Ontological Clustering: Battling with the Concept of Concept

Aurelie Herbelot

University of Cambridge

Ontology Extraction by Clustering

input → **clustering module** → **output** → **evaluation**

corpus → **clustering module**

hill
forest
lake
path

Web

WIKIPEDIA
The Free Encyclopedia

distributional similarity → clustering module

hill
forest
lake
path
cliff
river
wood
field

Clustering with Distributional Similarity

Distributional Similarity (Harris, 1968):

Terms found in similar contexts are semantically similar. (Context: bag of words, lexico-syntactic patterns, semantic patterns…)

Here, our patterns are RMRS features. (Robust Minimal Recursion Semantics, Copestake, 2003.)

Lemma: dig

ARG1: chipmunk   ARG2: burrow

Lemma: dig

ARG1: aardvark   ARG2: burrow

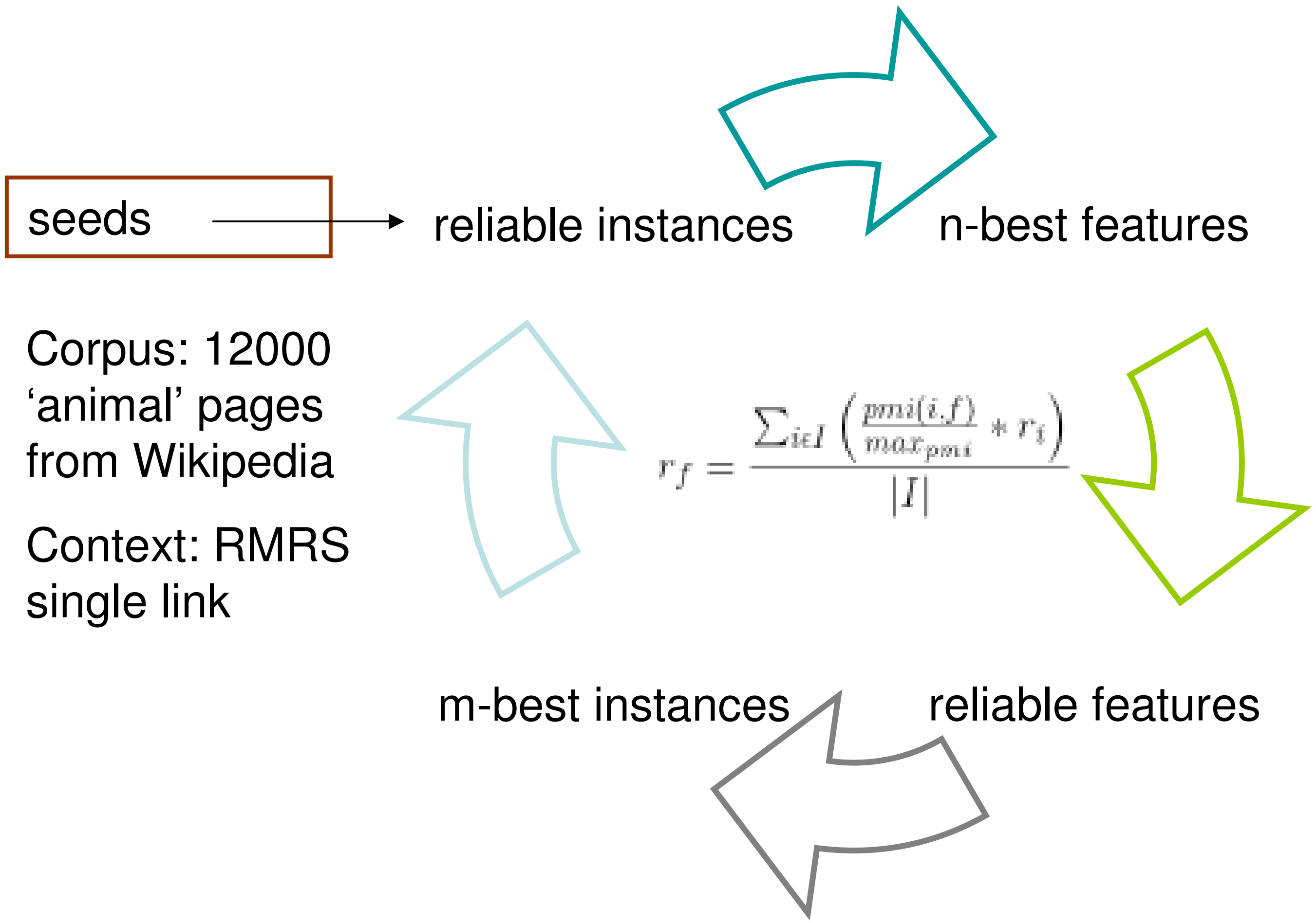
Common Issues with Distributional Similarity

1. defining context
2. patterns are not uniquely bound to concepts
   - Ex: ARG1 [prohibit] – by – ARG2 [hole_]
3. word sense ambiguity
   - Ex: company, society, group
4. amount of data required: co-ocurrence of patterns and seeds must be statistically reliable
5. defining the concept of ‘concept’: how to choose seeds, how to evaluate results (problems of definition and of specificity…)

Initial Investigations: Choice of Seeds on the Wikipedia Corpus

- 12000 pages on animals extracted from Wikipedia, parsed with RASP2 (Briscoe et al, 2006) and the RASP-to-RMRS converter (Ritchie & Copestake)
- Single-link features (only one argument considered for each pattern).

Boosting Recall with Bootstrapping



$$r_f = \frac{\sum_{i \in I} \left(\frac{pmi(i,f)}{max_{pmi}} * r_i \right)}{|I|}$$

Corpus: 12000 ‘animal’ pages from Wikipedia

Context: RMRS single link

calculate reliability of features using Pointwise Mutual Information and reliability of the instances they extract

Results on the Wikipedia Corpus

Four Queries:
animal names (1): animal, mammal, fish, bird, insect, cat, snake
animal names (2): angelfish, annelid, bat, fly, drosophilid, shrimp, kangaroo
body parts: whisker, hoof, bone, eye, fin, heart, wing
landscape features: cliff, forest, desert,lake, marshland, mountain, jungle

Query	Num Extractions	Recall	Minimal Precision
animal1	1551	98%	31%
animal2	531	34%	42%
body parts	1118	172%	15%
landscape	278	108%	19%

calculated against WordNet

Automatic Seed Selection (1)

- **Seeds in the middle of the conceptual hierarchy are better**

  Rosch (1976): The notion of 'basic level category' refers to the level in a conceptual hierarchy which best gathers the characteristic elements of a concept, that is, the categorical level of the best prototype. The notion usually refers to levels halfway through the hierarchy.

- **Seeds with a medium frequency are better**

Automatic Seed Selection (2)

1. Gather a set of potential seeds *Ws* from WordNet (look for a common ancestor to user seeds and record 10 levels of hyponyms *A1* to *A10* with at most 2 senses).
2. Find average frequency of terms in *Ws* and create four frequency brackets around the average. Run system on each bracket, compute automatic precision against WordNet.
3. Keep seeds in best frequency bracket. Run system on each conceptual level *A1* to *A10*. Keep level with best automatic precision.

Automatic Seed Selection: Results

Animals (1577 instances in corpus as per WordNet)			
Freq Range	Num Extractions	Recall	Precision
0-35	14	1%	7%
35-70	1192	76%	**36%**
70-105	1266	80%	35%
105-140	1823	116%	29%
Body Parts (651 instances in corpus as per WordNet)			
Freq Range	Num Extractions	Recall	Precision
0-15	1	0.2%	100%
15-30	137	21%	**46%**
30-45	551	85%	19%
45-60	1451	223%	13%
Landscape features (257 instances in corps as per WordNet)			
Freq Range	Num Extractions	Recall	Precision
0-45	53	21%	2%
45-90	386	150%	8%
90-135	202	79%	**20%**
135-180	428	167%	4%

Frequency plays a role in precision

Level in conceptual hierarchy only helps the animal query

Layer	Num Extractions	Recall	Precision
1	236	15%	22%
2	55	3%	31%
3	1823	116%	29%
4	1823	116%	29%
5	1823	116%	29%
6	1067	68%	39%
7	1192	76%	36%
8	1067	68%	39%
9	558	35%	40%
10	558	35%	40%

Further Investigations: Basic WSD and Weeding on TREC8

- 4000 pages subset of TREC8, parsed with RASP2 (Briscoe et al, 2006) and the RASP-to-RMRS converter (Ritchie & Copestake)
- Multiple-links features (several arguments considered for each pattern).

Some Basic Word Sense Disambiguation (1)

Choosing features that are linked to all seeds act as disambiguation:

approve pass adopt

Threshold	0.3	0.6	1
Precision	27%	53%	21%
Num Extractions	11	17	157

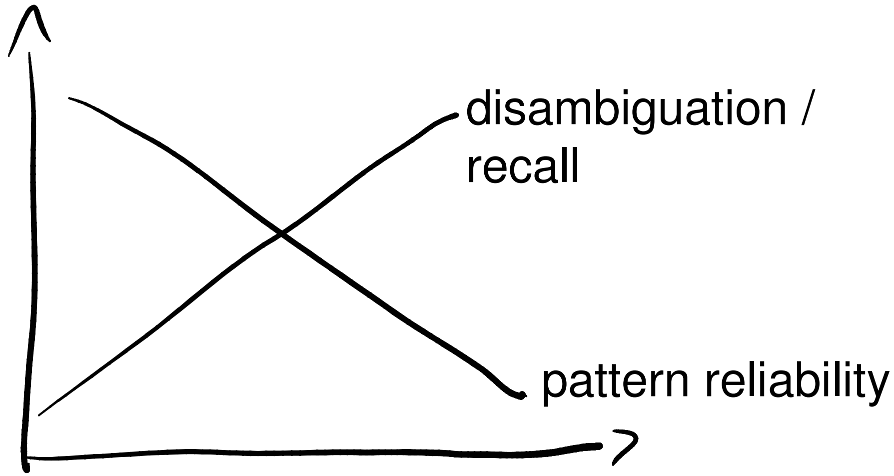
society company group

Threshold	0.3	0.6	1
Precision	100%	70%	27%
Num Extractions	3	20	210

Some Basic Word Sense Disambiguation (2)

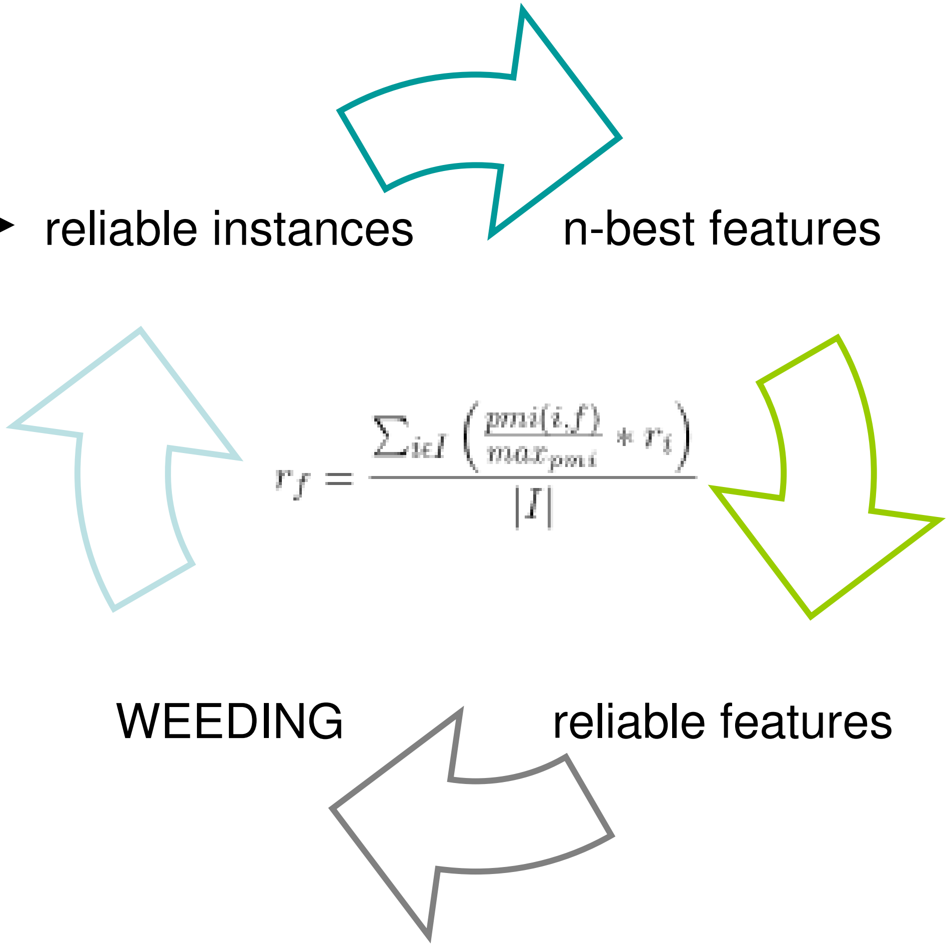
- Shared features do provide some disambiguation:
  1. Threshold 0.3: ARG1 [hole_] – through – ARG2 [arch]
  2. Threshold 0.6: ARG1 [senate] – hole_ – ARG2 [bill]
  3. Threshold 1: ARG1 [unanimously] – hole_

- But pattern reliability is inversely proportional to disambiguation level



Boosting Precision with Weeding

seeds

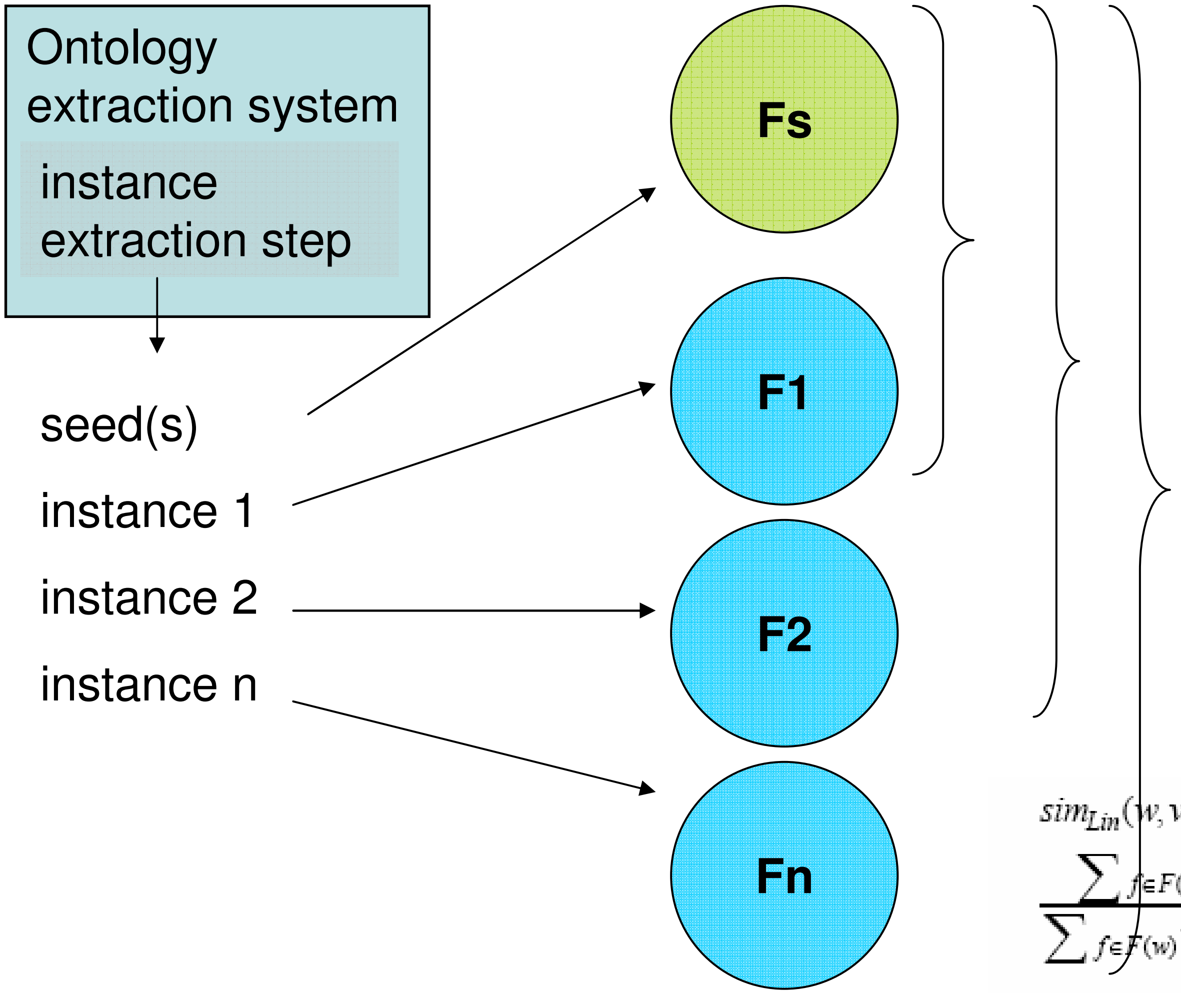


Corpus: 4000 pages from TREC8

Context: RMRS multiple links

calculate reliability of features using Pointwise Mutual Information and reliability of the instances they extract

Weeding Bad Instances



calculate Lin similarity measure between seeds' features and each instance's features. Only keep instances that have a similarity> threshold.

$$sim_{Lin}(w,v) = \frac{\sum_{f \in F(w) \cap F(v)} weight(w,f) + weight(v,f)}{\sum_{f \in F(w)} weight(w,f) + \sum_{f \in F(v)} weight(v,f)}$$

The Miller-Charles Experiment (1)

Pair	Miller Charles	Feature-based method	Pair	Miller Charles	Feature-based method
car-automobile	3.92	0.0563107	crane implement	1.68	0.00750327
gem-jewel	3.84	0.0850364	journey car	1.16	0.0508244
journey-voyage	3.84	0.115798	monk oracle	1.1	0.0259974
boy-lad	3.76	0.0256929	cemetery woodland	0.95	0.0397185
coast shore	3.7	0.0975351	food rooster	0.89	0.00298349
asylum madhouse	3.61	0.0159835	coast hill	0.87	0.0498394
magician wizard	3.5	0.0477247	forest graveyard	0.84	0.0112584
midday noon	3.42	0.0674808	shore woodland	0.63	0.0100002
furnace stove	3.11	0.0633645	monk slave	0.55	0.0298227
food fruit	3.08	0.102363	coast forest	0.42	0.058168
bird cock	3.05	0	lad wizard	0.42	0.048547
bird crane	2.97	0.0525795	chord smile	0.13	0.0179546
tool implement	2.95	0.0239168	glass magician	0.11	0.011844
brother monk	2.82	0.041539	rooster voyage	0.08	0.0150575
lad bother	1.66	0.0160828	noon string	0.08	0.0152741

The Miller-Charles Experiment (2)

- Correlation: 0.529165457

  When removing low frequency terms (freq < 100): 0.742468733

- Jarmasz & Szpakowicz (2003) report previous figures between 0.732 and 0.878 for systems using lexical resources such as WordNet or Roget's Thesaurus.

Results on the TREC Corpus

- Run extraction system for one iteration. Record number of extractions / precision before and after weeding. Threshold: 0.02

Company group society	Precision	Recall
Before	68%	22
After	88%	8

Scientist biologist researcher	Precision	Recall
Before	19%	53
After	57%	14

Black red yellow	Precision	Recall
Before	43%	37
After	53%	30

Gun revolver rifle	Precision	Recall
Before	13%	156
After	17%	115

Approve adopt pass	Precision	Recall
Before	59%	17
After	88%	8

Car bike bus	Precision	Recall
Before	33%	57
After	24%	34

Results on the TREC Corpus

Example: law bill constitution

1. appropriation
2. ban
3. cage
4. candidate
5. century
6. class
7. country
8. court order
9. decision
10. effort
11. fuel
12. function
13. glass case
14. hospital
15. industry
16. kray
17. legislation
18. letter
19. measure
20. name
21. network
22. paragraph
23. party
24. policy
25. road
26. ruling
27. territory
28. treaty
29. bill
30. constitution
31. law

before

1. law
2. constitution
3. bill
4. legislation
5. paragraph
6. treaty
7. measure
8. ban
9. ruling
10. kray
11. glass case

after

Evaluation Issues

- Human evaluation:
  - what is a concept?
  - how to deal with polysemous words?
  - manual recall / precision evaluation only possible on corpus subset
- Task-based evaluation
  - it must be possible for the task to be evaluated by humans!

In Summary…

- Issues inherent to distributional similarity make it difficult to control output (in particular the multiple correspondences between concepts and features).
- The choice of seeds can drastically affect results. Seed frequency help getting higher results but seed position in the hierarchy only affects concepts which are taxonomical in nature.
- WSD can be partially achieved by using several input seeds. Disambiguation and pattern reliability are inversely correlated.
- It is possible to a certain extent to ‘weed’ bad instances from results.
- Human evaluation requires a precise definition of the concept under investigation.