Mining Meaning From Wikipedia – part 2

PD Dr. Günter Neumann

LT-lab, DFKI, Saarbrücken

Outline

1. Namend Entity Disambiguation
2. Relation Extraction
3. Ontology Building and the Semantic Web

Disambiguation of named entities

- Wikipedia is recognized as the largest available resource of named entities!
- Approaches considered here:
  - Linking named entities appearing in text or in search queries to corresponding Wikipedia articles for disambiguation.

Bunescu & Pasca, 2006

- Disambiguate NE in search queries in order to group search results by the corresponding senses.
- Core idea: Use dictionary for identifying possible NE meanings, that can also be used for disambiguation.
- Use Wikipedia as a resource for creating such a dictionary:
  - A dictionary of 500.000 NE that appear in Wikipedia
  - Note that NE are mainly mentioned as titles in articles, and hence some preprocessing of articles are required, e.g., capitalization, multi words
  - add redirects and disambiguated names to each entry
    - Capture the many-to-many relationship between names and entities

Examples

TITLE	REDIRECT	DISAMBIG	CATEGORIES
John Williams (composer)	John Towner Williams	John Williams	Star Wars music, ... Film score composers, 20th century classical composers
John Williams (wrestler)	Ian Rotten	John Williams	Professional wrestlers, People living in Baltimore
John Williams (VC)	*none*	John Williams	British Army soldiers, British Victoria Cross recipients
Boston Pops Orchestra	Boston Pops, The Boston Pops Orchestra	Pops	American orchestras, Massachusetts musicians
United States	US, USA, ... United States of America	US, USA, United States	North American countries, Republics, United States
Venus (planet)	Planet Venus	Venus, Morning Star, Evening Star	Venus *Planets of the Solar System*, *Planets*, *Solar System*, ...

Table 1: Examples of Wikipedia titles, aliases and categories

Approach

- If a query contains a term that corresponds to 2 or more entries, chose the one whose Wikipedia article has greatest cosine similarity with the query.
- If similarity value is too low, use category to which the article belongs.
- Result: 55%-85% accuracies for Wikipedia's People by occupation category.

Cucerzan, 2007

- Disambiguate NE in free text
- Also uses a compiled dictionary from Wikipedia with two parts
  - Surface forms, s.a. Article titles, redirects, ...
  - Associated entities together with contextual information about them
  - ~1.4M entries, average of 2.4 surface forms each

Further extracted data structures

- <NE,tag> entries from Wikipedia list articles
  - Texas (band) → LIST_band name etymologies
  - Because it appears in a list with this title
  - 540.000 entries
- Categories assigned to Wikipedia articles decribing NE used as tag too
  - 2.65M entries
- A context for each NE is collected from its article
  - 38M entries <NE, context>

Identification of NE in text

- Capitalization rules indicate which phrases are surface forms of NE
- Co-occurrence statistics from web by a search engine are used to identify boundaries (Google n-grams)
  - „Whitney Museum of American Art" vs. „Whitney Museum in New York" (about 2.450.000 vs. 405.000 hits)
- Lexical analysis to collect identical NE (Mr. Brown & Brown)
- Disambiguation on basis of similarity of
  - the document in which the surface form appears
  - with Wikipedia articles that represent all NE that have been identified in it, and their context terms
- Result:
  - 88% accuracy/5000 Wikipedia entities
  - 91% accuracy on 750 news article entities

Disambiguating thesaurus & ontology terms

If
cardiovascular s. = circulatory s.
Then
add blood circulation to Agrovoc

- W
  a
  co
  of information as domain-specific thesaurus.
- Thus it can also be used to extend and improve resources.

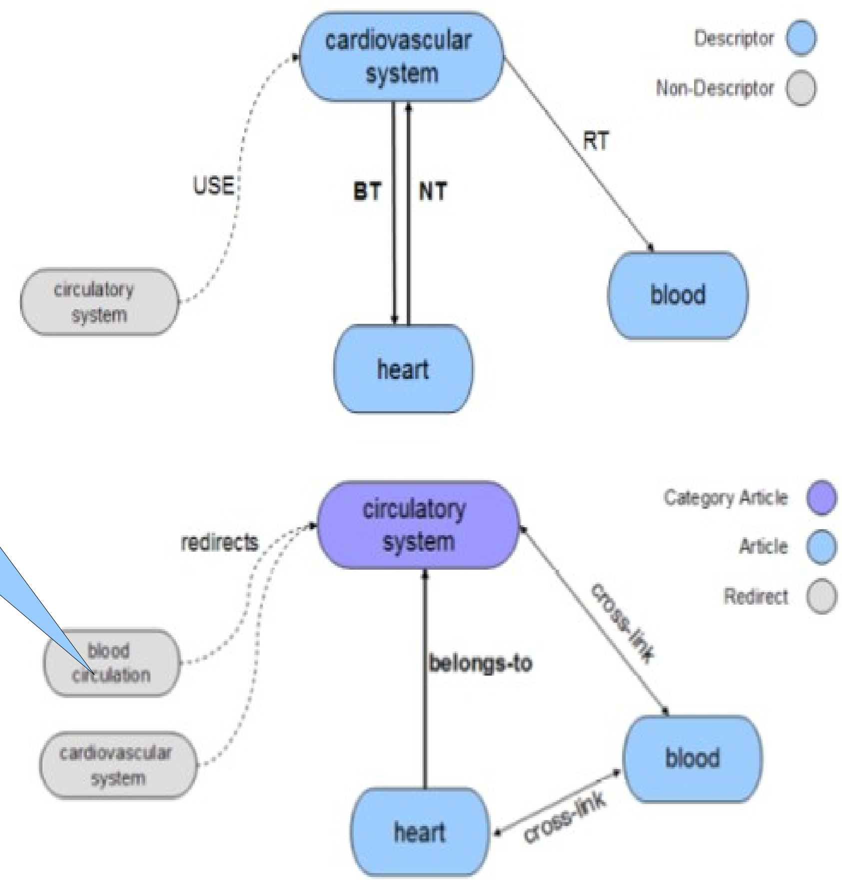


Figure 4. Comparison of organization structure in Agrovoc and Wikipedia.

Establishing Mappings between Wikipedia and other Resources

- Mapping Wikipedia articles to WordNet
- Idea, cf. Ruiz-Casado et al. 2005
  - If a Wikipedia article matches several WordNet synsets, the appropriate one is chosen by computing the similarity between the Wikipedia entry word-bag and the WordNet synset gloss.
- Gets 84% accuracy using dot product on stemmed texts.
- Problem: if (Simple) Wikipedia is growing
  - Ambiguities increase as well
  - Mapping gaps
    - We have observed similar problem when using WordNet for crosslingual QA!

Intermediate summary

http://mostpopularwebsites.net/

Today's Most Popular Websites on the Internet:

Ads by Google 2010 WWW Google com 31 12 2010 Best Bar 2010

The following list of the **Most Popular Websites** was updated on **Saturday, January 1st 2011**...

1.	Google.com	www.google.com
2.	Facebook.com	www.facebook.com
3.	Youtube.com	www.youtube.com
4.	Yahoo.com	www.yahoo.com
5.	Live.com	www.live.com
6.	Baidu.com	www.baidu.com
7.	Wikipedia.org	www.wikipedia.org
8.	Blogspot.com	www.blogspot.com
9.	Qq.com	www.qq.com
10.	Twitter.com	www.twitter.com
11.	Blogger.com	www.blogger.com
12.	Yahoo.co.jp	www.yahoo.co.jp
13.	Amazon.com	www.amazon.com
14.	Taobao.com	www.taobao.com
15.	Google.co.in	www.google.co.in
16.	Sina.com.cn	www.sina.com.cn
17.	Google.de	www.google.de
18.	Google.com.hk	www.google.com.hk
19.	Wordpress.com	www.wordpress.com
20.	Linkedin.com	www.linkedin.com
21.	Ebay.com	www.ebay.com
22.	Google.co.uk	www.google.co.uk
23.	Microsoft.com	www.microsoft.com
24.	Yandex.ru	www.yandex.ru
25.	Google.fr	www.google.fr
26.	Bing.com	www.bing.com
27.	Google.co.jp	www.google.co.jp
28.	163.com	www.163.com
29.	Google.com.br	www.google.com.br
30.	Fc2.com	www.fc2.com
31.	Craigslist.com	www.craigslist.com
32.	Conduit.com	www.conduit.com
33.	Google.it	www.google.it
34.	Flickr.com	www.flickr.com
35.	Vkontakte.ru	www.vkontakte.ru
36.	Charlotte.craigslist.org	www.charlotte.craigslist.org
37.	Apple.com	www.apple.com
38.	Googleusercontent.com	www.googleusercontent.com
39.	Craigslist.org	www.craigslist.org
40.	Google.es	www.google.es
41.	Go.com	www.go.com

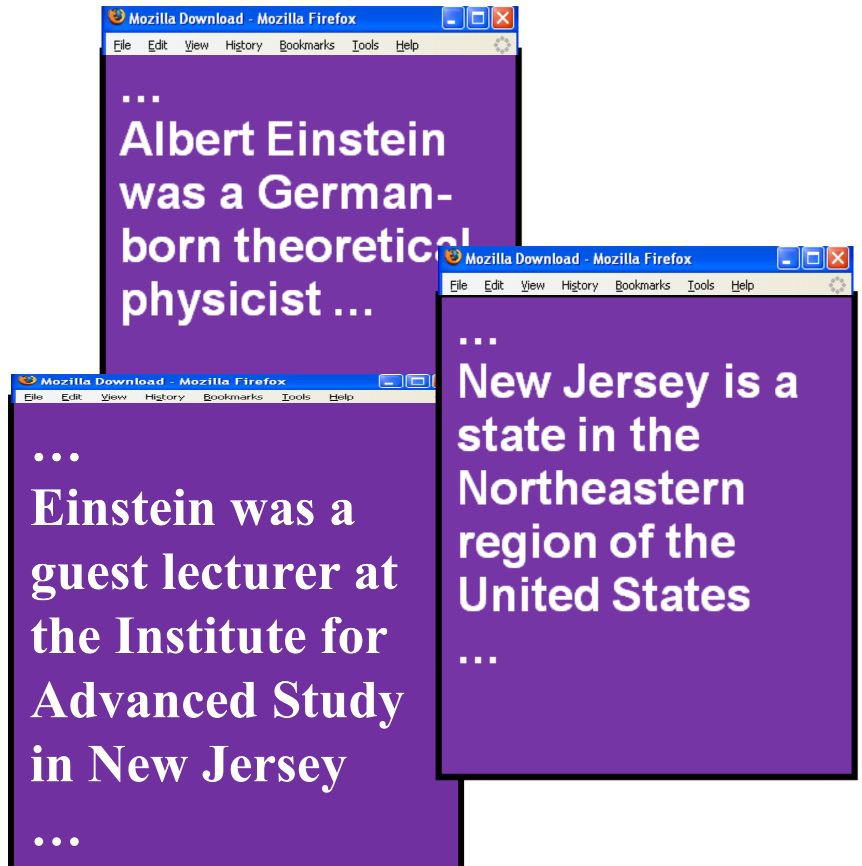
Benefit to Wikipedia: Tools

- Internal link maintenance
- Infobox Creation
- Schema Management
- Reference suggestion & fact checking
- Disambiguation page maintenance
- Translation across languages
- Vandalism Alerts

Motivating Vision

Next-Generation Search = Information Extraction
+ Ontology
+ Inference

Which German Scientists Taught at US Universities?



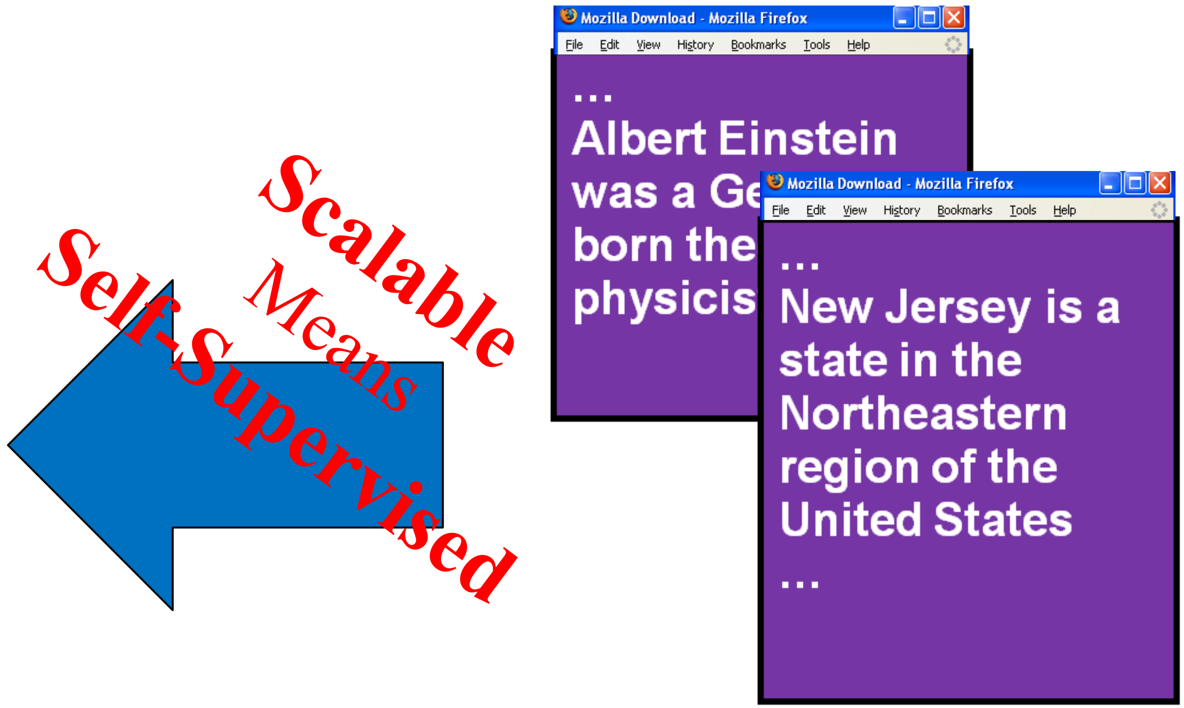
Next-Generation Search

- **Information Extraction**
  - <Einstein, Born-In, Germany>
  - <Einstein, ISA, Physicist>
  - <Einstein, Lectured-At, IAS>
  - <IAS, In, New-Jersey>
  - <New-Jersey, In, United-States>

    …
- **Ontology**
  - Physicist (x) ➔ Scientist(x)

    …
- **Inference**
  - Einstein = Einstein

    …



Extracting more Structure

Microsoft

From Wikipedia, the free encyclopedia

Coordinates: 47°38′22.55″N 122°7′42.42″W

Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses, and supports a wide range of products and services predominantly related to computing through its various product divisions. Established on April 4, 1975 to develop and sell BASIC interpreters for the Altair 8800, Microsoft rose to dominate the home computer operating system (OS) market with MS-DOS in the mid-1980s, followed by the Microsoft Windows line of OSs. Microsoft would also come to dominate the office suite market with Microsoft Office. The company has diversified in recent years into the video game industry with the Xbox and its successor, the Xbox 360 as well as into the consumer electronics market with Zune and the Windows Phone OS. The ensuing rise of stock in the company's 1986 initial public offering (IPO) made an estimated four billionaires and 12,000 millionaires from Microsoft employees.

Primarily in the 1990s, critics contend Microsoft used monopolistic business practices and anti-competitive strategies including refusal to deal and tying, put unreasonable restrictions in the use of its software, and used misrepresentative marketing tactics; both the U.S. Department of Justice and European Commission found the company in violation of antitrust laws. Known for its interviewing process with obscure questions, various studies and ratings were generally favorable to Microsoft's diversity within the company as well as its overall environmental impact with the exception of the electronics portion of the business.

Contents [hide]

1 History
1.1 1984–1994: Windows and Office
1.2 1995–2005: Internet and the 32-bit era
1.3 2006 on: Vista and Cloud computing
2 Product divisions
2.1 Windows & Windows Live Division, Server and Tools, Online Services Division
2.2 Business Division, Entertainment and Devices Division
3 Culture
4 Corporate affairs
4.1 Environment
4.2 Marketing
5 See also
6 References
7 External links

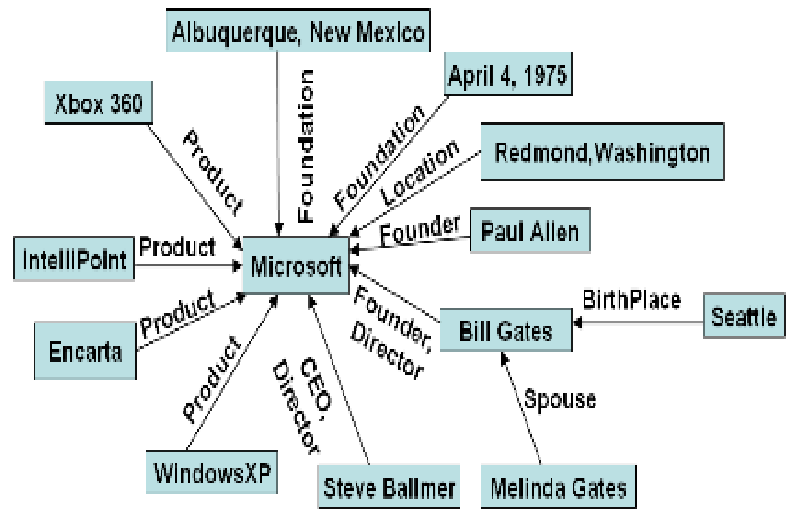
History

Main articles: History of Microsoft and History of Microsoft Windows

Paul Allen and Bill Gates, childhood friends with a passion in computer programming, were seeking to make a successful business utilizing their shared skills. The January 1975 issue of Popular Electronics featured Micro Instrumentation and Telemetry Systems's (MITS) Altair 8800 microcomputer.

Microsoft Corporation

Type	Public (NASDAQ: MSFT) Dow Jones Industrial Average Component S&P 500 Component
Industry	Computer software Consumer electronics Digital distribution Computer hardware Video games IT consulting Online advertising Retail stores Automotive software
Founded	Albuquerque, New Mexico April 4, 1975
Founder(s)	Bill Gates Paul Allen
Headquarters	One Microsoft Way Redmond, Washington, United States
Area served	Worldwide
Key people	Steve Ballmer (CEO) Brian Kevin Turner (COO) Bill Gates (Chairman) Ray Ozzie (CSA) Craig Mundie (CRSO)
Products	See products listing
Services	See services listing
Revenue	▲ $62.484 billion (2010)
Operating income	▲ $24.098 billion (2010)
Profit	▲ $18.760 billion (2010)
Total assets	▲ $86.113 billion (2010)
Total equity	▲ $46.175 billion (2010)
Employees	89,000 (2010)



Relation Extraction

- Basically means:
  - extract semantic relations from unstructured text
- Example:
  - Apple Inc.'s world corporate headquarters are located in the middle of Silicon Valley, at 1 Infinite Loop, Cupertino, California.
  - hasHeadquarters(Apple Inc., 1 Infinite Loop-Cupertino-California)
- Challenges:
  - Extract this relation from sentences expressing the same information about Apple Inc., regardless of the actual wording.
  - Same relation should be determined with different arguments from similar sentences
    - hasHeadquarters(Google Inc., Google Campus-Mountain View-California)

Wikipedia: Relation Extraction

- Semantic relations in Wikipedia's raw text
- Semantic relations in structured parts of Wikipedia
- Typing Wikipedia's named entities

Semantic relations in Wikipedia's raw text

- Standard approach:
  - Take known binary relations as seeds
  - Extract patterns from their textual representation, e.g.,
    - X's * headquarters are located in * at Y
  - Patterns are then applied to identify new relations as values from X and Y
- Known difficulties of this approach:
  - Enumerating over all pairs of entities yields a low density of correct relations even when restricted to a single sentence
  - Errors in the entity recognition stage create inaccuracies in relation classification.

Anchestors of Wikipedia-based relation extraction from text

- Brin, S.: Extracting patterns and relations from the World Wide Web, 1998
- Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections, 2002.
- Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system, 2002.
- There is also close relationship to Machine Reading

Snowball : Extracting Relations from Large Plain-Text Collections

Eugene Agichtein
Luis Gravano

Department of Computer Science
Columbia University,

ACM paper 2000

Example Task: Organization/Location

Redundancy

Microsoft's central headquarters in *Redmond* is home to almost every product group and division.

Brent Barlow, 27, a software analyst and beta-tester at *Apple Computer* headquarters in *Cupertino*, was fired Monday for "thinking a little too different."

Apple's programmers "think different" on a "campus" in *Cupertino, Cal*. *Nike* employees "just do it" at what the company refers to as its "World Campus," near *Portland, Ore*.

Organization	*Location*
Microsoft	Redmond
Apple Computer	Cupertino
Nike	Portland

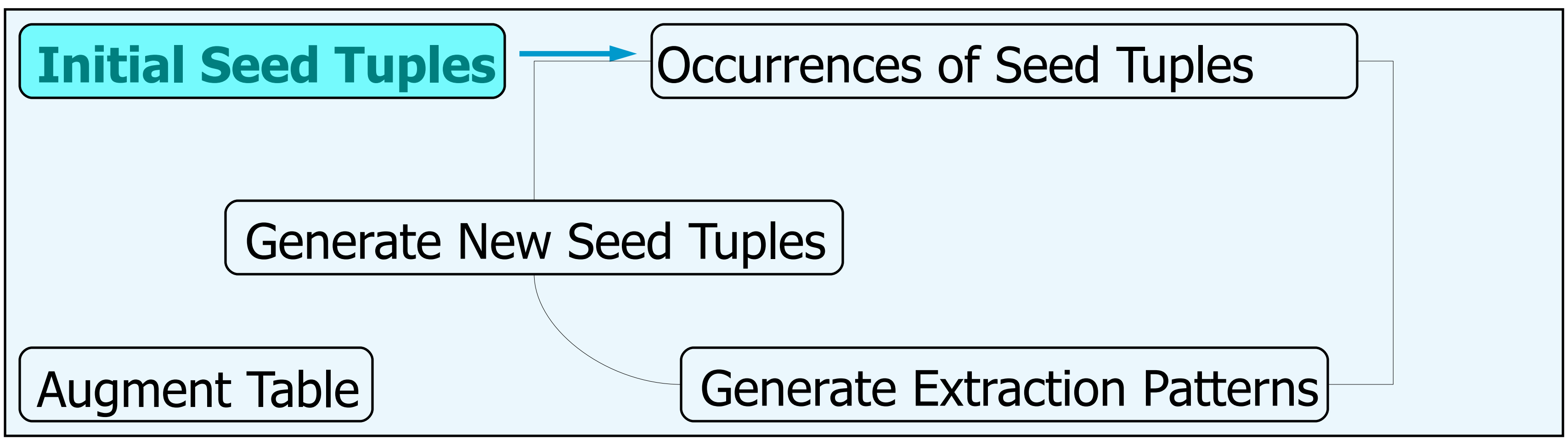
Related Work

- Bootstrapping
  - Riloff et al. ('99), Collins & Singer ('99)
    - (Named-entity recognition – see previous lectures of this course,)
  - Brin (DIPRE) ('98)

Extracting Relations from Text: DIPRE

Initial Seed Tuples:

ORGANIZATION	*LOCATION*
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA



Extracting Relations from Text: DIPRE

Occurrences of seed tuples:

ORGANIZATION	*LOCATION*
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA

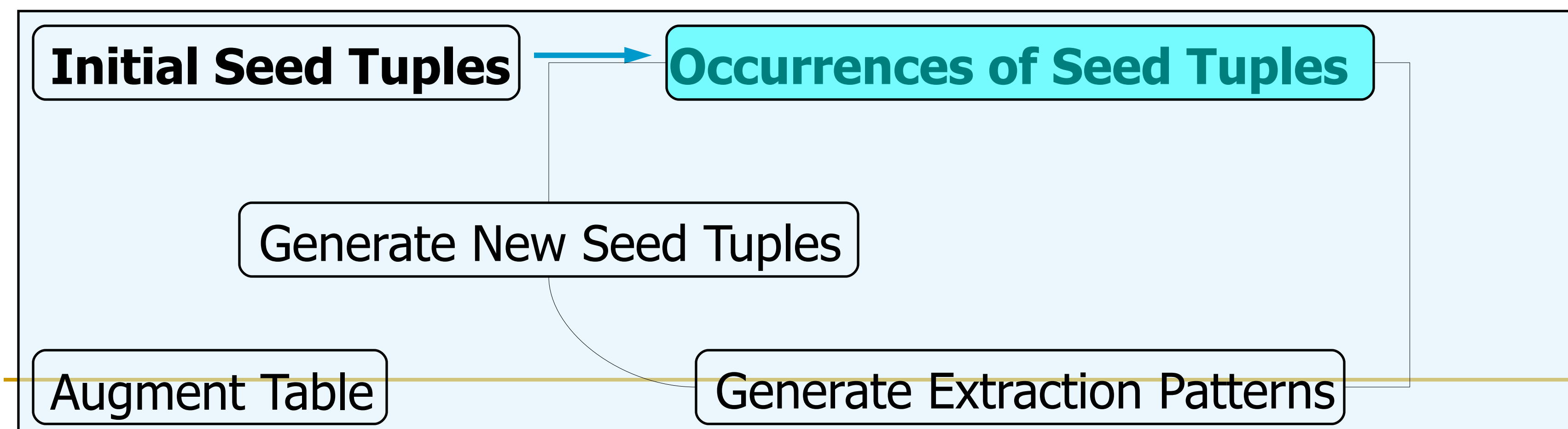
Computer servers at **Microsoft**'s headquarters in **Redmond**…

In mid-afternoon trading, share of **Redmond**-based **Microsoft** fell…

The **Armonk**-based **IBM** introduced a new line…

The combined company will operate from **Boeing**'s headquarters in **Seattle**.

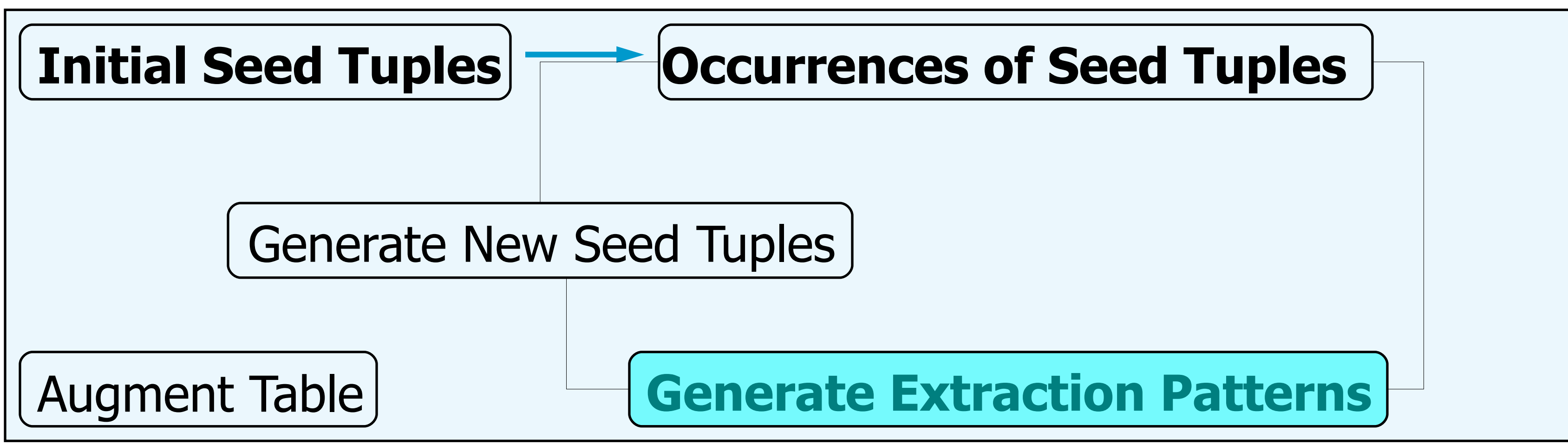
Intel, **Santa Clara**, cut prices of its Pentium processor.



Extracting Relations from Text: DIPRE

DIPRE Patterns:

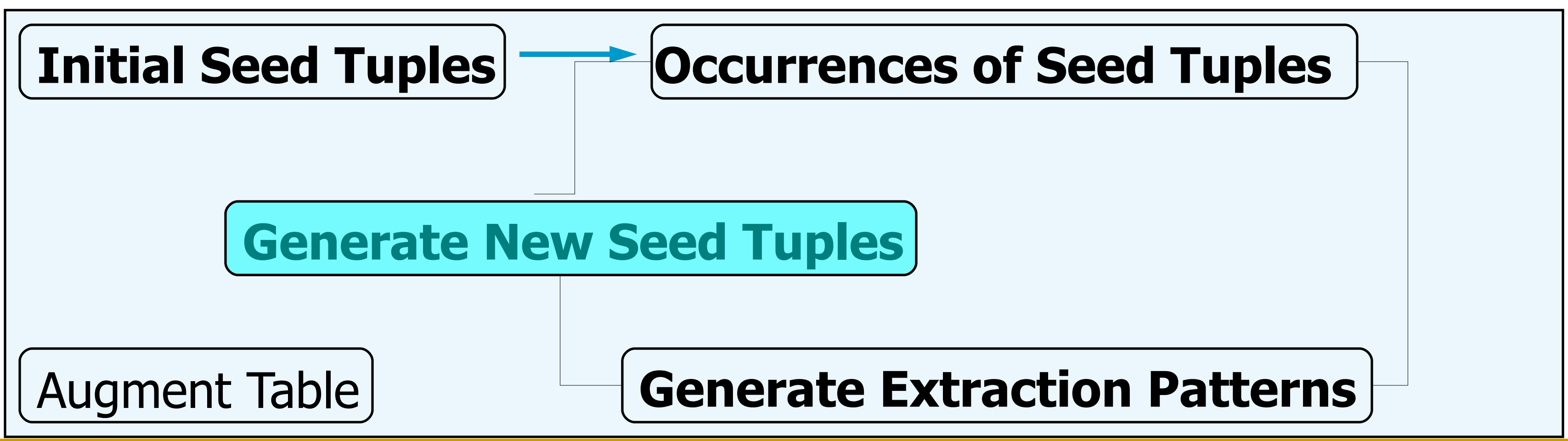
- <*STRING1*>'s headquarters in <*STRING2*>
- <*STRING2*> -based <*STRING1*>
- <*STRING1*> , <*STRING2*>



Extracting Relations from Text: DIPRE

Generate new seed tuples; start new iteration

ORGANIZATION	***LOCATION***
AG EDWARDS	ST LUIS
157TH STREET	MANHATTAN
7TH LEVEL	RICHARDSON
3COM CORP	SANTA CLARA
3DO	REDWOOD CITY
JELLIES	APPLE
MACWEEK	SAN FRANCISCO



Extracting Relations from Text: Potential Pitfalls

- Invalid tuples generated
  - Degrade quality of tuples on subsequent iterations
  - Must have automatic way to select high quality tuples to use as new seed
- Pattern representation
  - Patterns must generalize

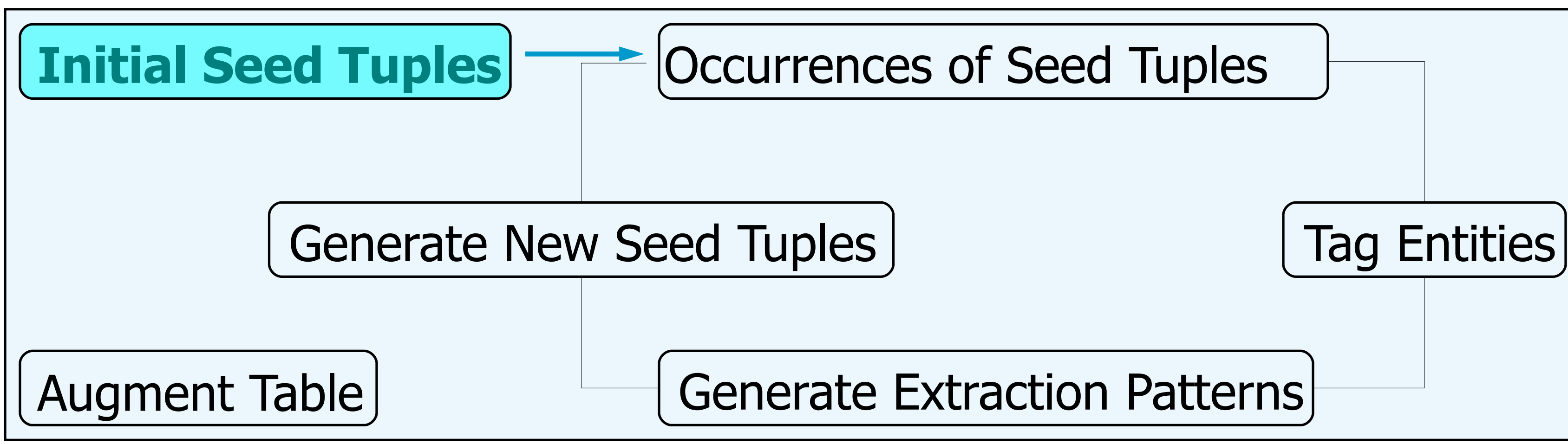
Extracting Relations from Text Collections

- Related Work
  - DIPRE
- The *Snowball* System:
  - Pattern representation and generation
  - Tuple generation
  - Automatic pattern and tuple evaluation
- Evaluation Metrics
- Experimental Results

Extracting Relations from Text: *Snowball*

Initial Seed Tuples:

ORGANIZATION	***LOCATION***
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA



Extracting Relations from Text: *Snowball*

Occurrences of seed tuples:

ORGANIZATION	***LOCATION***
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA

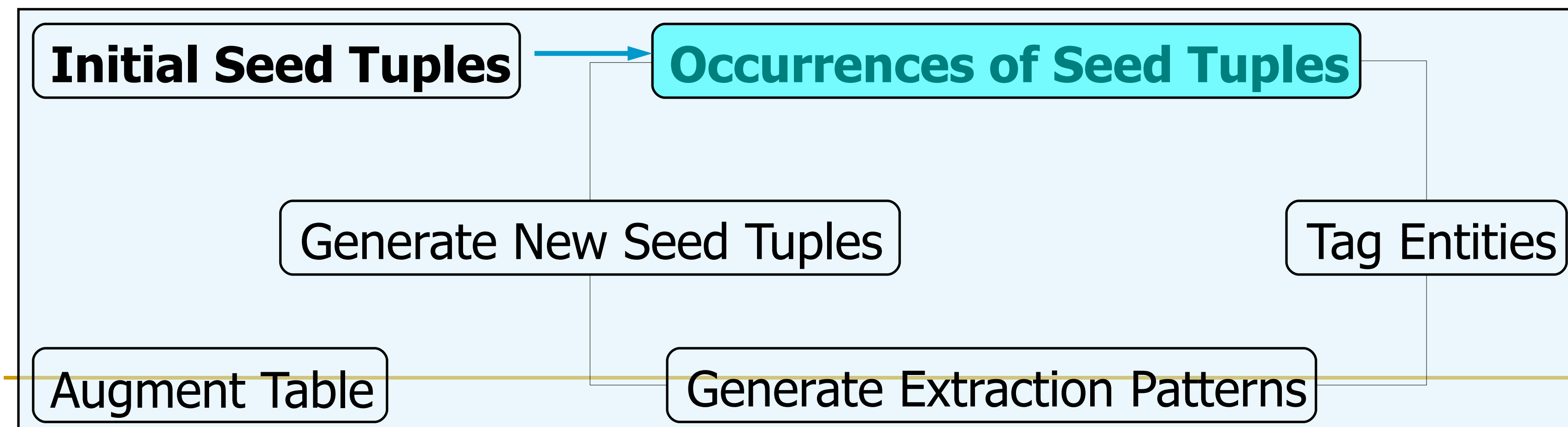
Computer servers at **Microsoft**'s headquarters in **Redmond**…

In mid-afternoon trading, share of **Redmond**-based **Microsoft** fell…

The **Armonk**-based **IBM** introduced a new line…

The combined company will operate from **Boeing**'s headquarters in **Seattle**.

Intel, **Santa Clara**, cut prices of its Pentium processor.



Problem: Patterns Excessively General

Pattern: <*STRING2*>-based <*STRING1*>

Today's merger with McDonnell Douglas positions ***Seattle*** ***-based*** ***Boeing*** to make major money in space.

…, a producer of ***apple-based jelly***, …

<***jelly***, ***apple***>

Incorrect!

Extracting Relations from Text: *Snowball*

Tag Entities

Use MITRE's Alembic Named Entity tagger

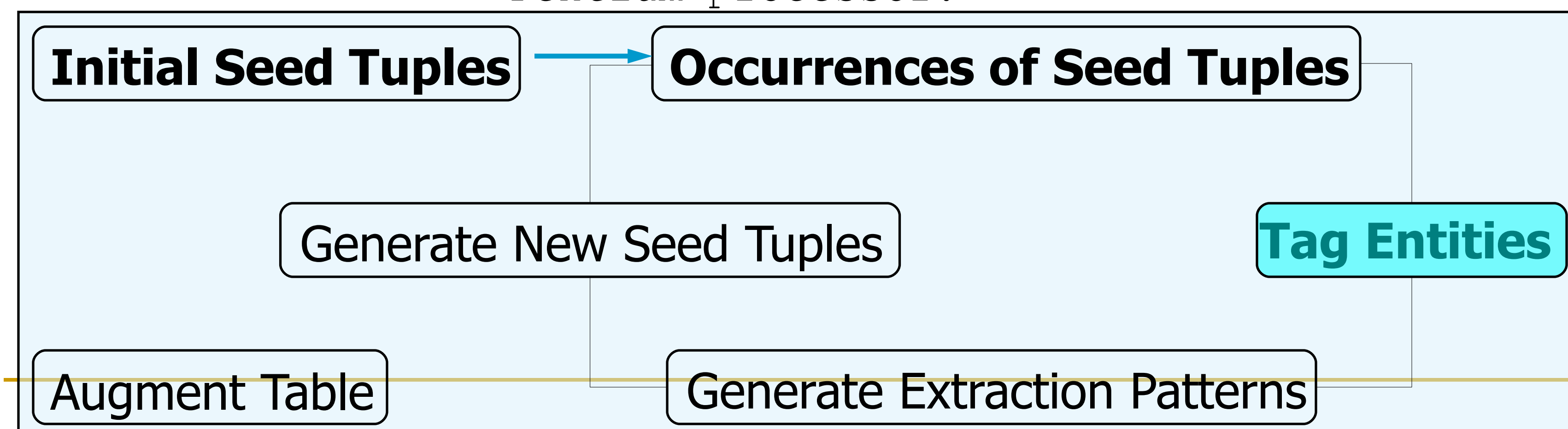
Computer servers at **Microsoft**'s headquarters in **Redmond**…

In mid-afternoon trading, share of **Redmond**-based **Microsoft** fell…

The **Armonk**-based **IBM** introduced a new line…

The combined company will operate from **Boeing**'s headquarters in **Seattle**.

Intel, **Santa Clara**, cut prices of its Pentium processor.



Extracting Relations from Text

- <*ORGANIZATION*>'s headquarters in <*LOCATION*>
- <*LOCATION*> -based <*ORGANIZATION*>
- <*ORGANIZATION*> , <*LOCATION*>

PROBLEM: Patterns too specific: have to match text *exactly*.

Initial Seed Tuples

Occurrences of Seed Tuples

Generate New Seed Tuples

Tag Entities

Augment Table

Generate Extraction Patterns

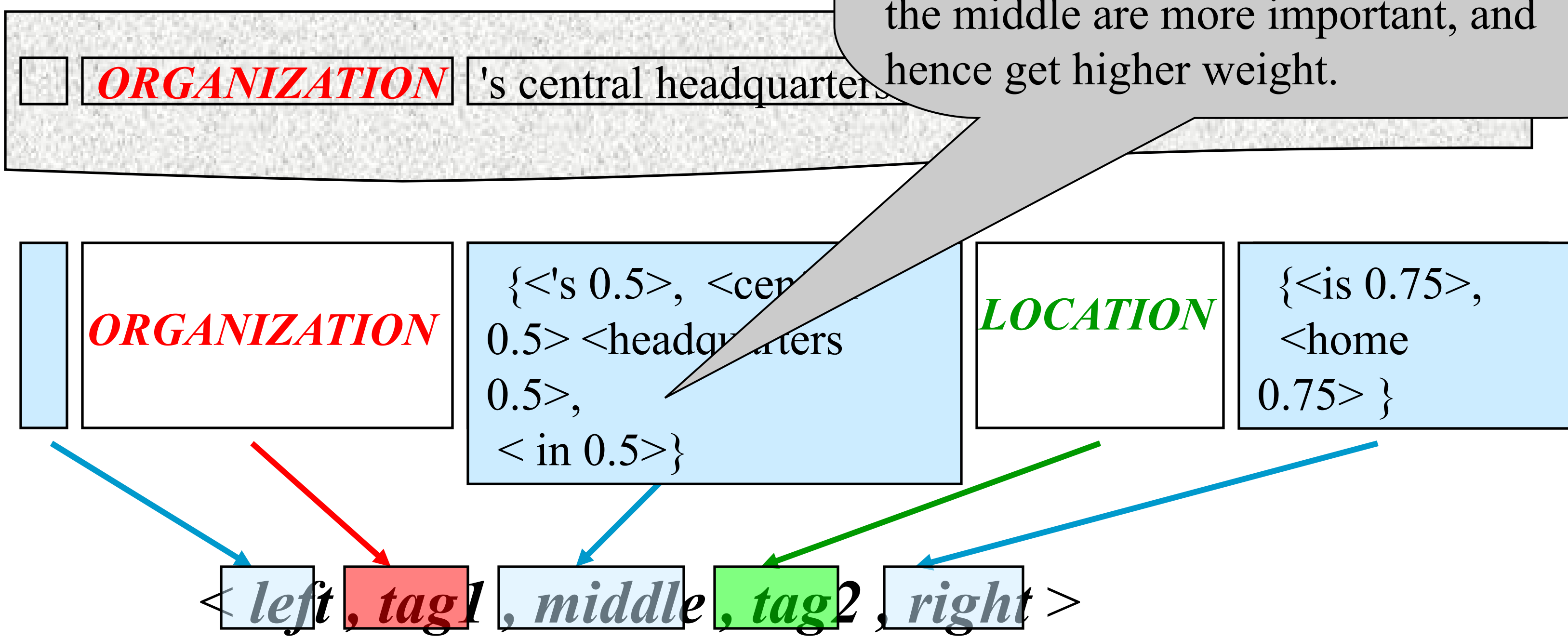
Snowball: Pattern Representation

A *Snowball* pattern vector is a 5-tuple

<***left, tag1, middle, tag2, right***>

- ***tag1***, ***tag2*** are named-entity tags
- ***left***, ***middle***, and ***right*** are vectors of

The weight of a term in each vector is a function of the frequency of the term in the corresponding context. These vectors are scaled so their norm is one. FInally, they are mulitplied by a scaling factor to indicate each vector's relative importance. From experiments, terms in the middle are more important, and hence get higher weight.



Snowball: Pattern Generation

Tagged Occurrences of seed tuples:

Computer servers at **Microsoft**'s central headquarters in **Redmond**…

In mid-afternoon trading, share of **Redmond**-based **Microsoft** fell…

The **Armonk** -based **IBM** introduced a new line…

The combined company will operate from **Boeing**'s headquarters in **Seattle**.

Snowball Pattern Generation: Cluster Similar Occurrences

Occurrences of seed tuples converted to *Snowball* representation:

{<servers 0.75> <at 0.75>}	ORGANIZATION	{<'s 0.5> <central 0.5> <headquarters 0.5> <in 0.5>}	LOCATION	
{<shares 0.75> <of 0.75>}	LOCATION	{<- 0.75> <based 0.75> }	ORGANIZATION	{<fell 1>}
{<the 1>}	LOCATION	{<- 0.75> <based 0.75> }	ORGANIZATION	{<introduced 0.75> <a 0.75>}
{<operate 0.75> <from 0.75>}	ORGANIZATION	{<'s 0.7> <headquarters 0.7> <in 0.7>}	LOCATION	

Similarity Metric

$P = \quad < \boxed{L_p}, \boxed{tag1}, \boxed{M_p}, \boxed{tag2}, \boxed{R_p} >$

$S = \quad < \boxed{L_s}, \boxed{tag1}, \boxed{M_s}, \boxed{tag2}, \boxed{R_s} >$

$$Match(P, S) = \begin{cases} \boxed{Lp} \cdot \boxed{Ls} + \boxed{Mp} \cdot \boxed{Ms} + \boxed{Rp} \cdot \boxed{Rs} & \text{if the tags match} \\ 0 & \text{otherwise} \end{cases}$$

Snowball Pattern Generation: Clustering

Cluster 1

{<servers 0.75> <at 0.75>}	ORGANIZATION	{<'s 0.5> <central 0.5> <headquarters 0.5> <in 0.5>}	LOCATION	
{<operate 0.75> <from 0.75>}	ORGANIZATION	{<'s 0.7> <headquarters 0.7> <in 0.7>}	LOCATION	

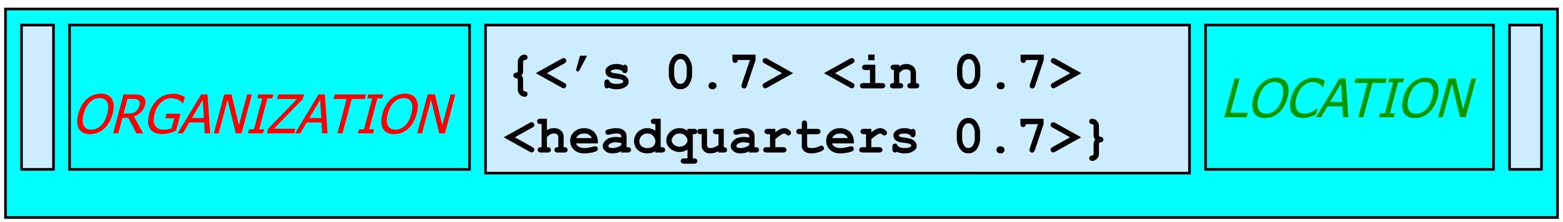
Cluster 2

{<shares 0.75> <of 0.75>}	LOCATION	{<- 0.75> <based 0.75> }	ORGANIZATION	{<fell 1>}
{<the 1>}	LOCATION	{<- 0.75> <based 0.75> }	ORGANIZATION	{<introduced 0.75> <a 0.75>}

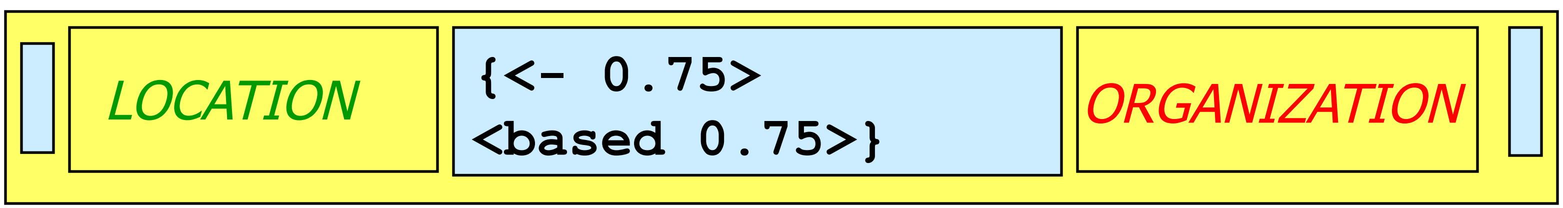
Snowball: Pattern Generation

Patterns are formed as ***centroids*** of the clusters.
Filtered by minimum number of supporting tuples.

Pattern1

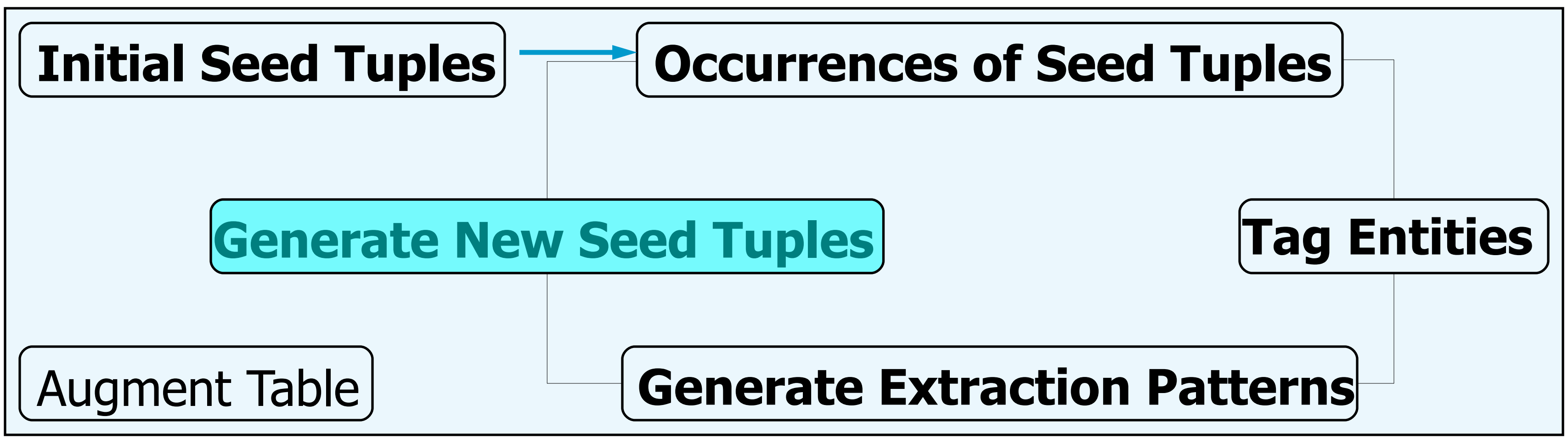


Pattern2



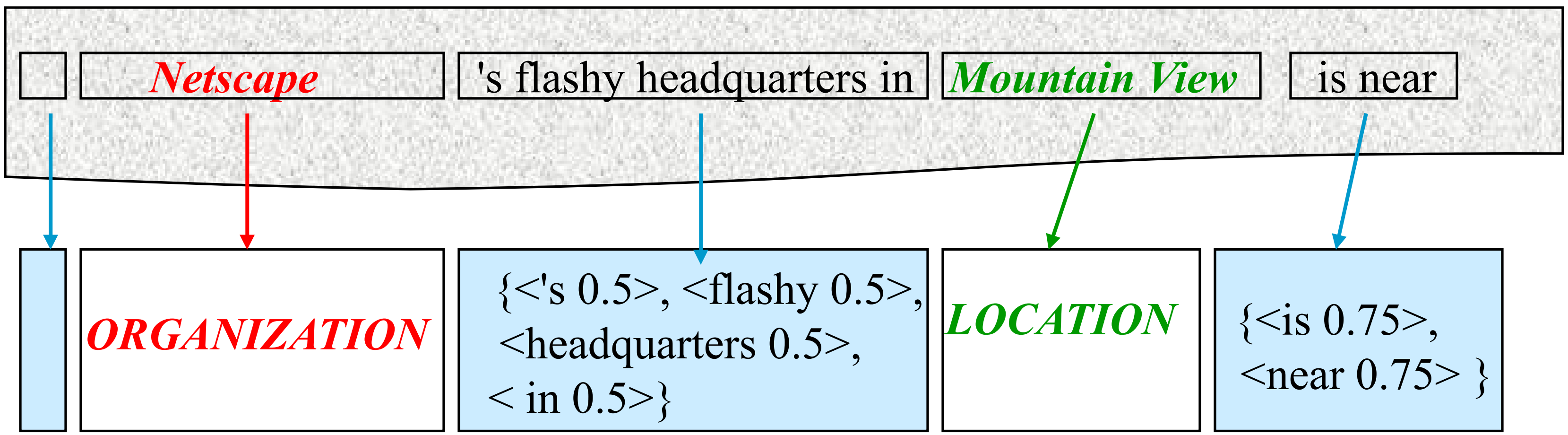
Snowball: Tuple Extraction

Using the patterns, scan the collection to generate new seed tuples:

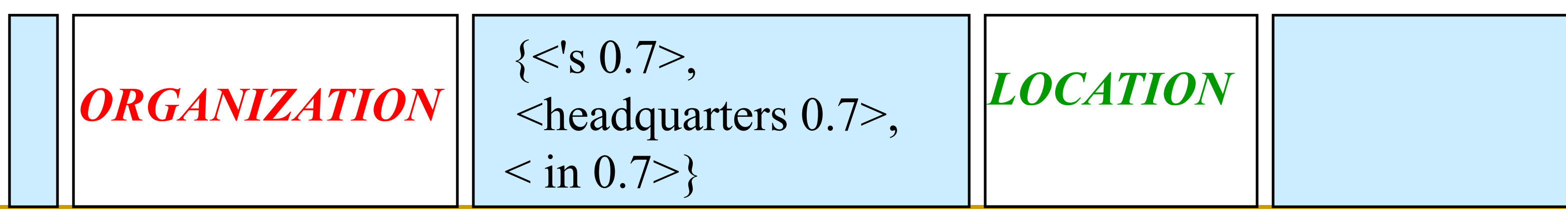


Snowball: Tuple Extraction

Represent each new text segment in the collection as the context 5-tuple:



Find most similar pattern (if any)



Snowball: Automatic Pattern Evaluation

Seed tuples

ORGANIZATION	*LOCATION*
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA

Pattern "ORGANIZATION, LOCATION" in action:

Boeing, **Seattle**, said…	**Positive**
Intel, **Santa Clara**, cut prices…	**Positive**
invest in **Microsoft**, **New York**-based analyst Jane Smith said	**Negative**

Pattern Confidence:

Automatically estimate probability of a pattern generating valid tuples:

$$\text{Conf(Pattern)} = \frac{\text{Positive}}{\text{Positive + Negative}}$$

e.g., Conf(Pattern) = 2/3 = 66%

Snowball: Automatic Tuple Evaluation

Brent Barlow, 27, a software analyst and beta-tester at ***Apple Computer*** headquarters in ***Cupertino***, was fired Monday for "thinking a little too different."

Apple's programmers "think different" on a "campus" in Cupertino, Cal.

<*Apple Computer*, *Cupertino*>

$$Conf(Tuple) = 1 - \quad (1 - Conf(P_i))$$

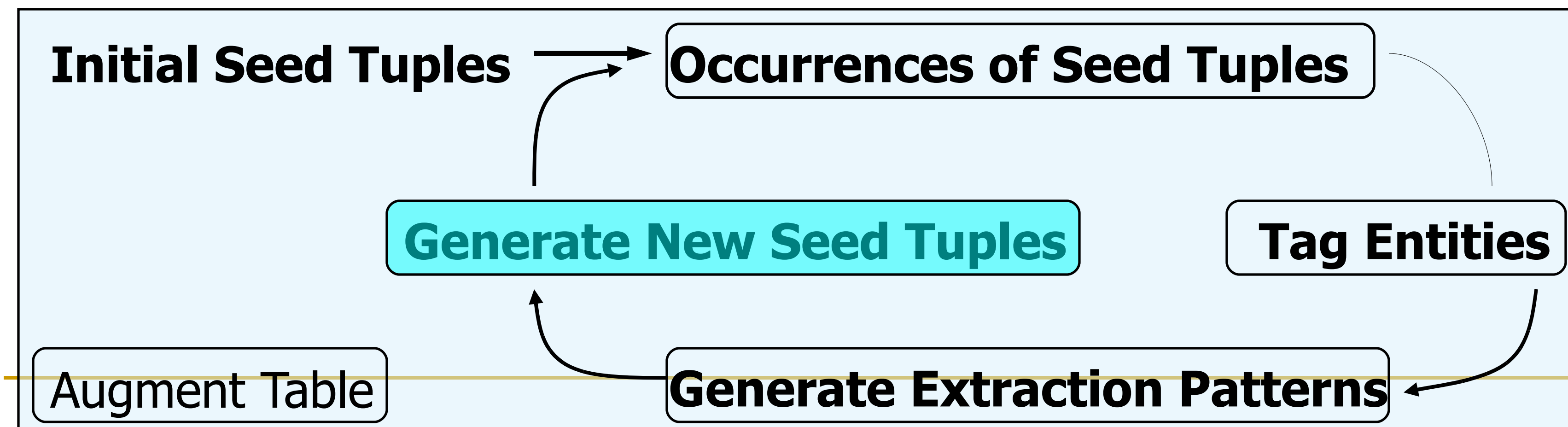
- Estimation of Probability (Correct (Tuple))
- A tuple will have high confidence if generated by multiple high-confidence patterns (P_i).

Similar to Yangarber et al (cf. Session number 6).

Snowball: Filtering Seed Tuples

Generate new seed tuples:

ORGANIZATION	*LOCATION*	*CONF*
AG EDWARDS	ST LUIS	0.93
AIR CANADA	MONTREAL	0.89
7TH LEVEL	RICHARDSON	0.88
3COM CORP	SANTA CLARA	0.8
3DO	REDWOOD CITY	0.8
3M	MINNEAPOLIS	0.8
MACWORLD	SAN FRANCISCO	0.7
157TH STREET	MANHATTAN	0.52
15TH CENTURY EUROPE	NAPOLEON	0.3
15TH PARTY CONGRESS	CHINA	0.3
MAD	SMITH	0.3



Extracting Relations from Text Collections

- Related Work
- The *Snowball* System:
  - Pattern representation and generation
  - Tuple generation
  - Automatic pattern and tuple evaluation
- Evaluation Metrics
- Experimental Results

Task Evaluation Methodology

- Data: Large collection, extracted tables contain many tuples (> 80,000)
- Need scalable methodology:
  - *Ideal* set of tuples
  - Automatic recall/precision estimation
- Estimated precision using sampling

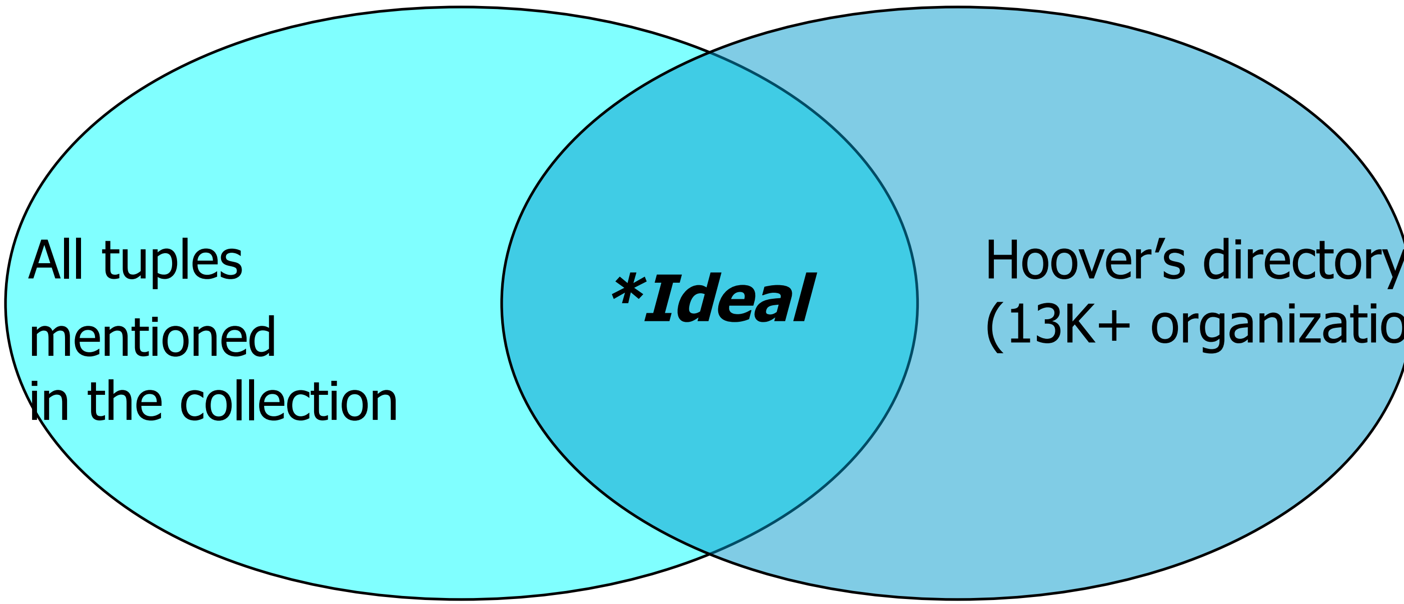
Collections used in Experiments

More than 300,000 real newspaper articles

Collection	*Source*	*Year*
Training	The New York Times	1996
	The Wall Street Journal	1996
	The Los Angeles Times	1996
Test	The New York Times	1995
	The Wall Street Journal	1995
	The Los Angeles Times	1995,’97

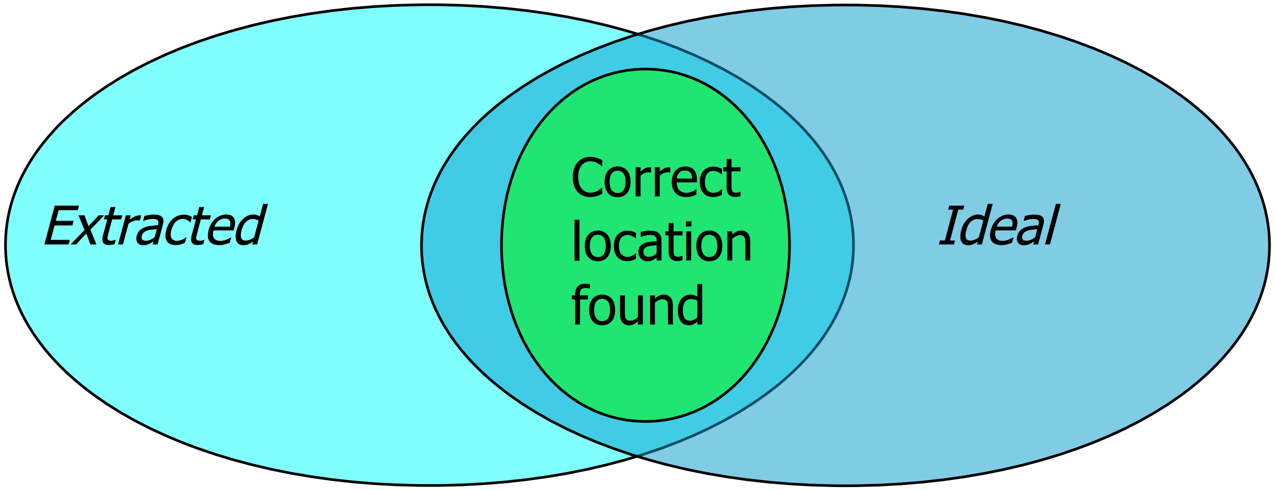
The *Ideal* Metric (1)

Creating the *Ideal* set of tuples



* A perfect, (*ideal*) system would be able to extract all these tuples

The *Ideal* Metric (2)



- Precision:

$$\frac{|\,\text{Correct}\ (\mathit{Extracted} \cap \mathit{Ideal})\,|}{|\,\mathit{Extracted} \cap \mathit{Ideal}\,|}$$

- Recall:

$$\frac{|\,\text{Correct}\ (\mathit{Extracted} \cap \mathit{Ideal})\,|}{|\,\mathit{Ideal}\,|}$$

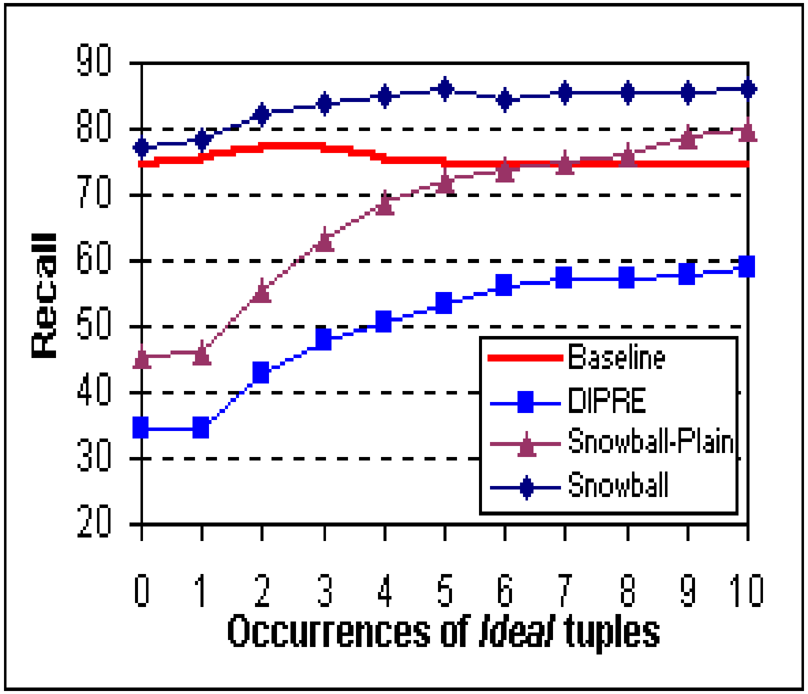
Estimate Precision by Sampling

- Sample extracted table
  - Random samples, each 100 tuples
- Manually check validity of tuples in each sample

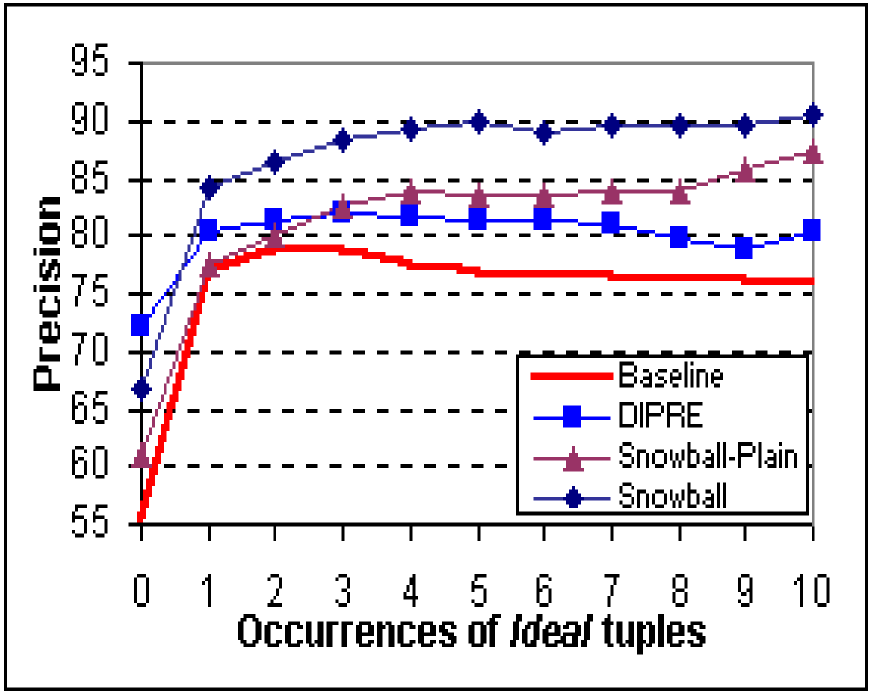
Extracting Relations from Text Collections

- Related Work
- The *Snowball* System:
  - Pattern representation and generation
  - Tuple generation
  - Automatic pattern and tuple validation
- Evaluation Metrics
- Experimental Results

Experimental results: Test Collection



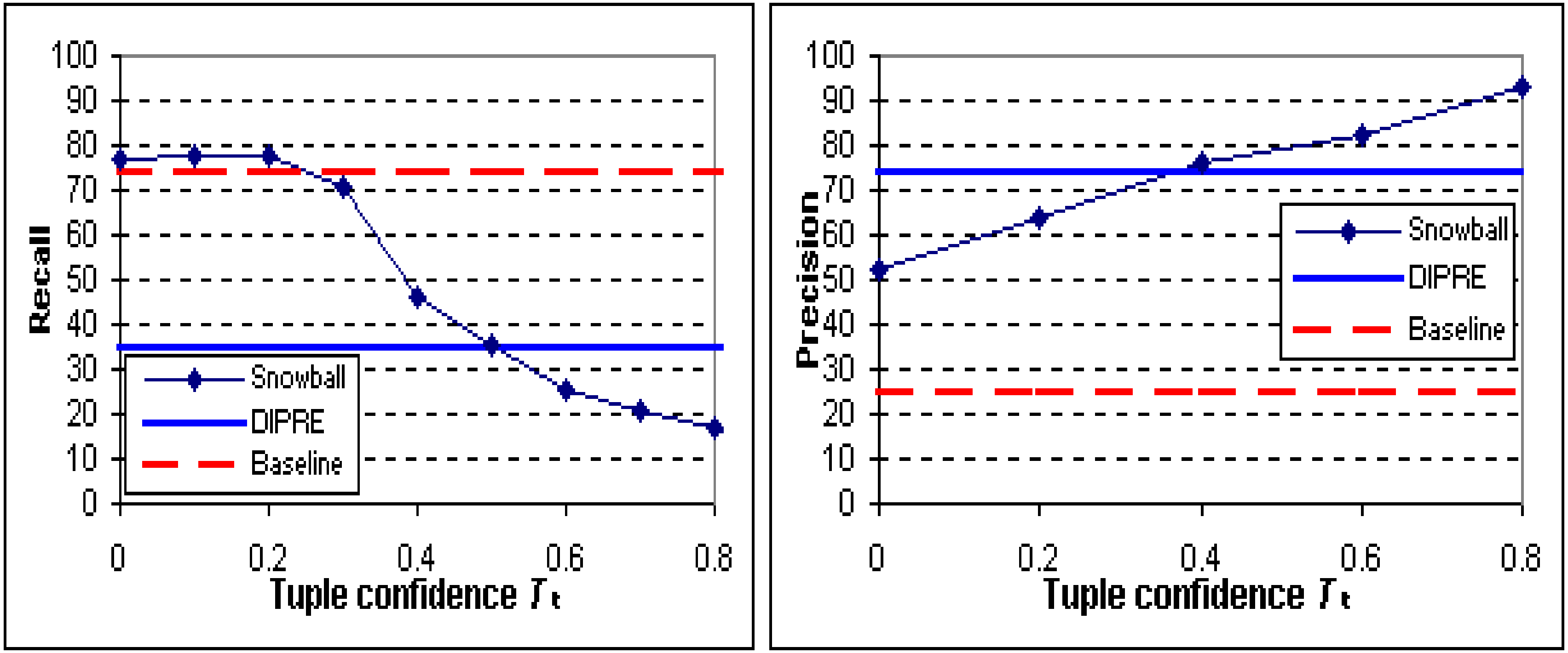
(a)



(b)

Recall (a) and precision (a) using the *Ideal* metric, plotted against the minimal number of occurrences of test tuples in the collection

Experimental results: Sample and Check



(a) (b)

Recall (a) and precision (b) for varying minimum confidence threshold T_t.
NOTE: Recall is estimated using the *Ideal* metric, precision is estimated by **manually checking random samples** of result table.

Conclusions

- *Snowball* system:
  - Requires minimal training (handful of seed tuples)
  - Uses a flexible pattern representation
  - Achieves high recall/precision
    - > 80% of test tuples extracted
- Scalable evaluation methodology