A report for 2011 Google Research Award

Latent Relational Search Engine

Nguyen Tuan Duc, Danushka Bollegala and Mitsuru Ishizuka

School of Information Science and Technology

THE UNIVERSITY OF TOKYO

We are interested especially in Relations between Entities toward Web Intelligence

1. Computing Relational Similarity between Two Word Pairs
   (1) Computing Relational Similarity
   (2) Open Relation Extraction employing Sequential Co-clustering

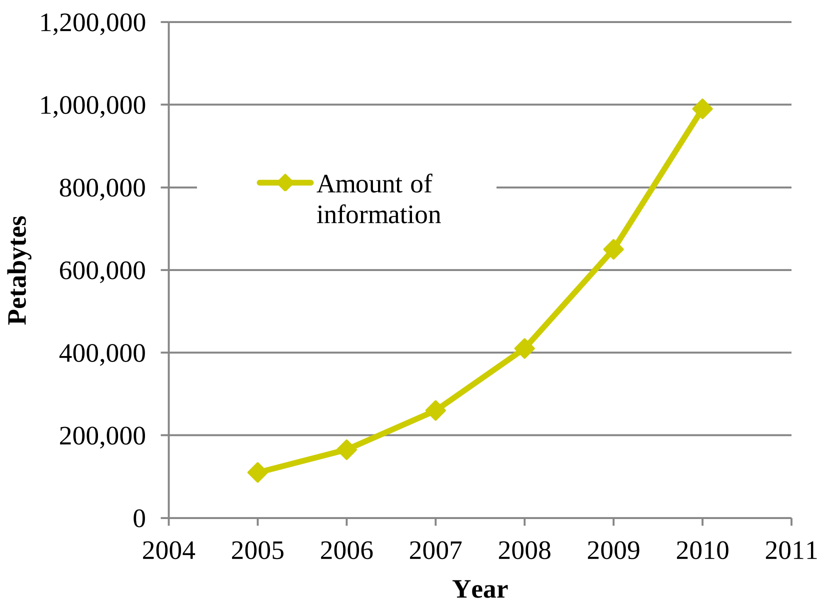
2. Latent Relational Search Engine

(2009 Japanese Patent Application)

3. Common and Universal Concept Description Language (CDL)
   as a Foundation of Semantic Computing

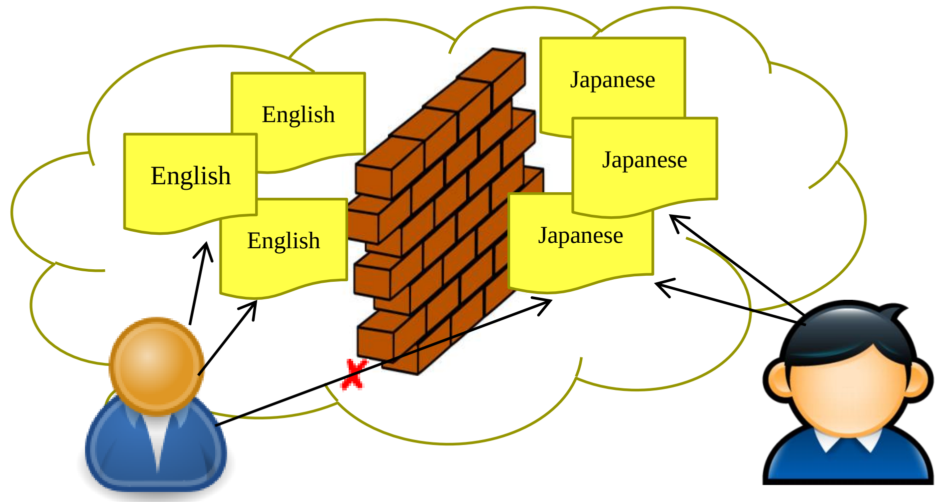
Challenges in Web Information Retrieval

- Huge amount of data



Source: IDC 2007

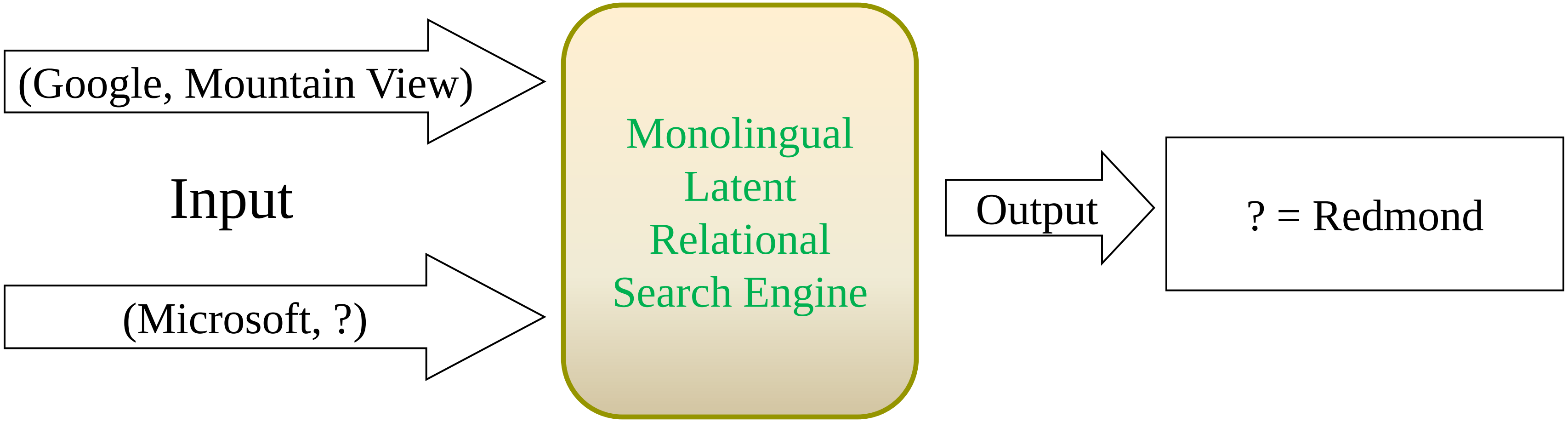
- Language barrier



- Hard to search for information in different languages

- Only keyword-based Web search? Numerous page retrieved
- Monolingual information retrieval ? → Could not search in other languages

Latent relational search



An entity retrieval paradigm based on the relational similarity between two entity pairs

- D. Bollegala et al. , Measuring the Similarity between Implicit Semantic Relations from the Web, Proc. of WWW2009
- T. Veale, The Analogical Thesaurus, IAAI 2003.

Demo (Monolingual LRS)

Word pair 1: Ganymede | Jupiter

Word pair 2: ? | Mars

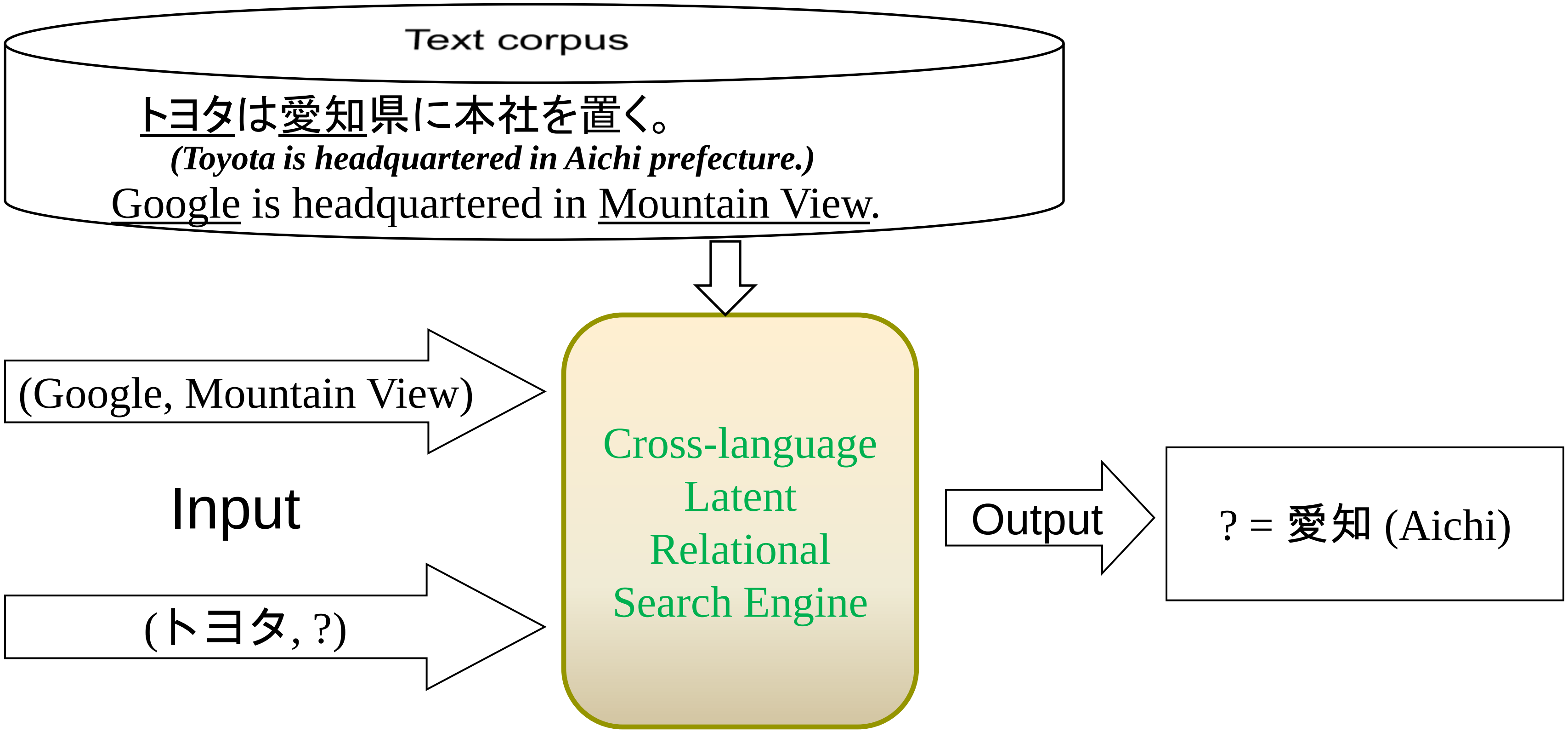
Search

Ganymede is to **Jupiter** as:

- **['phobos',]** is to **Mars** (Score = 0.411815075397) Hide evidence
  - Ganymede orbits Jupiter.
    http://wiki.answers.com/Q/FAQ/7527
  - Phobos orbiting Mars.
    http://www.helium.com/items/1860022-what-is-the-monolith-on-mars-moon-phobos
  - Earth colonists on Ganymede, the largest satellite of Jupiter, have discovered the existence of intelligent life on the planet's surface.
    http://en.wikipedia.org/wiki/Not_Final
  - Phobos, the inner satellite of Mars, appears to perform a secular acceleration in longitude.
    http://linkinghub.elsevier.com/retrieve/pii/001910356490048X
  - This chapter describes different methods that have been developed for the separation of external and internal source contributions, and their application to selected planets and one of Jupiter's moons, Ganymede.
    http://www.springerlink.com/index/k076148262404135.pdf

Supporting sentences

Cross-language latent relational search



We propose Cross-language latent relational search to utilize multilingual Web text

Screen shot (cross-language LRS)

Word pair 1: Ganymede | Jupiter

Word pair 2: ? | 火星

Search

Ganymede is to **Jupiter** as:

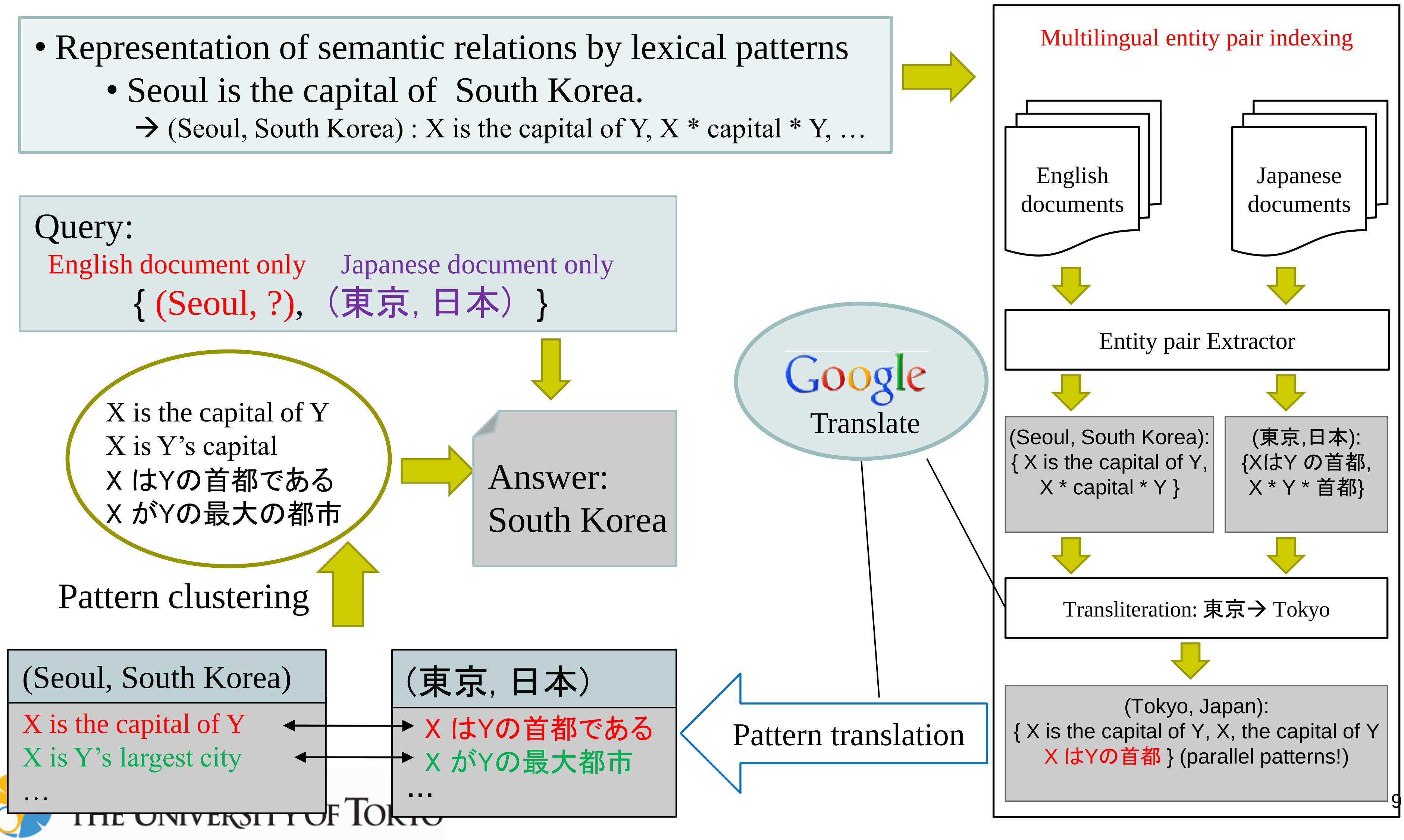
- **['フォボス',]** is to **火星** (Score = 0.170291277053) Hide evidence
  - フォボスは火星の最も大きい衛星だが、長い部分の直径で27km程度しかないため、火星の重力に捕捉された小惑星だと考えられている。
    http://news.searchina.ne.jp/disp.cgi?y=2011&d=0126&f=it_0126_009.shtml
  - Jupiter is so vast that it exerts an enormous ... One of Jupiter's moons, Ganymede, is the solar system's largest moon.
    http://www.trueknowledge.com/q/facts_about__ganymede
  - マーズ・エクスプレス撮影、火星の衛星フォボス。
    http://www.sorae.jp/031006/4274.html
  - Find information about its unique ... Facts about Jupiter's moon Ganymede including its surface and internal structure.
    http://www.trueknowledge.com/q/facts_about__ganymede
  - マーズ・エクスプレス撮影、火星の衛星フォボス。
    http://news.searchina.ne.jp/disp.cgi?y=2011&d=0126&f=it_0126_009.shtml

Outline of the presentation

- Introduction
- Method overview
- Proposal:
  - Entity pair and relation extraction, indexing method
  - Hybrid pattern clustering algorithm to alleviate data sparseness problem
- Evaluation and comparison
- Potential applications of latent relational search
- Conclusion

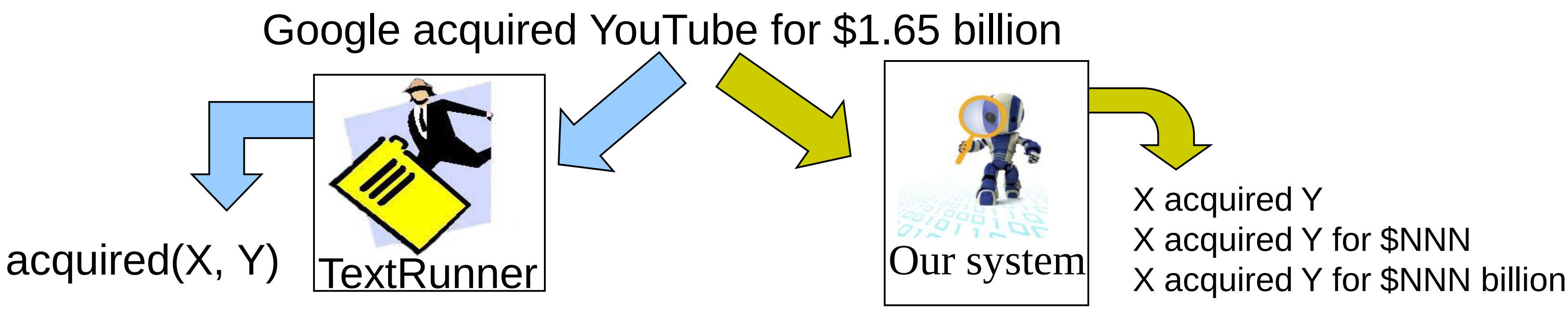
Method overview

- Representation of semantic relations by lexical patterns
  - Seoul is the capital of South Korea.
    - → (Seoul, South Korea) : X is the capital of Y, X * capital * Y, …



Entity pair and relation extraction

- For latent relational search, we don't need to explicitly extract predicates as relations



M. Banko et al. The Tradeoffs Between Open and Traditional Relation Extraction, ACL'08

- We use the n-grams of the context surrounding an entity pair to represent the relation
  - With this scheme, we can precisely measure the relational similarity
    - E.g., Microsoft acquired Powerset for $100 million → X acquired Y for $NNN

Example of relation extraction for the entity pair (Microsoft, Powerset)

It is now official : Microsoft acquires San Francisco based company Powerset for $ 100M .

Tagging (NER)

It is now official : <u>Microsoft</u> acquires San Francisco based company <u>Powerset</u> for $ 100M .

Cut (outside window size = 3)

It is $\underbrace{\text{now official:}}_{\text{3 words}}$ Microsoft acquires San Francisco based company Powerset $\underbrace{\text{for \$ 100M}}_{\text{3 words}}$.

Replace entities with variables

now official : <u>X</u> acquires San Francisco based company <u>Y</u> for $ 100M

Stemming

now offici : X acquir San Francisco base compani Y for $ 100M

Generate n-grams (lexical patterns)

X acquir * Y; offici : X acquir * Y; now offici : X acquir * Y; X * compani Y for $; …

THE UNIVERSITY OF TOKYO

Entity pair – Pattern co-occurrence matrix

- We represent co-occurrences between entity pairs and patterns in a matrix

pattern \ entity pair	(Google, Youtube)	(Microsoft, Powerset)	(Apple, Steve Jobs)	(Microsoft, Ballmer)
X acquires Y	10	9	0	0
X buys Y	8	5	0	0
Y CEO X	0	0	10	7
Y chief executive X	0	0	3	8

Number of Co-occurrences

THE UNIVERSITY OF TOKYO

For cross-lingual search

Multi-lingual entity pair and lexical pattern indexing

Patterns \ Entity pair	(Google, YouTube)	(Microsoft, Powerset)	(Rakuten, Infosiku)	(Guguru, YouChubu)
X ga Y wo baishu shita	30	10	400	350
X ga Y wo katta	20	8	250	190
X acquired Y	200	300	0	0
X purchased Y for * $	130	180	0	0
X buys Y	80	60	0	0

> 0

Number of co-occurrences

In Japanese, the transliteration of the named entity "Google" is "グーグル" (Guguru)
However, sometime Japanese use the identical surface form of an entity with English.

THE UNIVERSITY OF TOKYO

Measuring the relational similarity between two entity pairs

- Relational similarity ($pair_1$, $pair_2$) = cosine of their feature vectors
  - relsim((*Tokyo*, *Japan*), (*Paris*, *France*)) is expected to be high

- However, this trivial method does not work well because a semantic relation can be expressed by multiple lexical patterns
  - Tokyo *is the largest city in* Japan.
  - Paris *is the biggest city in* France.

LARGEST_CITY(X, Y)

Data sparseness problem!

Solution for monolingual case: pattern clustering

[D. Lin et al. KDD2001, Bollegala et al. WWW2009]

- Lexical patterns that co-occur with similar sets of entity pairs are semantically similar (Distributional hypothesis)

Similar
lexical pattern:
Similarity $\geq \theta$

X is the CEO of Y
(Steve Ballmer, Microsoft): freq = 50 (Larry Page, Google): freq = 75 (Steve Jobs, Apple) : freq = 80 …

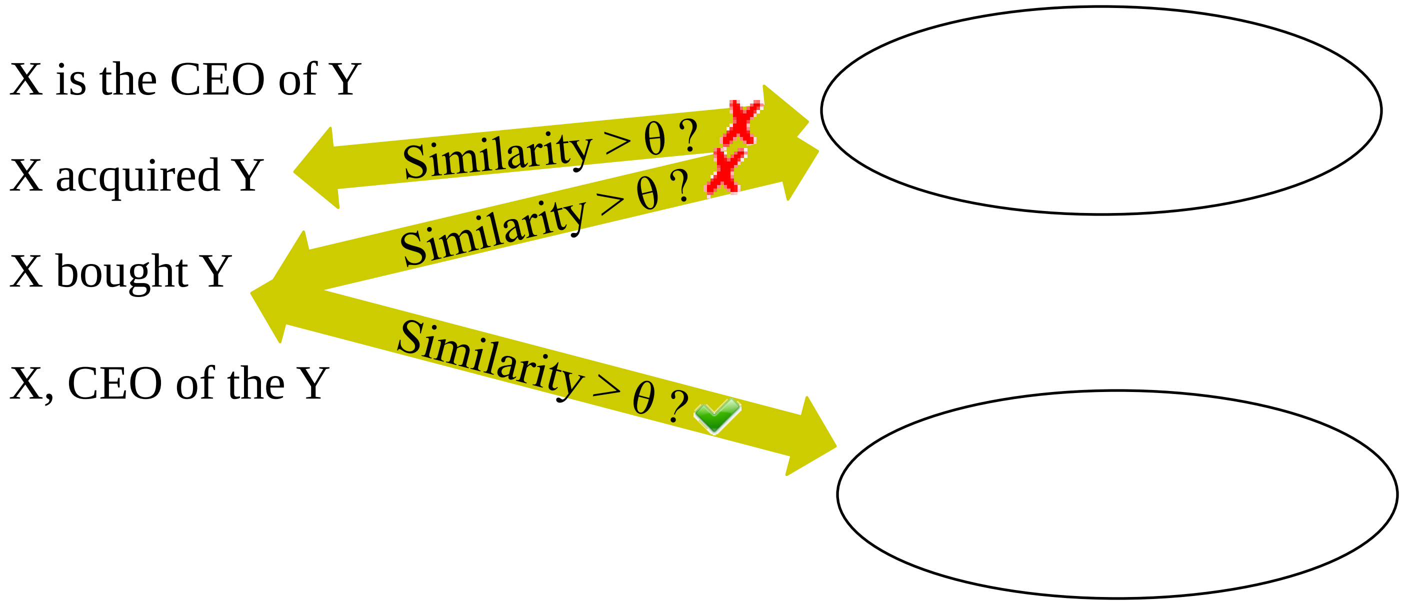
Y's CEO is X
(Steve Ballmer, Microsoft): freq = 82 (Larry Page, Google): freq = 60 (Steve Jobs, Apple) : freq = 90 …

- Group semantically similar patterns into a cluster and consider patterns in a cluster as identical when measuring the relational similarity between two entity pairs.

D. Lin, P. Pantel. DIRT - Discovery of Inference Rules from Text, KDD2001

D. Bollegala, Y. Matsuo, M. Ishizuka. Measuring the Similarity between Implicit Semantic Relations from the Web, WWW2009

The pattern hard clustering algorithm

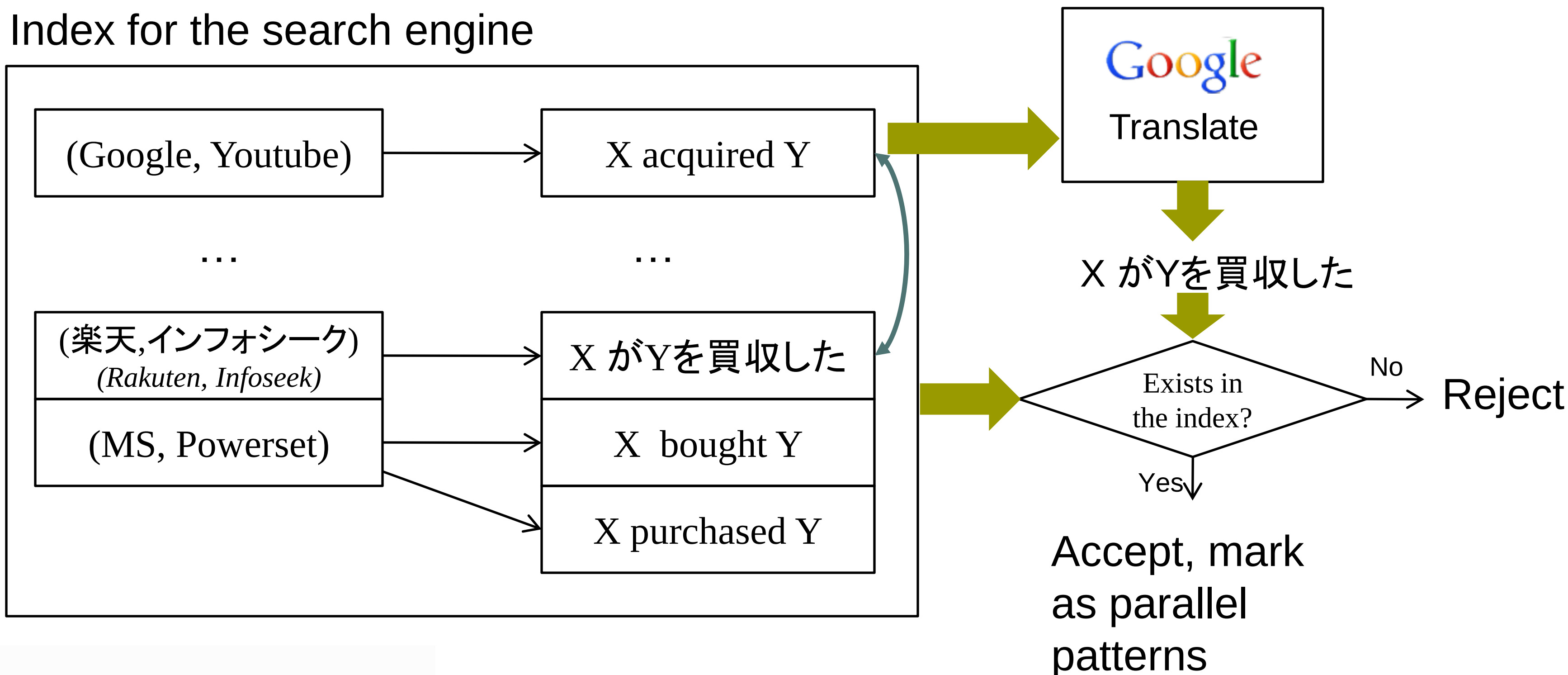


- Each pattern is assigned to only one cluster
  - Recognizing paraphrased lexical pattern in the same language

For cross-lingual search

Lexical pattern translation

- We use Google Translate for translation of entities and lexical patterns
  - Method to verify the translation result: look it up in the index



For cross-lingual search

Merging parallel patterns (rows), entity pairs (columns)

Patterns \ Entity pair	(Google, Youtube)	(Microsoft, Powerset)	(Rakuten, Infoseek)	(Guguru, YouChubu)
X ga Y wo baishu shita	30	10	400	350
X ga Y wo katta	20	8	250	190
X acquired Y	200	300	0	0
X purchased Y for * $	130	180	0	0
X buys Y	80	60	0	0

parallel

After merging, the cosine similarity between "X buys Y" and "X ga Y wo baishu shita" ("X acquired Y") is increased

Parallel pattern sparseness problem

- Very small number of parallel patterns

With machine translation, we can only find this

X acquired Y

Machine translation

Human writer

X が Y を買収した

X が Yを買収した
XがYを買った
YがXに買収された
…

- Exactly matched pattern sparseness problem
  - Many paraphrased patterns in the same language:
    X acquired Y, X bought Y, Y merged with X, …

A parallel pattern has a smaller similarity than non-parallel patterns

X bought Y ➔ { (Google, Youtube), (MS, Powerset), (Ebay, EachNet), (Apple, Emagic)}

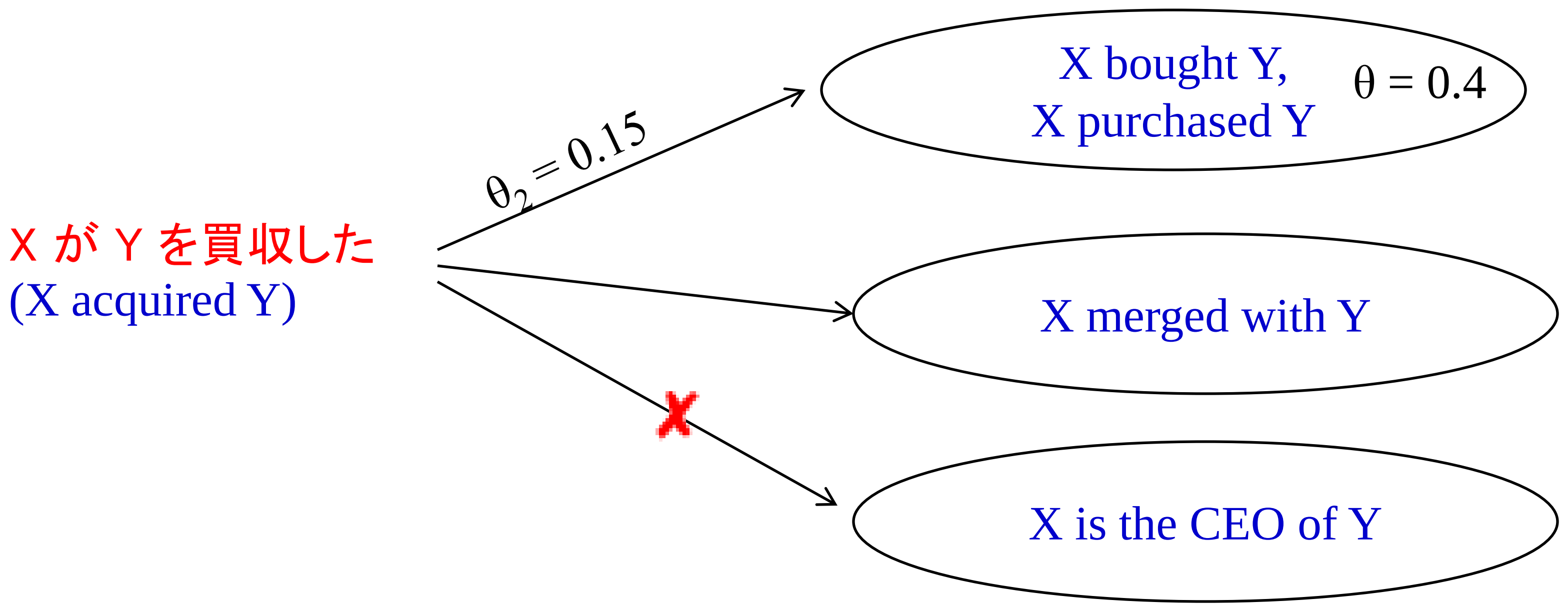
X purchased Y ➔ { (Google, YouTube), (MS, Powerset), (Ebay, EachNet)}

[X acquired Y, XがYを買収] ➔ { (Google, Youtube), (MS, Powerset), (Guguru, YouChubu)}

- Therefore, a pattern with parallel partners needs a smaller clustering similarity threshold θ_2 to be grouped into an appropriate cluster.

Proposal: use a soft-clustering step with smaller θ

- Purpose: associate as many paraphrased parallel patterns as possible to a cluster



The hybrid pattern clustering algorithm

X acquired Y
X bought Y
X purchased Y
X to acquire Y
X is the capital of Y
X is Y's capital
X は Y の首都
X がYの首都
XがYを買収した

Hard clustering
(θ large)

Recognizing paraphrased lexical pattern in the same language

X acquired Y
X to acquire Y

X bought Y
X purchased Y

X is Y's capital
X is the capital of Y

X は Y の首都
X がYの首都

XがYを買収した

The hybrid pattern clustering algorithm

X acquired Y
X bought Y
X purchased Y
X to acquire Y
X is the capital of Y
X is Y's capital
X は Y の首都
X がYの首都
XがYを買収した

Recognizing similar lexical patterns across languages

soft clustering of parallel patterns (small θ_2)

X acquired Y
X to acquire Y

X bought Y
X purchased Y
XがYを買収した

X is Y's capital
X is the capital of Y
X は Y の首都

X は Y の首都
X がYの首都
X is the capital of Y

XがYを買収した
X acquired Y

THE UNIVERSITY OF TOKYO

Naïve method for measuring the similarity between two lexical patterns

Patterns \ Entity pair	(Google, YouTube)	(Microsoft, Powerset)	(Rakuten, Inforseek)	(Google Inc., YouTube)
X ga Y wo baishu shita	30	10	400	10
X ga Y wo katta	20	8	250	5
X acquired Y	2	300	0	100
X purchased Y for * $	3	180	0	200
X buys Y	80	60	0	30

Cosine similarity

- Room for improvement:
  - Synonyms, similar words, similar entity pairs
    - If we can compress them into a dimension …

Use SVD here!

THE UNIVERSITY OF TOKYO

Candidate retrieval and ranking

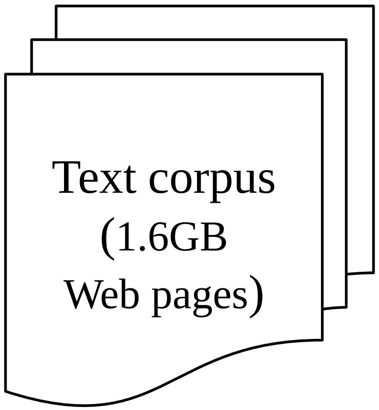
- Use cosine similarity, but lexical patterns in the same cluster are considered as in the same dimension

Patterns \ Entity pair	(**Ebay, EachNet**)	(**Microsoft, Ballmer**)	(**Rakuten, Infoseek**)	(**Guguru, YouChubu**)
X ga Y wo baishu shita, X acquired Y	100	0	400	350
X ga Y wo katta	0	0	250	190
X purchased Y for * $	130	0	0	0
X is the CEO of Y	0	60	0	0

cosine

Evaluation

- Data set



Relation type	Example
Capital	(Paris, France), (東京, 日本) …
CEO	(Apple, Steve Jobs), (トヨタ, 豊田章男) ..
Birthplace	(Albert Einstein, Ulm), (浅田真央, 愛知)
Headquarters	(Microsoft, Redmond), (任天堂, 京都) ...
Satellite	(Moon, Earth), (オベロン, 天王星) …
President	(Barack Obama, U.S), (李明博, 韓国) …
Prime Minister	(David Cameron, U.K), (菅直人, 日本) …
Acquisition	(Google, YouTube), (楽天, インフォシーク)

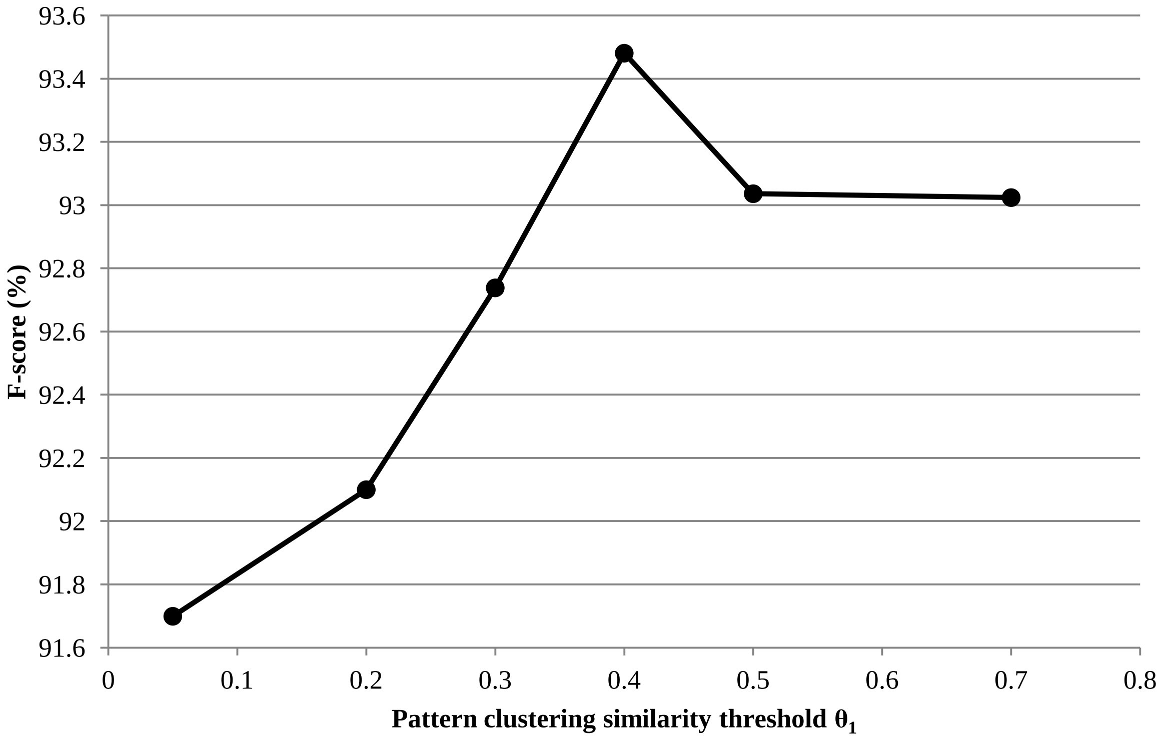
- Metric: MRR
  - Mean reciprocal rank
  - For a query set Q:

$$\mathrm{MRR}(Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r_q}$$

(r_q is the rank of the first answer of the query $q \in Q$)

Determine an appropriate value for θ_1

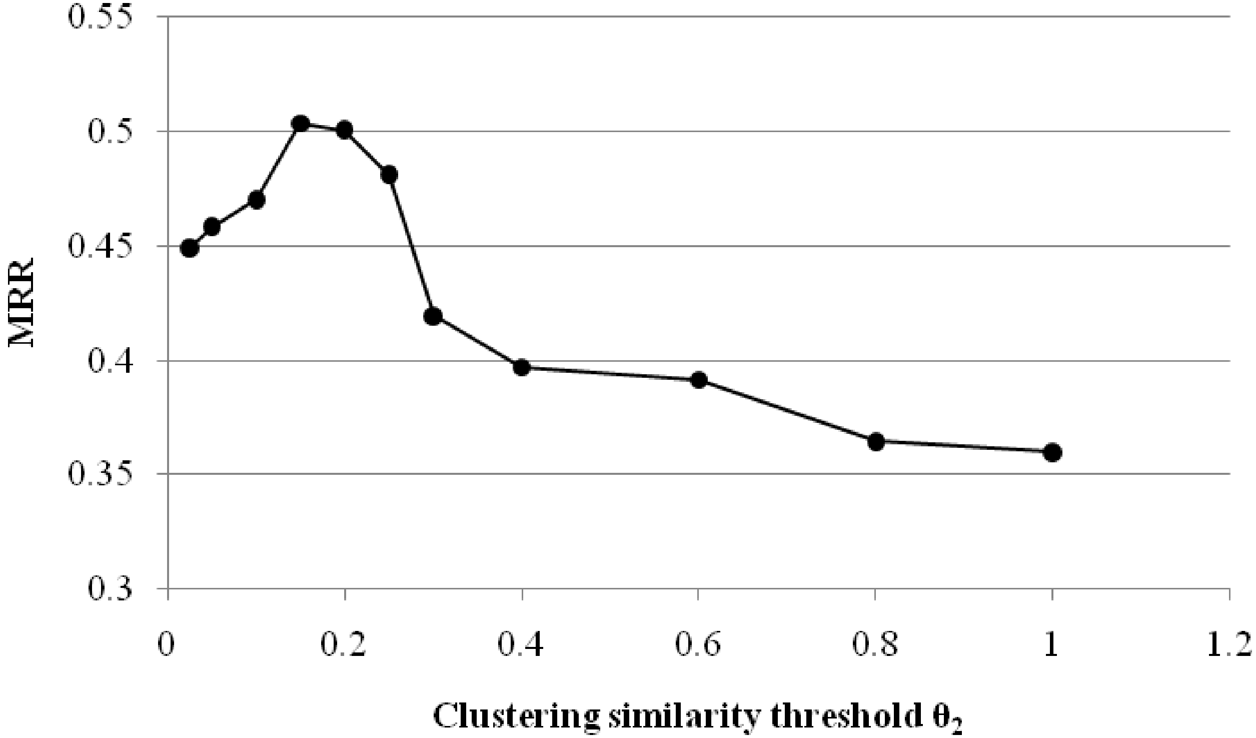
- θ_1 is the similarity threshold for the hard clustering step
  - To recognize paraphrased lexical patterns in the same language



- At $\theta_1 = 0.4$, we achieve the best result for monolingual query sets

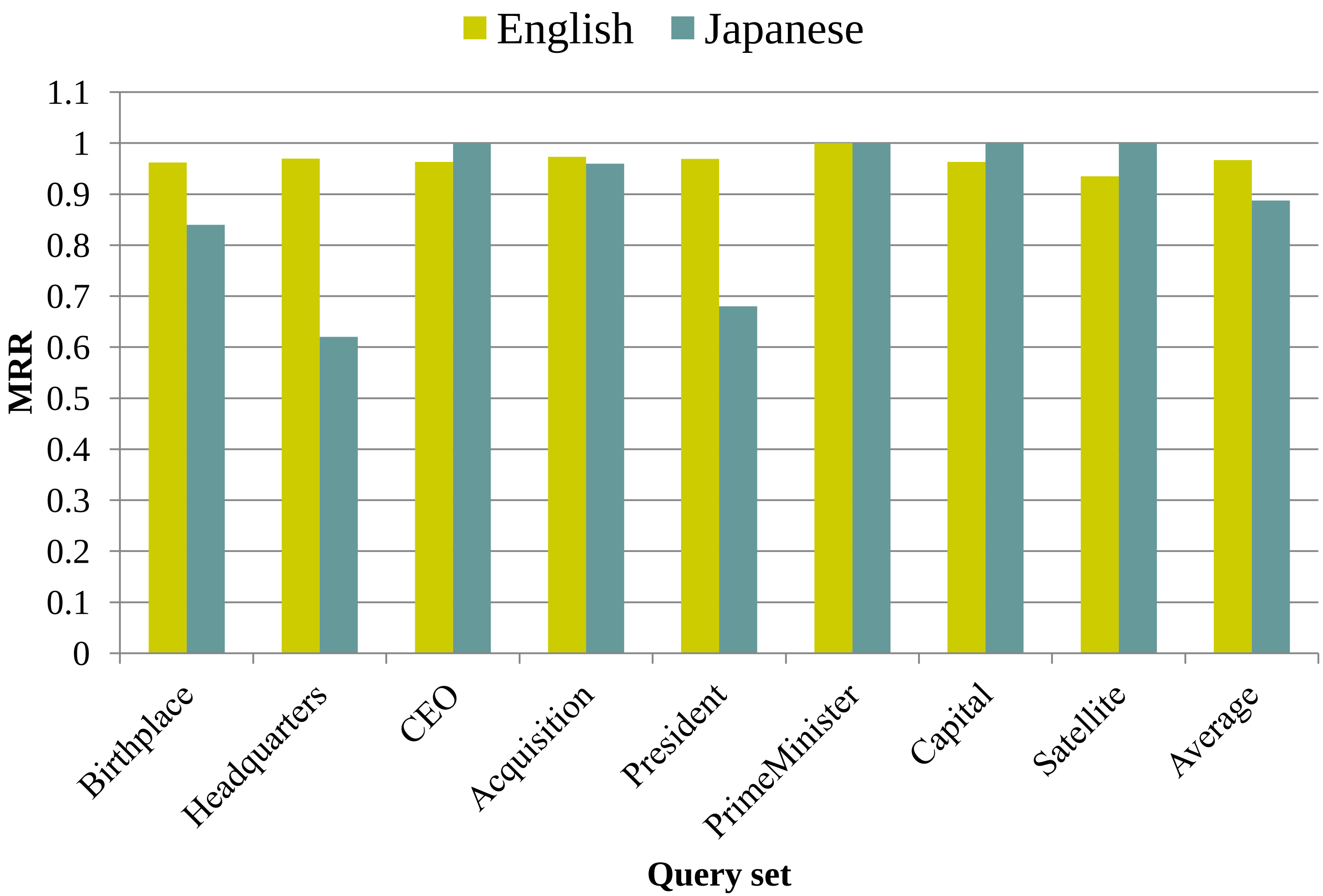
Adjusting the parameter θ_2

- θ_2 is the similarity threshold in the soft clustering step
  - To recognize paraphrased lexical patterns across languages



- At $\theta_2 = 0.15$, we achieve the best average performance for cross-language query sets

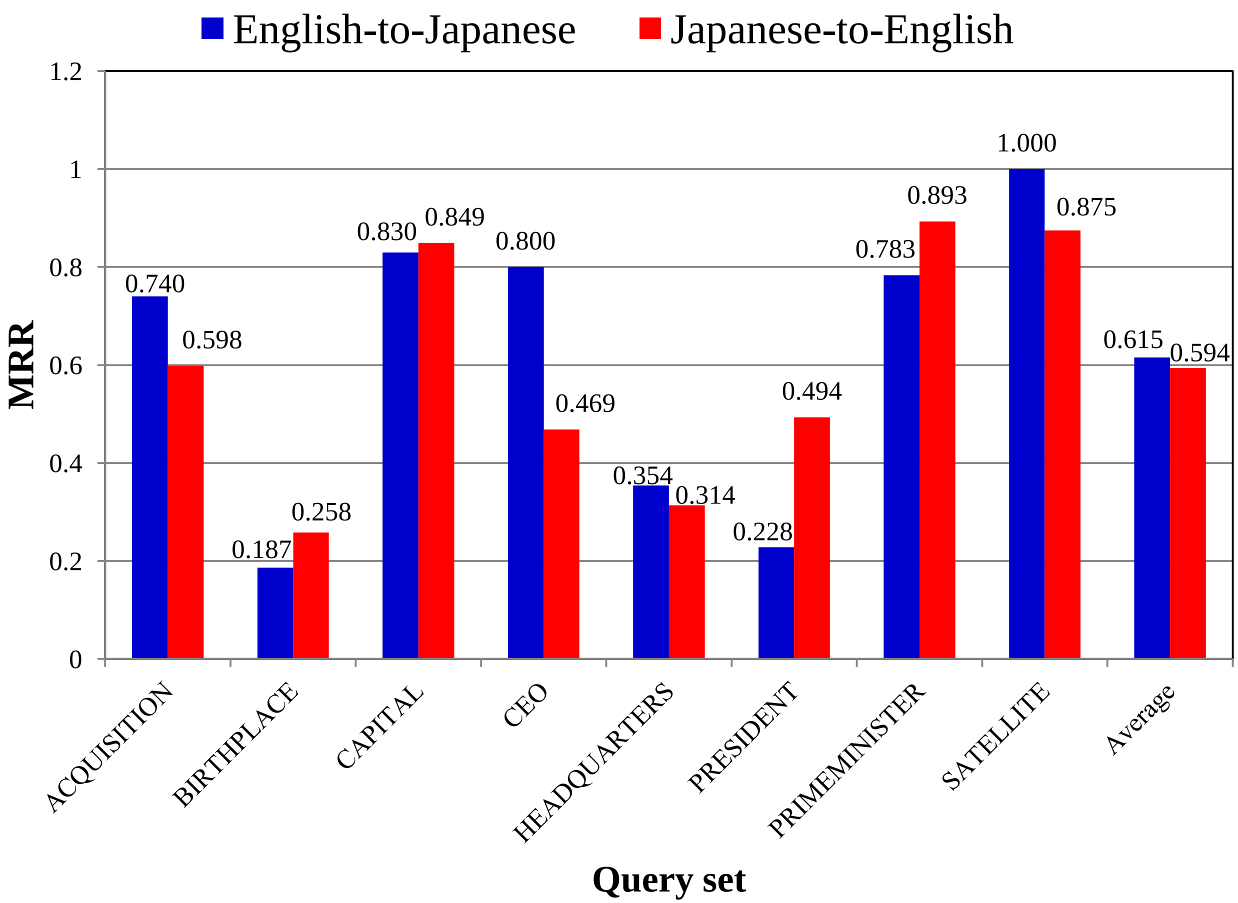
Performance on monolingual query sets



- We achieve very high MRR on monolingual query sets

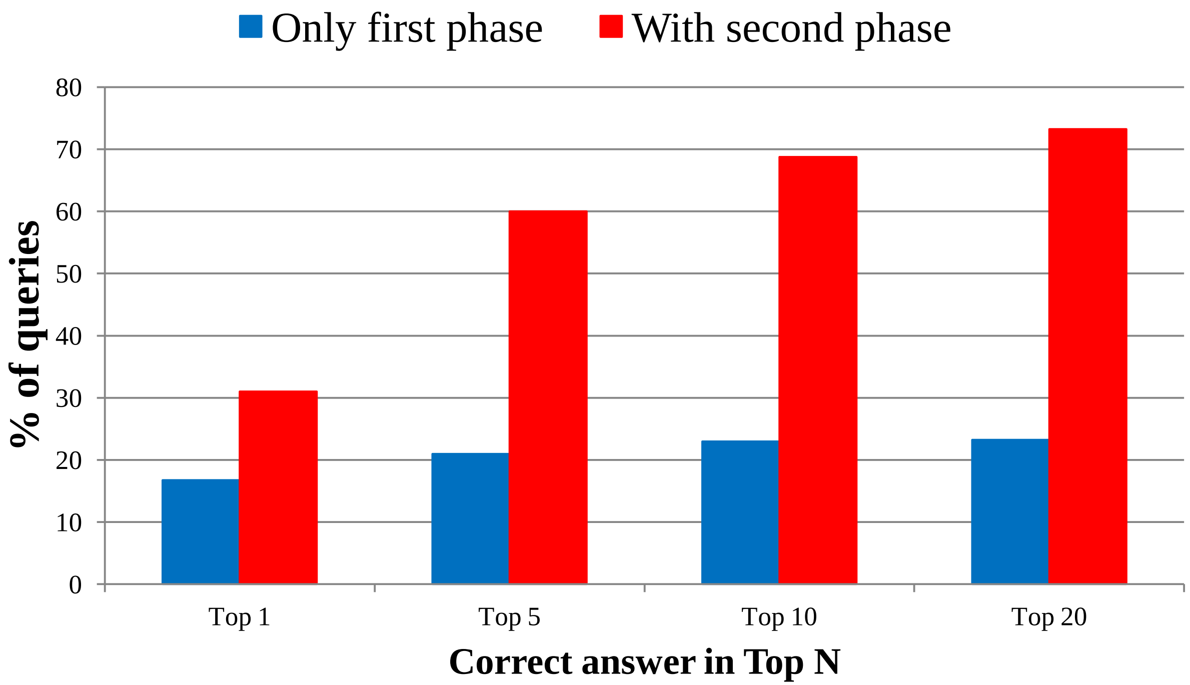
Performance on cross-language query sets

- An MRR of 0.6 on cross-language query sets



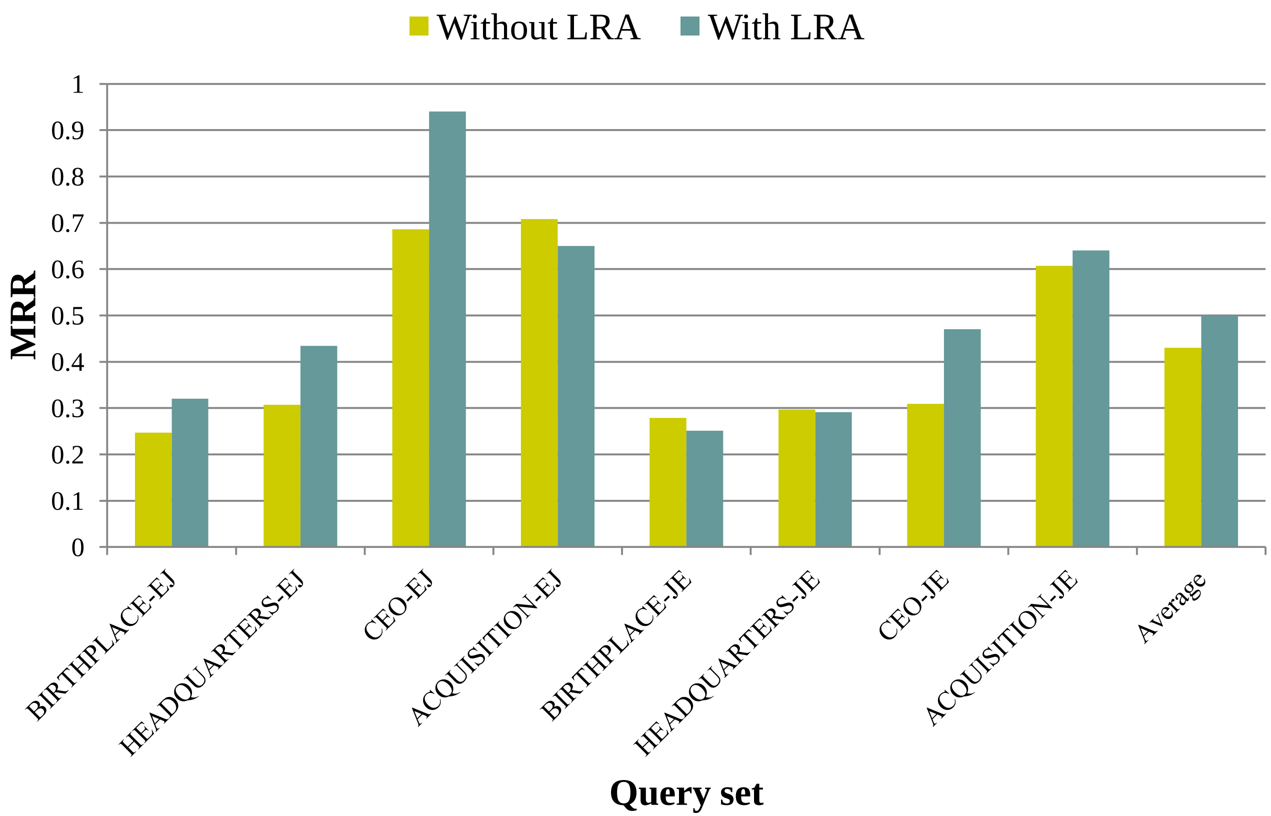
THE UNIVERSITY OF TOKYO

Effect of the soft clustering step



With the soft clustering phase (the second phase) : MRR = 0.430
Without the soft clustering phase : MRR = 0.186

Effect of LRA (SVD)



On average, the average MRR of eight query sets is improved from 0.43 to 0.50
(statistically significant under a paired t-test of 400 samples)

Performance of the proposed method and existing methods

Method	MRR	Top 1	Top 5	Top 10	Top 20
Kato et al. 2009 [JJ]	0.545	43.3	68.3	72.3	76.0
Proposed [EE]	0.971	94.9	99.9	100	100
Proposed [JJ]	0.889	87.0	91.0	91.0	91.0
Doc. Trans. (Baseline) [Cross]	0.345	30.5	39.3	40.8	42.0
Proposed [Cross]	0.605	49.8	74.5	78.5	82.0

Top N means the percentage of queries with correct answer in the Top N results.
JJ: Japanse-Japanese monolingual queries
EE: English-English monolingual queries

- Doc. Trans. (Baseline) : Translating all documents into English, then monolingual search

Kato et al. 2009 : Query by Analogical Example: Relational Search Using Web Search Engine Indices (CIKM '09)

Potential applications of latent relational search

Product search, Location search

- Very effective when a user does not know the exact keywords to formulate a query for keyword-based Web search engines.

No need to know keywords such as “Android” or “smart phone”, …

(Apple, iPhone) (Google, ?)

Japan

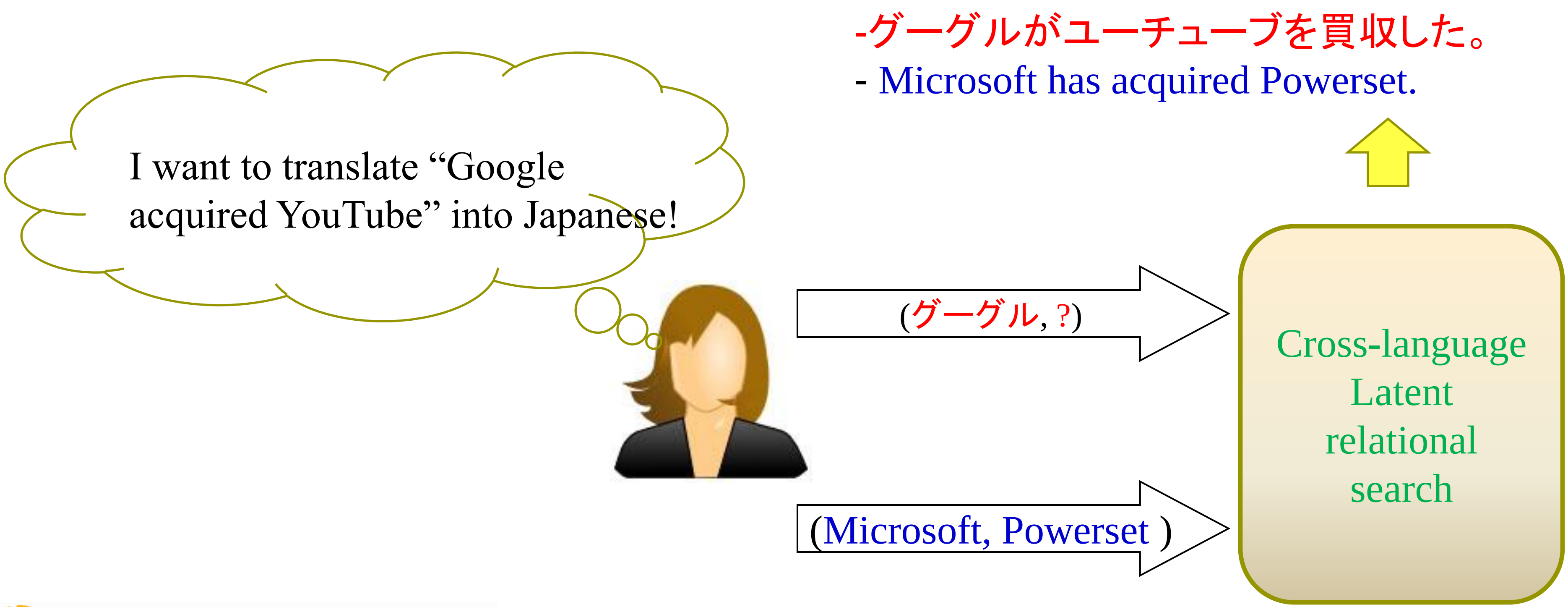
California

Mt. Fuji

?

Supporting human translators

- The evidences (supporting sentences) provide interesting human-quality examples sentences that mentioned the relation in multi-languages



Recommend friends in Social Networks

- {(Peter, Alice), (John, ?)} ⇨ Output: Anna

Recommend Anna for John!

- This kind of recommendations might be applied when John is viewing the profile of Peter.

Conclusion

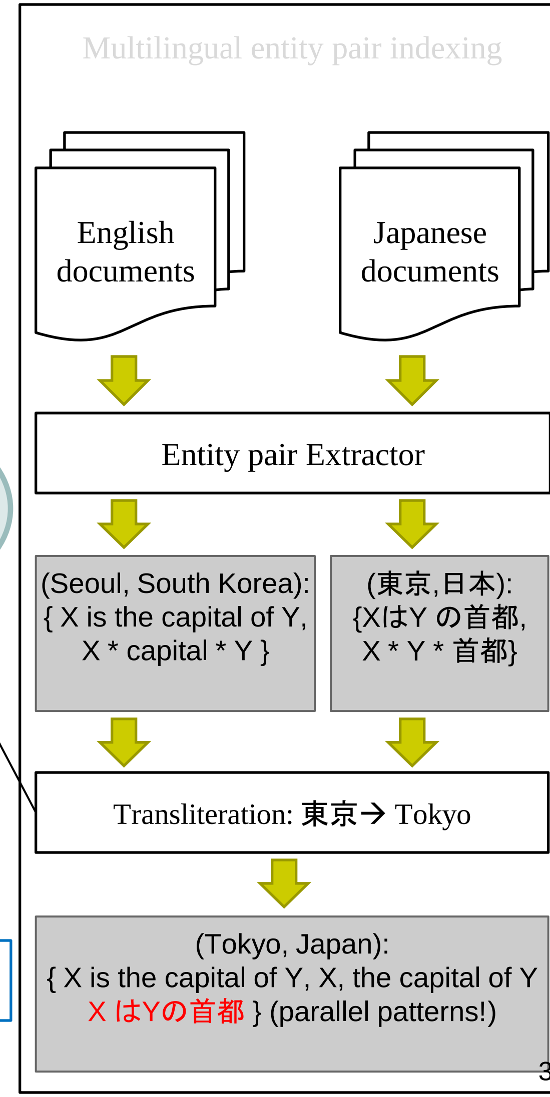
- We have presented Latent relational search, a new entity retrieval paradigm
  - Using relational similarity for ranking

- We achieve high MRR on monolingual latent relational search, an moderate performance on cross-language latent relational search

- We discuss many applications of latent relational search, such as product search or provide parallel sentences for human translators.

Help wanted:

- Representation of semantic relations by lexical patterns
  - Seoul is the capital of South Korea.
    - → (Seoul, South Korea) : X is the capital of Y, X * capital * Y, …



We are blocked if we issue a large number of queries to Google Translate!
We'd appreciate if you could allow us to freely access to Google Translate.

Google Translate

Pattern translation

(Seoul, South Korea)
X is the capital of Y
X is Y's largest city
…

(東京, 日本)
X はYの首都である
X がYの最大都市
…

Thank you!

For a live demo, please visit

http://www.miv.t.u-tokyo.ac.jp/duc/milresh/

or google for “latent relational search”!