Where do we begin?

» Humans!
  » Only learners of natural language we know
    » Relatively fast learners
    » Mostly reliable
    » Robust
  » ... but we don't know how we do it.
  » There are some clues
    » Learning process (e.g. verb over-regularisation)
    » Generalisation capabilities (unseen words)
    » Separation of syntactic/semantic knowledge (aphasia)

» Innateness
  » We don't seem to get a great deal of help

What can we learn?

» Linguistic knowledge has been divided many ways, but the popular distinction is:
  » syntax – relations between the symbols themselves
  » semantics – relations of symbols to the real world
  » pragmatics – reasoning about symbols and users
  » ... but computers don’t have access to the real world
    » we are limited to information that is easily represented in computers.
    » therefore, the majority of learning focuses on syntactic knowledge.

» So we try to learn language from text!

Overview

» Motivation
» Resources, supervision and evaluation
» Simple language models: n-grams
» Learning Context-Free Grammars
» Compression and Alignment
» Supervision

Why Learn Languages?

» Interesting!
  » close ties with machine learning, but a very hard problem.
  » insights into human cognition.
» ... but also practical
  » NLP requires some structural analysis
  » Manually constructing structure is costly
    » Annotation projects take time
    » linguists cost a lot!
  » For many languages, no resources exist.
    » Better to start from a partially annotated corpus than do it all from scratch.
  » Manually constructed grammars rely on linguists’ decisions.
Problems with $n$-grams

- How much history?
  - We are really unlikely to see some sequences, especially longer ones
    - $P(\text{mat} \mid \text{“the cat sat on the”})$
    - Storage? Efficient retrieval? Unseen words? Impractical for lengths $> 4$
  - Less history provides less information, but is
    - $P(\text{mat} \mid \text{“on the”})$
    - Here, our choice of “mat” is only likely because of “cat” and “sat”

- Is linear history enough?
  - No, there is much more structure (long-distance dependencies)
    - “I saw the man in the park with the telescope”
    - “telescope” is made likely by “saw”, but not the intervening words
  - We need something more structured...

Learning Syntax

- We want to learn relations between symbols only.
- As with any learning problem, we must be able to:
  - take some learning source
  - learn something
  - evaluate what we have learned
- We can make some immediate observations on the learning problem by examining resources and evaluation.

Learning Syntax From Text Alone

In aligning themselves with the author, however, they might also be victims of a certain irony the character of which I shall reveal shortly.

A contemporary philosopher invited to consider relevant difficulties raised by modern urban reeds/crane might think to approach the issues from the direction of either the well-established traditions of social philosophy or aesthetics.

Among other things, I shall be concerned to suggest that no adequate treatment of the problems is possible unless both perspectives are adopted, or, better still, merged.

Architecture being the paradigm of a public act, its philosophical examination is an exercise in social aesthetics.

That these matters are currently a topic of considerable public and professional discussion is due in large part to the several widely reported occasions on which the Prince of Wales has castigated modern architects for producing bogus grandiose monstrosities they...

- This is not a natural language learning task!
  - humans don’t learn from text alone
  - punctuation? Not present in speech...

The simple approach: $n$-grams

- Classical conditioning, or behaviourism
  - Decisions based on prior experience
- Words depend only on what we have already seen
  - “the cat sat on the”
  - predictive text? “I am nearly [good/home/gone/...]”
- Probability of a word conditioned on previous words:
  - $P(\text{mat} \mid \text{“the cat sat on the”})$
  - Estimate these probabilities by looking at a corpus:
    - $P(\text{mat} \mid \text{“the cat sat on the”}) = f(\text{“the cat sat on the”}) / f(\text{“the cat sat on the”})$
  - This works, up to a point:
    - $P(\text{mat} \mid \text{“the cat sat on the”}) >> P(\text{of} \mid \text{“the cat sat on the”})$
Evaluation [1]: Comparing Grammars

- Evaluate the grammar
  - against other grammars
    - existing grammar may be flawed
    - many potential grammars with same expressive capabilities
    - requires a grammar, so the target is simply to reproduce the existing grammar
    - grammatical units require symbol matching, and it's unlikely that the symbols of the learned structure would be the same as those in the original grammar
  - by human eye
    - contentious
    - really time consuming
    - requires knowledge of representation
      - All the problems we have with manual grammar construction!

Learning Resources: Corpora

- Corpora of language usage
  - Collections of natural language data ... as used by real people
  - Lots of data!
    - British National Corpus has 100 million words
    - Web has several billion documents
  - ... but it is very sparse data
    - Zipfian distribution
    - 486507 of the 938972 BNC word types occur once
  - Many corpora also have extra information:
    - part-of-speech (POS) tags are common (because taggers are cheap).
    - phrase-structure corpora (treebanks) also exist, but they are small and quite rare.
    - other annotation: prosody, emphasis, etc. Anything that cannot be automated is rare.

Evaluation [2]: Comparing Parser Output

- Evaluate the quality of parsed data (PARSEVAL)
  - obtain a parsed corpus
  - learn a grammar for that corpus
  - parse the corpus using the learned grammar.
    - compare the resulting phrase-structure
  - Good because...
    - removes the necessity to match unit labels
    - removes the peculiarities of particular grammar formalisms
  - But still quite bad:
    - expensive for computation (corpora are large; parsing is non-trivial)
    - requires a parsed corpus (and they are rare and expensive)
    - the correct interpretation requires parsing knowledge, as well as grammar knowledge

Learning Resources: Supervision

- Supervision should be a way to restrict learning
  - should be very useful, given the learning task at hand
    - directs us towards correct structures
    - reduces data sparsity
  - BUT, majority of supervision is a result of automated process – e.g. Part-of-speech tags.
    - These processes are not 100% accurate, and so we are learning flawed data.
    - Syntax is often integral to improving the accuracy of these processes (POS tags are defined by syntactic roles), so we should be trying to learn those features that are present in supervision!
  - Do humans get these forms of supervision?
Representing Syntax

- We need to find a representation that allows us to capture the intricacies of language.
  - should accord with human intuitions
- A reasonable starting point...
  - human-constructed grammars
    - Context-free grammars (discussed here)
    - Categorial grammars
    - Dependency grammars
- BUT... this is controversial!
  - no gold standard, little agreement between linguists, competing theories.

Context-Free Grammars (CFG)

- Example:
  - NP → DET NOUN
  - DET → the | a | an
  - NOUN → cat | mat | dog | wag
- In a treebank (Penn Treebank format):
  - (NP (DET the)(NOUN cat))
  - (S (NP (DET the)(NOUN cat)) (VP (V cat) (PP (PREP en) (NP (DET the) (NOUN mat))))
- Problems with representation:
  - There are a few peculiarities in natural language: in English, "Respectively"
  - Lexical information? Because we generalise without context, we lose information that is potentially useful for parsing.
  - ... but it's a common representation, and efficient to parse

Evaluation [3]: Generative Capacity

- Using the learned grammar, generate a large number of sentences, and ask native speakers to assess their correctness
  - Contentious (do the native speakers all agree?)
  - Time consuming
  - Humans tend to interpret structure, even in meaningless units. This can lead to biased results (a criticism of many animal language experiments).

Resources and Evaluation: Summary

- Evaluating learned structure is controversial, unreliable and computationally (or manually) expensive
- Corpora are invaluable, but suffer from sparse data.
- As a consequence:
  - General purpose, brute-force techniques are unlikely to be successful:
    - Genetic Algorithms have been applied to grammar learning, but only for very simple grammars, and with limited success. Few other Evolutionary techniques have been used.
    - Syntax-free learning is made very difficult.
- So...
  - We really need a tightly constrained model for learning
Learning by Compression [1]

- Principle:
  - A minimal grammar is the best

- Process:
  - Lossless compression – try to find the shortest description for a corpus
    - Minimum Description Length
  - Larger units are only retained if they provide a gain in compression
  - Replace combinations of symbols that co-occur with new symbols

- Syntax is typically linear (concatenative)

Learning a Context-Free Grammar

- We must decide where brackets should go:
  - 
  
  \[
  (NP (DET (NOUN cat))) \quad (VP (V sat) \quad (PP (PARP on) \quad (NP (DET the) \quad (NOUN mat))))
  \]

- why (on (the mat)), not ((on the) mat)

- ... but there are some problems
  - CFG allows recursive productions (e.g. NT1 → NT1)
  - CFG allows infinite unary productions (e.g. NT1 → NT2)
    - Both the above allow infinite number of expansions (very bad for learning)
  - CFG allows arbitrary numbers of symbols in RHS
    - Very large number of candidate structures even without recursion and unary productions!

- Brute force techniques will not work (sparse data)

Learning by Compression [2]

- Algorithm
  - Enumerate all adjacent word pairs in a corpus, compiling probability for each pair
  - Pick most probable pair
  - Merge pair into a new symbol

- Example:
  - "The cat sat on the mat" contains the frequent sequence "on the"
  - NT1 → on the
  - the cat sat NT1 mat

- Results:
  - Not encouraging. We can see that frequent unit combinations do not follow the boundaries of constituents.

Probabilistic Learning

- Idea: learn structures that we observe most often
  - Problem: word data is sparse, so there are few frequent structures

- We need to know \textbf{where} to look!
  - We cannot enumerate all possible structures.
  - Too many variables to learn by enumeration: constituent boundaries and constituent type must be simultaneously proposed.
  - This is a \textbf{BOOTSTRAPPING} problem: we need some structure to guide our examinations; however, there is no such structure.

- Some common approaches:
  - Compression (and Minimum Description Length)
  - Alignment (and interchangeability)
Learning by Alignment [2]

- An example:
  
  the boy **saw** the dog
  
  the girl with the telescope **saw** a woman on the beach

- Gives us the following classes/rules:
  
  $X \rightarrow \text{boy}$
  
  $X \rightarrow \text{girl with the telescope}$
  
  $Y \rightarrow \text{dog}$
  
  $Y \rightarrow \text{a woman on the beach}$

- And allows us to rewrite the sentence as:
  
  the $X$ **saw** $Y$

- We can then continue to learn from the new sentence.

Problems with Compression

- Good for...
  
  - Compressing data (not really related to NLP!)
  
  - Discovering word boundaries in unsegmented text
  
  - Finding morphemes and multi-word expressions

- But not... learning CFGs!
  
  - “the cat” and “the mat” are phrases of the same type, but compression will never allow this, because it requires information loss.
  
  - Equivalence is a crucial issue here: our grammar must provide a mechanism for equivalence.

- So, compression alone is not useful

Learning by Alignment [3]

- Benefits:
  
  - Designed for equivalence, so is capable of learning syntax

- Problems:
  
  - What counts as a context? (Bootstraping revisited)
  
  - Probable contexts are not necessarily good contexts
  
  - Sparse data problems – lots of infrequent potential contexts
  
  - Misalignment is possible, as words have different senses

Learning by Alignment [1]

- Principle:
  
  - Similar constituents can occur in similar contexts
  
  - (The reverse) If we can observe a “similar context”, we can extract “similar constituents”

- Process:
  
  - Identify context
  
  - Find every word sequence that occurs within that context
  
  - Propose an equivalence class to cover all those word sequences

- Syntax:

  
  LEFT  |  w1,w2,...,wn  |  RIGHT
Some ideas for the future...

- Hybrid models for learning
  - Alignment strategies that incorporate dependency information.
- Supervision that is more closely related to human learning
  - Prosody, phonology, speech limitations.
  - Situation, embodiment and attachment.
- Incremental syntax
  - Increasing syntactic complexity when it becomes insufficient
    - (Another bootstrapping problem)

Performance of learners

- Unsupervised systems do not perform well
  - EVALB: unlabelled precision / recall / cross-brackets
    - precision is proportion of learnt constituents that are correct (in treebank)
    - recall is proportion of correct constituents that are identified by the learner
    - cross-brackets is the number of overlapping structures per sentence, on average
  - ABL (van Zaanen): 43.6% / 35.6% / 2.12
  - EMILE (Adriaans): 51.6% / 16.8% / 0.84
  - Directed Alignment: 35% / 5.7% / 0.12

- This is not surprising: treebank parses are created by humans, using information outside syntax.

References


