Empirical work on (un)supervised learning

Learnability and Language Acquisition

Alexander Clark

Department of Computer Science
Royal Holloway, University of London

Friday

Outline

Introduction

POS induction

Word segmentation

Morphology

- Unsupervised learning

Models

- Implemented computational models
- Typically use heuristics
- Tested on naturally occurring data
- Evaluated:
    - Against gold standard annotations
    - Objectively

Motivations

Engineering motivations

- Build efficient language processing systems
- Annotation bottleneck

Cognitive modelling

Understand human language processing/acquisition

Do these overlap/interact at all ?

Tasks

- Word segmentation
- Morphology learning
- Phonology
- POS tagging
- Syntax – dependency parsing, and constituent structure.

Tasks

- Word segmentation
- Morphology learning
- Phonology
- POS tagging
- Syntax – dependency parsing, and constituent structure.

Two types

- Partially solved problems: try to find a model that matches developmental evidence
- Unsolved problems

Methodology

Supervised learning

Learning from labeled data
Sometimes appropriate: eg stress learning/inflectional morphology

Unsupervised learning

Children don't get parse trees, or segmented speech.

POS induction

- Historically the first attempts were made on this (Lamb, 1969) developmentally it is an early stage.
- Task is to determine the set of lexical categories for a language, and which words belong to which classes.
- Noun, Verb, Adj, etc.
- Often an initial phase before more sophisticated learning algorithms.

Semantic bootstrapping

Good case study in *What else* arguments.

- Pinker (1984): Children first learn the semantics of the words, and then use that, plus innate knowledge, to identify the syntax.
- Primary argument in favour of this: there is no other possible explanation: there are no other algorithms that could work.

Distributional learning

In fact there are many different algorithms that can learn this:

- Brown et al. (1992) “Class-based n-gram models of natural language”
- Ney, Essen, Kneser (1994) “On structuring probabilistic dependence in stochastic language modelling”

Distributional learning

In fact there are many different algorithms that can learn this:

- Brown et al. (1992) “Class-based n-gram models of natural language”
- Ney, Essen, Kneser (1994) “On structuring probabilistic dependence in stochastic language modelling”
- Distributional methods: Schütze, Finch and Chater (1992,1993)

Distributional learning

In fact there are many different algorithms that can learn this:

- Brown et al. (1992) “Class-based n-gram models of natural language”
- Ney, Essen, Kneser (1994) “On structuring probabilistic dependence in stochastic language modelling”
- Distributional methods: Schütze, Finch and Chater (1992,1993)
- Clark (2003) incorporates morphological information.

Ney Essen clustering

- Pick a number of clusters: say 32
- Randomly divide all of the words into 32 groups.
- For each word, move it to the class that would cause the largest increase in likelihood of a certain model.
  - Class based bigram model.
    $P(w_i|w_{i-1}) = P(w_i|c(w_i))P(c(w_i)|c(w_{i-1}))$
- Repeat until no word changes class.

Sample clusters

600K words of CHILDES data

you	the	a	it	that
we	your	some	me	this
they	my	very	him	her
Judy	his	Mommy's	them	those
littler	our	another	em	these
Tiggers	their	any	something	which
flies	Nina's	an	gone	wrong
ravens	Maggie's	many	careful	Mister
Jenny	Timmy's	Daddy's	ya	e
what'll	Mrs	kinda	yourself	whose

Sample clusters II

600K words of CHILDES data

to	what	I	with	and
doesn't	where	let's	for	so
hasn't	how	I'll	of	cause
ill	why	we'll	all	but
spoiled	who	lemme	from	if
shared	c	whatcha	isn't	o
punching	wha	you'll	about	or
nor	god	you've	eating	then
taxis	Jeremy	it'll	by	when
happily	Minoru	we've	putting	h

Sample clusters III

600K words of CHILDES data

come	gonna	see	up	one
look	not	have	down	baby
sit	just	want	out	boy
stand	gon	like	right	ball
stay	getting	know	back	book
lay	really	got	too	kitty
climb	still	think	off	girl
lie	feel	need	ere	house
slow	almost	try	again	water
calm	such	had	away	head

Evaluations

Three possibilities

- Subjective: looks good to me
- Objective but theoretically tied: comparison to gold standard part of speech tags (but which?)
- Objective and theoretically neutral: perplexity of a derived language model

Evaluations

Three possibilities

- Subjective: looks good to me
- Objective but theoretically tied: comparison to gold standard part of speech tags (but which?)
- Objective and theoretically neutral: perplexity of a derived language model
    - A measure of the ability of the model to predict the next word.
    - Perplexity of 189 means that the model can predict it as though there were only 189 equally likely words.

Incorporating some morphological information

Clark, EACL 2003

- Morphological information is key to determining which class something is in.
- We can augment these algorithms to use these pieces of information
- Tested on English and 6 Eastern European languages; written corpora.

Incorporating some morphological information

Clark, EACL 2003

- Morphological information is key to determining which class something is in.
- We can augment these algorithms to use these pieces of information
- Tested on English and 6 Eastern European languages; written corpora.
- Very accurate classes in English
  - classes that have the suffix -ed.
  - classes that have the suffix -ing.
  - classes that consist only of numbers.

Perplexity on WSJ data

Clusters	32	64	128	32	64	128
	Training			Test Data		
Baseline	854	760	673	890	795	711
D0	**479**	**380**	**316**	692	585	529
D5	502	417	355	556	469	412
DF	484	386	325	652	516	462
DM	494	406	335	620	523	464
DMF	495	392	338	**553**	**462**	**409**

What is wrong with Pinker's argument?

- Appeal to intuition about learning algorithms.
- Intuition of non-experts (machine learning experts) is of low value!

What is wrong with Pinker's argument?

- Appeal to intuition about learning algorithms.
- Intuition of non-experts (machine learning experts) is of low value!
- Intuition of experts is only slightly more valuable.
- If we do not know an algorithm to do X, this is no argument that there are no algorithms to do X.
- The poverty of the imagination/ *argumentum ad ignorantium*

Word segmentation

There are no word boundaries in continuous speech.

- Safran, Aslin and Newport demonstrated that infants can do statistical segmentation of nonsense syllables, using only transitional probabilities.
- Computational linguistics has huge numbers of papers in segmentation of Chinese and Japanese.
- Unsupervised algorithms work very well: (Goldwater, Griffiths and Tennenbaum, ACL 2006)

Unsupervised learning of Morphology

- Very popular area of research at the moment. (Morpho-challenge)
- Schone and Jurafsky, Gaussier, . . .
- Unsupervised segmentation of words : Goldsmith (2001)
    - Try to separate words into morphemes using just a list of words
    - *walk*, *walked*, *walking*,
    - *walk* plus *ed*, *ing* . . .
    - Using a Minimum description length method.

Unsupervised learning of Morphology

- Very popular area of research at the moment. (Morpho-challenge)
- Schone and Jurafsky, Gaussier, . . .
- Unsupervised segmentation of words : Goldsmith (2001)
  - Try to separate words into morphemes using just a list of words
  - *walk*, *walked*, *walking*,
  - *walk* plus *ed*, *ing* . . .
  - Using a Minimum description length method.
- View of words as concatenation of morphemes
- What about non-concatenative morphology?
  - Arabic broken plural, vowel harmony etc.

Supervised learning of morphology

- Input is a list of pairs of words

run	ran
walk	walked

Supervised learning of morphology

- Input is a list of pairs of words

run	ran
walk	walked

- Learning the transduction from uninflected to inflected form.
- Test on new words and see if it correctly generalises.

Basic approach

Clark (2001,2002)

- Start with a simple stochastic finite state transducer
- Randomise it
- Train it to convergence with the EM

Basic approach

Clark (2001,2002)

- Start with a simple stochastic finite state transducer
- Randomise it
- Train it to convergence with the EM
- Smoothing
- Model splitting

Basic approach

Clark (2001,2002)

- Start with a simple stochastic finite state transducer
- Randomise it
- Train it to convergence with the EM
- Smoothing
- Model splitting
- More advanced model with a MBL component performs better.

English past tense

The fruit fly of linguistics

- 20,000 tokens of phonetically transcribed data for training and test data
- Drawn from a natural distribution

English past tense

The fruit fly of linguistics

- 20,000 tokens of phonetically transcribed data for training and test data
- Drawn from a natural distribution
- 19991 correct which is (99.96%),
- 1571 types, 1567 correct (99.74%).
- 123 of these types were not in the training data.

English past tense

The fruit fly of linguistics

- 20,000 tokens of phonetically transcribed data for training and test data
- Drawn from a natural distribution
- 19991 correct which is (99.96%),
- 1571 types, 1567 correct (99.74%).
- 123 of these types were not in the training data.
- The four errors were *withhold*, *thrust*, *bind*, and *ring* .
- *withholded*, *thrusted*, *binded*, *rang*

English past tense

Classes derived

i	States	Words		λ
0	6	22	bet, shed	1.0
1	8	727	$+d$	0.90
2	8	281	$+t$	0.89
3	10	434	$+ed$	0.92
4	10	1	fly	1.0
5	12	26	break, draw	1.0
...				
24	20	2	sell, tell	1.0
25	20	2	go, undergo	1.0
26	22	1	leave	1.0
27	22	1	lose	1.0

Arabic broken plural

Singular	Plural	Type
babr	bubuur	broken
film	?aflaam	broken
sultan	salaatin	broken
sawwaaq	sawwaaqun	sound

Results on other languages

Data Set	CV	Models	CL	MBLSS
LING	10	1	61.3 (4.0)	85.8 (2.4)
	10	2	72.1 (2.0)	79.3 (3.3)
EPT	No	1	59.5 (9.4)	93.1 (2.1)
NAKISA	10	1	0.6 (0.8)	15.4 (3.8)
	10	5	9.2 (2.9)	31.0 (6.1)
	10	5	11.3 (3.3)	35.0 (5.3)
GP1	10	1	42.5 (0.8)	70.6 (0.8)
MCCARTHY	10	5	1.6 (0.6)	16.7 (1.8)
SLOVENE	No	1	63.6 (28.6)	98.9 (0.8)

Goldsmith (2001)

Task

Unsupervised segmentation of words into morphemes
Identifying sets of suffixes that form inflectional classes

Example

walk, walked, walks, walking . . .
Stem: walk
Suffixes -0, -s, -ed, -ing

Input

A list of words; no semantics

Closely related to word segmentation

Approach

Minimum Description Length (Rissanen, 1989)

Find a compact description of the data:
Size of a grammar plus size of the data wrt to the grammar
Closely related to Bayesian approaches

And it works!

Very well on concatenative languages
See paper.

Very good paper