Computational Linguistics

CSC 485/2501
Fall 2023

4

4. Word sense disambiguation

Gerald Penn
Department of Computer Science, University of Toronto

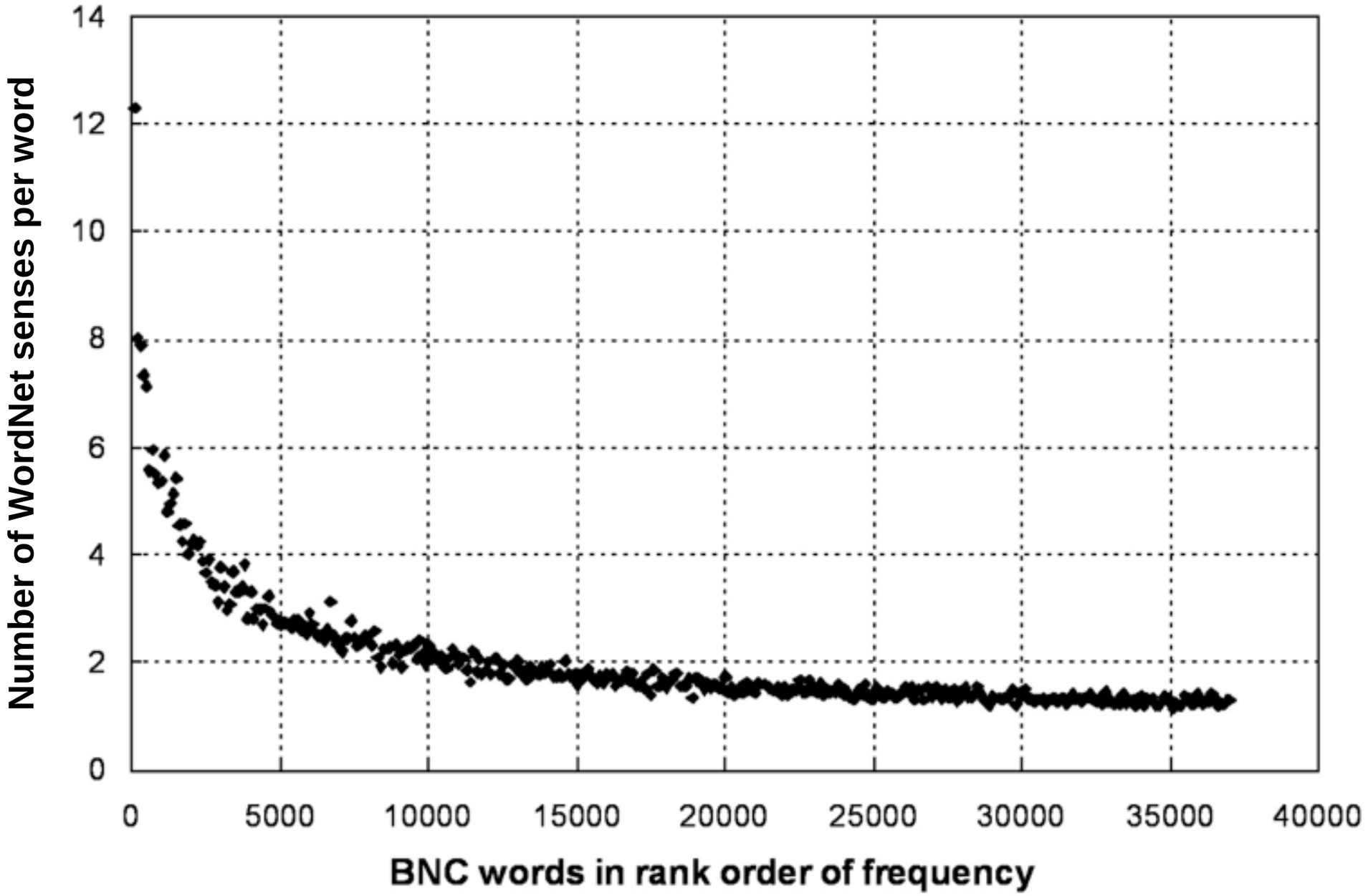
Based on slides by Lu Wang
Reading: Jurafsky & Martin: 20.1–5.

Word sense disambiguation

- Word sense disambiguation (WSD), lexical disambiguation, resolving lexical ambiguity, lexical ambiguity resolution.

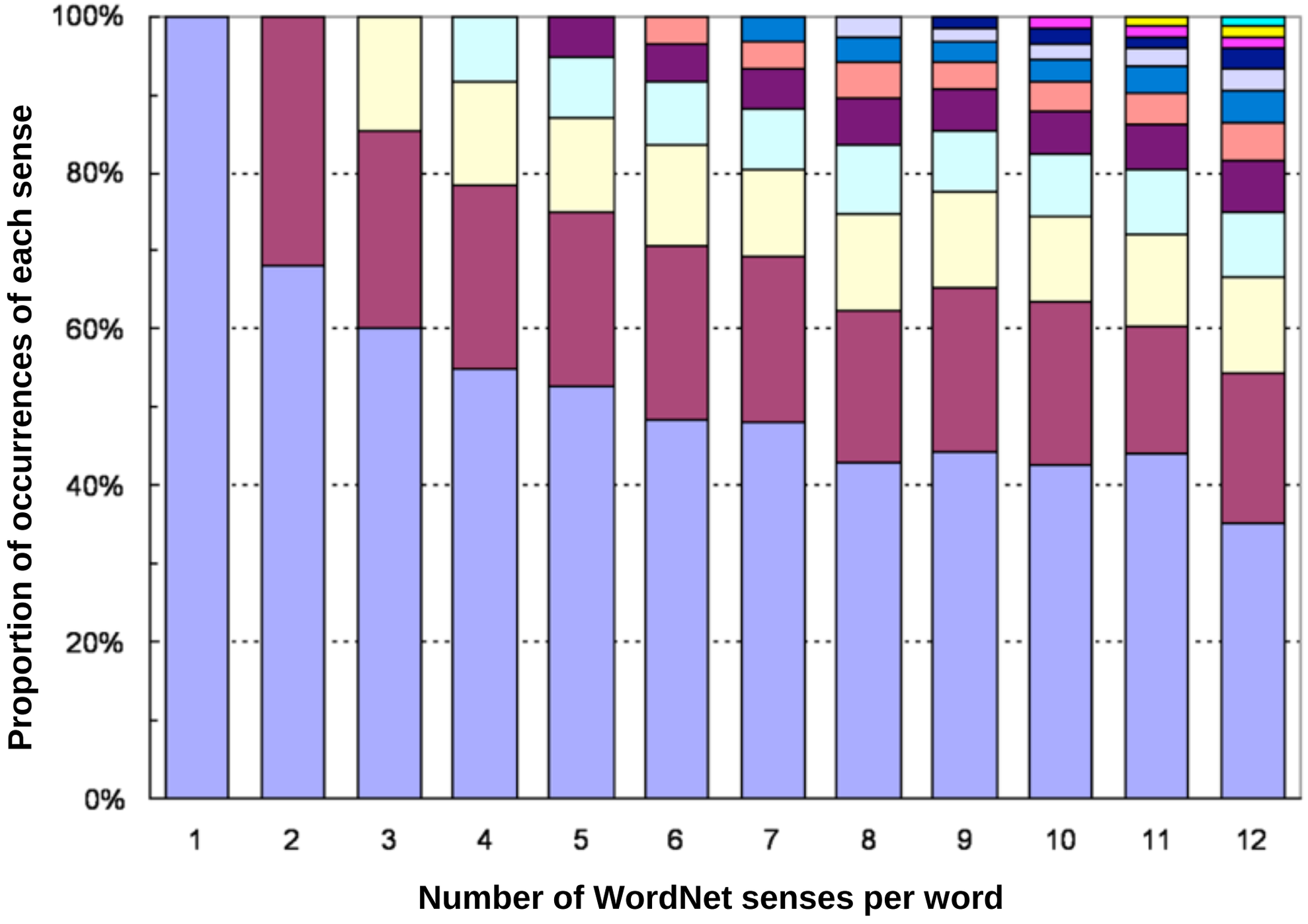
How big is the problem?

- Most words of English have only one sense.
  (62% in *Longman's Dictionary of Contemporary English*; 79% in WordNet.)
- But the others tend to have **several** senses.
  (Avg 3.83 in LDOCE; 2.96 in WordNet.)
- Ambiguous words are more frequently used
  (In British National Corpus, 84% of instances have more than one sense in WordNet.)
- Some senses are more frequent than others.



Words occurring in the British National Corpus are plotted on the horizontal axis in rank order by frequency in the corpus. Number of WordNet senses per word is plotted on the vertical axis. Each point represents a bin of 100 words and the average number of senses of words in the bin.

Edmonds, Philip. "Disambiguation, Lexical." *Encyclopedia of Language and Linguistics* (second edition), Elsevier, 2006, pp 607–623.



In each column, the senses are ordered by frequency, normalized per word, and averaged over all words with that number of senses.

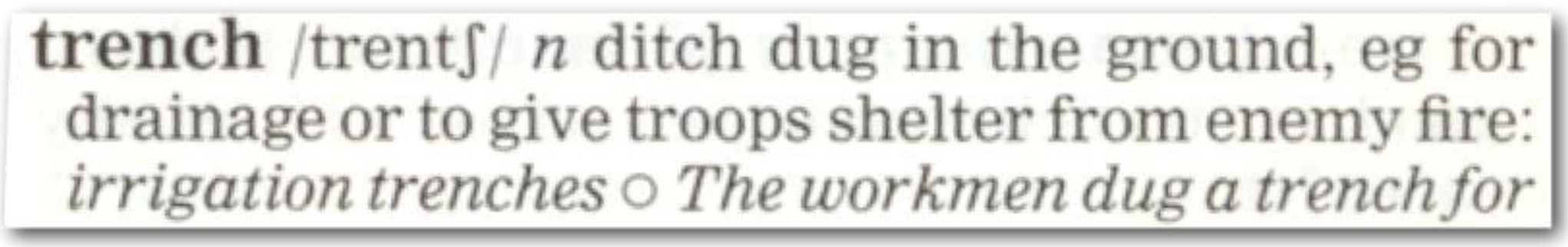
Edmonds, Philip. "Disambiguation, Lexical." *Encyclopedia of Language and Linguistics* (second edition), Elsevier, 2006, pp 607–623.

Sense inventory of a word

- Dictionaries, WordNet list senses of a word.
- Often, no agreement on proper sense-division of words.
- Don’t want sense-divisions to be too coarse-grained or too fine-grained.

Frequent criticism of WordNet

The American Heritage Dictionary of the English Language (3rd edition)



Oxford Advanced Learner's Dictionary (encyclopedic edition)

lit·ter (lĭt′ər) *n.* **1.a.** A disorderly accumulation of objects; a pile. **b.** Carelessly discarded refuse, such as wastepaper: *the litter in the streets after a parade.* **2.** The offspring produced at one birth by a multiparous mammal. See Synonyms at **flock**[1]. **3.a.** Material, such as straw, used as bedding for animals. **b.** An absorbent material, such as granulated clay, for covering the floor of an animal's cage or excretory box. **4.** An enclosed or curtained couch mounted on shafts and used to carry a single passenger. **5.** A flat supporting framework, such as a piece of canvas stretched between parallel shafts, for carrying a disabled or dead person; a stretcher. **6.** The uppermost layer of the forest floor consisting chiefly of fallen leaves and other decaying organic matter. **—litter**

AHDEL

litter /ˈlɪtə(r)/ *n* **1** (**a**) [U] light rubbish (eg bits of paper, wrappings, bottles) left lying about, esp in a public place: *Please do not leave litter.* ⇨ article at ENVIRONMENT. (**b**) [sing] state of untidiness: *Her desk was covered in a litter of books and papers.* ○ *His room was a litter of old clothes, dirty crockery and broken furniture.* **2** [U] straw, etc used as bedding for animals. **3** [CGp] all the young born to an animal at one time: *a litter of puppies.* **4** [C] (**a**) type of stretcher(1). (**b**) (formerly) couch carried on men's shoulders or by animals as a means of transport.

OALD

What counts as the right answer?

- Often, no agreement on which sense a given word-token is.
- Some tokens seem to have two or more senses at the same time.

Which senses are these? 1

- *image*

  1. a picture formed in the mind;
  2. a picture formed of an object in front of a mirror or lens;
  3. the general opinion about a person, organization, etc, formed or intentionally created in people's minds;

[*and three other senses*]

"… of the Garonne, which becomes an unforgettable ***image***. This is a very individual film, mannered, …"

Example from: Kilgarriff, Adam. "Dictionary word sense distinctions: An enquiry into their nature." *Computers and the Humanities,* 26: 365–387, 1993. Definitions from *Longman Dictionary of Contemporary English*, 2nd edition, 1987.

Which senses are these? 2

- *distinction*

  1. the fact of being different;
  2. the quality of being unusually good; excellence.

"… before the war, shares with Rilke and Kafka the ***distinction*** of having origins which seem to escape …"

Example from: Kilgarriff, Adam. "Dictionary word sense distinctions: An enquiry into their nature." *Computers and the Humanities,* 26: 365–387, 1993. Definitions from *Longman Dictionary of Contemporary English*, 2nd edition, 1987.

What counts as the right answer?

- Therefore, hard to get a definitive sense-tagged corpus.
- And hard to get human baseline for performance.
  - Human annotators agree about 70–95% of the time.
    [Depending on word, sense inventory, context size, discussions, etc.]

Baseline algorithms 1

- Assume that input is PoS-tagged. Why?
- *Obvious baseline algorithm:*
  Pick most-likely sense (or pick one at random).
  - Accuracy: 39–62%

Baseline algorithms 2

- *Simple tricks (1):*
  Notice when ambiguous word is in unambiguous fixed phrase.
  - ***private*** *school*, ***private*** *eye*.
    (But maybe not *right* in *all right*.)

Baseline algorithms 3

- *Simple tricks (2):*
  "One sense per discourse":
  A homonymous word is rarely used in more than one sense in the same text.
  - If word occurs multiple times, …
  - Not true for polysemy.
- *Simple tricks (3):*
  Lesk's algorithm (see below).

“Context” 1

- Meaning of word in use depends on (determined by) its context.
  - Circumstantial context.
  - Textual context.
    - Complete text.
    - Sentence, paragraph.
    - Window of *n* words.

“Context” 2

- Words of context are also ambiguous; need for mutual constraints; often ignored in practice.
- “One sense per collocation”.
- Collocation: words that *tend* to co-occur together.

Selectional preferences

- Constraints imposed by one word meaning on another—especially verbs on nouns.

  *Eagle Airways which has applied to **serve** New York …*
  *Plain old bean soup, **served** daily since the turn of the century …*

  *I don't mind washing **dishes** now and then.*
  *Sprouted grains and seeds are used in preparing salads and **dishes** such as chop suey.*

  *It was the most popular **dish served** in the Ladies' Grill.*

- Some words select more strongly than others.
  see (weak) — *drink* (moderate) — *elapse* (strong)

Examples from the *Brown University Standard Corpus of Present-Day American English.*

Limitations of selectional preferences

- Negation:
  - *You can't **eat** good intentions.*
    *It's nonsense to say that a **book** elapsed.*
    *I am not a **crook**.* (Richard Nixon, 17 Nov 1973)
- Odd events:
  - *Los Angeles secretary Jannene Swift **married** a 50-pound pet rock in a formal ceremony in Lafayette Park.* (Newspaper report)

Limitations of selectional preferences

- Metaphor:

  *The issue was acute because the exiled Polish Government in London, supported in the main by Britain, was still competing with the new Lublin Government formed behind the Red Army. More time was spent in trying to **marry** these incompatibles than over any subject discussed at Yalta. … The application of these formulae could not please both sides, for they really attempted to **marry** the impossible to the inevitable.*

Text from the *Brown Corpus*

Limitations of selectional preferences

- In practice, attempts to induce selectional preferences or to use them have not been very successful.
  - Apply in only about 20% of cases, achieve about 50% accuracy. (Mihalcea 2006, McCarthy & Carroll 2003)
  - At best, they are a coarse filter for other methods.

Lesk's algorithm 1

- Sense s_i of ambiguous word *w* is likely to be the intended sense if many of the words used in the dictionary definition of s_i are also used in the definitions of words in the context window.
- For each sense s_i of *w*, let D_i be the bag of words in its dictionary definition.
- *Bag of words*: unordered set of words in a string, excepting those that are very frequent (*stop list*).
- Let *B* be the bag of words of the dictionary definitions of *all* senses of all words $v \neq w$ in the context window of *w*. (Might also (or instead) include all *v* in *B*.)
- Choose the sense s_i that maximizes *overlap*(D_i,*B*).

Lesk's algorithm Example

- *… the keyboard of the **terminal** was …*

terminal

1. a point on an electrical device at which electric current enters or leaves.

2. where transport vehicles load or unload passengers or goods.

3. an input-output device providing access to a **computer**.

keyboard

1. set of keys on a piano or organ or typewriter or typesetting machine or **computer** or the like.

2. an arrangement of hooks on which keys or locks are hung.

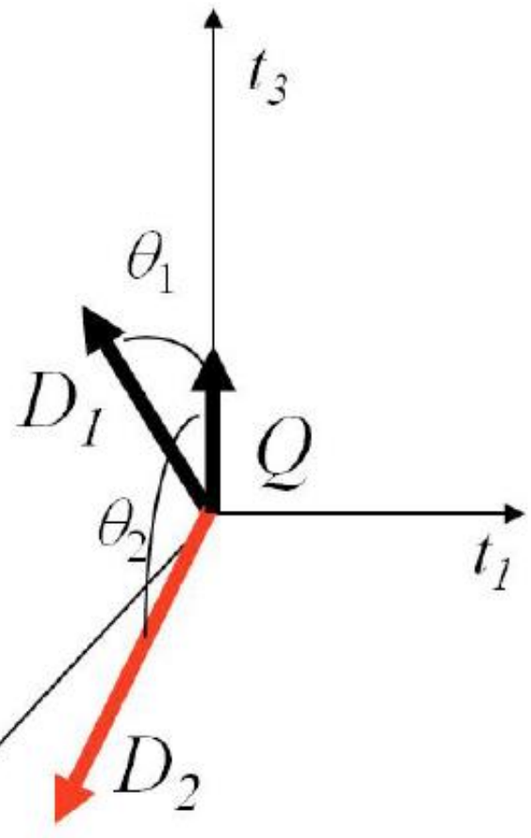
Lesk's algorithm 2

- Many variants of overlap score, but most common are based on cosine similarity of vectors that count occurrences of each word.
- ***Results:*** Simple versions of Lesk achieve accuracy around 50–60%; Lesk plus simple smarts gets to nearly 70%.
- Many variants possible on what is included in D_i and B.
  - *E.g.,* include the examples in dictionary definitions.
  - *E.g.,* include other manually tagged example texts.
  - PoS tags on definitions.
  - Give extra weight to infrequent words occurring in the vectors.

Cosine Similarity Score

- Cosine similarity measures the cosine of the angle between two vectors.
- Inner product normalized by the vector lengths.

$$\text{CosSim}(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$



$D_1 = 2T_1 + 3T_2 + 5T_3$ $\quad \text{CosSim}(D_1, Q) = 10 / \sqrt{(4+9+25)(0+0+4)} = 0.81$

$D_2 = 3T_1 + 7T_2 + 1T_3$ $\quad \text{CosSim}(D_2, Q) = 2 / \sqrt{(9+49+1)(0+0+4)} = 0.13$

$Q = 0T_1 + 0T_2 + 2T_3$

D_1 is 6 times better than D_2 using cosine similarity but only 5 times better using inner product.

Maths revision: Bayes's rule

- $$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$
- *Typical problem:* We have *B*, and want to know which *A* is now most likely.

$$\operatorname*{argmax}_A P(A \mid B) = \operatorname*{argmax}_A \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$= \operatorname*{argmax}_A P(B \mid A) \cdot P(A)$$

Supervised Bayesian methods 1

- Classify contexts according to which sense of each ambiguous word they tend to be associated with.
  - Bayes decision rule: Pick sense, s_j, that is most probable in given context, $j = \text{argmax}_i P(s_i \mid C)$.
- Bag-of-words model of context.
- For each sense s_k of w in the given context C, we know the ***prior probability*** $P(s_k)$ of the sense, but require its ***posterior probability*** $P(s_k|C)$.

Supervised Bayesian methods 2

- Want sense s' of word w in context C such that $P(s'|C) > P(s_k|C)$ for all $s_k \neq s'$.

$$
\begin{aligned}
s' &= \underset{s_k}{\operatorname{argmax}} P(s_k|C) \\
&= \underset{s_k}{\operatorname{argmax}} \frac{P(C|s_k)P(s_k)}{P(C)} \\
&= \underset{s_k}{\operatorname{argmax}} P(C|s_k)P(s_k) \\
&= \underset{s_k}{\operatorname{argmax}} P(s_k) \prod_{v_j \text{ in } C} P(v_j|s_k)
\end{aligned}
$$

where …

Supervised Bayesian methods 3

- ***Naïve Bayes assumption:*** Attributes v_j of context C of sense s_k of w are conditionally independent of one another. Hence

$$P(C|s_k) = \prod_{v_j \text{ in } C} P(v_j|s_k)$$

Supervised Bayesian methods 4

$$P(s_k) = \frac{c(s_k)}{c(w)}$$

$$P(v_j|s_k) = \frac{c(v_j,s_k)}{c(s_k)}$$

and $c(v_j, s_k)$ is the number of times v_j occurs in the context window of s_k.

Training corpora for supervised WSD

- ***Problem:*** Need ***large*** training corpus with each ambiguous word tagged with its sense.
  - Expensive, time-consuming human work.
  - "Large" for a human is small for WSD training.
- Some sense-tagged corpora:
  - SemCor: 700K PoS-tagged tokens (200K WordNet-sense-tagged) of Brown corpus and a short novel.
  - Singapore DSO corpus: About 200 "interesting" word-types tagged in about 2M tokens of Brown corpus and *Wall Street Journal*.

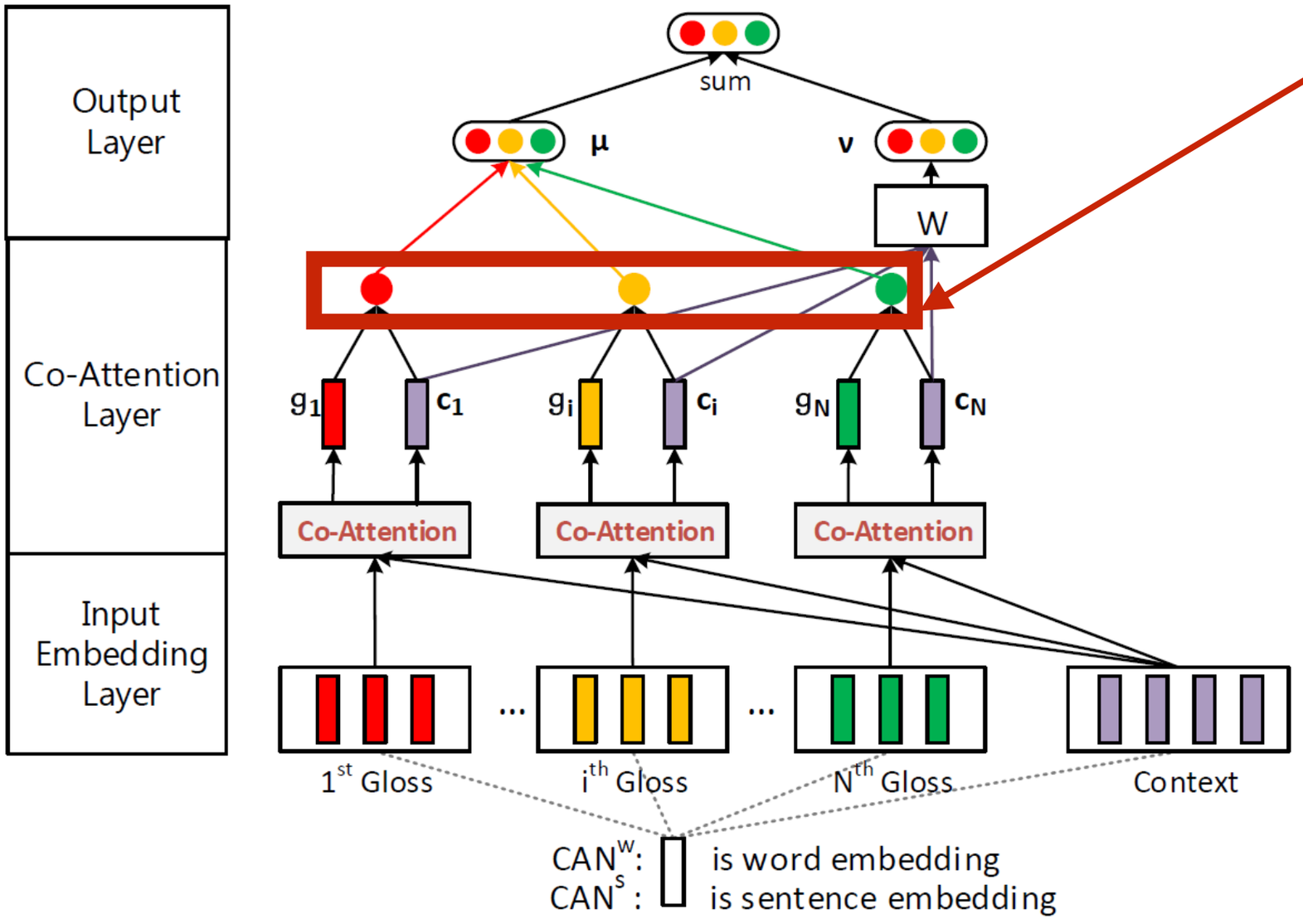
Evaluation

- Systems based on naïve Bayes methods have achieved 62–72% accuracy for selected words with adequate training data.
  (Màrquez etal 2006, Edmonds 2006)

Recent Neural Approaches

- There haven't been many successful ones, but the currently most influential is just an extension of Lesk's algorithm:
(Luo et al. 2018)
  - Instead of word counts, use lexical-semantic vector-space embeddings.
  - Use an attention mechanism to distort context-word vectors and dictionary-definition vectors with respect to each other.

Recent Neural Approaches



These dot-products measure the similarity between dictionary senses and context.

Evaluation

- Lesk's algorithm: 50-60%
- Naïve Bayes: 62–72%
- Neural Lesk: 65-72%
- Neural Lesk with "sentence" instead of word embeddings: 69-72%
- Neural hierarchical model with both: 68-73%

- Verbs are still particularly tough: 56-58%

Yarowsky 1995
Unsupervised decision-list learning

- ***Decision list:*** ordered list of strong, specific clues to senses of homonym.*

*Yarowsky calls them “polysemous words”.

Decision list for *bass*:

LogL	Context	Sense
10.98	*fish* in ±*k* words	FISH
10.92	*striped bass*	FISH
9.70	*guitar* in ±*k* words	MUSIC
9.20	*bass player*	MUSIC
9.10	*piano* in ±*k* words	MUSIC
8.87	*sea bass*	FISH
8.49	*play bass*	MUSIC
8.31	*river* in ±*k* words	FISH
7.71	*on bass*	MUSIC
5.32	*bass are*	FISH

Yarowsky 1995 Basic ideas

- Separate decision list learned for each homonym.
- Bootstrapped from seeds, very large corpus, heuristics.
  - One sense per ***discourse***.
  - One sense per ***collocation***.
- Uses supervised classification algorithm to build decision-list.
- Training corpus: 460M words, mixed texts.

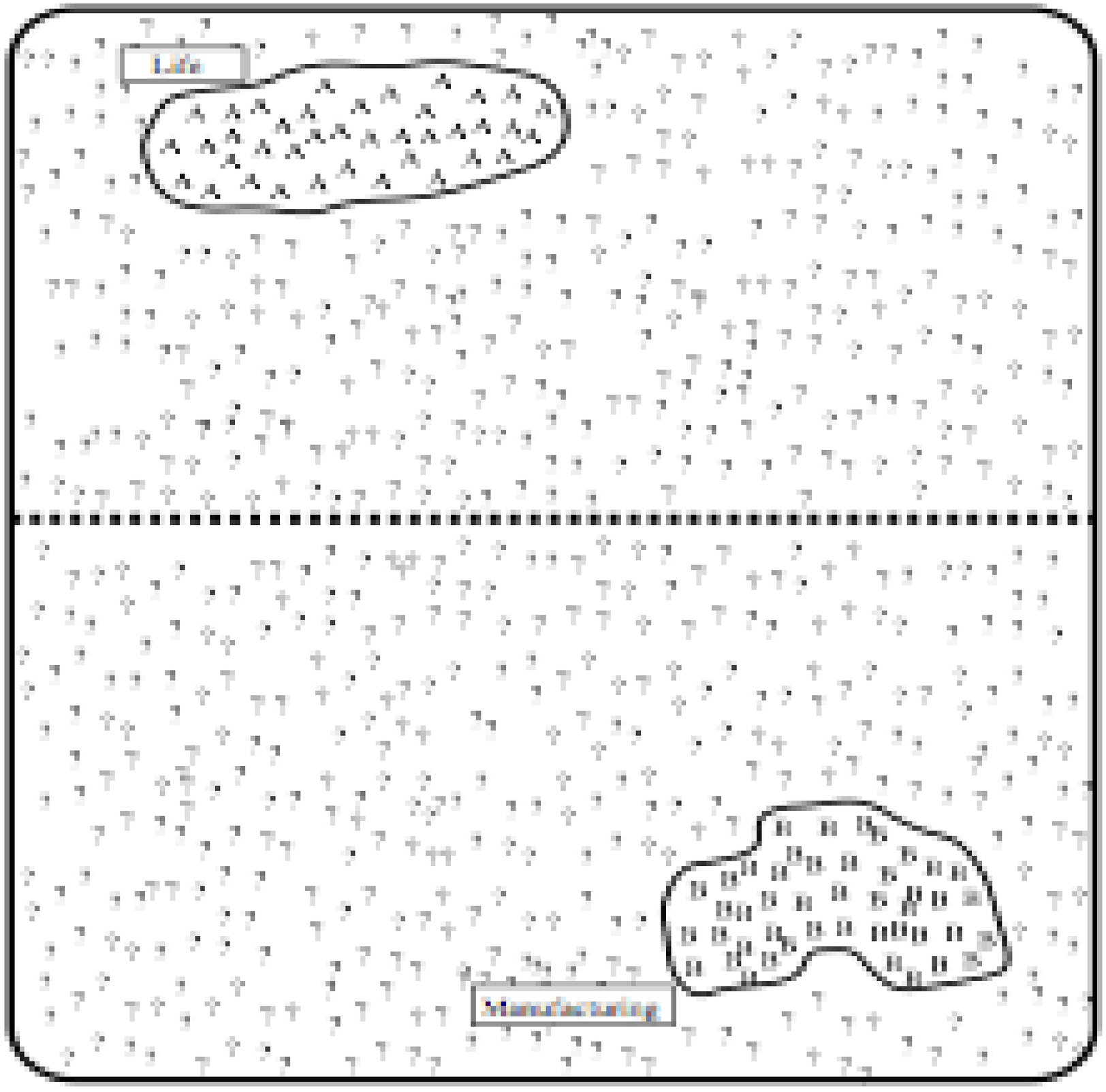
Yarowsky 1995 Method 1

- 1–2. Get data (instances of target word); choose seed rules; apply them.

used to strain microscopic plant life from the
zonal distribution of plant life .
close-up studies of plant life and natural
too rapid growth of aquatic plant life in water
the proliferation of plant and animal life
establishment phase of the plant virus life cycle
that divide life into plant and animal kingdom
many dangers to plant and animal life
mammals . Animal and plant life are delicately

automated manufacturing plant in Fremont
vast manufacturing plant and distribution
chemical manufacturing plant , producing viscose
keep a manufacturing plant profitable without
computer manufacturing plant and adjacent
discovered at a St. Louis plant manufacturing
copper manufacturing plant found that they
copper wire manufacturing plant , for example
s cement manufacturing plant in Alpena

vinyl chloride monomer plant , which is
molecules found in plant and animal tissue
Nissan car and truck plant in Japan is
and Golgi apparatus of plant and animal cells
union responses to plant closures .
cell types found in the plant kingdom are
company said the plant is still operating
Although thousands of plant and animal species
animal rather than plant tissues can be



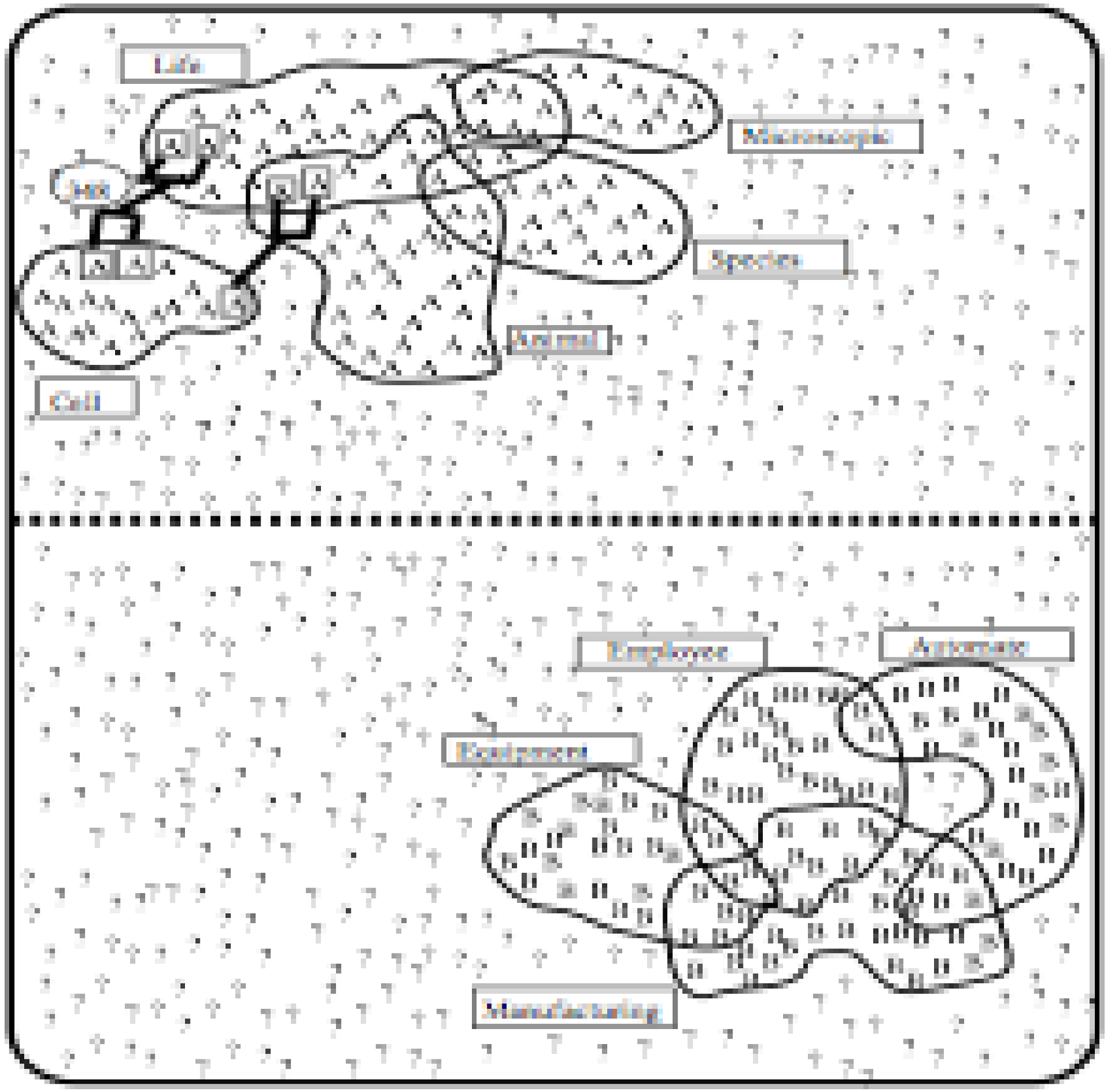
Initial state after use of seed rules

Figure from Yarowsky 1995.

Yarowsky 1995 Method 2

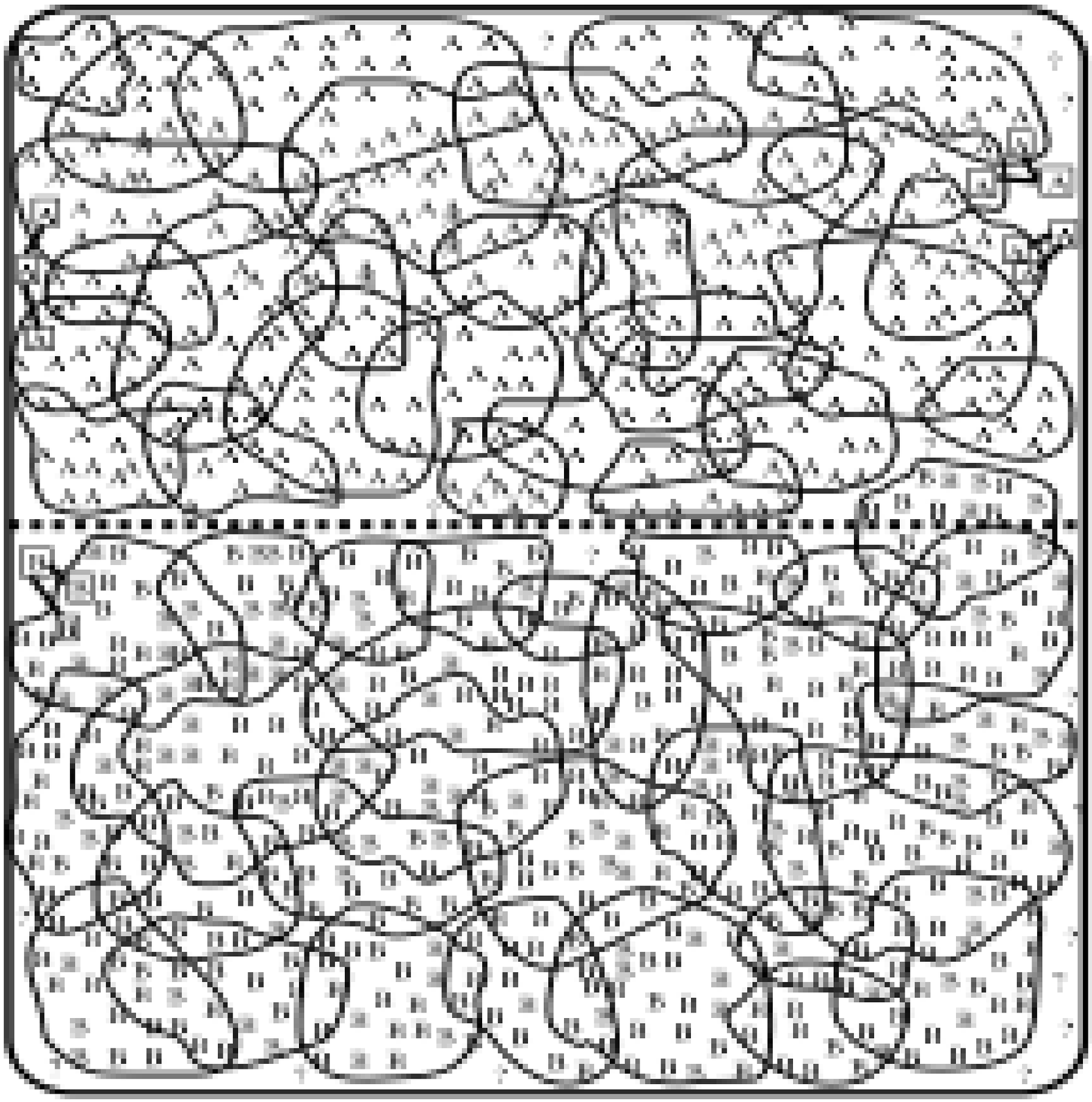
- 3. Iterate:
  - 3a. Create a new decision-list classifier: supervised training with the data tagged so far.

    Looks for collocations as features for classification.
  - 3b. Apply new classifier to whole data set, tag some new instances.
  - 3c. Optional: Apply one-sense-per-discourse rule wherever one sense now dominates a text.



Intermediate state

Figure from Yarowsky 1995.



Final state

Figure from Yarowsky 1995.

Yarowsky 1995: Method 3

- 4. Stop when converged. (Optional: Apply one-sense-per-discourse constraint.)
- 5. Use final decision list for WSD.

Yarowsky 1995 Evaluation

- Experiments: 12 homonymous words.
  - 400–12,000 hand-tagged instances of each.
  - Baseline (most frequent sense) = 63.9%.
- Best results, avg 96.5% accuracy.
  - Base seed on dictionary definition; use one-sense-per-discourse heuristic.
  - As good as or better than supervised algorithm used directly on fully labelled data.

Yarowsky 1995 Discussion 1

- Strength of method:
  - The one-sense heuristics.
  - Use of precise lexical and positional information.
  - Huge training corpus.
  - Bootstrapping: Unsupervised use of supervised algorithm.
- Disadvantages:
  - Train each word separately.
  - Homonyms only. Why?

Yarowsky 1995 Discussion 2

- Not limited to regular words; *e.g.,* in speech synthesis system:
  - / as fraction or date:
    3/4 → “three-quarters” or “third of April”.
  - Roman number as cardinal or ordinal:
    chapter VII → “chapter seven”;
    Henry VII → “Henry the seventh”.

Yarowsky, David. “Homograph disambiguation in speech synthesis.” In Jan van Santen, Richard Sproat, Joseph Olive and Julia Hirschberg (eds.), *Progress in Speech Synthesis*. Springer-Verlag, pp. 159–175, 1996.