

What good is computational linguistics?

John A Goldsmith
The University of Chicago
<http://linguistica.uchicago.edu>

9 January 2014



1 *Problems and Solutions in Natural Language Processing*

With the rise of the internet, a massive amount of data has become available in the form of texts and messages in English as well as in other natural languages. This information can be of great value, but some kind of analysis is always needed to allow the user to find, use, or understand it. The field that is concerned with this kind of work is called natural language processing.

Surprisingly, people who do not work in natural language processing rarely have a good intuition as to which of these categories their needs fall into. I will look at a range of examples, and explain why they fall into these categories, and what might change in years to come.

Problems that users would like to have their software deal with divide into these categories:

1. Software can be written to solve your problem.
2. It will be a long time before good software will be available to solve your problem.
3. If we redefine your problem a little bit, we can write software that will do an excellent job.
4. If we redefine your problem a little bit, we can write software that can at the very least be useful, and it is being improved with each passing year.



Public Sector BAO: Natural Language Processing Analyst

IBM - Chantilly, VA

Posted 58 days ago

Business Analytics and Optimization

Other Details

[View full job listing](#)

This is a preview of the Public Sector BAO: Natural Language Processing Analyst job at IBM. To view the full job listing, join LinkedIn - its free!

About this job



Job description

The Public Sector BAO Natural Language Processing (NLP) Analyst will be responsible for evaluating system performance and identifying steps to drive enhancements. This role is part analyst and part developer. The NLP Analysts will function independently to dive deep into system components, identify areas for improvement, and devise solutions. The NLP Analyst will also drive test and evaluation of the solutions, and empirically identify follow on steps to implement continuous system improvement. Natural Language Processing is an explosively dynamic field and NLP Analysts must expect ambiguity, and demonstrate the ability to develop courses of action on the basis of data driven analysis. Software development skills are necessary but not sufficient; successful candidates must demonstrate the ability to leverage data. Consultants will be located in an IBM office in the metropolitan area or at an IBM client site. Selected candidates may not need to travel for all projects outside of their metro area. However, all candidates must be able and willing to travel based on assigned project demand. Travel requirements may vary but could be up to 100%. Candidates are not able to refuse project based on travel. Work alongside some of the best minds in the industry as a Data Specialist at IBM. Want your skills to make a difference in how the world works? IBM is seeking talented IT Specialists who are interested in using their expertise to integrate, optimize and make available structured and/or unstructured data using database products, technologies and methods.

As a Data Specialist, you'll work in a collaborative team environment to deliver database designs, information models (logical, physical, dimensional, etc.) data migration plans and data warehouses. Get access to resources that only a global leader like IBM can provide.

In this position you will primarily apply your technical skills in an internal or external customer billable services and implementation environment.

p>Be a part of making the world more instrumented, interconnected - and intelligent. Join us.

Interested in learning more about IBM? Check out the IBM Global Careers newsletter .

2 *Computational Linguistics (CL) and Natural Language Processing (NLP)*

- A rough distinction is often made between CL and NLP. One way the distinction is understood reflects the difference between science (CL) and engineering (NLP), or between solving *theoretical* questions and solving *practical* problems.
- Another distinction that is sometimes made is between studying the *form = grammatical structure* of the corpus (text) and studying the *content* (meaning).
- Because of the large amount of data available today, most useful software contains a large element of *learning from training data*.

Our interest today is on *practical* questions bearing on *content*.

Terminology:

Corpus (plural: **corpora**) Computer readable English, French, Chinese (etc.) texts. Novels, web-pages, government reports, Twitter feeds, Yelp comments, internal emails, and many other things.

3 *Standard problems*

- Speech technology:
 - speech recognition
 - Text-to-speech (TTS)
- Automatic translation from one language to another (Machine translation, or MT)
- Miscellaneous
 - Information extraction: identifying and classifying entities referred to in texts. For example: [Named entity recognition](#). Many ways to identify the same person:
 - * President Kennedy, John Kennedy, John F. Kennedy.
 - * Osama Ben-Laden, OBL, Usama ..., Usamah Bin Ladin, Oussama Ben Laden, Osama Binladin.
 - * Is General Motors the same kind of entity as General Eisenhower? General Waters is a company in England, but General Waters was also General John K. Waters (1906-1889).
- Miscellaneous (continued)
 - Sentiment analysis: mapping textual customer response to a number from 1 to 10
 - Spell-checking.
 - Grammar-checking.
- Document retrieval: a problem with many sides to it.
- Using social media (crowd sourcing) to detect restaurants that *ought to be* inspected by city restaurant inspectors.

Any problem that really requires that the algorithm *understand* the text is unsolvable. But that turns out to be an unrealistically high bar.

4 *Bag of words model*

- Ignore linear order of words. This means giving up much of what makes language meaningful! E.g., occurrences of *not*.
 - *I am (not) in love with you.* That *not* really matters.
 - *Not that it matters (not that you care, not surprisingly), I am in love with you.* That *not* is much less important.
 - Or *I am in love with you, not with Sally.*

What is the following sentence about?

- *NYTimes* December 28, 2013: a a a about Agency among an and and balance big collects contribution courts data debate enormous era extraordinary federal Friday group how in is judge latest legal making National of of on phone presidential program privacy records review ruled security Security that that the the to to troves
- **Better:** Agency balance big collects contribution courts data debate enormous era extraordinary federal Friday group judge latest legal making National phone presidential program privacy records review ruled security Security troves

- It is an astonishing fact that a very large proportion of practical tasks can be accomplished using a *bag of words model*: just looking at the words in a sentence, and ignoring their serial order.
- It is often helpful to put greater weight on words that do *not* appear uniformly over all documents.
- Latent Dirichlet models. A statistical method that works hand-in-glove with *bag of words* models. Bags of words are naturally described as if they were generated by multinomial distributions. But documents that are *about* particular subjects will involve more use of words in a particular vocabulary (think *baseball, finance, politics,...*). Various statistical methods of modeling the relationship between word choices in a document have been explored over the last 20 years, and latent Dirichlet models have inspired a good deal of exploration.

A federal judge on Friday ruled that a National Security Agency program that collects enormous troves of phone records is legal, making the latest contribution to an extraordinary debate among courts and a presidential review group about how to balance security and privacy in the era of big data.

5 *Big data: Data everywhere*

- The World Wide Web (whose native language is *html*).
- Municipal, state, national agencies make a great deal of information public. Courts make bankruptcy declarations public in pdf form with a great deal of information.
- Social media.

6 Information Extraction

Extracting:

- Names
- Other specific entities (dates, diseases, proteins, countries)
- Pairs of objects entering into relationships
- Events: extract the key elements of an event (who, what, where, when, how...)

This was viewed as an important step towards *message understanding*, and was funded by the US Navy.

Hand-coded rules:

- (Capitalized word)+ "Inc." → organization
- Mr. ([Cap word]) (Cap letter .) [Cap Word] → person
- common-given-name (Cap letter .) [Cap Word] → person

Link this to entity recognition across alternative descriptions. *Prescott Adams announced the appointment of a new vice president for sales. Mr Adams explained...*

Beyond hand-coded rules:

- We know that Mozart lived from 1756 to 1791—and a lot of people know that. Can we search the web for paragraphs that include "Mozart" and also "1756" and "1791"? Are there formal *patterns* that can be discovered in which the dates are embedded?
- Yes—quite a few. The most common is: (1756-1791): that is, "(-)" or "(dddd1-dddd2)" where *dddd1* and *dddd2* are four digit sequences, and we can label such pairs as *date of birth* and *date of death*.
- Can we find *meta-patterns*? That is, constructions in text which can be used to identify useful relationships? One of these is **X, such as Y**: *non-profit publishers, such as The University of Chicago Press; third-world countries, such as Zambia and Haiti*.

Ralph Grishman 2010 "Information extraction"

7 Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews

Jun Seok Kang, Polina Kuznetsova (Stony Brook CS)
Michael Luca, Yejin Choi (Harvard Business School) July 2013

- A recent collaboration between computer scientists and business school researchers to measure the effectiveness of scraping on-line social media description of diners' experiences as a way to predict future failures of restaurants when visited by health inspectors.
- Data from Seattle restaurants 2006-2013: Yelp and Seattle municipal inspector records (public record). 13,000 inspections, 1756 restaurants, and 152,000 on-line reviews.
- Reviews chosen from 6-month period before inspection. Filtered out minor restaurant infractions.
- Goals: (i) detect and avoid spurious (fake, positive) restaurant reviews (ii) identify relevant words or word combinations (iii) determine if word-(language-) based experiments out-perform other methods (based, for example, on location or ethnicity of restaurant).

- They report some success in avoiding spurious reviews, based on detecting bimodal distributions of numerical ratings by customers and using results of other studies' text-based spurious-review detection (no details given).

- Inspectors' penalty scores appear to be on a scale from 0 to 60 (higher number is *worse*).

hygiene	gross, mess, sticky
service:neg.	door, student, sticker, the size
service:pos.	selection, atmosphere, attitude, pretentious
• food: pos	grill, toast, frosting, bento box
negative:	cheap, never, was dry
positive:	date, weekend, out, husband, evening
	lovely, yummy, generous, ambiance

Data	Accuracy
Number of reviews	50
Type of cuisine	66
Zip code	67
Average rating	58
Previous inspections	72
Unigram	78
Bigram	77
Unigram and bigram	83
Everything	81

8 Inexact String Matching

- This is an example of a real computer science problem whose solution (solutions) are of immediate interest to many real life tasks.

This problem has several variants. Here are two:

- Here is a list L_1 of the names of 100 banks. And here is a list L_2 of all of the banks in the world. For each bank in L_1 , find the best match in L_2 (or, find the n -best matches, ranked by goodness of match). (Names of all sorts of things are possible, of course.)
- Here is a large collection of texts. Consider all 100-letter strings (i.e., string that are 100 letters long) that appear twice, and I care about repetitions that are not perfect. Up to k letters may be different: that's good enough for my purposes.

- The first problem (bank names) can be attacked with the classic *string edit distance* or *Levenshtein distance* algorithm. It has two drawbacks: it is relatively slow, and it does not identify of letters (*linguistics* for *linguistics*).
- The second problem is a classic Big Data problem. A Big Data problem is:
 - One which is too big to be handled on a single processor;
 - One on which there is no upper bound to the amount of data the end-user wants to analyze. No matter what limit money and technology set on the amount of data handled today, the user wants to provide *more* data than that.

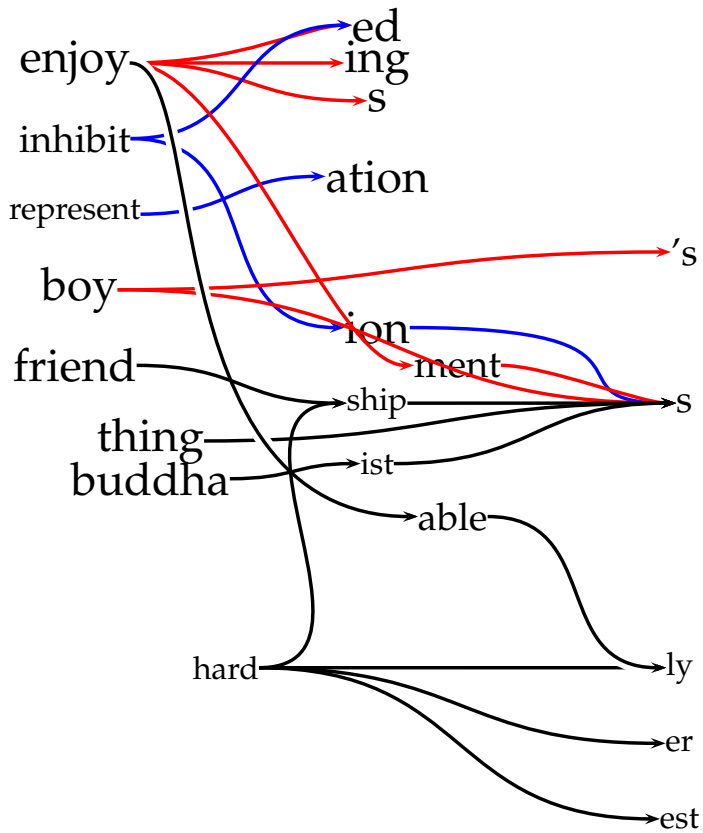
9 *Back to the kinds of problems we can take on:*

1. Software can be written to solve your problem.
 2. It will be a long time before good software will be available to solve your problem.
 3. If we redefine your problem a little bit, we can write software that will do an excellent job.
 4. If we redefine your problem a little bit, we can write software that can at the very least be useful, and it is being improved with each passing year.
- NLP progress often consists of shifting a problem from category 2 to categories 3 and 4, which may require considerable *domain expertise*: understanding what the end user needs and does not need—wants, and does not want.
 - The point at which imperfect solutions are acceptable has become lower because there is more useful information lurking in larger amounts of data, and because hardware is becoming less expensive — and also because we understand better how to divide large problems up into subpieces that can be computed in parallel, which better exploits the lower cost of computation.

10 A typical problem in computational linguistics

Develop an algorithm which will take in a large corpus in any human language, and will automatically (with no prior training) divide the words into prefixes, stems and suffixes.

Surprise application (1998): Microsoft's *Encarta*.



Word Sense Disambiguation

Problem:

The company said the *plant* is still operating ...

⇒ (A) Manufacturing plant or

⇒ (B) Living plant

Training Data:

Sense	Context
(1) Manufacturing	... union responses to <i>plant</i> closures
” ”	... computer disk drive <i>plant</i> located in ...
” ”	company manufacturing <i>plant</i> is in Orlando ...
(2) Living	... animal rather than <i>plant</i> tissues can be ...
” ”	... to strain microscopic <i>plant</i> life from the ...
” ”	and Golgi apparatus of <i>plant</i> and animal cells

Test Data:

Sense	Context
???	... vinyl chloride monomer <i>plant</i> , which is ...
???	... molecules found in <i>plant</i> tissue from the ...

Machine Translation

(English → Spanish)

Problem:

... He wrote the last **sentence** two years later ...
 ⇒ *sentencia* (legal sentence) or
 ⇒ *frase* (grammatical sentence)

Training Data:



Translation	Context
(1) sentencia	... for a maximum <i>sentence</i> for a young offender ...
” ”	... of the minimum <i>sentence</i> of seven years in jail ...
” ”	... were under the <i>sentence</i> of death at that time ...
(2) frase	... read the second <i>sentence</i> because it is just as ...
” ”	... The next <i>sentence</i> is a very important ...
” ”	... It is the second <i>sentence</i> which I think is at ...

Test Data:

Translation	Context
???	... cannot criticize a <i>sentence</i> handed down by ...
???	... listen to this <i>sentence</i> uttered by a former ...

sense-labeled training data?

- To do supervised WSD, need many examples of each sense in context

- have turned it into the hot dinner-party topic. The comedy is the
 - selection for the World Cup party, which will be announced on May 1
 - the by-pass there will be a street party. "Then," he says, "we are going
- ? 
- let you know that there's a party at my house tonight. Directions: Drive
- ? 
- in the 1983 general election for a party which, when it could not bear to
 - to attack the Scottish National Party, who look set to seize Perth and
 - number-crunchers within the Labour party, there now seems little doubt

Final decision list for *lead* (abbreviated)

To disambiguate a token of *lead*

:

- Scan down the sorted list
- The first cue that is found gets to make the decision all by itself
- Not as subtle as **combining** cues, but works well for WSD

Cue's score is its **log-likelihood**

ratio:

Position	Collocation	led	li:d
+1 L	lead level/N	219	0
-1 W	narrow lead	0	70
+1 W	lead in	207	898
-1w +1w	of lead in	162	0

LogL	Evidence	Pronunciation
11.40	<i>follow/V</i> + lead	⇒ li:d
11.20	<i>zinc</i> (in ± <i>k</i> words)	⇒ led
11.10	lead <i>level/N</i>	⇒ led
10.66	<i>of</i> lead <i>in</i>	⇒ led
10.59	<i>the</i> lead <i>in</i>	⇒ li:d
10.51	lead <i>role</i>	⇒ li:d
10.35	<i>copper</i> (in ± <i>k</i> words)	⇒ led
10.28	lead <i>time</i>	⇒ li:d
10.24	lead <i>levels</i>	⇒ led
10.16	lead <i>poisoning</i>	⇒ led
8.55	<i>big</i> lead	⇒ li:d
8.49	<i>narrow</i> lead	⇒ li:d
7.76	<i>take/V</i> + lead	⇒ li:d
5.99	lead , <i>NOUN</i>	⇒ led
1.15	lead <i>in</i>	⇒ li:d
	o o o	

Problem: Learning from Untagged Training Data

Sense	Training Examples (Keyword in Context)
?	... company said the <i>plant</i> is still operating ...
?	Although thousands of <i>plant</i> and animal species
?	... to strain microscopic <i>plant</i> life from the ...
?	vinyl chloride monomer <i>plant</i> , which is ...
?	and Golgi apparatus of <i>plant</i> and animal cells ...
?	... computer disk drive <i>plant</i> located in ...
?	... Nissan car and truck <i>plant</i> in Japan is ...
?	... the proliferation of <i>plant</i> and animal life ...
?	... keep a manufacturing <i>plant</i> profitable without ...
?	... animal rather than <i>plant</i> tissues can be ...
?	... union responses to <i>plant</i> closures
?	... molecules found in <i>plant</i> and animal tissue ...
?

plant ⇒ (A) manufacturing plant or
 ⇒ (B) living plant

very readable paper at <http://cs.jhu.edu/~yarowsky/acl95.ps>
 sketched on the following slides ...

Seed Words

- **Use words from dictionary definitions**
 - filtered for relevance by relative frequency and syntactic position
- **Use a single defining collocate for each class**
 - *crane* \Rightarrow BIRD or MACHINE
 - *plant* \Rightarrow LIFE or MANUFACTURING
- **Label salient corpus collocates**
 - co-occurrence analysis determines a small spanning set of collocates for hand labelling.

Example Initial State

Sense	Training Examples	(Keyword in Context)
A	used to strain microscopic	<i>plant</i> life from the ...
A	... rapid growth of aquatic	<i>plant</i> life in water ...
A	... that divide life into	<i>plant</i> and animal kingdom
A	beds too salty to support	<i>plant</i> life . River ...
A
?	... company said the	<i>plant</i> is still operating ...
?	... molecules found in	<i>plant</i> and animal tissue
?
?	... Nissan car and truck	<i>plant</i> in Japan is ...
?	... animal rather than	<i>plant</i> tissues can be ...
B
B	automated manufacturing	<i>plant</i> in Fremont ...
B	... vast manufacturing	<i>plant</i> and distribution ...
B	chemical manufacturing	<i>plant</i> , producing viscose
B	... keep a manufacturing	<i>plant</i> profitable without

1%

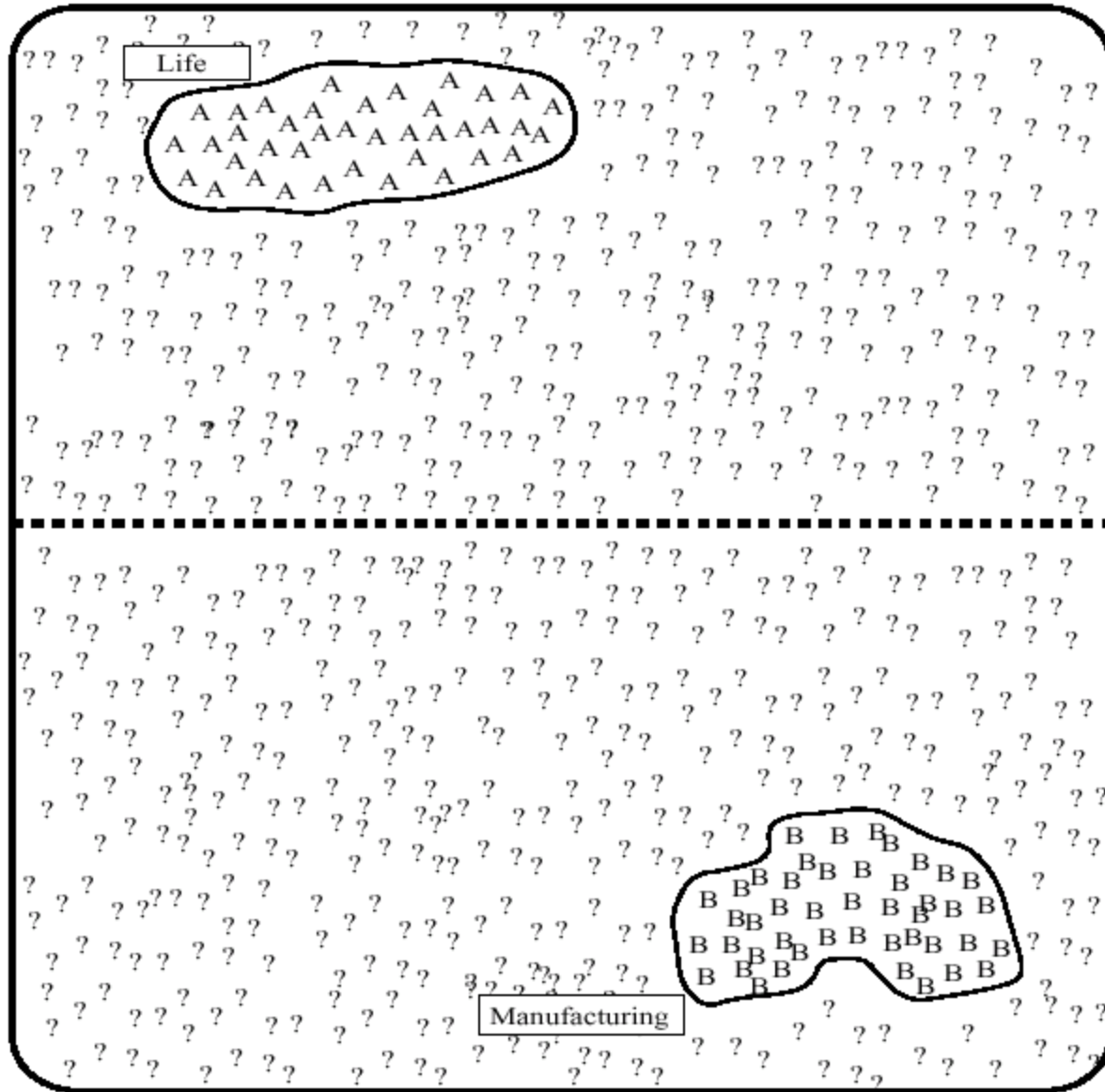
reasonably accurate

98%

1%

reasonably accurate

Example Initial State



Iteration Step

- Train a supervised sense tagger on the current seed sets

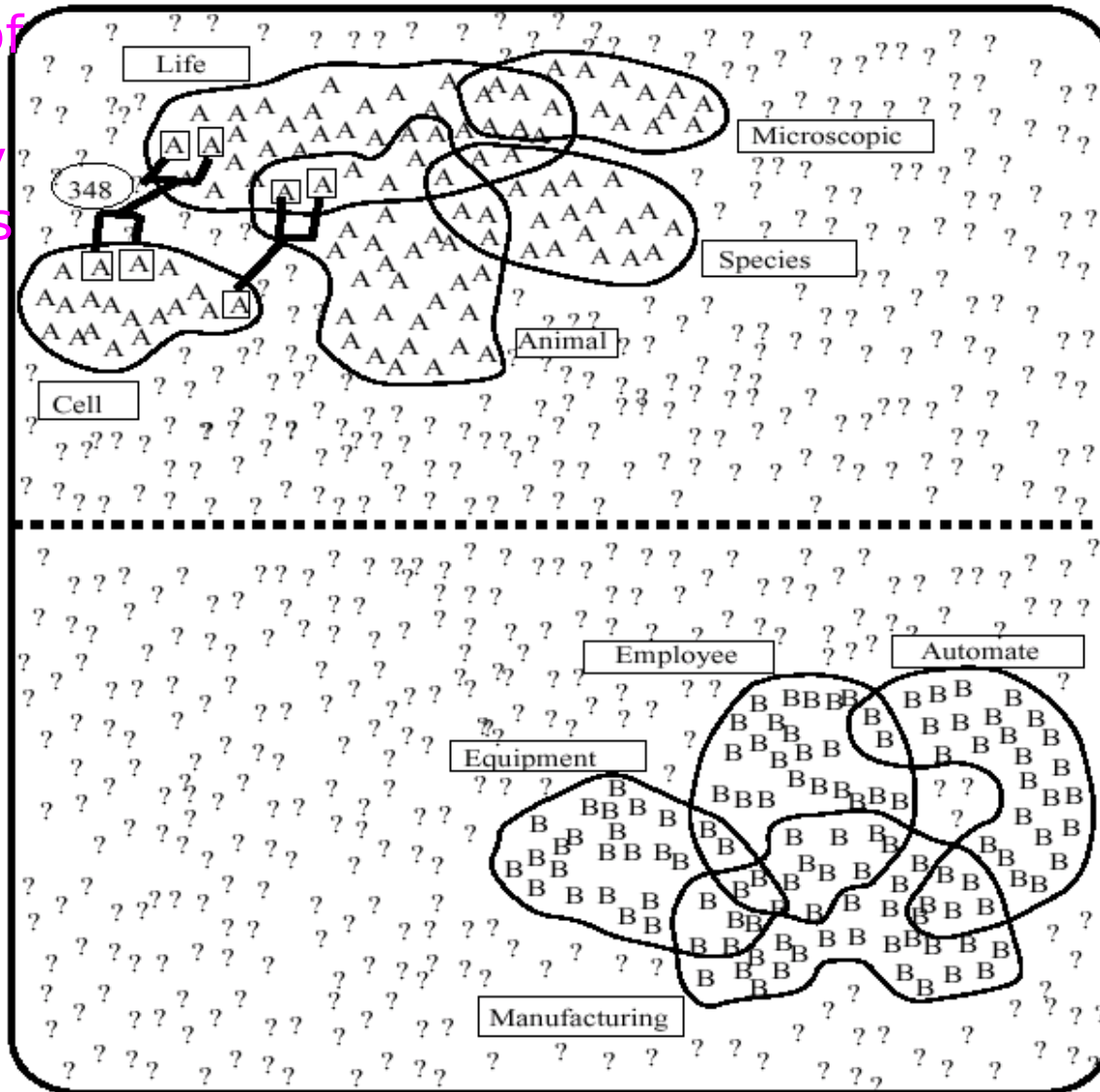
Initial decision list for <i>plant</i> (abbreviated)		
LogL	Collocation	Sense
8.10	<i>plant</i> life	⇒ A
7.58	manufacturing <i>plant</i>	⇒ B
7.39	life (within ±2-10 words)	⇒ A
7.20	manufacturing (in ±2-10 words)	⇒ B
6.27	animal (within ±2-10 words)	⇒ A
4.70	equipment (within ±2-10 words)	⇒ B
4.39	employee (within ±2-10 words)	⇒ B
4.30	assembly <i>plant</i>	⇒ B
4.10	<i>plant</i> closure	⇒ B
3.52	<i>plant</i> species	⇒ A
3.45	microscopic <i>plant</i>	⇒ A
	...	

no surprise
what the top
cues are

but other cues
also good for
discriminating
these seed
examples

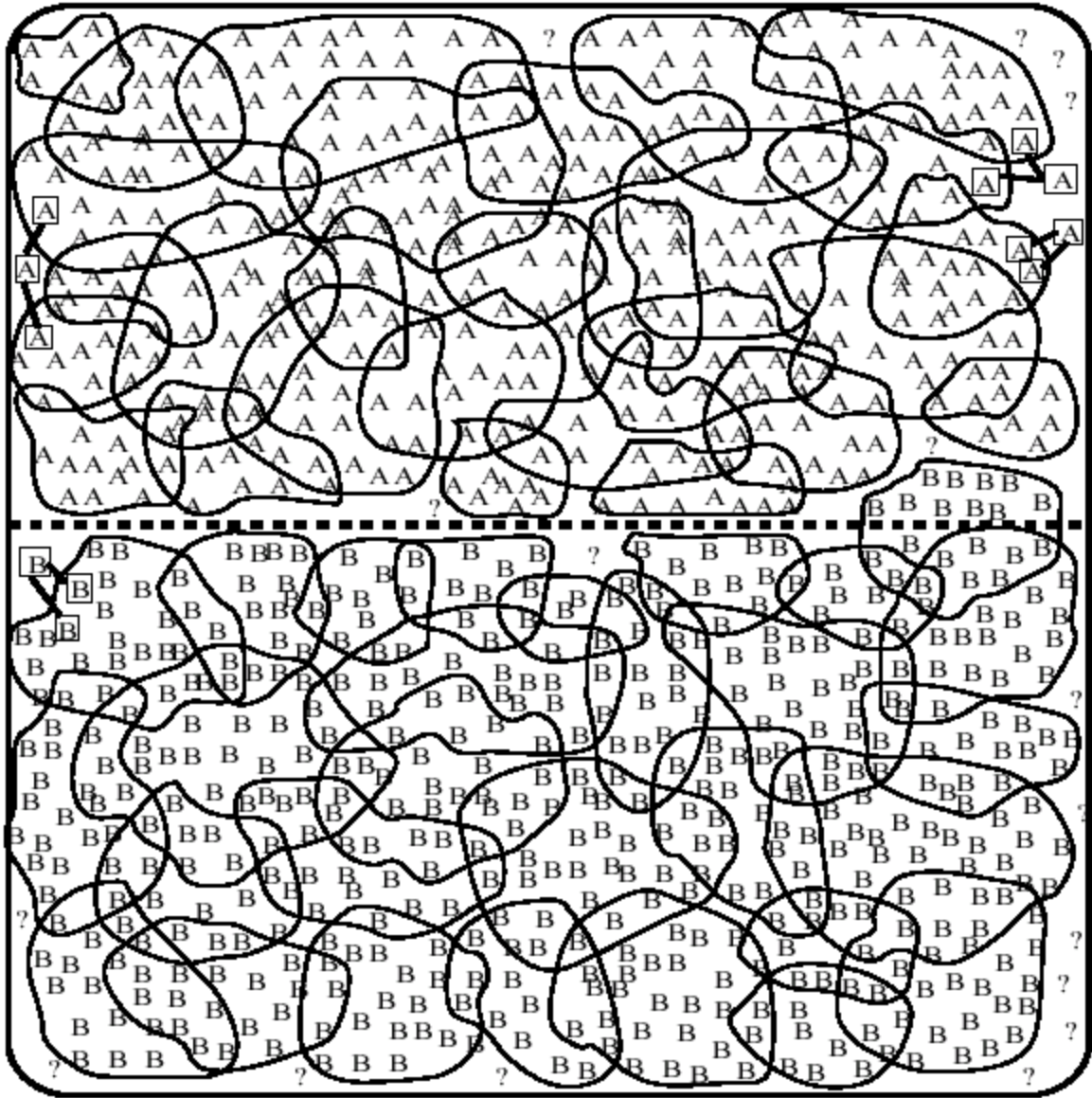
Example Intermediate State

the strongest of
the new cues
help us classify
more examples
...



from which
we can
extract and
rank even
more cues
that
discriminate
them ...

Final Training Iteration



Final Decision List

Final decision list for <i>plant</i> (abbreviated)		
LogL	Collocation	Sense
10.12	<i>plant</i> growth	⇒ A
9.68	car (within $\pm k$ words)	⇒ B
9.64	<i>plant</i> height	⇒ A
9.61	union (within $\pm k$ words)	⇒ B
9.54	equipment (within $\pm k$ words)	⇒ B
9.51	assembly <i>plant</i>	⇒ B
9.50	nuclear <i>plant</i>	⇒ B
9.31	flower (within $\pm k$ words)	⇒ A
9.24	job (within $\pm k$ words)	⇒ B
9.03	fruit (within $\pm k$ words)	⇒ A
9.02	<i>plant</i> species	⇒ A
...	...	

top ranked
cue
appearing in
this test
example

life and manufacturing are no longer even in the top cues!
many unexpected cues were extracted, without supervised
training

Now use the final decision list to classify **test** examples:

... the loss of animal and *plant* species through extinction ... ,