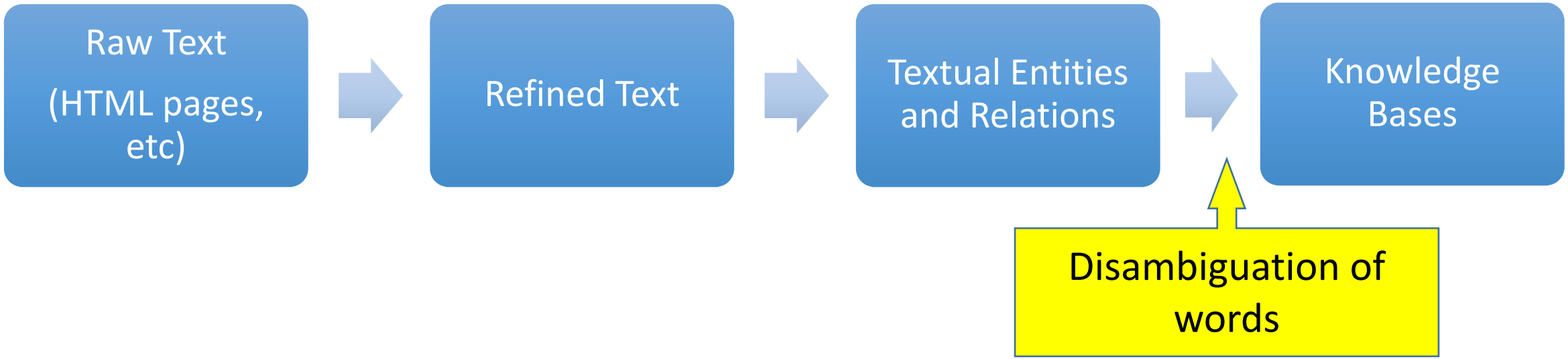
What is knowledge acquisition?

Knowledge acquisition: process to extract knowledge (to be integrated into knowledge bases) from unstructured text or other data



Word Sense Disambiguation

Word Sense Disambiguation (WSD)

- Given
  - A word in context
  - A fixed inventory of potential word senses
  - Decide which sense of the word this is
- Why? Machine translation, QA, speech synthesis
- What set of senses?
  - In general: the senses in a thesaurus like WordNet

Two variants of WSD task

- Lexical Sample task
  - Small pre-selected set of target words (*line, plant*)
  - And inventory of senses for each word
  - Supervised machine learning: train a classifier for each word
- All-words task
  - Every word in an entire text
  - A lexicon with senses for each word
  - Data sparseness: can't train word-specific classifiers

WSD Methods

- Supervised Methods
- Thesaurus/Dictionary Methods
- Semi-Supervised Learning

Supervised Methods

Supervised Machine Learning Approaches

- Supervised machine learning approach:
  - a training corpus of words tagged in context with their sense
  - used to train a classifier that can tag words in new text
- Summary of what we need:
  - the **tag set** (“sense inventory”)
  - the **training corpus**
  - A set of **features** extracted from the training corpus
  - A **classifier**

Supervised WSD 1: WSD Tags

- A tag = dictionary sense
- For example, for WordNet an instance of “bass” in a text has 8 possible tags or labels (bass^{1} through bass^{8})
- For instance, given the sentence:

> The curtain rises to the sound of angry dogs baying and ominous **bass** chord sounding.

- What is the sense of “bass”?

8 senses of “bass” in WordNet

1. bass - (the lowest part of the musical range)
2. bass, bass part - (the lowest part in polyphonic music)
3. bass, basso - (an adult male singer with the lowest voice)
4. sea bass, bass - (flesh of lean-fleshed saltwater fish of the family Serranidae)
5. freshwater bass, bass - (any of various North American lean-fleshed freshwater fishes especially of the genus Micropterus)
6. bass, bass voice, basso - (the lowest adult male singing voice)
7. bass - (the member with the lowest range of a family of musical instruments)
8. bass - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Supervised WSD 2: Get a corpus

- Lexical sample task:
  - *Line-hard-serve* corpus - 4000 examples of each
  - *Interest* corpus - 2369 sense-tagged examples
- All words:
  - **Semantic concordance**: a corpus in which each open-class word is labeled with a sense from a specific dictionary/thesaurus.
    - SemCor: 234,000 words from Brown Corpus (1961), manually tagged with WordNet senses
    - SENSEVAL-3 competition corpora - 2081 tagged word tokens

SemCor

<wf pos=PRP>**He**</wf>

<wf pos=VB lemma=recognize wnsn=4 lexsn=2:31:00::>**recognized**</wf>

<wf pos=DT>**the**</wf>

<wf pos=NN lemma=gesture wnsn=1 lexsn=1:04:00::>**gesture**</wf>

<punc>.</punc>

VU VRIJE UNIVERSITEIT AMSTERDAM

Supervised WSD 3: Extract feature vectors Intuition from Warren Weaver (1955):

“If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words…

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word…

The practical question is : ``What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?”

Two kinds of features in the vectors

- If we want to train a model, then we must represent each word as a vector of features
- **Collocational** features and **bag-of-words** features
  - **Collocational**
    - Features about words at **specific** positions near target word
      - Often limited to just word and POS
  - **Bag-of-words**
    - Features about words that occur anywhere in the window (regardless of position)
      - Typically limited to frequency counts

Examples

Collocational features

- Example text (WSJ):

  An electric guitar and **bass** player stand off to one side not really part of the scene

- Assume a window of +/- 2 from the target

Collocational features

- word 1,2,3 grams in window of ±3 is common

$$[w_{i-2}, \text{POS}_{i-2}, w_{i-1}, \text{POS}_{i-1}, w_{i+1}, \text{POS}_{i+1}, w_{i+2}, \text{POS}_{i+2}, w_{i-2}^{i-1}, w_{i}^{i+1}]$$

```
[guitar, NN, and, CC, player, NN, stand, VB, and guitar, player stand]
```

Bag-of-words features

- “an unordered set of words” – position ignored
- Counts of words occur within the window
- First choose a vocabulary. Typically it is some pre-labeled corpus
- Then count how often each of those terms occurs in a given window
  - sometimes just a binary “indicator” 1 or 0

Co-Occurrence Example

- Assume we've settled on a possible vocabulary of 12 words in "bass" sentences:

[*fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band*]

- The vector for:

  guitar and bass player stand

  [0,0,0,1,0,0,0,0,0,0,1,0]

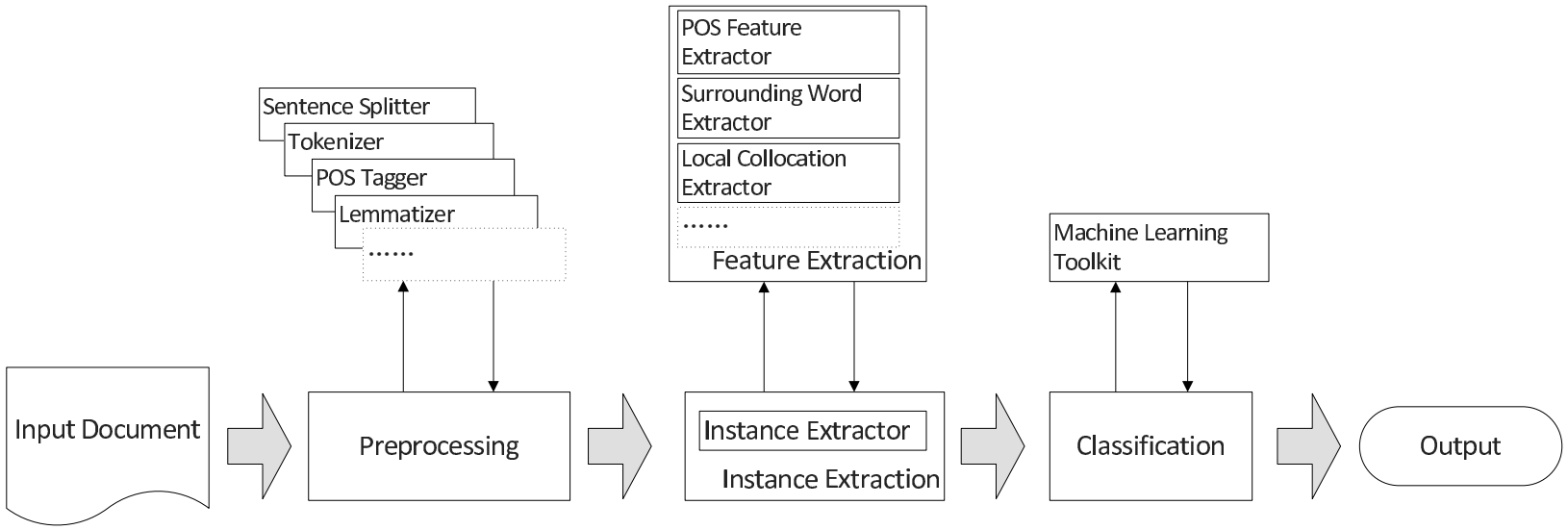
Classification Methods: Supervised Machine Learning

- *Input:*
  - a word *w in a text window d (which we'll call a "document")*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled text windows again called "documents" $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output:*
  - a learned classifier $\gamma: d \rightarrow c$

VU VRIJE UNIVERSITEIT AMSTERDAM

IMS

- IMS (It makes sense) [1] is the most popular supervised algorithm
  - Allows the usage of different features
- Uses a SVM classifier



VU VRIJE UNIVERSITEIT AMSTERDAM

[1] Z. Zhong and H. T. Ng, “It Makes Sense: A Wide-coverage Word Sense Disambiguation System for Free Text,” in *ACL*, 2010, pp. 78–83.

WSD Evaluations and baselines

- Best evaluation: **extrinsic (‘end-to-end’, `task-based’) evaluation**
  - Embed WSD algorithm in a task and see if you can do the task better!
- What we often do for convenience: **intrinsic evaluation**
  - Exact match **sense accuracy**
    - % of words tagged identically with the human-manual sense tags
  - Usually evaluate using **held-out data** from same labeled corpus
- Baselines
  - Most frequent sense

Most Frequent Sense

- WordNet senses are ordered in frequency order
- So "most frequent sense" in WordNet = "take the first sense"
- Sense frequencies come from the *SemCor* corpus

Freq	Synset	Gloss
338	plant[1], works, industrial plant	buildings for carrying on industrial labor
207	plant[2], flora, plant life	a living organism lacking the power of locomotion
2	plant[3]	something planted secretly for discovery by another
0	plant[4]	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

Dictionary and Thesaurus Methods

VU VRIJE UNIVERSITEIT AMSTERDAM

The Simplified Lesk algorithm

- Let's disambiguate "**bank**" in this sentence:

  The **bank** can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

- given the following two WordNet senses:

bank[1]	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank[2]	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

The Simplified Lesk algorithm

Choose sense with most word overlap between gloss and context
(not counting function words)

The **bank** can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

bank[1]	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank[2]	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

The Corpus Lesk algorithm

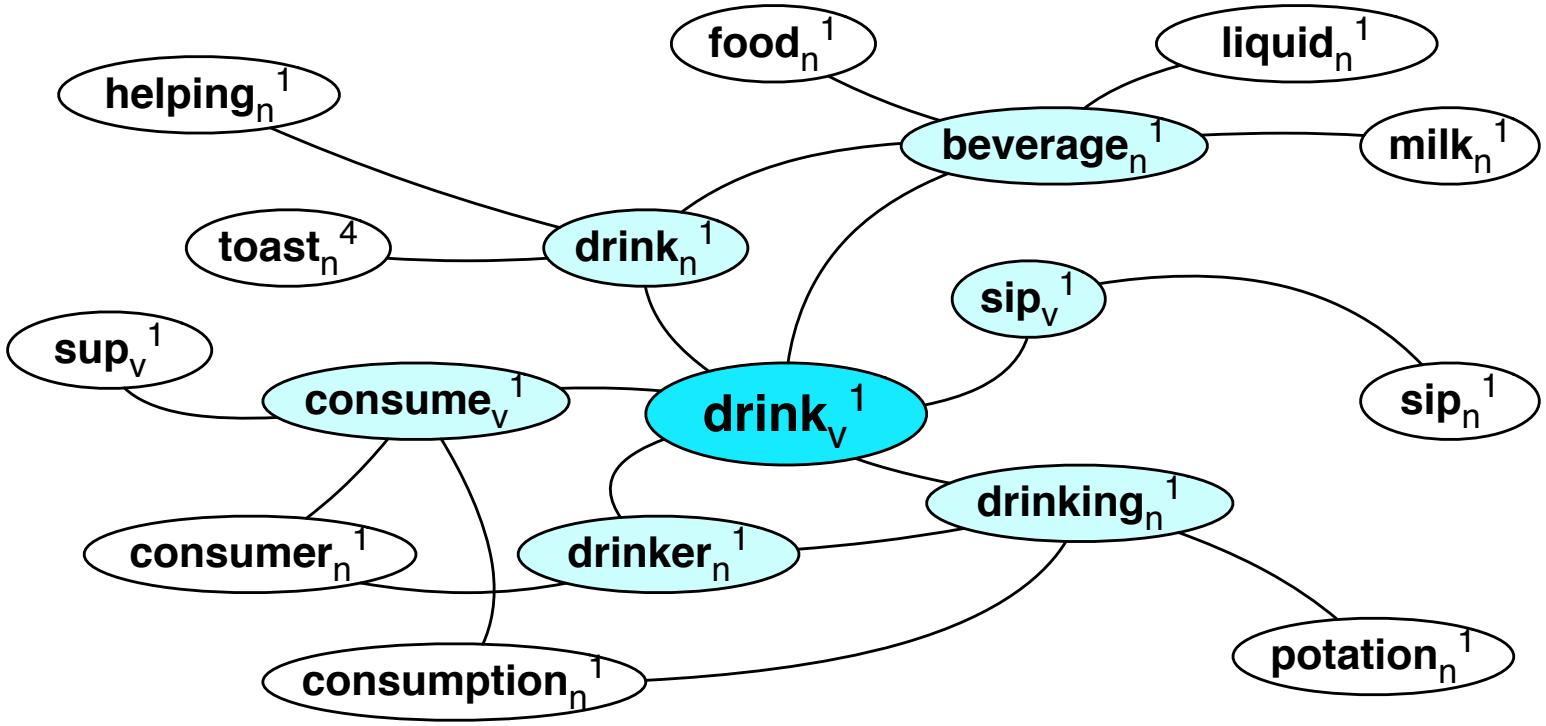
- Assumes we have some sense-labeled data (like SemCor)
- Take all the sentences with the relevant word sense:

  *These short, "streamlined" meetings usually are sponsored by local **banks**[1], Chambers of Commerce, trade associations, or other civic organizations.*

- Now add these to the gloss + examples for each sense, call it the “signature” of a sense.
- Choose sense with most word overlap between context and signature.

Graph-based methods

- First, WordNet can be viewed as a graph
  - senses are nodes
  - relations (hypernymy, meronymy) are edges
  - Also add edge between word and unambiguous gloss words



How to use the graph for WSD

- Insert target word and words in its sentential context into the graph, with directed edges to their senses

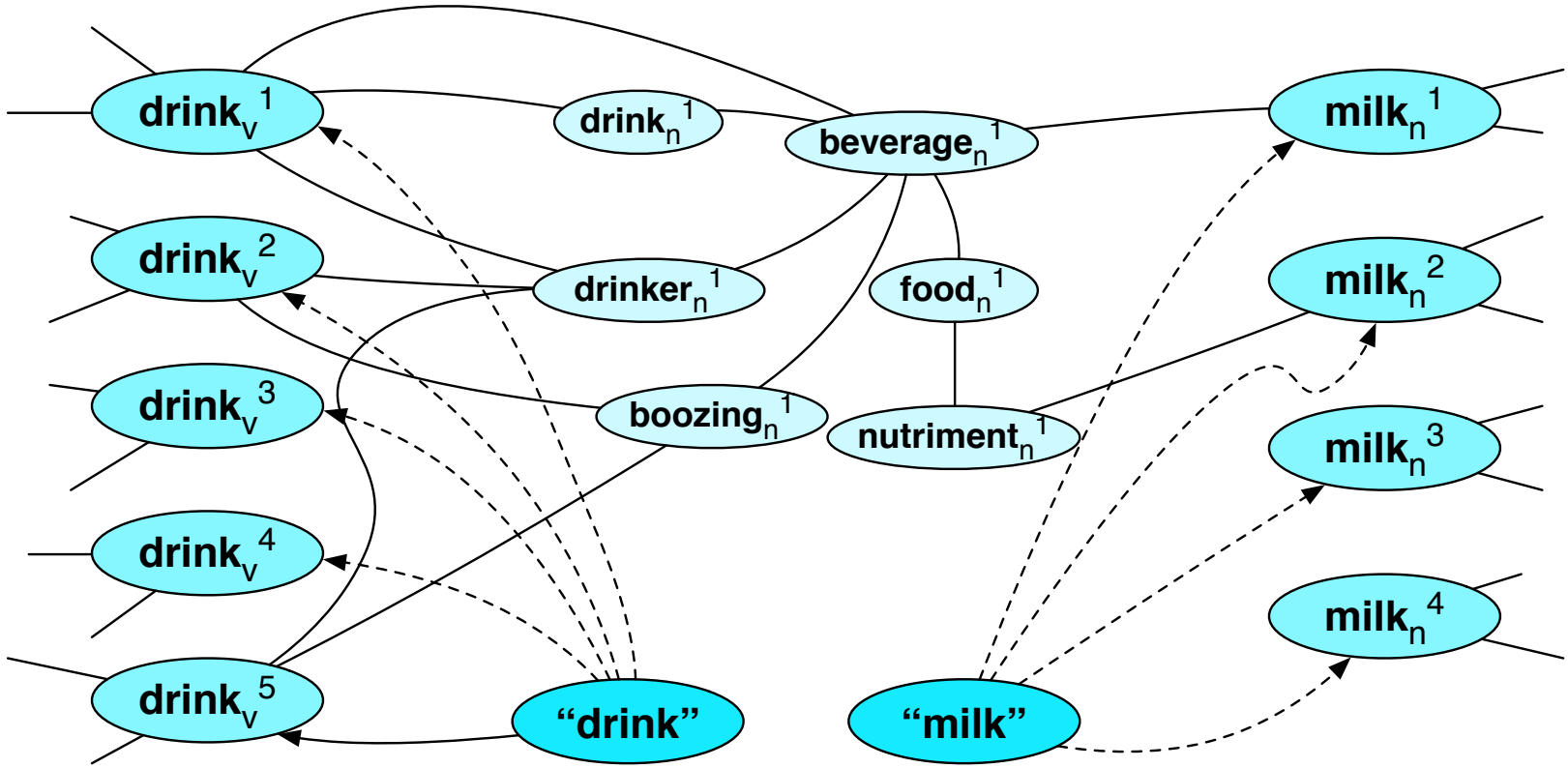
"She drank some milk"

- Now choose the

  most central sense

Pick nodes with

highest "pagerank" or similar measures



VU VRIJE UNIVERSITEIT AMSTERDAM

Semi-supervised Methods

Semi-Supervised Learning

Problem: supervised and dictionary-based approaches require large hand-built resources. What if you don't have so much training data?

Solution: Bootstrapping. Generalize from a very small hand-labeled seed-set

Yarosky's algorithm: Learns a classifier for a target word. Uses bootstrapping to create more training examples

- Intuition: Let's take the word bass
  - the word `play` occurs with the music sense of bass
  - the word `fish` occurs with the fish sense of bass

Sentences extracting using “fish” and “play”

We need more good teachers – right now, there are only a half a dozen who can **play** the free **bass** with ease. An electric guitar and **bass play**er stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.
The researchers said the worms spend part of their life cycle in such **fish** as Pacific salmon and striped **bass** and Pacific rockfish or snapper. And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny.

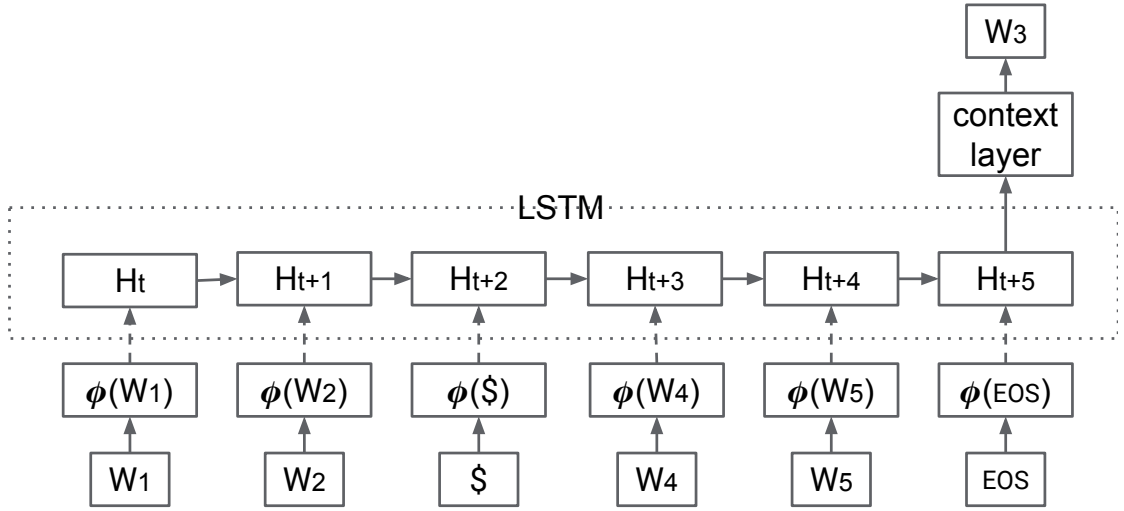
Generating seeds

1) “One sense per collocation”:
   - A word reoccurring in collocation with the same word will almost surely have the same sense.
2) “One sense per discourse”:
   - The sense of a word is highly consistent within a document
   - (At least for non-function words, and especially topic-specific words)

The Google's way

- Recently Google proposed a new method to perform WSD using deep neural networks

1. Use a LSTM to create context embeddings
2. Use manually annotated data to construct *sense embeddings.* They are the **average** of all context vectors where the sense appears
3. Given a sentence, calculate the embedding of the target word, then compare it to all sense embeddings (cosine similarity)



VU VRIJE UNIVERSITEIT AMSTERDAM

[1] Semi-supervised Word Sense Disambiguation with Neural Models. Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, Eric Altendorf. COLING 2016. LINK

The Google's way

	Senseval2		Senseval3		SemEval7		SemEval7-Coarse		SemEval13
model	**all**	**n.**	**all**	**n.**	**all**	**n.**	**all**	**n.**	**n.**
IMS + Word2Vec (T:SemCor)	0.634	0.742	0.653	0.701	0.578	0.686			
IMS + Word2Vec (T:OMSTI)	0.683	0.777	0.682	0.741	0.591	0.715			
Taghipour and Ng (2015b)			0.682						
Chen et al. (2014)							0.826	**0.853**	
Weissenborn et al. (2015)				0.688		0.660		**0.855**	**0.728**
Word2Vec (T:SemCor)	0.678	0.737	0.621	0.714	0.585	0.673	0.795	0.814	0.661
LSTM (T:SemCor)	0.736	0.786	0.692	0.723	0.642	**0.723**	0.828	0.834	0.670
LSTM (T:OMSTI)	0.724	0.777	0.643	0.680	0.607	0.673	0.811	0.820	0.673
LSTMLP (T:SemCor, U:OMSTI)	0.739	0.797	0.711	0.748	**0.637**	0.704	**0.843**	0.834	0.679
LSTMLP (T:SemCor, U:1K)	0.738	0.796	**0.718**	**0.763**	0.635	0.717	0.836	0.831	0.695
LSTMLP (T:OMSTI, U:1K)	**0.744**	**0.799**	0.710	0.753	0.633	0.717	0.833	0.825	0.681

Non reproducible

ctors trained using **100 billion tokens** and a vocabulary of **1M** words.

VU VRIJE UNIVERSITEIT AMSTERDAM

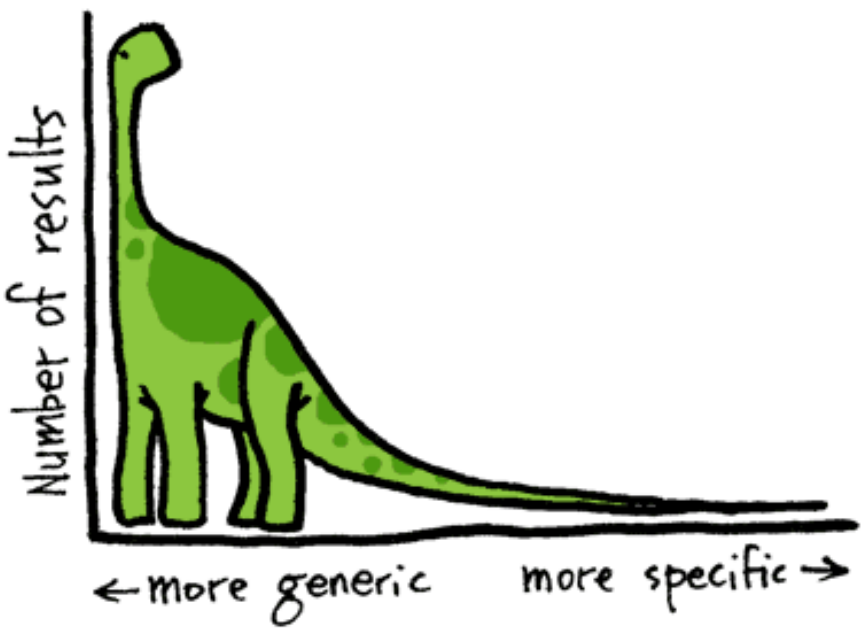
[1] Semi-supervised Word Sense Disambiguation with Neural Models. Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, Eric Altendorf. COLING 2016. LINK

General problem of WSD

- In general, WSD is hard!
- Human inter-annotator agreement
  - Compare annotations of two humans, Given same tagging guidelines
- Human agreements on all-words corpora with WordNet style senses
  - 75%-80%

Performance state-of-the-art

- State-of-the-art reaches 60-80% (depends on the task and repo.) [1]
  - IMS and Personalized Pagerank are the most popular ones
- No existing method can handle the “Long tail problem”



VU VRIJE UNIVERSITEIT AMSTERDAM

[1] R. Navigli, “Word Sense Disambiguation: A Survey,” *ACM Comput. Surv.*, vol. 41, no. 2, p. 10:1–10:69, Feb. 2009.