
Toward Semantic Information Integration

Eduard Hovy

Information Sciences Institute
University of Southern California

Why do I need info integration?

- NLP projects:
 - MT—connect multiple languages
 - QA, summarization, IR, etc.—better quality using a large resource of concepts and facts
- Data transfer and database integration projects:
 - EDC—provide access to heterogeneous databases about gasoline for EIA, Census, BLS, CEC
 - EPA-NEISGEI—help EPA agencies transfer data

The problem with automated QA...

- Where do lobsters like to live?
— *on the table*
- Where are zebras most likely found?
— *in the dictionary*
- How many people live in Chile?
— *nine*
- What is an invertebrate?
— *Dukakis*

Webclopedia
(Hovy et al., 01)

...need commonsense semantic, numerical info
...need an ontology

Language technology application limits

- **How to improve QA?**

TREC 99–02: around 65%

- ★ Understand Q and A; match their meanings; know common info

- **How to improve accuracy of IR / web search?**

TREC 98–01: around 40%

- ★ Understand user query; expand query terms by meaning

- **How to achieve conceptual summarization?**

Never been done yet, at non-toy level

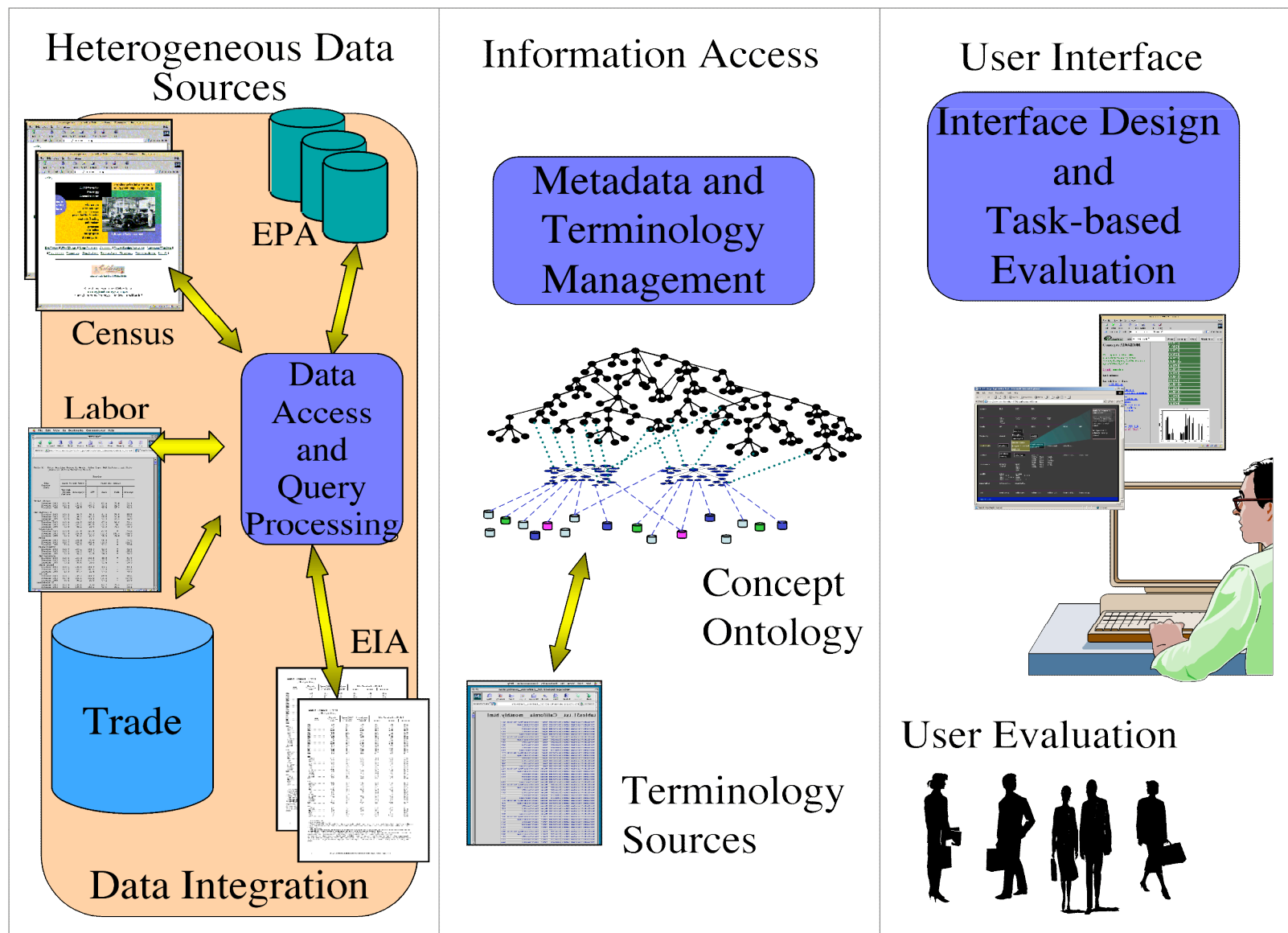
- ★ Interpret topic, fuse concepts according to meaning; re-generate

- **How to improve MT quality?**

MTEval 94: ~70%, depending on what you measure

- ★ Disambiguate word senses to find correct meaning

Example application: EDC project



Two principal paths for info integration

1. Using a central structure as ‘interlingua’

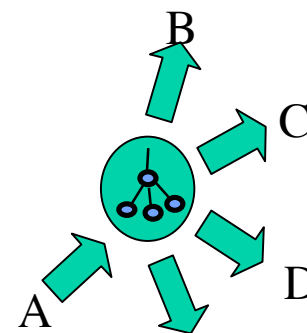
- Language (MT): Interlingua systems
- Database integration: SIMS model

– Problems:

- Creating the central structure (coverage, consistency, updating)
- Linking sources and targets to it (automatically?)

– Benefits:

- Linear ($2N$) in number of sources/targets



2. Creating individual source-to-target mappings

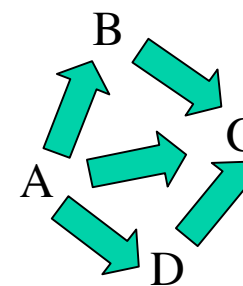
- Language (MT): Transfer systems

– Problems:

- Creating and updating the mappings (automatically?)
- N^2 in number of sources/targets

– Benefits:

- Doesn't require general one-size-fits-all model/structure



Credo and methodology

Ontologies (and even concepts) are too complex to build all in one step...

...so build them bit by bit, testing each new (kind of) addition empirically...

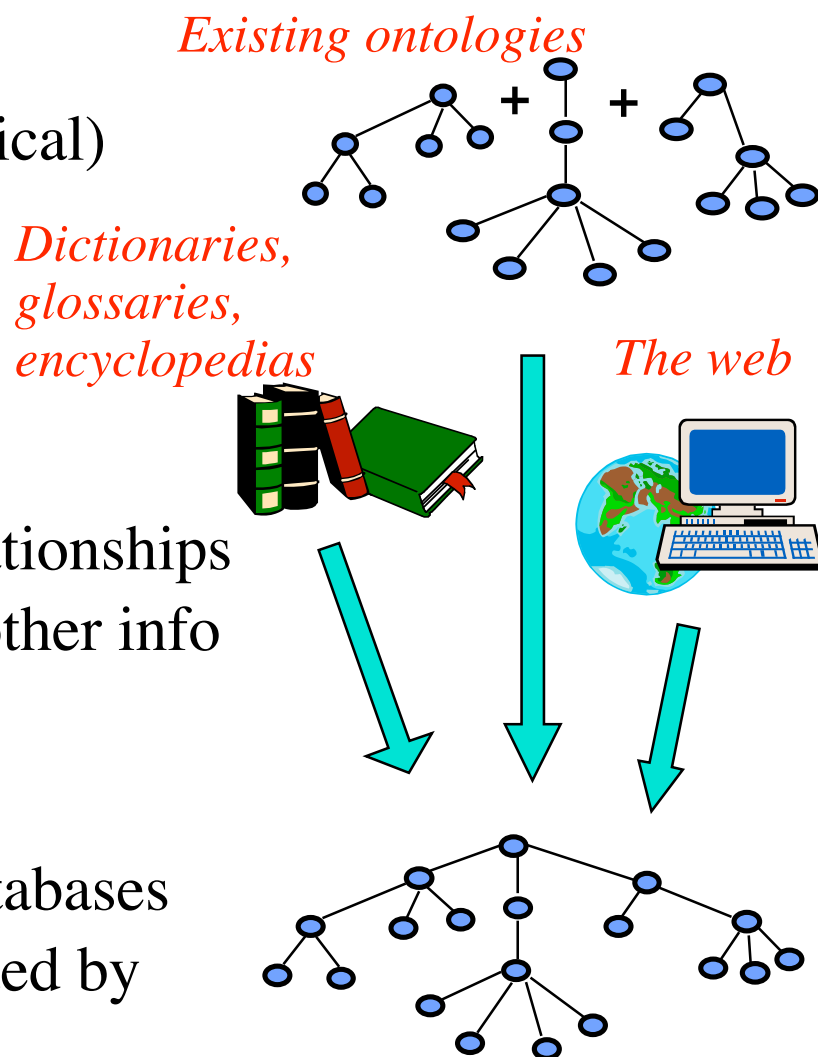
...and develop appropriate learning techniques for each bit, so you can automate the process...

...so next time (since there's no ultimate truth) you can build a new one more quickly.

Note: I am interested in content, not formalism.

Plan: stepwise accretion of knowledge

- Initial framework:
 - Start with existing (terminological) ontologies
 - Weave them together
- Build concepts:
 - Define/extract concept ‘cores’
 - Extract/learn inter-concept relationships
 - Extract/learn definitional and other info
- Build (large) instance base:
 - Extract instance ‘cores’
 - Link into ontology; store in databases
 - Extract more information, guided by parent concept



Talk overview

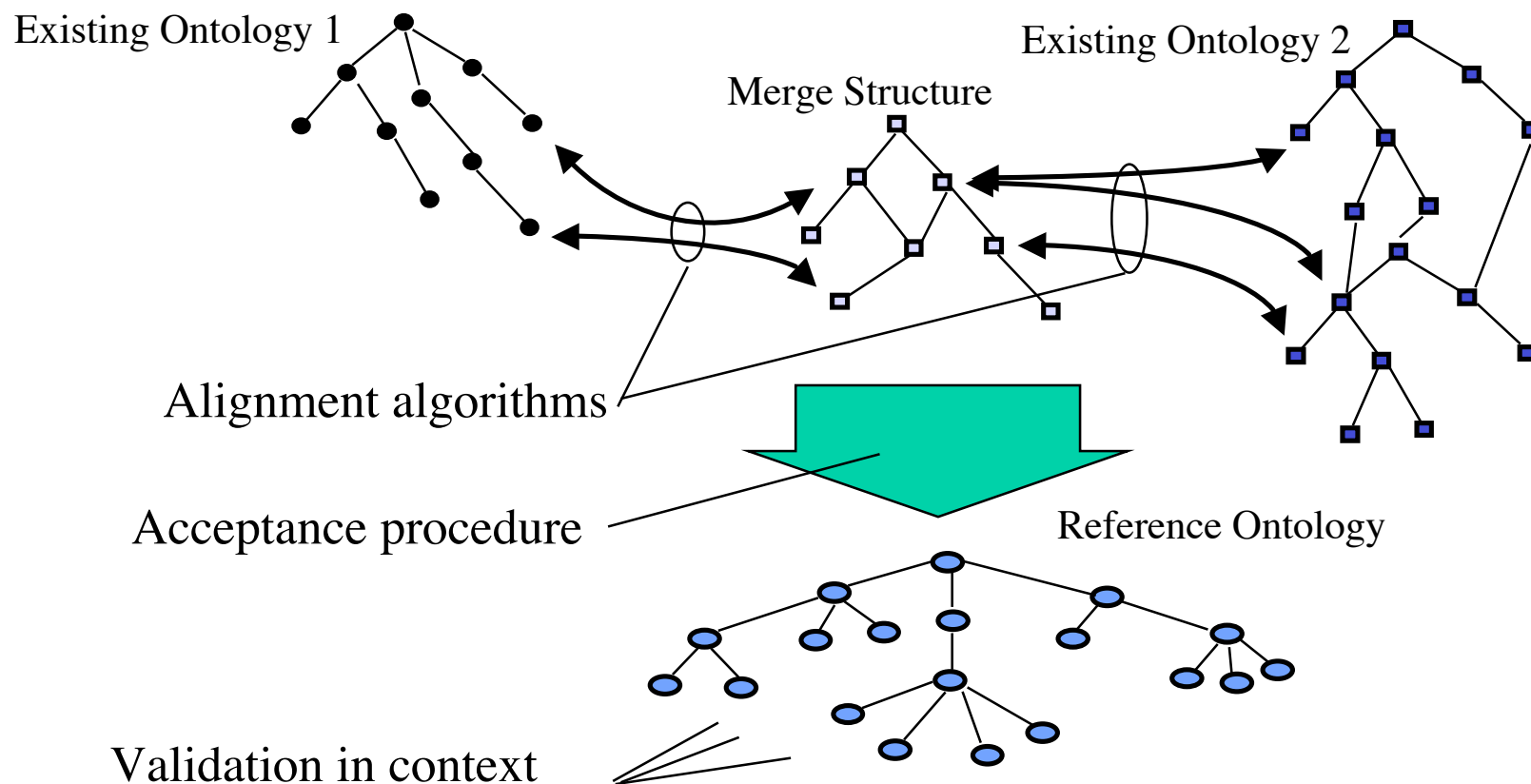
1. Introduction
2. The Interlingua route: toward a merged ontology
 - Concept alignment experiments
 - Problems and Shishkebobs
3. The Transfer route: toward learning individual mappings
4. Conclusion

The Interlingua route: Toward a merged Ontology

General alignment and merging problem

Goal: **find attachment point(s) in ontology** for node/term from somewhere else (ontology, website, metadata schema, etc.)

It's hard to do manually; very hard to do automatically — system needs to understand semantics of entities to be aligned



Three ingredients for automation

1. Basic source material to be integrated
 - Two ontologies / models / sets of terms
2. Additional information that may help
 - Text descriptions, databases, etc.
3. Alignment/merging/integration algorithms
 - Cyclic process:
 - Automated alignment suggestion
 - Manual accept/reject decisions
 - Automated validation wrt master material



Alignment procedure

- Procedure:
 - 1. For the new term/concept, **extract and format**: name, definition, associated text, local taxonomy cluster, etc.
 - 2. apply **alignment suggestion heuristics** (NAME, DEFINITION, HIERARCHY, DISPERSAL, etc. match) against big ontology, to get proposed attachment points with strengths (Hovy 98) — test with numerous parameter combinations, see <http://edc.isi.edu/alignment/> (Hovy et al. 01)
 - 3. automatically **combine** proposals (Fleischman et al. 03)
 - 4. apply **verification** checks
 - 5. **accept or reject** proposals manually
- Process developed in early 1990s: (Agirre et al. 94; Knight & Luk 94; Okumura & Hovy 96; Hovy 98; Hovy et al. 01)
- Not stunningly accurate, but can speed up manual alignment markedly

Alignment suggestion heuristics: 3 types

1. Text matches

- concept names: reward cognates; reward delimiter (start, end, hyphen) confluence... (Hovy 98; Hovy et al. 01)
- concept (text) definitions: string matches, w or w/o demorphing, stop words, ngram lang models... (Knight & Luk 94; Dalianis & Hovy 98; Hovy 98; Hovy et al. 01)

2. Hierarchy matches

- shared superconcepts, to filter ambiguity (Knight & Luk 94)
- semantic distance (Resnik 93; Agirre et al. 94; Hovy 98)
- cluster dispersal (Hovy et al. 01)

3. Data item and form matches

- inter-concept relations (Ageno et al. 94; Rigau & Agirre 95)
- slot-filler restrictions (Okumura & Hovy 94)
- entity slot frameset

1996–1998 alignment work

(Hovy 98)

- **Ontologies:**

- SENSUS Upper Model (350) (Bateman et al. 89)
 - CYC top region (2400) (Lenat; Lehmann 96)
 - MIKROKOSMOS (4790 concepts) (Mahesh 96)
 - SENSUS top region (6768) (Knight & Luk 94)
- } 1996
- } 1997

- **Link suggestion heuristics:**

NAME: $score = (match-length(N1,N2))^2 + 20.if-equal + 10.if-same-end$

DEF: $score = (shared(D1,D2) / min(D1,D2)) * shared(D1,D2)$

TAX: $score = 1 / link-distance$ (for superc and subc links only)

- **Combination function:**

$$score = \sqrt{namescore} * defscore * (10 * taxscore)$$

Cross-Ontology alignment results

- SENSUS top (6700) ↔ MIKROKOSMOS (4900)
- 3 algorithms (NAME, DEF, TAX); 5 cycles
- **Recall** (*how many good links were missed?*):
difficult to count! ... 32.4 mill pairs
- **Precision** (*how many suggested links were correct?*):

- 0.252 (strict)
- 0.517 (lenient)

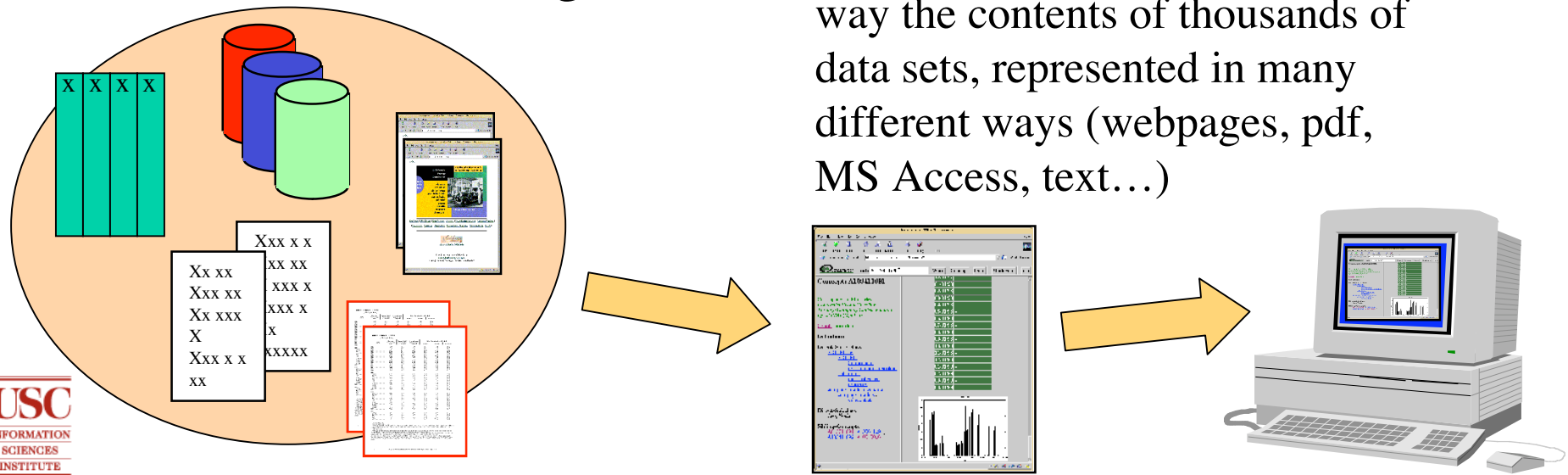
cutoff	1.4	10	7.8	12	15
new heuristic	NAME/DEF/TAX	TAX	TAX	TAX	TAX
total	187	151	170	218	241
correct	73	11	18	36	106
near	51	92	51	60	2
wrong	63	48	101	122	39

- **After 5 runs:**

- 883 suggestions (= 13% of SENSUS candidates)
- *correct*: 244 (= 3.6%)
- *near miss*: 256 (= 3.8%)
- *wrong*: 383 (= 5.6%)

1999–2002: EDC database access project

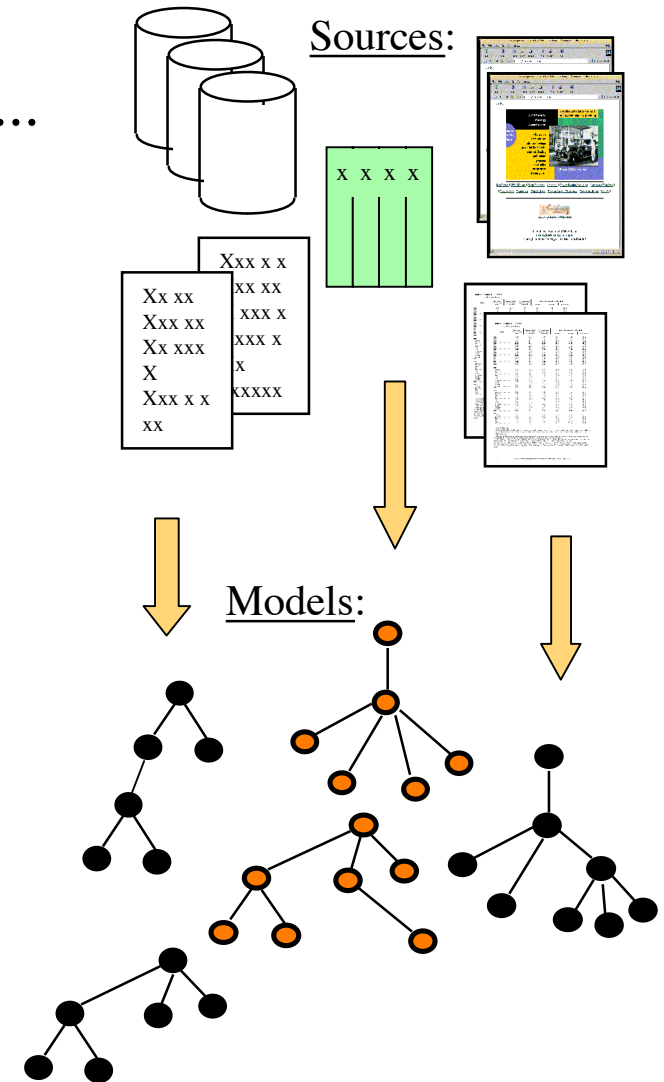
- EDC members
 - Information Sciences Institute, USC
 - Dept of CS, Columbia University
- Government partners
 - Energy Information Admin. (EIA)
 - Bureau of Labor Statistics (BLS)
 - Census Bureau
- Research challenge
 - Make accessible in standardized way the contents of thousands of data sets, represented in many different ways (webpages, pdf, MS Access, text...)



The idea behind SIMS

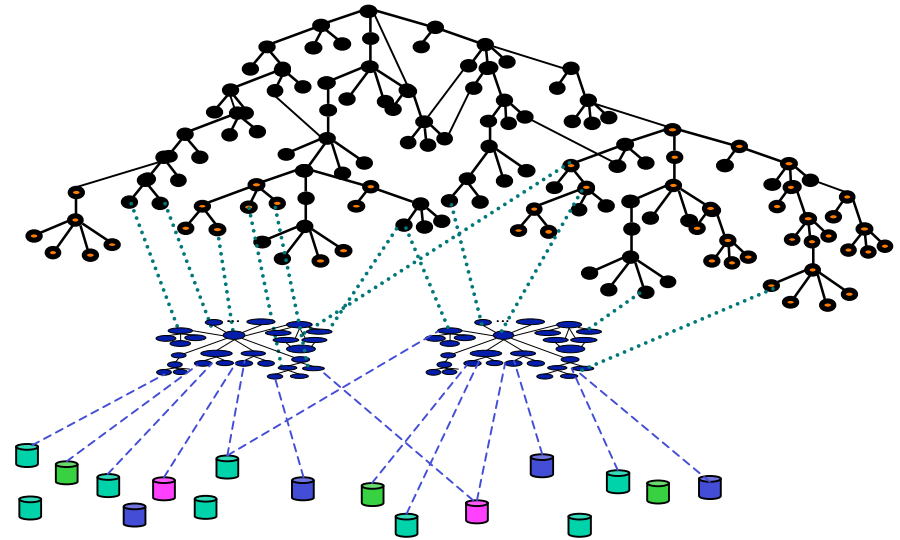
(This work with Jose Luis Ambite and Andrew Philpot)

- There are many types of data sources: databases, pdf files, text files, html files...
- The user doesn't want to know this!
- Solution:
 1. 'Wrap' each source in software that handles access to its data
 2. Record the types of info in each source in a 'source model'
 3. Arrange all source models together in the same space—the Domain Model
 4. Use a data access planner (SIMS) to transform a user's request for data into a set of individual access queries that extracts the right data from the appropriate sources
- Current dataset: 53689 different time series, spread across 171 different sources



A super domain model: the Ontology

- Given many domain models—how can you integrate them consistently, without overlap or redundancy?
- Solution:
 1. Use a large general-purpose concept network to provide the background—the SENSUS Ontology (90,000 nodes)
 2. Embed the Domain Models inside it, linking to the appropriate concepts
 3. This makes sure that different sources with the same kind of information can share the appropriate Domain Model nodes
 4. The Ontology provides a (formal) modeling language



Linking DMs to the Ontology

Match concepts against all of SENSUS
(1: domain model; 2: 3% subset of SENSUS)

1. Name Match

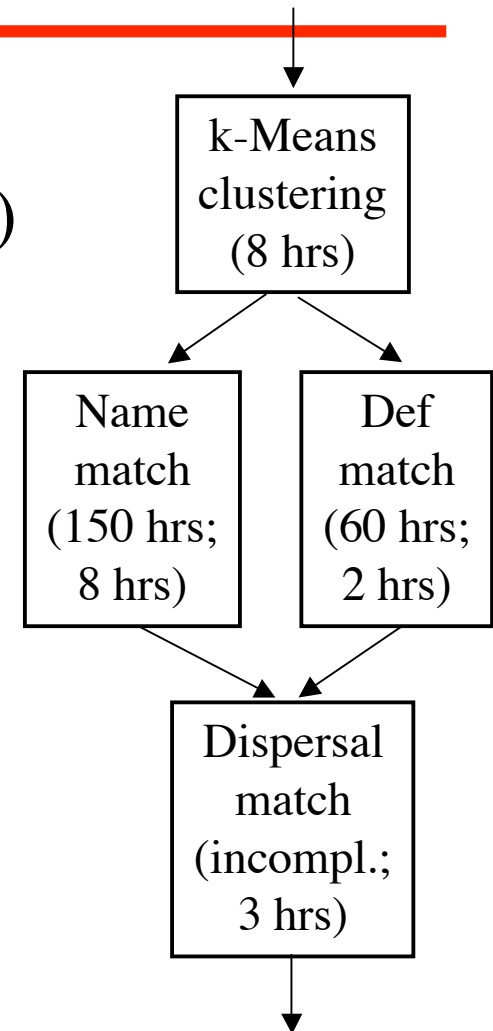
- Variations: substrings; reward for contiguous endings; stemming; variant word forms; case; etc.
- Implemented string match inspired by molecular biology DNA matching

2. Definition Match

- Variations: stemming; stop words; words or letter trigrams; word weights (*itf*, *tf.idf*), etc.
- Used IR vector space; cosine measure

3. Dispersal Match

- Def: Given a cluster of concepts, match them all and then select one match for each concept to form tightest cluster in SENSUS
- Variations: greedy search; number of candidates; weightings



Result example

(Hovy et al. 01)

- Tested with many parameter combinations, see <http://edc.isi.edu/alignment/>
- Example: Concept **Premium Gasoline**
 - **Manual**: 1 alignment: "*product.gasoline.premium*" (confidence 1.0)
 - **CharBag** 5th among 15 matches. (conf 0.65; range [.5806–.7059])
□ *high char overlap*
 - **TrigramOverlap** (tied for) 1st of 15 matches. (conf .5909; range [.3556–.5909])
□ *best score*
 - **StopStemToken** (tied for) 1st of 15 matches. (conf 0.5; range [.2326–0.5])
 - **NoncontigSub** (tied for) 6th of 15 matches. (conf 0.45; range [.4118–.5625])
 - **ContigSub3** (tied for) 5th of 15 matches. (conf .3542; range [.2812–.5937])
 - **ContigSub2** (tied for) 5th of 15 matches. (conf .3542; range [.2812–.5937])
 - **ContigSub1** (tied for) 5th of 15 matches. (conf .3542; range [.2812–.5937])
 - **FiltContigSub4** (tied for) 5th of 15 matches. (conf .3542; range [.2812–.5937])
 - **FiltContigSub3** (tied for) 5th of 15 matches. (conf .3542; range [.2812–.5937])
 - **FiltContigSub2** (tied for) 5th of 15 matches. (conf .3542; range [.2812–.5937])
 - **EditDist** (tied for) 6th of 15 matches. (conf .3333; range [.2778–0.5])
 - **FiltContigSub1** not found □ *bug*

Improving alignment by enriching content by adding definitional material

(This work with Uli Germann and Andrew Philpot (ISI)
and Judith Klavans and colleagues (Columbia))

Finding additional info to help

Example: EIA glossary page

Energy Glossary, Energy Definitions, Terms - Netscape

http://www.eia.doe.gov/glossary/glossary_main_page.htm

Search

Home | Energy Glossary |

Search EIA: by FIRSTGOV GO

Energy Glossary

This glossary provides energy terms and definitions as used in EIA reports, on EIA survey forms, and by the energy industry.

A	B	C	D	E	F	G	H
I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X-Z

Contact:
Questions about definitions: Renee Miller, renee.miller@eia.doe.gov

Other Topics
[EIA Publishing Style Guide](#)
[Energy Quiz](#)
[How Are We Doing?](#)
Please click here and give us your feedback.

What's New
Energy A-Z
Publications
[Sign Up for E-mail Updates](#)
Contact Experts
Privacy/Security
Featured Publications:
Recent Monthly Statistics
Annual Data from 1949
Projections to 2020

Home | [Petroleum](#) | [Alternative Fuels](#)

Page last modified: Tuesday, 12/12/2000
URL: http://www.eia.doe.gov/glossary/glossary_a.htm
U.S. Energy Information Administration

Glossary - Energy Glossary - A page - Netscape

http://www.eia.doe.gov/glossary/glossary_a.htm

Glossary - Energy Glossary - A p...

other water uses.

Affiliate: An entity which is directly or indirectly owned, operated, or controlled by another entity. See **firm**.

Afforestation: Planting of new forests on lands that have not been recently forested.

Aftermarket converted vehicle: A standard conventionally fueled, factory-produced vehicle to which equipment has been added that enables the vehicle to operate on alternative fuel.

Aftermarket vehicle converter: An organization or individual that modifies OEM vehicles after first use or sale to operate on a different fuel (or fuels).

Agglomerating character: Agglomeration describes the caking properties of coal. Agglomerating character is determined by examination and testing of the residue when a small powdered sample is heated to 950 degrees Centigrade under specific conditions. If the sample is "agglomerating," the residue will be coherent, show swelling or cell structure, and be capable of supporting a 500-gram weight without pulverizing.

Document: Done (0.27 secs)

Getting ontology defs from online terms

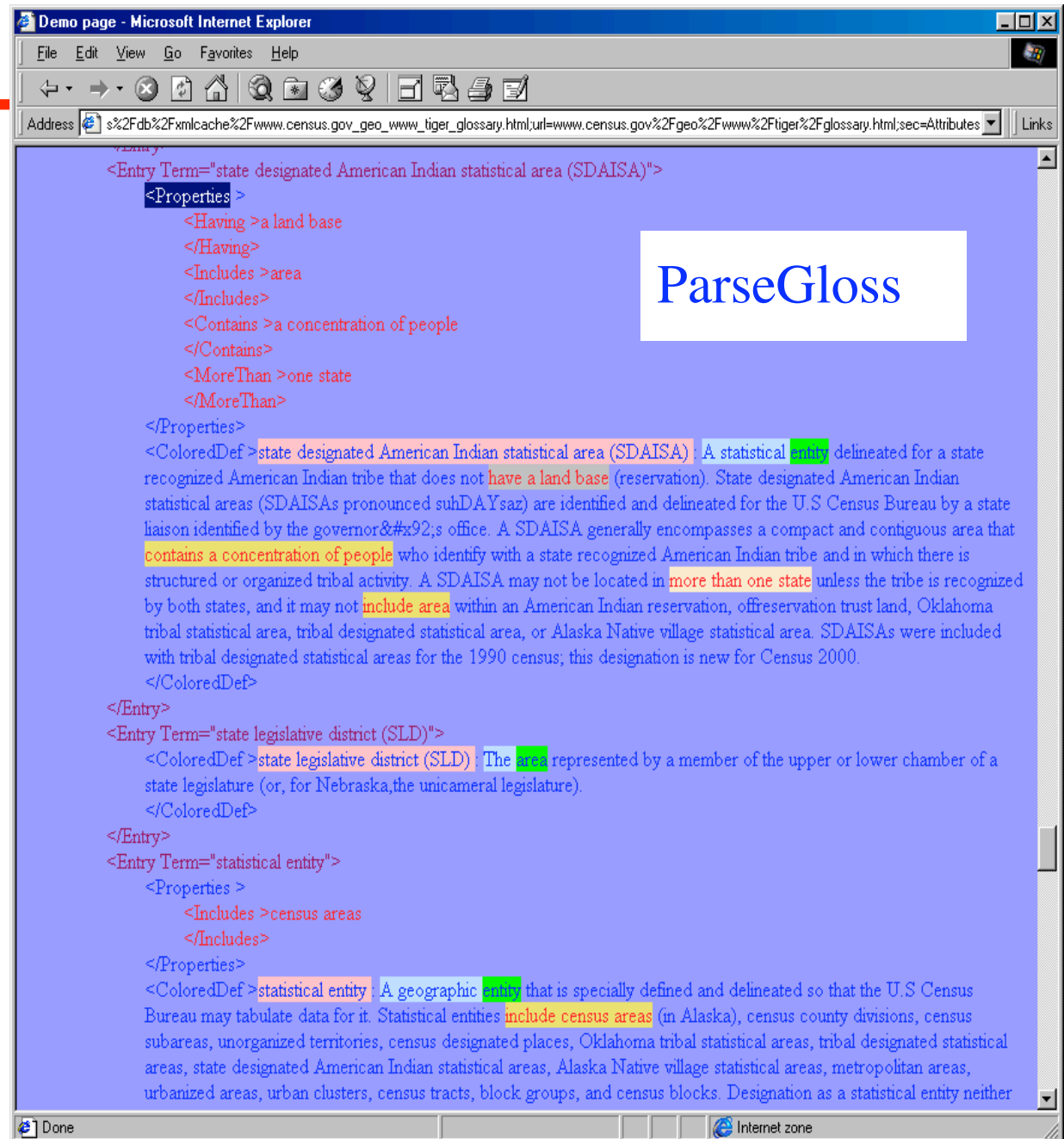
(Past work, joint with Klavans et al. at Columbia U)

Three steps:

1. Collect all glossary pages for a site
 - GetGloss (Klavans et al.)
 - Google “glossary site: eia.doe.gov” (Germann, ISI)
2. Extract the terms and their definitional and other information
 - ParseGloss (Klavans et al.)
 - Charniak parser (Germann, ISI; Charniak 99)
3. Locate their appropriate positions in the domain model or ontology and embed them there
 - various methods (Hovy et al. dg.o 2002)

Columbia work:

1. GetGloss: given a URL, find all its glossary files
 - Glossary identification rules consider format tags, etc.
 - Treat as categorization problem: glossary or not?
2. ParseGloss: given a set of NL glossary definitions, extract and format the important information
 - Identify term, def, head noun, etc. using POS patterns
 - Reformat into KR rep



ISI procedure

1. Given a URL, collect all glossary files
2. Clean up html
3. Extract all terms and definitions
4. Parse them
5. Identify head terms and other major properties, using parse tree
6. Identify likely superconcepts from Omega
7. Reformulate into Omega formalism
8. Define within Omega
9. Redisplay or use

Steps 1–3: Setup

Apply Google: “glossary site: <http://www.eia.doe.gov>”

Find glossaries
Download and clean
entries
Convert to XML

Example HTML fragment:

```
<p>
<b><a name="after_market_conv_veh">Aftermarket converted vehicle</a>:</b>
  A standard conventionally fueled, factory-produced vehicle to which
  equipment has been added that enables the vehicle to operate on
  alternative fuel.
<p>
```

Resultant XML:

```
<ENTRY>
<DEFINED>Aftermarket converted vehicle</DEFINED>
<DEFINITION>
  A standard conventionally fueled, factory-produced vehicle to which
  equipment has been added that enables the vehicle to operate on
  alternative fuel.
</DEFINITION>
</ENTRY>
```

Head term: “standard”? “vehicle”?

Step 4,5: Parse tree

```
<ENTRY>
<TERM>Aftermarket converted vehicle</TERM>
<DEF1>An Aftermarket converted vehicle is a standard conventionally fueled , factory-
produced vehicle to which equipment has been added that enables the vehicle to
operate on alternative fuel . </DEF1>
<PARSED>
(S1 (S (NP (DT An) (NNP Aftermarket) (JJ converted) (NN vehicle))
      (VP (AUX is)
           (NP (NP (DT a)
                  (JJ standard)
                  (ADJP (RB conventionally) (VBN fueled))
                  (, ,)
                  (JJ factory-produced)
                  (NN vehicle))
           (SBAR (WHPP (TO to) (WHNP (WDT which) (NN equipment))))
           (S (VP (AUX has)
                  (VP (AUX been)
                      (VP (VBN added)
                          (SBAR (IN that)
                              (S (VP (VBZ enables)
                                    (S (NP (DT the) (NN vehicle))
                                        (VP (TO to)
                                            (VP (VB operate)
                                                (PP (IN on) (NP (NN alternative) (NN fuel))))))))))))))
           (. .)))
</PARSED>
</ENTRY>
```

Use Charniak parser
Restate def as sentence
Identify head noun

Steps 6–8: Reformulation

```
;; =====  
;; Aftermarket converted vehicle  
;; =====  
;; TERM HEAD = "vehicle"  
;; DEFINITION HEAD = "vehicle"
```

Identify likely superconcepts
Formalize into Omega notation
Introduce into Omega

```
(DEFCONCEPT EIA@::|Aftermarket converted vehicle|  
:direct-superclass (o@::|vehicle>sled|)  
:direct-superclass (o@::|vehicle|)  
:lex ((EIA@EN::|Aftermarket converted vehicle| NOUN 1))  
:properties (|factory-produced| |conventionally fueled| |standard|)  
:definition "A standard conventionally fueled, factory-produced  
vehicle to which equipment has been added that enables the  
vehicle to operate on alternative fuel."  
)
```

Results

- Columbia: It's easy to find glossaries:
 - GetGloss test data: 2608 files collected under www.census.gov
 - Used Rainbow package categorization algorithms (McCallum 96)

Algorithm	Recall	Precision	F1	Recall	Precision	F1	Accuracy
GetGloss	1.000	0.517	0.682	0.994	1.000	0.997	0.994
Naïve Bayes	1.000	0.162	0.279	0.957	1.000	0.978	0.958
SVM	0.944	0.895	0.919	0.999	0.999	0.999	0.999
KNN	0.000	0.000	0.000	1.000	0.992	0.996	0.992
ProbIndex	0.222	0.400	0.286	0.997	0.994	0.995	0.991

- Columbia: ParseGloss: it's harder to parse and place entries
 - EIA glossary: 2182 entries; postulated superconcepts for 1447 (66.3%)
 - 284 of these superconcept terms already defined in Omega (= 17.2%)
 - 102 of these matched correctly
 - precision 35.9% (on concepts already in Omega), 4.6% (on all entries)
- ISI: Charniak parser: get many more correct head terms
 - For website with appropriate glossaries, get placement >80%
 - Problems: some defs badly written, and too many superconcept options

Improving alignment by clustering entities and aligning clusters

(This work with Andrew Philpot)

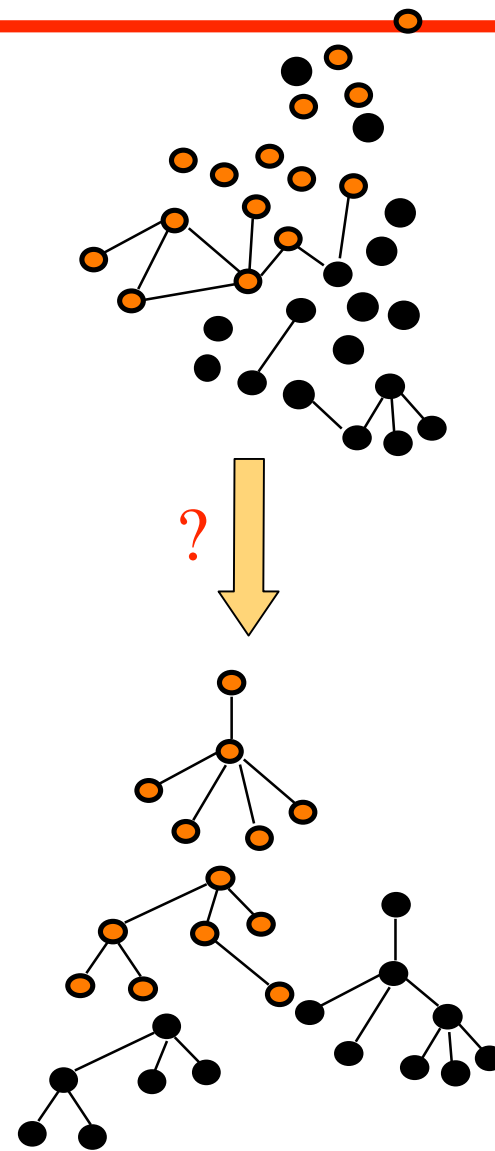
Creating domain models

Manually: extract from database metadata

Automatically, given terms parsed from glossaries: cluster concepts in word vector space, using definitions' text:

- tried ISI's clustering package (CLINK, SLINK, Ward's Method, etc.)
- implemented new version of k-Means (Euclidean and spherical distance measures)
- tried variations (stemming, etc.)

☹ Results: not great



Dispersal match: pilot experiment

To get best parameterization, try with ‘perfect’ input:
clustering + dispersal match on 3 clusters taken from SENSUS

k-Means found
3 clusters...

...but they are
somewhat
‘smeared’...

though this one
is ok

	Tools	Beverages	Furniture	Totals
Global % distribution				
Cluster 1	5.90%	5.76%	6.91%	18.56%
Cluster 2	35.83%	4.03%	17.84%	57.70%
Cluster 3	0.72%	22.73%	0.29%	23.74%
Totals	42.45%	32.52%	25.04%	100.00%
Row distribution				
Cluster 1	13.90%	17.70%	27.59%	
Cluster 2	84.41%	12.39%	71.26%	
Cluster 3	1.69%	69.91%	1.15%	
Totals	100.00%	100.00%	100.00%	
Column distribution				
Cluster 1	31.78%	31.01%	37.21%	100.00%
Cluster 2	62.09%	6.98%	30.92%	100.00%
Cluster 3	3.03%	95.76%	1.21%	100.00%

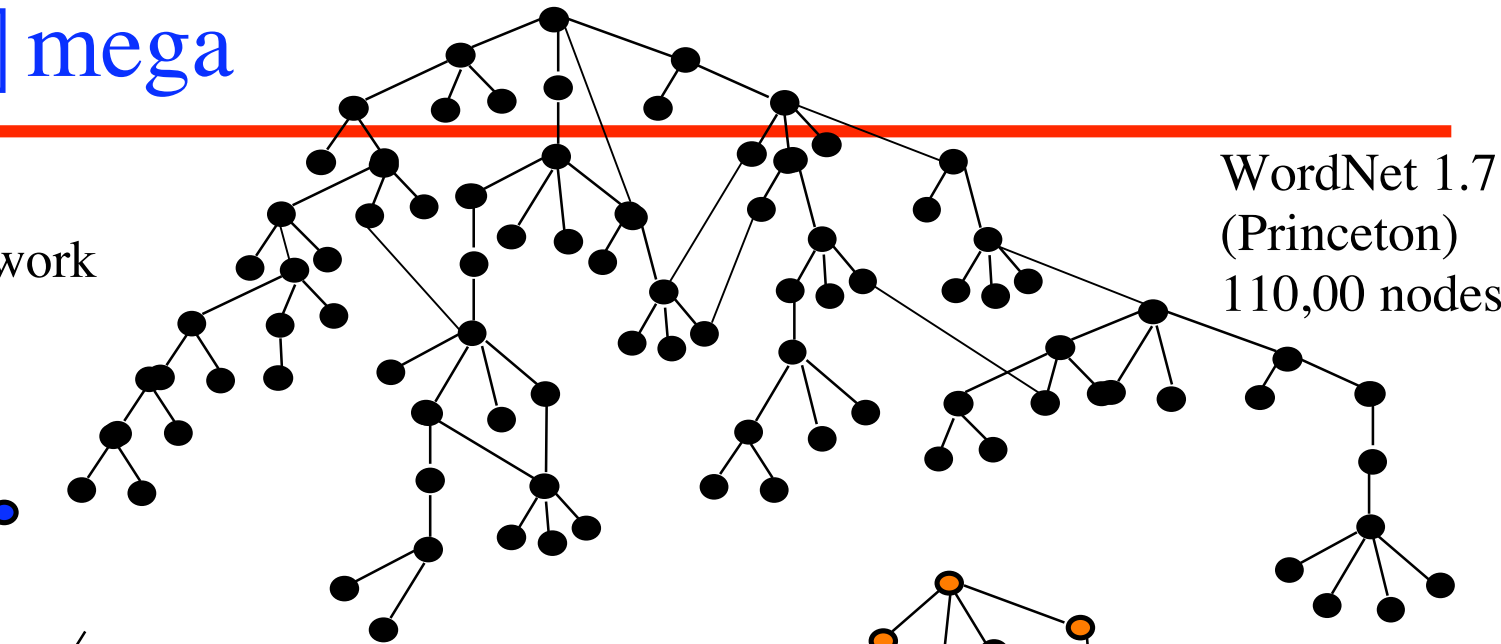
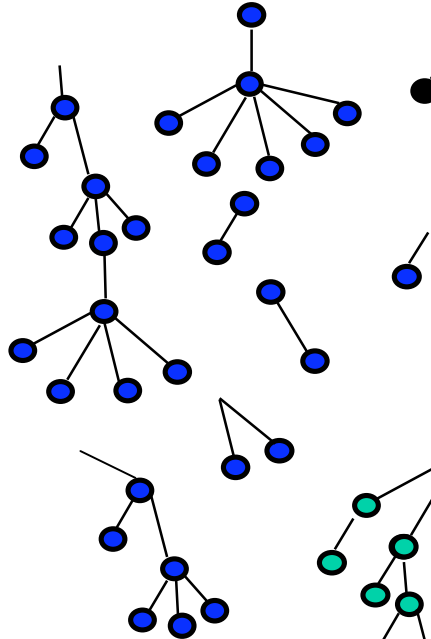
Conclusion:
More work needed

When the alignment is done: problems in semantics

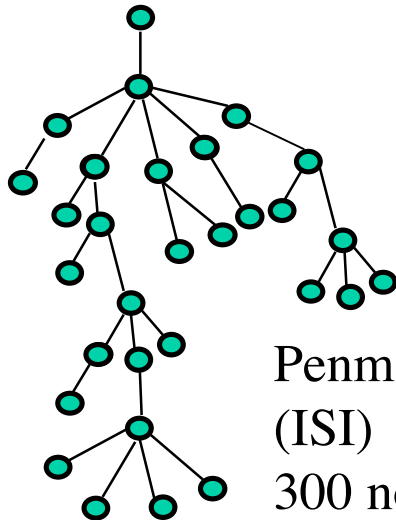
(This work with Michel Fleischman,
Andrew Philpot, and Jerry Hobbs)

2003: □ mega

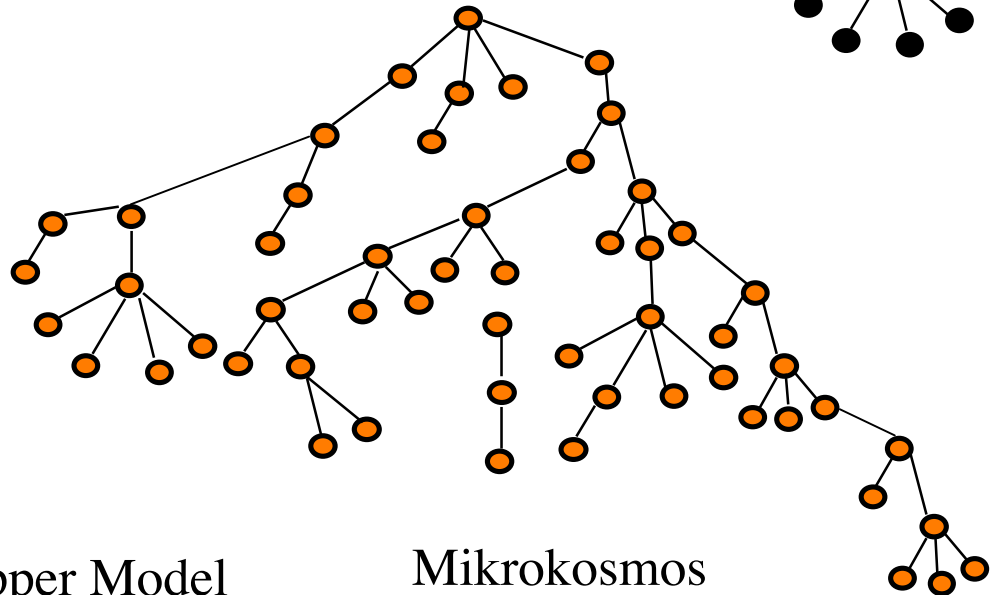
Our own new work
(ISI)
400 nodes



WordNet 1.7
(Princeton)
110,000 nodes



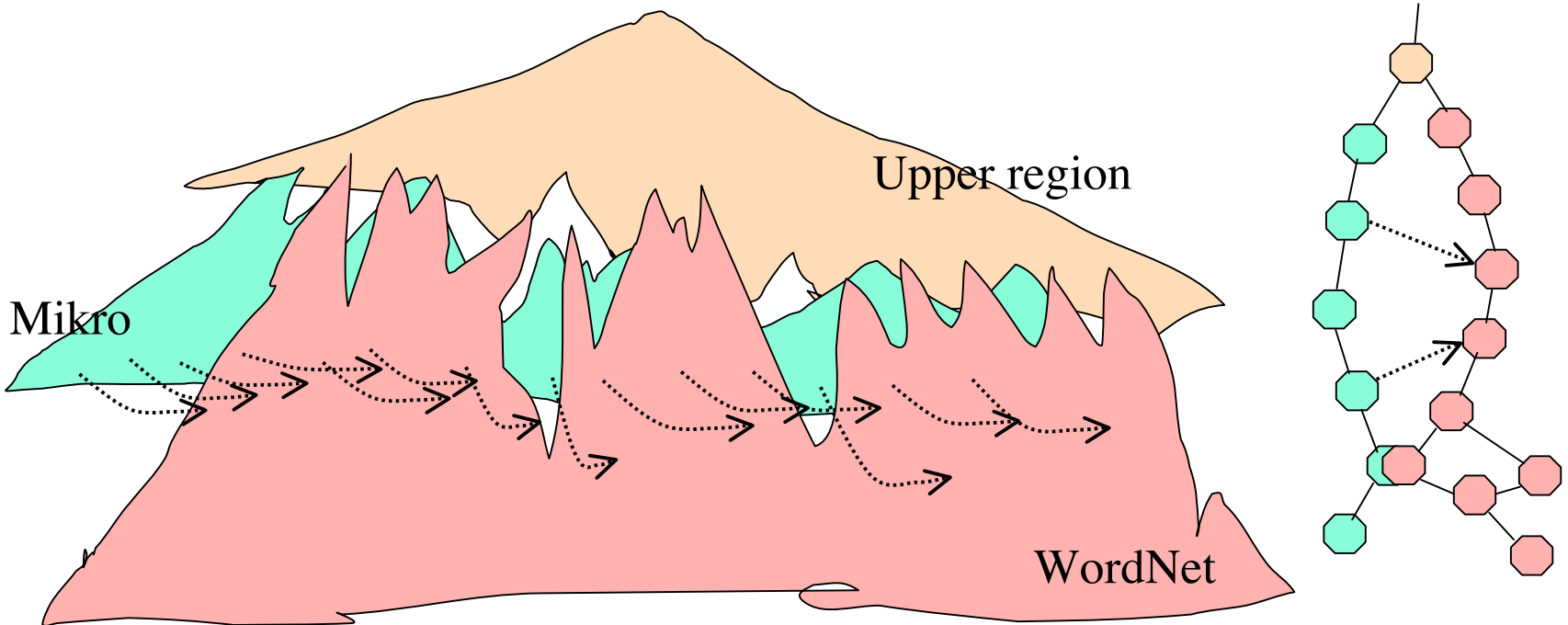
Penman Upper Model
(ISI)
300 nodes



Mikrokosmos
(New Mexico State U)
6,000 nodes

Alignment for \square mega

- Created Upper Region (400 nodes) manually
- Manually snipped tops off Mikro and WordNet, then attached them to fringe of Upper Region
- Automatically aligned bottom fringe of Mikro into WordNet
- Automatically aligned sides of bubbles



Netscape
File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print

Bookmarks Netsite: <http://www.isi.edu/sims/philpot/ontology/fr>

Instant Message WebMail Radio People Yellow Pages

Dynamically generated sewing up page

For merge point [U@:|VERTEBRATE|](#), mikro leaf [M25@:|BOTTLENOSE|](#)

U@: VERTEBRATE 	
U@: MAMMAL 	S@: craniate
M25@: SEA-MAMMAL 	S@: mammal
M25@: DOLPHIN 	S@: placental>bat
	S@: aquatic mammal
	S@: cetacean
	S@: whale
	S@: toothed whale
	S@: dolphin>orca
M25@: BOTTLENOSE 	S@: bottlenose dolphin

For 24 pairings, there were 6 possible matches:

- [\[ACCEPT\]M25@:|DOLPHIN|](#) <-> [S@:|dolphin>orca|](#) : 0.9756098
- [\[ACCEPT\]M25@:|SEA-MAMMAL|](#) <-> [S@:|cetacean|](#) : 0.8333333
- [\[ACCEPT\]M25@:|SEA-MAMMAL|](#) <-> [S@:|placental>bat|](#) : 0.8333333
- [\[ACCEPT\]U@:|MAMMAL|](#) <-> [S@:|cetacean|](#) : 0.8333333
- [\[ACCEPT\]U@:|MAMMAL|](#) <-> [S@:|placental>bat|](#) : 0.8333333
- [\[ACCEPT\]U@:|MAMMAL|](#) <-> [S@:|mammal|](#) : 0.9756098

The maximal consistent subset size is 3. There were 3 subsets found of that size:

- [\[ACCEPT\]](#)
 - [M25@:|DOLPHIN|](#) <-> [S@:|dolphin>orca|](#) : 0.9756098
 - [M25@:|SEA-MAMMAL|](#) <-> [S@:|placental>bat|](#) : 0.8333333
 - [U@:|MAMMAL|](#) <-> [S@:|mammal|](#) : 0.9756098
- [\[ACCEPT\]](#)
 - [M25@:|DOLPHIN|](#) <-> [S@:|dolphin>orca|](#) : 0.9756098
 - [M25@:|SEA-MAMMAL|](#) <-> [S@:|cetacean|](#) : 0.8333333

Document: Done

Netscape
File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Netsite: <http://www.isi.edu/sims/philpot/ontology/frame/frame-tag-2522.html> What's Related

Instant Message WebMail Radio People Yellow Pages Download Calendar Channels

Dynamically generated sewing up page

For merge point [U@:|TANGIBLE-OBJECT|](#) (orig merge point was [U@:|NATURAL-OBJECT|](#)), mikro leaf [M25@:|AMBER|](#) / sensor interior concept [S@:|amber<<resin|](#)

U@: TANGIBLE-OBJECT 	
U@: DECOMPOSABLE-OBJECT 	U@: NONDECOMPOSABLE-OBJECT
U@: TANGIBLE-NONVOLITIONAL-OBJECT 	U@: STATE-NON-SPECIFIC-OBJECT
U@: NONBIOLOGICAL-OBJECT 	S@: compound
U@: NONVOLITIONAL_NONBIOLOGICAL-OBJECT 	S@: stuff
U@: NATURAL-OBJECT 	S@: organic compound
M25@: EARTH-MATERIAL 	S@: plant material
M25@: MINERAL 	S@: resin
M25@: GEMSTONE 	S@: plant product
	S@: natural resin
M25@: AMBER 	S@: amber<<resin

For 72 pairings, there were 3 possible matches:

- [\[ACCEPT\]M25@:|MINERAL|](#) <-> [S@:|compound|](#) : 0.7916666
- [\[ACCEPT\]M25@:|EARTH-MATERIAL|](#) <-> [S@:|stuff|](#) : 0.8333333
- [\[ACCEPT\]U@:|DECOMPOSABLE-OBJECT|](#) <-> [U@:|NONDECOMPOSABLE-OBJECT|](#) : 0.9756098

The maximal consistent subset size is 3. There were 1 subsets found of that size:

- [\[ACCEPT\]](#)
 - [M25@:|MINERAL|](#) <-> [S@:|compound|](#) : 0.7916666
 - [M25@:|EARTH-MATERIAL|](#) <-> [S@:|stuff|](#) : 0.8333333
 - [U@:|DECOMPOSABLE-OBJECT|](#) <-> [U@:|NONDECOMPOSABLE-OBJECT|](#) : 0.9756098

[Next Page](#)

Document: Done

- Is Amber Decomposable or Nondecomposable?
- The 'stone' sense (Mikro) is; the 'resin' sense (WordNet) isn't...
- What to do??

[Next](#)

Shishkebobs

- Library ISA Building (and hence can't buy things)
Library ISA Institution (and hence can buy things)
SO: Building \square Institution \square Location ...a Library is *all* these
- Also: Field-of-Study \square Activity \square Result-of-Process:
(Science, Medicine, Architecture, Art...)
- Allowing shishkebobs makes merging ontologies easier
(possible?): you respect each ontology's perspective
- Continuum: from on-the-fly shadings to metonymy
(see Guarino's *identity conditions*; Pustejovsky's *qualia*)

Current status

- **Omega:**
 - Approx. 110,000 concepts
 - Approx. 1.1 mill instances
 - Subject information from TAP (Guha et al.)
 - Additional information from various sources
- **Tools:**
 - Alignment algorithms
 - Concept spotting, glossary parsing algorithms
 - Instance harvesting algorithm
 - Work on learning inter-concept/instance relations
- **Infrastructure:**
 - Busy putting instances into RDF (also database form?)
 - Concepts in Lisp-like notation; where next? — toward Semantic Web?
KR? RDF?

RestartOntoLoad Find: Word Concept Match Home: EIA

[About](#) [AskCal](#) [Legend](#) [XML](#)

Matches of "vehicle" in ontology EIA

Matching Concepts:
[Aftermarket converted vehicle](#)

Matching Words:
[Aftermarket converted vehicle](#)

Matches in other ontologies

Matching Concepts:
vehicle, VEHICLE, 4WD<motor vehicle, AIR-VEHICLE, AIR-VEHICLE\$NOUN, AIR-VEHICLE-MANUFACTURING-CORPORATION, AIR-VEHICLE-PART, AIR-VEHICLE-PART\$NOUN, ANIMAL-PROPELLED-VEHICLE, ANIMAL-PROPELLED-VEHICLE\$NOUN, armored vehicle, caisson<<vehicle, CATERPILLAR-VEHICLE, CATERPILLAR-VEHICLE\$NOUN, craft<vehicle, drawn(of vehicles), ENGINE-PROPELLED-VEHICLE, ENGINE-PROPELLED-VEHICLE\$NOUN, FLY-AIR-VEHICLE, FLY-AIR-VEHICLE\$NOUN, (48 more matches)

Matching Words:
vehicle, amphibious vehicle, armored combat vehicle, armored vehicle, armoured combat vehicle, armoured vehicle, automotive vehicle, military vehicle, motor vehicle, passenger vehicle

Concept: Aftermarket converted vehicle

Definition:
(none recorded)

Direct-Superclass:
vehicle
medium<means
means>medium
instrumentality>arms
artifact
whole>sum
object>lot
NONVOLITIONAL_NONBIOLOGICAL-OBJECT
NONBIOLOGICAL-OBJECT
DECOMPOSABLE-OBJECT
TANGIBLE-OBJECT
OBJECT
Summum Genus
TANGIBLE-NONVOLITIONAL-OBJECT

vehicle>sled
conveyance>tram
instrumentality>arms*

Direct-Subclass:
(Leaf Node)

Ontology (nicknames)	Details	Concepts	Words
EDC	Concepts: 428; EN words: 841; ES words: 1073	EDC@	EDC@EN EDC@ES
EIA	Concepts: 3; EN words: 3	EIA@	EIA@EN
O	Concepts: 121411; EN words: 148237; ES words: 26085	O@	O@EN O@ES
SIMS	Concepts: 2; EN words: 2	SIMS@	SIMS@EN

<http://omega.isi.edu>

The other route: aligning databases directly

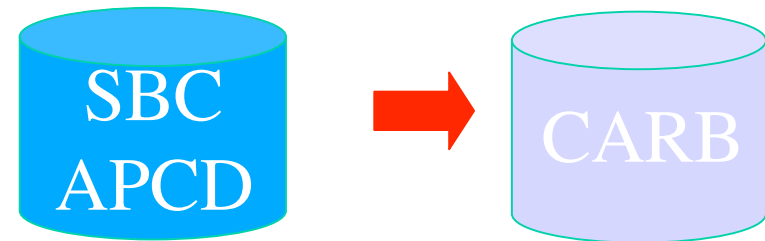
(This work with Andrew Philpot)

Cross-source data mapping using statistical MT

Recent advances in MT have allowed the automatic induction of **cross-natural language** correspondences from large multi-lingual corpora

We are investigating applying these techniques to learn **cross-database** correspondences, based on features such as:

- Declared or detected metadata: *e.g., field names, database schema, table headers, footnotes*
- Learned data patterns: *e.g., domain, range, formats, orthography*
- Topological relationships: *e.g., foreign key/subset discovery*
- Terminological reference via ontology or thesaurus



Statistical MT (1): Weight training

FR: Il y a un crayon jaune sous le grand camion.

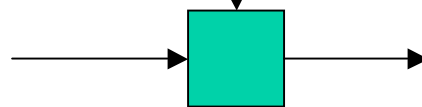
EN: There is a yellow pencil under the big truck.

Large corpus of parallel French/English sentences
(e.g., from Canadian Parliamentary Records)

Conditional item-item probabilities

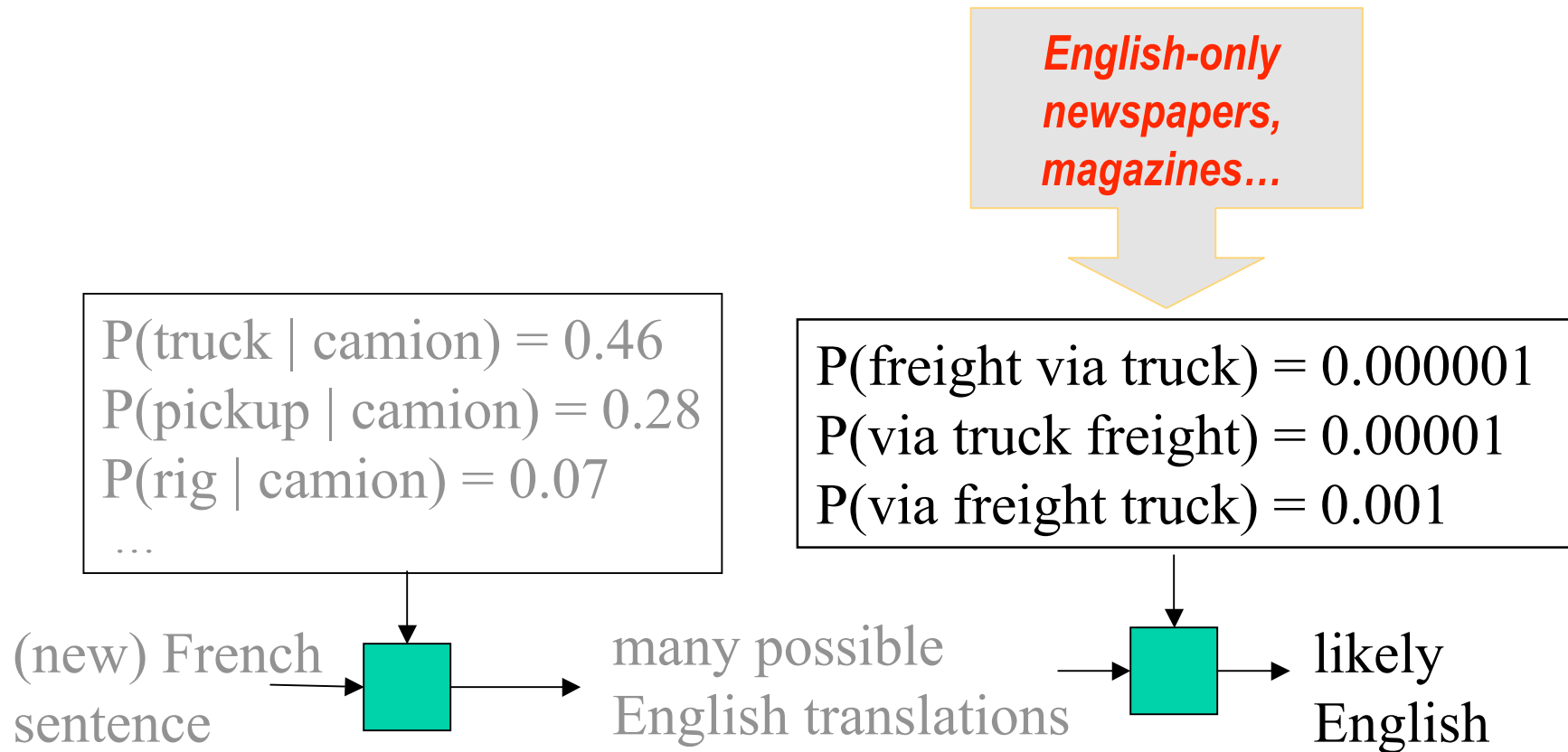
$P(\text{truck} \mid \text{camion}) = 0.46$
 $P(\text{pickup} \mid \text{camion}) = 0.28$
 $P(\text{rig} \mid \text{camion}) = 0.07$
...

(new) French sentence

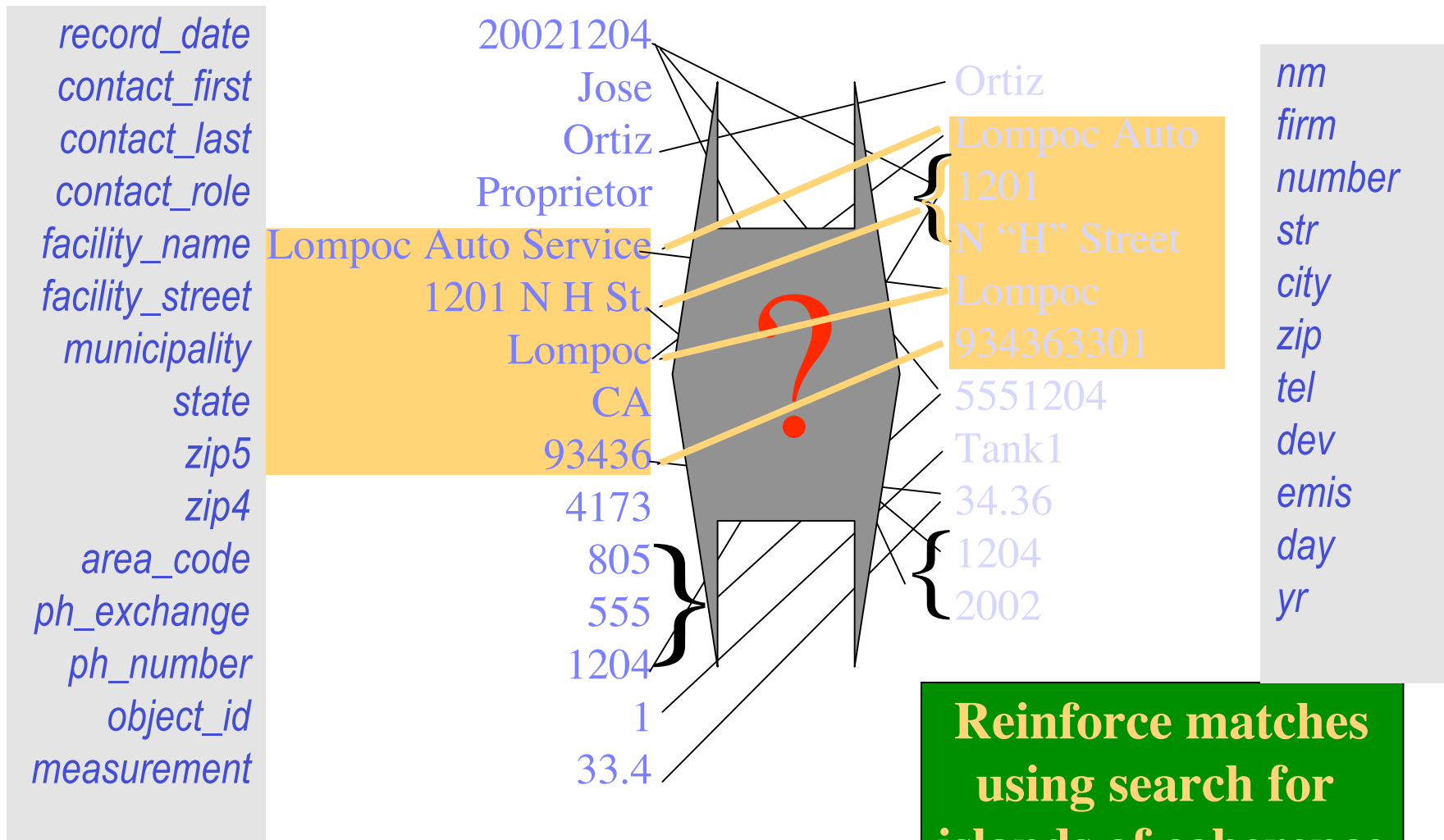


many possible English translations

Statistical MT (2): Coherence filtering



MT for automatic record linkage



**Reinforce matches
using search for
islands of coherence**

Conclusion

Next steps

- Collect **instances** of many other entities (not only locations, organizations, and people)
- Learn **more details** about each person, location, organization, etc., using patterns: date of birth, nationality, occupation, spouse, etc.
- Into Omega, incorporate **WordNet extensions** (inference rules) from (Moldovan 03)
- Merge **OpenCyc** and perhaps SIC code terms into Omega
- Build **access tools** and inline access methods to support QA, summarization, etc.

Vision

- Many people could use something like Omega:
 - The Semantic Web needs a large standardized well-organized multi-lingual termset
 - MT systems need a language-independent (or at least neutral) ontology
 - Many HLT systems can use the semantic and instancial information in Omega for better performance
 - Database integration and access systems might use something like Omega
 - AI systems might take subsets of it
- People should be encouraged to build their own!

Thank you