Unsupervised Translation Sense Clustering

Mohit Bansal
UC Berkeley

John DeNero
Google

Dekang Lin
Google

Motivation

Manual curation

Lexicographers

Bilingual dictionaries

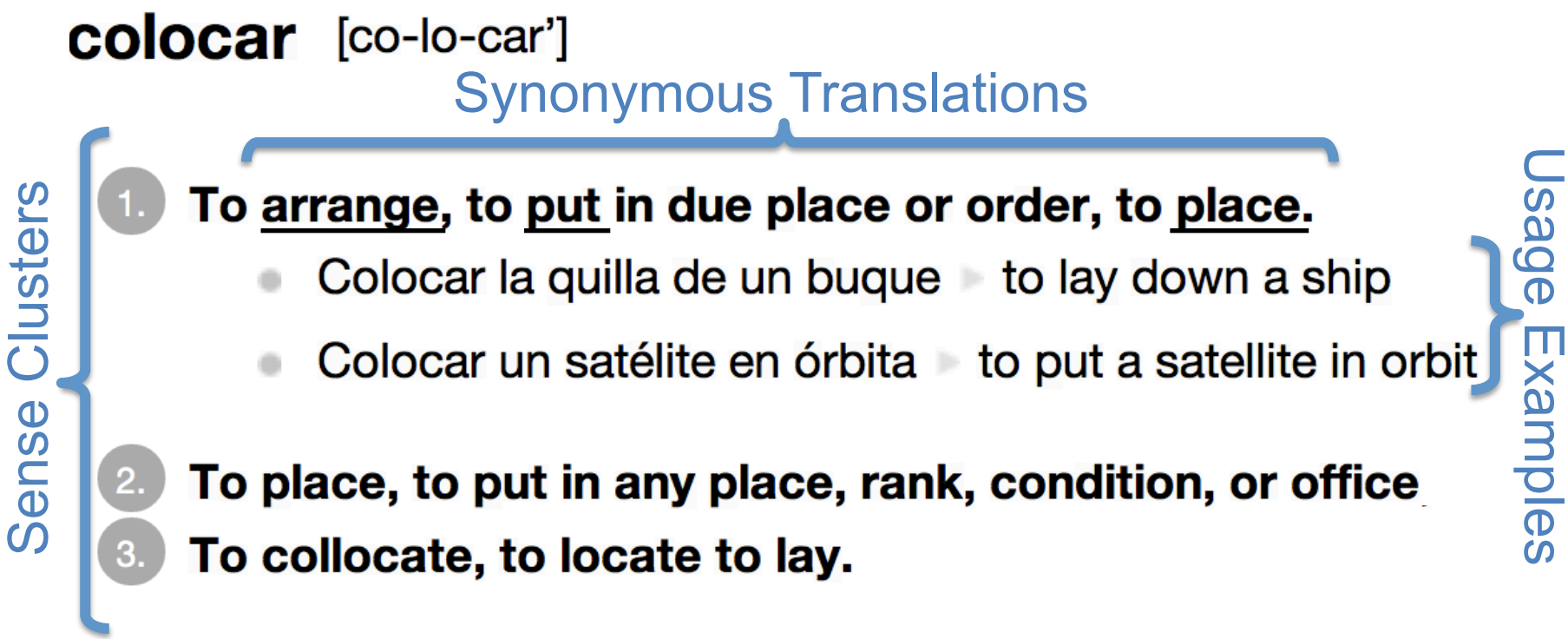
Motivation

Statistical approach

Text and bitext

Bilingual dictionaries

An Example Dictionary Entry



Velazquez Spanish and English Dictionary

Task Description

3-step Pipeline

1) Identify high-quality target-side translations for source lexical items
   - Well-studied problem, e.g., Brown et al. 1993

Human-curated translations

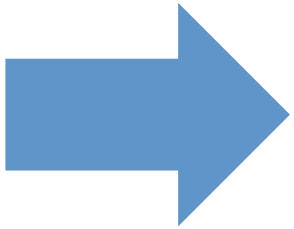
2) Cluster translations of each source word according to common word senses

3) Annotate clusters with usage examples

Example 1

ganar

win
gain
earn
make
beat
save

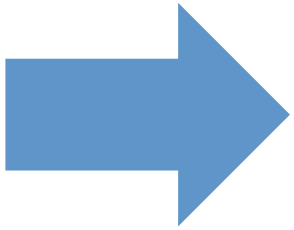


1. earn, gain, make
   - *ganar algo de dinero (earn some money)*
2. win, beat
   - *ganar la competencia (win the competition)*
3. save
   - *ganar tiempo (gain time)*

Example 2

colocar

place
put
position
locate
invest
collocate



1. place, position, put
   - *colocar en un lugar (put in a place)*
2. invest, place, put
   - *capitales para colocar (capital to invest)*
3. locate, place
   - *colocar el número de serie (locate the serial number)*
4. collocate
   - *colocar juntas todas los libros (collocate all the books)*

Translation Sense Clustering

Given a source word s and a set T_s of target translations, generate clusters such that

- Each $t \in T_s$ appears in at least one cluster
- Each cluster $G \subseteq T_s$ is:
  - ***coherent*** –all target words in G share some common sense
  - ***complete*** – for any sense B shared by all target words in G, there is no word in T_s but not in G that also shares that sense

Translation Sense Clustering

Coherence:

(earn, gain, make, beat)

Translation Sense Clustering

Coherence:

(earn, gain, make, beat)

Translation Sense Clustering

Coherence:

(earn, gain, make, beat)

Completeness:

(earn) (gain, make)

Translation Sense Clustering

Coherence:

(earn, gain, make, beat)

Completeness:

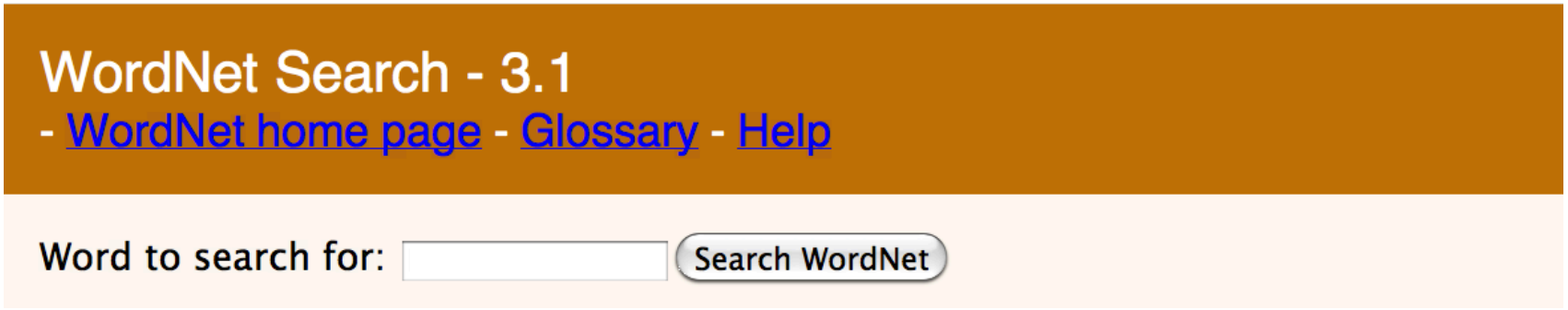
(earn, gain, make)

Building the Test Set

Clean Translation Lists

Source	Translations
colocar	collocate
	invest
	locate
	place
	position
	put

Reference (Gold) Clusters



Reference (Gold) Clusters

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: locate Search WordNet

Verb

Synsets

- S: (v) **locate**, turn up (discover the location of; determine the place of; find by searching or examining) *"Can you locate your cousins in the Midwest?"; "My search turned up nothing"*
- S: (v) situate, **locate** (determine or indicate the place, site, or limits of, as if by an instrument or by a survey) *"Our sense of sight enables us to locate objects in space"; "Locate the boundaries of the property"*
- S: (v) **locate**, place, site (assign a location to) *"The company located some of their agents in Los Angeles"*
- S: (v) settle, **locate** (take up residence and become established) *"The immigrants settled in the Midwest"*

Cluster Projection

Source	Translations	Synsets
colocar	collocate	collocate collocate, lump, chunk
	invest	invest, put, commit, place invest, clothe, adorn invest, vest, enthrone …
	locate	locate, turn up situate, locate locate, place, site …
	…	

Cluster Projection

Source	Translations	Synsets
colocar	collocate	collocate collocate, lump, chunk
	invest	invest, put, commit, place invest, clothe, adorn invest, vest, enthrone …
	locate	locate, turn up situate, locate locate, place, site …
	…	

Cluster Projection

Source	Translations	Synsets
colocar	collocate	collocate ~~collocate, lump, chunk~~
	invest	invest, put, commit, place ~~invest, clothe, adorn~~ ~~invest, vest, enthrone~~ …
	locate	~~locate, turn up~~ ~~situate, locate~~ locate, place, site …
	…	

Cluster Projection

Source Translations Synsets

colocar

Translation Sense Clusters

collocate

invest, place, put

locate, place

place, position, put

…

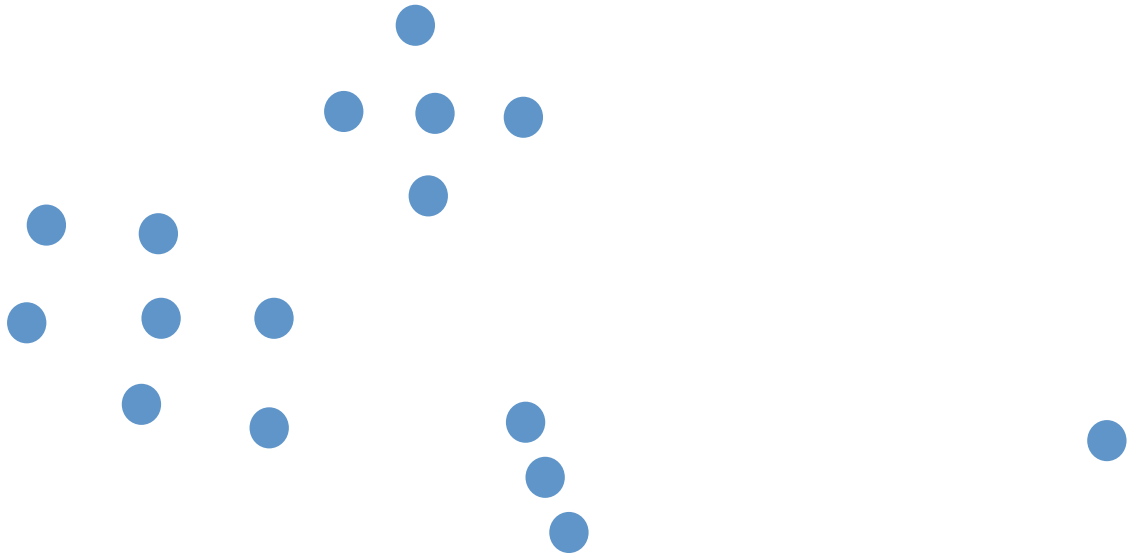
Approach

Unsupervised Sense Clustering

2-step process:

- Induce the clustering on the whole vocabulary
- Apply our Cluster Projection algorithm to project vocabulary-level clusters to source-specific clusters

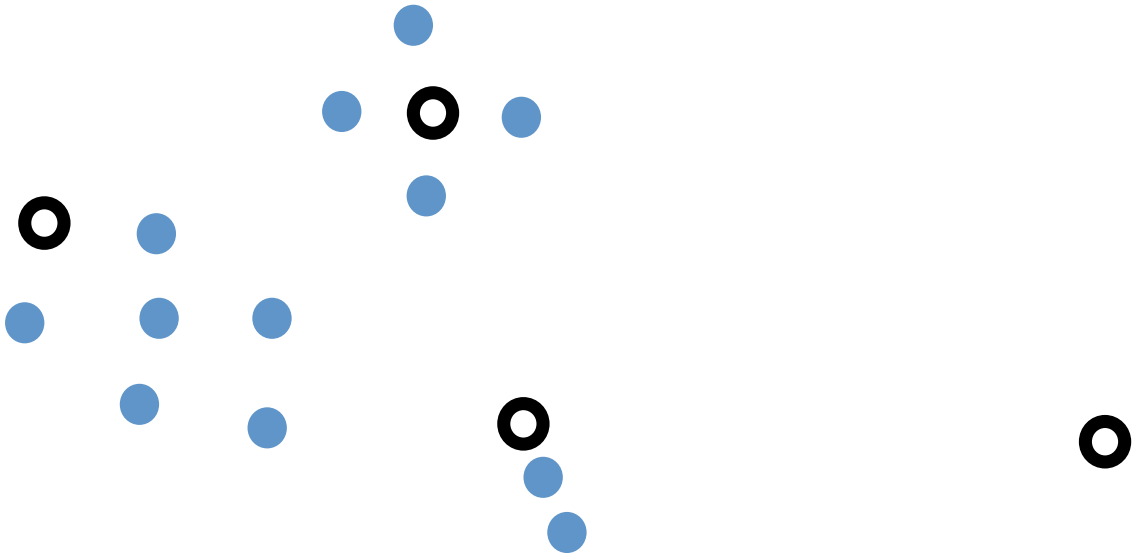
K-Means Clustering



MacQueen, 1967

K-Means Clustering

1. Select *K* initial centroids



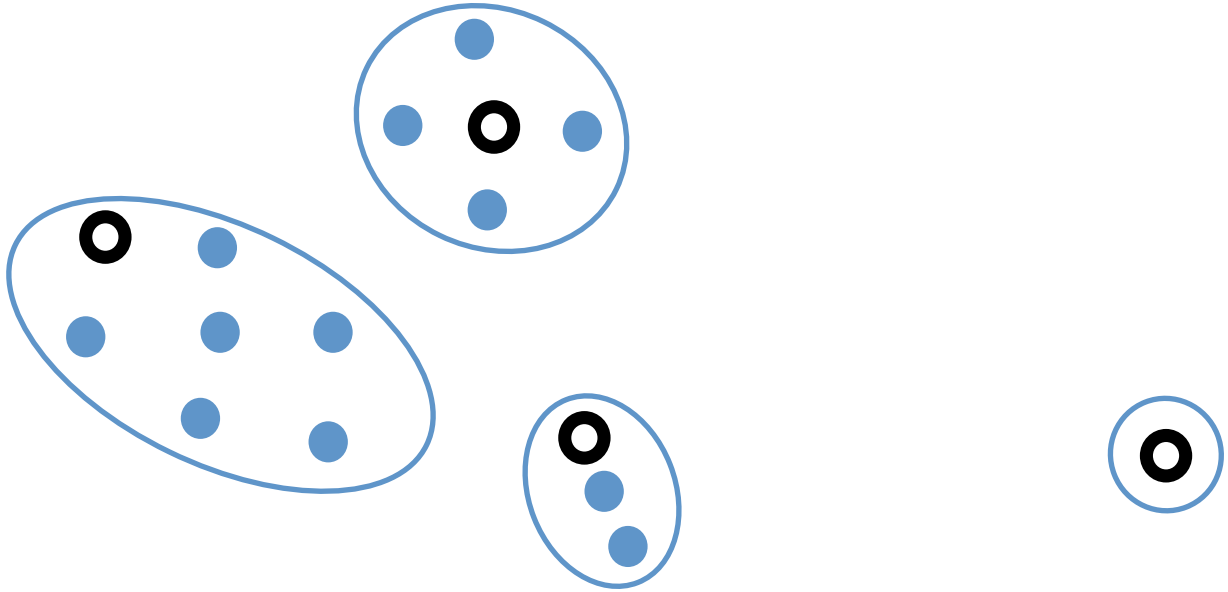
MacQueen, 1967

K-Means Clustering

1. Select *K* initial centroids

repeat

2. Assign each element to the nearest cluster



MacQueen, 1967

K-Means Clustering

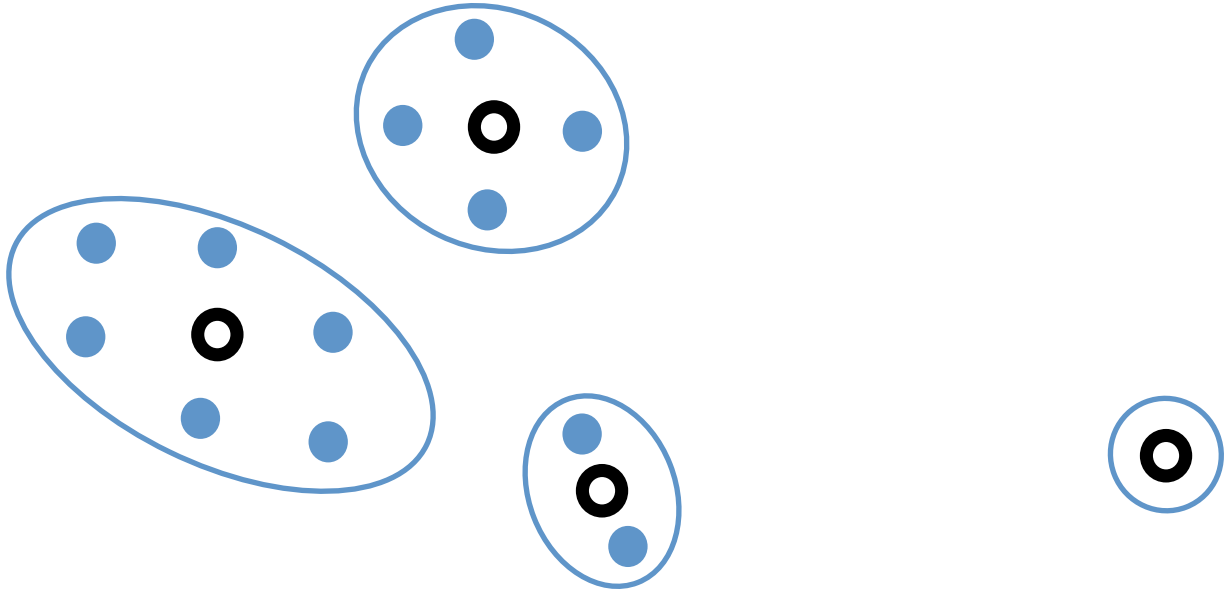
1. Select *K* initial centroids

repeat

2. Assign each element to the nearest cluster

3. Recompute centroids by averaging members

until convergence



MacQueen, 1967

M-Soft *K*-Means Overlapping Clusters

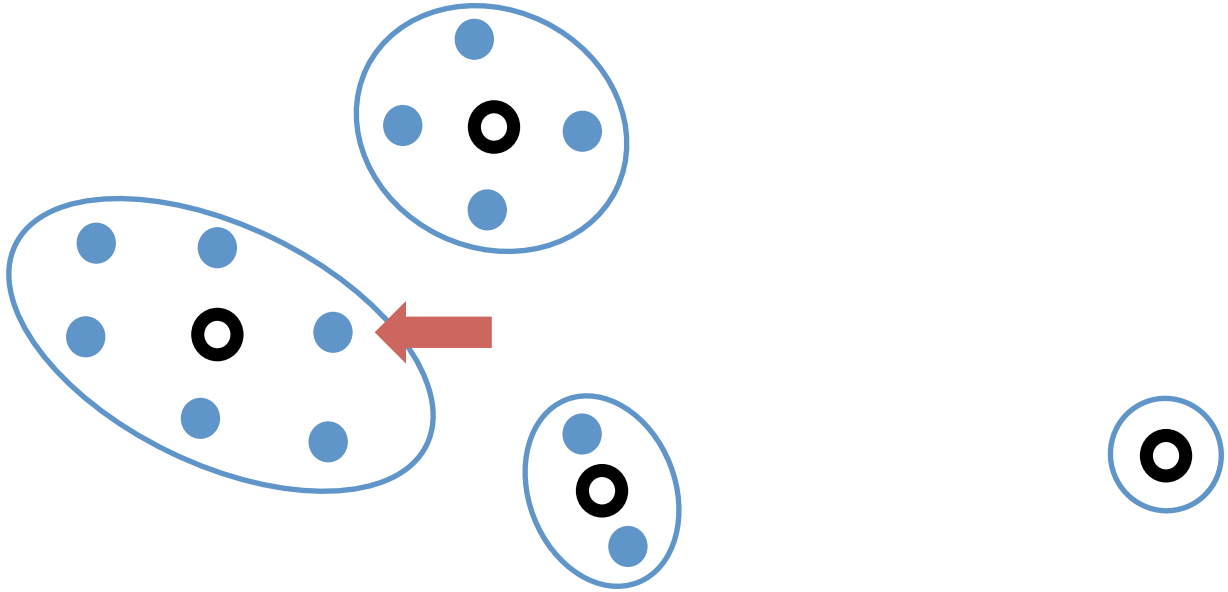
1. Select *K* initial centroids

repeat

2. Assign each element to the *M* nearest clusters

3. Recompute centroids by averaging members

until convergence



Lin and Wu, 2009

M-Soft *K*-Means Overlapping Clusters

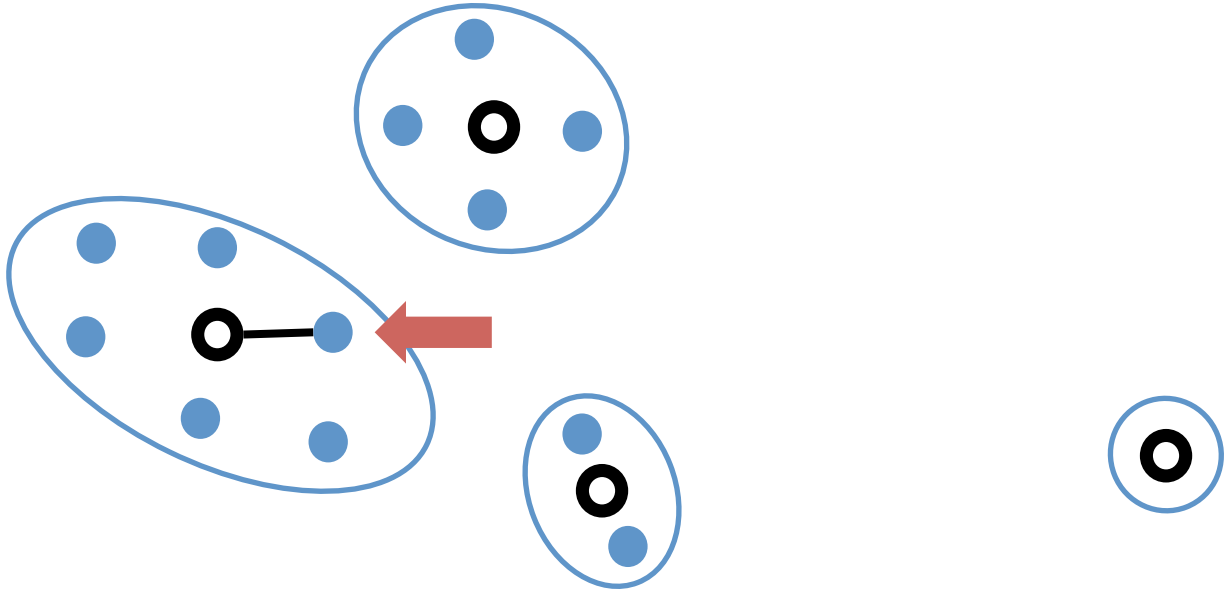
1. Select *K* initial centroids

repeat

2. Assign each element to the *M* nearest clusters

3. Recompute centroids by averaging members

until convergence



Lin and Wu, 2009

M-Soft *K*-Means Overlapping Clusters

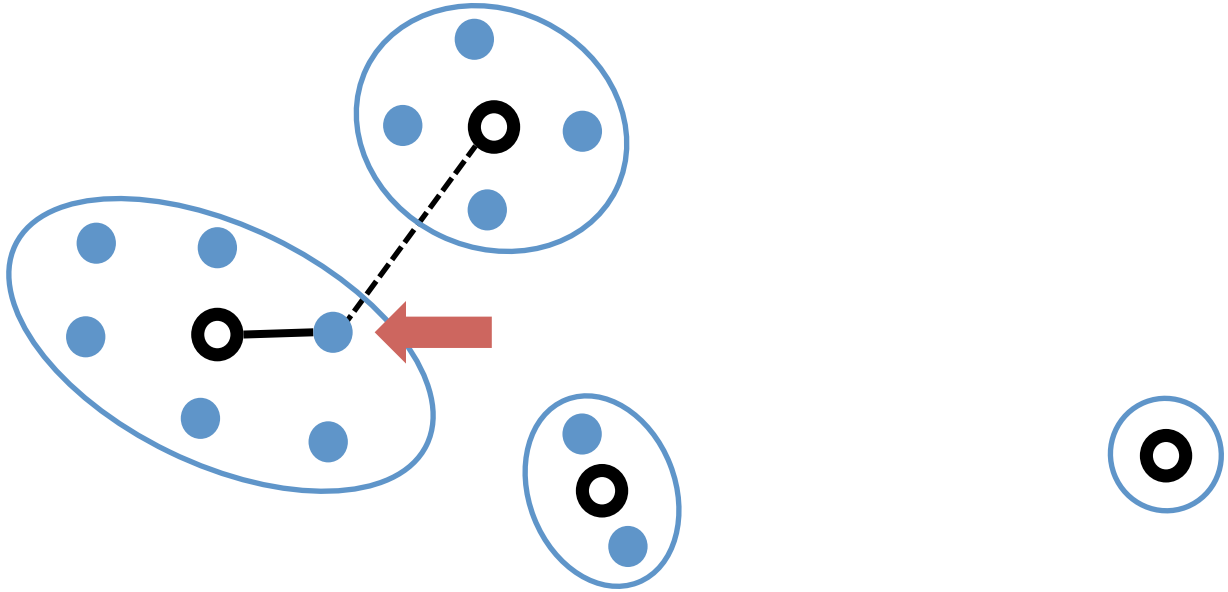
1. Select *K* initial centroids

repeat

2. Assign each element to the *M* nearest clusters

3. Recompute centroids by averaging members

until convergence



Lin and Wu, 2009

M-Soft *K*-Means Overlapping Clusters

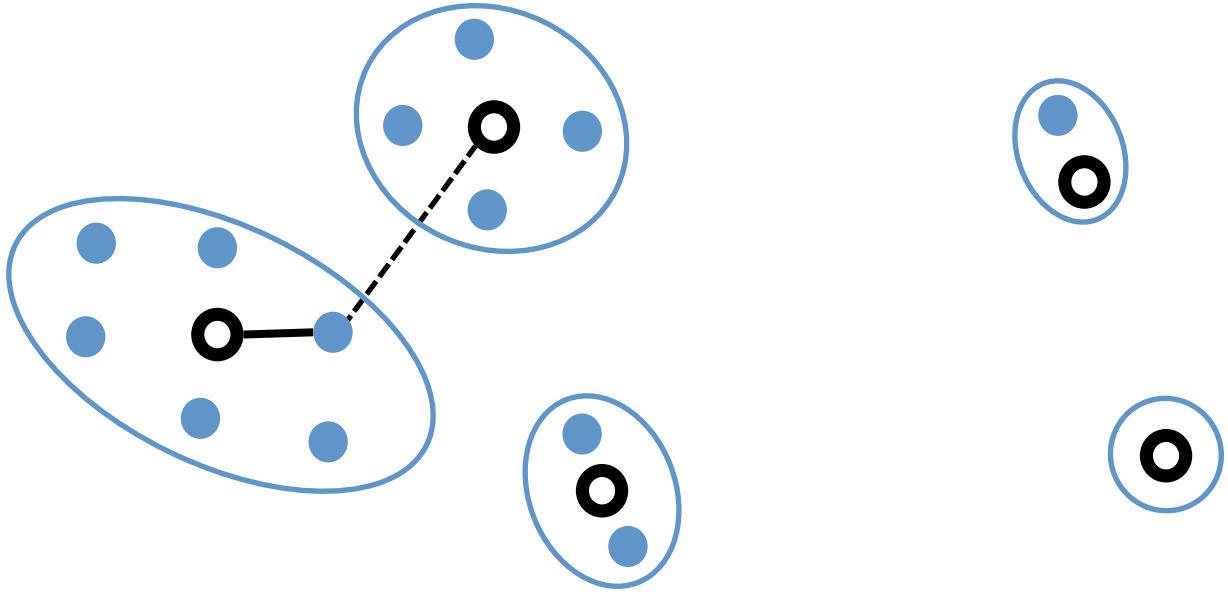
1. Select *K* initial centroids

repeat

2. Assign each element to the *M* nearest clusters

3. Recompute centroids by averaging members

until convergence



Lin and Wu, 2009

M-Soft *K*-Means Overlapping Clusters

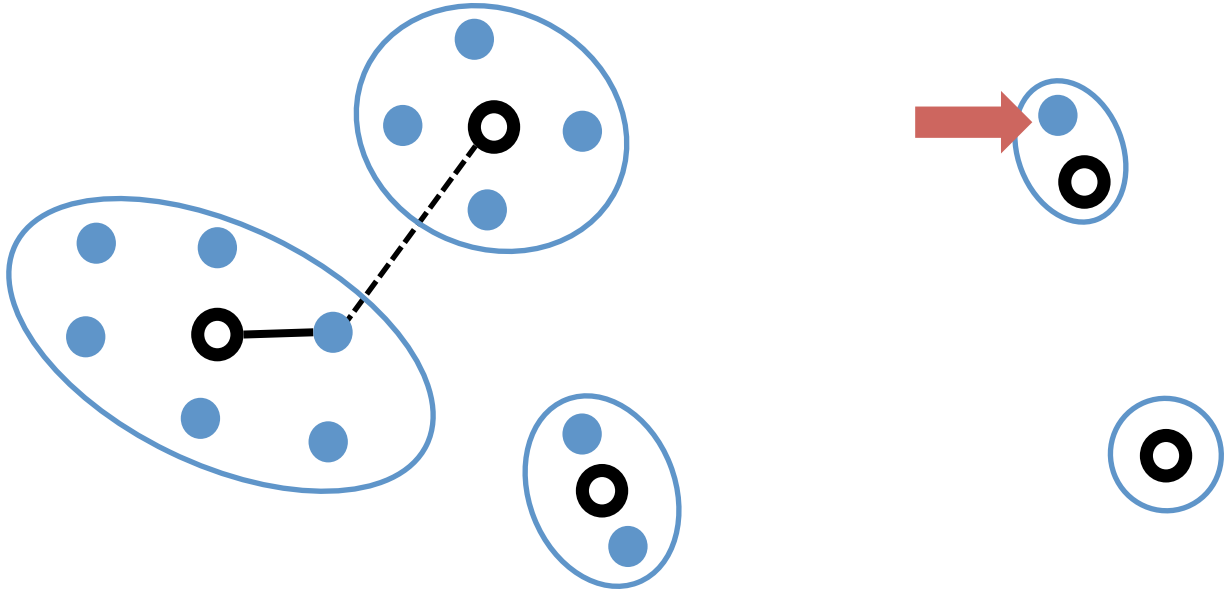
1. Select *K* initial centroids

repeat

2. Assign each element to the *M* nearest clusters

3. Recompute centroids by averaging members

until convergence



Lin and Wu, 2009

M-Soft *K*-Means Overlapping Clusters

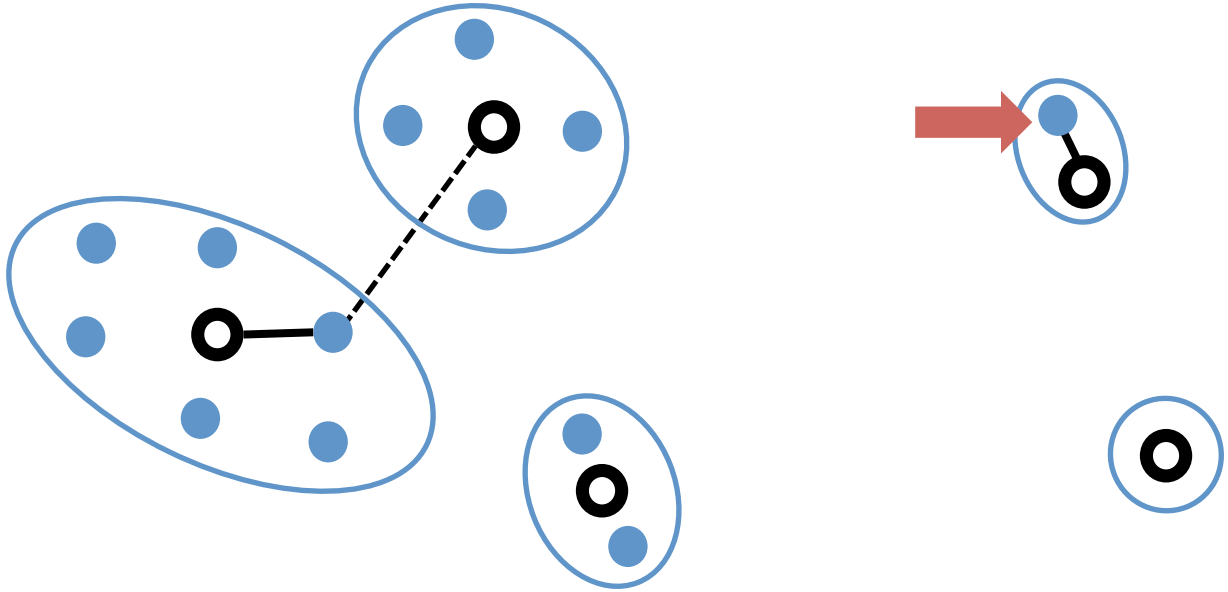
1. Select *K* initial centroids

repeat

2. Assign each element to the *M* nearest clusters

3. Recompute centroids by averaging members

until convergence



Lin and Wu, 2009

Cluster Projection

Number of clusters in corpus $K = 2^3, 2^4, \ldots, 2^{12}$
Number of clusters per word $M = 1, 2, 3, 4, 5$

For example, K = 2^8, M = 3:

ablate	→	C151	C124	
ablated	→	C151	C132	C124
ablates	→	C151	C118	C124
ablaze	→	C250	C29	
able	→	C208	C124	C255
...				
…				
zymogenetics	→	C200	C129	
zymogens	→	C129	C96	C168
zymographic	→	C87	C246	
zymography	→	C87	C151	C81

top-3 cluster indices per word

Cluster Projection

Only consider the clusters of the target words in T_s

colocar

collocate	→	C131	C114	C12
invest	→	C73	C124	C44
locate	→	C145	C36	
place	→	C14	C145	C73
position	→	C14	C138	C112
put	→	C14	C138	

Cluster Projection

Again remove redundant and subset clusters

colocar

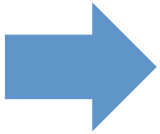
collocate	→	C131	C114	C12
invest	→	C73	C124	C44
locate	→	C145	C36	
place	→	C14	C145	C73
position	→	C14	C138	C112
put	→	C14	C138	

Cluster Projection

Again remove redundant and subset clusters

colocar

collocate	→	C131	C114	C12
invest	→	C73	C124	C44
locate	→	C145	C36	
place	→	C14	C145	C73
position	→	C14	C138	C112
put	→	C14	C138	



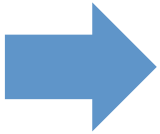
C131	→	collocate, locate
C114	→	collocate
C12	→	collocate
C73	→	invest, place
C124	→	invest
C44	→	invest
C145	→	locate, place
C36	→	locate
C14	→	place, position, put
C138	→	position, put
C112	→	position

Cluster Projection

Again remove redundant and subset clusters

colocar

collocate	→	C131	C114	C12
invest	→	C73	C124	C44
locate	→	C145	C36	
place	→	C14	C145	C73
position	→	C14	C138	C112
put	→	C14	C138	



C131 → collocate, locate

~~C114 → collocate~~

~~C12 → collocate~~

C73 → invest, place

~~C124 → invest~~

~~C44 → invest~~

C145 → locate, place

~~C36 → locate~~

C14 → place, position, put

~~C138 → position, put~~

~~C112 → position~~

Cluster Projection

Again remove redundant and subset clusters

Translation Sense Clusters

collocate, locate

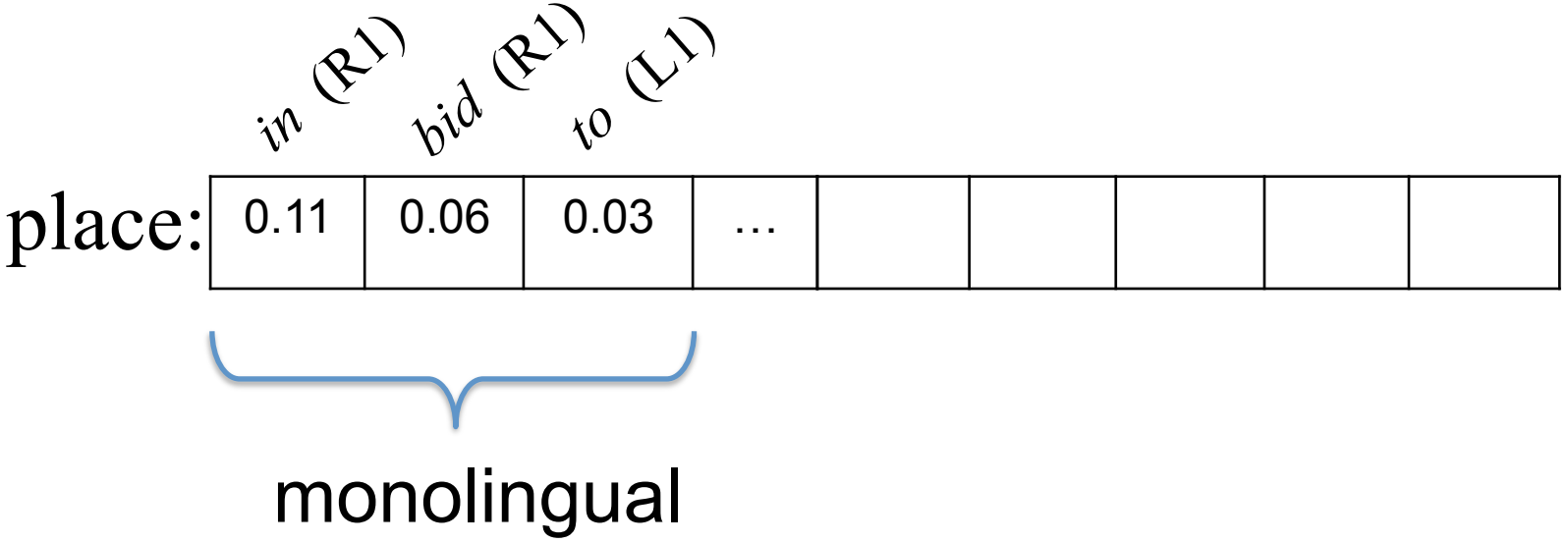
invest, place

locate, place

place, position, put

Monolingual and Bilingual Features

Monolingual and Bilingual Features

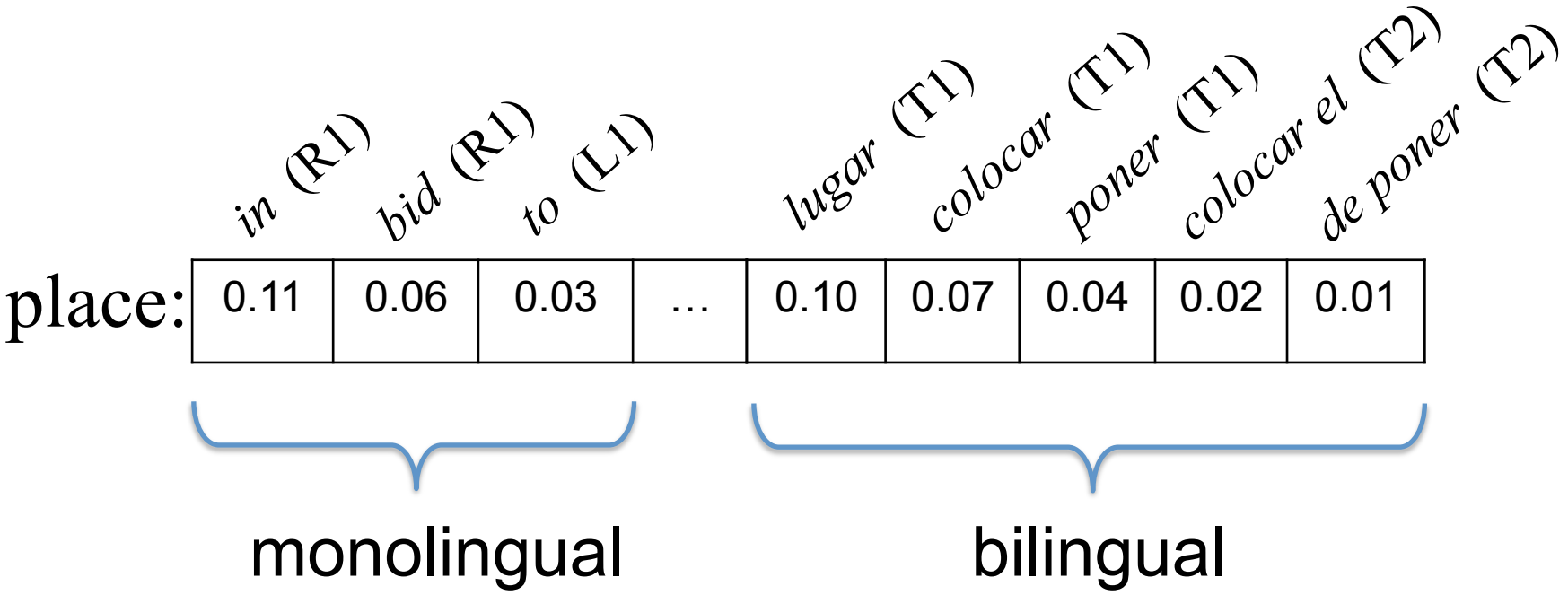


L1: word to the left (monolingual context)

R1: word to the right (monolingual context)

Lin and Wu, 2009

Monolingual and Bilingual Features



T1: word translation (bilingual context)

T2: bigram translation (bilingual context)

Results

Data

Language-pairs

Japanese-English, Spanish-English

Monolingual features

English corpus of Web documents with 700B tokens of text

Bilingual features

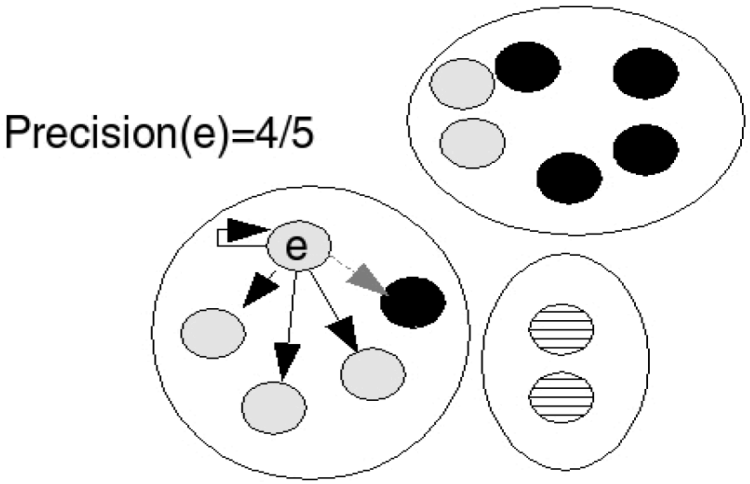
0.78B (S-E) and 1.04B (J-E) tokens of parallel text

Test sets

Dataset	# source words	Total # target words
Japanese-English	369 (319 NN, 38 VB, 12 ADV)	1639
Spanish-English	52 (38 NN, 10 VB, 4 ADV)	230

BCubed Metric

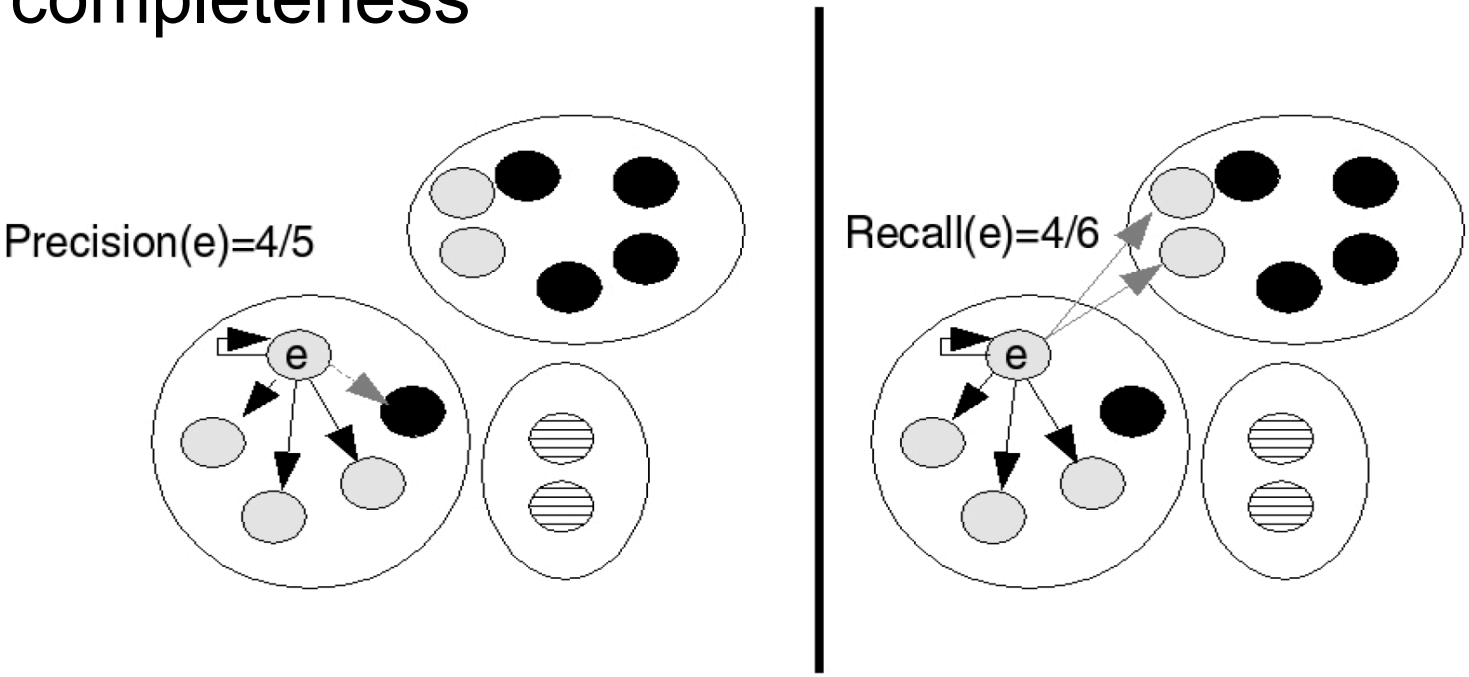
- Handles overlapping clusters
- Decomposes evaluation into P and R associated to each item, which reflect coherence and completeness



Amigo et al., 2009

BCubed Metric

- Handles overlapping clusters
- Decomposes evaluation into P and R associated to each item, which reflect coherence and completeness



Amigo et al., 2009

BCubed Metric

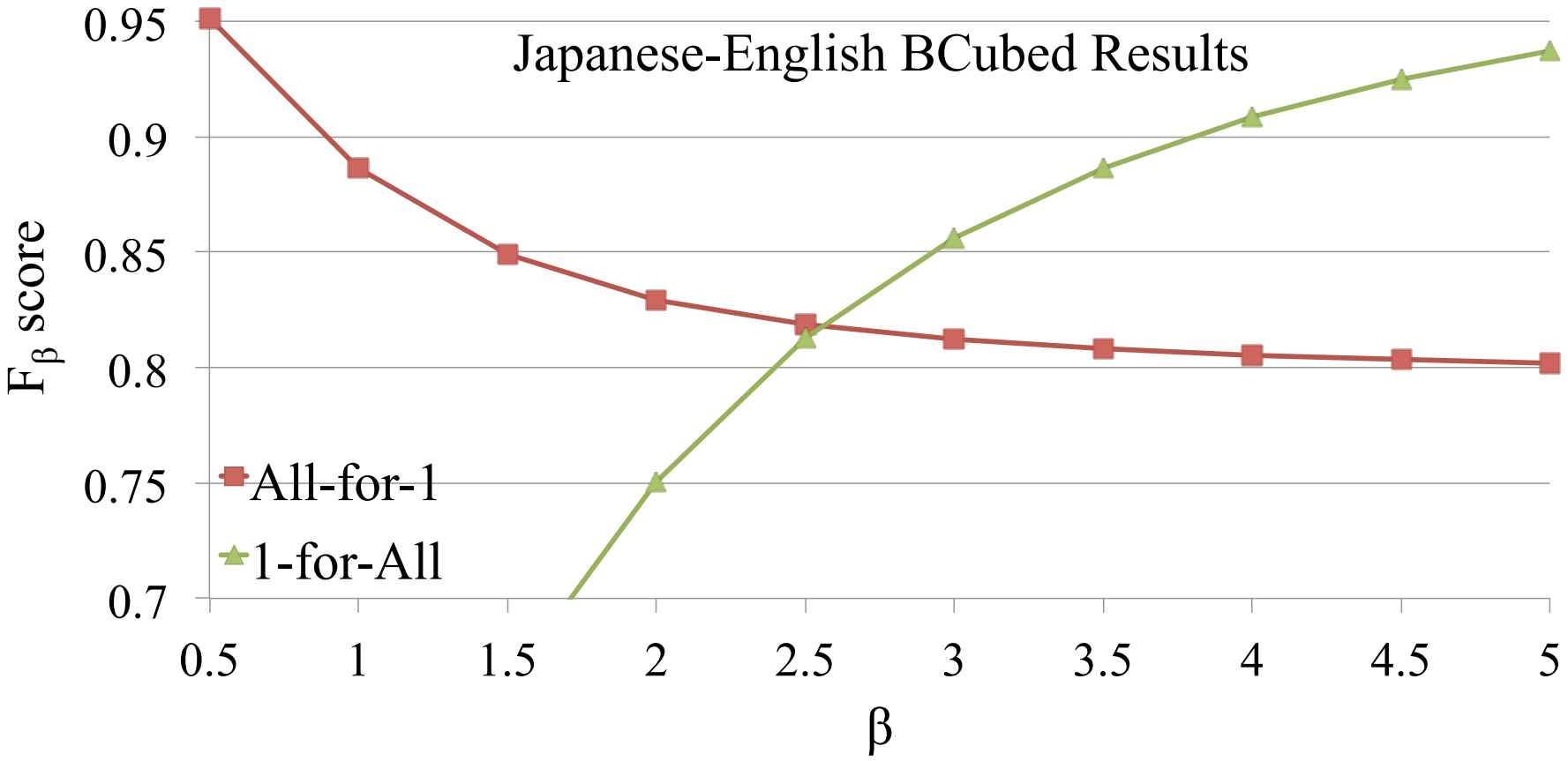
$$\mathrm{P_{B3}} = \mathrm{Avg}_e[\mathrm{Avg}_{e' s.t. C(e) \cap C(e') \neq \emptyset}[\mathrm{P}(e, e')]]$$

$$\mathrm{R_{B3}} = \mathrm{Avg}_e[\mathrm{Avg}_{e' s.t. L(e) \cap L(e') \neq \emptyset}[\mathrm{R}(e, e')]]$$

$$\mathrm{F}_\beta = \frac{(1 + \beta^2) \cdot \mathrm{P_{B3}} \cdot \mathrm{R_{B3}}}{\beta^2 \cdot \mathrm{P_{B3}} + \mathrm{R_{B3}}}$$

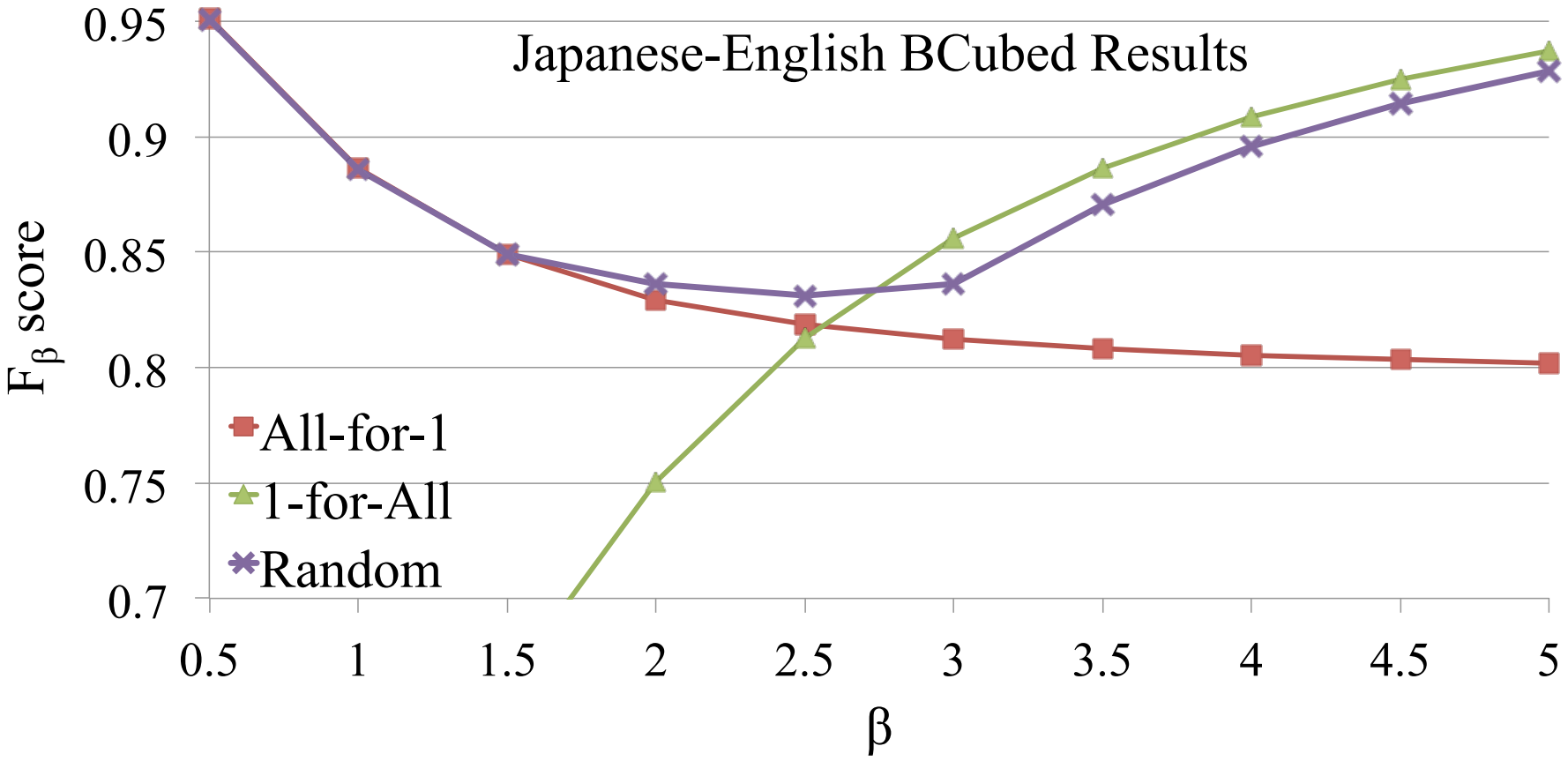
Amigo et al., 2009

F_β Results per Clustering Approach



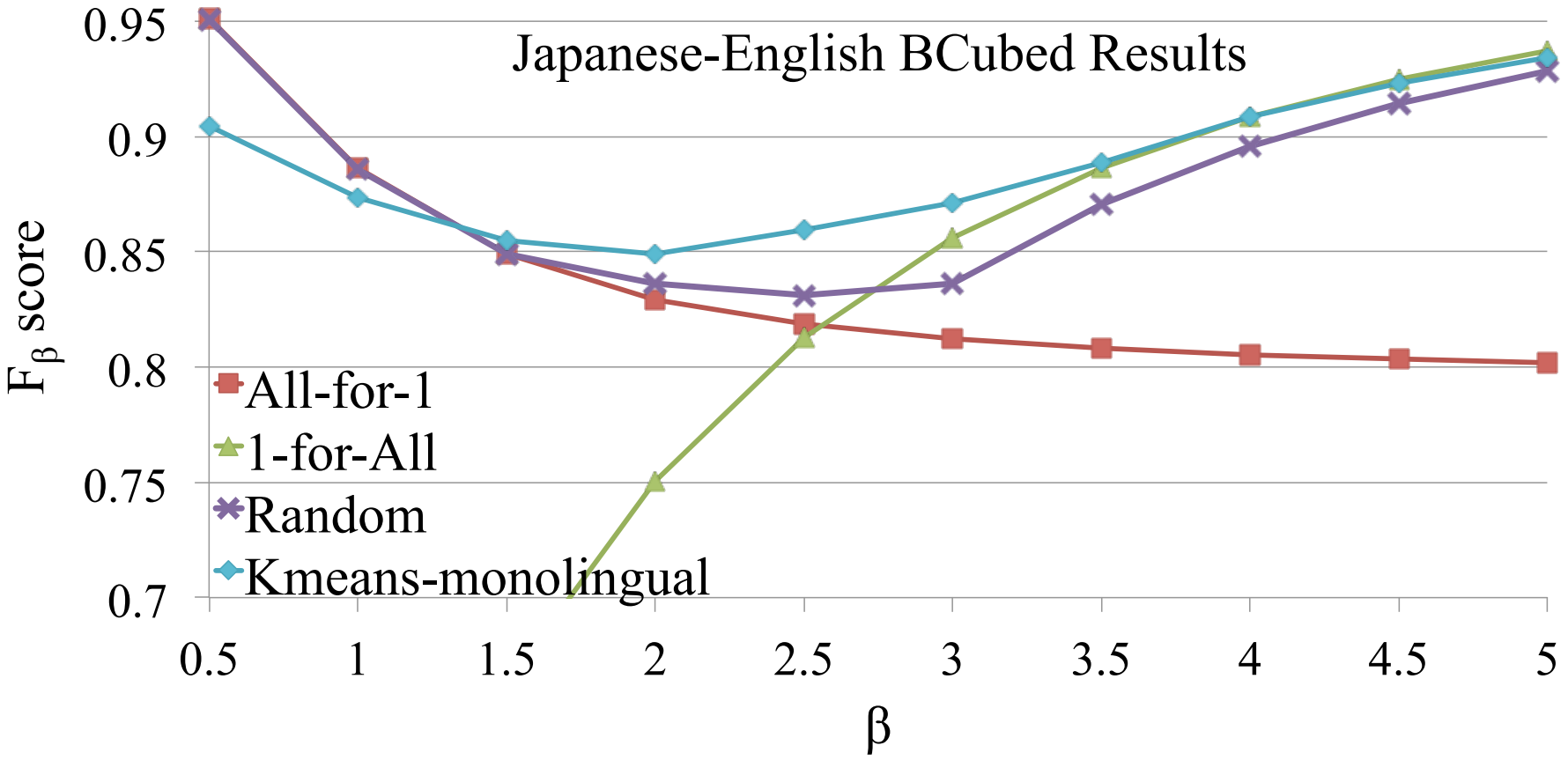
All-for-1: Each word in its own cluster
1-for-All: All words in one cluster

F_β Results per Clustering Approach



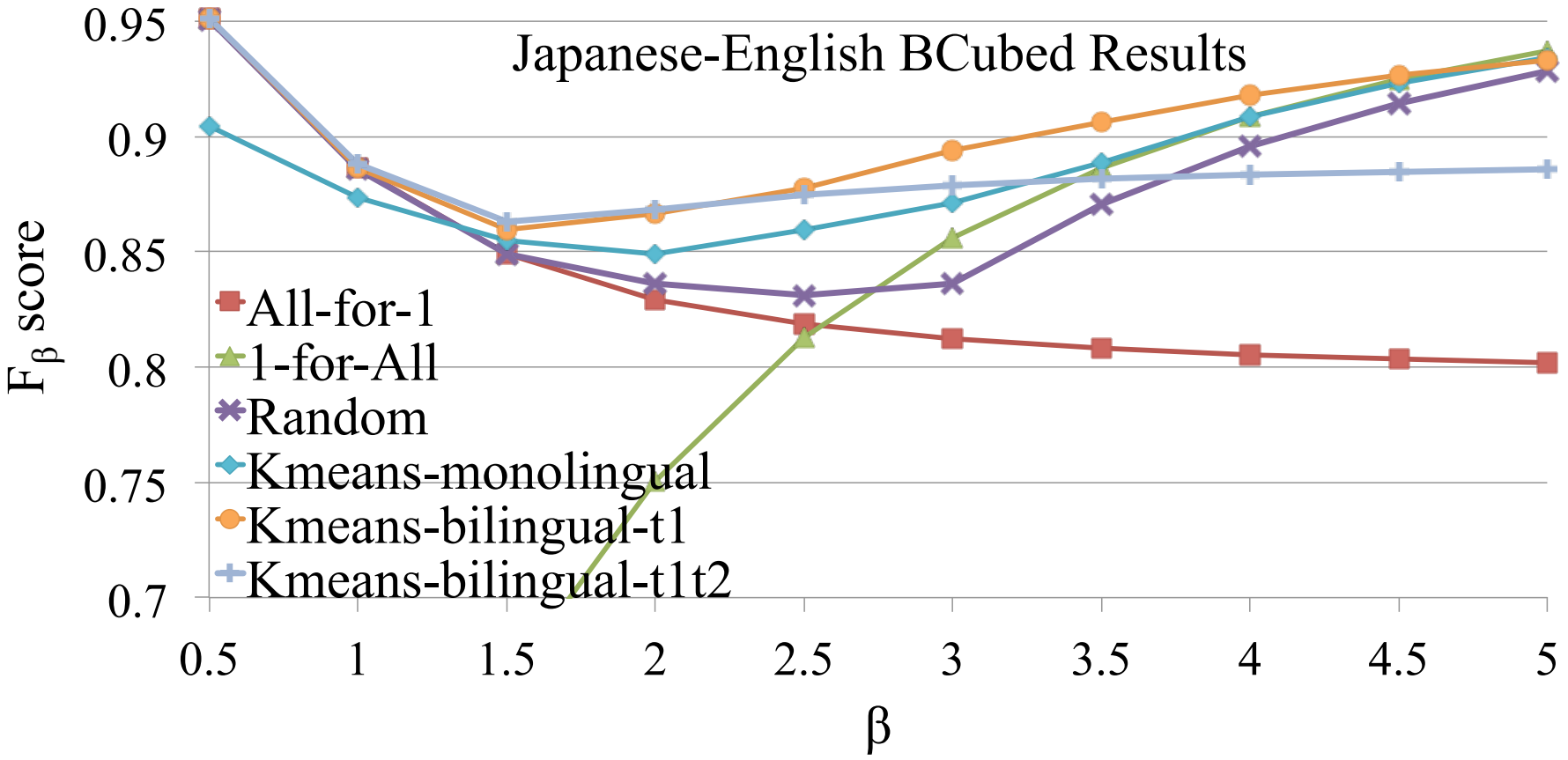
Random: Each word assigned *M* random cluster id's in *1* to *K*

F_β Results per Clustering Approach



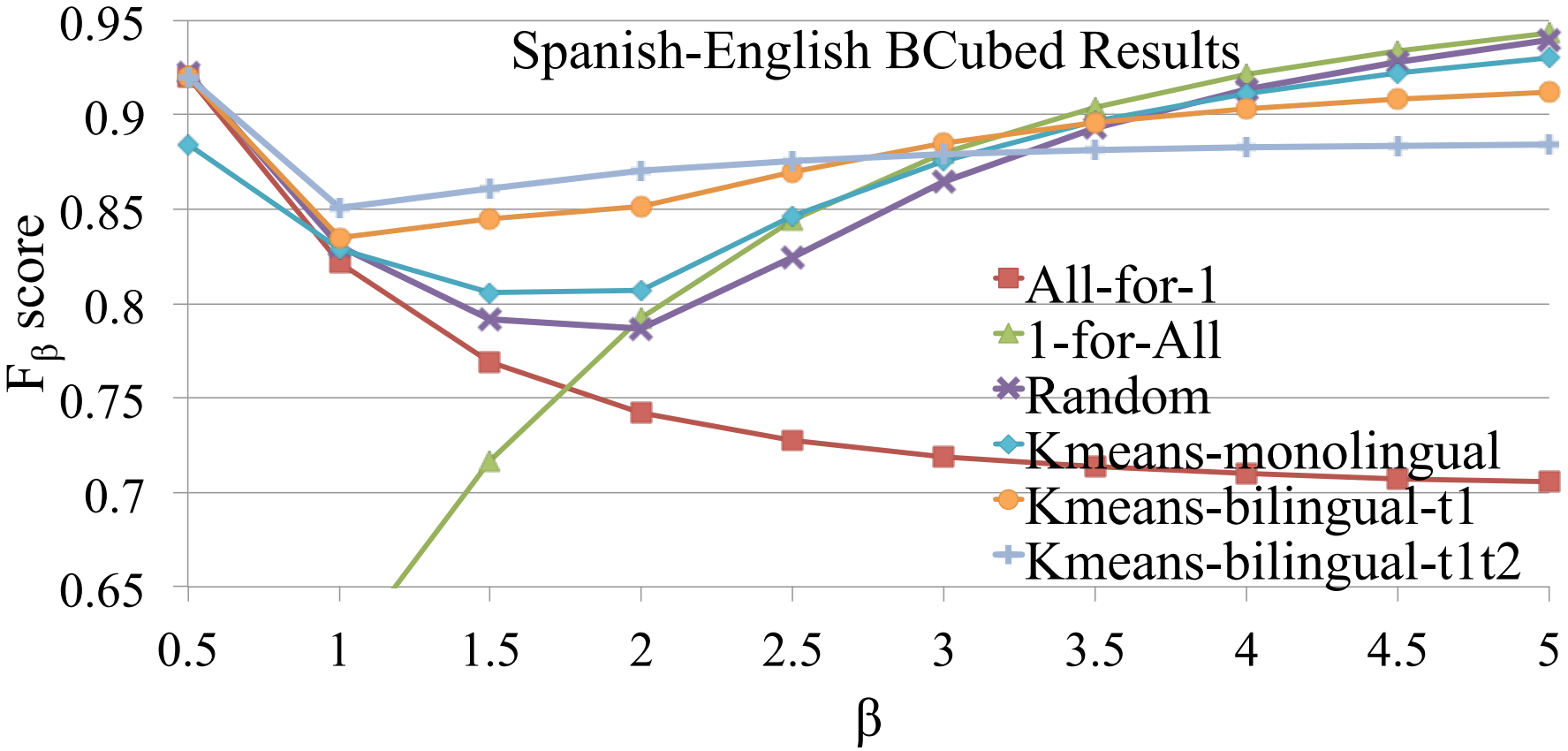
Kmeans-monolingual: Uses monolingual features

F_β Results per Clustering Approach

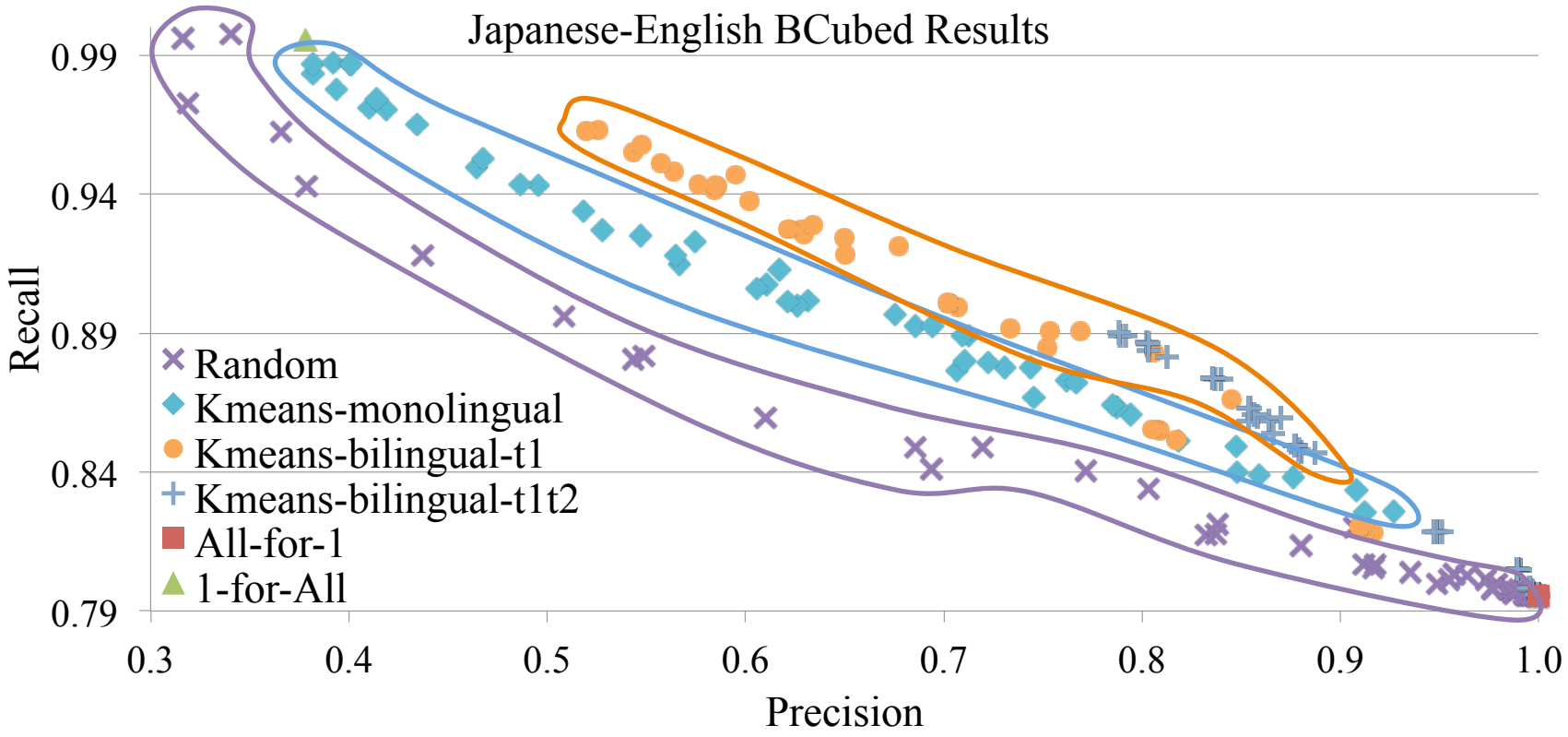


Kmeans-bilingual-t1: Uses bilingual unigram features
Kmeans-bilingual-t1t2: Uses bilingual unigram and bigram features

F_β Results per Clustering Approach



Precision-Recall Scatter Plot



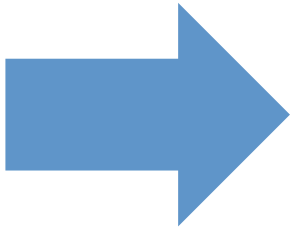
Each point corresponds to a setting of *K* and *M* for *M*-best *K*-means

Usage Examples

Example 1

ganar

win
gain
earn
make
beat
save



1. earn, gain, make
   - *ganar algo de dinero (earn some money)*
2. win, beat
   - *ganar la competencia (win the competition)*
3. save
   - *ganar tiempo (gain time)*

Output Examples

debajo

["below","beneath"]	→	debajo de la superficie (*below the surface*)
["below","under"]	→	debajo de la línea (*below the line*)
["underneath"]	→	debajo de la piel (*under the skin*)

休養

["break"]	→	一生懸命 働い た から 休養 する の は 当然 です . (*I worked hard and I deserve a good break.*)
["recreation"]	→	従来 の 治療 や 休養 方法 (*Traditional healing and recreation activities*)
["rest"]	→	ベッド で 休養 する だけ で 治り ます . (*Bed rest is the only treatment required.*)

Output Examples

利用

["application"]	→	コンピューター 利用 技術
		(*Computer-aided technique*)
["use","utilization"]	→	土地 の 有効 利用 を 促進 する
		(*Promote effective use of land*)

引く

["draw","pull"]	→	カーテン を 引く
		(*Draw the curtain*)
["subtract"]	→	A から B を 引く
		(*Subtract B from A*)
["tug"]	→	袖 を ぐいと 引く
		(*Tug at someone's sleeve*)

Thank you!

Other Details

Cluster Projection Algorithm

Notation

T_s : The set of target-language translations (given)
$\mathcal{D}_t$: The set of synsets in which t appears
C : A synset; a set of target-language words
B : A source-specific synset; a subset of T_s
$\mathcal{B}$: A set of source-specific synsets
$\mathcal{G}$: A set of correct sense clusters for T_s

The Cluster Projection Algorithm:

$\mathcal{B} \leftarrow \left\{ C \cap T_s \ : \ C \in \bigcup_{t \in T_s} \mathcal{D}_t \right\}$
$\mathcal{G} \leftarrow \emptyset$
for $B \in \mathcal{B}$ **do**
  if $\nexists B' \in \mathcal{B}$ such that $B \subset B'$ **then**
    add B to $\mathcal{G}$
return $\mathcal{G}$

BCubed Metric

$$\mathrm{P}(e, e') = \frac{\min(|C(e) \cap C(e')|,\ |L(e) \cap L(e')|)}{|C(e) \cap C(e')|}$$

$$\mathrm{R}(e, e') = \frac{\min(|C(e) \cap C(e')|,\ |L(e) \cap L(e')|)}{|L(e) \cap L(e')|}$$

$$\mathrm{P}_{\mathrm{B3}} = \mathrm{Avg}_e[\mathrm{Avg}_{e' s.t. C(e) \cap C(e') \neq \emptyset}[\mathrm{P}(e, e')]]$$

$$\mathrm{R}_{\mathrm{B3}} = \mathrm{Avg}_e[\mathrm{Avg}_{e' s.t. L(e) \cap L(e') \neq \emptyset}[\mathrm{R}(e, e')]]$$

$$\mathrm{F}_\beta = \frac{(1 + \beta^2) \cdot \mathrm{P}_{\mathrm{B3}} \cdot \mathrm{R}_{\mathrm{B3}}}{\beta^2 \cdot \mathrm{P}_{\mathrm{B3}} + \mathrm{R}_{\mathrm{B3}}}$$

Amigo et al., 2009

Example Extraction Procedure

For source word s and target sense cluster G:

P_s : set of source phrases containing source word s

A_t : set of source phrases that align to target phrases containing target word t

Collect the set of phrases that contain s and translate to all words in G, i.e.,

$$P_s \cap \bigcap_{t \in G} A_t$$