *tezgüino* is on the table.
- Everyone likes *tezgüino*.
- *Tezgüino* makes you drunk.
- We make *tezgüino* out of vine.



*tezgüino* is something ...

- its bottle can be on the table.
- liked by everyone.
- makes you drunk.
- made out of vine.

*Context*

*tezgüino = wine?*



*wine* is something ...

- A bottle of *wine* is on the table.
- Everyone likes *wine*.
- *Wine* makes you drunk.
- We make *wine* out of vine.

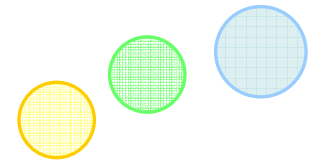


- Its bottle can be on the table.
- liked by everyone.
- makes you drunk.
- made out of vine.

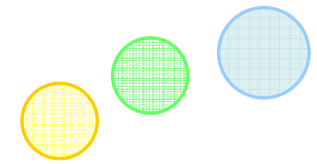
*Context*

- What kind of contexts should be used?

# Contextual information



- Word-based context
  - Surrounding words
  - Low extraction cost, simplicity
- Dependency structure
  - Words having syntactic relations
  - Better performance with smaller semantic space
- Dependency path
  - [Lin and Pantel 01][Pado and Lapata 07][Hagiwara et al. 07]
  - Words having *indirect* dependency relations
  - Best performance



# Word-based context

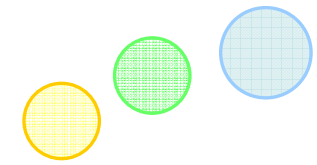
- Words which are surrounding the target word
- Consider a window, and extract words located within



*The investigators were still looking for **witnesses** and the motive of the attack.*

(Target word)	(Context)
witness	L1:for
witness	L2:look
witness	L3:still
witness	R1:and
witness	R2:the
witness	R3:motive

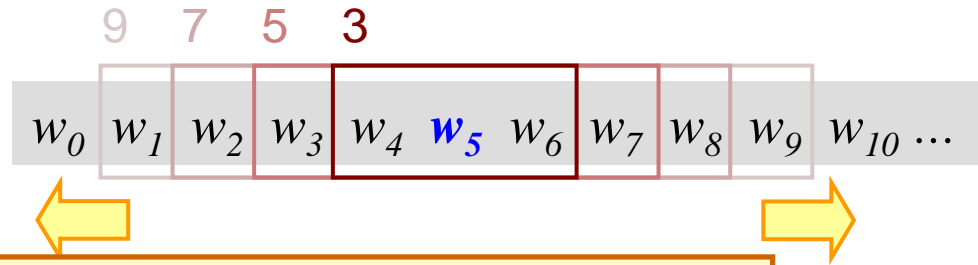
Window range : 3 tokens on both sides



# Effective window range

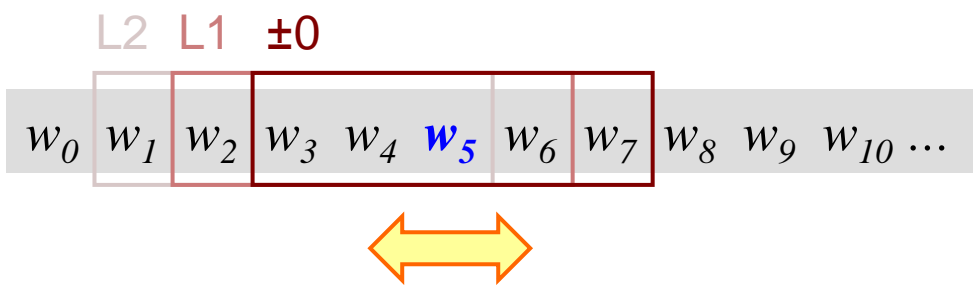
Not Fully Discussed

- Window width

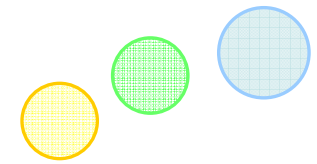


The effective window range for word-based context should be investigated

- Window offset



- Most of the previous studies use symmetric window

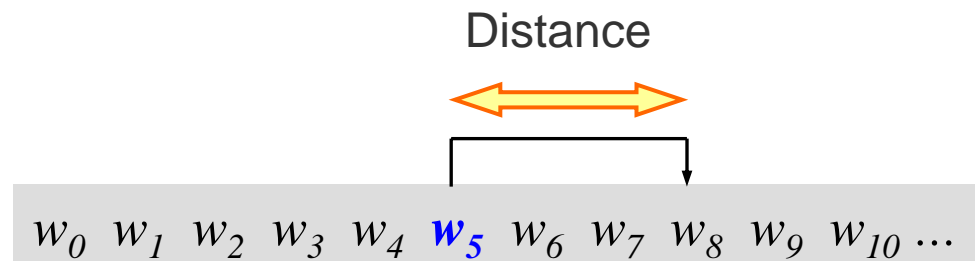


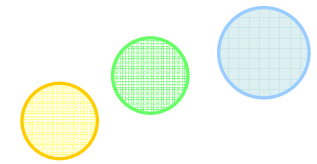
# Approach

- Use dependency structure as a clue
  - Effective contextual information for featuring words
  - Word-based context as an approximation of dep.

Windows which cover a large portion of dependency structure are effective?

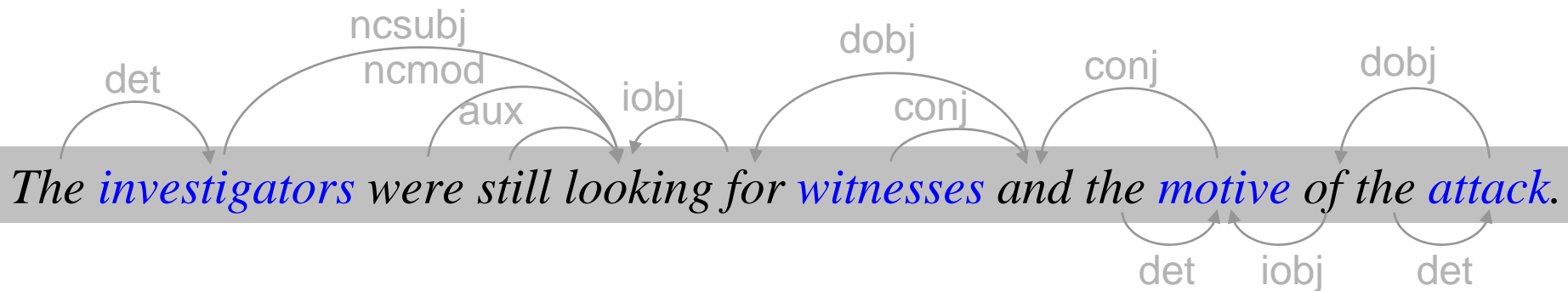
- 1 . Investigate the distance distribution of dependency
- 2 . Evaluate the performance on changing the window range
  - Automatic synonym acquisition was used as a task
  - Evaluation based on the existing thesauri such as WordNet

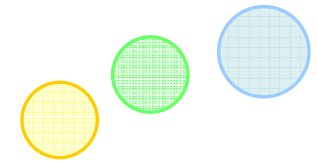




# Distance distribution of dependency

- Investigate the distance distribution of dependency between nouns and syntactically related words

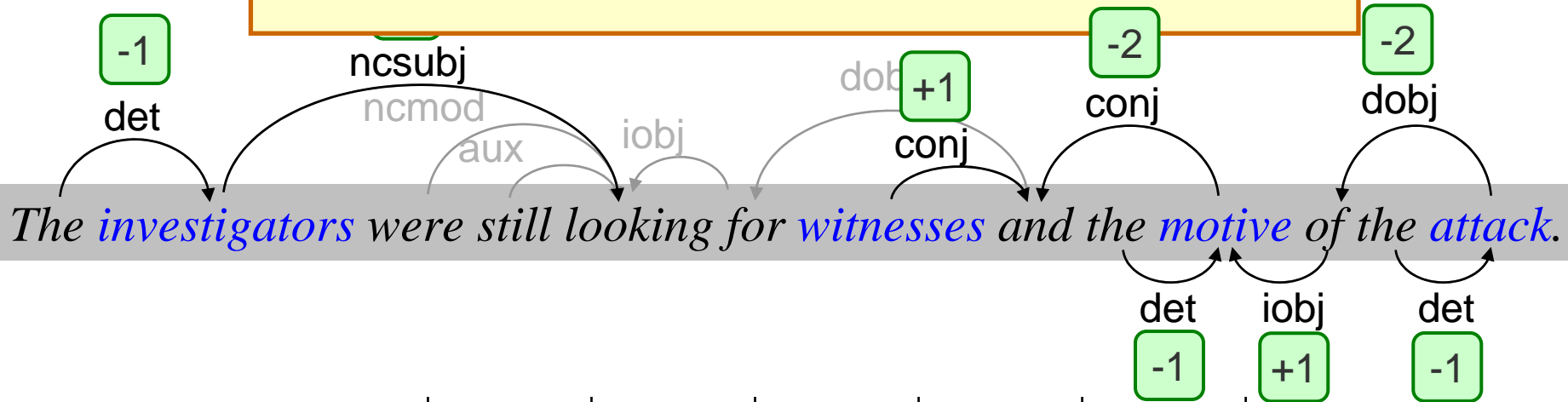




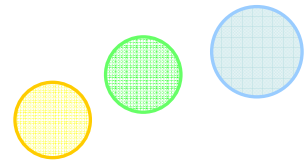
# Distance distribution of dependency

- Investigate the distance distribution of dependency between nouns and syntactically related words

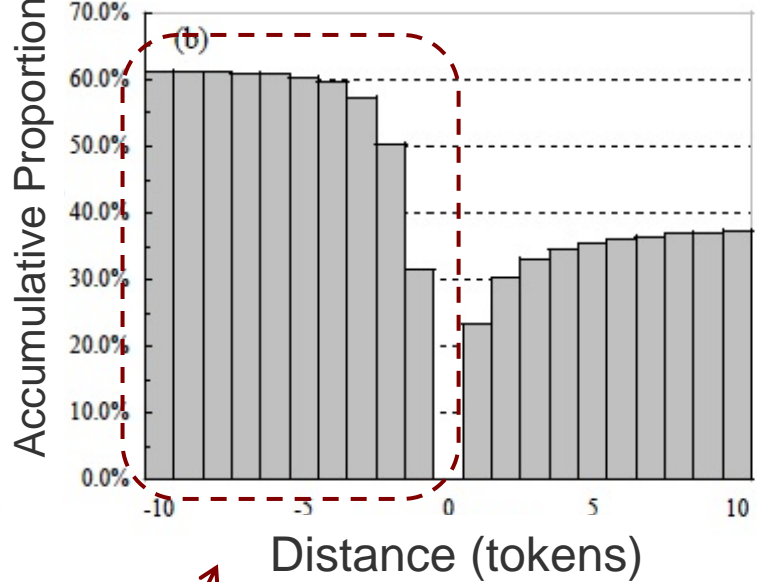
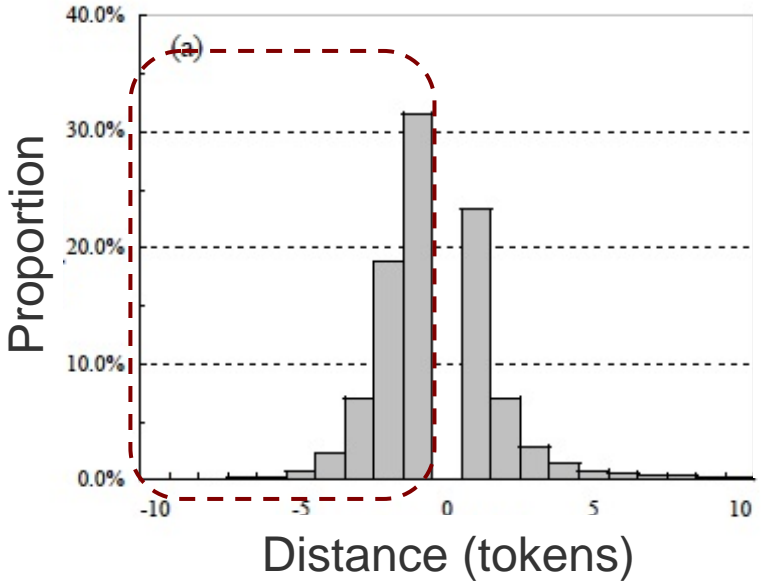
- Dependency extraction: RASP Toolkit 2  
- Corpus : WordBank (approx. 3.5M words)



Distance	-2	-1	0	+1	+2	+3
Count	2	3	0	2	0	1

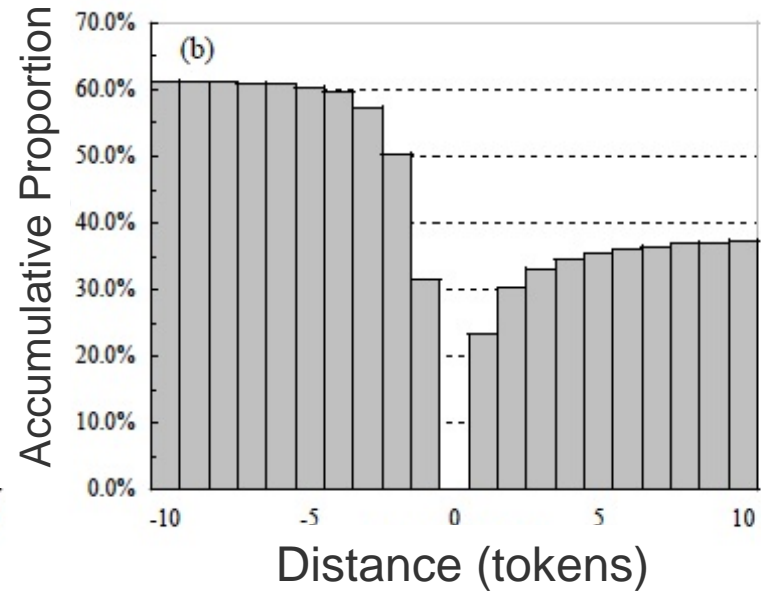
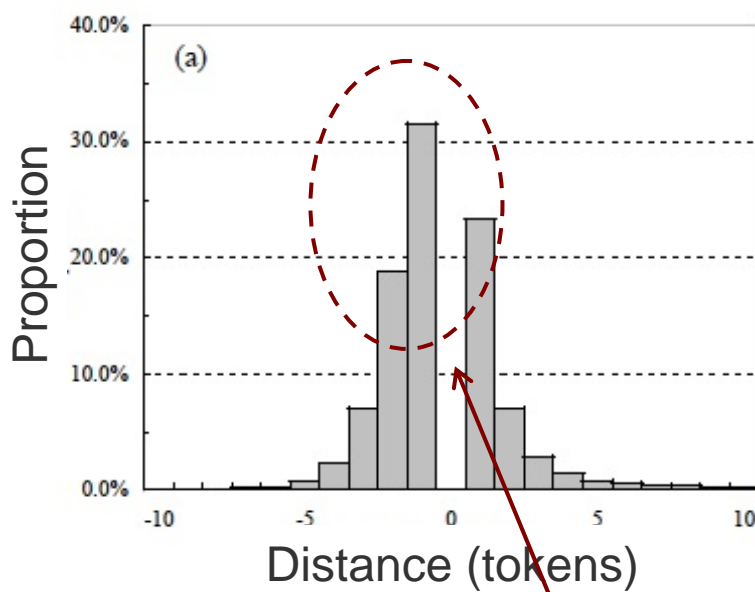
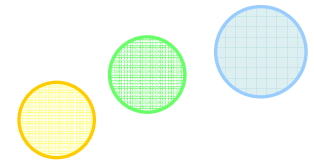


# Distance distribution results



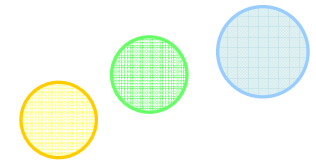
Much larger numbers of relations on the left side

# Distance distribution results

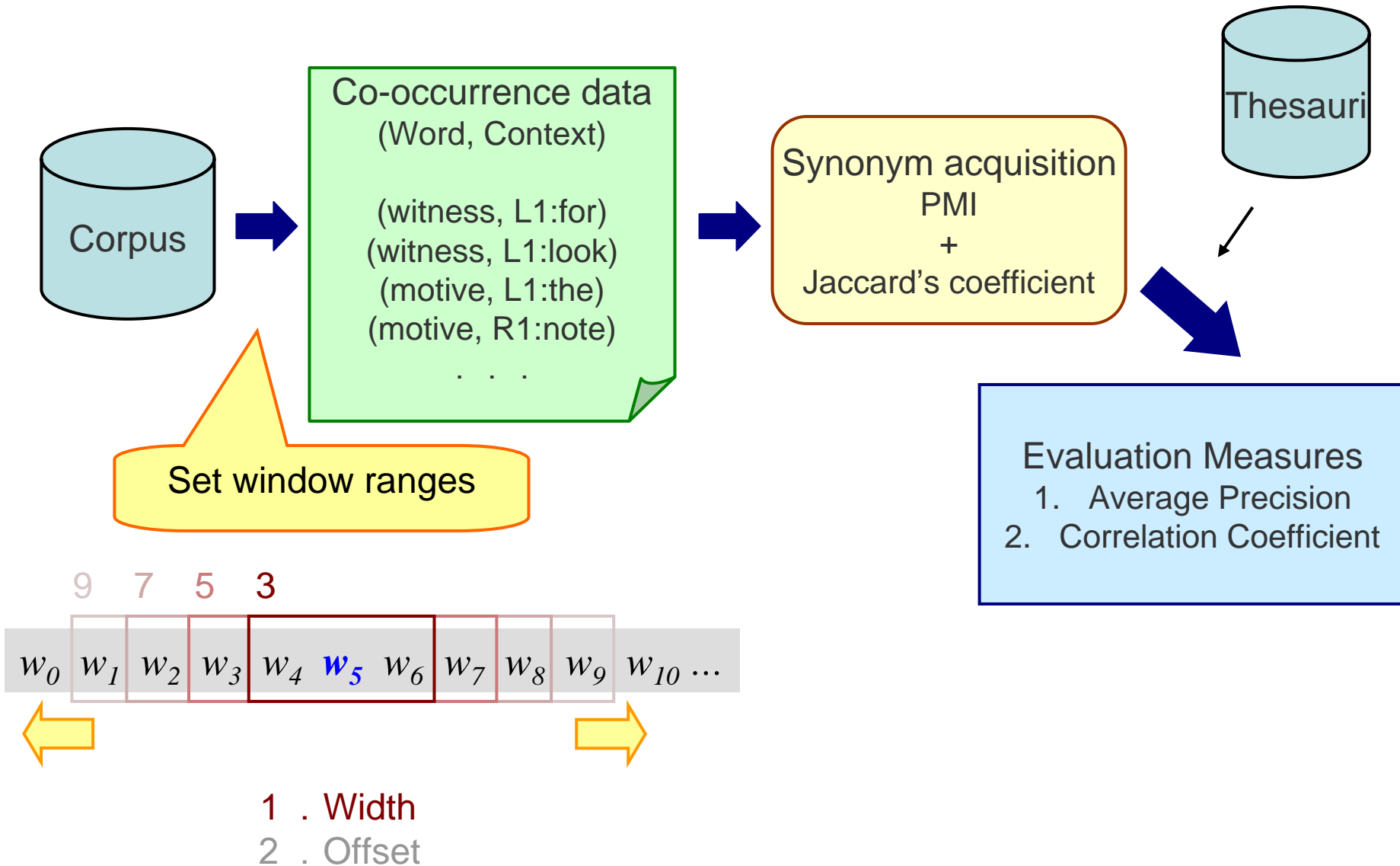


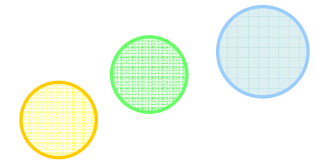
Peak at -1  
↓  
Because of  
articles and adjectives

The effective word-based contexts are also biased to the left?

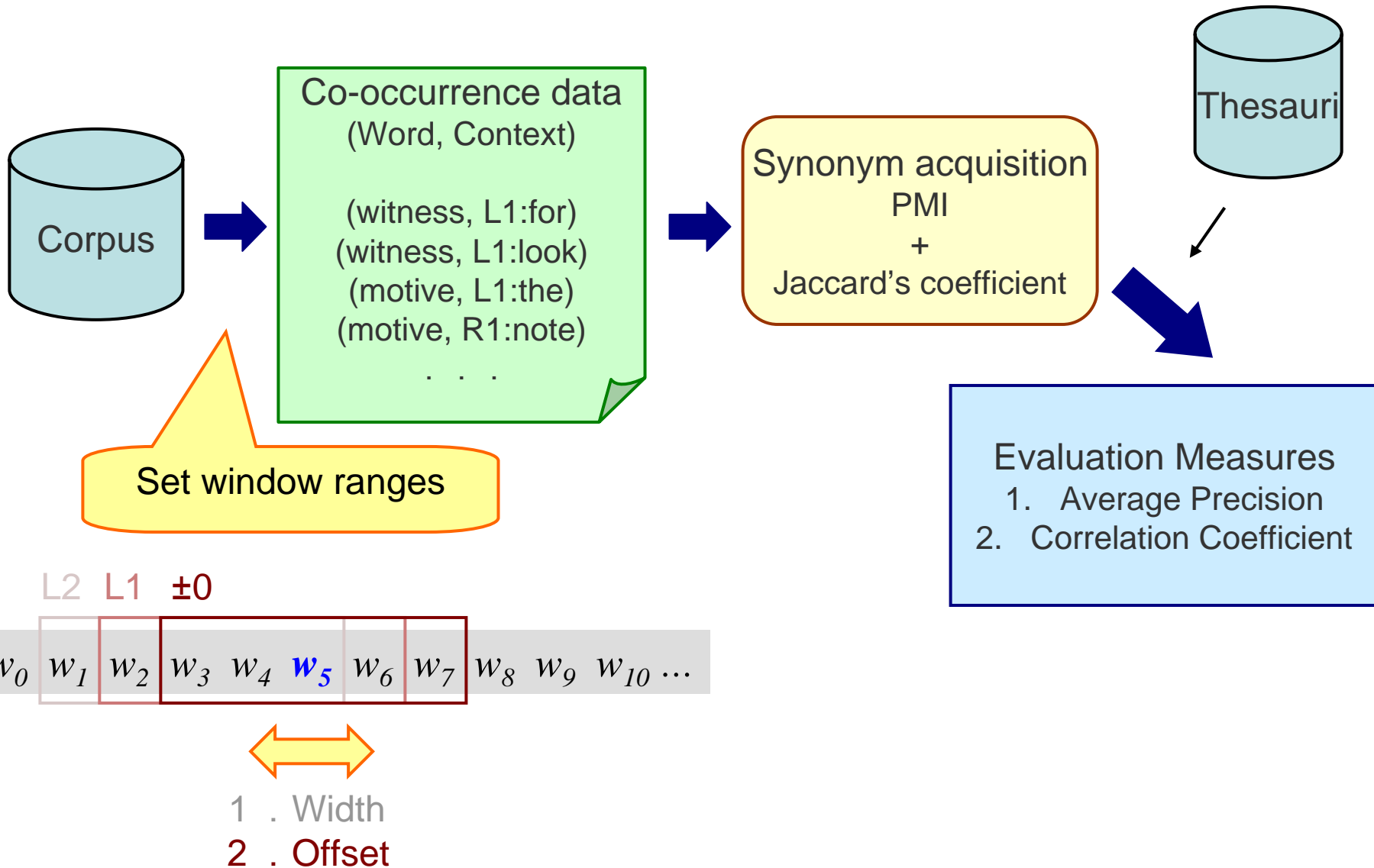


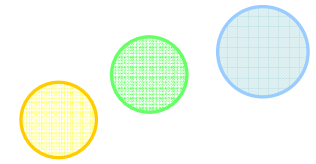
# Performance evaluation



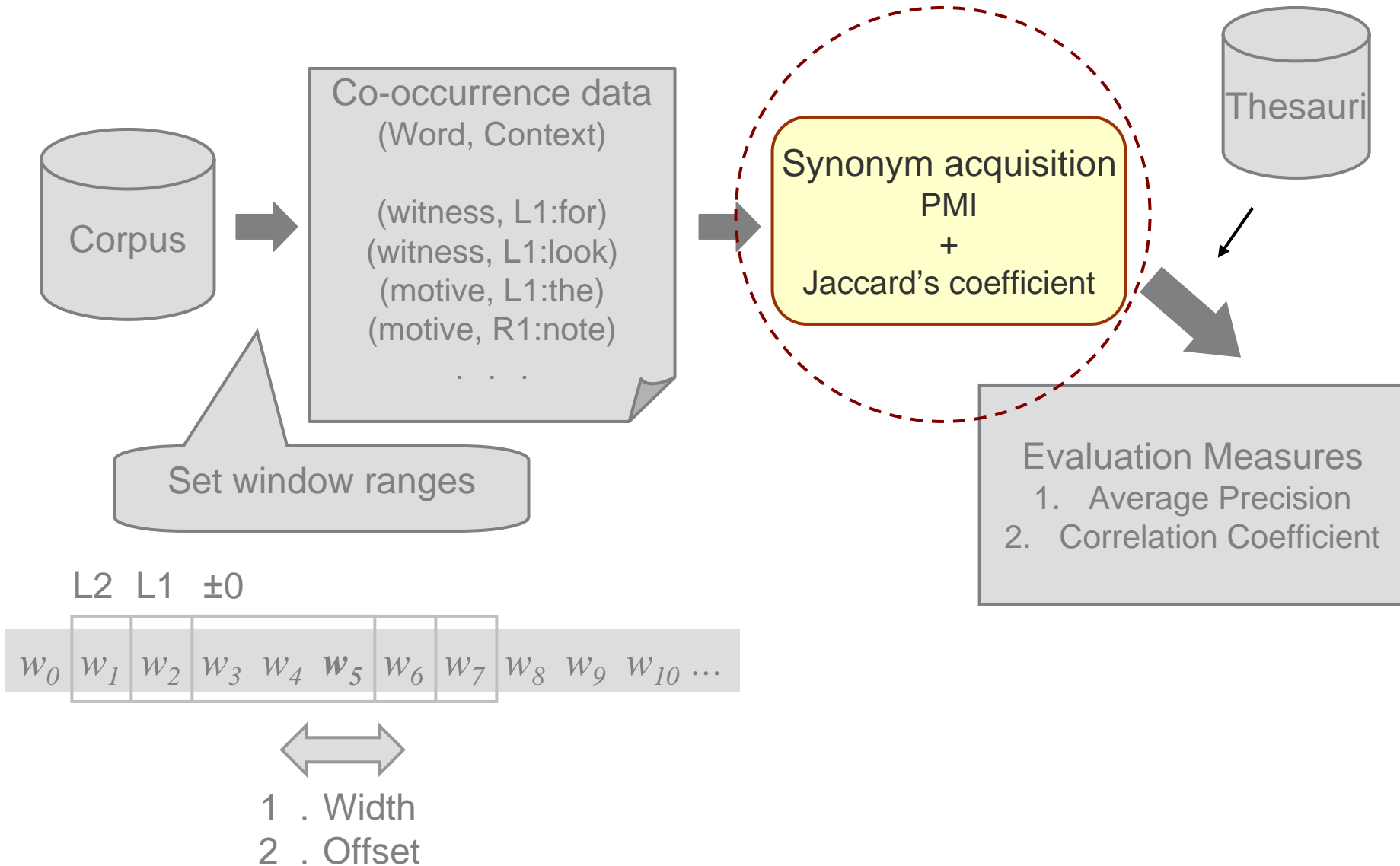


# Performance evaluation

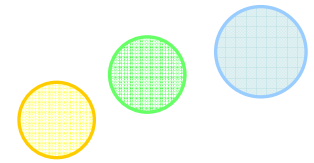




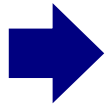
# Performance evaluation



# Synonym acquisition model



Language models or similarity measures are not within the scope of this study



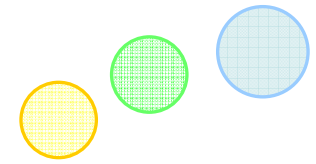
Used combination of pointwise mutual information and Jaccard coefficient

Vector construction :

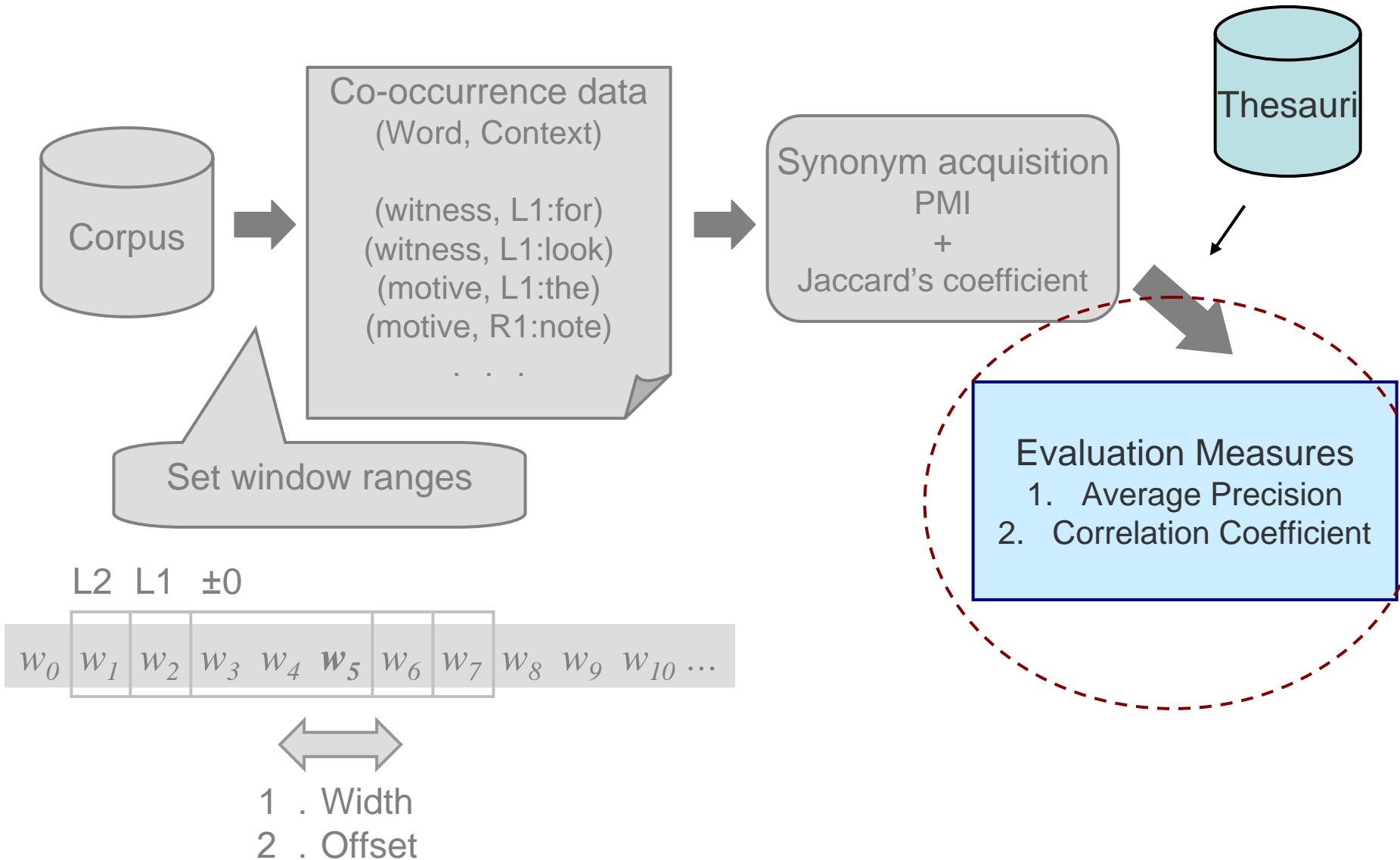
$$\mathbf{w}_i = \begin{pmatrix} pmi(w_i, c_1) \\ \vdots \\ pmi(w_i, c_M) \end{pmatrix}, pmi(w, c) = \log \frac{P(w, c)}{P(w)P(c)}$$

Similarity calculation :

$$sim(w_1, w_2) = \frac{\sum_{c \in C(w_1) \cap C(w_2)} \min(pmi(w_1, c), pmi(w_2, c))}{\sum_{c \in C(w_1) \cup C(w_2)} \max(pmi(w_1, c), pmi(w_2, c))}$$

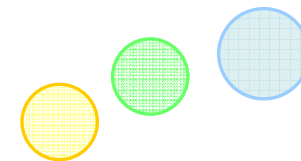


# Performance evaluation



# Evaluation measures

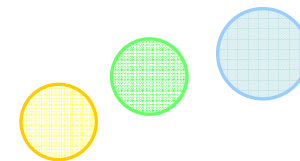
---



- Measure(1) – Average precision (AP)
  - Averaged precision values over 11 recall points
  - Based on the “reference set” created from three existing thesauri: WordNet, Roget’s, and COBUILD thesaurus
- Measure(2) – Correlation coefficient (CC)
  - Correlation between “reference similarity” and target similarity
  - Reference similarity: calculated based on the depth of word nodes in WordNet tree structure [Wu and Palmer 94]

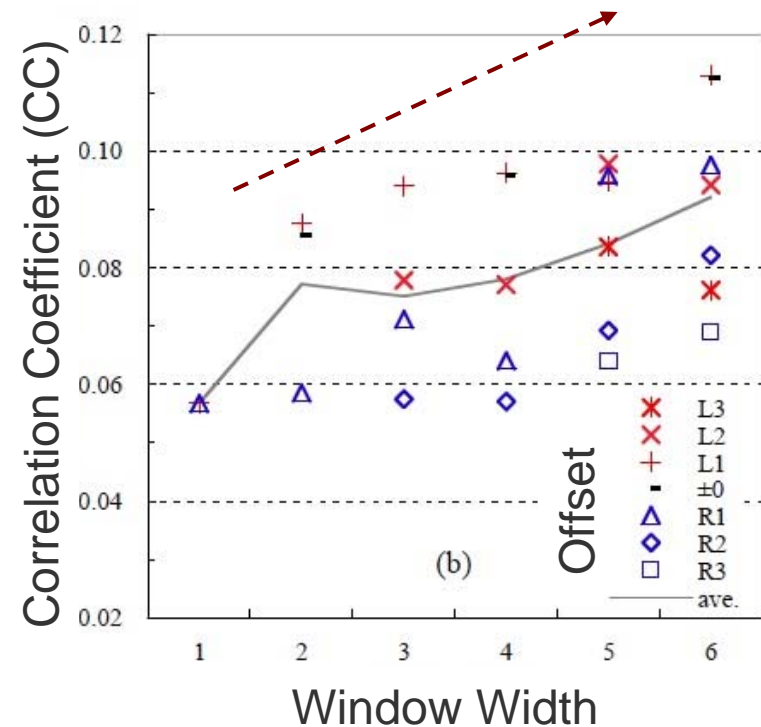
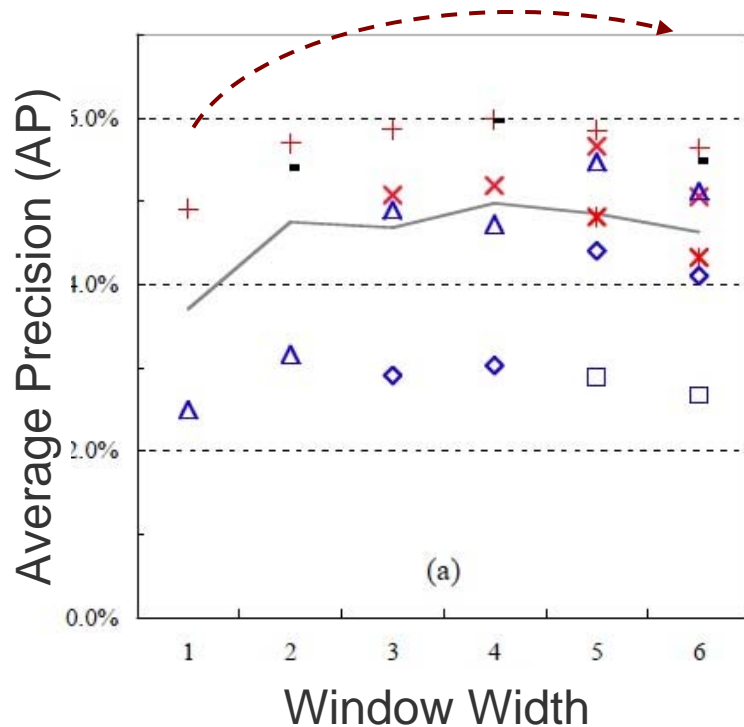
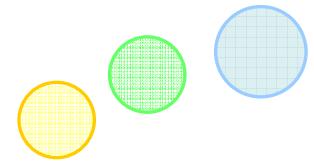
# Experiment – Conditions

---



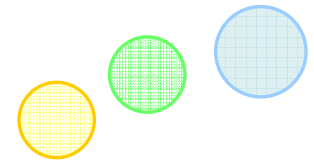
- Corpus
  - WordBank (approx. 190,000 sentences and 3.5M words)
- Limited to noun synonyms
- Frequency cutoff: removed words and contexts appearing less than  $\theta_f$  times
  - $\theta_f = 15$

# Evaluation results

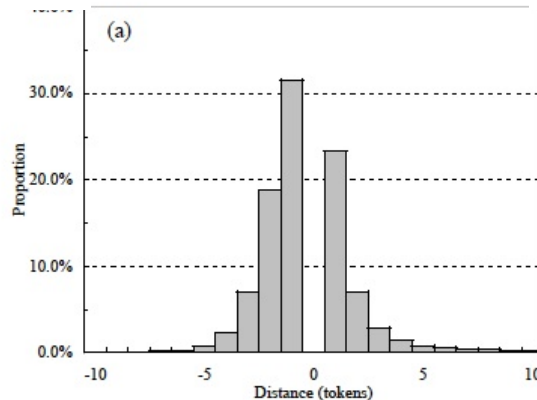
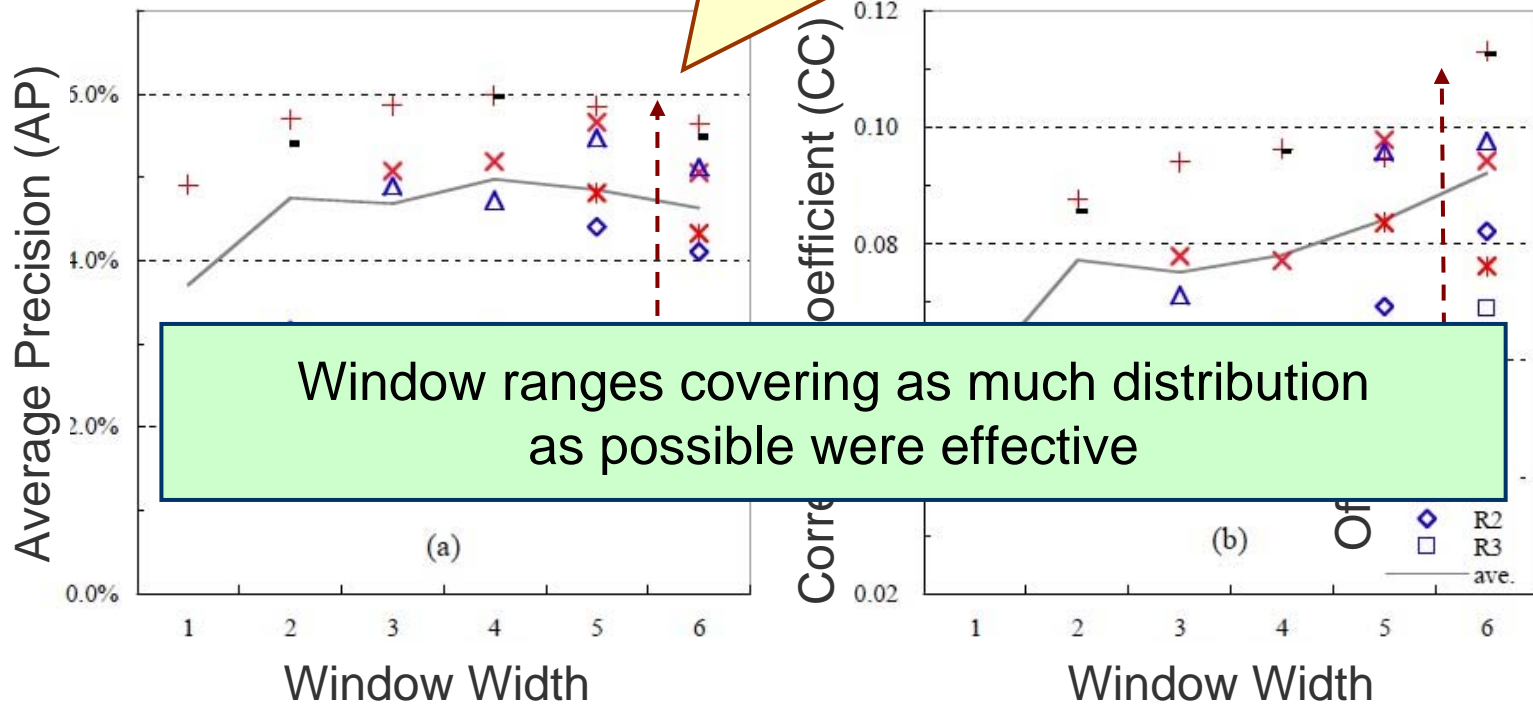


Wider window = Higher performance,  
But relatively small effect for AP

# Evaluation result

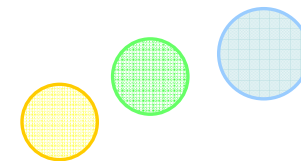


Asymmetric windows shifted to the left showed better performance



# Conclusion

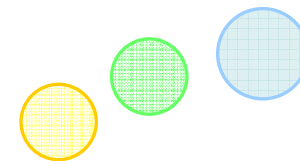
---



- Investigated the dependency distribution
  - As a clue for optimal window settings for word-based context
  - Biased to the left side of the target words
- Performance evaluation through synonym acquisition
  - Effectiveness of left-shifted asymmetric windows
  - Possible relation with the dependency distribution

# Future Works

---



- Confirm the relation between dependency and word-based context
  - Use different corpora, parsers, and/or languages
- Investigate on the representation of word-based context
  - POS tags, semantic role tags, offset marks, etc.