NATURAL LANGUAGE PROCESSING

(COM4513/6513)

WORD SENSE DISAMBIGUATION

Andreas Vlachos
a.vlachos@sheffield.ac.uk

Department of Computer Science
University of Sheffield

SO FAR

- part-of-speech tagging
- syntactic dependency parsing
- and algorithms for learning models from data

IN THIS LECTURE

Word senses and their disambiguation

- definitions
- resources
- three ways to disambiguate:
  - supervised
  - knowledge-based
  - unsupervised

WHAT IS A WORD SENSE?

A discrete representation of an aspect of a word's meaning:

- sense1: a *bank* is a financial institution and a financial intermediary...
- sense2: a raised portion of seabed or sloping ground...

Usually, by word in this contex we mean **lemma**, i.e. how the word is found in dictionaries:

- banks -> bank
- sung -> sing

Words with different parts of speech tags (verb, noun, etc.) have different lemmas

WORDS AND THEIR SENSES

- I deposited the cheque at my *bank*.

versus:

- Fishing from the river *bank* is prohibited.

Homonymy: same spelling and pronunciation, different unrelated senses (their translations are usually different)

Homographs: same spelling (*bass* guitar vs sea *bass*)

Homophones: same pronunciation (*right* and *write*)

POLYSEMY

- I deposited the cheque at my *bank*.
- The *bank* is around the corner.
- I volunteer at the blood *bank*.

Polysemy: like **homonymy**, but with related senses.

- *Beethoven* wrote the Moonlight Sonata.
- I like *Beethoven*.

Metonymy: special type of **polysemy** when senses are aspects of the same concept (composer vs pieces by).

DETECTING POLYSEMY

- Which of those flights *serve* breakfast?
- Does Midwest Express *serve* Philadelphia?
- Does Midwest Express *serve* breakfast and Philadelphia?

The last construnction is called **zeugma**. It is ill-formed, suggesting that *serve* has two distinct senses.

Polysemy vs. *homonymy*: helpful distinction but not black & white. How (un-)related are blood *banks* and financial *banks*?

SYNONYMY

Words with (nearly) identical meaning, e.g. *couch* and *sofa*. More formally, the can substitute each other in a sentence without changing its truth value. (propositional meaning)

Remember:

- synonymy is among senses not words:
  - How *big/large* is this plane?
  - My *big/large* brother is lying.

  big has the sense of being older, *large* doesn't.
- No absolute synonymy; there are reasons why we say *H2O* instead of *water*: text genre, politeness, slang, etc.

OTHER RELATIONS BETWEEN SENSES

antonyms: opposites with respect to one aspect of meaning: *hot/cold*, *rise/fall*, *up/down*, etc.

hyponym/hypernym: one sense more/less specific than the other:

- *car* is a **hyponym** of *vehicle*
- *fruit* is a **hypernym** of *mango*

meronymy: part-whole relation: *wheel/car*, *handle/door*, etc.

WORDNET

A publicly available database of words (lemmas) annotated with senses and relations among them, e.g.:

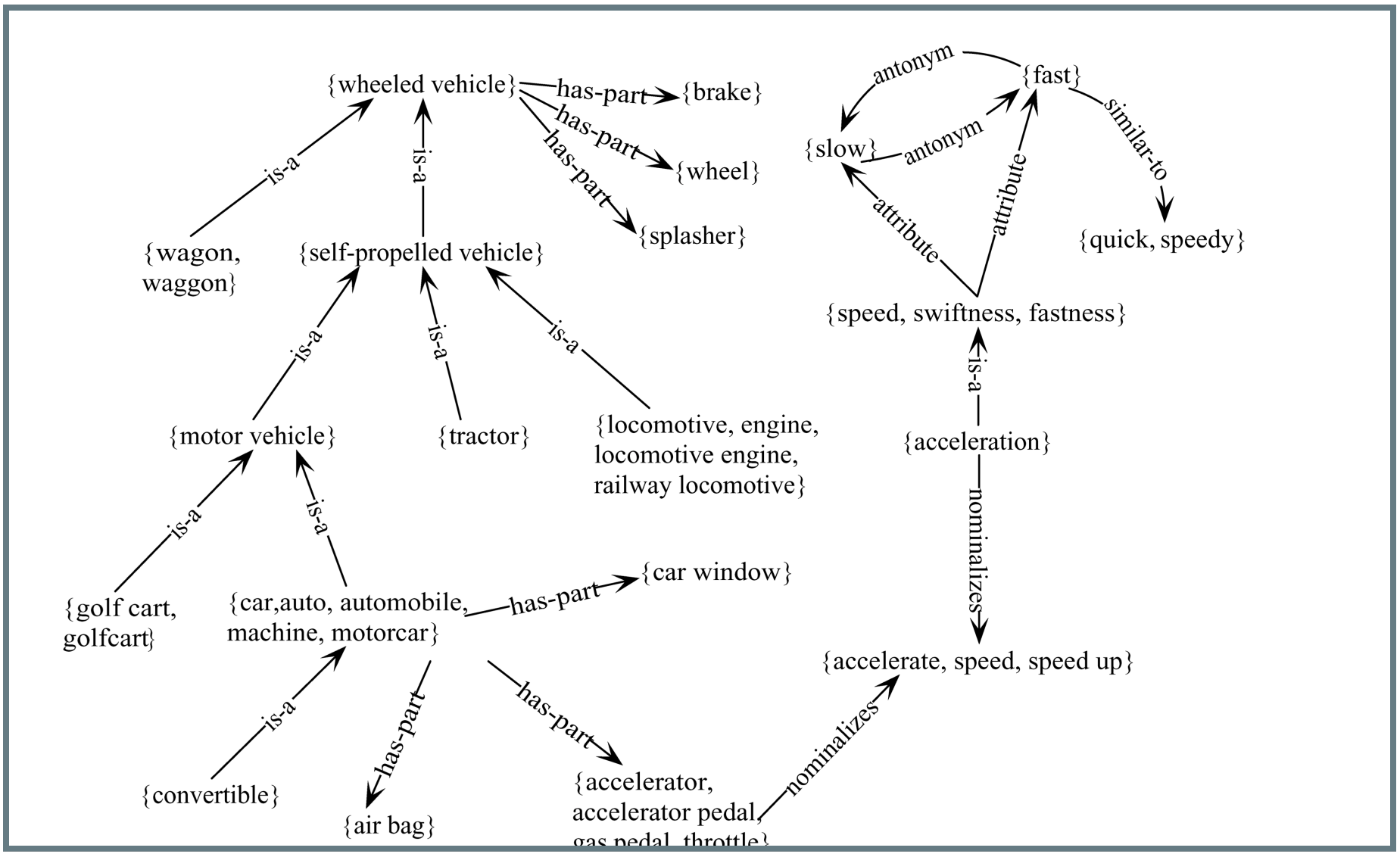
The noun *bass* has 8 senses in WordNet:

1. $bass^1$: the lowest part of the musical range
2. $\{sea\ bass^1, bass^4\}$: the lean flesh of a saltwater fish of the family Serranidae, etc.

gloss: a dictionary-style definition of a sense, e.g. *the lowest part of the musical range*

synset: a set of near-synonymous senses, e.g. $\{sea\ bass^1, bass^4\}$

WORDNET GRAPH



WORDNET

Most commonly used resource for senses and their relations

Originally for English, but now available for many languages: http://globalwordnet.org/

Coverage issues:

- 147,278 unique words, is that all the words?
- what about domain-specifc usage, e.g. *gene expression*?

SUPERVISED WORD SENSE DISAMBIGUATION

Given a set of words in context annotated with senses, e.g.:

WordNet Sense	Spanish Translation	Roget Category	Target Word in Context
bass^4	lubina	FISH/INSECT	...fish as Pacific salmon and striped **bass** and...
bass^4	lubina	FISH/INSECT	...produce filets of smoked **bass** or sturgeon...
bass^7	bajo	MUSIC	...exciting jazz **bass** player since Ray Brown...
bass^7	bajo	MUSIC	...play **bass** because he doesn't have to solo...

learn a model (one per word/lemma) to predict the sense.

SUPERVISED WSD

Use you favourite classifier: perceptron, naive Bayes, etc.

Feature vector with:

- collocational: previousWord, nextWord, previous2Words, next2Words, PoS tags, etc.
- bag-of-words: as above, but without encoding their position relative to the word

Intuition: like text classification, but focusing on a word

KNOWLEDGE-BASED WSD

Supervised WSD assumes annotated examples for each word for all its senses, will not scale to 1000s of words.

Each word sense has a definition:

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"

Given the word in a new sentence:

"The *bank* can guarantee deposits which eventually cover..."

find the sense with the max word overlap (simplified Lesk)

UNSUPERVISED WSD

New word senses emerge, dictionaries get out of date.

Solution: Unsupervised WSD, a.k.a. word sense **induction**

- take words in context and extract their feature vector as in supervised WSD
- use any clustering algorithm
- words are now clustered according to their context

The clusters are related to the senses.

But no labels in, no labels out: $cluster^1$ is not $bass^1$.

CLUSTERING WITH K-MEANS

Input: $\mathcal{X} = \{\mathbf{x}^1 \ldots \mathbf{x}^{\mathcal{N}} \in \Re^D\}$, *clusters* $\mathcal{K}$

initialize cluster means $\mu_1, \ldots \mu_K \in \Re^D$

while *not_converged* **do**

  for $\mathbf{x}^n \in \mathcal{X}$ **do**

    assign $\mathbf{x}^n$ to cluster $c_n = \arg\max_{k=1...K} cosine(\mu_k, \mathbf{x}_n)$

  for $k \in 1...K$ **do**

    set $\mu_k = mean(\{\mathbf{x_n} | c_n = k\})$

return $c_1 \ldots c_N$

means initialization matters, but random works

convergence ≈ assignments don't change

INTRINSIC WSD EVALUATION

For supervised and knowledge-based use accuracy.

For unsupervised we cannot use accuracy:

- clusters are not labels, how to compare
- assess how well cluster "correspond" to labels
- variety of clustering evaluation measures exist, no agreement on which one is the best

Baseline: most frequent sense for each word.

EXTRINSIC WSD EVALUATION

Use the word senses predicted to improve performance in a different task, e.g.:

- bag-of-word-senses in addition to bag-of-word features for text classification
- word-senses in addition to words for syntactic parsing
- etc.

Applicable to supervised and unsupervised WSD, since sense-ids and cluster-ids are only used as vector indices.

BIBLIOGRAPHY

- Jurafsky & Martin Chapter 18

COMING UP NEXT

Distributed word representations

(a.k.a. *word2vec*)

Is it always a good idea to use a discrete representation of meaning?