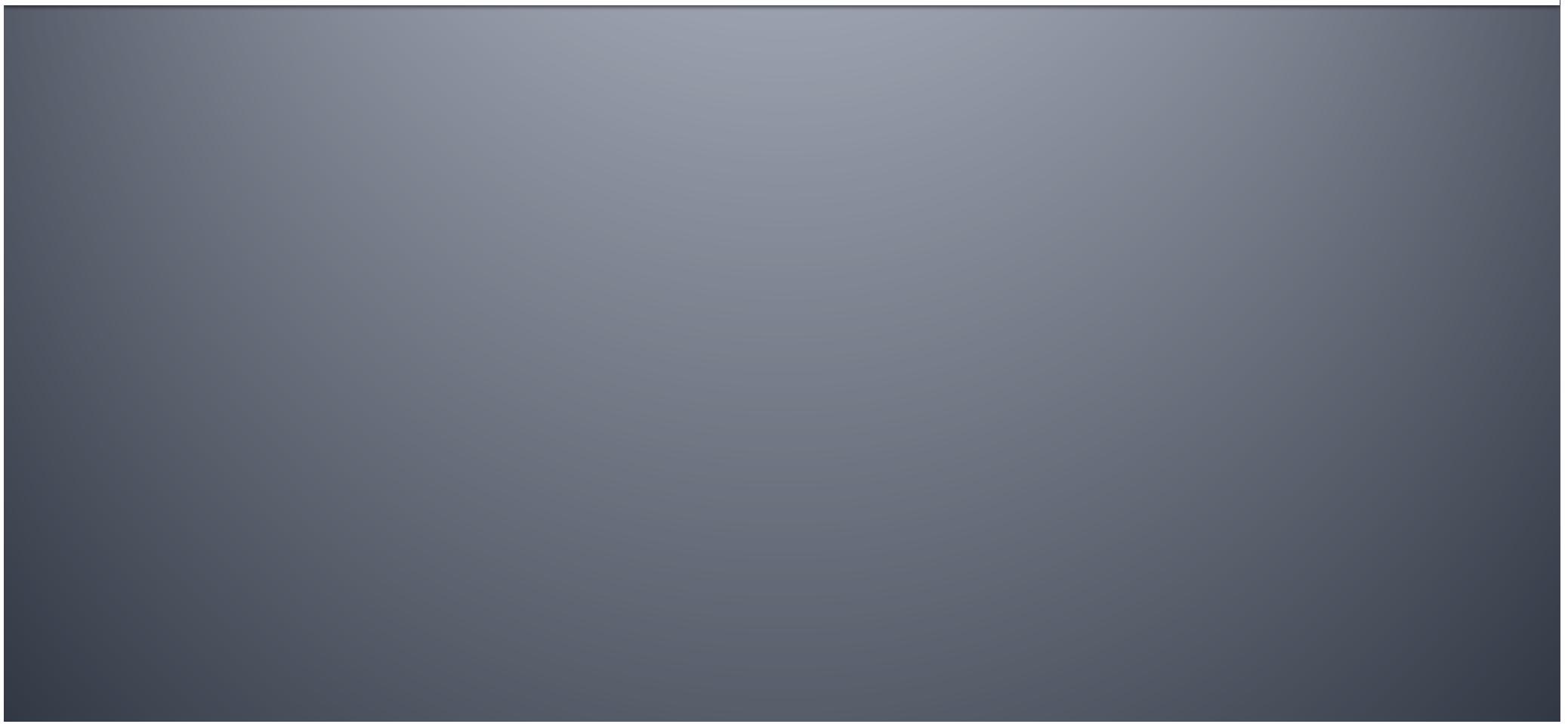
Week 1 – Lexical Distributional Semantics



What would you do ….?

- Imagine you have been given a set of 1000 documents, each of which have been annotated as ***relevant*** or ***irrelevant*** to a particular topic. Your task is to build a classifier which can assign the correct label (***relevant*** or ***irrelevant***) to a previously unseen document. How would you go about doing this?

Group discussion (5-10 mins)

- Groups of approximately 4-5 students
- Discuss for ~5 mins
- Report back to main group afterwards

Document classification using Naïve Bayes

- Represent each document as a bag-of-words. Features are observed words.
- The best class for a document is the one which maximises the probability of the class given the feature vector for the document.

$$\hat{l} = \underset{l \in L}{\operatorname{argmax}} P(l \mid \underline{f})$$

- We can't collect statistics about this directly so….

Apply Bayes Rule

Revision?

$$\hat{l} = \underset{l \in L}{\operatorname{argmax}} \frac{P(\underline{f} \mid l) \times P(l)}{P(\underline{f})}$$

- We can ignore the denominator (as this is independent of l)
- The prior probabilities P(l) can easily be estimated from the training data.
- How do we estimate P(f|l)?
- We make the **naive** assumption that the features are independent given the label so:

$$P(\underline{f} \mid l) = \prod_{j=1}^{n} P(f_j \mid l)$$

Towards more intelligent NLP

- Word tokens often viewed as atomic building blocks of language
- But "bag-of-word" models or even "word sequence" models miss a vital ingredient of human language understanding
- What is the meaning of a word?
- Meaning is generally inferred through
  - relationships with other words
  - similarity to other words

lexical semantics

Lecture 1 Overview

PART 1

- Lexical semantics
  - word senses
  - semantic relationships
  - WordNet
  - semantic similarity measures based on WordNet
  - evaluation

This might be largely revision if you have taken Applied NLP!

PART 2

- Distributional Semantics
  - bootstrapping semantics from context
  - cosine similarity
  - (positive) pointwise mutual information
  - evaluation
  - word ambiguity
  - semantic relationships
  - sparsity

Lecture 1 Questions?

- Use the module discussion forum for questions
- Or ask them now

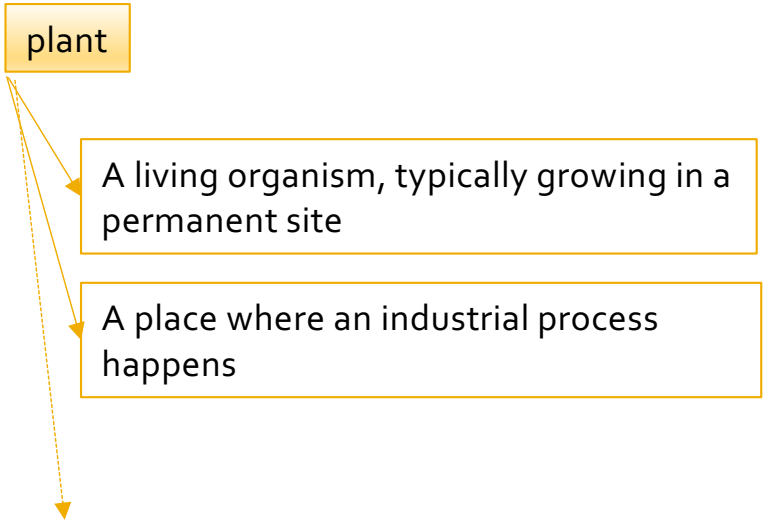
Lecture 1.1 questions (10 mins)

1. Give an example of lexical ambiguity
2. Give an example of lexical variation
3. What is a WordNet synset?
   - What does the number of synsets that a word form occurs in tell us?
   - What does the size of a synset tell us?
   - How are synsets connected?
4. Describe 2 ways WordNet can be used to calculate the similarity of 2 concepts?
   - Which is the best way and how do you know?

1.1.1 Lexical ambiguity: Sense distinctions

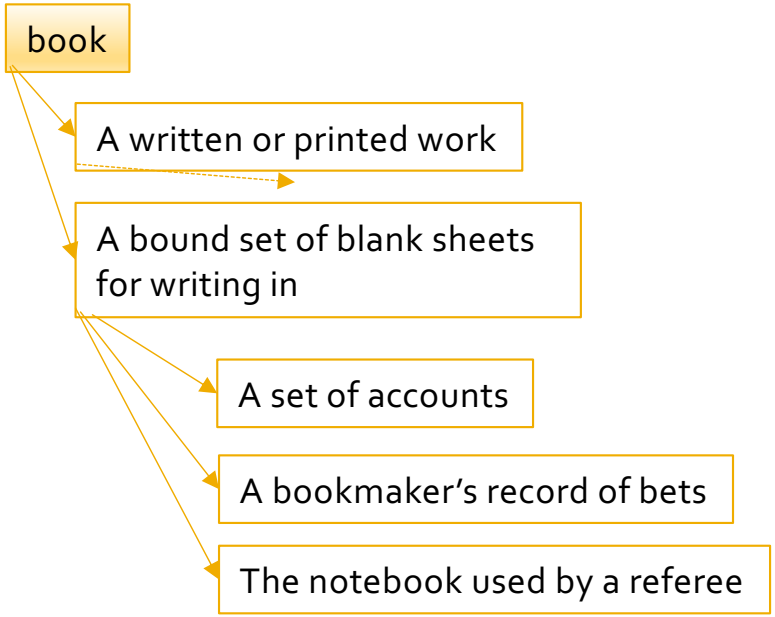
HOMONYMY

- Broad distinctions



POLYSEMY

- fine-grained distinctions



1.1.2 Lexical variation: Synonymy

fast == quickly

- Words which mean the same thing
- *Two words are **synonymous** if they can be substituted in all possible contexts without changing the meaning of the utterance.*
- True synonyms are very rare
- Choice of synonym usually gives us some extra information about the situation or speaker e.g., *car* vs *automobile*
- It is often defined as a relationship between word senses rather than between words. e.g., *plan*t == *spy* ?

1.1.3 WordNet

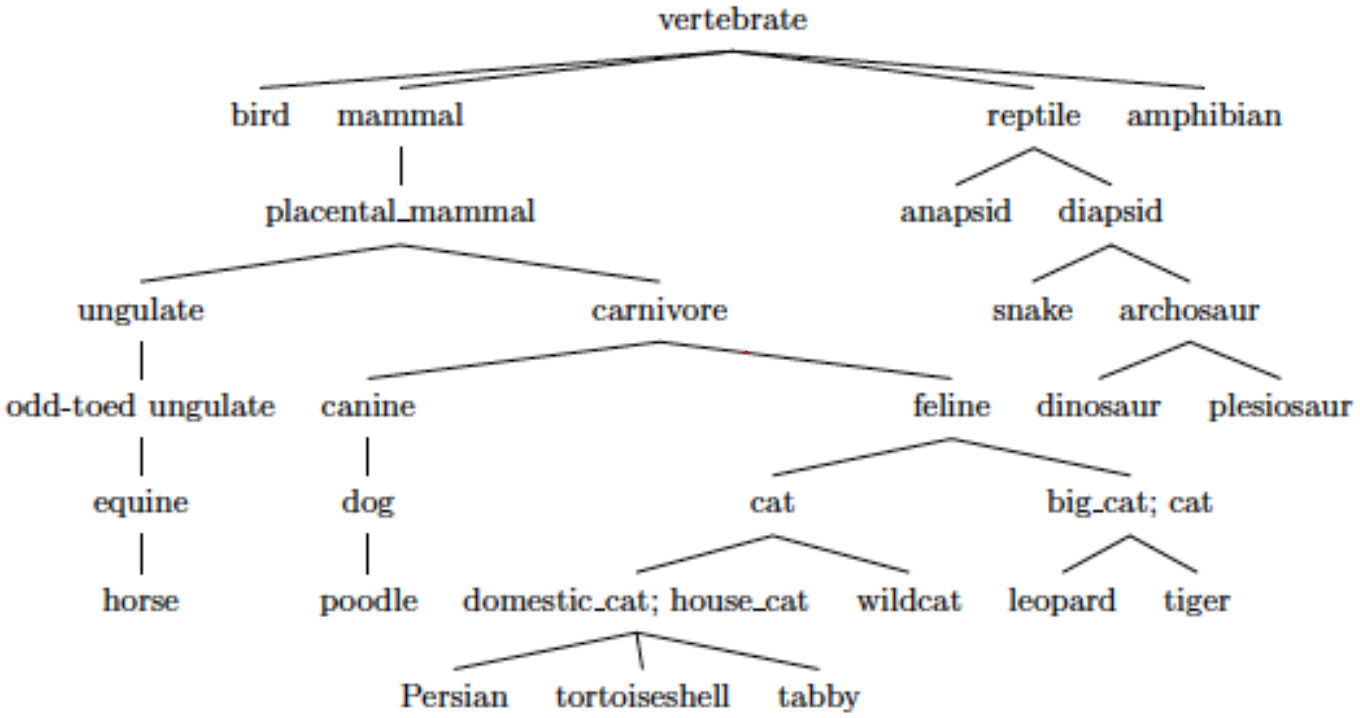
- A linguistic network organized around synonymy and hyponymy
- Core unit is the **synset**
  - a set of synonymous word senses
  - a set may contain a single word
  - synset items may be bigrams (e.g., "plant life") as well as unigrams
  - each synset is also associated with a single definition
- Polysemous words appear in multiple synsets
  - One for each sense
- Synsets are then connected via hyponymy…..

{**plant, flora, plant life**} = a living organism lacking the power of locomotion

{plant} = something planted secretly for discovery by another

{**plant, works, industrial plant**} = buildings for carrying on industrial labour

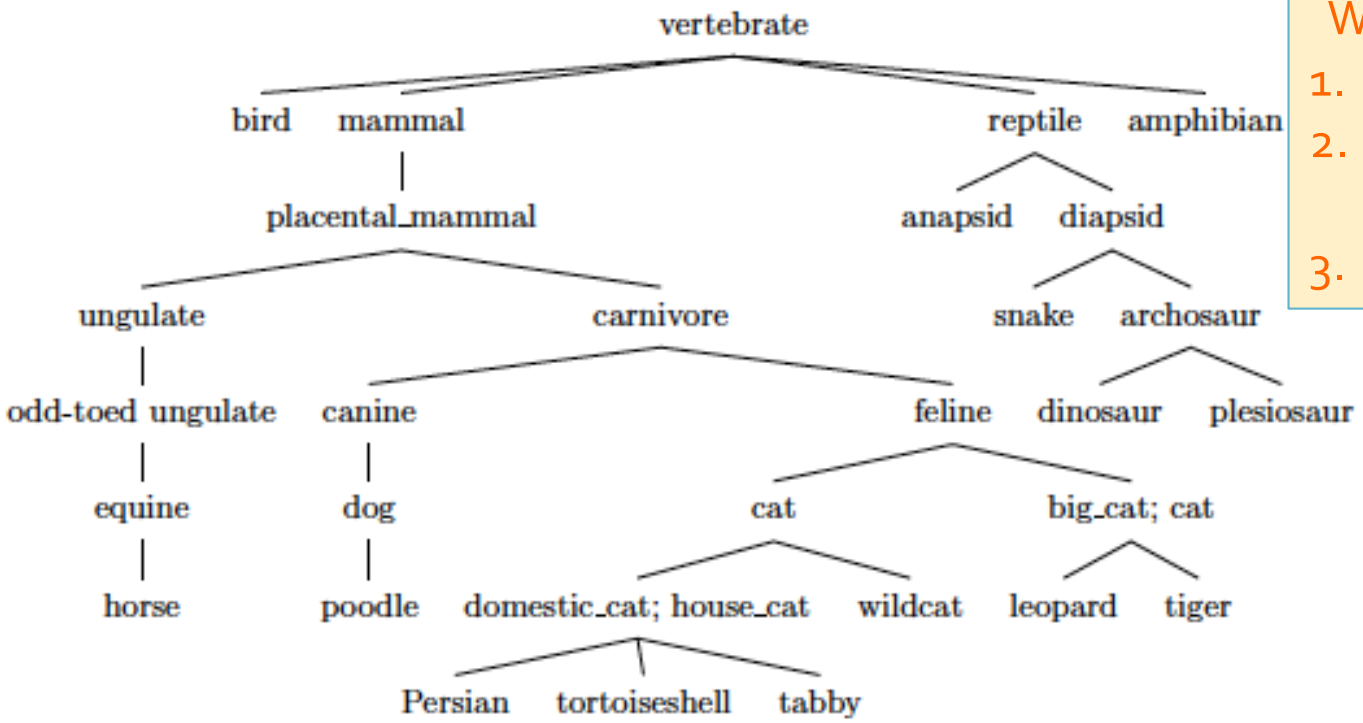
1.1.4 Path length: shorter path -> greater similarity



$$\text{sim}_{\text{path}}(c_1, c_2) = \frac{1}{1 + \text{pathlen}(c_1, c_2)}$$

1.1.4 Lowest common subsumer:

similarity based on what two concepts share



What is the LCS of:

1. tabby and tiger?
2. poodle and carnivore?
3. poodle and tiger?

Lecture 1.2 questions (10 mins)

1. What is the distributional hypothesis?
2. Explain how distributional semantics might help us in another application e.g., document classification
3. In traditional distributional semantics (aka vector semantics), how is the **association** between 2 words often measured?
4. In traditional distributional semantics, how is the **similarity** between 2 words often measured?

1.2.1 Distributional Hypothesis

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

Firth (1957)

The Distributional Hypothesis: "Words that occur in the same contexts tend to have similar meanings."

Harris (1954)

1.2.2 Distributional semantics in document classification

- Imagine we have built a Naïve Bayes document relevancy classifier using a relatively small training sample (e.g., 500 documents)
- A test document contains the word *tezguino* which has not been seen in the training sample

  so it cannot contribute to the relevancy classification
- But by applying distributional semantics to a very large unlabeled corpus (e.g., the web), we know that *tezguino* is very similar to *beer*

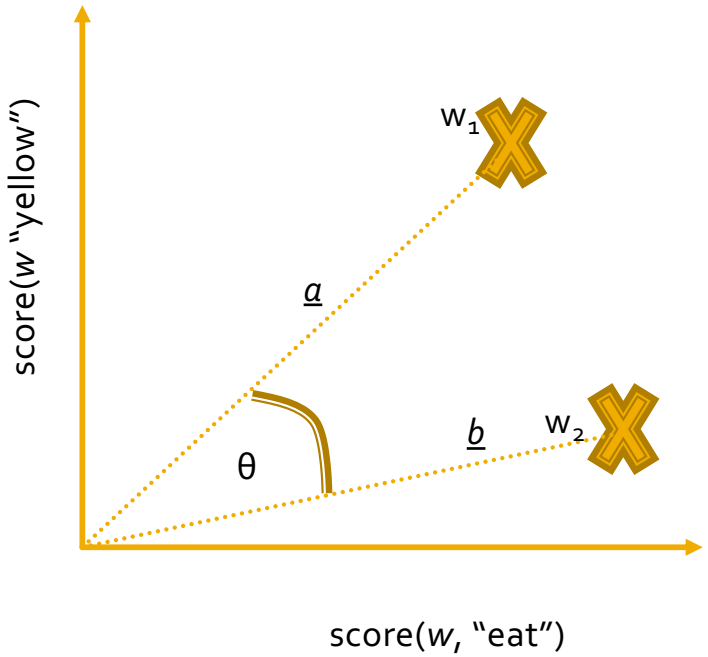
  beer has been seen in the training sample

  Assume **P(*tezguino*|class) ≈ P(*beer*|class)**

1.2.3 Association: Pointwise Mutual information (PMI)

- Frequency and/or simple conditional probability do not capture the intuition that some features are more informative than others
- *the* and *is* appear relatively frequently with all of the words
  - so their contribution to similarity should be smaller
- PMI measures the amount of information gained by seeing a word and a feature together
- A feature which co-occurs with a target word more than we would expect (if words and features occurred independently) has more weight in the similarity calculation

1.2.4 Similarity: cosine



- The more similar two words are, the smaller the angle θ between their vectors will be.
- So:

$$sim(w_1, w_2) = \cos(\theta)$$

dot product

$$= \frac{\underline{a}.\underline{b}}{\sqrt{\underline{a}.\underline{a} \times \underline{b}.\underline{b}}}$$

Where:

m=number of dimensions

$$\underline{a}.\underline{b} = \sum_{i}^{m} a_i b_i$$

Discussion of paper (20-30 mins)

- Ted Pedersen. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In Proceedings of NAACL
- See separate question sheet

Figure 1 shows part of the WordNet noun hypernym hierarchy. This is an **ISA** hierarchy where each concept in the tree **IS A** type of its parent. The parent concept is referred to as a hypernym (of the child) and the child concept is referred to as a hyponym (of the parent). Pedersen (2010) presents an empirical comparison of similarity measures for pairs of concepts in WordNet based on Information Content. Read the paper and answer the following questions.

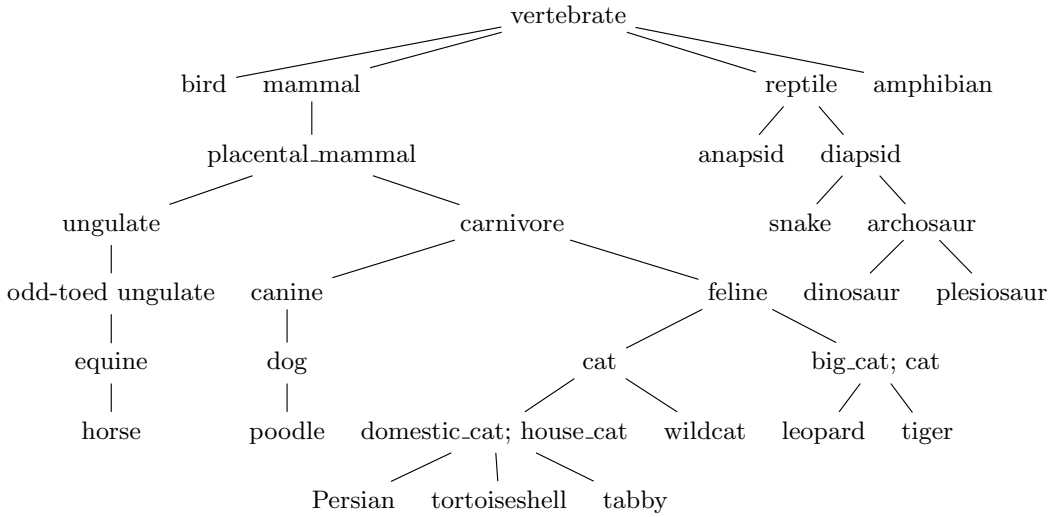


Figure 1: Part of the WordNet noun hypernym hierarchy

1. With reference to Figure 1, what concept is the hypernym of `ungulate`? How many hyponyms does `carnivore` have? Give an example. Why do you think the word `cat` appears twice in the hierarchy?

2. What do you understand by *path length*? Give some examples of pairs of words which have a *path length* of 2. What limitations can you think of in using *path length* as a measure of semantic similarity?

3. How is information content for a WordNet concept computed from a sense-tagged corpus? How can information content for a WordNet concept be estimated from untagged data?

4. What is the *lowest common subsumer (LCS)* of `dog` and `big_cat`? What is the LCS of `mammal` and `reptile`? What is the LCS of `poodle` and `tabby`? Which of these three pairs would have the greatest similarity according to the *res* measure? What about if you used the *lin* measure? Or a measure based on *path length*?

5. What is the main experimental conclusion of the paper? Are you convinced?

Information Content Measures of Semantic Similarity Perform Better Without Sense-Tagged Text

Ted Pedersen
Department of Computer Science
University of Minnesota, Duluth
Duluth, MN 55812
tpederse@d.umn.edu
http://wn-similarity.sourceforge.net

Abstract

This paper presents an empirical comparison of similarity measures for pairs of concepts based on Information Content. It shows that using modest amounts of untagged text to derive Information Content results in higher correlation with human similarity judgments than using the largest available corpus of manually annotated sense–tagged text.

1 Introduction

Measures of semantic similarity based on WordNet have been widely used in Natural Language Processing. These measures rely on the structure of WordNet to produce a numeric score that quantifies the degree to which two concepts (represented by a sense or synset) are similar (or not). In their simplest form these measures use path length to identify concepts that are physically close to each other and therefore considered to be more similar than concepts that are further apart.

While this is a reasonable first approximation to semantic similarity, there are some well known limitations. Most significant is that path lengths between very specific concepts imply much smaller distinctions in semantic similarity than do comparable path lengths between very general concepts. One proposed improvement is to augment concepts in WordNet with *Information Content* values derived from sense–tagged corpora or from raw unannotated corpora (Resnik, 1995).

This paper shows that Information Content measures based on modest amounts of unannotated corpora have greater correlation with human similarity judgements than do those based on the largest corpus of sense-tagged text currently available.[1] The key to this success is not in the specific type of corpora used, but rather in increasing the number of concepts in WordNet that have counts associated with them. These results show that Information Content measures of semantic similarity can be significantly improved without requiring the creation of sense–tagged corpora (which is very expensive).

1.1 Information Content

Information Content (IC) is a measure of specificity for a concept. Higher values are associated with more specific concepts (e.g., *pitch_fork*), while those with lower values are more general (e.g., *idea*). Information Content is computed based on frequency counts of concepts as found in a corpus of text. The frequency associated with a concept is incremented in WordNet each time that concept is observed, as are the counts of the ancestor concepts in the WordNet hierarchy (for nouns and verbs). This is necessary because each occurrence of a more specific concept also implies the occurrence of the more general ancestor concepts.

When a corpus is sense–tagged, mapping occurrences of a word to a concept is straightforward (since each sense of a word corresponds with a concept or synset in WordNet). However, if the text has not been sense–tagged then all of the possible senses of a given word are incremented (as are their ancestors). For example, if *tree* (as a plant) occurs in a sense–tagged text, then only the concept associated

[1] These experiments were done with version 2.05 of WordNet::Similarity (Pedersen et al., 2004).

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 329–332,
Los Angeles, California, June 2010.

Question 1

- With reference to Figure 1, what concept is the hypernym of ungulate?
- How many hyponyms does carnivore have?
- Give an example.
- Why do you think the word cat appears twice in the hierarchy?

Figure 1 shows part of the WordNet noun hypernym hierarchy. This is an **ISA** hierarchy where each concept in the tree **IS A** type of its parent. The parent concept is referred to as a hypernym (of the child) and the child concept is referred to as a hyponym (of the parent). Pedersen (2010) presents an empirical comparison of similarity measures for pairs of concepts in WordNet based on Information Content. Read the paper and answer the following questions.

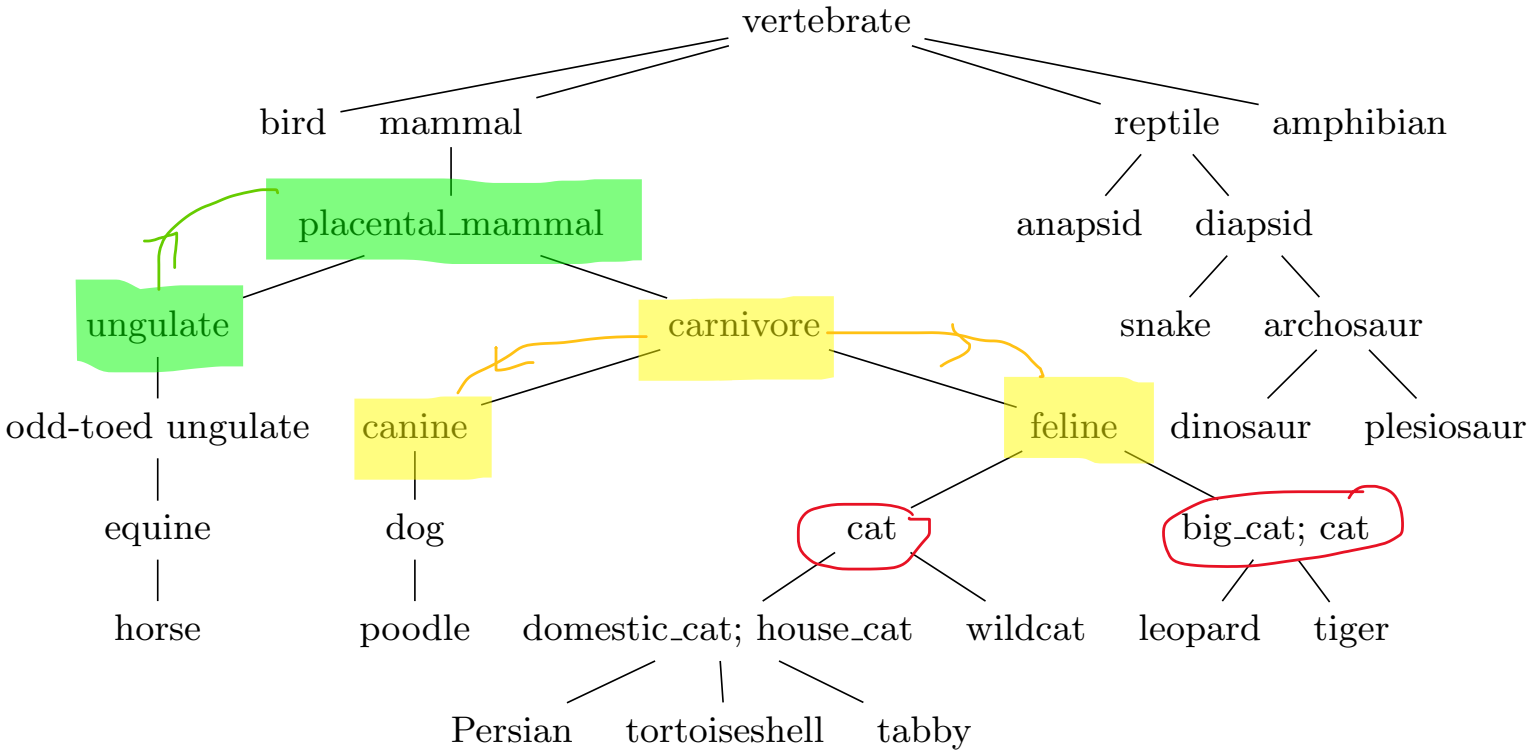
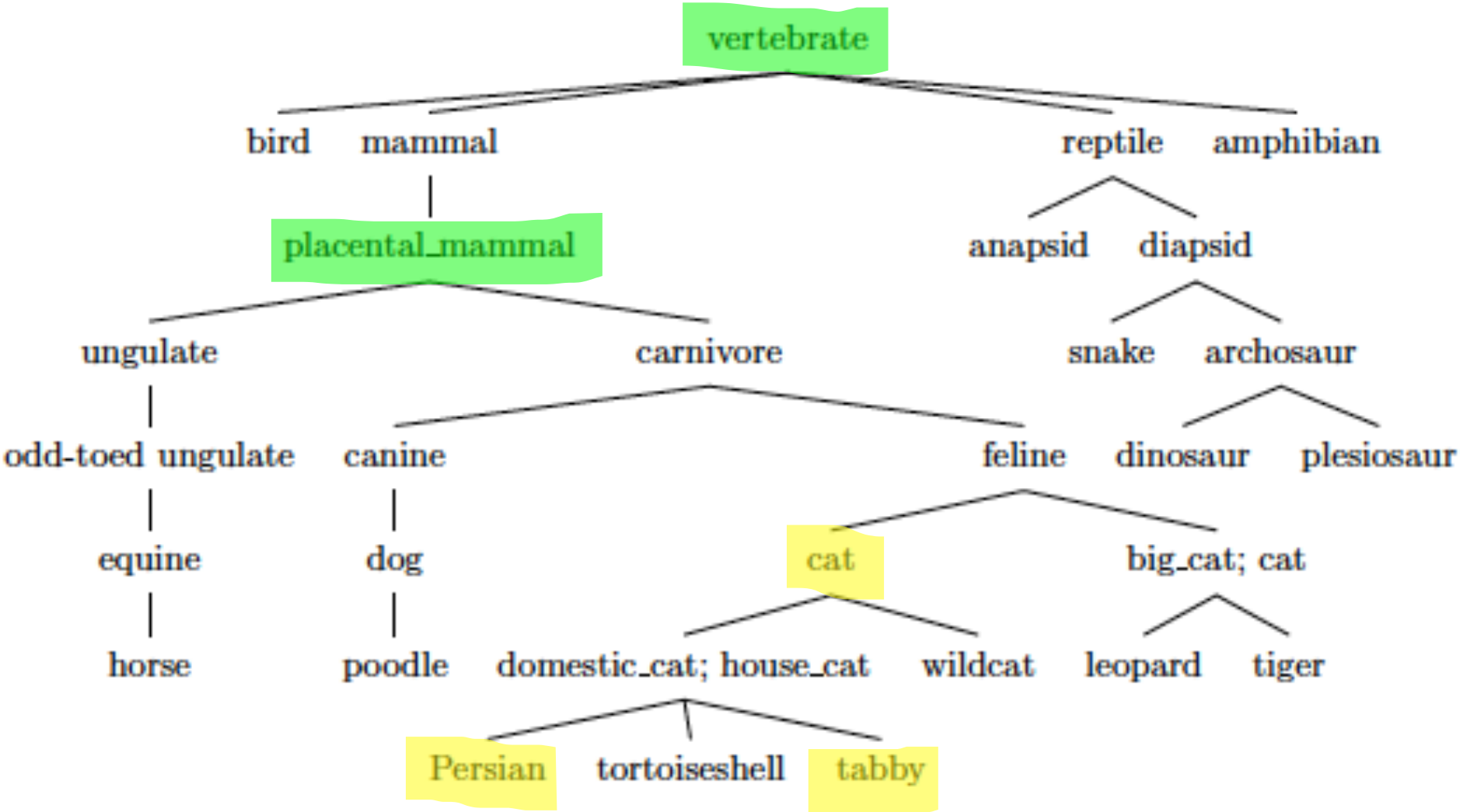


Figure 1: Part of the WordNet noun hypernym hierarchy

1. With reference to Figure 1, what concept is the hypernym of `ungulate`? How many hyponyms does `carnivore` have? Give an example. Why do you think the word `cat` appears twice in the hierarchy?

Question 2

- What do you understand by path length?
- Give some examples of pairs of words which have a path length of 2
- What limitations can you think of in using path length as a measure of semantic similarity?


vertebrate
bird
mammal
reptile
amphibian
placental_mammal
anapsid
diapsid
ungulate
carnivore
snake
archosaur
odd-toed ungulate
canine
feline
dinosaur
plesiosaur
equine
dog
cat
big_cat; cat
horse
poodle
domestic_cat; house_cat
wildcat
leopard
tiger
Persian
tortoiseshell
tabby

1 Introduction

Measures of semantic similarity based on WordNet have been widely used in Natural Language Processing. These measures rely on the structure of WordNet to produce a numeric score that quantifies the degree to which two concepts (represented by a sense or synset) are similar (or not). In their simplest form these measures use path length to identify concepts that are physically close to each other and therefore considered to be more similar than concepts that are further apart.

While this is a reasonable first approximation to semantic similarity, there are some well known limitations. Most significant is that path lengths between very specific concepts imply much smaller distinctions in semantic similarity than do comparable path lengths between very general concepts. One proposed improvement is to augment concepts in WordNet with *Information Content* values derived from sense–tagged corpora or from raw unannotated corpora (Resnik, 1995).

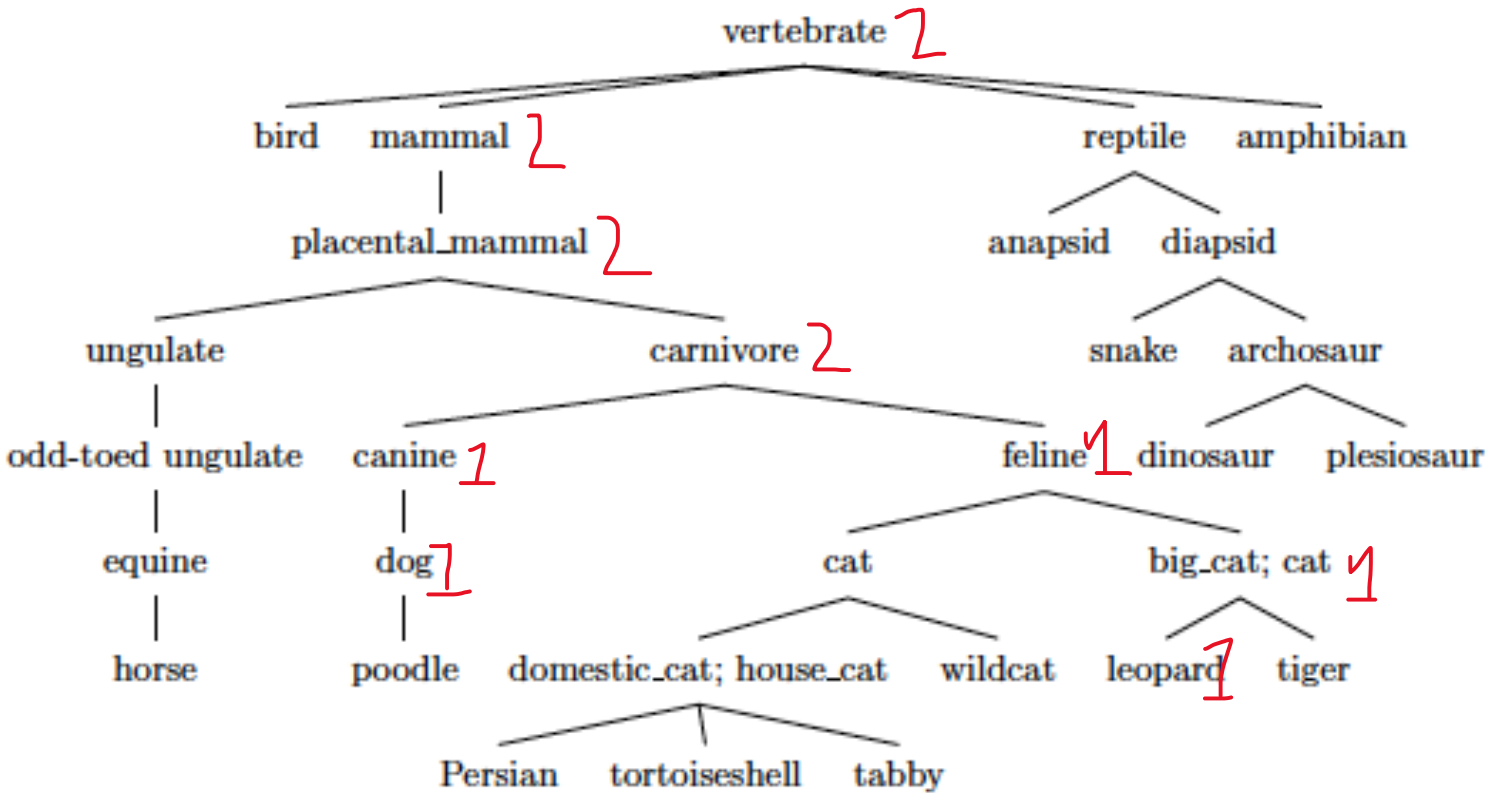
Question 3

- How is information content for a WordNet concept computed from a sense-tagged corpus?
- How can information content for a WordNet concept be estimated from untagged data?

Information Content from sense-tagged data

- The frequency associated with a concept, c, is incremented each time that concept is observed in the corpus
- Also increment the ancestor counts, since each occurrence of a more specific concept implies the occurrence of the more general ancestor concepts

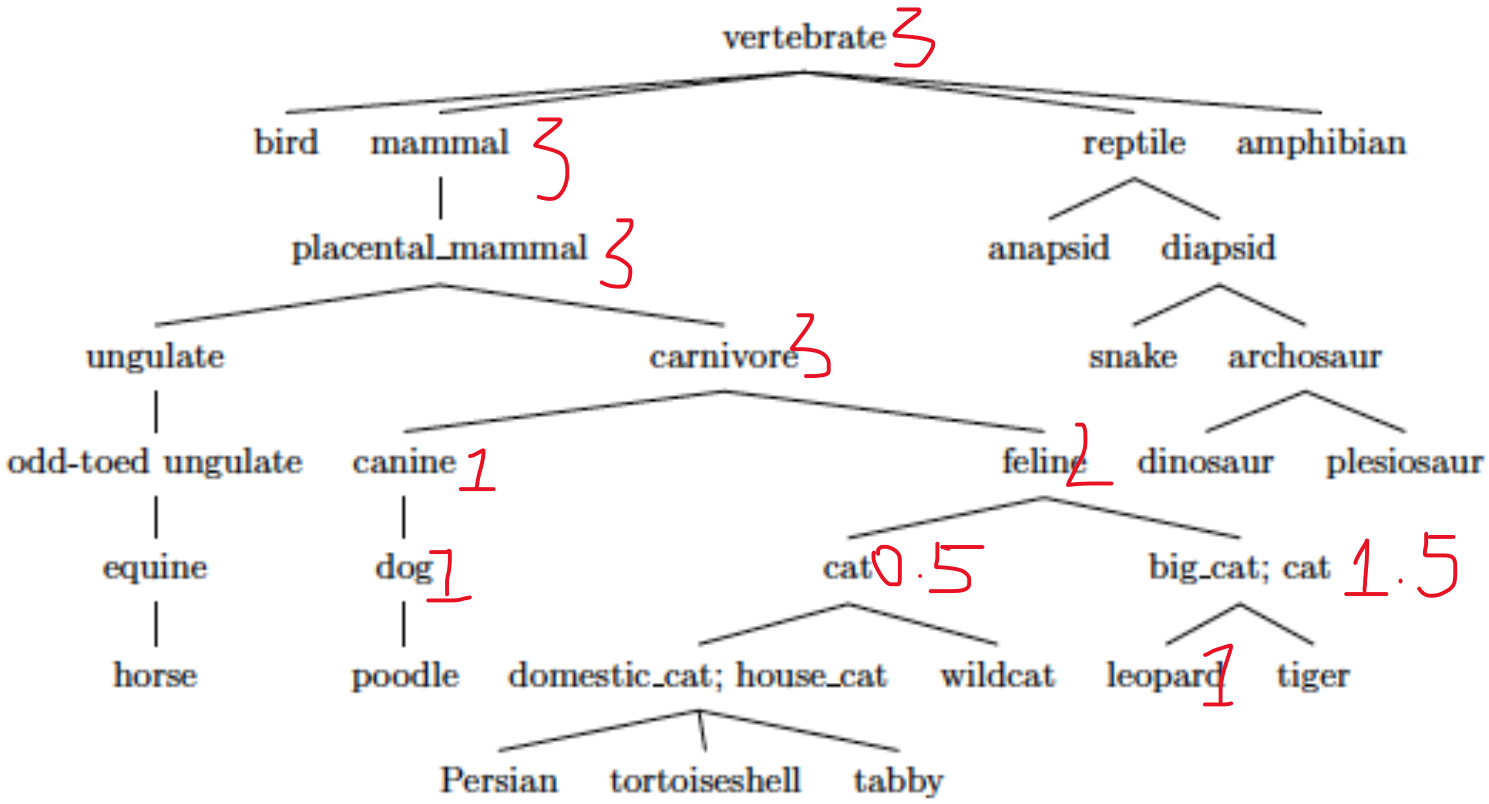
$$IC(c) = -\log P(c)$$



"I saw a leopard and a dog."

Information content from untagged data

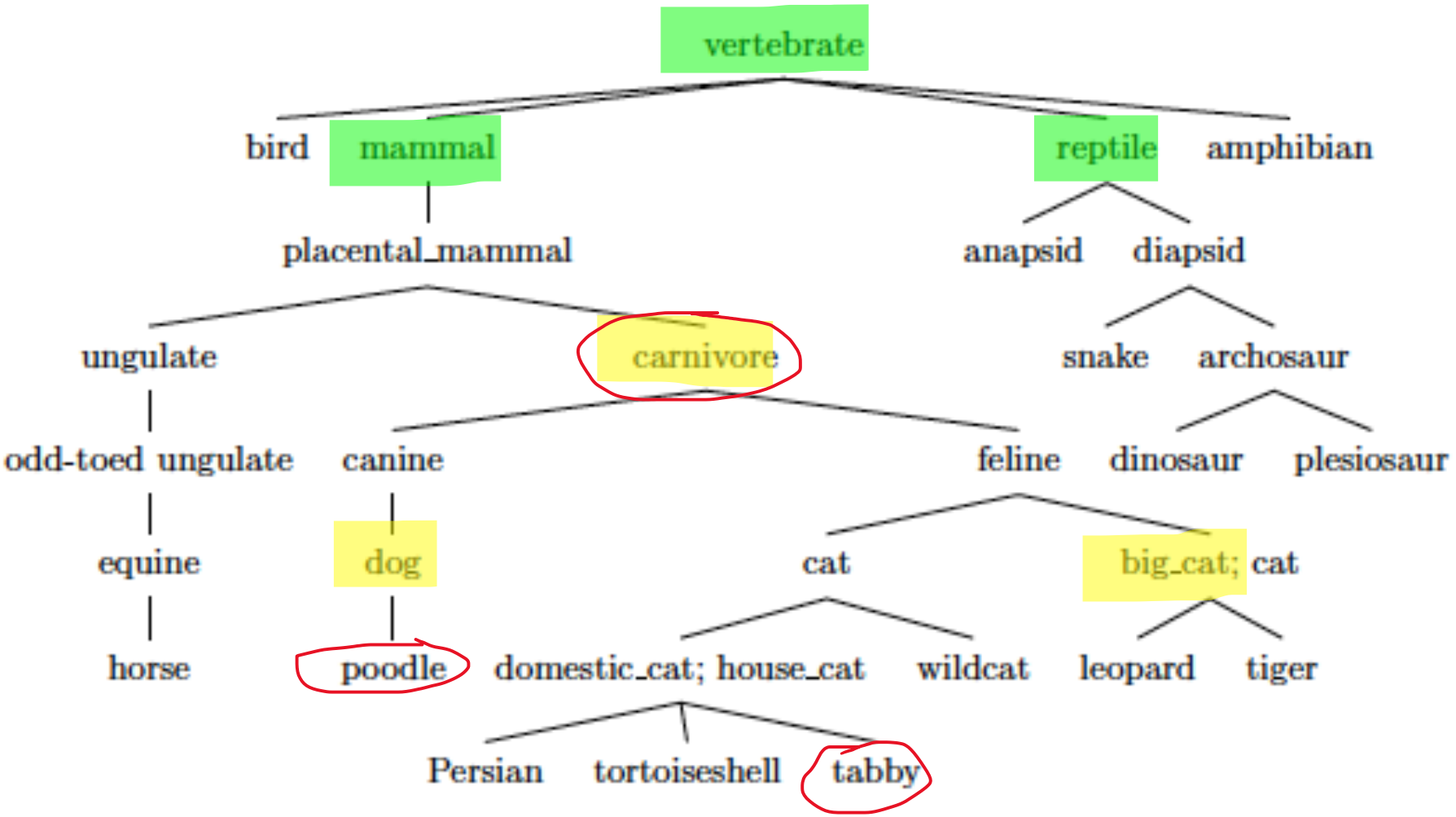
- If the text is untagged, increment counts for all of the possible senses of a word
- Divide the frequency of all the occurrences between the different possible senses



"I saw a leopard and a dog and a cat."

Question 4

- What is the lowest common subsumer (LCS) of dog and big_cat?
- What is the LCS of mammal and reptile?
- What is the LCS of poodle and tabby?
- Which of these 3 pairs would have the greatest similarity according to the Resnik measure?
- What about if you used the Lin measure?
- Or a measure based on path length?



$sim_{res}(poodle,tabby) = sim_{res}(dog,big\_cat)$

$IC(vertebrate) < IC(carnivore)$

$sim_{res}(mammal,reptile) < sim_{res}(dog,big\_cat)$

WordNet similarity measures based on information content (IC)

$$\mathrm{IC}(c) = -\log P(c)$$

Information content in a concept

$$\mathrm{sim}_{\mathrm{res}}(c_1, c_2) = \mathrm{IC}(\mathrm{LCS}(c_1, c_2))$$

Information content in what the concepts share (their lowest common subsumer)

See Resnik, 1995

$$\mathrm{sim}_{\mathrm{lin}}(c_1, c_2) = \frac{2 \times \mathrm{sim}_{\mathrm{res}}(c_1, c_2))}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)}$$

Ratio of shared information content to total information content

See Lin 1998b

$$\mathrm{sim}_{\mathrm{lin}}(c_1, c_2) = \frac{2 \times \mathrm{sim}_{\mathrm{res}}(c_1, c_2))}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)}$$

sim_{res}(poodle,tabby) = sim_{res}(dog,big_cat)

IC(poodle) > IC(dog)

IC(tabby) > IC(big_cat) ??

IC(poodle)+IC(tabby) > IC(dog) +IC(big_cat) ??

sim_{Lin}(poodle,tabby) < sim_{Lin}(dog,big_cat)

Question 5

- What is the main experimental conclusion of the paper?
- Are you convinced?

Paper conclusions

- Semantic similarity measures based on IC can be significantly improved by increasing coverage of the frequency counts
- Increased coverage can come from:
  - annotated data
  - unannotated data
  - smoothing
- Quantity of data more important than quality!