Statistical NLP: Lecture 10

Lexical Acquisition

(Ch 8)

Goal of Lexical Acquisition

- ***Goal:*** To develop algorithms and statistical techniques for filling the holes in existing machine-readable dictionaries by looking at the occurrence patterns of words in large text corpora.
- Acquiring collocations and word sense disambiguation are examples of lexical acquisition, but there are many other types.
- ***Examples of lexical acquisition problems***: selectional preferences, subcategorization frames, semantic categorization.

Why is Lexical Acquisition Necessary?

- Language evolves. i.e., new words and new uses of old words are constantly invented.
- Traditional Dictionaries were written for the needs of human users. Lexicons are dictionaries formatted for computers. In addition, lexicons can be useful if they contain quantitative information. Lexical acquisition can provide such information.
- Traditional Dictionaries draw a sharp boundary between lexical and non-lexical information. In NLP it may be useful to erase this distinction.

Lecture Overview

- Methodological Issues: Evaluation Measures
- Verb Subcategorization
- Attachment Ambiguity
- Selectional Preferences
- Semantic Similarity

Evaluation Measures

- Precision and Recall
- F Measure
- Precision and Recall versus Accuracy and Error
- Fallout
- Receiver Operating Characteristic (ROC) Curve

Verb Subcategorization (I)

- Verbs express their semantic categories using different syntactic means. A particular set of syntactic categories that a verb can appear with is called a subcategorization frame.
- Most dictionaries do not contain information on subcategorization frame.
- (Brent, 93)'s subcategorization frame learner tries to decide based on corpus evidence whether verb ***v*** takes frame ***f***. It works in 2 steps.

Verb Subcategorization (II)

Brent's Lerner system:

- ***Cues:*** Define a regular pattern of words and syntactic categories which indicates the presence of the frame with high certainty. For a particular cue c^j we define a probability of error ε_j that indicates how likely we are to make a mistake if we assign frame f to verb v based on cue c^j.
- ***Hypothesis Testing:*** Define the null hypothesis, H_0, as: "the frame is not appropriate for the verb". Reject this hypothesis if the cue c^j indicates with high probability that our H_0 is wrong.

Verb Subcategorization (III)

- Brent's system does well at precision, but not well at recall.
- (Manning, 93)'s system addresses this problem by using a tagger and running the cue detection on the output of the tagger.
- Manning's method can learn a large number of subcategorization frames, even those that have only low-reliability cues.
- Manning's results are still low and one way to improve them is to use prior knowledge.

Attachment Ambiguity (I)

- When we try to determine the syntactic structure of a sentence, there are often phrases that can be attached to two or more different nodes in the tree. Which one is correct?
- A simple model for this problem consists of computing the following likelihood ratio: $\lambda(v, n, p) = \log (P(p|v)/P(p|n))$ where $P(p|v)$ is the probability of seeing a PP with p after the verb v and $P(p|n)$ is the probability of seeing a PP with p after the noun n.
- Weakness of this model: it ignores the fact that other things being equal, there is a preference for attaching phrases "low" in the parse tree.

Attachment Ambiguity (II)

- The preference bias for low attachment in the parse tree is formalized by ***(Hindle and Rooth, 1993)***
- The model asks the following questions:
- Va_p: Is there a PP headed by p and following the verb v which attaches to v ($Va_p=1$) or not ($Va_p=0$)?
- Na_p: Is there a PP headed by p and following the noun n which attaches to n ($Na_p=1$) or not ($Na_p=0$)?
- We compute $P(Attach(p)=n|v,n)=P(Na_p=1|n)$ and $P(Attach(p)=v|v,n)=P(Va_p=1|v)\ P(Na_p=0|n)$.

Attachment Ambiguity (III)

- P(Attach(p)=v) and P(Attach(p)=n) can be assessed via a likelihood ratio λ where $\lambda(v, n, p) = \log (P(Va_p=1|v)\ P(Na_p=0|n))/ P(Na_p=1|n)$
- We estimate the necessary probabilities using maximum likelihood estimates:
- $P(Va_p=1|v)=C(v,p)/C(v)$
- $P(Na_p=1|n)=C(n,p)/C(n)$

General Remarks on PP Attachment

- There are some limitations to the method by Hindle and Rooth:
- Sometimes information other than v, n and p is useful.
- There are other types of PP attachment than the basic case of a PP immediately after an NP object.
- Other types of attachments: N N N or V N P. The Hindle and Rooth formalism is more difficult to apply in these cases because of data sparseness.
- In certain cases, there is ***attachment indeterminacy.***

Selectional Preferences (I)

- Most verbs prefer arguments of a particular type (e.g., the things that bark are dogs). Such regularities are called ***selectional preferences*** or ***selectional restrictions***.
- Selectional preferences are useful for a couple of reasons:
  - If a word is missing from our machine-readable dictionary, aspects of its meaning can be inferred from selectional restrictions.
  - Selectional preferences can be used to rank different parses of a sentence.

Selectional Preferences (II)

- ***Resnik (1993, 1996)***'s idea for Selectional preferences uses the notions of ***selectional preference strength*** and ***selectional association***. We look at the ***<Verb, Direct Object>*** problem.
- ***Selectional preference strength***, ***S(v)*** measures how strongly the verb constrains its direct object.
- S(v) is defined as the ***KL divergence*** between the prior distribution of direct objects (for verbs in general) and the distribution of direct objects of the verb we are trying to characterize.
- We make 2 assumptions in this model: 1) only the head noun of the object is considered; 2) rather than dealing with individual nouns, we look at classes of nouns.

Selectional Preferences (III)

- The ***Selectional Association*** between a verb and a class is defined as the proportion that this contributes to the overall preference strength S(v).
- There is also a rule for assigning association strengths to nouns as opposed to noun classes. If a noun is in a single class, then its association strength is that of its class. If it belongs to several classes, then its association strength is that of the class it belongs to that has the highest association strength.
- Finally, there is a rule for estimating the probability that a direct object in noun class c occurs given a verb v.

Semantic Similarity

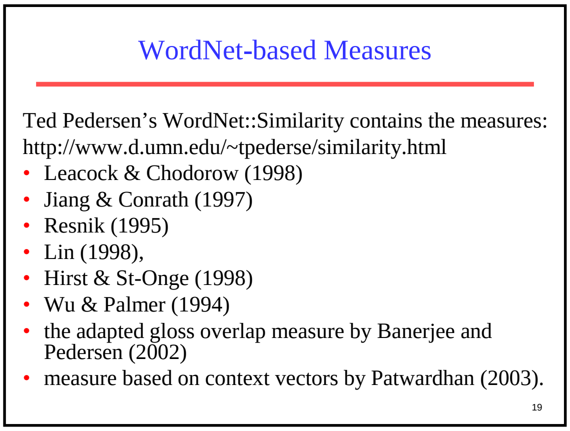
- Text Understanding or Information Retrieval could benefit much from a system able to acquire meaning.
- Meaning acquisition is not possible at this point, so people focus on assessing ***semantic similarity*** between a new word and other already known words.
- Semantic similarity is not as intuitive and clear a notion as we may first think: synonymy? Same semantic domain? Contextual interchangeability?
- Vector Space versus Probabilistic Measures

Vector Space Similarity

- Words can be expressed in different spaces: document space, word space and modifier space.
- Similarity measures for binary vectors: matching coefficient, Dice coefficient, Jaccard (or Tanimoto) coefficient, Overlap coefficient and cosine.
- Similarity measure for the real-valued vector space: cosine (normalized correlation coefficient, Euclidean Distance)

Probabilistic Similarity Measures

- The problem with vector space based measures is that, aside from the cosine, they operate on binary data. The cosine, on the other hand, assumes a Euclidean space which is not well-motivated when dealing with word counts.
- A better way of viewing word counts is by representing them as probability distributions.
- Then we can compare two probability distributions using the following dissimilarity measures (semantic distance): ***KL Divergence***, ***Information Radius*** (***Irad***) and ***L_1 Norm***.



WordNet-based Measures

Ted Pedersen’s WordNet::Similarity contains the measures:
http://www.d.umn.edu/~tpederse/similarity.html

- Leacock & Chodorow (1998)
- Jiang & Conrath (1997)
- Resnik (1995)
- Lin (1998),
- Hirst & St-Onge (1998)
- Wu & Palmer (1994)
- the adapted gloss overlap measure by Banerjee and Pedersen (2002)
- measure based on context vectors by Patwardhan (2003).