PRICAI 2010 Plenary Talk

Exploiting Macro and Micro Relations Toward Web Intelligence

Mitsuru Ishizuka

School of Information Science and Technology

THE UNIVERSITY OF TOKYO

In this Talk

- Relations between entities are basic elements for representing knowledge, such as in semantic net, logic, etc.
- In Web intelligence, the extraction or mining of meaningful knowledge and the utilization of the knowledge for intelligent services are key issues.
- I will present some of our researches related to these issues, ranging from macro relations to micro ones.

Outline

1. Social Relation Extraction
2. Relational Similarity between Two Word Pairs
   - (2.1) Computing Relational Similarity
   - (2.2) Latent Relational Search Engine
   - (2.3) Open Relation Extraction employing Sequential Co-clustering
3. Common and Universal Concept Description Language as a Foundation of Semantic Computing

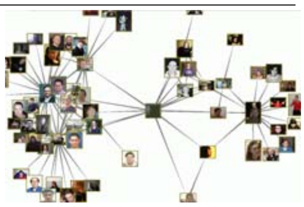
---

1. Social Relation Extraction

Message from Social Network Study

- Attribute data

  and
- Relational data

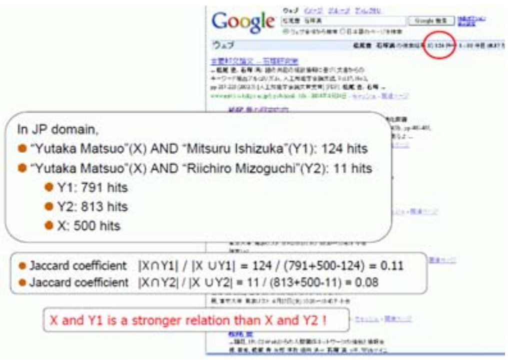


- Relational data is important as well as Attribute data, to assess the role/characteristics of an entity (a person) in a social network.
- Relational data are represented as ties/connections, and reveal neighbors (friends), a position in the network structure, etc.

Human Relation Mining from the Web

Basic Idea : the Use of Co-occurrence

Comparing Co-occurrence (Hits)



In JP domain,
- "Yutaka Matsuo"(X) AND "Mitsuru Ishizuka"(Y1): 124 hits
- "Yutaka Matsuo"(X) AND "Riichiro Mizoguchi"(Y2): 11 hits
  - Y1: 791 hits
  - Y2: 813 hits
  - X: 500 hits

- Jaccard coefficient $|X \cap Y1| / |X \cup Y1| = 124 / (791+500-124) = 0.11$
- Jaccard coefficient $|X \cap Y2| / |X \cup Y2| = 11 / (813+500-11) = 0.08$

X and Y1 is a stronger relation than X and Y2 !

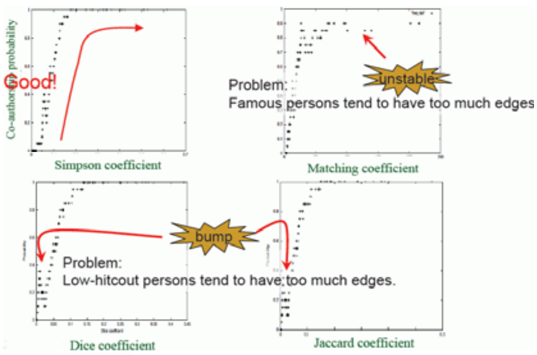
Measures of Co-occurrence

- Matching coefficient $|X \cap Y|$
- Mutual Information $\log N|X \cap Y| / |X||Y|$
- Dice coefficient $2|X \cap Y| / (|X|+|Y|)$
- Jaccard coefficient $|X \cap Y| / |X \cup Y|$
- Cosine $|X \cap Y| / (\sqrt{|X||Y|})$
- Simpson coefficient (overlap coefficient) $|X \cap Y| / \min(|X|,|Y|)$

- with a cutoff threshold on |X| and |Y|.

$$f(X,Y)=\begin{cases}|X\cap Y|/\min(|X|,|Y|) & if\ |X|>k\ and\ |Y|>k\\ 0 & otherwise\end{cases}$$

Comparing Co-occurrence Measures (in the case of Co-authorship relation)



Relation Type Recognition

- For the case of JSAI conf. participants, we defined four types (classes) of relations:
  - Coauthor : co-author of a paper
  - Lab : members of the same lab. or research institute
  - Proj : members of the same project or committee
  - Cof : participants of the same conf. or workshop

- We designed a decision-tree (C4.5) classifier for these relation types.

Attribute Features for C4.5 Classifier

- No. of Co-occurrence. (one, more_than_two)
- Two names appear in one line more than once. (yes, no)
- The strength of the relationship is larger than a threshold. (yes, no)
- Occurrence of Name-1. (one, more_than_two)
- Occurrence of Name-2. (one, more_than_two)
- A word in the word cluster A～F appears in the title. (zero, more_than_one)
- A word in the word cluster A～F appears in the first 5 lines. (zero, more_than_one)

- **Word Clusters**
  - **Cluster A:** "publication papers" "publications" "achievement" "research activities" "publication themes" "award" "authors"
  - **Cluster B:** "laboratory members" "group" "team members"
  - **Cluster C:** "project" "committee"
  - **Cluster D:** "conference" "symposium" "workshop" "seminar" "research meeting" "co-sponsors"
  - **Cluster E:** "society" "program" "journal" "session" "contexts"
  - **Cluster F:** "professor" "lecture" "teaching staff" "research student"

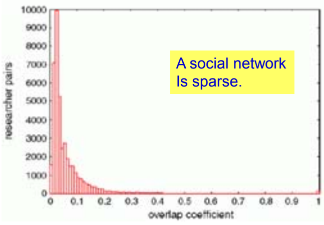
The Performance of the Classifier

- 275 training samples and 200 evaluation samples from JSAI2003 participant data are used for the performance evaluation.

Class	Error rate	precision	recall
Coauthor	4.1%	91.8% (90/98)	97.8% (90/92)
Lab	25.7%	70.9% (73/103)	86.9% (73/84)
Proj	5.8%	74.4% (67/90)	91.8% (67/73)
Conf	11.2%	89.7% (87/97)	67.4% (87/129)

Scalability Issue

- Too many queries are issued to a search engine.
  - Assume we have n names. Then, ${}_nC_2$ or $O(n^2)$ queries become necessary.
  - For 500 people, 124,750 queries….
    - cf) Google API (1,000 queries/day, Yahoo! API (5,000 queries/day)
- Distribution of Simpson Coefficient
  - 0: approx. 67%
  - 0〜0.2: approx. 98%

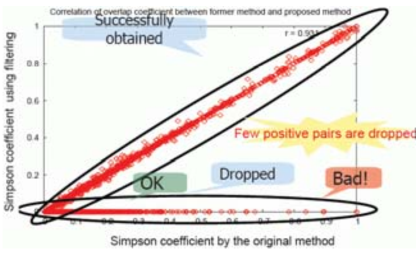


Idea for Scalability

- Filter out pairs of persons that seem to have no relation.
- Apply Google Co-occurrence Count only for promising pars after investigating the texts of Google top 10 pages.

Evaluation of the Filtering
(JSAI conf. participants case)

- Originally 126,253 queries ($O(n^2)$) for 504 researches.
- By applying filtering, 19,182 queries: $O(n)$



Keyword Extraction for a Person

- Word-to-Person Co-occurrence

	agent	mining	communication	audio	cognition	…
Mitsuru Ishizuka	454	143	414	382	840	…
Koiti Hasida	412	156	1020	458	1150	…
Yutaka Matsuo	129	112	138	89	58	…
Nobuaki Minematsu	227	22	265	648	138	…
Yohei Asada	6	6	6	2	0	…
…	…	…	…	…	…	…

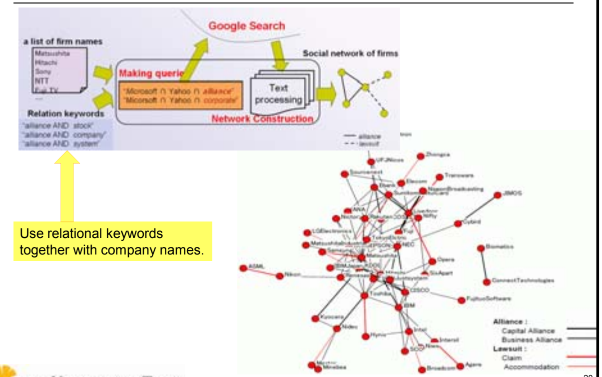
- **Keyword Extraction for a Person**
- **Clustering of Persons**

POLYPHONET *for a research community*

SPYSEE *for public* (operated by Ohma Inc., Tokyo)

SPYSEE

Company Relation Extraction
(Alliance and Lawsuit Relation cases)



Some Other Related Work in Social Network Mining

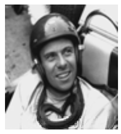
- Referral Web (H. Kautz et al, 1997)
  - A name is given as input
  - Retrieve the name, and extract other names.
  - Measure co-occurrence (by Jaccard coefficient), and invent edges.
  - Ego-centric network within 2-3 radius
    - E.g, find a path from Henry Kautz to Marvin Minsky
- Flink (P. Mika, 2004)
  - Email messages, publications, FOAF documents, and Web mining
  - Web mining part
    - Similar to Referral Web
    - Jaccard coefficient
    - "Semantic Web OR Ontology" is added to a query for disambiguation.
- A. McCallum et al. (2004-)
  - Identify people in e-mail messages, and find homepages
    - Links are placed between the owner of Web page and persons discovered on the page.
  - They also use co-occurrence on the entire Web
    - with name-disambiguation probability model
- Other studies using co-occurrence information
  - [Harada04] [Faloutsos04] [Kees04]…

Some Related Technologies

- Namesake disambiguation

Jim Clark(s)

Founder, Silicon Graphics and Netscape



F1 racer

- Alias Name detection

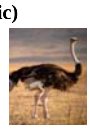
Hideki Matsui
Godzilla (Matsui)

2. Relational Similarity between Two Word Pairs

(2.1) Computing Relational Similarity
(2.2) Latent Relational Search Engine
(2.3) Open Relation Extraction employing Sequential Co-clustering

Attributional vs. Relational Similarity

- **Attributional Similarity:**
  - Correspondence between attributes of two words/entities
  - e.g., automobile vs. car — sim(X,Y)
- **Relational Similarity:**
  - Correspondence between relations between word/entity pairs — sim(A,B, X,Y)
  - e.g., (Ostrich, Bird) vs. (Lion, Cat)
    - *X is a large Y*
  - (word, language) vs. (note, music)
    - *Y is composed using X*



Analogy in AI

- **Structure Mapping Theory (SMT)** (Gentner, *Cognitive Science*, 1983)
  - Analogy is a mapping of knowledge from one domain (the base) into another (the target) which conveys that a system of relations known to hold in the base also holds in the target.
- **Mapping rules:** M: $b_i \rightarrow t_i$
  - Attributes of objects are dropped
    - RED(b_i) RED(t_i)
  - Certain relations between objects in the base are mapped to the target
    - REVOLVES(EARTH,SUN) →REVOLVES(ELECTRON,NEUCLEUS)
  - ***Systematicity principle***: base predicate that belongs to a mappable system of mutually constraining interconnected relations is more likely to be mapped to the target domain.
    - CAUSE[PUSH(b_i,b_j), COLLIDE(b_j,b_k)] → CAUSE[PUSH(t_i,t_j), COLLIDE(t_j,t_k)]

Computing Relational Similarity

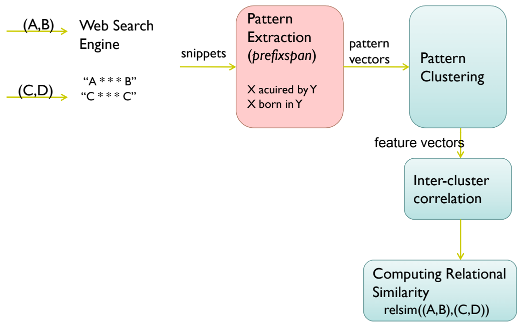
- **Turney's Work using LSA (Latent Semantic Analysis)**
  (Turney, ACL 2006)
  - **(traffic, road) vs. (water, river)**

X *flows in* Y

Challenges in Computing Relational Similarity and Our Approach

How to explicitly state the relation between two entities? / **How to extract the multiple relations between two entities?**	• **Extract lexical patterns from contexts where the two entities co-occur**
A single semantic relation can be expressed by multiple patterns.	• E.g. "ACQUISITION": X *acquires* Y, Y *is bought by X* • **Cluster the semantically related lexical patterns into separate clusters.**
Semantic Relations might not be independent.	• E.g. IS-A and HAS-A. Ostrich is a bird, Ostrich has feathers • **Measure the correlation between various semantic relations** **Mahalanobis Distance vs. Euclidian Distance**
The contribution of different semantic relations towards relational similarity is unknown	• **Learn the contribution of different semantic relations using training data** Information Theoretic Metric Learning (ITML) (Davis 2008)

Outline of the proposed method



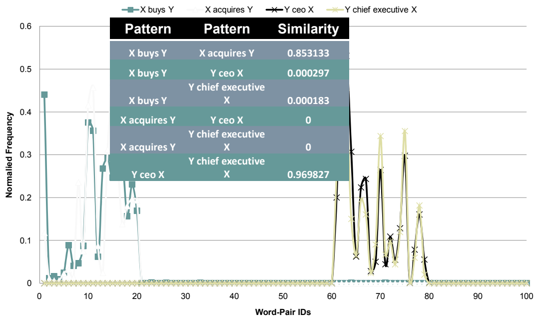
Pattern Extraction

- We use **prefix-span**, a sequential pattern mining algorithm, to extract patterns that describe various relations, from text snippets returned by a web search engine.
- query = lion * * * * * * * cat
- snippet = .. **lion**, a large heavy-built social **cat** of open rocky areas in Africa ..
- patterns = X, *a large* Y / X a large Y / X a Y / X a large Y of
- **Prefix-span algorithm is used to extract patterns:**
  - Efficient
  - Considers gaps
- **Extracted patterns can be noisy:**
  - misspellings, ungrammatical sentences, fragmented snippets

Clustering the Lexical Patterns

- We have approx. 150,000 patterns that express various semantic relations.
- However, a single semantic relation is expressed by more than one lexical patterns.
- How to identify the patterns that express a particular semantic relation?
  - **Distributional Hypothesis** **(Harris 1957)**
    Patterns that are equally distributed among word-pairs are semantically similar.
- We can **cluster** the patterns according to their distribution in word-pairs.
  - Pair-wise comparison is computationally expensive.
  - **Propose a greedy sequential pattern clustering algorithm**.

Distribution of patterns in word-pairs



Pattern	Pattern	Similarity
X buys Y	X acquires Y	0.853133
X buys Y	Y ceo X	0.000297
X buys Y	Y chief executive X	0.000183
X acquires Y	Y ceo X	0
X acquires Y	Y chief executive X	0
Y ceo X	Y chief executive X	0.969827

Greedy Sequential Clustering
for large lexical pattern data (approx.150,000)

1. **Sort the patterns according to their total frequency in all word-pairs.**
2. **Select the next pattern:**
   1. Measure the similarity between each of the existing clusters and the pattern.
   2. If the similarity with the most similar cluster is greater than **a threshold θ**, then add to that cluster, otherwise form a new cluster with this pattern.
   3. Repeat until all patterns are clustered.
3. **We view each cluster as a vector of word-pair frequencies and compute the cosine similarity between the centroid vector and the pattern.**

- Properties of the clustering algorithm
  - Scales linearly with the number of patterns $O(n)$
  - More general clusters are formed ahead of the more specific clusters
  - Only one parameter to be adjusted (clustering threshold θ)
  - No need to specify the number of clusters
  - Does not requite pair-wise comparisons, which are computationally costly
  - A greedy clustering algorithm

Computing Relational Similarity

- The formed pattern clusters might not be independent because,
  - Semantic relations can be mutually dependent.
  - The Greedy Sequential Clustering algorithm might split a semantic relation into multiple clusters.
- Euclidean distance (Cosine similarity) cannot reflect the correlation between pattern clusters.
  - We use **Mahalanobis distance** to measure the relational similarity.
  - Mahalanobis distance between two vectors **x** and **y** is defined by,

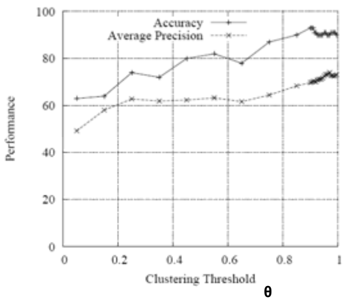
$$(x-y)^t A (x-y)$$

  where A is the covariance matrix.
  - Using a labeled dataset of positive and negative instances, we learn the Mahalanobis distance metric.
    - Information Metric Learning algorithm [Davis et. al. 2007]

Dataset-1 for experiments

- **ENT dataset**
  - We created a dataset that has 100 entity-pairs covering five relation types. (20x5 = 100)
  - **ACQUIRER-ACQUIREE** (e.g. [*Google, YouTube*])
  - **PERSON-BIRTHPLACE** (e.g. [*Charlie Chaplin, London*])
  - **CEO-COMPANY** (e.g. [*Eric Schmidt, Google*])
  - **COMPANY-HEADQUARTERS** (e.g. [*Microsoft, Redmond*])
  - **PERSON-FIELD** (e.g. [*Einstein, Physics*])
- approx. 100,000 snippets are downloaded for each relation type.
  - 473,910 lexical patterns were extracted.
  - From these patterns, we selected 148,655 patterns that occur at least twice.

Setting the threshold θ in the Clustering



When θ=0.905, we obtain 6354 non-singleton clusters, and 4093 singletons (10,447 in total).

Pattern Clusters

cluster 1 (2868)	X acquires Y	X has acquired Y	X's Y acquisition	X, acquisition, Y	Y goes X
cluster 2 (2711)	Y legend X was	X's championship Y	Y star X was	X autographed Y ball	Y star X robbed
cluster 3 (2615)	Y champion X	world Y champion X	X teaches Y	X's greatest Y	Y players like X
cluster 4 (2008)	X to buy Y	X and Y confirmed	X buy Y is	Y purchase to boost X	X is buying Y
cluster 5 (2002)	Y founder X	Y founder and ceo X	X, founder of Y	X says Y	X talks up Y
cluster 6 (1364)	X revolutionized Y	X professor of Y	in Y since X	ago, X revolutionized Y	X's contribution to Y
cluster 7 (845)	X and modern Y	genius: X and modern Y	Y in DDDD, X was	on Y by X	X's lectures on Y
cluster 8 (280)	X headquarters in Y	X offices in Y	past X offices in Y	the X conference in Y	X headquarters in Y on
cluster 9 (144)	X's childhood in Y	X's birth in Y	Y born X	Y born X introduced the	sobbing X left Y to
cluster 10 (49)	X headquarters in Y	X's Y headquarters	Y - based X	X works with the Y	Y office of X

- **clusters 1 and 4:** (acquire - acquiree)
- **cluster 2,3, 6 and 7:** (person –field)
- **cluster 5:** (ceo – company)
- **cluster 8 and 10:** (company – headquarter)
- **cluster 9:** (person – birthplace)

Relation Classification

- We use the proposed relational similarity measure to classify entity-pairs according to the semantic relations between them.
- We compute the relational similarity between a word-pair and the remaining 99 word-pairs. Then, sort word-pairs in the descending order of the relational similarity, and select the most similar k word-pairs.
- We use k-nearest neighbor classification (k=10)

- Evaluation measures

$$\text{Accuracy} = \frac{\text{No. of correctly classified pairs}}{\text{Total no. of pairs}}$$

Evaluation of top most similar k word-pairs

1, 2/2, 3/4 ….

$$\text{Average Precision} = \frac{\sum_{r=1}^{k} \text{Precision}(r) \times \text{Relevant}(r)}{\text{No. of relevant pairs}}$$

THE UNIVERSITY OF TOKYO

37

Results - Average Precision

Relation	VSM	LRA	EUC	CORR
ACQUIRER-ACQUIREE	92.7	92.24	91.47	94.15
COMPANY-HEADQUARTERS	84.55	82.54	79.86	86.53
PERSON-FIELD	44.70	43.96	51.95	57.15
CEO-COMPANY	95.82	96.12	90.58	95.78
PERSON-BIRTHPLACE	27.47	27.95	33.43	36.48
OVERALL	68.96	68.56	69.46	74.03

- **Comparison with baselines and previous work**
  - **VSM**: Vector Space Model (cosine similarity between pattern frequency vectors)
  - **LRA**: Latent Relational Analysis (Turney '06 ACL, Based on LSA)
    - 4,000 lexical patterns → 300 patterns
  - **EUC**: Euclidean distance between cluster vectors
  - **CORR:** Mahalanobis distance between entity-pairs (**PROPOSED METHOD**)

THE UNIVERSITY OF TOKYO

38

Results – Accuracy
in 10 Nearest Neighbor Classification

Relation	VSM	LRA	EUC	CORR
ACQUIRER-ACQUIREE	100	100	100	100
COMPANY-HEADQUARTERS	100	100	100	100
PERSON-FIELD	80	80	95	95
CEO-COMPANY	100	100	100	100
PERSON-BIRTHPLACE	50	60	55	70
OVERALL	86	88	90	93

Comparison with baselines and previous work
- **VSM**: Vector Space Model (cosine similarity between pattern frequency vectors)
- **LRA**: Latent Relational Analysis (Turney '06 ACL, Based on LSA)
  - 4,000 lexical patterns → 300 patterns
- **EUC**: Euclidean distance between cluster vectors
- **CORR:** Mahalanobis distance between entity-pairs (**PROPOSED METHOD**)

THE UNIVERSITY OF TOKYO

39

Dataset-2: SAT Word Analogy Questions
{SAT: Scholastic Assessment Test}

- **SAT Analogy Questions have been used as a baseline to evaluate relational similarity measures.** (Turney RANLP 2003)
  - SAT question: Ostrich - Bird (Each question has five choices; one is correct.)
    - Lion – Cat
    - Goose – Flock
    - Ewe – Sheep
    - Cub – Bear
    - Primate – Monkey

    correct answer
  - **374 SAT word analogy questions (2178 word pairs).**
- **Average SAT score by native senior high school students: 57%**
- **WordNet-based approaches** (Veale, ECAI 2004) [43%]
- **Vector Space Model** (Turney, *Machine Learning* 2005) [47%]
- **Latent Relational Analysis** (Turney, *Computational Linguistics* 2006) [56%]

THE UNIVERSITY OF TOKYO

40

Results for SAT Dataset

Algorithm	SAT score	Algorithm	SAT score
Random guessing	0.200	LSA+Predictation	0420
Jiang & Conrath	0.273	Veale (WordNet)	0.430
Lin	0.273	Bicici & Yuret	0.440
Leacock & Chodrow	0.313	VSM	0.470
Hirst & St.-Onge	0.321	**PROPOSED**	0.511
Resnik	0.332	Pertinence	0.535
PMI-IR (Turney 2003)	0.35	**LRA** (Turney 2006)	0.561
SVM (Bollegala ECAI)	0.401	**Human**	0.570

less than 6 hours

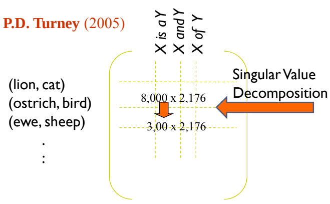
8 days!!!

THE UNIVERSITY OF TOKYO

41

Latent Relational Analysis vs. The Proposed Method

P.D. Turney (2005)



- To compute relational similarity between two word-pairs using N number of lexical patterns, LRA requires $2N$ web-queries ($N \approx 4000$)
- Proposed method requires only two web-queries and is independent of the number of patterns!

- **In LRA, for each new word-pair, we must repeat SVD**
- **No SVD is required**

THE UNIVERSITY OF TOKYO

42

Summary of our Computing Method for Relational Similarity

- **Distributional hypothesis** is useful to identify semantically similar lexical patterns.
- **Clustering lexical patterns** prior to measuring similarity improves performance.
- **Our Greedy Sequential Clustering Algorithm** efficiently produces pattern clusters for common semantic relations.
- **Mahalanobis distance** outperforms Euclidean distance when measuring similarity between semantic relations.

2. Relational Similarity between Two Word Pairs

(2.1) Computing Relational Similarity
(2.2) Latent Relational Search Engine
(2.3) Open Relation Extraction employing Sequential Co-clustering

Latent Relational Search Engine

[Japan Patent Application: 2009/12/03]

Web (text corpus)

Japan's highest mountain is Mt. Fuji.
Mt. Fuji is the highest mountain in Japan.
Germany's highest mountain is Zugspitze.

Input: (Japan, Mt. Fuji) (Germany, ?) → Relational Search Engine → Output: ? = Zugspitze

Screen Shots (1)

Query:

Word pair 1: steve jobs | apple
Word pair 2: ? | microsoft

Search

steve jobs is to **apple** as:

- ['ballmer', 'steve ballmer'] is to **microsoft** (Score = 295) Show evidence
- ['bill gates'] is to **microsoft** (Score = 52) Show evidence
- ['danny thorpe'] is to **microsoft** (Score = 27) Show evidence

Screen Shots (2)

Query:

Word pair 1: steve jobs | apple
Word pair 2: ? | microsoft

Search

steve jobs is to **apple** as:

- ['ballmer', 'steve ballmer'] is to **microsoft** (Score = 295) Hide evidence
  - Steve Jobs is the CEO of Apple, which he co-founded in 1976.
    http://www.apple.com/pr/bios/jobs.html
  - Steve Jobs is the CEO of Apple, which he co-founded in 1976.
    http://jobsearchtech.about.com/od/historyoftechindustry/a/SteveJobs.htm
  - But nothing as mundane would prompt Steve Jobs, Chief Executive, Apple, to predict that cities will be built around it.
    http://www.rediff.com/netguide/2001/dec/03ginger.htm
  - Bangalore: Steve Jobs, Chief Executive, Apple has admitted that a mechanism exists within the iPhone that enables the company to unilaterally remove software from the iPhone.
    http://www.siliconindia.com/shownews/Steve_Jobs_confirms_application_-nid-45489.html
  - In a abbreviated broadcast, SEGuru reports on the announcement that Steve Ballmer, CEO of Microsoft will deliver the keynote at SMX West 2010 in Santa Clara.
    http://www2.webmasterradio.fm/seo-rockstars/2009/microsoft-ceo-steve-ballmer-to-keynote-smx-west/
  - Steve Ballmer is the CEO of Microsoft.
    http://qna.rediff.com/questions-and-answers/who-is-the-ceo-of-microsoft/8431179/answers
  - Steve Ballmer is the CEO of Microsoft and one of the richest men in the world -LRB- $ 25 Billion -RRB- so I guess he's allowed to do pretty much what he wants including looking like a fat sweaty dancing monkey.
    http://www.sonicbomb.com/modules.php?name=News&file=article&sid=32&mode=thread&order=0&thold=0
  - The news broke after a meeting in Seattle on Saturday at which Steve Ballmer, chief executive of Microsoft, sought to persuade Jerry Yang, co-founder of Yahoo, to yield to the software group by raising his offer from $ 31 a share to $ 33.
    http://business.timesonline.co.uk/tol/business/markets/mergers_and_acquisitions/article3872866.ece

Screen Shots (3)

Query:

Word pair 1: steve jobs | apple
Word pair 2: ? | oracle

Search

steve jobs is to **apple** as:

- ['larry ellison'] is to **oracle** (Score = 20) Hide evidence
  - Steve Jobs is the CEO of Apple, which he co-founded in 1976.
    http://www.apple.com/pr/bios/jobs.html
  - Steve Jobs is the CEO of Apple, which he co-founded in 1976.
    http://jobsearchtech.about.com/od/historyoftechindustry/a/SteveJobs.htm
  - But nothing as mundane would prompt Steve Jobs, Chief Executive, Apple, to predict that cities will be built around it.
    http://www.rediff.com/netguide/2001/dec/03ginger.htm
  - Steve Jobs, the chief executive of Apple, shocked shareholders and the tech community last night by stepping down from his role while he fights a " hormone imbalance" that has made him lose weight rapidly.
    http://technology.timesonline.co.uk/tol/news/tech_and_web/article5519684.ece
  - Larry Ellison is CEO of Oracle, an integrated database software company.
    http://www.businessweek.com/mediacenter/coi/content.htm
  - Show Debug Info

Main Tasks: Relation Extraction and Relational Similarity Measurement

- **Relation Extraction from Contextual Lexical Patterns**
  - Tokyo is Japan's capital. →
    (Tokyo, Japan) : X is Y's capital, X is Y's, is Y's capital, ..
  - **Indexing of these Relational Properties of Possible Entity Pairs for efficient search.**
- **Relational Similarity Measurement based on the Distributional Hypothesis:**
  - (Tokyo, Japan) ≈ (Paris, France)

Lexical Pattern Extraction for Indexing in Relational Search

- **In the earlier researches of measuring relational similarity, such as Turney'06, Bollegala et al.'09, the entity pairs are given.**

"Microsoft **** PowerSet" ⇒ Google ⇒

Microsoft's acquisition of PowerSet will ...
Microsoft to buy PowerSet
Microsoft, PowerSet confirm deal
Microsoft to acquire PowerSet for $100M

Microsoft = X
PowerSet = Y

X's acquisition of Y, X to buy Y,
X, Y confirm deal, X to acquire Y for $NNN

- **At the time of Indexing of our relational search system, entity pairs are not given.**
- **Thus, we find possible entity pairs which co-occur in a sentence more than a certain count, and make the index of their properties. (At present, we find only the pairs of nouns from the Wikipedia texts.)**

Clustering Lexical Patterns (2)

[Davidov ACL'07, Bollegala et. al WWW'09]

- **Clustering based on the Distributional Hypothesis**
  - Y's CEO X :
    - (Jobs, Apple) : 50 occurrences
    - (Ballmer, MS) : 10 occurrences
  - X, CEO of Y :
    - (Jobs, Apple) : 20 occurrences
    - (Ballmer, MS) : 30 occurrences

similar!

- **Patterns in the same cluster become the same feature vector component.**
- **This clustering is effective in order to solve the problem of data sparseness in high dimensions.**

Dmitry Davidov et al. Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining, ACL'07
D. Bollegala, Y. Matsuo, M. Ishizuka. Measuring the Similarity between Implicit Semantic Relations from the Web, WWW'09

Entity Clustering

- **United States, U.S., US, U.S, ...indicate the same entity. They should be clustered into an entity.**

- Steve Ballmer, Microsoft
  - (Steve Ballmer, Microsoft): 50 occurrences
  - (Steve Ballmer, Bill Gates) : 10
  - (Steve Ballmer, Microsoft Corp) : 8
  - ...
- Ballmer:
  - (Ballmer, Microsoft) : 20
  - (Ballmer, Bill Gates) : 15
  - (Ballmer, Gates) : 10
  - ...

There is a high similarity between "Steve Ballmer" and "Ballmer", which can be clustered.

Index (Lexical Patterns) Table

Showing rows 30 - 59 (~380,621 total, Query took 0.0009 sec)

```
SELECT *
FROM `rs_lex_pattern_ngrams`
LIMIT 30, 30
```

id	contents	freq	size size of this n-gram (n)
128115	melvyn bragg ^x with ^y	1	4
128116	^x with ^y professor	1	4
128117	by melvyn bragg ^x with ^y	1	5
128118	melvyn bragg ^x with ^y professor	1	5
128119	^x with ^y professor polit	1	5
128120	^x professor polit at ^y	1	5
128121	at ^x ^y	759	3
128122	^x ^y senior	2	3
128123	polit at ^x ^y	2	4
128124	at ^x ^y senior	1	4
128125	^x ^y senior lectur	1	4
128126	professor polit at ^x ^y	1	5

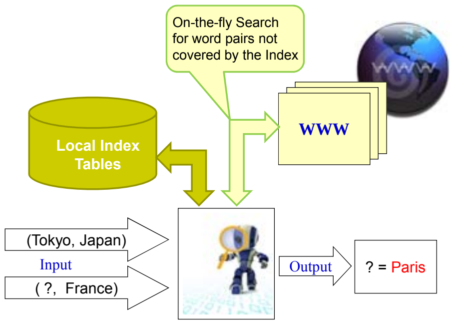
Index (POS Patterns) Table

Showing rows 60 - 89 (~47,102 total, Query took 0.0010 sec)

```
SELECT *
FROM `rs_pos_pattern_ngrams`
LIMIT 60, 30
```

id	contents	freq	size size of this n-gram (n)
28602	^x VBG ^y	422	3
28603	JJR ^x VBG ^y	1	4
28604	^x VBG ^y PERSON	24	4
28605	VB JJR ^x VBG ^y	1	5
28606	JJR ^x VBG ^y PERSON	1	5
28607	^x VBG PERSON ^y	22	4
28608	JJR ^x VBG PERSON ^y	1	5
28609	VBG ^x ^y	679	3
28610	LOCATION VBG ^x ^y	39	4
28611	JJR LOCATION VBG ^x ^y	1	5
28612	^x ^y IN NNP	253	4
28613	^x ^y ORGANIZATION	1777	3

On-the-fly Search for word pairs not covered by the Index



Preliminary Performance Evaluation

- Indexing from a corpus contains 12,000 Web pages
  - Articles mostly on company acquisition, headquarters, CEO and person birthplaces
  - ~100MB of text
  - No. of Entity pairs: ~ 113,000 (occurrences> 4 : ~ 4000)
  - No. of Lexical patterns : ~ 2,000,000
    - 17 lexical patterns for one entity pair on average
- Relational Search (A, B), (C, ?) for the case of entity pairs with occurrence counts more than 4.
- The accuracy of Top10 outputs is about 81% at present.
- Average mean reciprocal rank (MRR) is 0.963
- On the process of improvements and detailed analyses.

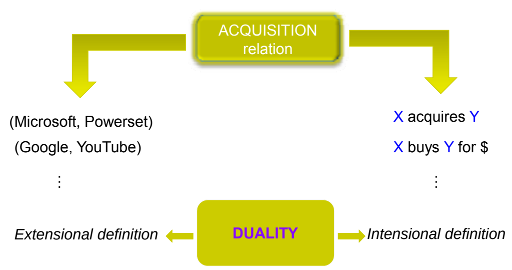
Current Issues

- **An Efficient Implementation of the Remote Corpus (On-the-fly access to the Web)**
- **A way of removing erroneous outputs with no/small relational similarity.**
- **Errors in the Clustering**
- **Entities other than nouns**
  - **Verbs, Adjectives, Adverbs, …….**
- **The Facts with Time.**
  - Eg., Bill Gate is CEO of Microsoft. (in the article before 2000.)
  - Bill Gate was CEO of Microsoft. ← no problem

2. Relational Similarity between Two Word Pairs

(2.1) Computing Relational Similarity
(2.2) Latent Relational Search Engine
(2.3) Open Relation Extraction employing Sequential Co-clustering

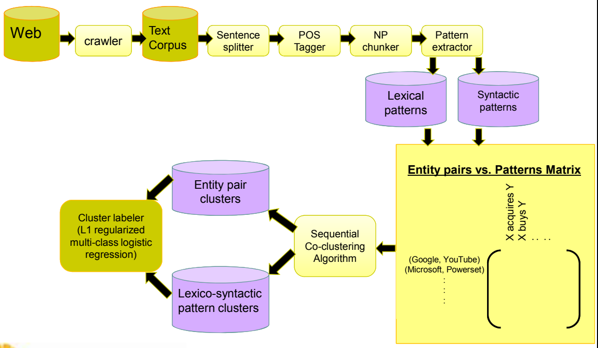
Relational Duality



Open Relation Extraction from the Web

- **Problem definition**
  - *Given a crawled corpus of Web text, identify all the different semantic relations that exist between entities mentioned in the corpus.*
- **Challenges**
  - The number or the types of the relations that exist in the corpus are not known in advance
  - Costly, if not impossible to create training data
  - Entity name variants must be handled
    - Will Smith vs. William Smith vs. fresh prince,…
  - Paraphrases of surface forms must be handled
    - acquired by, purchased by, bought by,…
  - Multiple relations can exist between a single pair of entities

Overview of the proposed method

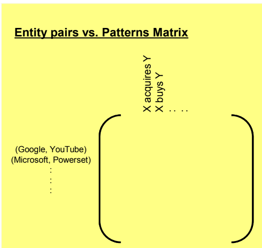


Lexico-Syntactic Pattern Extraction

- Replace the two entities in a sentence by X and Y
- Generate subsequences (over tokens and POS tags)
  - A subsequence must contain both X and Y
  - The maximum length of a subsequence must be L tokens
  - A skip should not exceed g tokens
  - Total number of tokens skipped must not exceed G
  - Negation contractions are expanded and are not skipped
- Example
  - … merger/NN is/VBZ software/NN maker/NN [Adobe/NNP System/NN] acquisition/NN of/IN [Macromedia/NNP]
  - X acquisition of Y, software maker X acquisition of Y
  - X NN IN Y, NN NN X NN IN Y

Entity pairs vs. Lexico-Syntactic Pattern Matrix

- Select the most frequent entity pairs and patterns, and create an entity-pair vs. pattern matrix.



Sequential Co-clustering Algorithm

1. Input: A data matrix, row and column clustering thresholds
2. Sort the rows and columns of the matrix in the descending order of their total frequencies.
3. for rows and columns do:
   - Compute the similarity between current row (column) and the existing row (column) clusters
   - If maximum similarity < row (column) clustering threshold:
     - Create a new row (column) cluster with the current row (column)
   - else:
     - Assign the current row (column) to the cluster with the maximum similarity
   - repeat until all rows and columns are clustered
4. return row and column clusters

Sequential Co-clustering Algorithm

	X acquired Y	Y CEO X	X buys Y for $	X of Y	Y head X	
(Jobs, Apple)	0	5	0	2	1	=8
(Balmer, Microsoft)	0	8	0	3	2	=13
(Microsoft, Powerset)	5	0	1	1	0	=7
(Google, YouTube)	6	0	8	2	0	=16

Row clustering threshold = column clustering threshold = 0.5

Sequential Co-clustering Algorithm

	X acquired Y	Y CEO X	X buys Y for $	X of Y	Y head X
(Google, YouTube)	6	0	8	2	0
(Balmer, Microsoft)	0	8	0	3	2
(Jobs, Apple)	0	5	0	2	1
(Microsoft, Powerset)	5	0	1	1	0
	=11	=13	=9	=8	=3

Sequential Co-clustering Algorithm

	Y CEO X	X acquired Y	X buys Y for $	X of Y	Y head X
(Google, YouTube)	0	6	8	2	0
(Balmer, Microsoft)	8	0	0	3	2
(Jobs, Apple)	5	0	0	2	1
(Microsoft, Powerset)	0	5	1	1	0
	=13	=11	=9	=8	=3

Sequential Co-clustering Algorithm

	Y CEO X	X acquired Y	X buys Y for $	X of Y	Y head X
(Google, YouTube)	0	6	8	2	0
(Balmer, Microsoft)	8	0	0	3	2
(Jobs, Apple)	5	0	0	2	1
(Microsoft, Powerset)	0	5	1	1	0

Sequential Co-clustering Algorithm

	Y CEO X	X acquired Y	X buys Y for $	X of Y	Y head X
[(Google, YouTube)]	0	6	8	2	0
(Balmer, Microsoft)	8	0	0	3	2
(Jobs, Apple)	5	0	0	2	1
(Microsoft, Powerset)	0	5	1	1	0

Sequential Co-clustering Algorithm

	[Y CEO X]	X acquired Y	X buys Y for $	X of Y	Y head X	
[(Google, YouTube)]	0	6	8	2	0	0.067 < 0.5
(Balmer, Microsoft)	8	0	0	3	2	
(Jobs, Apple)	5	0	0	2	1	
(Microsoft, Powerset)	0	5	1	1	0	

Sequential Co-clustering Algorithm

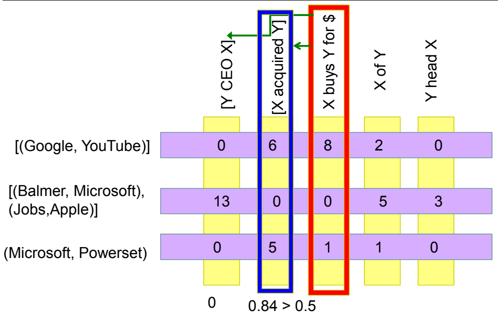
0<0.5

	[Y CEO X]	X acquired Y	X buys Y for $	X of Y	Y head X
[(Google, YouTube)]	0	6	8	2	0
[(Balmer, Microsoft)]	8	0	0	3	2
(Jobs, Apple)	5	0	0	2	1
(Microsoft, Powerset)	0	5	1	1	0

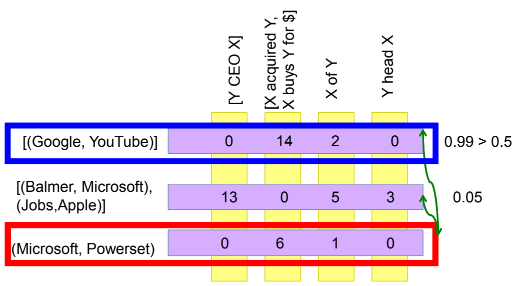
Sequential Co-clustering Algorithm

	[Y CEO X]	[X acquired Y]	X buys Y for $	X of Y	Y head X	
[(Google, YouTube)]	0	6	8	2	0	0.071
[(Balmer, Microsoft)]	8	0	0	3	2	0.998 >0.5
(Jobs, Apple)	5	0	0	2	1	
(Microsoft, Powerset)	0	5	1	1	0	

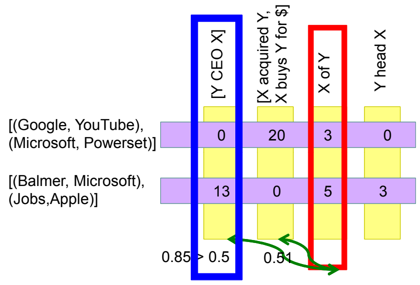
Sequential Co-clustering Algorithm



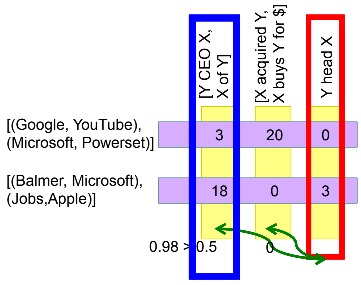
Sequential Co-clustering Algorithm



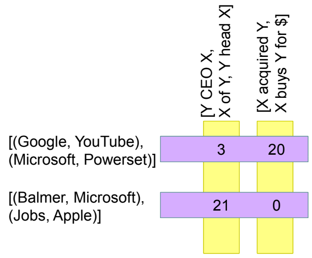
Sequential Co-clustering Algorithm



Sequential Co-clustering Algorithm



Sequential Co-clustering Algorithm



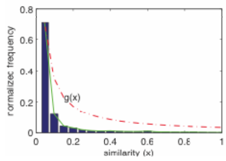
Lexical-syntactic pattern clusters

Entity pair clusters

- ✓ A greedy clustering algorithm
- ✓ Alternates between rows and columns
- ✓ Complexity $O(n \log n)$
- ✓ Common relations are clustered first
- ✓ The no. of clusters is not required
- ✓ Two thresholds to determine

Estimating the Clustering Thresholds

- Ideally each cluster must represent a unique semantic relation
- Number of clusters = Number of semantic relations
- Number of semantic relations is unknown
- Thresholds can be either estimated via cross-validation (requires training data) OR approximated using the similarity distribution.



Similarity distribution is approximated using a Zeta distribution (Zipf's law)
Ideal clustering:
inter-cluster similarity = 0
→ intra-cluster similarity = mean
with a large number of data points:
average similarity in a cluster ≥ threshold
→ threshold ≈ distribution mean

$$\hat{\theta} = E_g(x) = \int_\delta^1 x g(x)\,dx = \frac{a(1-\delta^{2-k})}{2-k}$$

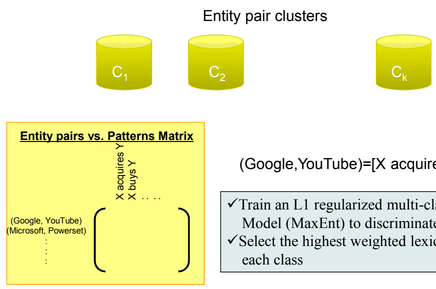
Measuring Relational Similarity

- Empirically evaluate the clusters produced
  - Use the clusters to measure relational similarity (Bollegala, WWW 2009)
  - Distance = $(\mathbf{f}_{(a,b)} - \mathbf{f}_{(c,d)})^{\mathbf{T}}\mathbf{T}^{-1}(\mathbf{f}_{(a,b)} - \mathbf{f}_{(c,d)})$
  - ENT dataset: 5 relation types, 100 instances
  - Task: query using each entity pair and rank using relational distance

Relation	VSM	LRA	EUC	RELSIM	Proposed
ACQUSITION	0.92	0.92	0.91	0.94	0.89
HEADQUARTERS	0.84	0.82	0.79	0.86	0.97
FIELD	0.44	0.43	0.51	0.57	0.42
CEO	0.95	0.96	0.90	0.95	0.99
BIRTHPLACE	0.27	0.27	0.33	0.36	0.53
Overall Average Precision	0.68	0.68	0.69	0.74	0.76

Self-supervised Relation Detection

- What is the relation represented by a cluster?
  - Label each cluster with a lexical pattern selected from that cluster.



(Google,YouTube)=[X acquired Y:10,…]

✓Train an L1 regularized multi-class logistic regression Model (MaxEnt) to discriminate the k-classes.
✓Select the highest weighted lexical patterns from each class

Subjective Evaluation of Relation Labels

- Baseline
  - Select the most frequent lexical pattern in a cluster as its label
- Ask three human judges to assign grades
  - A: baseline is better
  - B: proposed method is better
  - C: both equally good
  - D: both bad

Relation	A	B	C	D
ACQUSITION	16.7%	40%	40%	3.3%
HEADQAURTERS	20%	40%	23.3%	16.7%
CEO	6.7%	53.3%	20%	20%
FIELD	13.3%	56.7%	23.3%	6.7%
BIRTHPLACE	13.3%	36.7%	10%	40%
Overall	14%	45.3%	23.3%	17.3%

Open Information Extraction

- SENT500 dataset (Banko and Etzioni, ACL 2008)
- 500 sentences, 4 relation types
- Lexical patterns 947, Syntactic patterns 384
- 4 row clusters, 14 column clusters

Method	Precision	Recall	F
O-NB	0.866	0.232	0.366
O-CRF	0.883	0.452	0.598
MLN	0.798	0.733	0.764
PROP (lexical)	0.943	0.647	0.767
PROP (syntactic)	0.752	0.860	0.802
PROP (lexical + syntactic)	0.751	0.857	0.801

Classifying Relations in a Social Network

spysee.jp

Relation Classification

- Dataset
  - 790,042 nodes (people), 61,339,833 edges (relations)
  - Randomly select 50,000 edges and manually classify into 53 classes
  - 11,193 lexical patterns, 383 pattern clusters, 664 entity pair clusters

Relation	P	R	F	Relation	P	R	F
colleagues	0.76	0.87	0.81	friends	0.58	0.77	0.66
alumni	0.83	0.68	0.75	co-actors	0.75	0.74	0.74
fan	0.91	0.50	0.64	teacher	0.83	0.73	0.78
husband	0.89	0.57	0.74	wife	0.67	0.34	0.45
brother	0.79	0.60	0.68	sister	0.90	0.52	0.66
Micro	0.72	0.68	0.70	Macro	0.78	0.52	0.63

Summary of Open Relation Extraction employing Sequential Co-clustering

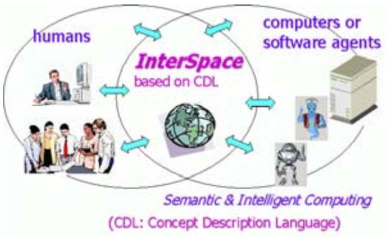
- Dual representation of semantic relations leads to a natural co-clustering algorithm.
- Clustering both entity pairs and lexico-syntactic patterns simultaneously helps to overcome data sparseness in both dimensions.
- Co-clustering algorithm scales $n\log(n)$ with data
- Clusters produced can be used to:
  - Measure relational similarity with performance comparable to supervised approaches
  - Open Information Extraction Tasks
  - Classify relations found in a social network.

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

THE UNIVERSITY OF TOKYO

In cooperation with ISeC (Institute of Semantic Computing)

We need a Common and Universal Language of Representing Concept Meaning toward Semantic Computing on the Web



The aims of CDL are
1) to realize machine understandability of Web text contents, and
2) to overcome language barrier on the Web.

Major Differences from Semantic Web

Semantic Web

- Target of representation: Meta-data extracted from Web contents.
- Domain-dependent ontologies (which cause the difficulty of wide inter-boundary usage)
- RDF / OWL (description logic is hard for ordinary people to understand)

Tim Berners-Lee says that:
"Data Web" or "Linked Data" is more adequate rather than "the Semantic Web". (2007)

Semantic Computing Initiative

- Target of representation: Semantic concepts expressed in texts.
- Universal vocabulary (+ additional specific vocabulary in a domain if necessary), and pre-defined relation set.
- CDL.nl (richer than RDF)

Main body:
Institute of Semantic Computing (ISeC) in Japan
Int'l Standardization Activity:
W3C Common Web Language(CWL)-XG

Incubator Group Activity at W3C from Oct. 2006 to May 2008

Common Web Language Incubator Group Charter

End date	15 October 2007
Confidentiality	Proceedings are public
Initial Chairs	Hiroshi Uchida (ISeC)
Initiating Members	Institute of Semantic Computing (ISeC); National Institute of Advanced Industrial Science and Technology (AIST); Justsystem Corporation

Scope

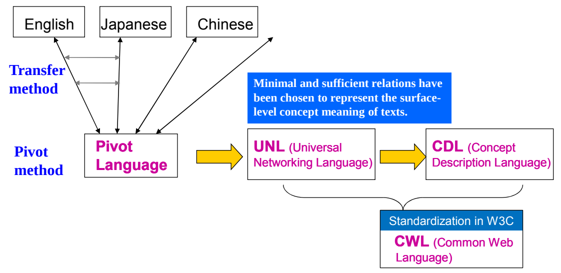
2nd Incubator Group at W3C from June 2008

Common Web Language Evaluation and Installation Incubator Group Charter

End date	20 May 2009
Confidentiality	Proceedings are public
Initial Chairs	Hiroshi Uchida (ISeC)
Initiating Members	Institute of Semantic Computing (ISeC); National Institute of Advanced Industrial Science and Technology (AIST); Keio University; JustSystems Corporation
Usual Meeting Schedule	Teleconferences: No teleconferences Face-to-face: 2 per year

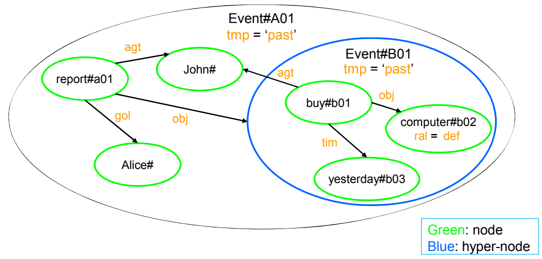
Scope

From Machine Translation



CDL Representation

- Text example:
  "John reported to Alice that he bought a computer yesterday."
- CDL graph notation:



CDL Representation

- Text example:
  "John reported to Alice that he bought a computer yesterday."
- CDL text notation:

```
{#A01 Event tmp='past';
        {#B01 Event tmp='past';
                <#b01:buy;>
                <#b02:computer ral='def';>
                <#b03:yesterday;>
                [#b01 agt #John]
                [#b01 obj #b02]
                [#b01 tim #b03]
        }
        <#John:John;>
        <#Alice:Alice;>
        <#a01:report;>
        [#a01 agt #John]
        [#a01 gol #Alice]
        [#a01 obj #B01]
}
```

Orange: entity
Blue: relation

CDL (UNL) Relations – 44 labels

Semantic Roles: Intra-Event		Logical: Inter-Entity	Restrictive: Restrictive
[Agent Relations]	[Instrument Relations]	[Logical Relations]	cnt (content, namely)
agt (agent)	ins (instrument)	and (conjunction)	fmt (range, from-to)
cag (co-agent)	met (method, means)	orr (disjunction, alternative)	fmr (origin)
aoj (thing w/ attribute)	[State Relations]	[Concept Relations]	mod (modification)
cao (co-thing w/ attribute)	src (source, initial state)	equ (equivalent)	nam (name)
ptn (partner)	gol (goal, final state)	icl (included)	per (proportion, rate)
[Object Relations]	via (interm. place or state)	iof (an instance of)	pof (part of)
obj (affected thing)	[Time Relations]	**Intra- and Inter-Event** (Discourse)	pos (possessor)
cob (affected co-thing)	tim (time)	[Cause Relations]	qua (quantity)
opl (affected place)	tmf (initial time)	con (condition)	tto (destination)
ben (beneficiary)	tmt (final time)	pur (purpose, objective)	
[Place Relations]	dur (duration)	rsn (reason)	
plc (place)	[Manner Relations]	[Sequence Relations]	
plf (initial place)	man (manner)	coo (co-occurence)	
plt (final place)	bas (basis for a standard)	seq (sequence)	
scn (scene)			

Semantic Role Labels in PropBank

The focus is on Predicate-Argument Structure.

- **Arg0** (prototypical agent)
- **Arg1** (prototypical patient)
- **Arg2** (indirect object/benefactive/instrument/attribute/end state)
- **Arg3** (start point/benefactive/instrument/attribute)
- **Arg4** (end point)
- **Arg5** (   )
- **TMP** (time)
- **LOC** (location)
- **DIR** (direction)
- **MNR** (manner)
- **PRP** (purpose)
- **CAU** (cause)
- **MOD** (modal verb)
- **NEG** (negative marker)
- **ADV** (general-purpose modifier)
- **DIS** (discourse particle and clause)
- **PRD** (secondary predication)

These are defined wrt each word sense.

Ex) **buy**::
Arg0: buyer
Arg1: thing bought
Arg2: seller (bought-from)
Arg3: price paid
Arg4: benefactive (bought-for)

This set is not sufficient for representing every concept expressed in natural language texts. It cannot be used for every language due to its language (English) dependency.

Rich Attributes in UNL and CDL

- **Express subjectivity evaluation of the writer/speaker for the sentence.**
- **Ex.) tense, aspect, mood, etc.**

- Time with respect to writer
  @past @present @future
- Writer's view on aspect of event
  @begin @complete @continue @custom @end @experience @progress @repeat @state
- Writer's view of reference
  @generic @def @indef @not @ordinal
- Writer's view of emphasis, focus and topic
  @emphasis @entry @qfocus @theme @title @topic
- Writer's attitudes
  @affirmative @confirmation @exclamation @imperative @interrogative @invitation @politeness @respect @vocative
- Writer's view of reference
  @generic @def @indef @not @ordinal
- Writer's feeling and judgements
  @ability @get-benefit @give-benefit @conclusion @consequence @sufficient @grant @grant-not @although @discontented @expectation @wish @insistence @intention @want @will @need @obligation @obligation-not @should @unavoidable @certain @inevitable @may @possible @probable @rare @regret @unreal @admire @blame @contempt @regret @surprised @troublesome
- Describing logical characters and properties of concepts
  @transitive @symmetric @identifiable @disjoint
- Modifying attribute on aspect
  @just @soon @yet @not
- Attribute for convention
  @passive @pl @angle_bracket @brace @double_parenthesis @double_quote @parenthesis @single_quote @square_bracket

The defining method of one unique sense of a word in *UW* (Patent of UN Univ.)

- **Defining category**

swallow(icl>bird) — the bird
"One swallow does not make a summer"
swallow(icl>action) — the action of swallowing
"at one swallow"
swallow(icl>quantity) — the quantity
"take a swallow of water"

- **Defining possible case relations**

spring(agt>thing,obj>wood) — bending or dividing something
spring(agt>thing,obj>mine)) — blasting something
spring(agt>thing,obj>person, src>prison)) — escaping (from) prison
spring(agt>thing,gol>place) — jumping up
"to spring up"
spring(agt>thing,gol>thing) — jumping on
"to spring on"
spring(obj>liquid) — gushing out
"to spring out"

UW (*Universal Words*) in UNL

Universal Word
uw{(equ>Universal Word)}
adjective concept{(icl>uw)}
uw(aoj>thing{,and>uw,ben>thing,cao>thing,cnt>uw,cob>thing,con>uw,coo>uw,dur>period,man>how,obj>thing,or>uw(aoj>thing),plc>thing,plf>thing,plt>thing,rsn>uw(aoj>thing),rsn>do,icl>adjective concept})
Achaean({icl>uw(}aoj>thing{)})
Afghan({icl>uw(}aoj>thing{)})
African({icl>uw(}aoj>thing{)})
African-American({icl>uw(}aoj>thing{)})
Ainu({icl>uw(}aoj>thing{)})
Alaskan({icl>uw(}aoj>thing{)})
Albanian({icl>uw(}aoj>thing{)})
Aleutian({icl>uw(}aoj>thing{)})
Alexandrian({icl>uw(}aoj>thing{)})
Algerian({icl>uw(}aoj>thing{)})
Altaic({icl>uw(}aoj>thing{)})
American({icl>uw(}aoj>thing{)})
Anglian({icl>uw(}aoj>thing{)})
Anglo-American({icl>uw(}aoj>thing{)})
Anglo-Catholic({icl>uw(}aoj>thing{)})
Anglo-French({icl>uw(}aoj>thing{)})
Anglo-Indian({icl>uw(}aoj>thing{)})
Anglo-Irish({icl>uw(}aoj>thing{)})
Anglo-Norman({icl>uw(}aoj>thing{)})
Arab({icl>uw(}aoj>thing{)})
Arab-Israeli({icl>uw(}aoj>thing{)})
Arabian({icl>uw(}aoj>thing{)})
Arabic({icl>uw(}aoj>thing{)})

40,000 lexicons are open to public.

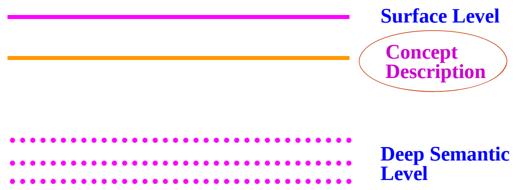
The full vocabulary includes 200,000 lexicons as of 2007.

Discourse (Inter-sentence) Relations are missing in current CDL.nl

Discourse Relations at ISO/TC37/SC4/TDG3 (34 types)

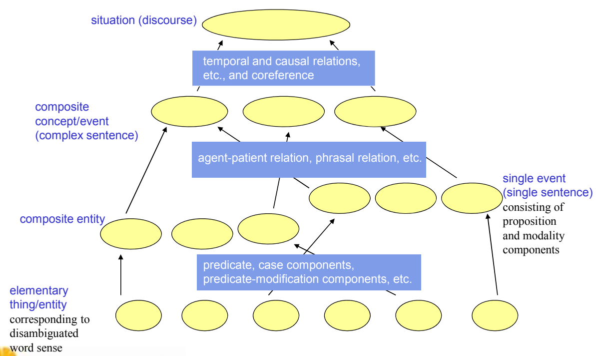
- derivation
- causes
- conditional
- inference
- purpose
- trigger
- compromise
- conflict
- contrast
- unconditional
- comparison
- disjunction
- dissimilar
- manner
- otherwise
- proportion
- similar
- strongComparison
- detail
- element
- example
- extraction
- general-specific
- minimum
- part
- process-step
- restatement
- constraint
- supplement
- background
- content
- evaluation

Concept Description Levels



- **There are several choices for the deep semantic-level description depending on applications. On the other hand, a certain consensus has been made wrt "Concept Description" which is slightly below the surface level, through decades-long researches on NLP, machine translation and electric dictionaries.**
- **Whereas a complete consensus has not been achieved yet regarding the Concept Description level and its description scheme, it is meaningful to set up a common concept description format as an international standard today.**

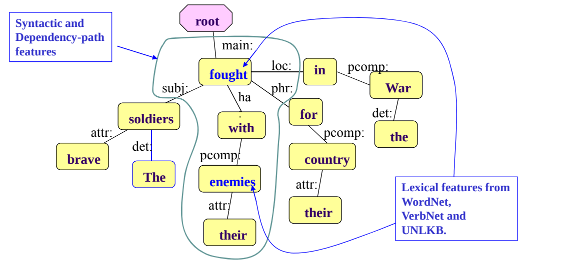
Hierarchical Construction of Concept Representation in CDL



Approaches for Generating CDL Data

- Manual Coding & Editing
  - Even in this case, a graphical input editor is necessary.
- Graphical Input & Editing (**Hasida's Semantic Authoring**)
- Some Manual Tagging to Text, then Conversion into CDL.
- Semi-automatic Conversion from Text (1) ⟸ **Our current approach**
  - Automatic and Manual Word Sense Disambiguation, then Conversion into CDL.
- Semi-automatic Conversion from Text (2)
  - Post editing of converted CDL data with a GUI.
- Full Automatic Conversion (**ultimate goal**)

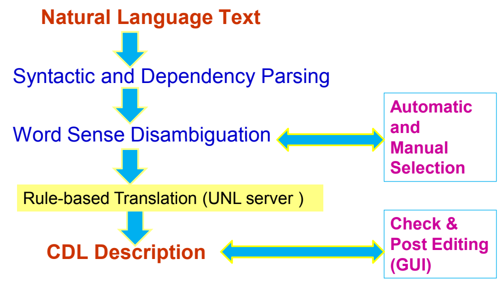
Recognition of CDL Relations from dependency-analyzed text



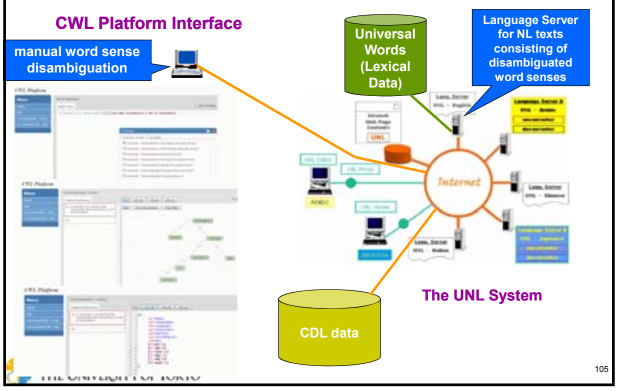
Some labels of Connexor Machinese Analyser:
ha (prepositional phase attachment), **phr** (verb particle),
pcomp (subject complement)

Performance for frequent 36 relations (out of 44)
Precision 87.3% Recall 88.1% F-value 87.1%

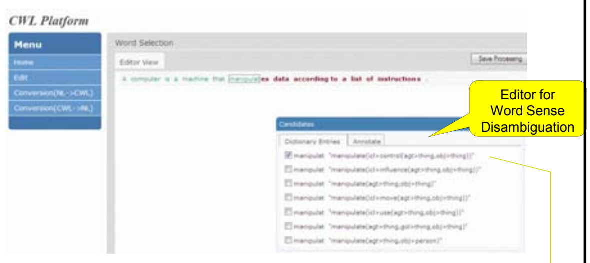
A Semi-automatic Conversion from NL Text to CDL



Semi-automatic Conversion from NL Texts to CDL

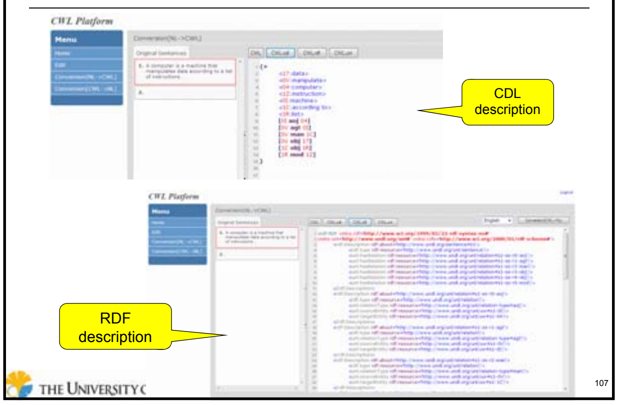


CWL Platform Interface (1)

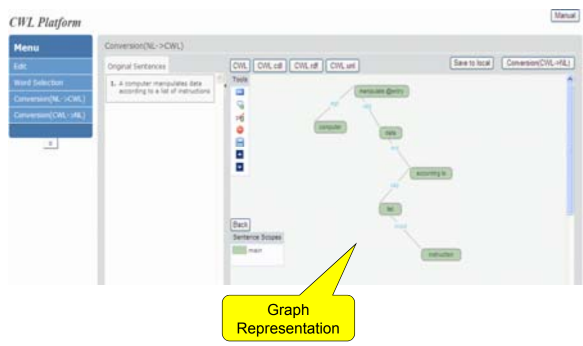


manipulat "manipulate(icl>control(agt>thing, obj>thing))"

CWL Platform Interface Screenshots (2)

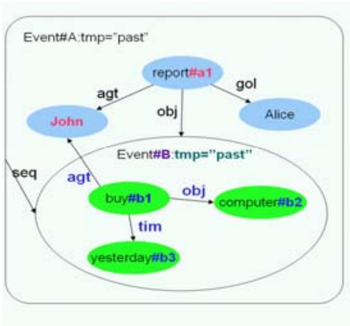


CWL Platform Interface (3)



CDL Data Retrieval via CDQL

(an Extended SPARQL)



Query::
What did John report?

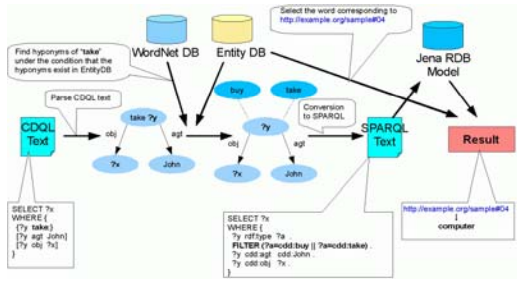
```
SELECT *y *z
WHERE {
 [report agt John]
 [report obj ?x]
 {?x Event: *y, *z}
}
```

result

```
*y = {#b1 buy;}
     {#b2 computer;}
     {#b3 yesterday;}
*z = [#b1 agt John]
     [#b1 obj #b2]
     [#b1 tim #b3]
```


Semantic Retrieval of CDL data

- **CDQL: SQL-like query language for CDL data**



Summary of the Talk

Exploiting Macro and Micro Relations Toward Web Intelligence

1. Social Relation Extraction

2. Relational Similarity between Two Word Pairs
   - (2.1) Computing Relational Similarity
   - (2.2) Latent Relational Search Engine
   - (2.3) Open Relation Extraction employing Sequential Co-clustering

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

Thank You

Mitsuru Ishizuka

School of Information Science and Technology

THE UNIVERSITY OF TOKYO