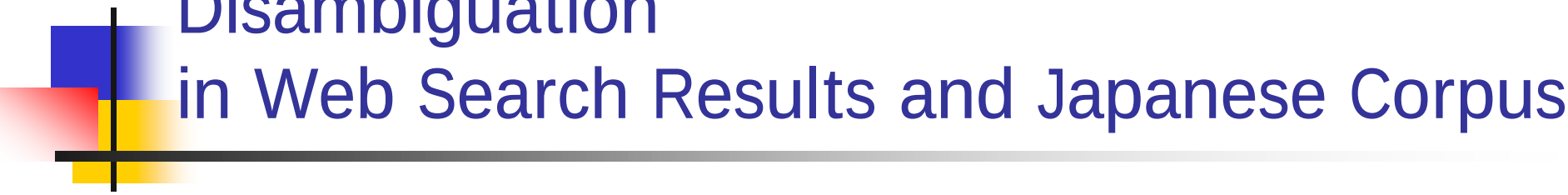
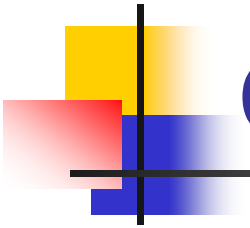


Disambiguation in Web Search Results and Japanese Corpus

Kazunari Sugiyama
School of Computing,
National University of Singapore



Outline

I. Personal Name Disambiguation
in Web Search Results

II. Word Sense Disambiguation
in Japanese Texts

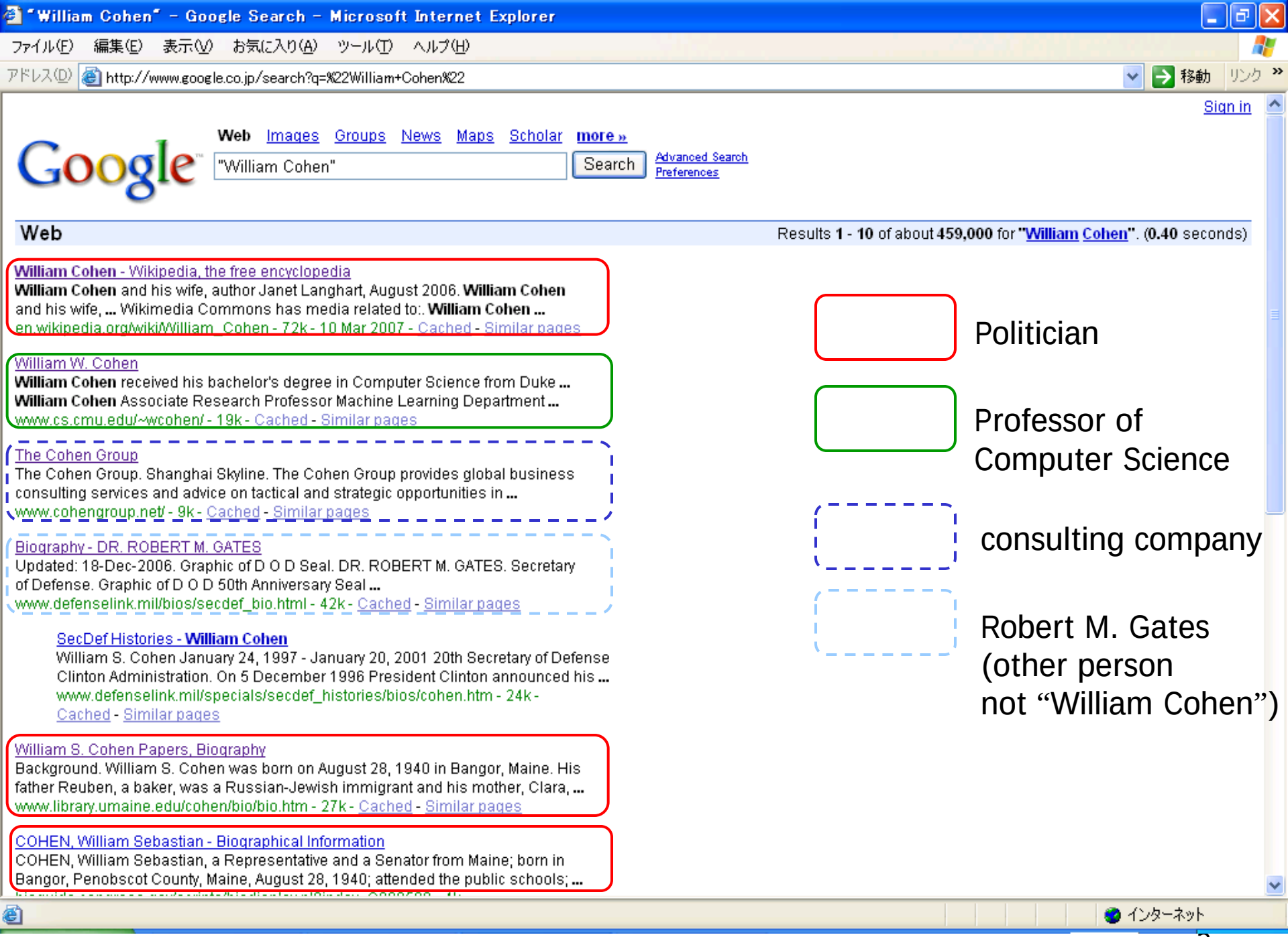
I. Personal Name Disambiguation in Web Search Results

[Outline]

1. Introduction
2. Related Work
3. Our Proposed Method
4. Experiments
5. Conclusion
6. Future Work

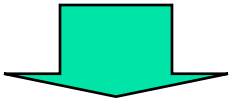
1. Introduction (1/2)

- The amount of information on the World Wide Web (WWW) continues to increase.
  - It has become increasingly difficult to find relevant information on the WWW.
- Web Search Engines
  - Useful tools to find relevant information on the WWW
  - However, they return a long list of search results.
    - Users must go through the titles or snippets of search results sequentially.
    - In the search results, multiple topics are mixed together.
      - Especially, if a user submits a personal name as a query to a search engine, Web pages about several namesakes are contained in the search results.


"William Cohen" - Google Search - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://www.google.co.jp/search?q=%22William+Cohen%22
移動 リンク
Sign in
Google
Web Images Groups News Maps Scholar more »
"William Cohen"
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 459,000 for "William Cohen". (0.40 seconds)
William Cohen - Wikipedia, the free encyclopedia
William Cohen and his wife, author Janet Langhart, August 2006. William Cohen and his wife, ... Wikimedia Commons has media related to:. William Cohen ...
en.wikipedia.org/wiki/William_Cohen - 72k - 10 Mar 2007 - Cached - Similar pages
William W. Cohen
William Cohen received his bachelor's degree in Computer Science from Duke ...
William Cohen Associate Research Professor Machine Learning Department ...
www.cs.cmu.edu/~wcohen/ - 19k - Cached - Similar pages
The Cohen Group
The Cohen Group. Shanghai Skyline. The Cohen Group provides global business consulting services and advice on tactical and strategic opportunities in ...
www.cohengroup.net/ - 9k - Cached - Similar pages
Biography - DR. ROBERT M. GATES
Updated: 18-Dec-2006. Graphic of D O D Seal. DR. ROBERT M. GATES. Secretary of Defense. Graphic of D O D 50th Anniversary Seal ...
www.defenselink.mil/bios/secdef_bio.html - 42k - Cached - Similar pages
SecDef Histories - William Cohen
William S. Cohen January 24, 1997 - January 20, 2001 20th Secretary of Defense Clinton Administration. On 5 December 1996 President Clinton announced his ...
www.defenselink.mil/specials/secdef_histories/bios/cohen.htm - 24k - Cached - Similar pages
William S. Cohen Papers, Biography
Background. William S. Cohen was born on August 28, 1940 in Bangor, Maine. His father Reuben, a baker, was a Russian-Jewish immigrant and his mother, Clara, ...
www.library.umaine.edu/cohen/bio/bio.htm - 27k - Cached - Similar pages
COHEN, William Sebastian - Biographical Information
COHEN, William Sebastian, a Representative and a Senator from Maine; born in Bangor, Penobscot County, Maine, August 28, 1940; attended the public schools; ...
Politician
Professor of Computer Science
consulting company
Robert M. Gates (other person not "William Cohen")
インターネット

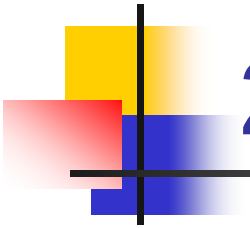
1. Introduction (2/2)

- Most of the previous works on disambiguating personal names in Web search results
  - Employ several types of unsupervised agglomerative clustering approaches
    - It is difficult to guide the clustering process appropriately.



Therefore, if some Web pages (seed page) that describe the entity of a person are introduced in a semi-supervised manner, the clustering for personal name disambiguation would be much more accurate.

→ We apply "*semi-supervised clustering*" to personal name disambiguation in Web search results.



2. Related Work

2.1 Personal Name Disambiguation in Web Search Results Based on Agglomerative Clustering

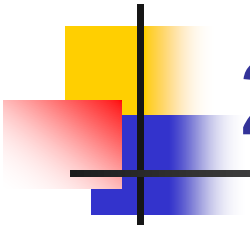
2.2 Semi-supervised Clustering

2.1 Personal Name Disambiguation in Web Search Results Based on Agglomerative Clustering

- [Mann and Yarowsky, CoNLL'03]
  - Extract biographical information such as birthdates, birthplaces, occupations and so on and generate feature vector combining TF-IDF value, then clustering is performed
  - "*WebHawk*" is a system that employ similar approaches.

    [Wan et al., CIKM'05]
- [Pedersen et al., CICLing'05]
  - Extract the context of each instance of an ambiguous name and generate second-order context vectors by using significant bi-grams, then the vectors are clustered such that similar instances are grouped into the same clusters
- [Bekkerman and McCallum, WWW'05]
  - Employ agglomerative/conglomerative double clustering (A/CDC)
  - Assume social network of people or lists of people to be disambiguated are previously known.
- [Bollegala et al., CICLing'06]
  - Cluster a target set of documents to ex
  - Also extract key phrases from the targe
  - Merge the clusters according to the similarity between (a) and (b)

These works cannot guide clustering process appropriately due to the lack of relevant clustering criteria.



2.2 Semi-supervised Clustering

2.2.1 Constraint-based Approach

2.2.2 Distance-based Approach

2.2.1 Constraint-based Approach

- Use label or constraints that user assign to guide clustering process appropriately
  - [Wagstaff and Cardie, ICML'00]

    [Wagstaff et al., ICML'01]
    - K-means clustering approach that introduce the following two constraints:
      - "must-link" (Two instances must be in the same cluster),
      - "cannot-link" (Two instances must be in different cluster).
  - [Basu et al., ICML'02]
    - Semi-supervised K-means clustering that uses labeled data to generate initial seed clusters and to guide the more appropriate clustering process.

2.2.2 Distance-based Approach

- A particular clustering measure is trained to satisfy the labels or constraints in the supervised data.
  - [Klein et al., ICML'02]
    - Euclidean distance is modified by a shortest path algorithm.
  - [Xing et al., NIPS'03]
    - Mahalanobis distance is trained using convex optimization.

Problems of Related Work

Personal Name Disambiguation
in Web Search Results
Based on Agglomerative Clustering

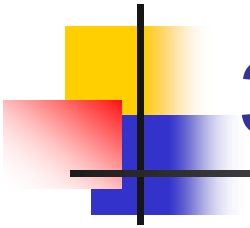
It is difficult to guide clustering process appropriately due to the lack of relevant clustering criteria.

Semi-supervised Clustering

- The purpose of existing methods is to improve *K*-means algorithm that needs to be set the number of clusters *K* in advance.
- Existing algorithms focus on introducing constraints and learning distances and overlook the fluctuation of the centroid of a cluster.

(a) Introducing constraints into seed pages

(b) It is important to control the fluctuation of the centroid of a cluster that contains a seed page when a certain cluster is merged into another cluster.



3. Our Proposed Method

3.1 Preparation

3.2 Our Proposed Semi-Supervised Clustering

3.1 Preparation

- Feature vector of search-result Web page *p*

$$\mathbf{w}^p = (w_{t_1}^p, w_{t_2}^p, \cdots, w_{t_m}^p)$$

m : the number of distinct terms

t_k : each term $(k = 1,2, \cdots, m)$

- Each term weight of $\mathbf{w}^p$
  - Adopt "*gain*" scheme after comparing "*TF*," "*IDF*," "*residual IDF*," "*TF-IDF*," "χ^I-*measure*," and "*gain*"

$$w_{t_k}^p = \frac{df(t_k)}{N}\left(\frac{df(t_k)}{N} - 1 - \log\frac{df(t_k)}{N}\right)$$

$df(t_k)$: the number of Web pages that term t_k appears

N : the total number of search-result Web pages

- The centroid vector of a cluster

$$\mathbf{G} = (g_{t_1}, g_{t_2}, \cdots, g_{t_m})$$

g_{t_k} : the weight of each term in the centroid vector of a cluster

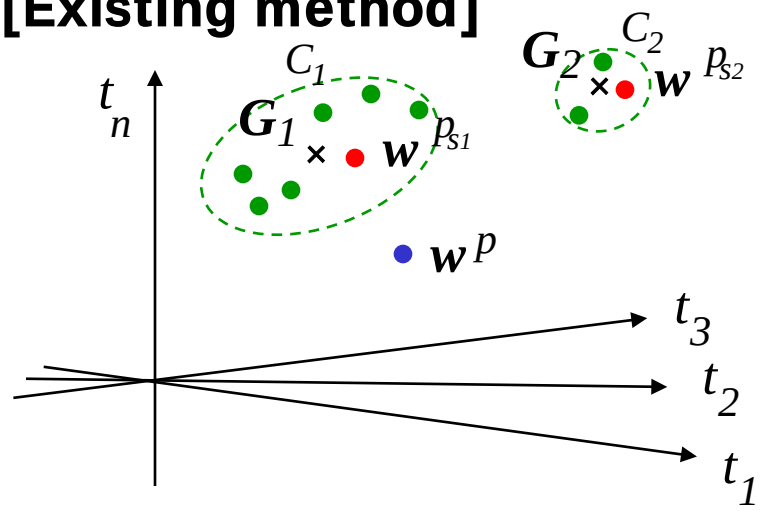
t_k : each term $(k = 1,2, \cdots, m)$

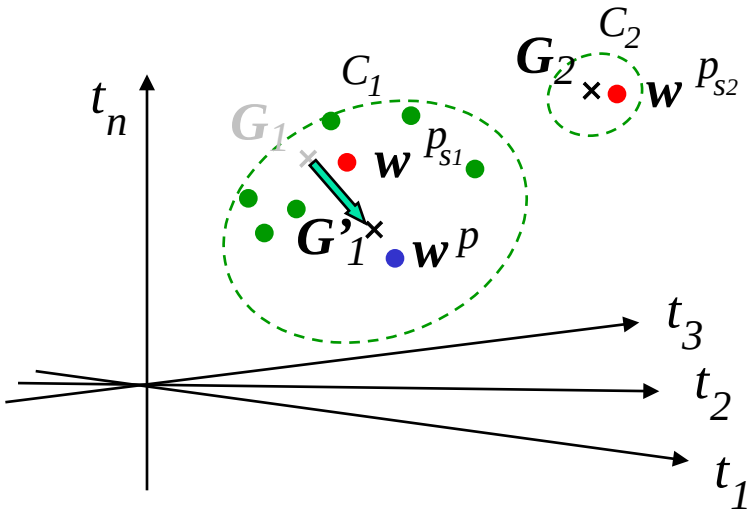
3.2 Our Proposed Semi-supervised Clustering (1/7)

- Introduce the following two types of seed page that describes a person
  - an article on each person in Wikipedia
  - the top-ranked Web page in the Web search results
- Control the fluctuation of the centroid of a cluster
  - Weight each element $\mathbf{w}^p$ by the distance between $\mathbf{G}$ and $\mathbf{w}^p$

3.2 Our proposed semi-supervised clustering (2/7)

[Existing method]





$$\mathbf{G}^{new} = \frac{\left(\sum_{\mathbf{w}^{p^{(\mathbf{G})}} \in \mathbf{G}}^{n} \mathbf{w}^{p^{(\mathbf{G})}} + \mathbf{w}^{p}\right)}{n_1 + n_2}$$

p : search-result Web page

p_{s_i} : seed page

$\boldsymbol{w}^{p}$: feature vector of p

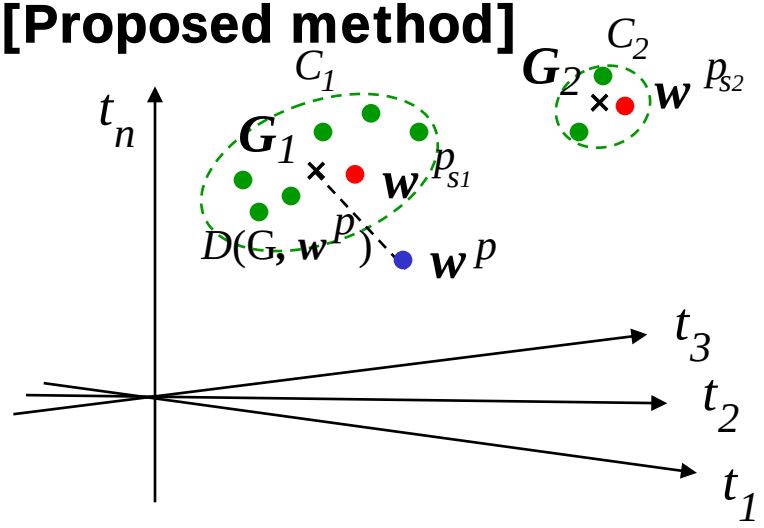
$\boldsymbol{G}_i$: the centroid vector of a cluster

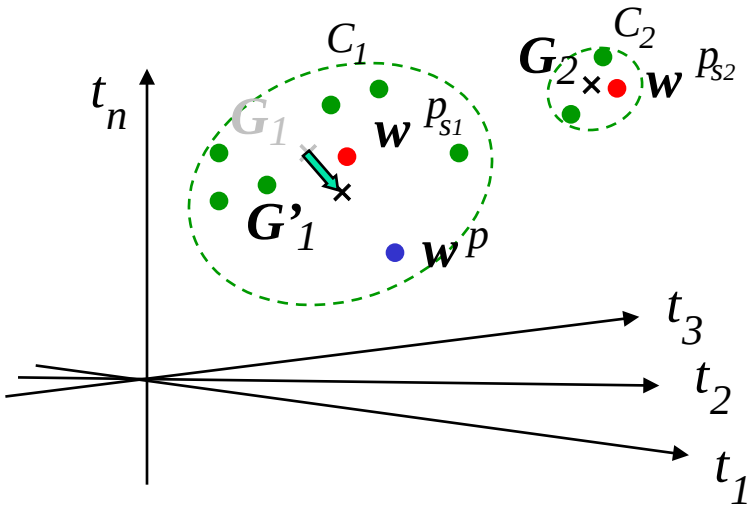
$\boldsymbol{w}^{p^{(G)}}$: feature vector of a Web page contained in a cluster that has centroid $\boldsymbol{G}$

The fluctuation of the centroid is large.

3.2 Our proposed semi-supervised clustering (3/7)

[Proposed method]





p : search-result Web page

p_{s_i} : seed page

$\mathbf{w}^p$: feature vector of p

$\mathbf{G}_i$: the centroid vector of a cluster

$\mathbf{w}^{p^{(G)}}$: feature vector of a Web page contained in a cluster that has centroid $\mathbf{G}$

$$\mathbf{G}^{new} = \frac{\left(\sum_{\mathbf{w}^{p^{(\mathbf{G})}} \in \mathbf{G}}^{n} \mathbf{w}^{p^{(\mathbf{G})}} + \frac{\mathbf{w}^p}{D(\mathbf{G}, \mathbf{w}^p)} \right)}{n_1 + n_2}$$

Control the fluctuation of the centroid of a cluster

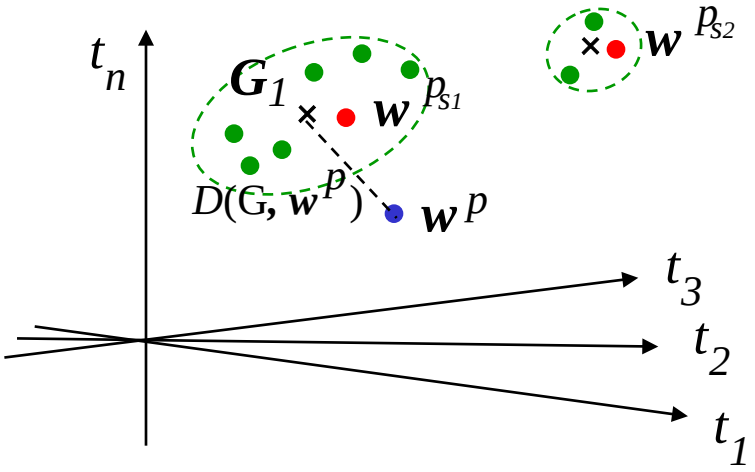
3.2 Our proposed semi-supervised clustering (4/7)

- Distance measure
  - (i) Euclidean distance
  - (ii) Mahalanobis distance
  - (iii) Adaptive Mahalanobis distance

3.2 Our proposed semi-supervised clustering (5/7)

(i) Euclidean distance

- the distance between 2 points

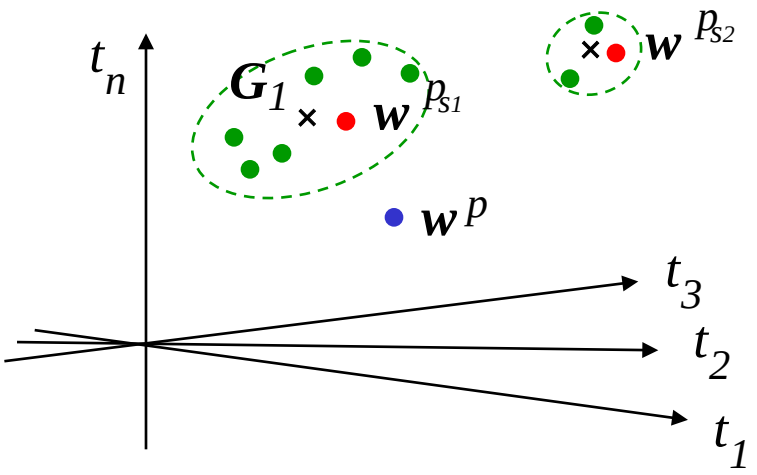


$$D(\mathbf{G}, \mathbf{w}^p) = \sqrt{\sum_{k=1}^{m} (g_{t_k} - w_{t_k}^p)^2}$$

3.2 Our proposed semi-supervised clustering (6/7)

(ii) Mahalanobis distance

- Distance between a group and a point
- Consider the correlations of the data set



$$D(\mathbf{G}, \mathbf{w}^p) = \sqrt{(\mathbf{w}^p - \mathbf{G})^T \sum^{-1} (\mathbf{w}^p - \mathbf{G})}$$

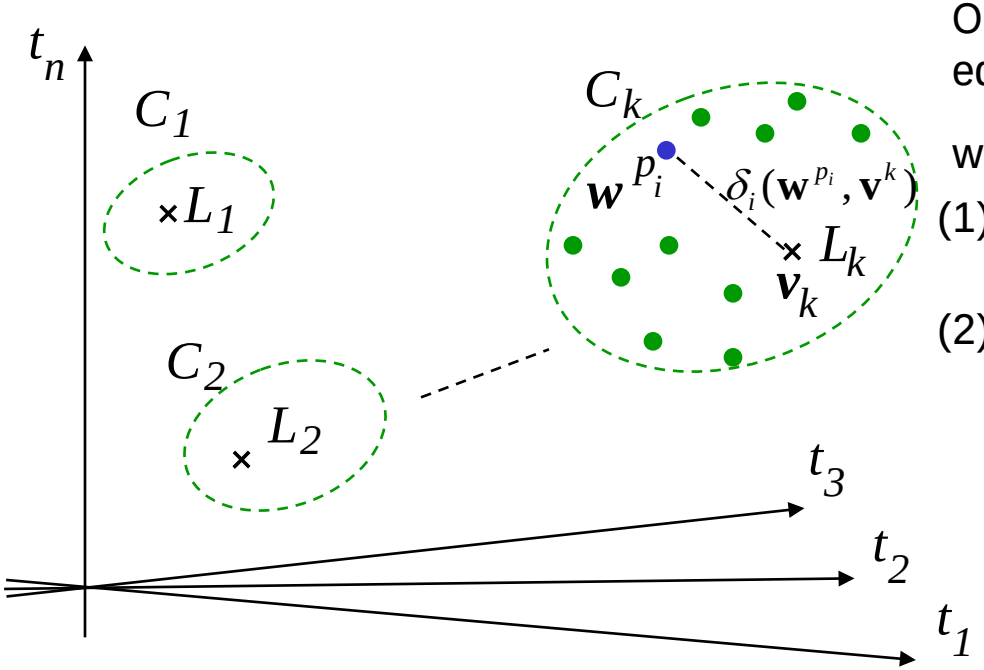
$\sum$: covariance matrix defined by the members of a cluster

3.2 Our proposed semi-supervised clustering (7/7)

(iii) Adaptive Mahalanobis distance

- Drawback of Mahalanobis distance
  - Covariance tends to be large when the number of documents in a cluster is small.

Optimize covariance matrix regarding each cluster



Optimize W that minimize the following equation: $W = \sum_{k=1}^{K} \Delta_k^2(L_k, \delta_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \delta_k(\mathbf{w}^{p_i}, \mathbf{v}^k)$

where $\delta_k(\mathbf{w}^{p_i}, \mathbf{v}^k) = (\mathbf{w}^{p_i} - \mathbf{v}^k)^T \mathbf{M}_k^{-1} (\mathbf{w}^{p_i} - \mathbf{v}^k)$ —(*)

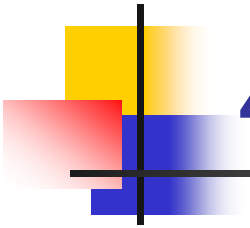
(1) Firstly, regarding cluster C_k, compute prototype L_k that locally minimize equation(*).

(2) Then fix the prototype computed in (1), and compute the distance δ_k in each cluster C_k that locally minimize $\Delta_k^2(L_k, \delta_k)$ with $\det(\mathbf{M}_k) = 1$

$\mathbf{M}_k = (\det \mathbf{Q}_k)^{1/p} \mathbf{Q}_k^{-1}$ ([Diday and Govaert, Infomatique Comp. Sci. '77])

$\mathbf{Q}_k$: Adapted covariance matrix of cluster C_k

$D(\mathbf{G}_k, \mathbf{w}^p) = \sqrt{(\mathbf{w}^p - \mathbf{G}_k)^T \mathbf{M}_k^{-1} (\mathbf{w}^p - \mathbf{G}_k)}$



4. Experiments

4.1 Experimental Data

4.2 Evaluation Measure

4.3 Experimental Results

4.1 Experimental Data

- WePS Corpus
  - Established for “Web People Search Task” at SemEval-2007 in the Association for Computational Linguistics (ACL) conference
    - Web pages related to 79 personal names
      - Sampled from
        - participants in conferences on digital libraries and computational linguistics,
        - bibliographic articles in the English Wikipedia,
        - the U.S. Census
      - The top 100 Yahoo search results via its search API for a personal name query
        - Training sets：49 names, Test sets：30 names（7,900 Web pages in total）
- Pre-processing for WePS corpus
  - Eliminate stopwords, and perform stemming
  - Determine the optimal parameter for merging similar clusters using the training set in the WePS corpus and then apply it to the test set in the WePS corpus

4.2 Evaluation Measure

(1) Purity

(2) Inverse purity

(3) *F* (harmonic mean of (1) and (2))

[Hotho et al., GLDV Journal'05]

(These are standard evaluation measures employed in the "Web People search task.")

4.3 Experimental Results

4.3.1 Experimental Results

Using Full Text in the Documents

4.3.2 Experimental Results

Using Fragments in the Documents

4.3.3 Comparison Our Results with Other Methods

4.3.1 Experimental Results Using Full Text in the Documents (1/4)

Purity: Large
Inverse purity: Small
The agglomerative clustering tends to generate clusters that contain only one document.

[Exp ne seed page]

		Purity	Inverse purity	F
Agglomerative clustering	(No seed page)	0.66	0.49	0.51
Semi-supervised clustering				
(i) Euclidean distance	(a) Wikipedia article	0.39	0.90	0.54
	(b) Top-ranked Web page	0.40	0.82	0.54
(ii) Mahalanobis distance	(a) Wikipedia article	0.44	0.96	0.55
	(b) Top-ranked Web page	0.47	0.81	0.60
(iii) Adaptive Mahalanobis distance	(a) Wikipedia article	0.48	0.88	**0.62**
	(b) Top-ranked Web page	0.50	0.78	*0.61*

4.3.1 Experimental Results Using Full Text in the Documents (2/4)

purity: Worse than agglomerative clustering
inverse purity: Better than agglomerative clustering
The effect of introducing a seed page,
and controlling the fluctuation of the centroid of a cluster.

[Exp

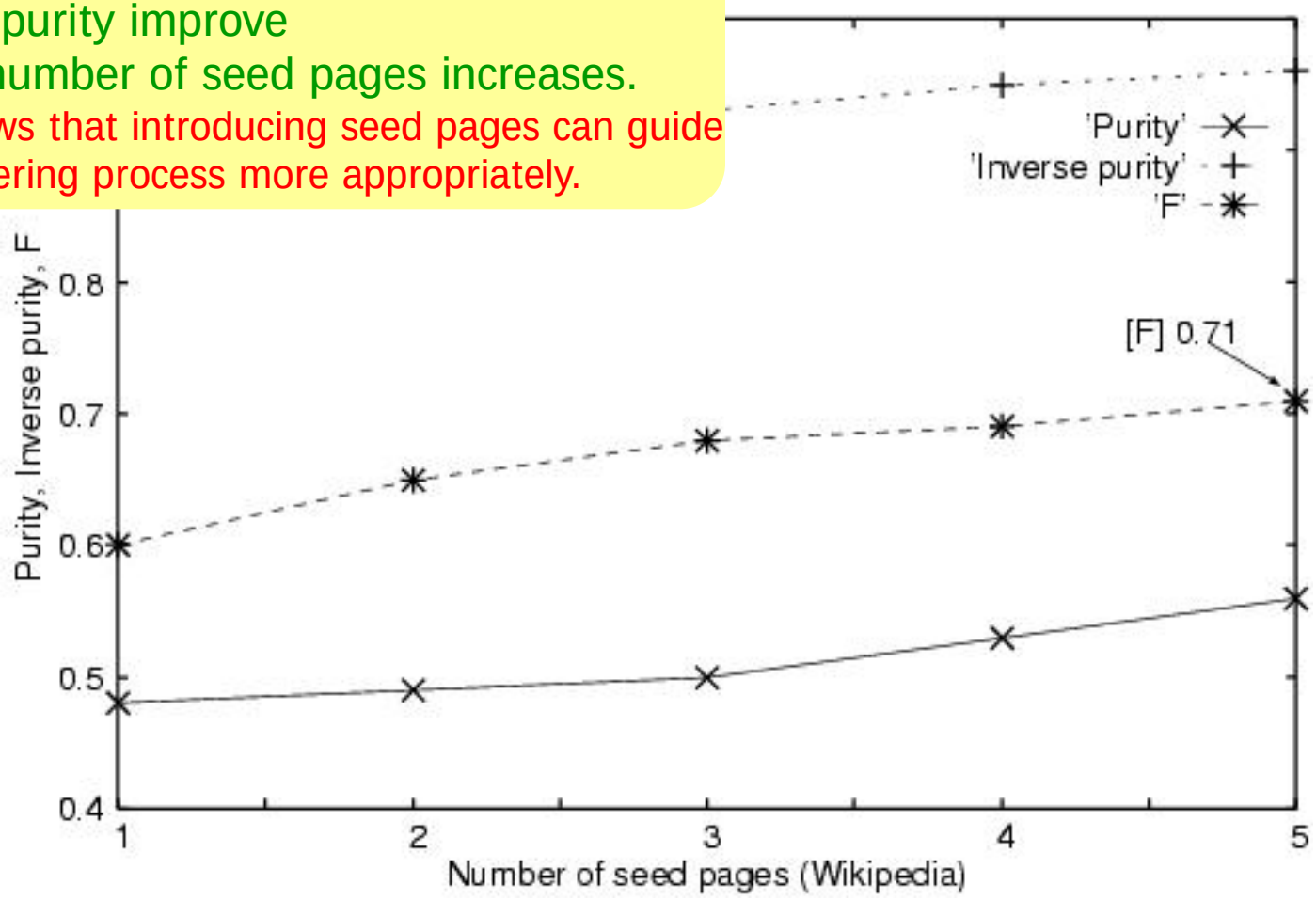
			inverse purity	F
Agglomerative clustering	(No seed page)	0.66	0.49	0.51
Semi-supervised clustering				
(i) Euclidean distance	(a) Wikipedia article	0.39	0.90	0.54
	(b) Top-ranked Web page	0.40	0.82	0.54
(ii) Mahalanobis distance	(a) Wikipedia article	0.44	0.96	0.55
	(b) Top-ranked Web page	0.47	0.81	0.60
(iii) Adaptive Mahalanobis distance	(a) Wikipedia article	0.48	0.88	**0.62**
	(b) Top-ranked Web page	0.50	0.78	*0.61*

4.3.1 Experimental Results Using Full Text in the Documents (3/4)

[Using multiple seed pages (seed page: Wikipedia article)]

The values of both purity and inverse purity improve as the number of seed pages increases.

This shows that introducing seed pages can guide the clustering process more appropriately.

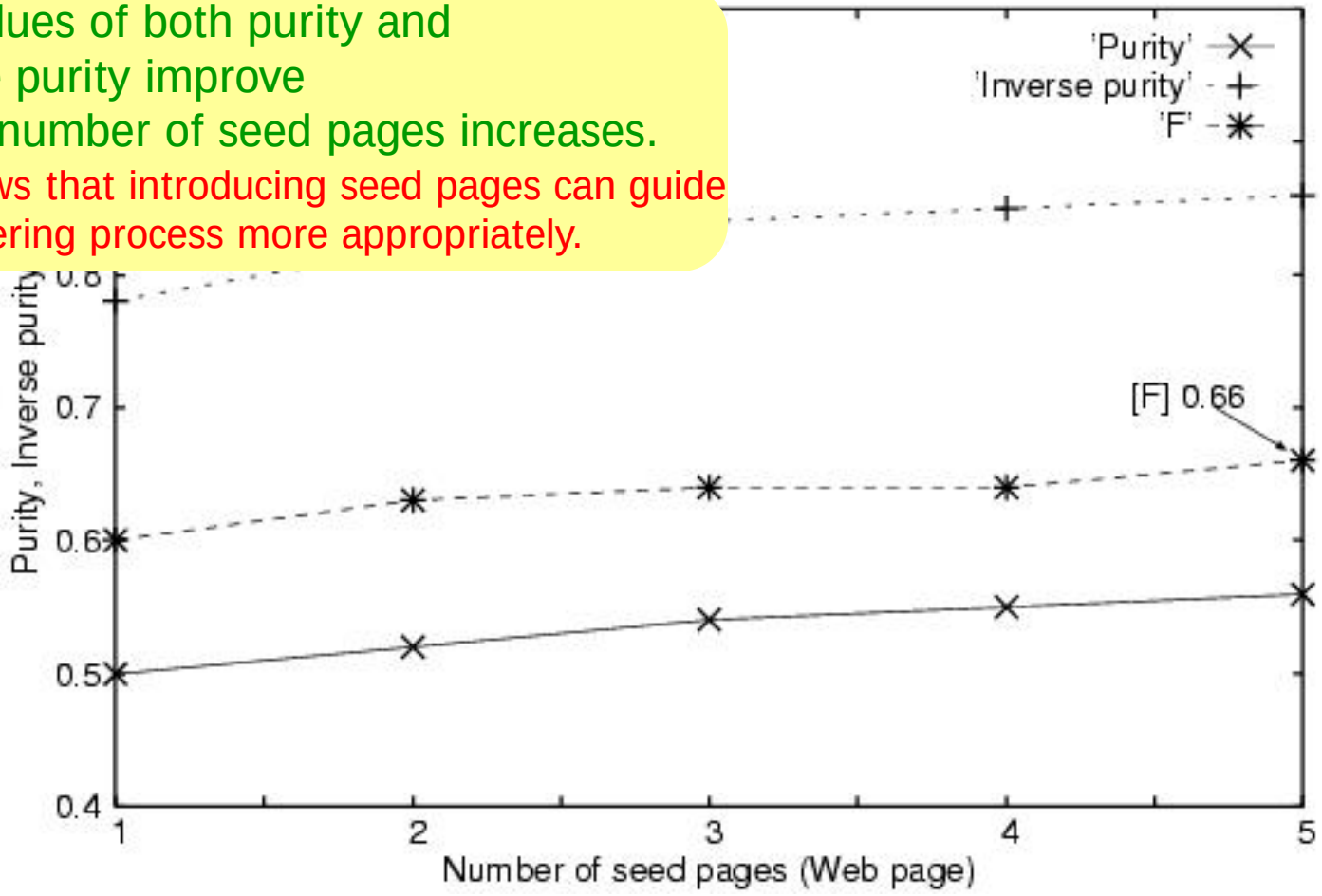


4.3.1 Experimental Results Using Full Text in the Documents (4/4)

[Using multiple seed pages
(seed page: Web pages ranked up to the top 5)]

The values of both purity and inverse purity improve as the number of seed pages increases.

This shows that introducing seed pages can guide the clustering process more appropriately.

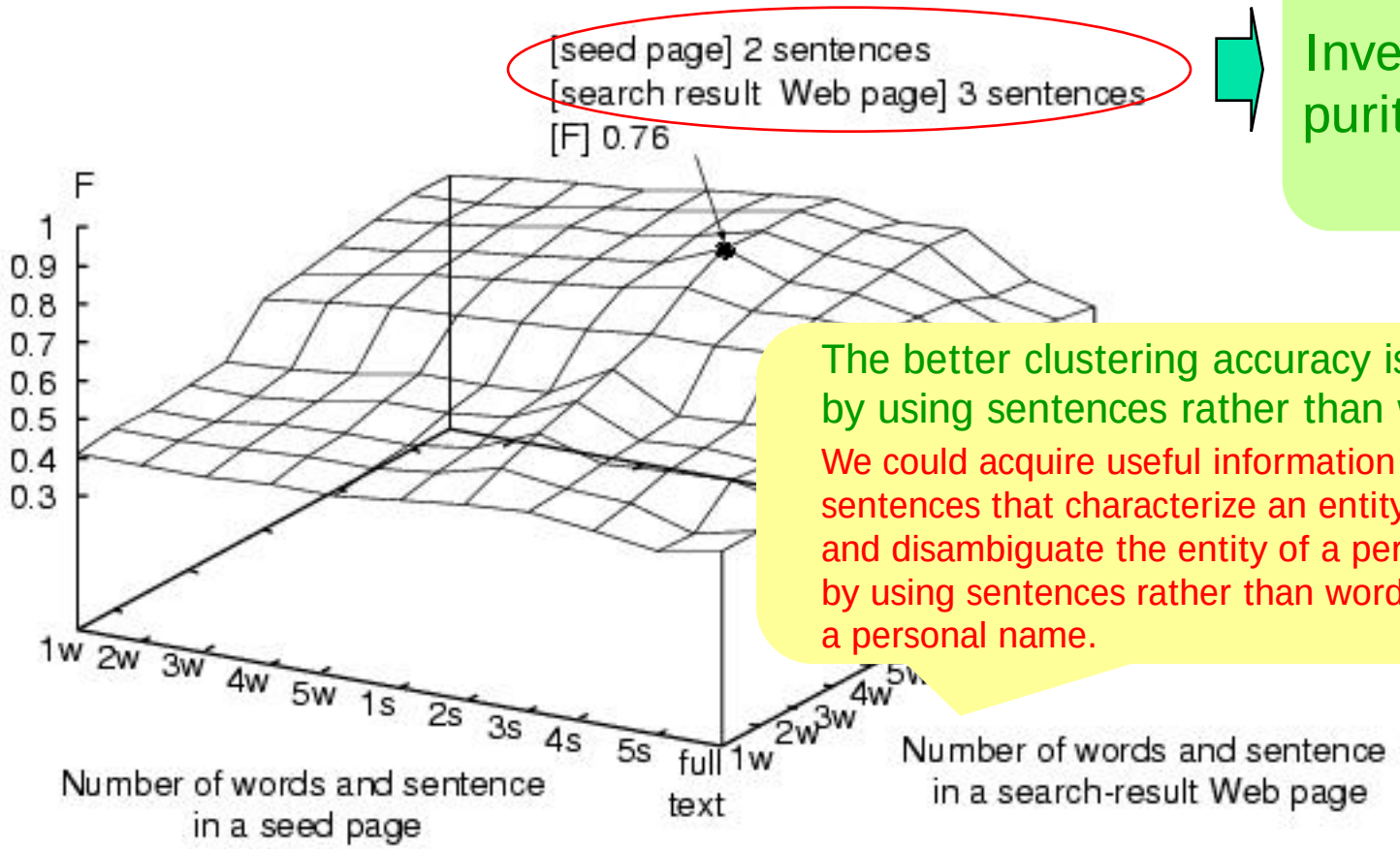


4.3.2 Experimental Results Using Fragments in the Documents (1/2)

- The words that characterize the person often appear around a personal name.
- We conducted the following experiments in the case where the best value of F (0.71) is obtained (in other words, 5 Wikipedia articles are used) 
  - Vary the number of words and sentences backward and forward from a personal name in seed pages and search-result Web pages using the training set of WePS corpus
  - Then apply these numbers of sentences around a personal name to the test set of WePS corpus

4.3.2 Experimental Results Using Fragments in the Documents (2/2)

• Search for the optimal number of words and sentences backward and forward from a personal name



Purity: 0.72
Inverse purity: 0.81
F: 0.76

The better clustering accuracy is obtained by using sentences rather than words.

We could acquire useful information from sentences that characterize an entity of a person and disambiguate the entity of a person effectively by using sentences rather than words around a personal name.

4.3.3 Comparison with Other Methods

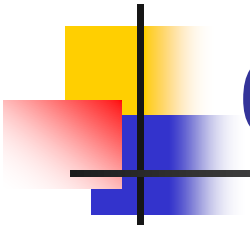
Team ID of "Web People Search Task"	Purity	Inverse Purity	F
CU_COMSEM	0.72	0.88	**0.78**
IRST-BP	0.75	0.80	0.75
PSNUS	0.73	0.82	0.75
UVA	0.81	0.60	0.67
SHEF	0.60	0.82	0.66
Our proposed method			
2 and 3 sentences in 5 Wikipedia seed pages and search result Web page, redpectively	0.72	0.81	**0.76**

This result is comparable to the top results (0.78) among the top 5 participants in "Web People Search Task."

We could acquire useful information from sentences that characterize an entity of a person and disambiguate the entity of a person effectively.

5. Conclusion

- We proposed a semi-supervised clustering approach to disambiguate personal names in Web search results.
  - Control the fluctuation of the centroid of a cluster that contains a seed page
    - The adaptive Mahalanobis distance is effective.
    - The better accuracy is obtained compared with agglomerative clustering.
  - The clustering accuracy is further improved by using fragments in the documents:
    - 2 sentences backward and forward from an ambiguous name in a seed page,
    - 3 sentences backward and forward from an ambiguous name in a search-result Web page



6. Future Work

We plan to:

- Use Web pages hyperlinked from a target page to disambiguate personal names in Web search results,
- Extend our approach to disambiguate geographical names.

Recent Works on Personal Name Disambiguation in Web Search Results

[1] Javier Artiles, Satoshi Sekine, Julio Gonzalo:
"Web people search: results of the first evaluation and the plan for the second" (WWW'08 poster)
→ WePS2 (2nd Web People Search Evaluation Workshop Co-located with the WWW2009 conference, Apr 21st, http://nlp.uned.es/weps/)

[2] D. V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra:
"Towards breaking the quality curse.: a web-querying approach to Web People Search" (SIGIR'08)

[3] D. V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra:
"Resolving Personal Names in Email Using Context Expansion"(ACL-08:HLT)

[4] O. Popescu and B. Magnini:
"Alleviating the Problem of Coreferences in Web Person Search" (CICLing'09)

[5] T. Pedersen:
"Improved Unsupervised Name Discrimination with Very Wide Bigrams and Automatic Cluster Stopping" (CICLing'09)

II. Word Sense Disambiguation in Japanese Texts

[Outline]

1. Introduction
2. Related Work
3. Proposed Method
4. Experiments
5. Conclusion
6. Future Work

1. Introduction (1/2)

- Word Sense Disambiguation (WSD)
  - Determining the meaning of ambiguous word in its context

    "run":

    (1) Bob goes running every morning.

    "to move fast by using one's feet"

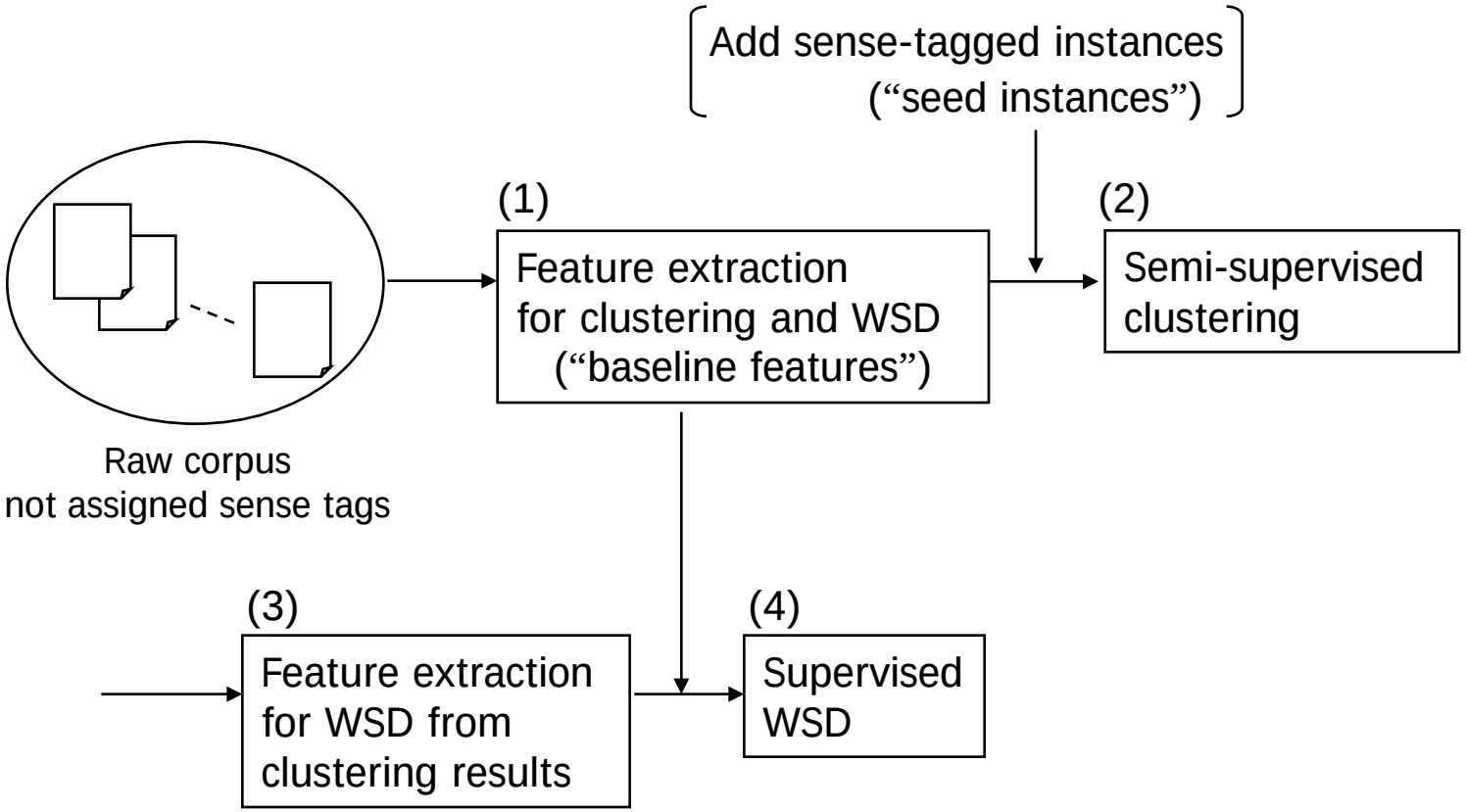
    (2) Mary runs a beauty parlor.

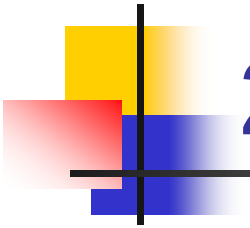
    "to direct or control"

1. Introduction (2/2)

- Our approach for WSD
  - Basically, supervised WSD
  - Applying semi-supervised clustering

    by introducing sense-tagged instances to supervised WSD





2. Related Work

2.1 Related Work on Semi-supervised Clustering

2.2 Related Work on Word Sense Disambiguation

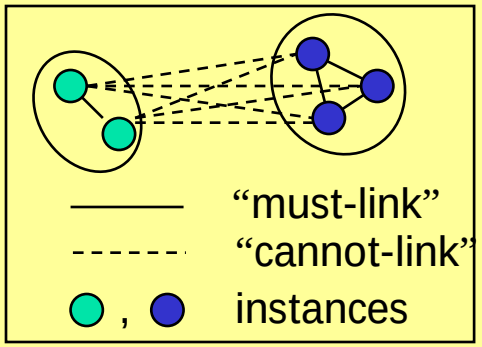
2.1 Related Work on Semi-supervised Clustering

Semi-supervised Clustering

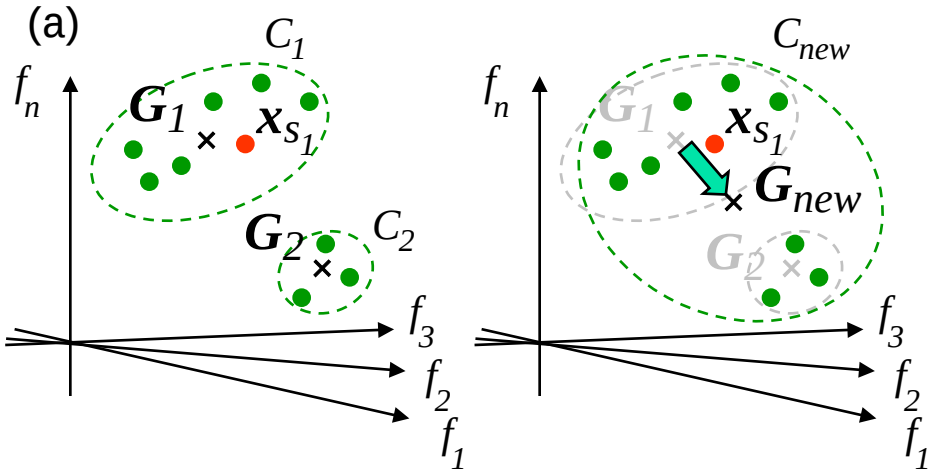
(1) Constraint-based Approach
[Wagstaff and Cardie, ICML'01], [Wagstaff et al., ICML'01], [Basu et al., ICML'02]

(2) Distance-based Approach
[Klein et al., ICML'02], [Xing et al, NIPS'03], [Bar-Hillel et al., ICML'02]

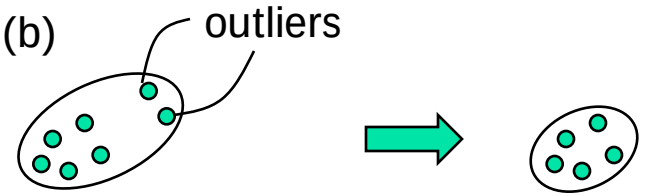
[Constraint-based Approach]



(a) Existing algorithms focus on introducing constraints or learning distances and overlook the fluctuation of the centroid of a cluster.



(b) Especially, word instances with the same sense may be distant each other.



2.2 Related Work on Word Sense Disambiguation (WSD)

WSD systems

(1) Adding features in order to improve WSD accuracy

[Agirre et al., OL'00], [Specia et al., ACL'07], [Cai et al., EMNLP'07]

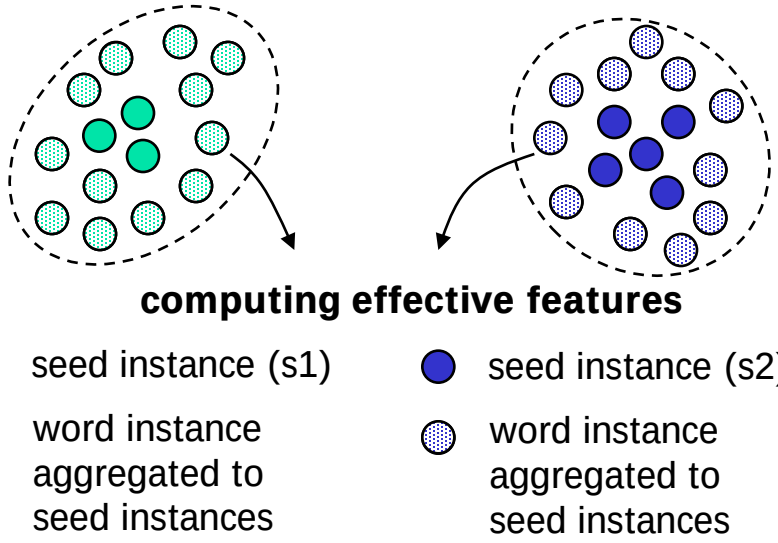
(2) Alleviate sparseness of feature space

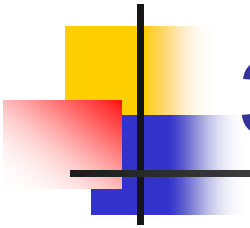
[Niu et al., HLT/EMNLP'05]

(3) Recent works on WSD

[Davidov et al., ACL'08:HLT], [Zhong et al., EMNLP'08]

If we introduce sense-tagged seed instances and perform clustering, we can directly compute effective features from word instances clustered to seed instances.





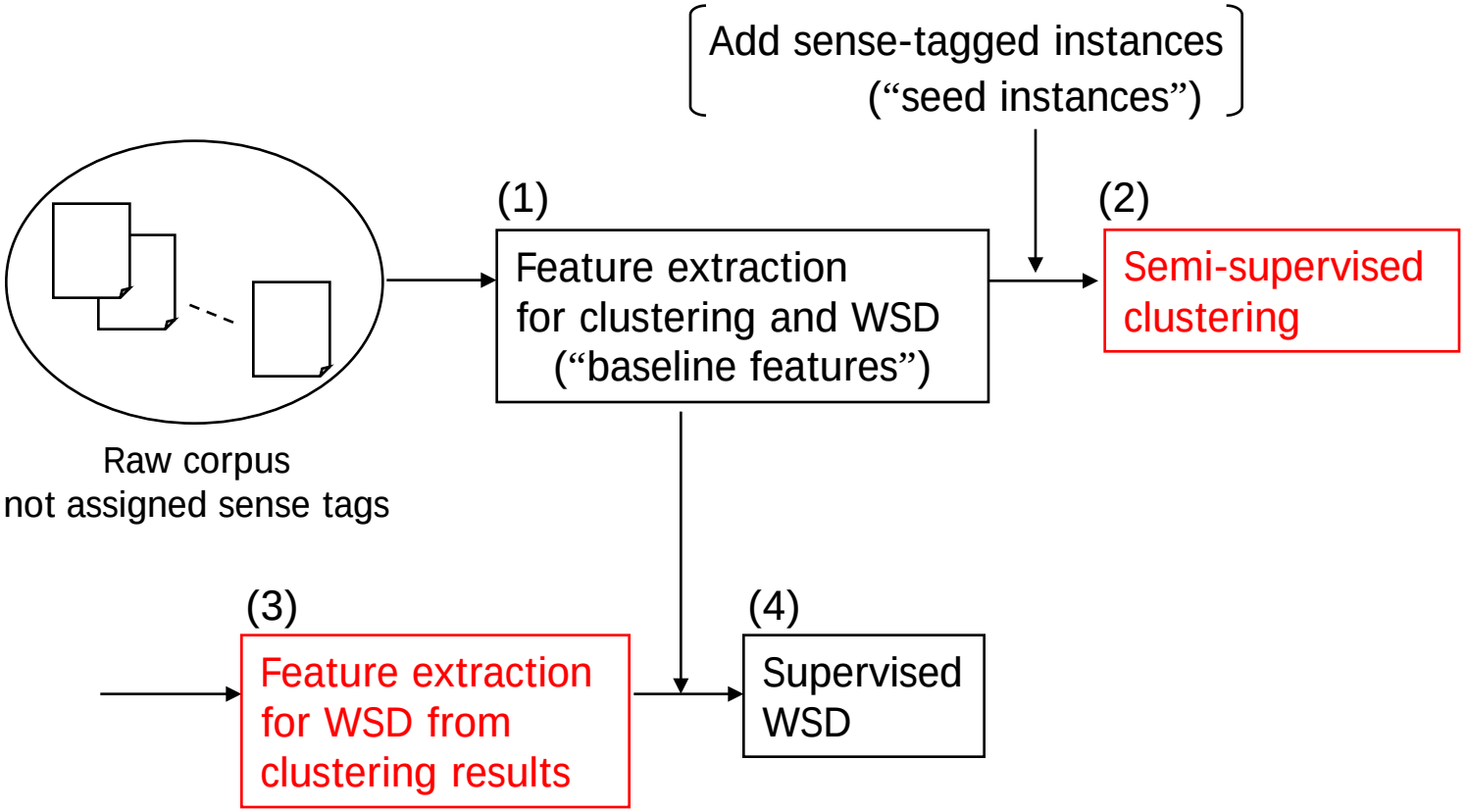
3. Proposed Method

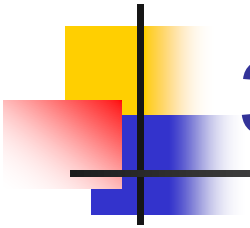
3.1 System Architecture

3.2 Semi-supervised Clustering

3.3 Features for WSD
obtained Using Clustering Results

3.1 System Architecture





3.2 Semi-supervised Clustering

3.2.1 Features for Clustering

3.2.2 Semi-supervised Clustering

3.2.3 Seed Instances and Constraints for Clustering

3.2.1 Features for Clustering and WSD ("baseline features")

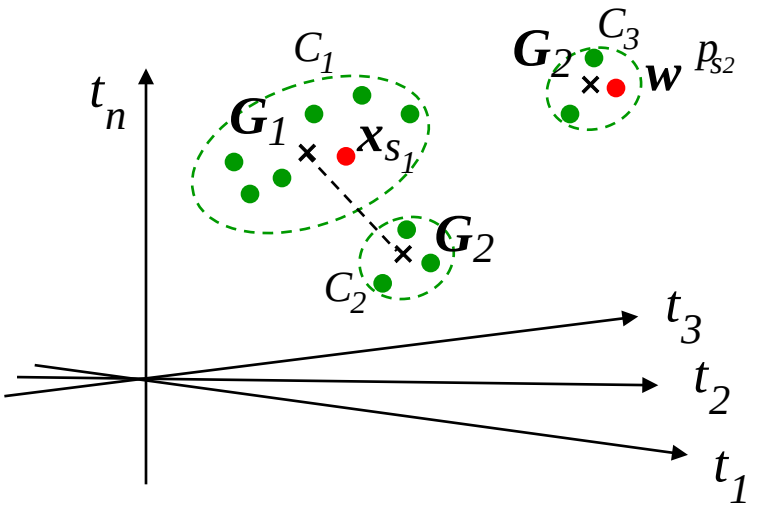
- Morphological features
  - Bag-of-words (BOW), Part-of-speech (POS), and detailed POS classification
    - Target word itself and the two words to its right and left.
- Syntactic features
  - If the POS of a target word is a noun, extract the verb in a grammatical dependency relation with the noun.
  - If the POS of a target word is a verb, extract the noun in a grammatical dependency relation with the verb.
- Figures in Bunrui-Goi-Hyou
  - 4 and 5 digits regarding the content word to the right and left of the target word.
    - 地域 ("community") 社会 ("society")

      "1.1720,4,1,3" → 1172 (as 4 digits), 11720 (as 5digits)
- 5 topics inferred on the basis of LDA
  - Compute the log-likelihood of a word instance ("soft-tag" approach [Cai et al., EMNLP'07])

3.2.2 Semi-supervised Clustering

[Proposed method]

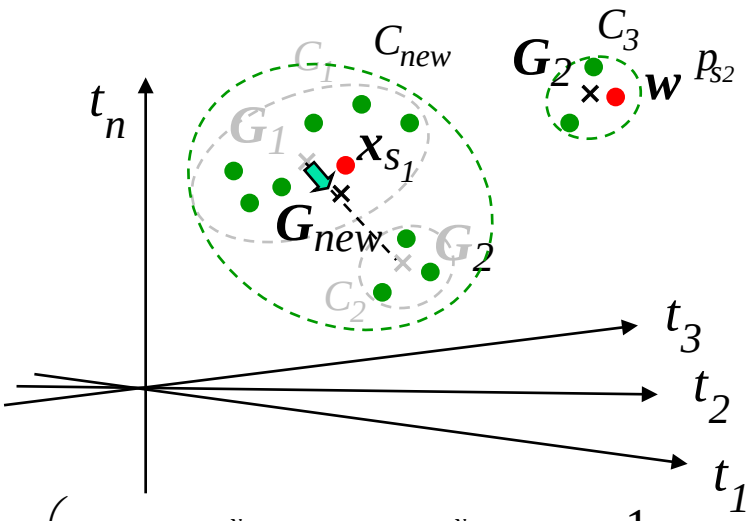
Refine [Sugiyama and Okumura, ICADL'07] (about Web People Search Task) for word instances



x : word instance

x_{s_i} : seed instance

$\boldsymbol{G}_i$: the centroid of a cluster



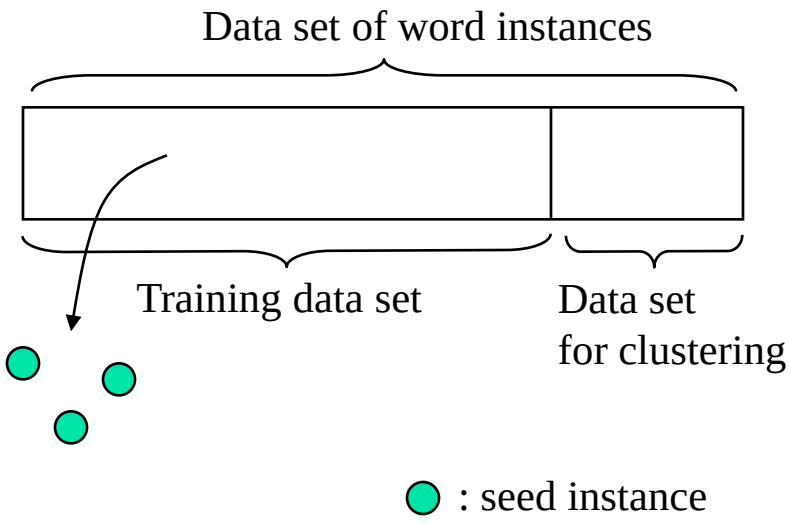
$$\mathbf{G}^{new} = \frac{\left(\sum_{\mathbf{f}^x \in C_1}^{n} \mathbf{f}^x + \sum_{\mathbf{f}^x \in C_2}^{n} \mathbf{f}^x \frac{1}{D(\mathbf{G}^{C_1}, \mathbf{G}^{C_2}) + c} \right)}{n_1 + n_2 \times \frac{1}{D(\mathbf{G}^{C_1}, \mathbf{G}^{C_2}) + c}}$$

$D(\mathbf{G}^{C_1}, \mathbf{G}^{C_2})$: adaptive Mahalanobis distance

Control the fluctuation of the centroid of a cluster

3.2.3 Seed Instances and Constraints for Clustering (1/4)

[Method I]



Select initial seed instances:

(I-1) randomly,

(I-2) “KKZ”

(I-3) centroid of a cluster generated by K-means

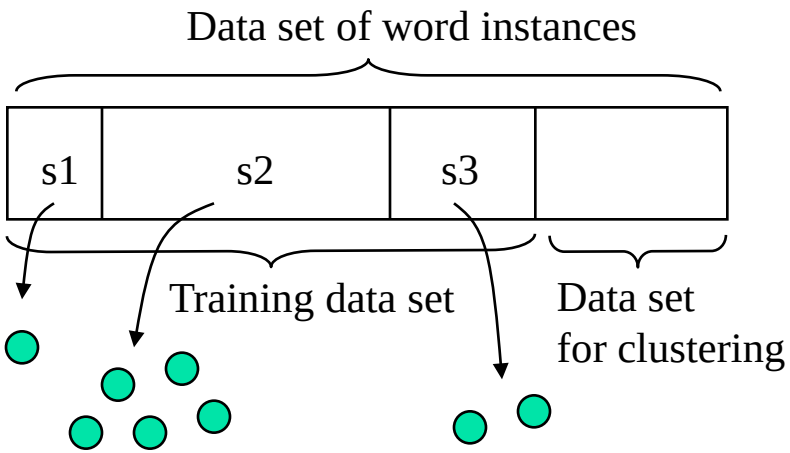
initial instances for K-means

* randomly,

* “KKZ” [Katsavounidis et al., IEEE Signal Processing Letters, ‘94]

3.2.3 Seed Instances and Constraints for Clustering (2/4)

[Method II]



s1, s2, s3: word senses of target word

: seed instance

Select initial seed instances:

(II-1) By considering the frequency of word senses

(II-1-1) randomly,

(II-1-2) “KKZ”

(II-1-3) centroid of a cluster generated by K-means

initial instances for K-means

* randomly,

* “KKZ”

(II-2) In proportion to the frequency of word senses (D’Hondt method)

(II-2-1) randomly,

(II-2-2) “KKZ”

(II-2-3) centroid of a cluster generated by K-means

initial instances for K-means

* randomly,

* “KKZ”

3.2.3 Seed Instances and Constraints for Clustering (3/4)

[Applying D'Hondt method to WSD]

[D'Hondt method]

	Party A (1600)	Party B (700)	Party C (300)
seat 1 (/1)	1600 (1)	700 (3)	300 (8)
seat 2 (/2)	800 (2)	350 (6)	150
seat 3 (/3)	533 (4)	233 (9)	
seat 4 (/4)	400 (5)	175 (10)	
seat 5 (/5)	320 (7)		

Selecting word instances using D'Hondt method

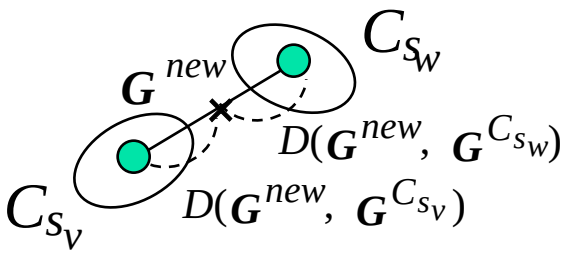
	s2 (50)	s1 (20)	s3 (15)
seed 1 (/1)	50 (1)	20 (3)	15 (5)
seed 2 (/2)	25 (2)	10 (8)	8 (9)
seed 3 (/3)	17 (4)	7 (10)	3
seed 4 (/4)	13 (6)		
seed 5 (/5)	10 (7)		

* s1, s2, s3: word senses

3.2.3 Seed Instances and Constraints for Clustering (4/4)

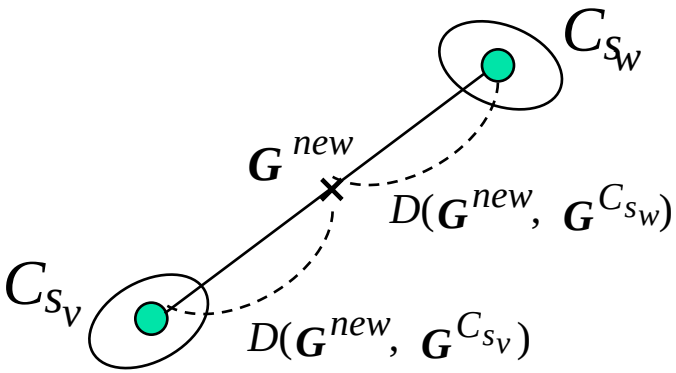
- Constraints
  - "cannot-link" only,
  - "must-link" only,
  - Both constraints,
  - "cannot-link" and "must-link" without outliers

["must-link" without outliers]



$$D(G^{new}, G^{C_{s_v}}) < Th_{dis},$$
$$D(G^{new}, G^{C_{s_w}}) < Th_{dis}$$

Put "must-link" constraints between C_{s_v} and C_{s_w}



$$D(G^{new}, G^{C_{s_v}}) > Th_{dis},$$
$$D(G^{new}, G^{C_{s_w}}) > Th_{dis}$$

C_{s_w} is an outlier,
so a "must-link" constraint is not added.
$(Th_{dis} = 0.388)$

3.3 Features for WSD Obtained Using Clustering Results

(a) Inter-cluster information

- TF in a cluster (TF)
- Cluster ID (CID)
- Sense frequency (SF)

(b) Context information regarding adjacent words

$w_i w_{i+1}\ (i = -2,...,1)$

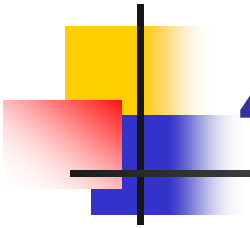
- Mutual information (MI)
- T-score (T)
- χ^2 (CHI2)

(c) Context information regarding two words to the right and left of the target word

$w_{-2} w_{-1} w_0 w_{+1} w_{+2}$

- Information gain (IG)

*We employ (b) and (c) to reflect the concept of *"one sense per collocation."* [Yarowsky, ACL'95]



4. Experiments

4.1 Experimental Data

4.2 Semi-supervised Clustering

4.3 Word Sense Disambiguation

4.1 Experimental Data

- RWC corpus from

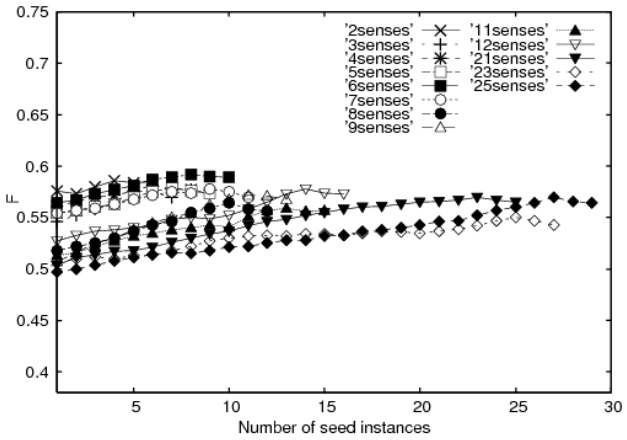
  the “SENSEVAL-2 Japanese Dictionary Task”

  - 3000 Japanese newspaper articles issued in 1994
    - Sense tags in Japanese Dictionary, “Iwanami Kokugo Jiten” were manually assigned to 148,558 ambiguous words.
  - 100 target words
    - 50 nouns, and 50 verbs

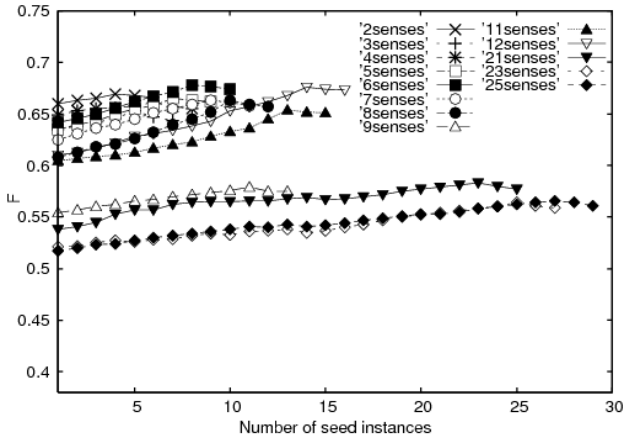
4.2 Semi-supervised Clustering (1/2)

- Experimental Results

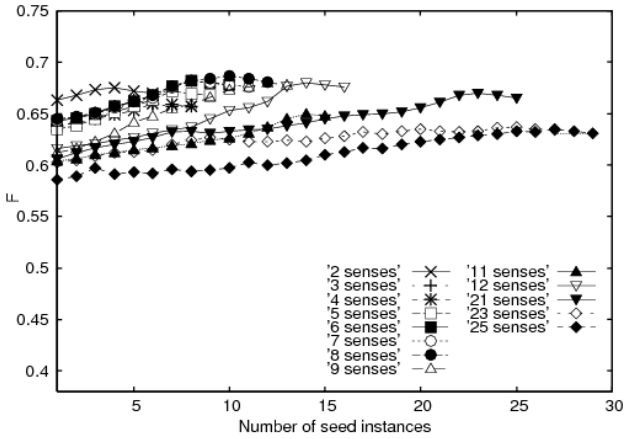
(I-3KKZ)



(II-1-3rnd)



(II-2-1)



[Observations]

- Method I (I-3KKZ)
  The representative instances tend to be selected by *K*-means after selecting distant initial instances with KKZ.
- Method II
  Selecting seed instances in proportion to the frequency of word senses (II-2) is better than considering frequency of word senses (II-1).
- Other findings
  In most cases, the best clustering accuracy is obtained when two more seed instances are added to the original number of word senses.

⇒ New senses can be discovered.

4.2 Semi-supervised Clustering (2/2)

- Comparison of distance-based approaches

	Method I	Method (II-1)	Method (II-2)
Proposed method	**0.543** (I-3KKZ)	**0.592** (II-1-3rnd)	**0.646** (II-2-1)
Distance-based approaches: Bar-Hillel et al. [1]	0.516	0.570	0.608
Distance-based approaches: Xing et al. [9]	0.494	0.539	0.591
Distance-based approaches: Klein et al. [5]	0.448	0.504	0.570
Fixed centroid	0.385	0.402	0.514
Agglomerative clustering	0.380	0.389	0.471

[Observations]

- Our semi-supervised clustering approach outperforms other distance-based approaches.
  - Our method locally adjusts the centroid of a cluster

4.3 Word Sense Disambiguation

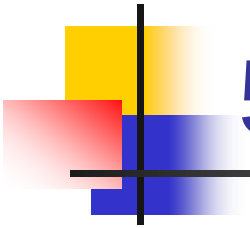
- Experimental Results

Features	SVM	NB	ME	Features	SVM	NB	ME
OURS (not clustered)	0.663	0.667	0.662	CRL (not clustered)	0.775	0.778	0.773
OURS + MI (not clustered)	0.666	0.669	0.664	CRL + MI (not clustered)	0.776	0.780	0.775
OURS + CID + MI + IG	**0.780**	**0.782**	**0.779**	CRL + CID + MI + IG	**0.778**	**0.783**	**0.780**
OURS + CID + T + IG	0.768	0.777	0.764	CRL + CID + T + IG	0.778	0.779	0.777
OURS + CID + CHI2 + IG	0.762	0.765	0.757	CRL + CID + CHI2 + IG	0.776	0.779	0.775
TITECH (not clustered)	0.661	0.663	0.660	NAIST (not clustered)	0.745	0.747	0.743
TITECH + MI (not clustered)	0.663	0.665	0.662	NAIST + MI (not clustered)	0.747	0.748	0.745
TITECH + CID + MI + IG	**0.767**	**0.770**	**0.764**	NAIST + CID + MI + IG	**0.765**	**0.767**	**0.764**
TITECH + CID + T + IG	0.765	0.767	0.759	NAIST + CID + T + IG	0.756	0.760	0.755
TITECH + CID + CHI2 + IG	0.756	0.759	0.751	NAIST + CID + CHI2 + IG	0.752	0.754	0.747

[Observations]

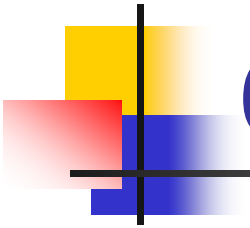
• The best accuracy is obtained, when we add features from clustering results, CID, MI, and IG to the baseline features.

• According to the results of OURS, TITECH, NAIST, WSD accuracy is significantly improved by adding features computed from clustering results.



5. Conclusion

- We proposed a semi-supervised clustering for word instances and its effect on WSD.
  - The best way of selecting seed instances and constraints for semi-supervised clustering
    - Selecting word sense with D'Hondt method and randomly selecting seed instances from ones that belong to the word sense.
  - Effective features for WSD obtained from clustering results
    - CID, MI, IG



6. Future Work

We plan to:

- Develop a much more accurate semi-supervised clustering approach,
- Look for features that can lead to higher accuracy for WSD.

Recent Works on Word Sense Disambiguation

[1] S. Brody and M. Lapata:

"Good Neighbors Make Good Senses: Exploiting Distributional Similarity for Unsupervised WSD" (Coling'08)

→ The title has "unsupervised." But I think supervised WSD rather than unsupervised WSD because unsupervised process is introduced to obtain enough training examples.

[2] R. Guzman-Cabrera, P. Rosso, M. Montes-y-Gomez, L. Villasenor-Pineda, and D. Pinto-Avendano:

"Semi-supervised Word Sense Disambiguation Using the Web as Corpus" (CICLing'09)

→ This work is similar to "topic signatures." [Agirre et al., Ontology Learning'00]

[3] R. Izquierdo, A. Suarez, and G. Rigau:

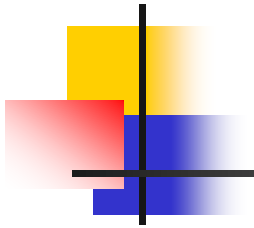
"An Empirical Study on Class-based Word Sense Disambiguation" (EACL'09)

→ The senses in WordNet for WSD are very fine-grained or very coarse-grained. Therefore, this work select appropriate level of sense.

[4] E. Agirre and O. Lopez de Lacalle:

"Supervised Domain Adaptation for WSD" (EACL'09)

→ Their domain adaptation approach allows to obtain high WSD accuracy for domain-related words.



Thank you very much!