Computational Semantics and Pragmatics

Autumn 2012

Raquel Fernández
Institute for Logic, Language & Computation
University of Amsterdam

Where we are / Where we go

Last Friday:

- Supervised WSD: it assumes that every words has a finite set of discrete senses and in each context one sense is activated; we can use a sense-tagged corpus to learn to predict the right sense.

Today:

- What do psychologists tell us about senses / concepts? main psychological theories of concepts and word meaning
- Papers by Kilgarriff and Hanks
- Brief look at unsupervised at WSD

Friday:

- Distributional semantic models (DSMs)
- Choose a paper on DSMs you'd like to present (list of some possible papers given today)

Next Week:

- Paper presentations

Concepts and Word meaning

- Psycholinguists typically consider that a word gets it significance by being connected to a mental representation – a concept.
- This contrasts with traditional views in linguistics and philosophy of language, which are "externalists" or referential.
- According to cognitive psychologists, all the features that have been found to be true for concepts also apply to words.
- Does the private nature of concepts prevent them from being the basis for communication?
  - No, if we assume a collaborative and feedback-based model of communication.

Gregory L. Murphy (2002) *The Big Book of Concepts*, MIT Press.

Classical View of Concepts

The classic view of concepts, wingspread until the 1970's, is definitional:

- concepts are mentally represented as definitions: a definition gives characteristics that are necessary and jointly sufficient for membership in the category.
- every object either belongs or does not belong to the category (*law of excluded middle*)
- no distinction between category members: anything that meets the definition is an equally good member of the category.

Problems for the Classical View

There are theoretical arguments and empirical evidence against the classical, definitional view of concepts.

- Wittgenstein argues that most concepts can't be defined
  - if the classic view is correct, it should be possible to come up with the defining features of, say, games. But, is it?
- Empirical problems
  - Category membership is not discretely determined
    - borderline cases
    - members and non-members form a continuum
  - Not all category members are perceived equally:
    - Typical category members are the good examples - what you normally think of when you think of the category.

Typicality Effects

Differences in typicality are one of the most robust and reliable effects in categorization research.

- high reliability of typicality judgements, with over 95% agreement
- correlation between inconsistent category membership judgements and typicality ratings
- easier and faster identification and production of typical category members
- artificial category learning: typical items are learned to be members of a category earlier than atypical ones
- typicality influences the likelihood of drawing inferences
- ...

Typical vs. Atypical Items

Typicality is a graded phenomenon: typical items, moderately typical, atypical, borderline category members.

What makes items typical?

- Frequency? there isn't a simple correlation
- Family resemblance. Typical items. . .
  - tend to have the properties of other category members.
  - tend not to have properties of category nonmembers.

  Experiments have shown that
  - there is a correspondence between high typicality ratings and items with most common features in the category
  - items with greater overlap of features with other categories are harder to learn and rated less typical.

Rosch & Mervis (1975). Family Resemblances: Studies in the Internal. Structure of Categories, *Cognitive Psychology*, 7(4):573–605.

Alternatives to the Classical View

Two main theories that arose after the downfall of the classical view of concepts and which aim to explain typicality effects:

- Prototype theory
- Exemplar theory

Prototype Theory

Eleanor Rosch was one of the main critics of the classical view of concepts and the proponent of an early alternative.

According to this alternative (family of theories) the representation of a category is based on the notion of prototype.

- a prototype can be thought of as a *summary representation*
  - features that are usually found in the category members, weighted
  - "contradictory" features may be included with different weights
  - categorization criterion based on feature weights
  - no feature is required to be present
- this view can explain the lack of definitional features, borderline cases, faster categorization of typical items, etc.

Exemplar View

The exemplar view rejects the idea that there is a representation that encompasses an entire concept.

According to this view, a concept is just the set of instances of that concept that one person remembers.

To categorise new items, we weight them by how similar they are to the items in our memory.

- the most typical items are those that are more similar to many category members
- borderline cases are those that are almost equally similar to remembered category members and non-members
- typical items would be categorised faster because it is easier to find evidence

Summing Up

None of these theories suffers from the problems of the classical view:

- category membership is a matter of degree - the theories rely on the idea of similarity, which is inherently continuous
- this gradation of similarity leads to typicality differences

Prototype theory does not deny that some exemplars may be kept in memory, but it proposes that in general people rely on summary representations of the entire category.

Exemplar theory rejects the existence of a summary representation, but must agree to the fact that information from remembered exemplars interacts with general knowledge that may not have been acquired via direct experience.

Concepts and Word meaning (again)

Even if we accept a conceptual view of word meaning, the relationship between concepts and words is complex.

- learning can happen in both directions: first concept then word for it, first word then right concept for it.
- Polysemy is challenging: the mapping between words and concepts is not 1-to-1 and can be dynamic (with stored and derived meanings)

What about distributional semantic models? Are they (in)compatible with a conceptual view of meaning?

Readings

Adam Kilgarriff (1997) I don't believe in word senses.
Computers and the Humanities, 31:91-113.

Patrick Hanks (2000) Do Word meanings exist?
Computers and the Humanities, 34:205–215.

Unsupervised WSD

Why use unsupervised learning for WSD?

- It is expensive and difficult to build hand-labelled corpora.
- Hand-labelled senses may not be theoretically sound.
  Recall Kilgarriff's arguments:
  - defining a fix set of word senses may be impossible, and would at any rate be a domain-dependent task.
  - word senses should be reduced to abstractions over clusters of word usages.

In unsupervised WSD we do not start with a set of human-defined senses – the "senses" are created automatically from the instances of each word in the training set.

$\Rightarrow$ we can use a version of a DSM where we compute context vectors for each token of interest, i.e. for each usage, instead of computing vectors for *types* of target terms.

Unsupervised WSD

Training: creating "senses" from usages

- For each token t_w of word w in a corpus, compute a context vector $\mathbf{c}_{t_w}$
- Use a clustering algorithm to cluster the vectors into groups or clusters; each cluster defines a sense of w
- Compute the vector centroid (the average or arithmetic mean) of each cluster; each centroid is a vector $\mathbf{s}_{w_i}$ representing that sense of w

Prediction: disambiguating a token t_w of w by assigning it a sense

- Compute a context vector $\mathbf{v}_{t_w}$ for t_w
- Retrieve all sense vectors for w
- Assign to t_w the sense represented by the sense vector $\mathbf{s}_{w_i}$ that is closest to $\mathbf{v}_{t_w}$

This procedures requires a clustering algorithm and a distance metric to compare vectors.

Clustering

Clustering is a general term referring to the task of classifying a set of objects into groups (clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters.

Several clustering algorithms exist. Two common techniques are:

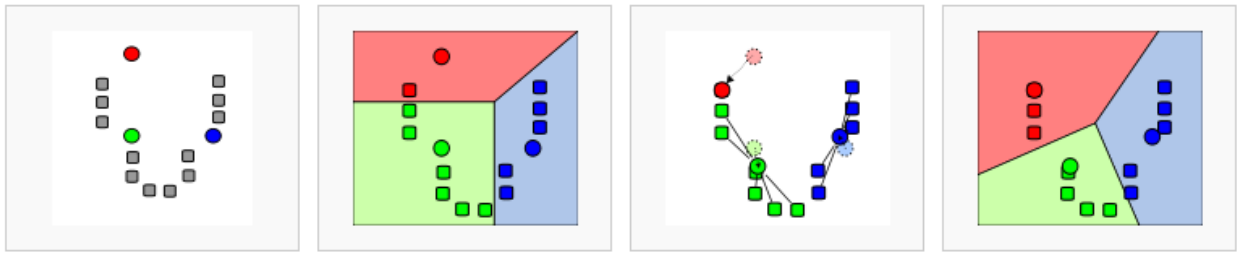
- k-means clustering
- Agglomerative hierarchical clustering

We will briefly review the basic steps involved in these two types of algorithms. For further details, you can consult these reference:

Manning & Schütze (1999) Foundations of Statistical Natural Language Processing, ch. 14: *Clustering*, MIT Press.
Jain, Murty & Flynn (1999) Data Clustering: A Review, *ACM Computing Surveys*, 31:264-323.

k-means Clustering: Basics

1. assume a certain number k of clusters;
2. select k objects that are as distant as possible from each other; these are the starting centroids of the clusters;
3. assign each remaining object to the cluster whose centroid is the closest;
4. when all objects have been assigned, recalculate the positions of the k centroids.
5. Repeat Steps 3 and 4 until the centroids are stable.

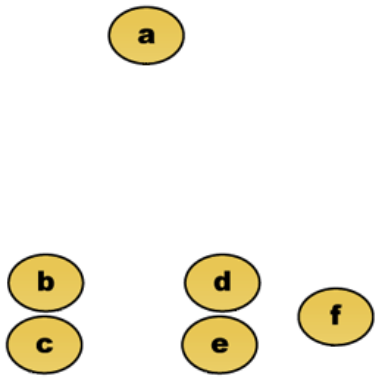


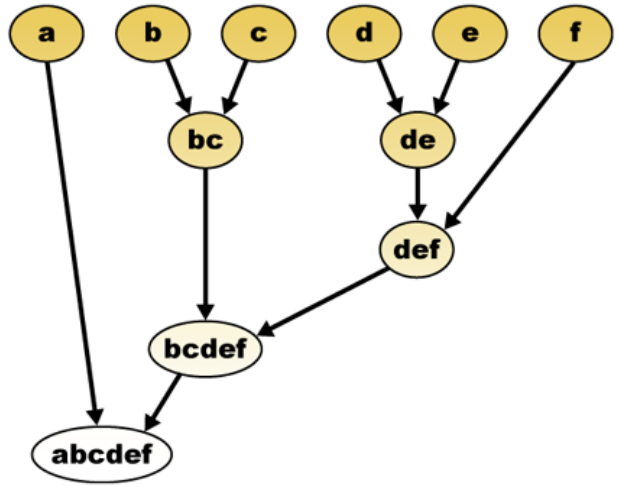
Picture from Wikipedia `http://en.wikipedia.org/wiki/K-means_algorithm`
There seems to be a mistake with red cluster, but good enough for illustration

Agglomerative Clustering: Basics

1. assign each training instance to its own cluster
2. compute the distance between the clusters and merge the most similar pair of clusters
   * similarity between clusters can be computer by taking the shortest, the longest, or the average distance
3. repeat step 2 until either a specified number of clusters is reached or the clusters have some desired property.

By repeating step 2 until all items belong to the same cluster we end up with a tree that can be cut at the desired level of specificity.





Evaluation of Unsupervised Predictions

In unsupervised learning we don't have a gold standard or ground truth against which we can compare the output of our system. Therefore evaluation can be tricky...

Some possibilities include:

- extrinsic evaluation: is the system's output positively evaluated by human judgements?
- in vivo evaluation: does the output of the system improve the performance of a larger task? (e.g. does unsupervised WSD improve machine translation?)
- if an annotated corpus exists, we can also do an intrinsic evaluation (such as those in supervised learning). For instance, for WSD:
  - map each cluster (induced sense) to the predefined sense that in the training set has most word tokens overlapping with the cluster; or
  - for all pairs of usages of a word in the test set, test whether the system and the hand-labels consider the pairs to have the same sense or not.

Distributional Semantic Models

Two overview papers:

Alessandro Lenci (2008) Distributional Semantics in Linguistic and Cognitive Research, *Italian Journal of Linguistics*, 20(1):1–30.

P. Turney and P. Pantel (2010) From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, 37:141–188.

Possible papers for presentations

- Erk & Padó (2010) Exemplar-Based Models for Word Meaning In Context, ACL.
- Baroni & Zamparelli (2010) Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, EMNLP.
- Grefenstette & Sadrzadeh (2011), Experimental support for a categorical compositional Distributional model of meaning, EMNLP. [**Phong**]
- Boleda, Padó, & Utt (2012) Regular polysemy: a distributional model. *SEM.
- Bruni, Boleda, Baroni, & Tran (2012) Distributional semantics in technicolor. ACL.
- Socher, Huval, Manning, & Ng (2012) Semantic Compositionality Through Recursive Matrix-Vector Spaces. EMNLP. [**Philip**]
- Huang, Socher, Manning, & Ng (2012) Improving Word Representations via Global Context and Multiple Word Prototypes. ACL

Where to look for further papers:

ACL, NAACL, EACL, EMNLP, IWCS, *SEM, plus workshops at these conferences
Web pages of authors and their research groups

Come up with a proposal of a paper you want to present by Friday.