Relation Extraction from the Web
ウェブからエンティティ間の意味的関係抽出

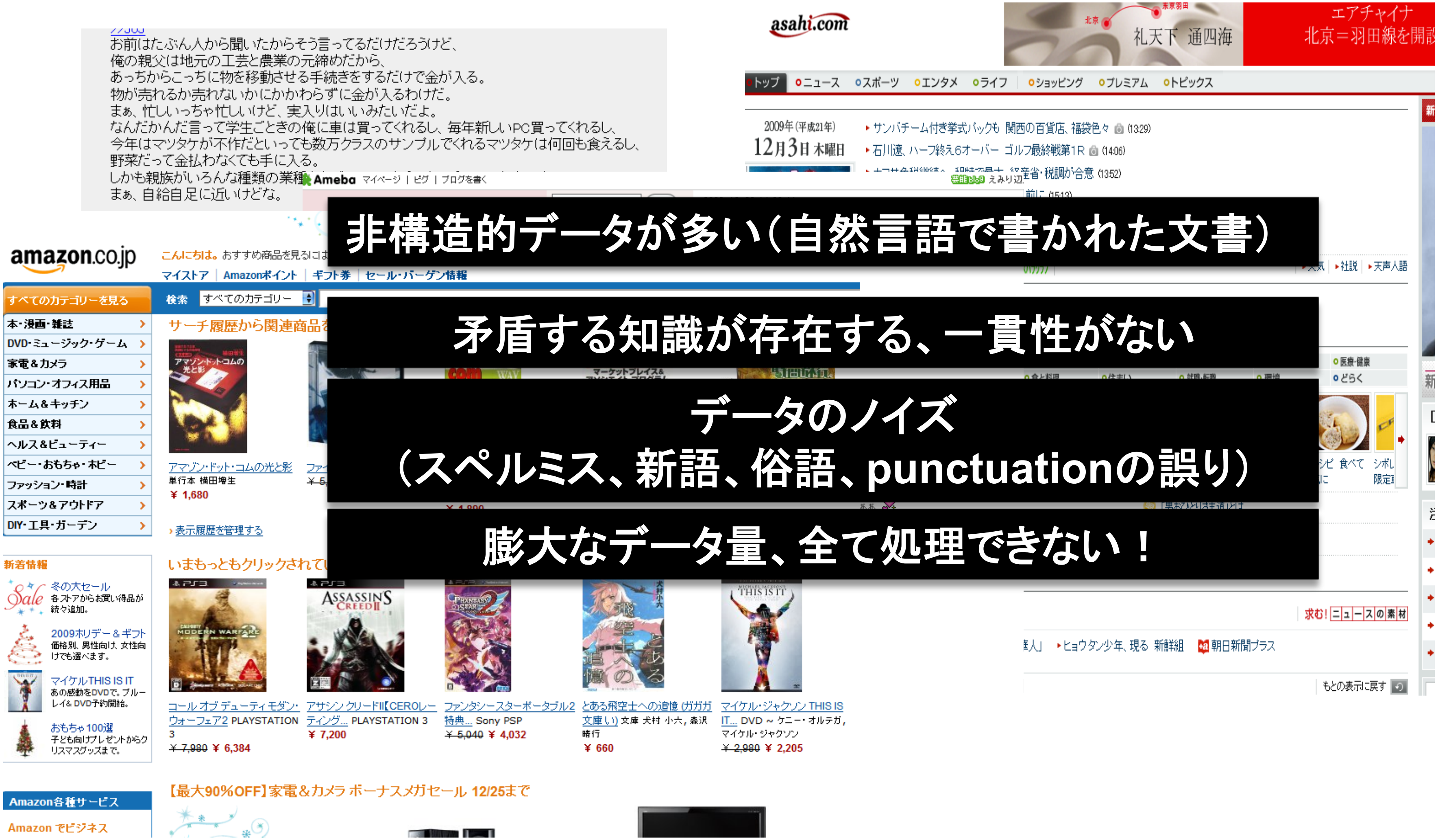
ダヌシカ　ボレガラ
東京大学

THE UNIVERSITY OF TOKYO

自己紹介

- 氏名：ダヌシカ　ボレガラ (Danushka Bollegala)
- 国籍：スリランカ
- 研究分野：自然言語処理、Web Intelligence
- WWW, ACL, EMNLP, IJCAI, AAAIを中心に論文発表
- 経歴：
  - 2000年：文部科学省国費留学生として来日
  - 2005年：東京大学工学部電子情報工学科卒業
  - 2007年：東京大学大学院情報理工学系研究科電子情報学修士課程修了
  - 2009年：東京大学大学院情報理工学系研究科電子情報学博士課程修了
  - 2010~現在：東京大学大学院情報理工学系研究科助教

Webから関係抽出の課題



非構造的データが多い（自然言語で書かれた文書）

矛盾する知識が存在する、一貫性がない

データのノイズ
（スペルミス、新語、俗語、punctuationの誤り）

膨大なデータ量、全て処理できない！

お前はたぶん人から聞いたからそう言ってるだけだろうけど、
俺の親父は地元の工芸と農業の元締めだから、
あっちからこっちに物を移動させる手続きをするだけで金が入る。
物が売れるか売れないかにかかわらずに金が入るわけだ。
まぁ、忙しいっちゃ忙しいけど、実入りはいいみたいだよ。
なんだかんだ言って学生ごときの俺に車は買ってくれるし、毎年新しいPC買ってくれるし、
今年はマツタケが不作だといっても数万クラスのサンプルでくれるマツタケは何回も食えるし、
野菜だって金払わなくても手に入る。
しかも親族がいろんな種類の業種
まぁ、自給自足に近いけどな。

2009年（平成21年）
12月3日 木曜日

- サンバチーム付き挙式パックも 関西の百貨店、福袋色々 (13:29)
- 石川遼、ハーフ終え6オーバー ゴルフ最終戦第1R (14:06)
- 省・税調が合意 (13:52)

こんにちは。おすすめ商品を見るには

サーチ履歴から関連商品を

アマゾン・ドット・コムの光と影
単行本 横田増生
¥ 1,680

表示履歴を管理する

新着情報

冬の大セール
各ストアからお買い得品が続々追加。

2009ホリデー＆ギフト
価格別、男性向け、女性向けでも選べます。

マイケル THIS IS IT
あの感動をDVDで。ブルーレイ＆DVD予約開始。

おもちゃ100選
子ども向けプレゼントからリスマスグッズまで。

いまもっともクリックされて

コール オブ デューティ モダン・ウォーフェア2 PLAYSTATION 3
~~¥ 7,980~~ ¥ 6,384

アサシン クリードII【CEROレーティング... PLAYSTATION 3
¥ 7,200

ファンタシースターポータブル2 特典... Sony PSP
~~¥ 5,040~~ ¥ 4,032

とある飛空士への追憶 (ガガガ文庫 い) 文庫 犬村 小六, 森沢 晴行
¥ 660

マイケル・ジャクソン THIS IS IT... DVD ～ ケニー・オルテガ, マイケル・ジャクソン
~~¥ 2,980~~ ¥ 2,205

【最大90%OFF】家電＆カメラ ボーナスメガセール 12/25まで

求む！ニュースの素材

ヒョウタン少年、現る 新鮮組

Webから関係抽出の課題

Jim Clark

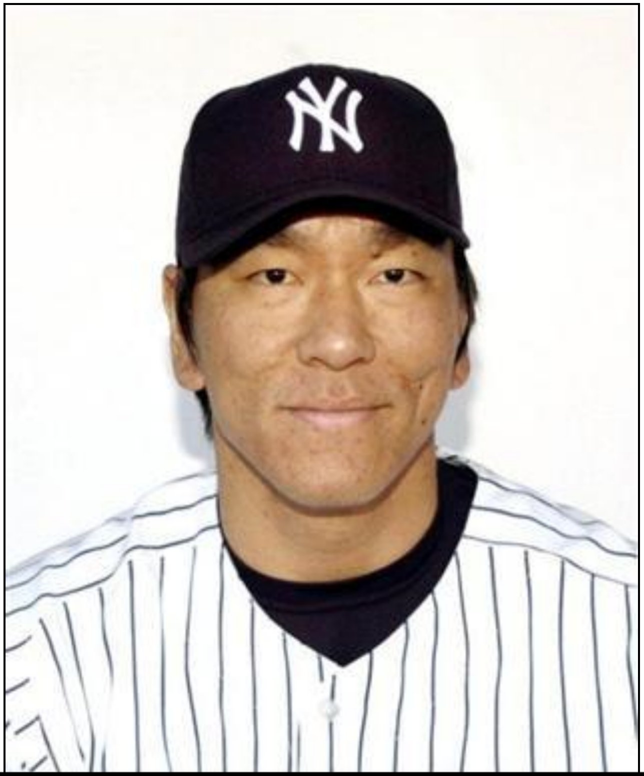
複数のentityが同一の名称で参照される（同姓同名問題）

D. Bollegala, Y. Matsuo, M. Ishizuka,
Disambiguating Personal Names on the Web using Automatically Extracted Keyphrases, ECAI 2006

Netscape創業者

F1チャンピオン

Webから関係抽出の課題



松井秀喜

松井秀

Godzilla

ゴジラ

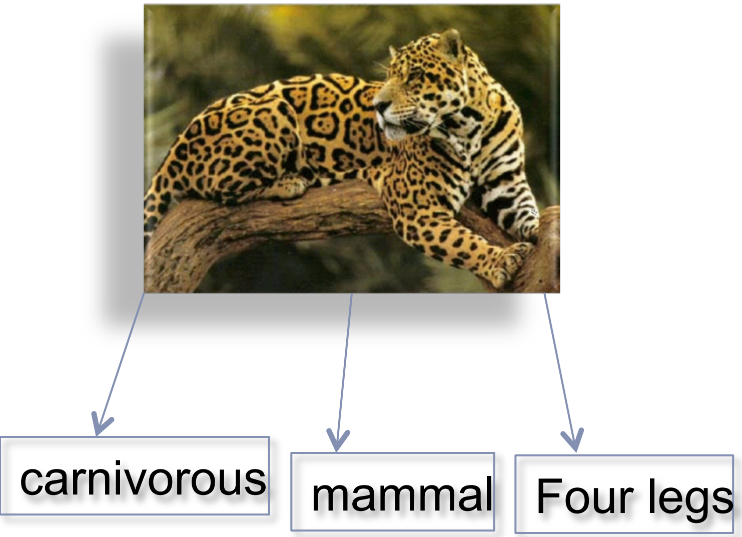
Hideki Matsui

同一のentityが複数の名称で参照される（別名問題）
D. Bollegala, Y. Matsuo, M. Ishizuka,
Automatic Discovery of Personal Name Aliases from the Web, IEEE TKDE 2010.

属性類似性と関係類似性

２つのエンティティが持つ属性が似ていればそれらのエンティティ間には高い属性類似性があると言える。

Jaguar

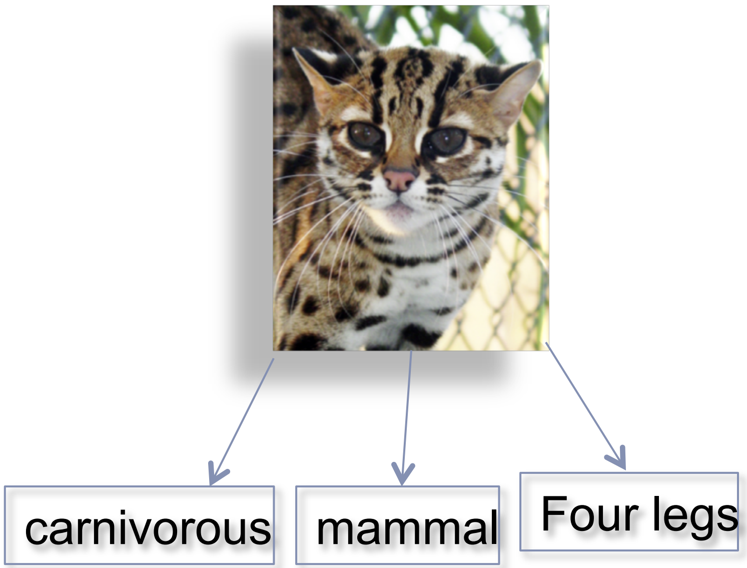


carnivorous

mammal

Four legs

cat



carnivorous

mammal

Four legs

属性類似性関数は２変数関数となる：　sim(X,Y)

属性類似性と関係類似性

２つのエンティティ間の関係が別の２つのエンティティ間の関係に似ていれば、
それのエンティティ対間では高い関係類似性があると言える。

(ostrich, bird)

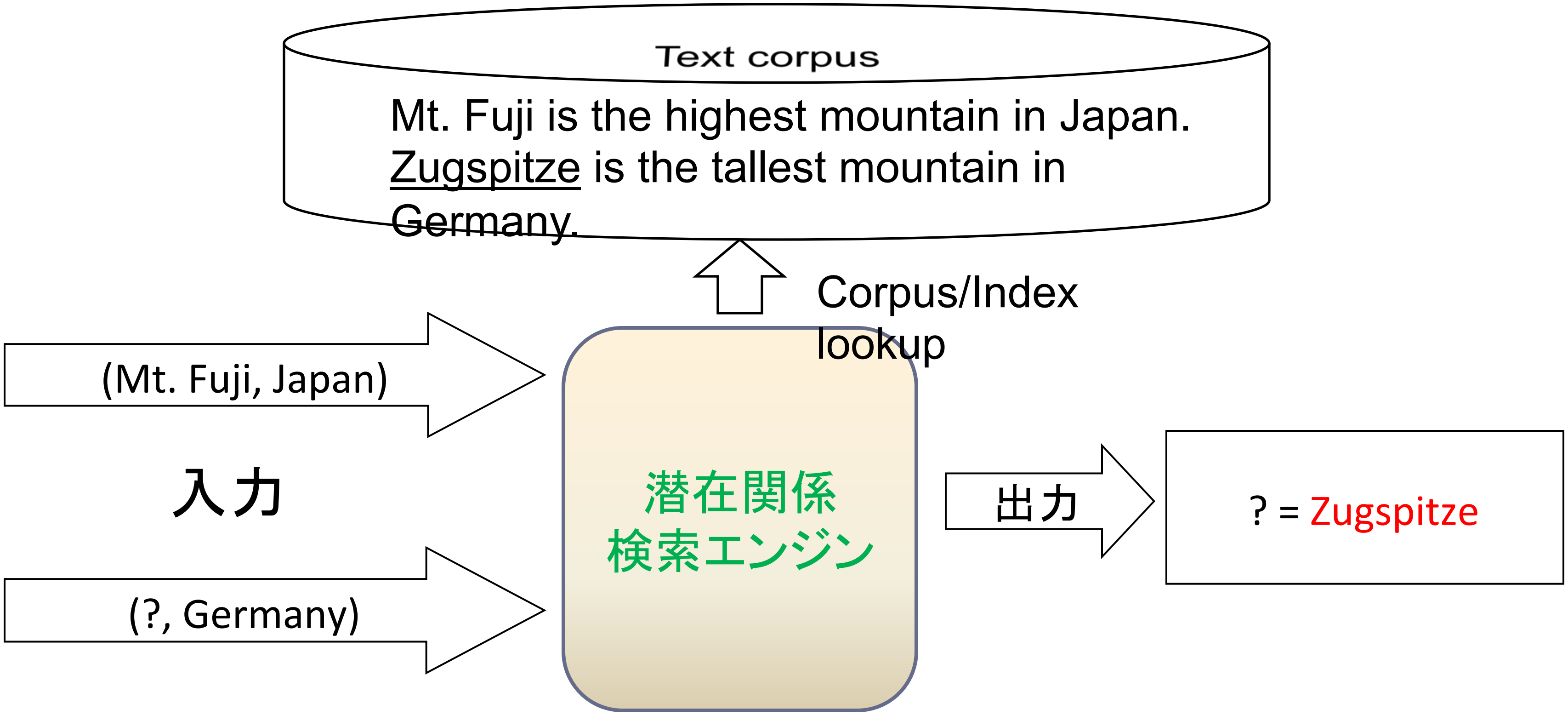
Ostrich is a large bird

(lion, cat)

Lion is a large cat

関係類似性関数は４変数の関数となる：sim(A,B,X,Y)

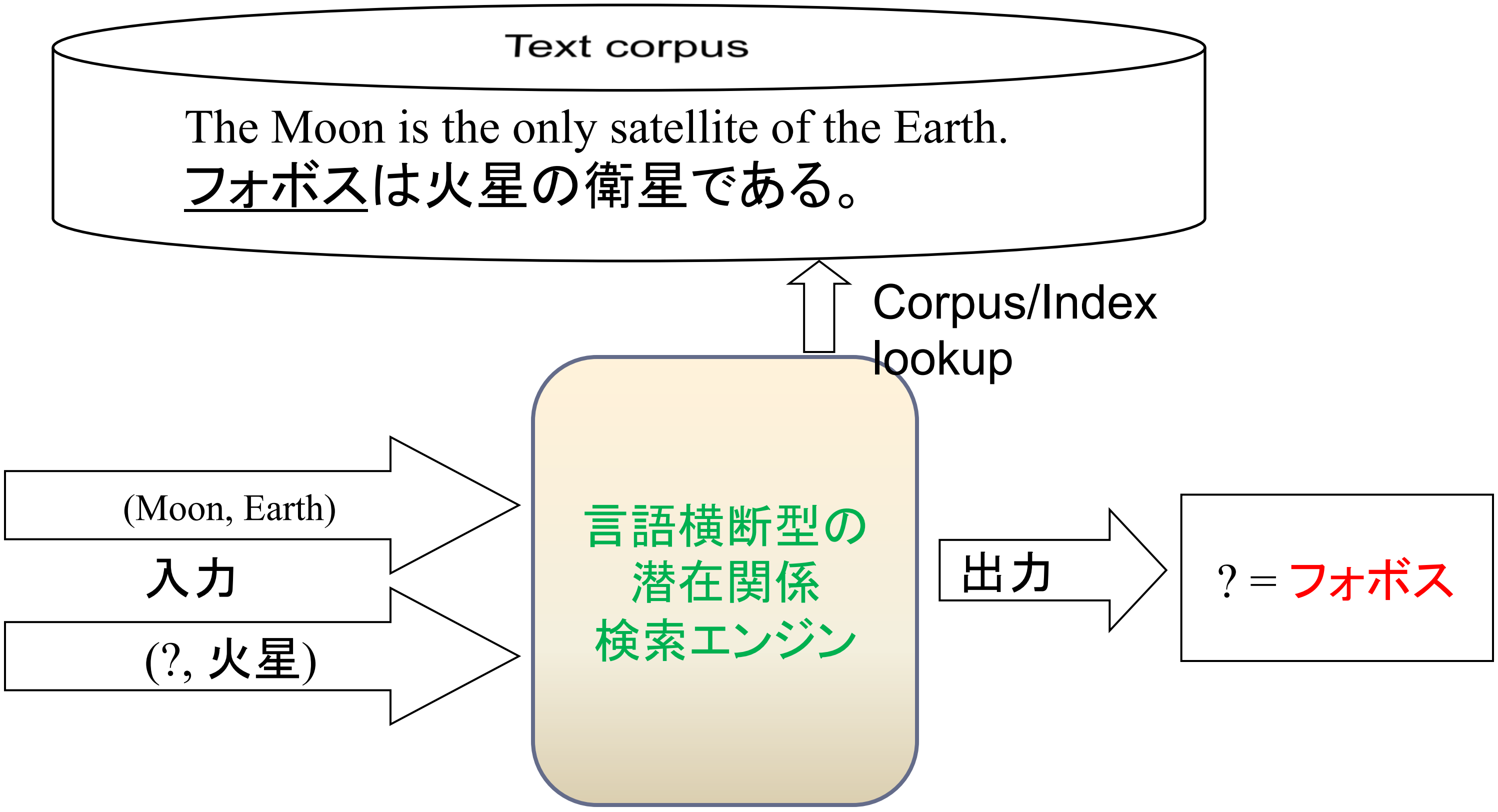
Entity oriented search の一つの形式: 潜在関係検索 (Latent Relational Search)



エンティティ (地名、人名など) 間の関係を利用するエンティティ検索手法

- M.P. Kato et al., Query by Analogical Example: Relational Search using Web Search Engine Indices. CIKM2009
- D. Bollegala et al. , Measuring the Similarity between Implicit Semantic Relations from the Web, WWW2009
- T. Veale, The Analogical Thesaurus, IAAI 2003.

言語の壁への対応: 言語横断型の潜在関係検索


Text corpus

The Moon is the only satellite of the Earth.

フォボスは火星の衛星である。

Corpus/Index lookup

(Moon, Earth)

入力

(?, 火星)

言語横断型の潜在関係検索エンジン

出力

? = フォボス

異なる言語のテキストから結果を検索:

Web空間における言語の壁を越える cross-lingual latent relational search

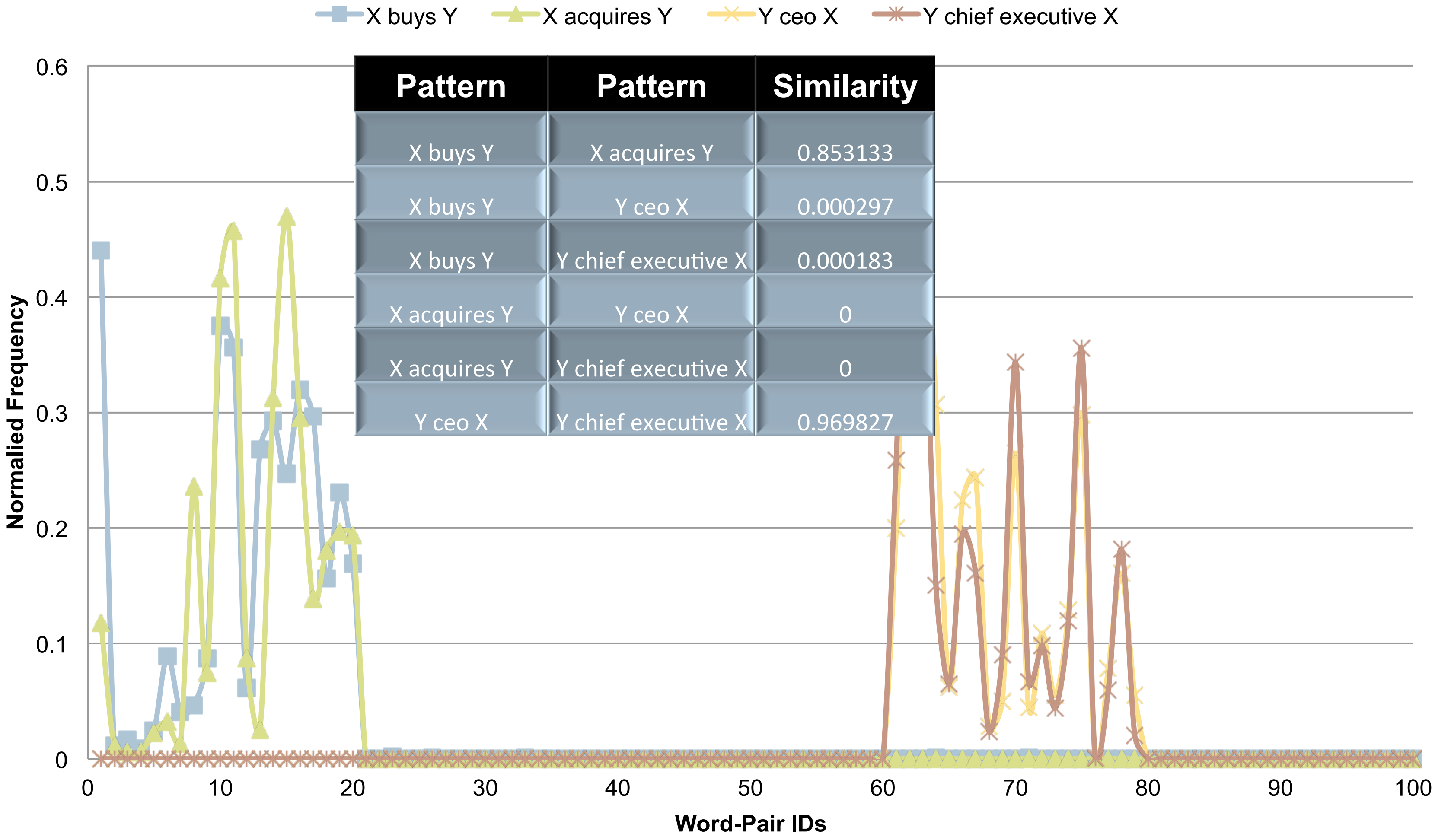
関係類似性計測のチャレンジ

How to explicitly state the relation between two entities?	• Extract lexical patterns from contexts where the two entities co-occur
How to extract the multiple relations between two entities?	
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)

Pattern Extraction

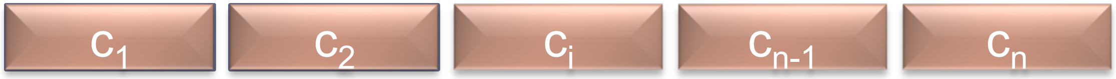
- We use prefix-span, a sequential pattern mining algorithm, to extract patterns that describe various relations, from text snippets returned by a web search engine.
- query = lion * * * * * * * cat
- snippet = .. lion, a large heavy-built social cat of open rocky areas in Africa ..
- patterns = X, *a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

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

Sequential Pattern Clustering Algorithm



C_1 C_2 C_i C_{n-1} C_n

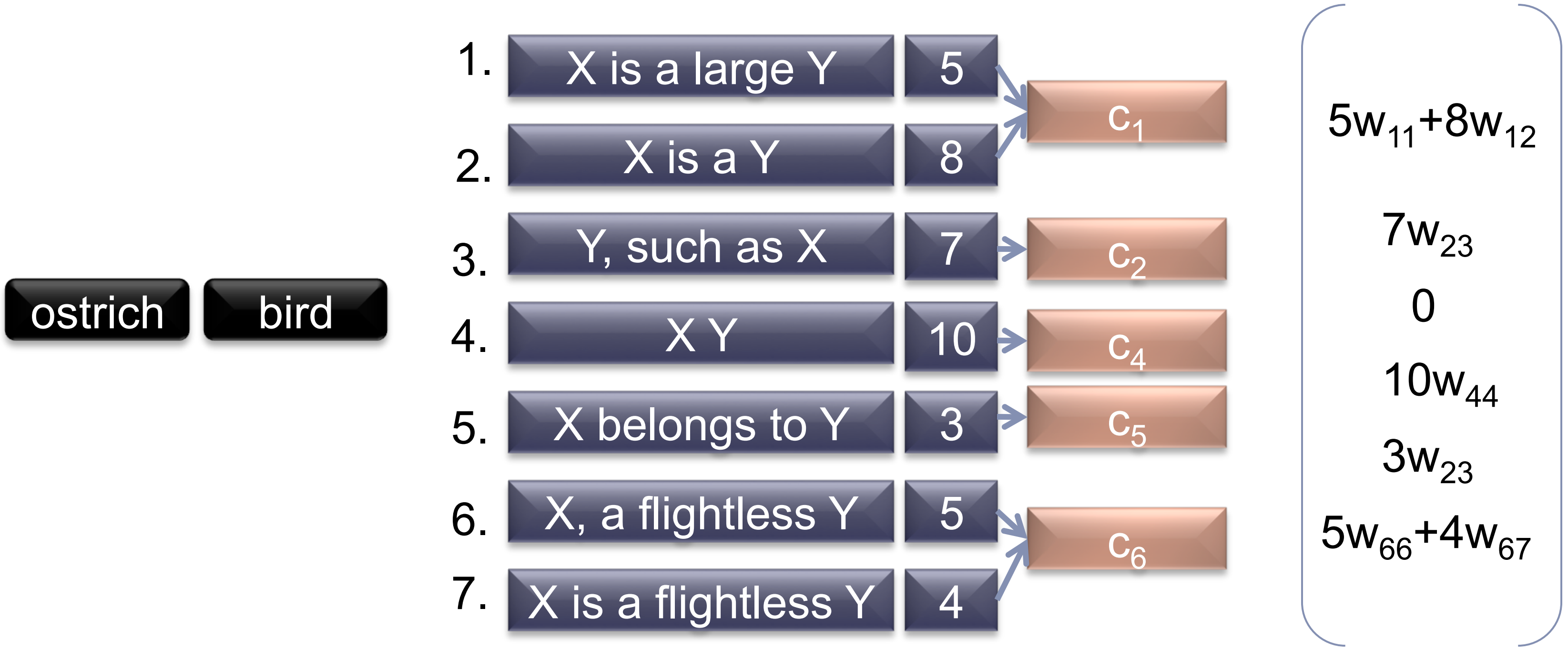
$\mathrm{sim}(p_1,p_2)$⧖Θ

INPUT:
A sorted list of pattern-frequency tuples
$[(p_1,f_1),\ldots(p_N,f_N)]$
$f_1>\ldots>f_N$
Clustering Threshold Θ

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 require pair-wise comparisons, which are computationally costly
- A greedy clustering algorithm

Feature Vector Generation



ostrich bird

1. X is a large Y 5
2. X is a Y 8
3. Y, such as X 7
4. X Y 10
5. X belongs to Y 3
6. X, a flightless Y 5
7. X is a flightless Y 4

c_1 c_2 c_4 c_5 c_6

$$\begin{pmatrix} 5w_{11}+8w_{12} \\ 7w_{23} \\ 0 \\ 10w_{44} \\ 3w_{23} \\ 5w_{66}+4w_{67} \end{pmatrix}$$

$$w_{ij} = \frac{\text{(Frequency of pattern } i \text{ in all word pairs)}}{\sum_{t \in \text{cluster j}} \text{(Frequency of pattern } t \text{ in all word pairs)}}$$

Computing Relational Similarity

- We represent each word pair by an *N* dimensional feature vector
  - *N*: Total number of clusters
  - *feature value*: total frequency of patterns that belong to a cluster
  - feature vectors are normalized to unit length
- Using a labeled dataset of positive and negative instances, we learn a Mahalanobis distance metric.
  - Mahalanobis distance between two vectors **x** and **y** is defined by,

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

where A is the Mahalanobis matrix.

- We use the Information Theoretic Metric Learning algorithm proposed by Davis et al. 2007.
  - No eigenvalue or eigenvector computations are required
  - Scalable to large datasets via lower rank approximations
  - Can incorporate slack variables

ENT Dataset

ENT

- 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*])

Relation Classification Task

- For each word pair (P,Q) in the ENT dataset:
  - Measure the relational similarity between (P,Q) and the remaining 99 word pairs.
  - Rank the most similar *k* word pairs . (k=10)
  - Use average precision to measure the ranking.

Google	You Tube
ACQUIRER-ACQUIREE	

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

Microsoft	Powerset	ACQUIRER-ACQUIREE
Yahoo	Inktomi	ACQUIRER-ACQUIREE
Gauss	Mathematics	PERSON-FIELD
Einstein	Physics	PERSON-FIELD
Microsoft	Redmond	COMPANY-HEADQUARTERS
Eric Schmidt	Google	CEO-COMPANY

Results – Relation Classification Task

Relation	VSM	LRA	EUC	PROPOSED
ACQUIRER-ACQUIREE	92.7	92.24	91.47	94.15
COMPANY-HEADQARTERS	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)
- **EUC**: Euclidean distance between cluster vectors
- **PROPOSED**: Proposed method (Learned Mahalanobis distance between entity-pairs)

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** start **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**

Acquisition Relation

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** start **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**

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** start **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**

PERSON-FIELD Relation

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** start **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**	[illegible]	[illegible]	ference 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**

PERSON-BIRTHPLACE Relation

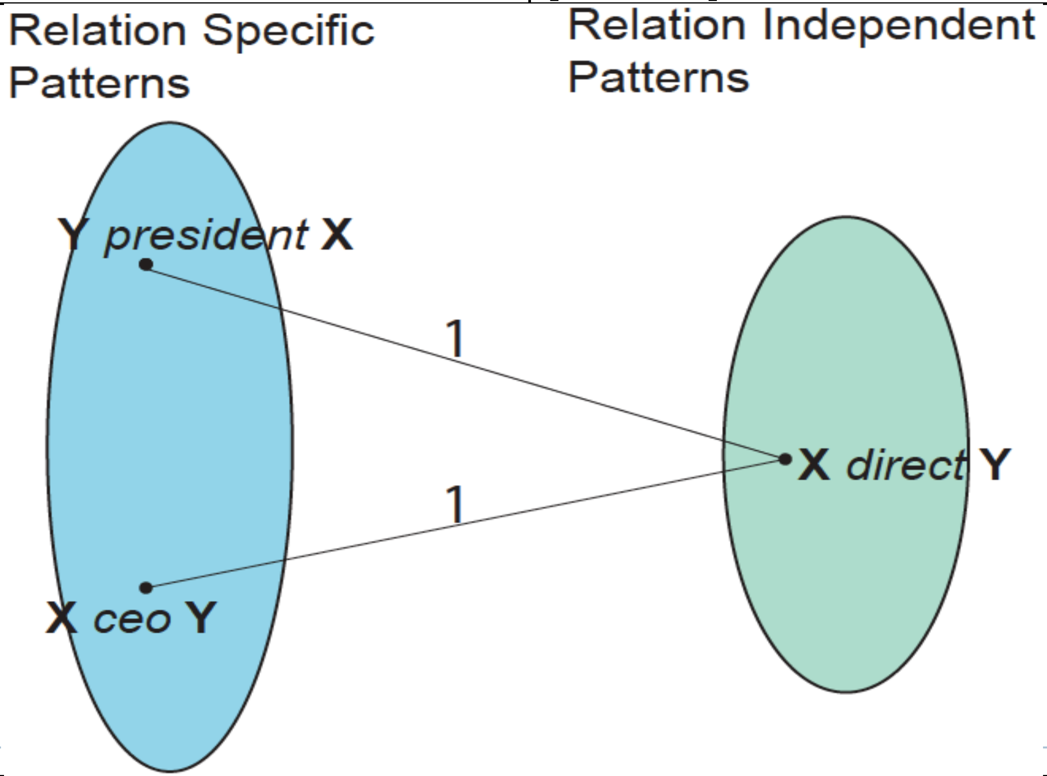
関係の分野適応 - Relation Adaptation

- Given training instances for some **source** relations $S_1,\ldots,S_k$ and some **seed** instances for a **target** relation T, learn a classifier to extract target relation.
- Characteristics of relation adaptation
  - Multiple source relation types
  - Many training instances for the source relations
  - Only a few (seeds) for the target relation type
  - We are only interested in obtaining good performance on the target relation type

Relation Extraction + Domain Adaptation = Relation Adaptation

Relational Mapping

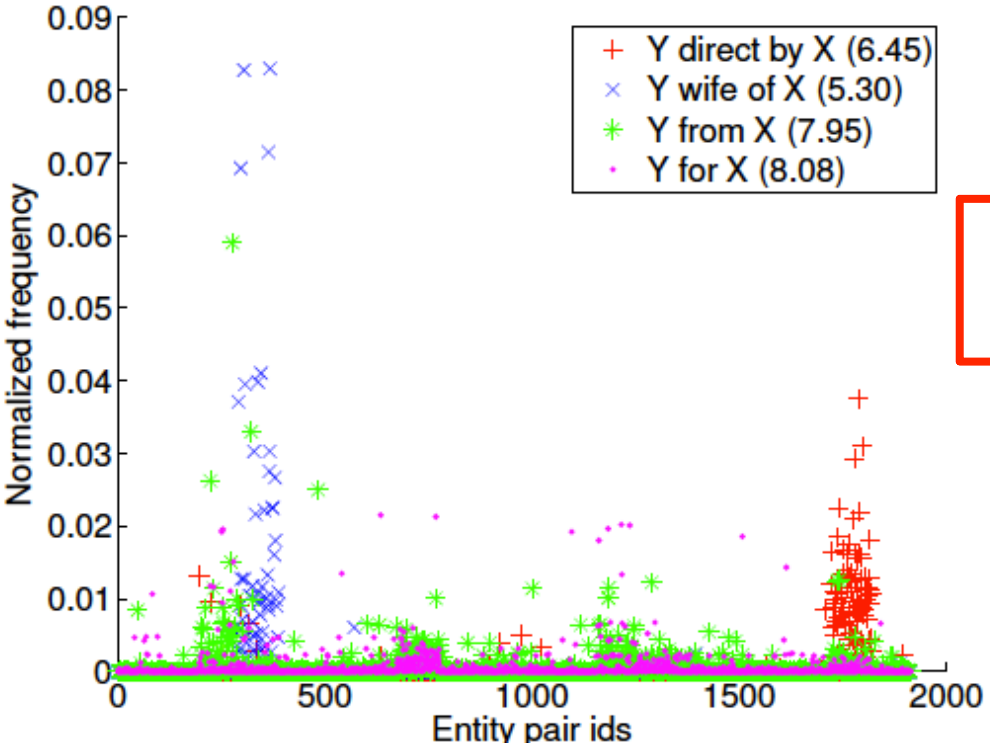
leaderOf (source relation)	ceoOf (target relation)
President George Bush directed U.S. to an unnecessary war against Iraq. [X *direct* Y]	Steve Jobs personally directs Apple and make final decisions on various UI designs. [X *direct* Y]
U.S. president George Bush attended the G8 summit last month. [Y *president* X]	Steve Jobs is the CEO of Apple, which he co-founded in 1976. [X *ceo* Y]



Relational Duality: Unsupervised Extraction of Semantic Relations between Entities, WWW 2010.

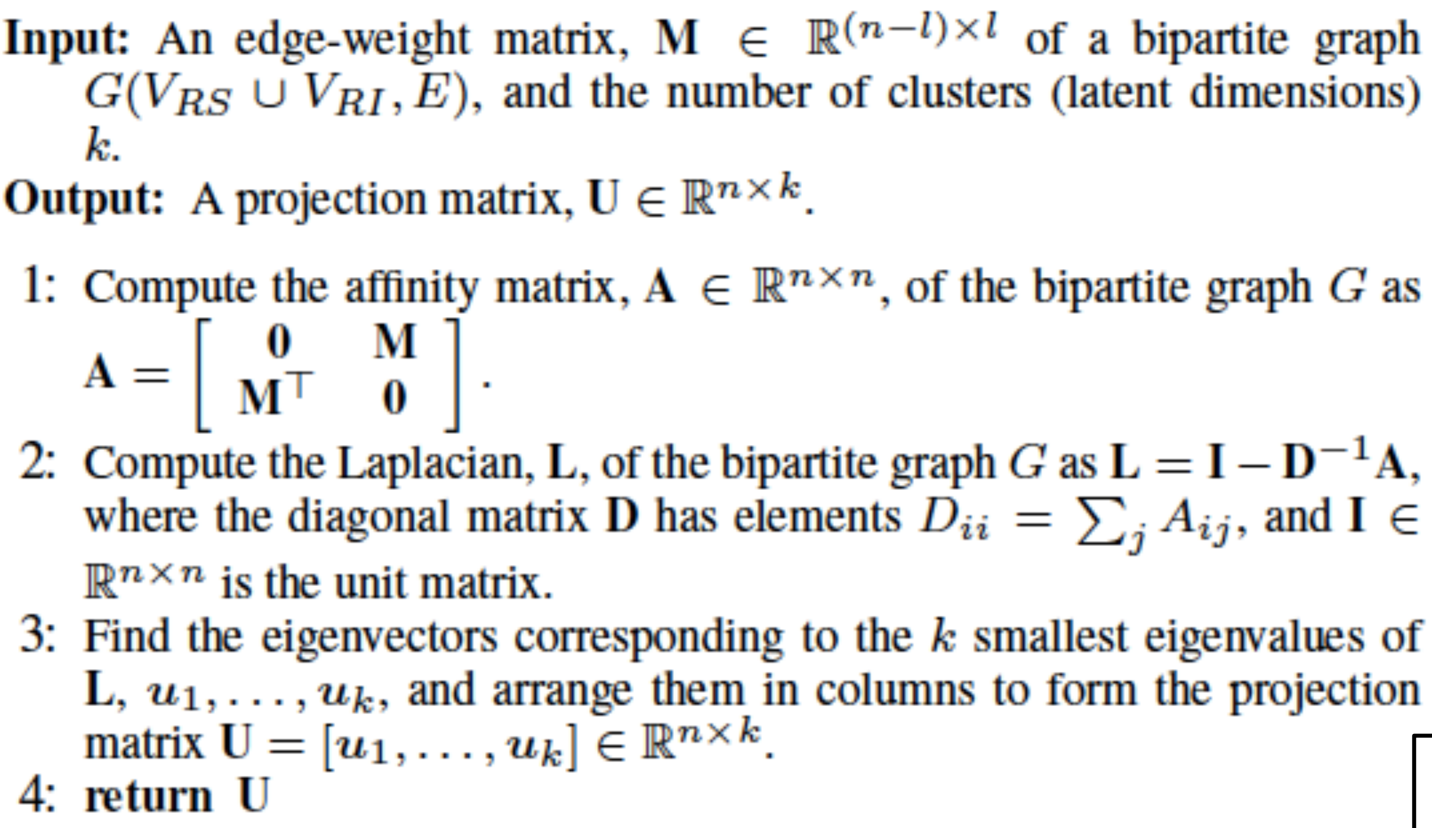
Recognizing Relation Independent Patterns

- Entropy of a pattern as a measure of independence
  - Hypothesis
    - If a pattern co-occurs with numerous entity pairs that have different relation types, then that pattern is relation independent.



$$\boxed{H(\rho) = \sum_{\mathcal{R}\in\Omega} \sum_{(A,B)\in\mathcal{R}} p(\rho, A, B) \log_2 p(\rho, A, B).}$$

Relational Mapping Algorithm



```
Input: An edge-weight matrix, $\mathbf{M} \in \mathbb{R}^{(n-l)\times l}$ of a bipartite graph
    $G(V_{RS} \cup V_{RI}, E)$, and the number of clusters (latent dimensions)
    $k$.
Output: A projection matrix, $\mathbf{U} \in \mathbb{R}^{n\times k}$.
  Compute the affinity matrix, $\mathbf{A} \in \mathbb{R}^{n\times n}$, of the bipartite graph $G$ as
    $\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M}^\top & \mathbf{0} \end{bmatrix}$.
  Compute the Laplacian, $\mathbf{L}$, of the bipartite graph $G$ as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$,
    where the diagonal matrix $\mathbf{D}$ has elements $D_{ii} = \sum_j A_{ij}$, and $\mathbf{I} \in$
    $\mathbb{R}^{n\times n}$ is the unit matrix.
  Find the eigenvectors corresponding to the $k$ smallest eigenvalues of
    $\mathbf{L}$, $u_1, \ldots, u_k$, and arrange them in columns to form the projection
    matrix $\mathbf{U} = [u_1, \ldots, u_k] \in \mathbb{R}^{n\times k}$.
  return $\mathbf{U}$
```

creating a bi-partite graph

spectral clustering

lower dimensional mapping

今後の課題と展望

- 関係をどう表現するか
  - それぞれのエンティティの属性間の対応として表現する
  - 関係の特徴（属性）として表現する
- 関係の間の関係をどう表現するか
  - ４次のテンソルとして表現可能？
  - Webのような膨大なデータの場合はどう計算するか
- 多項関係(multinomial relation)をどう抽出するか
- 可変多項関係を（テンソルで）どう表現するか
- 関係の分野適応
  - どんな関係ならば分野適応可能か (negative transfer)
- SemEval 2012で関係類似性計測タスクをやります

Thank You

Contact:
Email: danushka@iba.t.u-tokyo.ac.jp
Web: www.iba.t.u-tokyo.ac.jp/~danushka
Twitter: @Bollegala