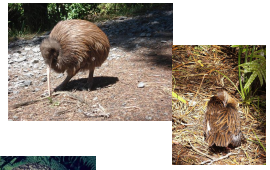
Multi-Modal Word Synset Induction

Jesse Thomason and Raymond Mooney
University of Texas at Austin

Word Synset Induction

“kiwi”



Word Synset Induction

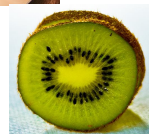
“kiwi”

“chinese grapefruit”

“kiwi vine”

Word Synset Induction















Word Sense Induction + Synonymy Detection

- First finding senses, then merging those senses through synonymy detection
- We call this *synset induction*, the task of finding synonymous sets of word senses
- Synsets used in WordNet [Fellbaum, 1998] and analogous ImageNet [Deng et al., 2009] corpora
  - Represent hierarchical collections of synonymous noun phrases
  - e.g. “kiwi”, “chinese grapefruit”, “kiwi vine”

Word Synset Induction

- WordNet is a handcrafted resource that required lots of human annotation
- ImageNet also utilizes human annotation
- For a new language or specialized domain, would be ideal to induce synsets in an unsupervised fashion
- We show this can be done, and is most effective when both textual and visual context are considered

Multi-modal Perception

- An instance of a concept is an image and contextual text about that image
- Textual and image data both give evidence of multiple word senses

Bat

"... most of the oldest known, definitely identified bat fossils were already very similar to modern microbats … "

Bat

"... a baseball bat is divided into several regions …"

Bat

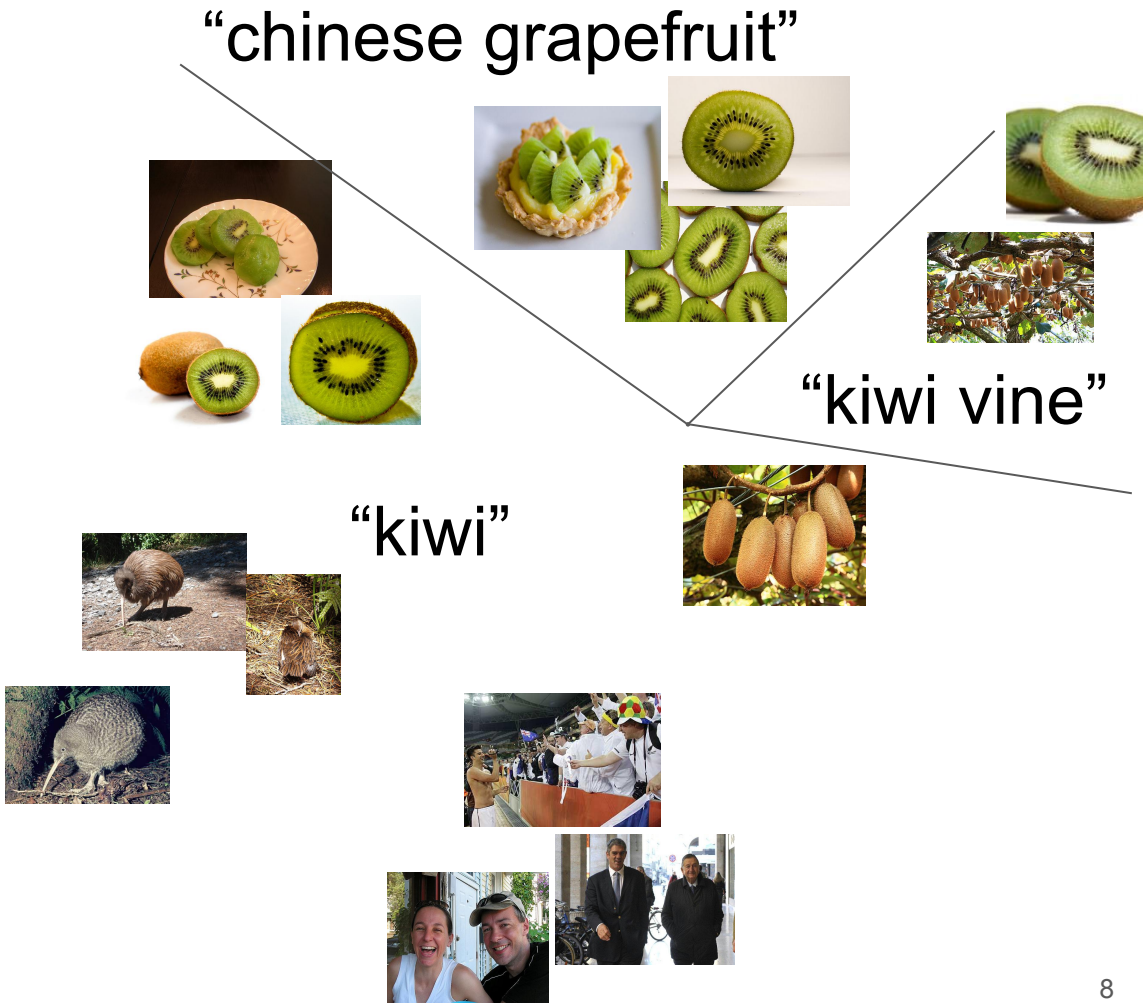
"... about 70% of bat species are insectivores … "

Bat

"... hickory has fallen into disfavor over its greater weight, which slows down bat speed … "

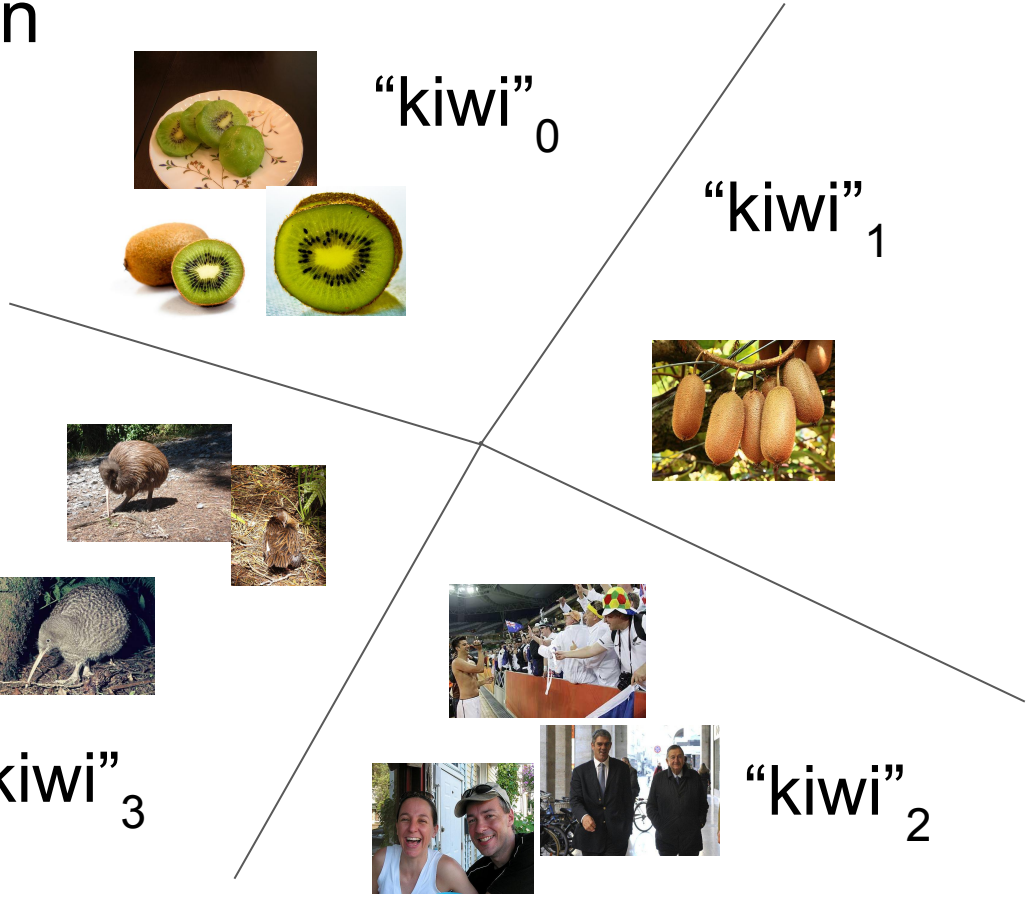
Task

- Take instances of noun phrases (images paired with text)
- Perform synset induction to gather underlying senses of noun phrases



Word Sense Induction

- Take instances of noun phrases (images paired with text)
- Perform synset induction to gather underlying senses of noun phrases



+Synonymy Detection

- Take instances of noun phrases (images paired with text)
- Perform synset induction to gather underlying senses of noun phrases

Dataset

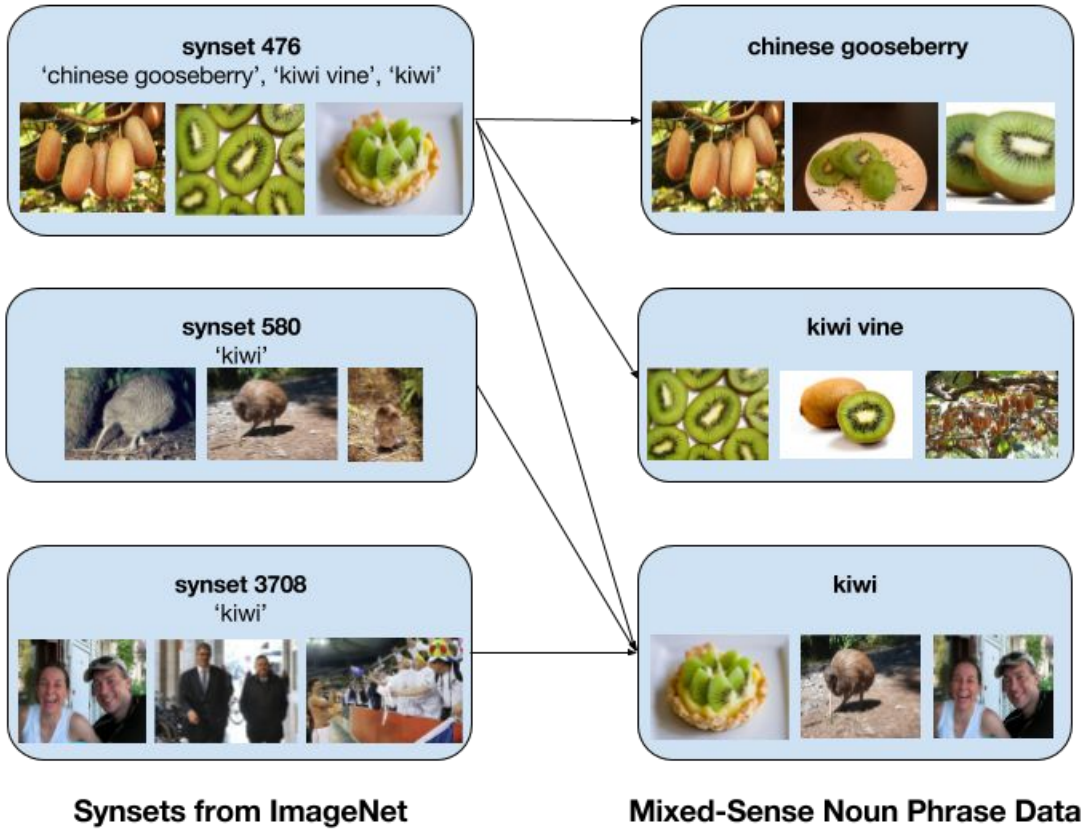
- Gather many leaf-level synsets (6710) and images from ImageNet
- Get a mix of noun phrase types (8426 total)
  - Many past works assume all words are polysemous

    (e.g. [Loeff et al., 2006; Saenko and Darrell, 2008])

Noun phrase relationships			
synonymous	**polysemous**	**both**	**neither**
4019	804	1017	2586

- Provides “gold” synsets we aim to construct from image-level instances
- Hold out validation noun phrases for hyperparameter tuning

Dataset


synset 476
'chinese gooseberry', 'kiwi vine', 'kiwi'
synset 580
'kiwi'
synset 3708
'kiwi'
chinese gooseberry
kiwi vine
kiwi
Synsets from ImageNet
Mixed-Sense Noun Phrase Data

Pairing Images w/ Text Data

[sentences]

[sentences]

[sentences]

text corpus

[bag of words]

[bag of words]

[bag of words]

LSA

256-dimensional

text feature space

Pairing Images w/ Text Data

Text features for image

“about 70% of bat species are insectivores”
“most of the oldest known, definitely identified bat fossils were already very similar to modern microbats”
….

LSA emb

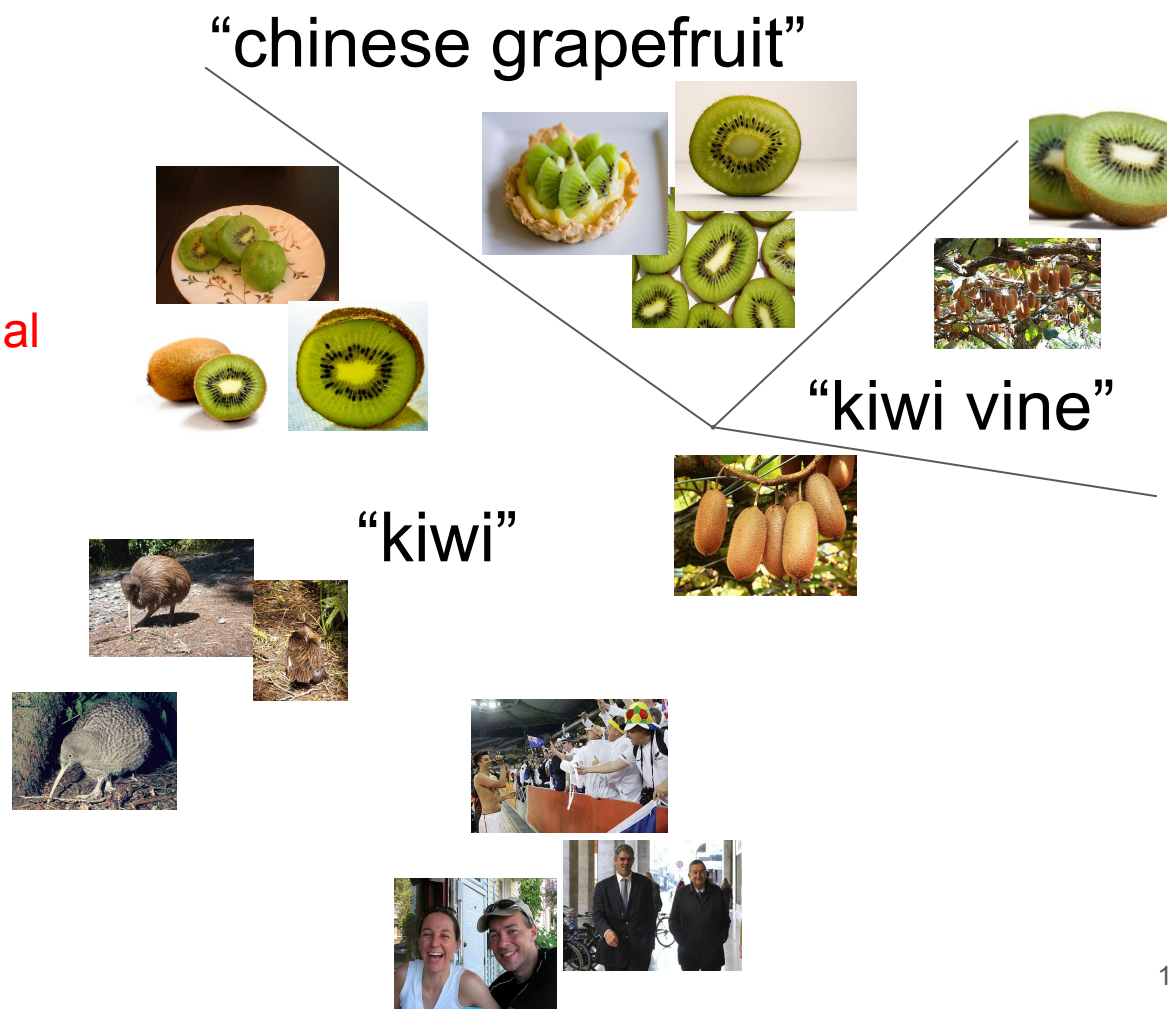
Extract Image Features

VGG network
[Simonyan and Zisserman, 2014]

Visual features for image (penultimate 4,096 unit layer of VGG network)

Dataset

- Each instance has associated text and visual features
- Features used to find distances between instances



Related Work - Word Sense Induction

- Task of discovering word senses [Pedersen and Bruce, 1997]
- “Bat”
  - Baseball, animal
- “Light”
  - Weight, color
- “Kiwi”
  - Fruit, bird, people



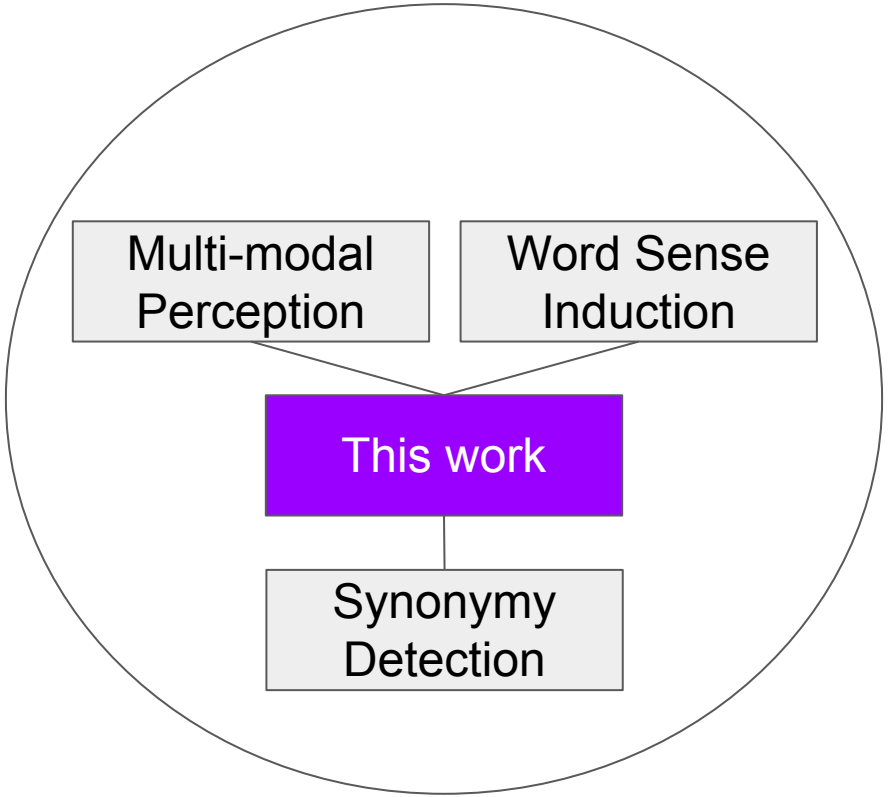




- Represent instances as vectors of their context; cluster to find senses
  - [Yarowsky, 1995; Pedersen and Bruce, 1997; Schutze, 1998; Bordag, 2006; Navigli, 2009; Manandhar et al., 2010; Di Marco and Navigli, 2013]

Related Work - Synonymy Detection

- Given words or word senses, find synonyms
- “Ball” and “sphere”
- “Mobile” and “phone” (for one sense of “mobile”)
- “Kiwi” and “New Zealander” (for one sense of “kiwi”)
- In text space, represent instances as vectors of their context; cluster means to find synonyms
    - Related to synonym detection [Turney, 2001] and lexical substitution [McCarthy and Navigli, 2009]

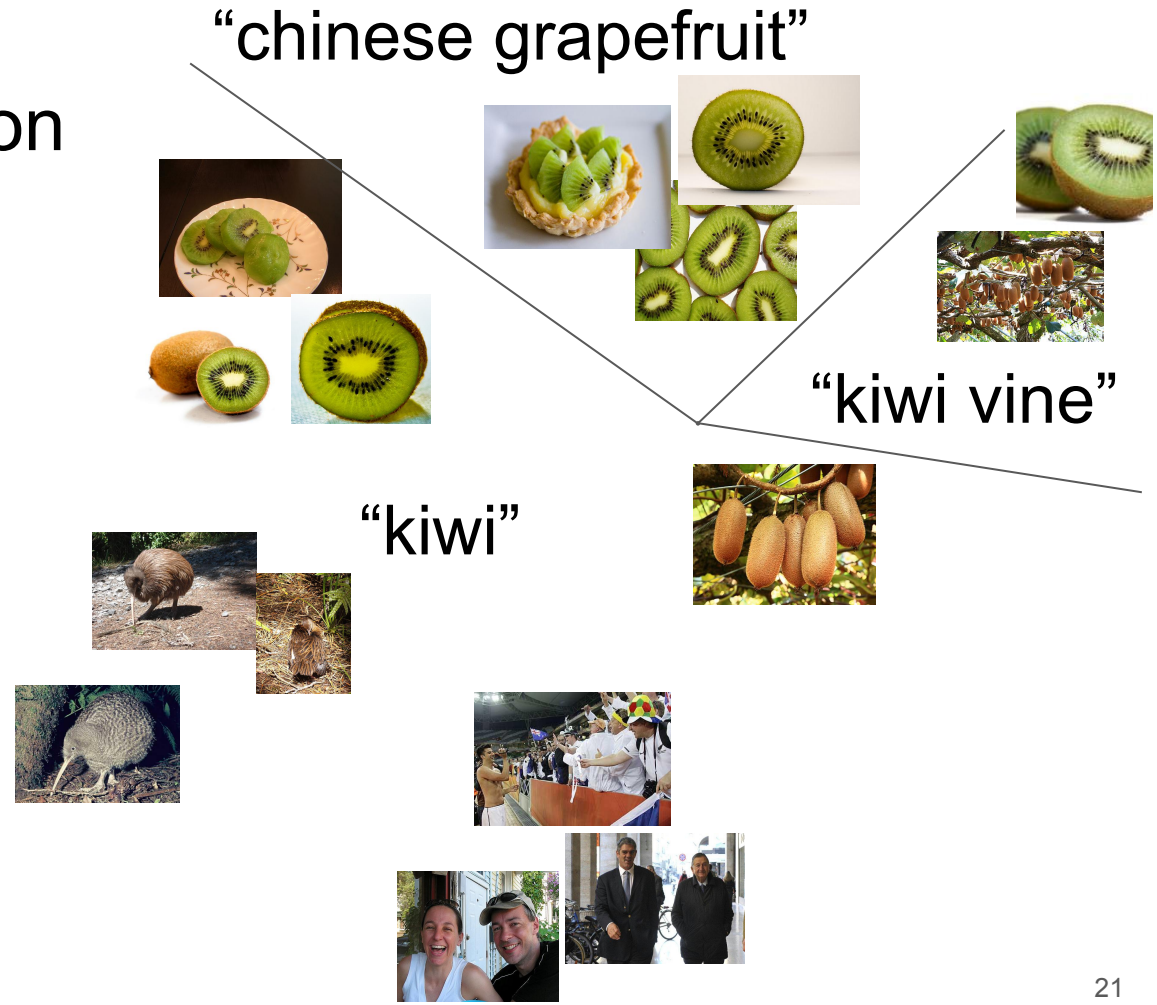

Multi-modal Perception
Word Sense Induction
This work
Synonymy Detection

Goal

- Induce ImageNet-like synsets from images labeled with just noun phrase
- First perform word-sense induction on mixed-sense noun phrase inputs
- Given induced word senses, perform synonymy detection to form synsets
- Compare induction considering text-only, visual-only, and multi-modal features
- For multi-modal space, interpolate distance calculations in text and visual spaces

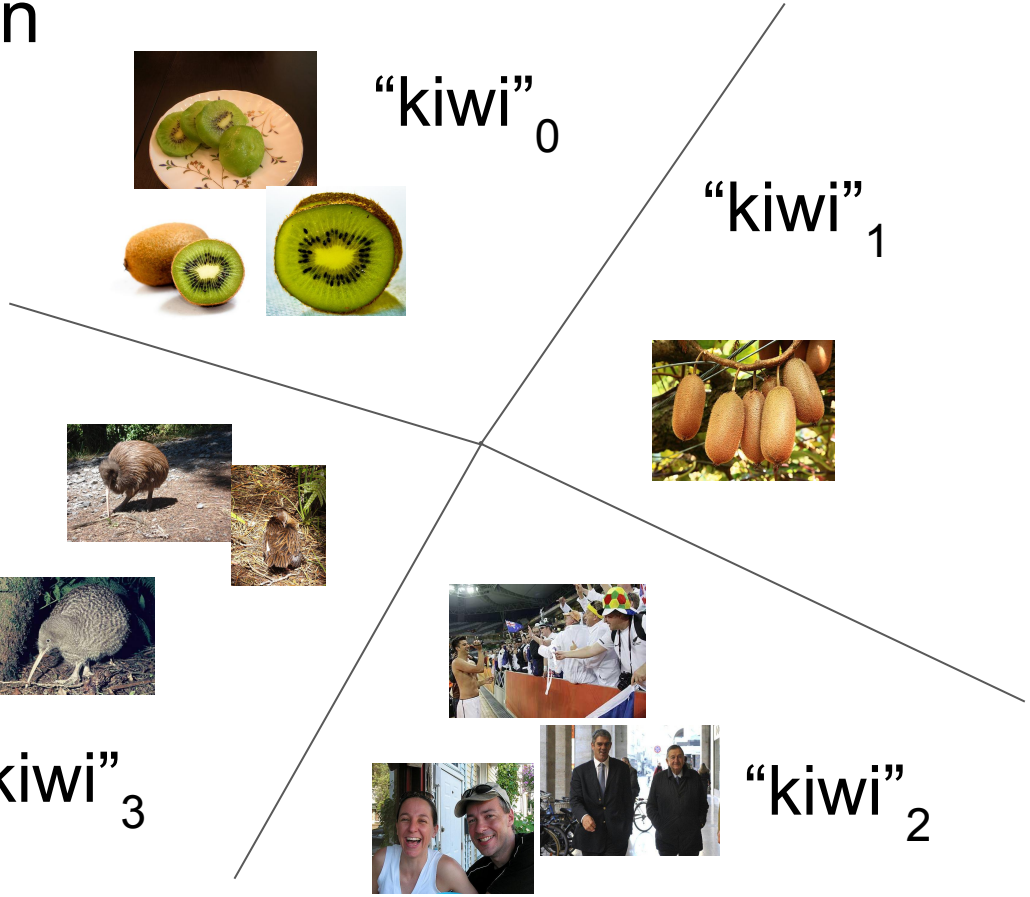
Word Sense Induction

- For every noun phrase, we perform k-means clustering to find senses
- Determine k by the gap statistic [Tibshirani et al., 2001]



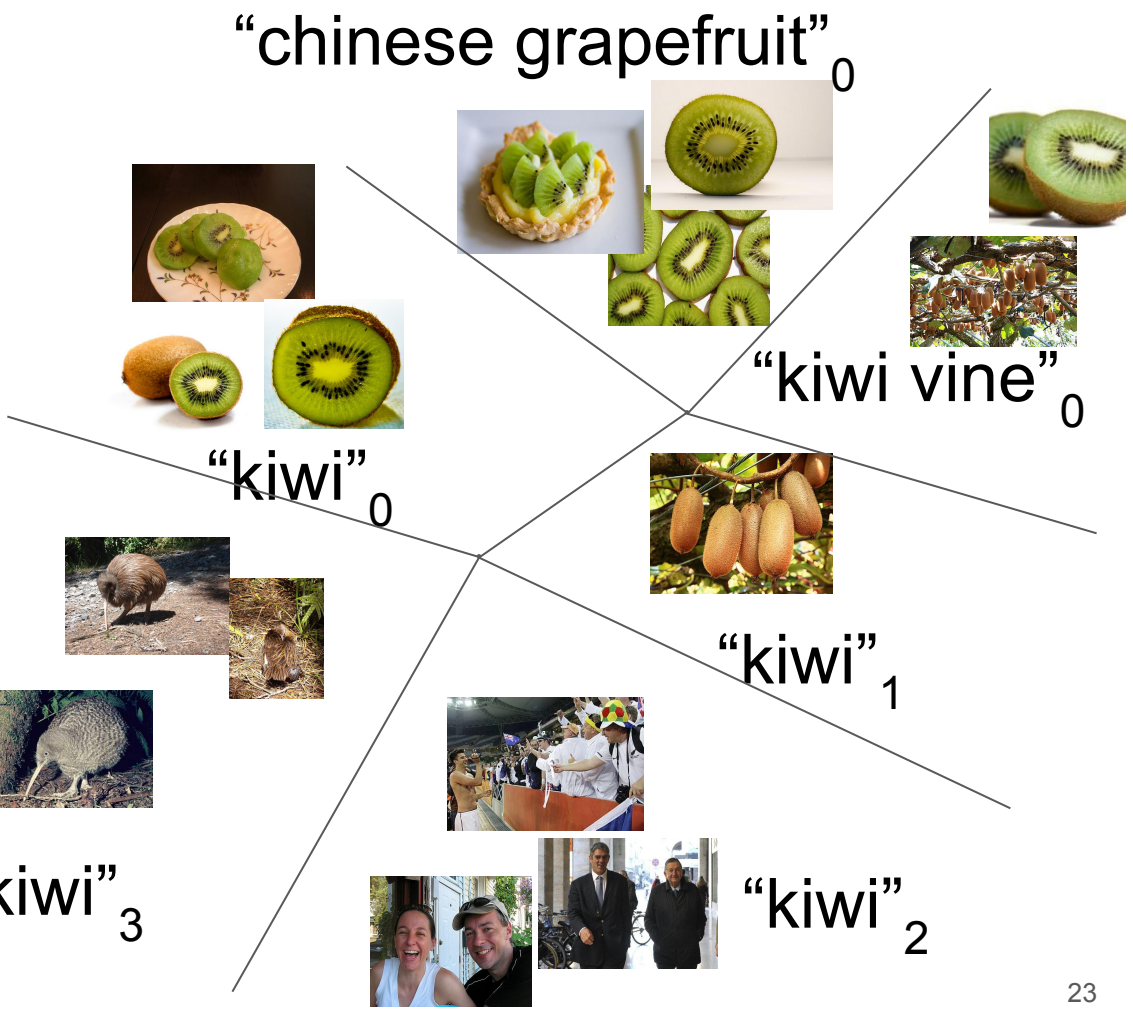
Word Sense Induction

- For every noun phrase, we perform k-means clustering to find senses
- Determine k by the gap statistic [Tibshirani et al., 2001]



Synonymy Detection

- Greedily merge nearest neighboring clusters by means
- Cap maximum merged senses (20, in our experiments)
- Results in synsets



Synonymy Detection

- Greedily merge nearest neighboring clusters by means
- Cap maximum merged senses (20, in our experiments)
- Results in synsets

Experiments

- Measure homogeneity, completeness, and their harmonic mean between induced synsets and ImageNet synsets
  - Analogous to precision, recall, and *f*-measure for sets of sets [Manandhar et al., 2010],
- Perform qualitative human evaluation of synset sensibility

Results

ImageNet

splashboard, washboard

washboard

Text-only

psaltery, washboard, dulcimer, cithern, headstock

king post, dugout, washboard, catapult, knothole

Vision-only

washboard

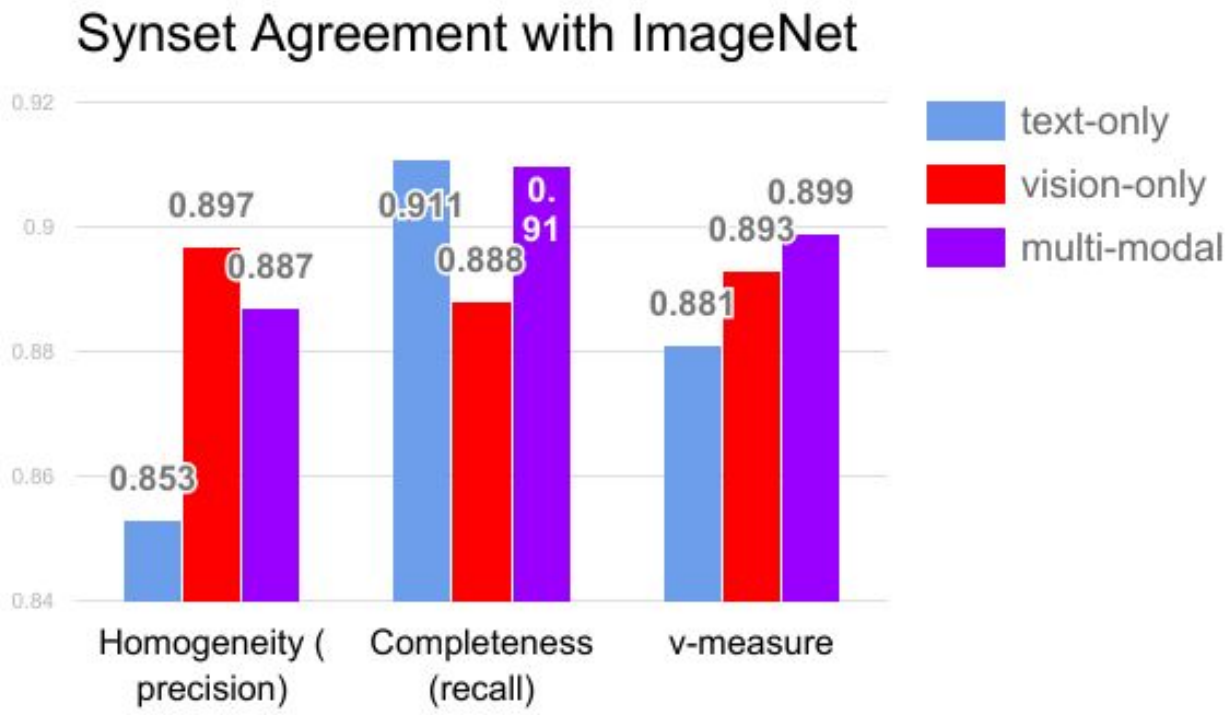
washboard, splashboard

splashboard, washboard

Multi-modal

splashboard, washboard

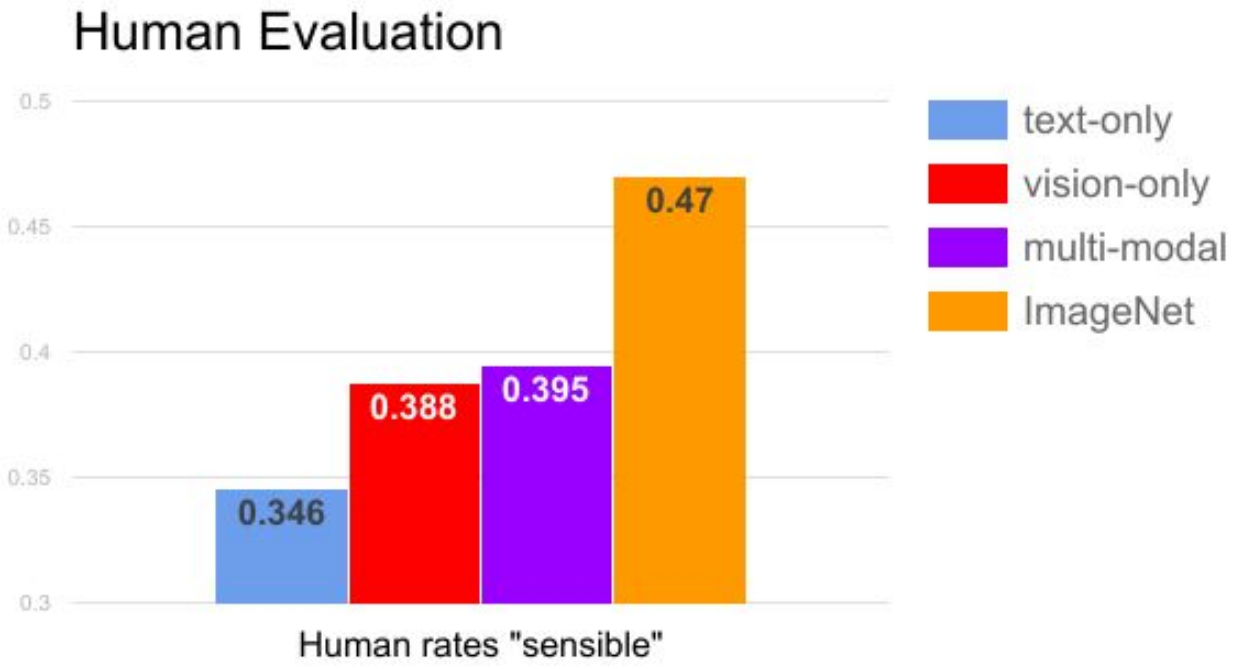
washboard


Synset Agreement with ImageNet
0.92
0.9
0.88
0.86
0.84
0.897
0.887
0.853
0.911
0.888
0.91
0.881
0.893
0.899
text-only
vision-only
multi-modal
Homogeneity (precision)
Completeness (recall)
v-measure

Human Evaluations

- Synset induction tends to join things ImageNet separates
- ImageNet separates people by nationality (e.g. “Austrian” and “Croatian”)
- ImageNet has odd categories for describing people (e.g. “energizer”)
- We evaluate induced synsets and ImageNet synsets by human judgements of sensibility
    - Humans shown all synsets a sampled noun phrase ended up in for each system
- Use paired t-test to determine whether humans statistically significantly favor ImageNet over induced synsets

Human Evaluations



- Text-only and vision-only statistically significantly less favored versus ImageNet
- Multi-modal difference not significant; 84% of ImageNet score

Conclusions

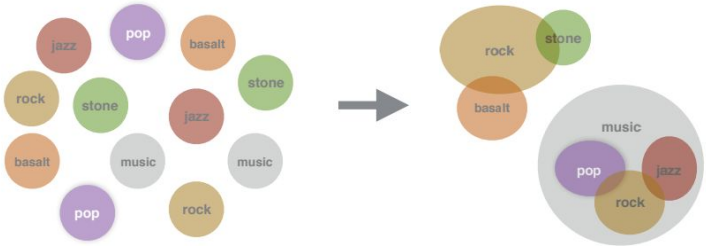
- Synset induction can be used to create ImageNet-like resource at leaf level from instances tagged with noun phrase labels
  - Substantially cheaper than human annotation-assisted ImageNet
  - Could be used for non-English ImageNet resource or specialized domains
- Image and text features together lead to synsets that more closely match ImageNet’s
- Human annotators rate multi-modal synsets sensible 84% as often as ImageNet synsets

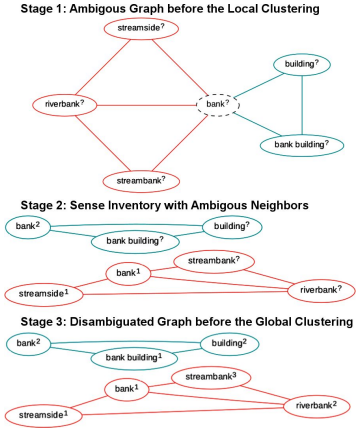
Multi-Modal Word Synset Induction

Jesse Thomason and Raymond Mooney
University of Texas at Austin

Contemporary Work on Synset Induction

- Watset: Automatic Induction of Synsets from a Graph of Synonyms (Ustalov *et al.*, ACL 2017)
  - Similar WSI + synonym clustering steps
  - Uses only textual information - we will use images as well





- Multimodal Word Distributions (Athiwaratkun and Wilson, ACL 2017)
  - Distributional WSI captures something like synonymy as well
  - Uses a fixed number of senses per word; we deduce from data