Mitigating Feature Exclusion to Improve Hypernymy Recognition

Max Lotstein

PROBLEM:

Find more pairs of senses like <u, v>, such that u is a kind of v

For example, dog_1 is a kind of *$animal_1$*

PROBLEM:

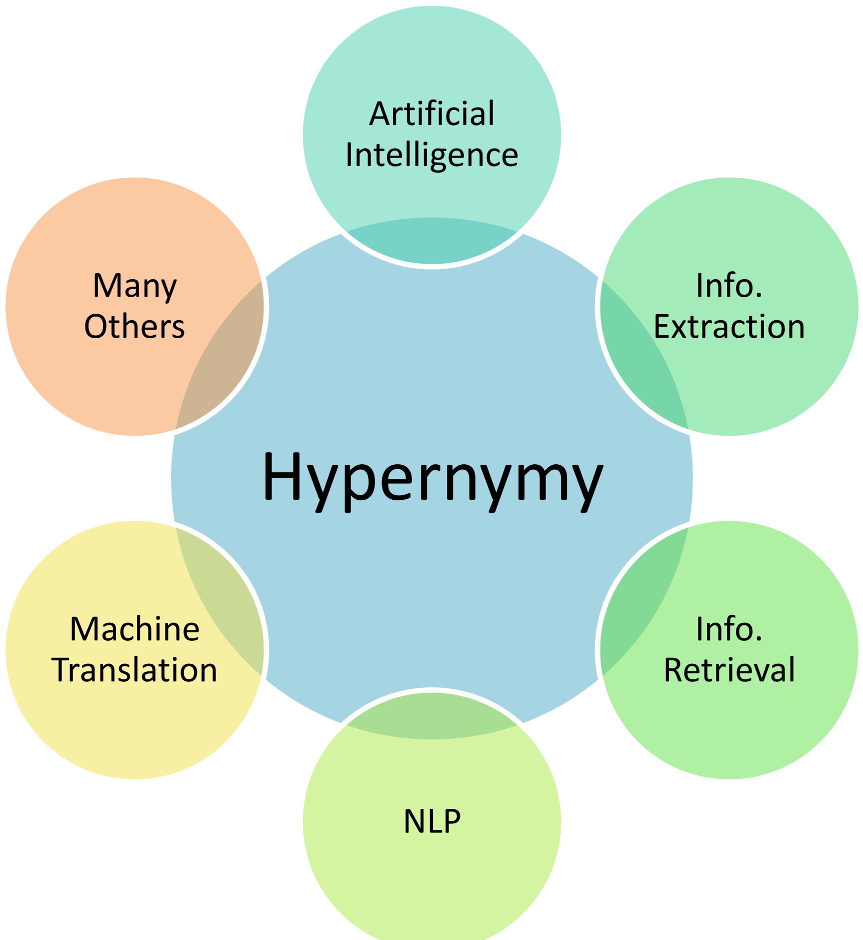
Find more pairs of sense like <dog_1, $animal_1$>

PROBLEM:

Find more pairs of sense like <dog$_1$, animal$_1$>

hyponym hypernym

Hmmm… So, what?

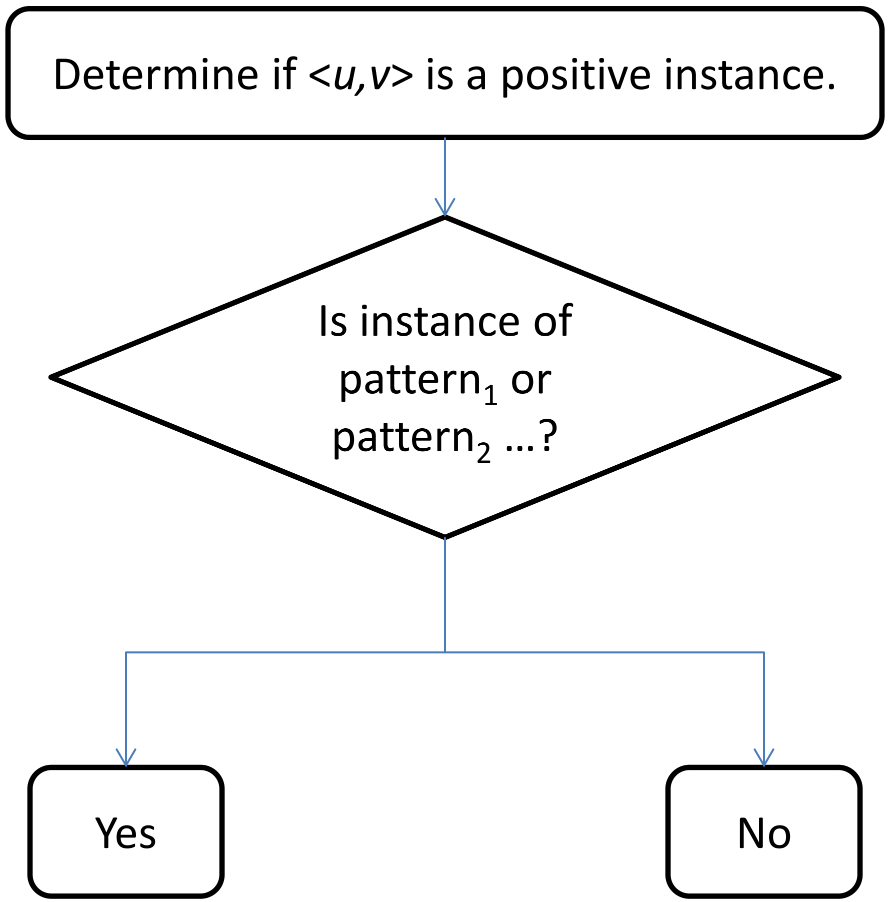

Artificial Intelligence
Info. Extraction
Many Others
Hypernymy
Machine Translation
Info. Retrieval
NLP

Idea 1: Use shallow patterns

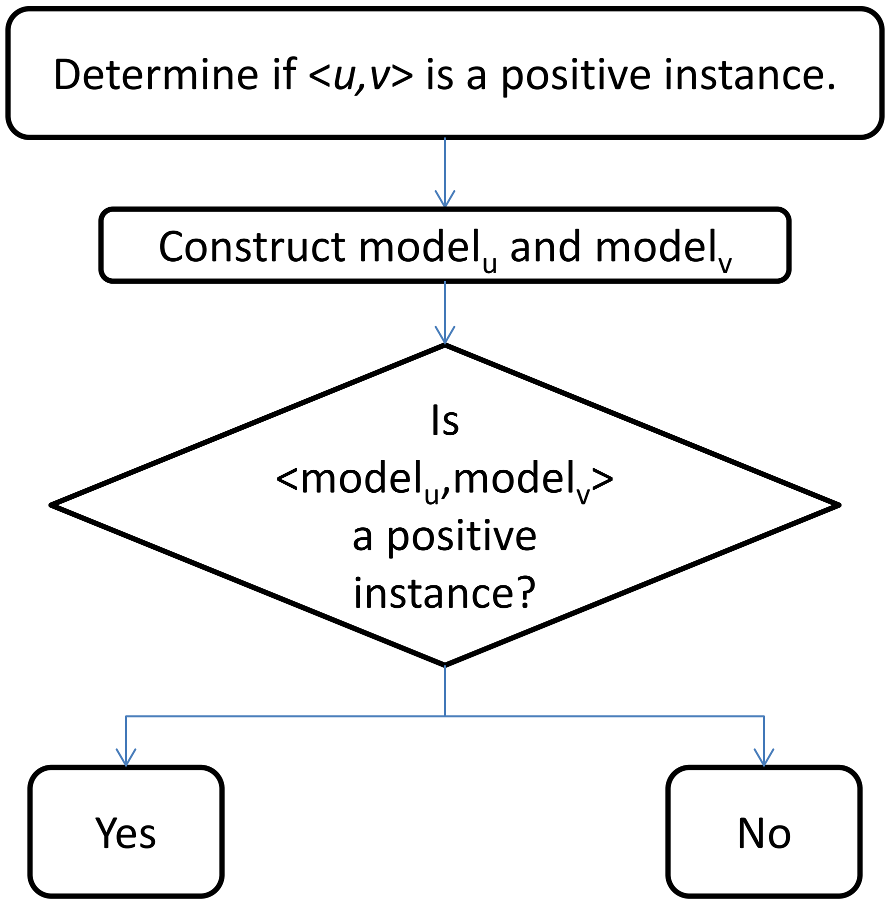
Relation	Example
NP_0 *such as* $\{NP_1, NP_2 \ldots, (and \mid or)\}\ NP_n$	"The bow lute, such as the Bambara ndang, ..."
such NP as $\{NP\ ,\}^*\ \{(or \mid and)\}\ NP$	"... works by such authors as Herrick, Goldsmith, and Shakespeare."
$NP\ \{, NP\}^*\ \{,\}$ *or other NP*	"Bruises, wounds, broken bones or other injuries ..."
$NP\ \{, NP\}^*\ \{,\}$ *and other NP*	"... temples, treasuries, and other important civic buildings."
$NP\ \{,\}$ *including* $\{NP\ ,\}^*\ \{or \mid and\}\ NP$	"All common-law countries, including Canada and England ..."
$NP\ \{,\}$ *especially* $\{NP\ ,\}^*\ \{or \mid and\}\ NP$	"... most European countries, especially France, England, and Spain."

Hearst, M. a. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora.
Herger, P. (2014). Learning semantic relations with distributional similarity.

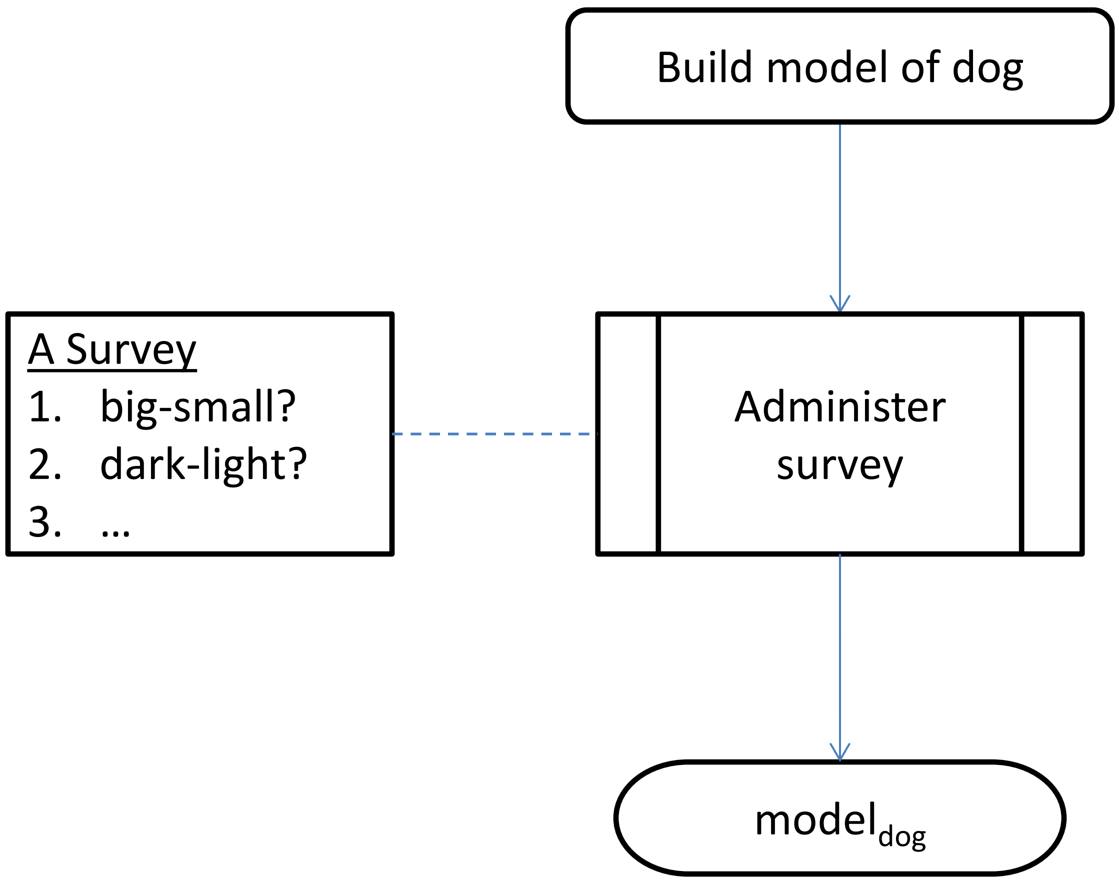
Shallow Patterns



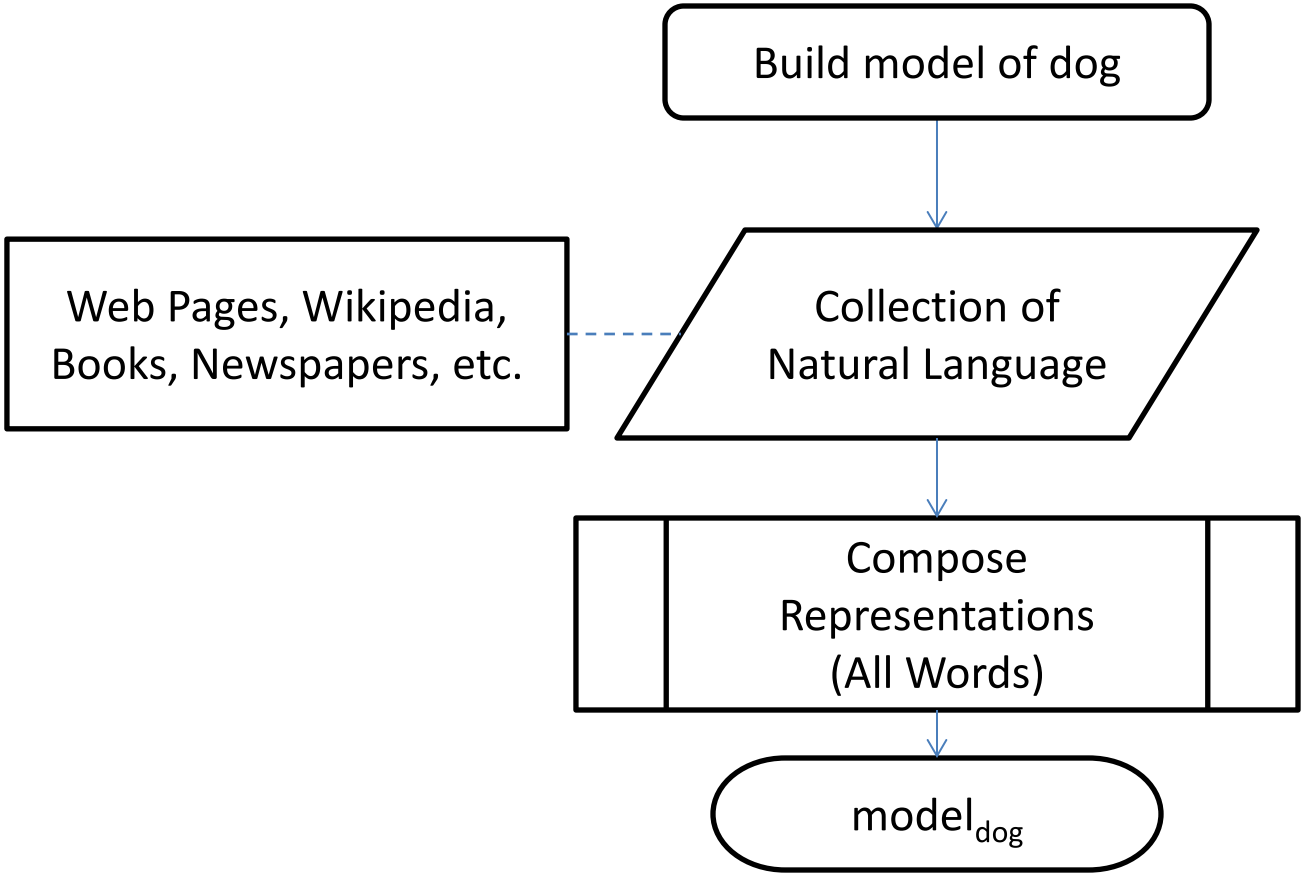
Idea 2: Modeling Word Meaning



Modeling Meaning: Psychology



Modeling Meaning: Distributional Semantics



Contextual Clues to Meaning

He filled the **wampimuk**, passed it around and we all drunk some.

We found a little, hairy **wampimuk** sleeping behind the tree.

McDonald, S & Ramscar, M (2001). [Testing the Distributional Hypothesis: The Influence of Context on Judgements of Semantic Similarity.]

Distributional Hypothesis

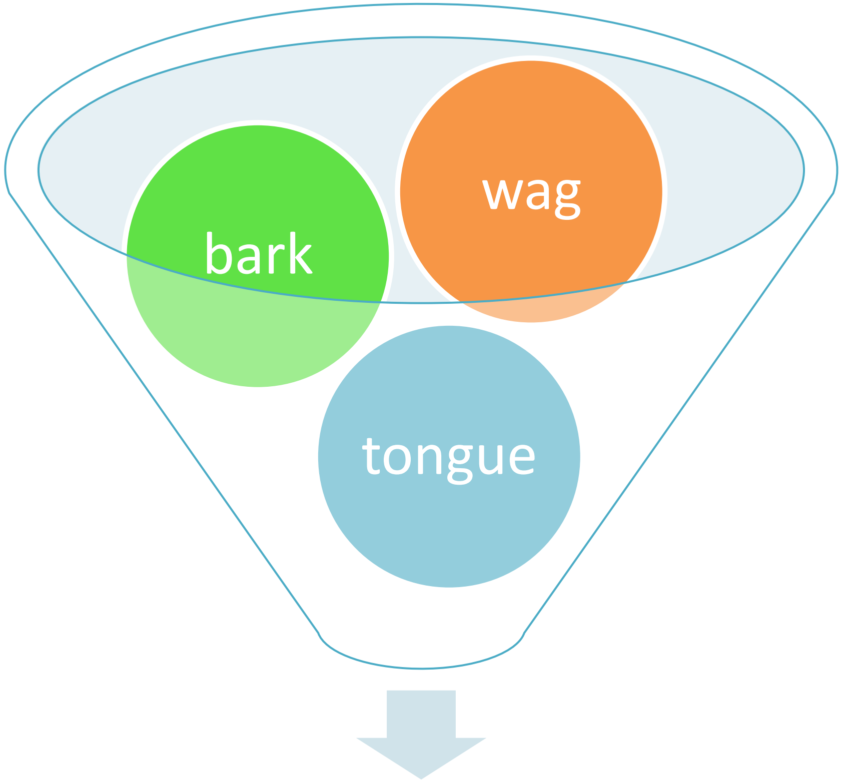
"You shall know a word by the company it keeps!"

J.R. Firth. A synopsis of linguistic theory 1930-55.

"The degree of semantic similarity between two words (or other linguistic terms) can be modeled as a function of the degree of overlap among their linguistic contexts."

M. Baroni and A. Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics

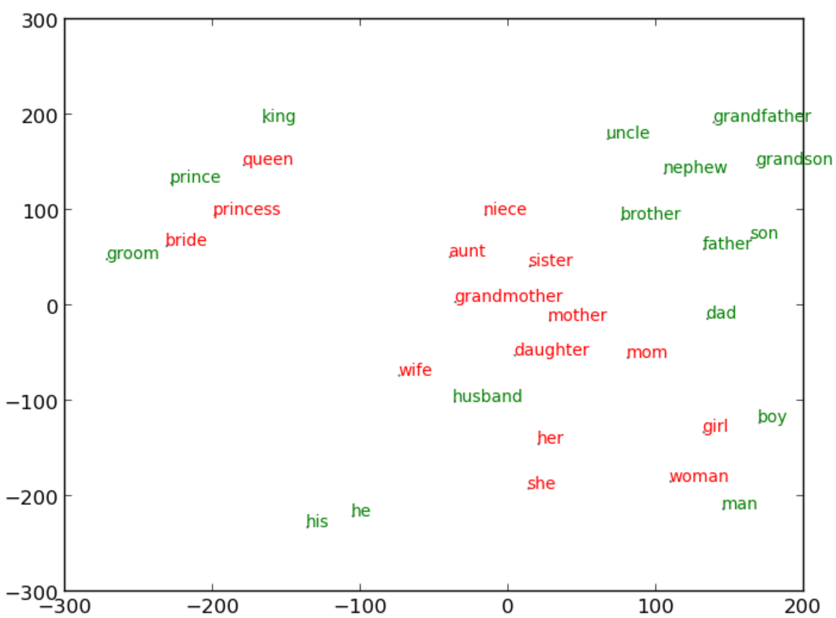
For example



$\text{model}_{\text{dog}}$

Sparse Feature-Vector Models

Geometric Metaphor of Meaning



http://wordvectors.org/demo.php

Sahlgren, M. (2006).

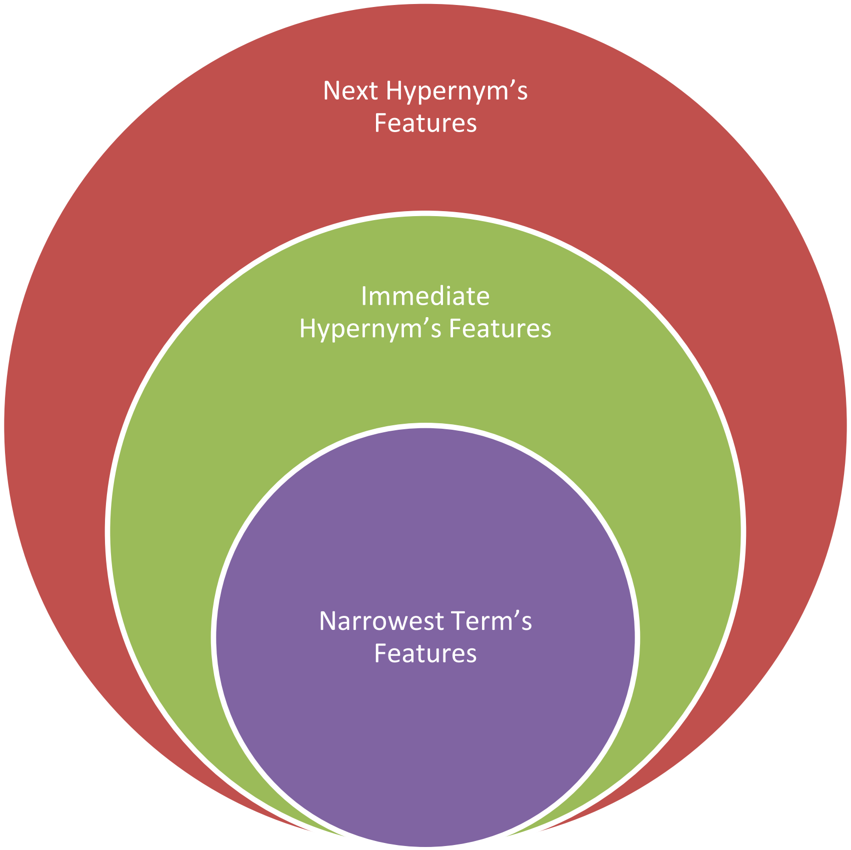
Sparse Representation

$$
\begin{array}{l}
 \\
dog \\
animal \\
cup \\
\dots
\end{array}
\begin{array}{c}
\begin{array}{ccccc} wag & bark & fly & fill & \dots \end{array} \\
\begin{bmatrix} 5 & 4 & & & \\ 3 & 2 & 1 & & \\ & & & 4 & \\ & & & & \end{bmatrix}
\end{array}
$$

$$\vec{v} \in S_{R \times C}$$

$$\vec{v} \mapsto \{f_1, f_2, \dots\} \Big| \left|\{f_1, f_2, \dots\}\right| \ll C$$

Feature Inclusion



<u>Substitutability</u>

The dog barked.

The animal barked.

The animal flew.

The dog flew.*

Distributional Inclusion Hypothesis (DIH)

Given words $\langle A, B \rangle$, and $f(w)$, a function that determines for a sense its most important features, and that $A \rightarrow B$ denotes that A entails B, then $A \rightarrow B \equiv f(A) \subset f(B)$.

Maayan Geffet and Ido Dagan. The distributional inclusion hypotheses and lexical entailment. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL ’05

Feature Inclusion As Precision

$$F(\vec{u}) = \{u_1, u_2, u_3, \ldots u_n\}$$

$$w(f \in F(u), u) = PPMI(u_f)$$

$$\mathrm{P}_{\mathrm{Weeds}}(\vec{u}, \vec{v}) = \frac{\sum\limits_{f \in F(u) \cap F(v)} w(f, u)}{\sum\limits_{f \in F(u)} w(f, u)}$$

Weeds, Julie, et al. "Learning to distinguish hypernyms and co-hyponyms." *Proceedings of COLING*. 2014.

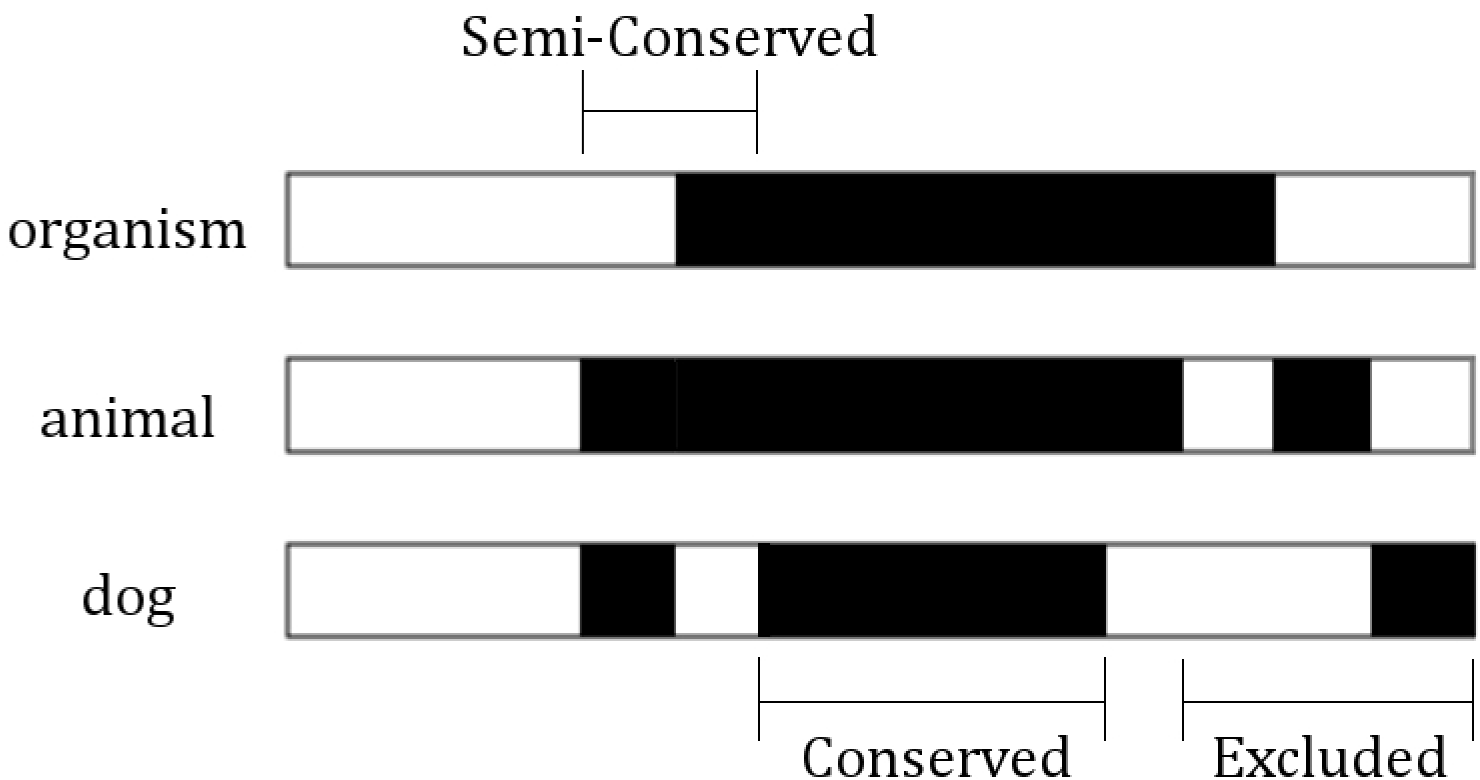
State-of-the-Art Performance

Accuracy for Model vs. Dataset

	Baseline Models					DIH-based Models		
dataset	most freq	cosineP	linP	widthdiff	singlewidth	CRdiff	invCLP	balAPincP
$hyponym_{\mathrm{BLESS}}$	0.54	0.53	0.54	0.56	0.58	0.52	0.54	0.54
$cohyponym_{\mathrm{BLESS}}$	0.61	**0.79**	**0.78**	-	-	-	-	-
$hyponym_{\mathrm{WN}}$	0.50	0.53	0.52	0.70	0.65	0.70	0.66	0.53
$cohyponym_{\mathrm{WN}}$	0.50	0.50	0.55	-	-	-	-	-

Weeds, Julie, et al. "Learning to distinguish hypernyms and co-hyponyms." *Proceedings of COLING*. 2014.

Feature Exclusion Problem



How Bad Is It?

Percentage of Features for Degree of Conservation vs. DSM

Space	Conserved	Semi-Conserved			Excluded		
		110	011	101	001	100	010
U	0.027	0.097	0.331	0.060	2.853	1.093	2.712
Y	0.001	0.002	0.012	0.002	0.067	0.009	0.054

Table 5.1: The percentage of features as a function of conservation type for **U** and **Y** for words $\{w_1, w_2, w_3 | w_1 \rightarrow w_2 \wedge w_2 \rightarrow w_3\}$. The percentage of features that are zero in all three words is omitted.

How Bad Is It?

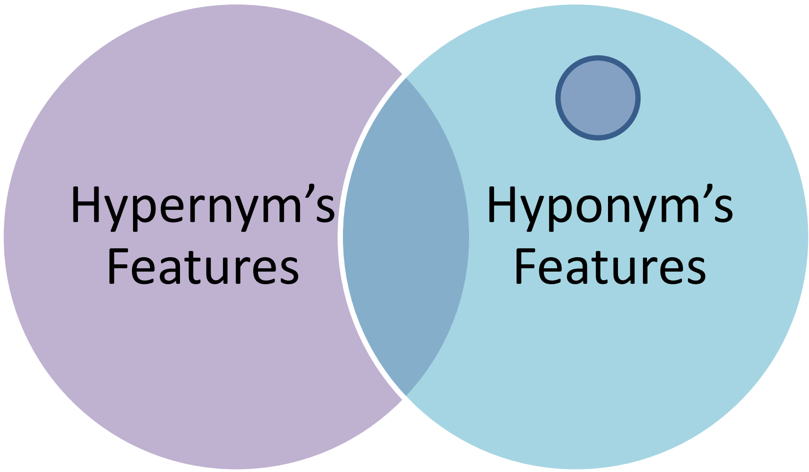
Percentage of Feature Weight for Degree of Conservation vs. DSM

Space	Conserved	Semi-Conserved 110	Semi-Conserved 101	Excluded 100
U	3.6	9.5	5.2	81.7
Y	9.2	15.3	12.5	63.0

Table 5.2: The percentage of feature weight with respect to w_1 as a function of conservation type for $\mathbf{U}$ and $\mathbf{Y}$ for words $\{w_1, w_2, w_3 | \, w_1 \rightarrow w_2 \wedge w_2 \rightarrow w_3\}$.

Why Do We Observe This?

1. Human Communication
2. Representation Design



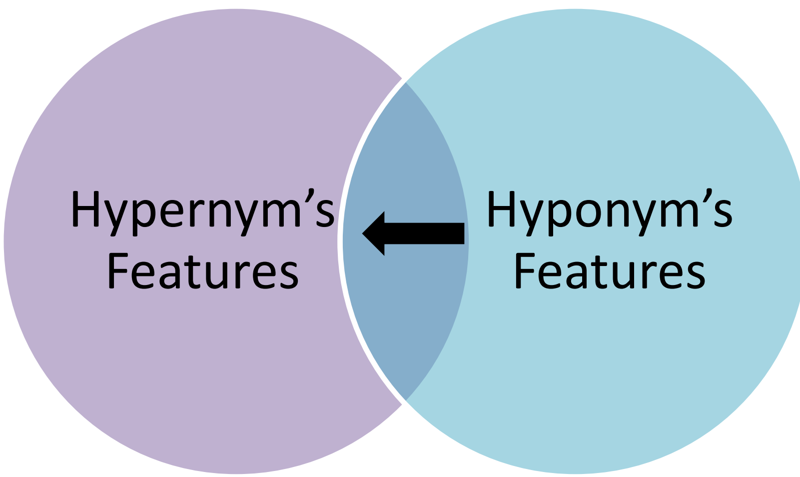
Causes of the FEP: Human Communication

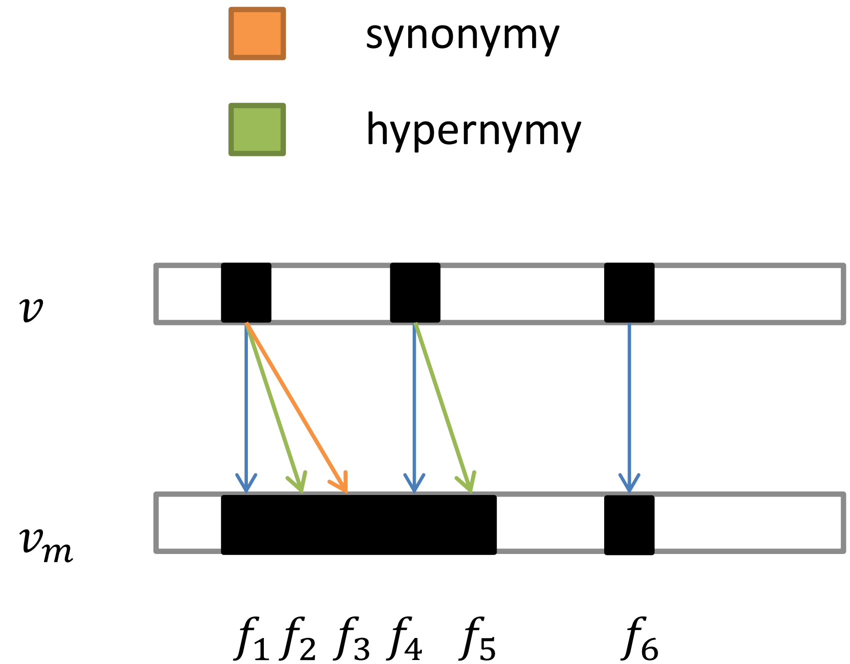
Type	Example
Non-Compositional	red herring
Compositional Non-Entailing	mane, hair
Compositional Entailing	gullet, esophagus
Spelling Variation	
Capitalization	colour, color
Case-Folding	
Gricean	*I went to the zoo and saw the entities.

Table 5.3: Types of Feature Exclusion

Causes of the FEP: Representation Design

A One-to-Many Map Using Entailment





$f_1 \mapsto \{f_1, f_2, f_3\}$ $f_6 \mapsto \{f_6\}$

$f_4 \mapsto \{f_4, f_5\}$

Effects: Generality ∝ Rank^{-1}

Rank	Noun		Verb	
	$\mathbf{U}$	$\mathbf{U}_m$	$\mathbf{U}$	$\mathbf{U}_m$
1	sledding	animal	bark	bark
2	sniffer	dog	foul	breed
3	sled	mammal	yap	sterilize
4	musher	food	kennel	treat
5	turd	carnivore	wag	walk
6	rehome	waste	muzzle	feed
7	whelk	bird	sledge	frighten
8	leash	ungulate	salivate	foul
9	crossbreed	disease	rehomed	jiggle
10	cat	canine	rehoming	eat
11	Alsatian	hound	neuter	sleep
12	kennel	guardian	yelp	yelp
13	mongrel	primate	cross-breed	yap
14	poo	goody	snarl	kennel
15	prairie	guard	herd	wag

Table 7.1: The top ranked features by POS in both $\mathbf{U}$ and $\mathbf{U}_m$

Effects: More Feature Conservation

Space	Conserved	Semi-Conserved			Excluded		
		110	011	101	001	100	010
$\mathbf{U}$	0.03	0.10	0.33	0.06	2.85	1.09	2.71
$\mathbf{U}_m$	0.18	0.18	0.60	0.13	3.31	1.22	3.13
$\frac{\mathbf{U}_m}{\mathbf{U}}$	6.75	1.83	1.81	2.22	1.16	1.12	1.15
$\mathbf{Y}$	0.001	0.002	0.012	0.002	0.067	0.009	0.054
$\mathbf{Y}_m$	0.003	0.003	0.018	0.002	0.073	0.009	0.058
$\frac{\mathbf{Y}_m}{\mathbf{Y}}$	2.598	1.287	1.479	1.393	1.089	1.023	1.079

Table 7.6: Percent of features by degree of feature conservation.

Effects: More Feature Weight Conservation

Space	Conserved	Semi-Conserved 110	Semi-Conserved 101	Excluded 100
$\mathbf{U}$	0.04	0.10	0.05	0.82
$\mathbf{U}_m$	0.35	0.10	0.07	0.48
$\frac{\mathbf{U}_m}{\mathbf{U}}$	9.80	1.05	1.34	0.58
$\mathbf{Y}$	0.09	0.15	0.12	0.63
$\mathbf{Y}_m$	0.32	0.14	0.13	0.41
$\frac{\mathbf{Y}_m}{\mathbf{Y}}$	3.48	0.92	1.05	0.65

Table 7.7: Proportion of feature weight as a function of degree of feature conservation.

Experimental Results

Model	$\mathbf{U}$	$\mathbf{U}_m$	$\mathbf{U}_m^{th}$	$\mathbf{Y}$	$\mathbf{Y}_m$	$\mathbf{Y}_m^{th}$
BalAPInc	0.78	0.64	0.65	0.71	0.50	0.50
WeedsDiff	0.67	0.69	0.69	0.70	0.72	0.72
Cosine	0.75	0.71	0.71	0.74	0.64	0.64
InvCL	0.73	0.65	0.66	0.82	0.79	0.80
$\mathrm{P}_{\mathrm{Set}}$	0.77	0.77	0.77	0.82	0.80	0.80
$\mathrm{R}_{\mathrm{Set}}$	0.65	0.54	0.56	0.50	0.50	0.50
SingleWidth	0.64	0.64	0.63	0.68	0.68	0.68
$\mathrm{P}_{\mathrm{Weeds}}$	0.78	0.76	0.75	0.82	0.80	0.80
$\mathrm{R}_{\mathrm{Weeds}}$	0.67	0.57	0.59	0.50	0.50	0.50
WidthDiff	0.67	0.67	0.67	0.71	0.71	0.71

Table 8.1: Accuracy of models for Experiments 1 and 2 using the Entailment dataset

Possible Explanation

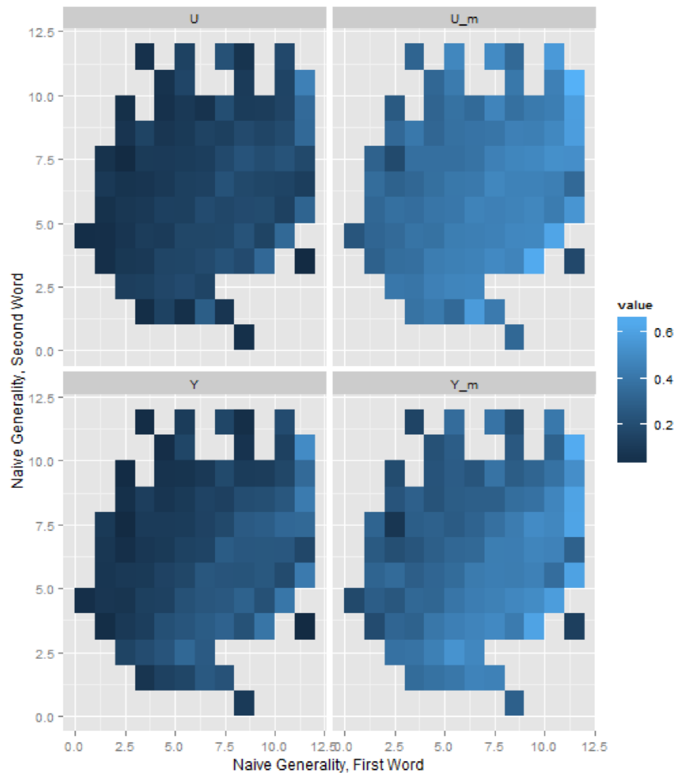


Figure 8.2: Heatmaps for all spaces of the output of the P_{Weeds} model for the Weeds dataset

Conclusion

- Feature Exclusion is a seriously bad.
- Method could be improved
  - Change feature weight re-distribution
  - Choose which features to map
- The HR task, as currently formalized, has no way of incorporating known information.