Mutual Constraint in Bayesian Learning of Word Reference and Word Meaning

Luke Maurits, Amy Perfors, Dan Navarro

Australasian Mathematical Psychology Conference 2009

Slides available from http://www.luke.maurits.id.au/research/presentations/ampc09.pdf

THE UNIVERSITY OF ADELAIDE
AUSTRALIA

Outline

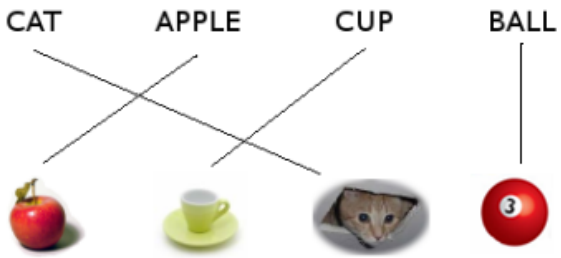
Outline

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Semantic and syntactic learning

- Learning a language requires solving two broad classes of problem.
- *Semantic* learning involves learning the manner in which words correspond to concepts or items.
- *Syntactic* learning involves learning the acceptable ways to combine words to convey thoughts.

Semantic Learning



Syntactic Learning

Ball with plays the cat ✗

The cat plays with the ball ✓

Cat the plays ball with ✗

With ball cat plays the ✗

Word order: a syntactic subproblem

- In many languages, a substantial and early syntactic learning problem is learning word order.
- “Dog bites man” and “Man bites dog” are very different!
- There are 6 possible word orders (e.g. subject-verb-object for English, subject-object-verb for Japanese, etc). All of which occur in natural languages.
- No innate or genetic predisposition to one word order - children must learn it (aside: the also need to learn that it is actually important!).

Are these problems really that hard?

- Word meaning learning is much harder than one expects, because the correct mappings of words to objects or concepts is logically underconstrainted.
- E.g. if an infant is shown an apple and hears the word “blick”, does “blick” mean apple? Does it mean red apple? The colour red? Fruit? Food in general? Round thing?
- This uncertainty combined with there being multiple words for each object or concept, and some words having multiple meanings, makes for a challenging learning task.
- Word order learning seems much harder when one considers how early children learn it - children understand word order before even combining words!

Yes, they are!

- These and other language-related learning tasks are so hard to solve in the available timespan using the available data that psycholinguists often argue for innate or "hard-wired", domain-specific modules of the mind to explain how quickly infants solve them.
- See also: Chomsky, "poverty of the stimulus", universal grammar, etc.... (but don't take it too seriously!)

Outline

One problem good, two problems bad?

- We are interested in investigating the extent to which the problems of learning word meanings and learning word order can be made more tractible by considering the single problem of acquiring them jointly and simultaneously instead of as distinct problems:
- If a learner's knowledge about one problem can be used to constrain their hypotheses about solutions to the other, will they learn substantially faster?
- This may reduce the need for assuming innate modules, or at least reduce their number/size/complexity.

Mutual constraint as a general principle?

- If the identification and use of mutual constraint works to improve learning in this situation, we might do well to ask:
- Could this be a general principle? Might many of the difficult problems in cognitive development which are often attributed to domain-specific innate abilities actually be solvable by more domain-general learning abilities which exploit mutual constraint?

Two Bayesian Models

- We present two models, based on *Bayesian inference*, for learning the meanings of words.
- One model attempts to *only* learn word meaning. It serves as our baseline model.
- The second model attempts to learn both word meaning and word order. It serves as a comparison to our baseline model to see if adding the second goal actually makes learning easier.
- Both models have a lot of details in common.

Outline

THE UNIVERSITY OF ADELAIDE AUSTRALIA

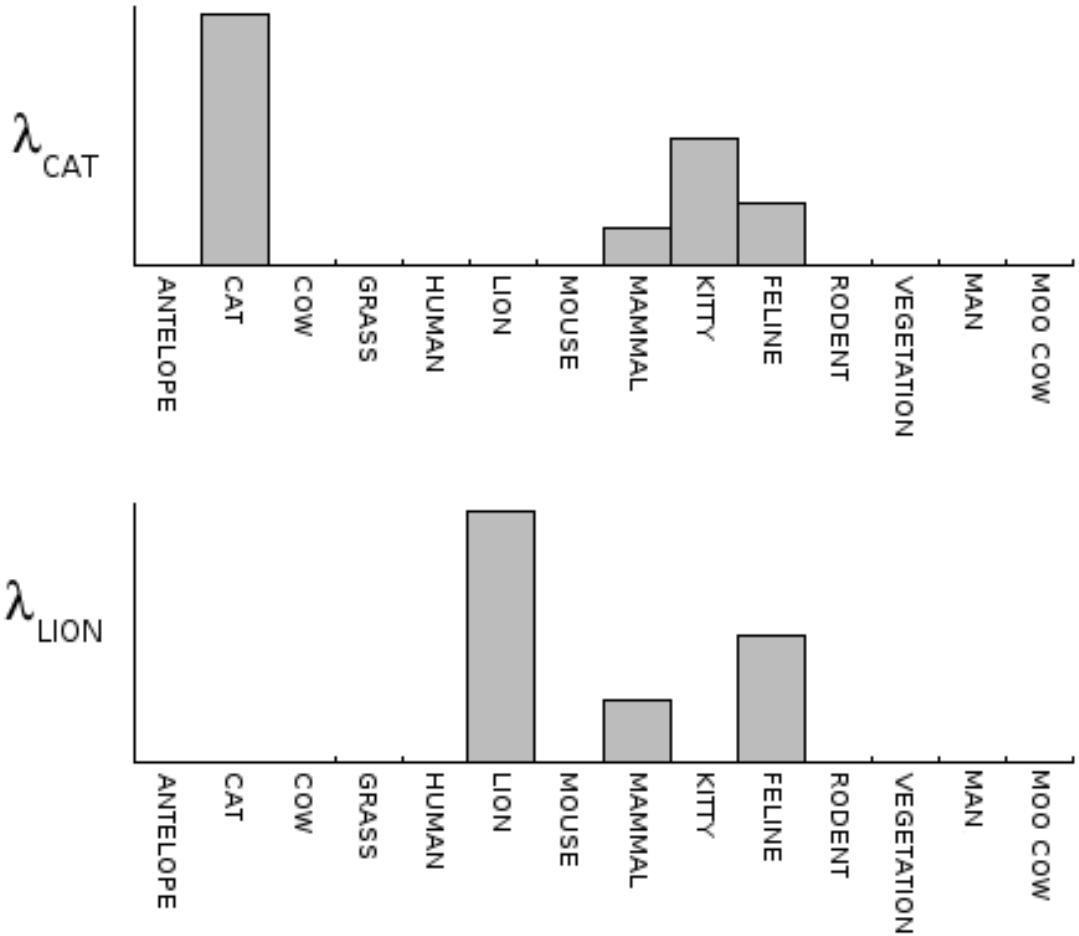
A Simple World

- We define a *world* which consists of a set $\mathcal{O}$ of *objects* and a set $\mathcal{R}$ of *relations*.
- The set of all defined (r, o, o) triples is equipped with a *probability distribution* Φ.

A Simple Language

- We define a finite *vocabulary* $\mathcal{V}$ of words.
- Every object $o \in \mathcal{O}$ and every relation $r \in \mathcal{R}$ has a *naming distribution* associated with it, which we call λ_x.
- The naming distributions are probability distributions over $\mathcal{V}$.
- High probability is associated to words most likely to be used as names for that object/relation.

An Example Naming Distribution



THE UNIVERSITY OF ADELAIDE AUSTRALIA

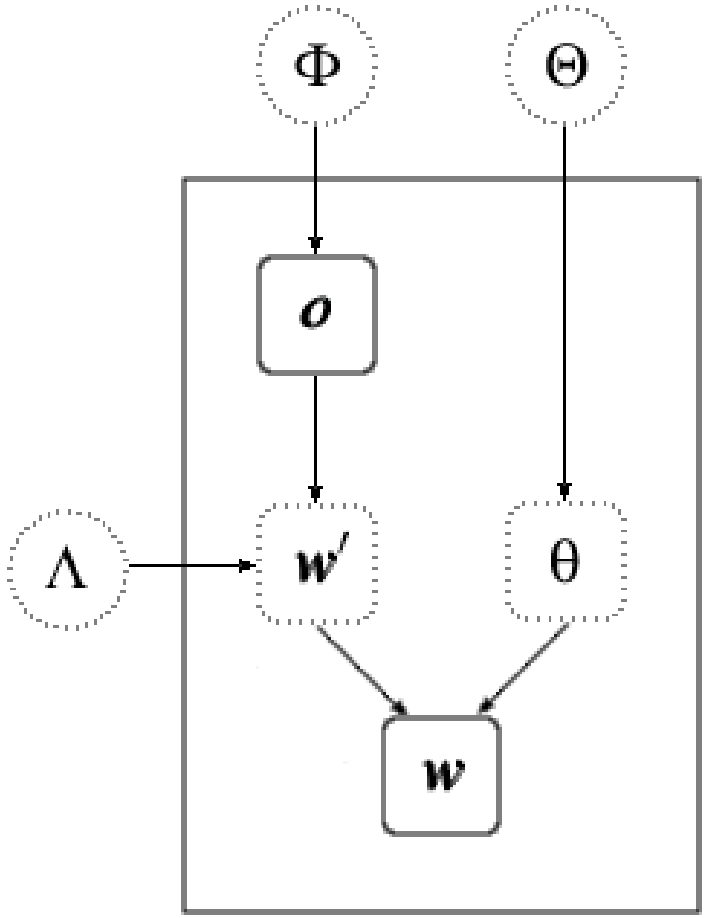
Incorporating word order

- Along with $\mathcal{V}$ and $\Lambda = \{\lambda_x | x \in \mathcal{R} \cup \mathcal{O}\}$ we give our language a *word order distribution*, Θ.
- This is a probability distribution over the 6 possible word orders which is intended to represent which orders are permissible / common.
- E.g. English may assign probability 0.8 to SVO (active voice - the cat sat on the mat), 0.2 to OVS (passive voice - the mat was sat on by the cat) and probability 0.0 to all other options (sat the cat on the mat is wrong).
- Languages with extremely strict word order may have all their probability assigned to one order, in which case word ordering is a deterministic process.

A Generative Process

- To generate a dataset $\mathcal{D}$ of size n, we perform the following steps n times:
  - Select a relation $z = r(o_1, o_2)$ at random from Φ.
  - Select words w_1 at random from λ_r, w_2 from λ_{o_1} and w_3 from λ_{o_2}.
  - Select a word order θ at random from Θ and set $w = \theta(w_1, w_2, w_3)$.
  - Return the data point (z, w).

Plate Diagram



Sample Data

Relational observation (z)	Linguistic utterance (w)
EAT(*cat*, *mouse*)	"cat eat rodent"
CHASE(*lion*, *antelope*)	"lion chase prey"
EAT(*cow*, *grass*)	"cow consume grass"
EAT(*antelope*, *grass*)	"antelope eat grass"

Outline

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Learning Word Order

- We model learning word order by computing Bayesian posterior estimates for each of the naming distributions in Λ.
- Our posterior distributions are computed numerically using Gibbs sampling.
- We will omit technical details related to the inference here (but feel free to bug me for them later!).
- Central point: to each data point (i.e. each row of our data table) we probabilistically assign a word order θ, each assignment being dependent on all others.

Baseline Model

- In our baseline model, the probability of assigning a particular word order θ to a datapoint (z, w) is dependent only upon the consistency of the induced naming with the current estimate of the relevant naming distributins.
- Example: We have the observation SAT(*cat*, *mat*) and the pair utterance “cat sat mat”. We have never encountered the words “sat” or “mat” before, but we have seen “cat” many times and have a good idea that it corresponds to the object CAT.
- We assign either SVO or SOV word order after tossing a coin - our knowledge about cats tells us that S must come first and we are completely naive about the other words.

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Baseline Model (contd)

- Maximising this consistency is the baseline model's only concern!
- There is no tendency whatsoever toward consistent preference for one or two word orders, beyond that encoded in the consistent concurrence of words and objects/relations.

Joint Learning Model

- In our joint learning model, the probability of assigning a particular word order θ to a datapoint (z, w) is dependent upon both the consistency of the induced naming (as before) *and* the consistency of θ with other word order assignments (i.e. with the current estimate of Θ).
- Roughly: We are most likely to assign the most commonly used word order so far *unless* doing so would be sufficiently inconsistent with established namings to warrant assuming otherwise.

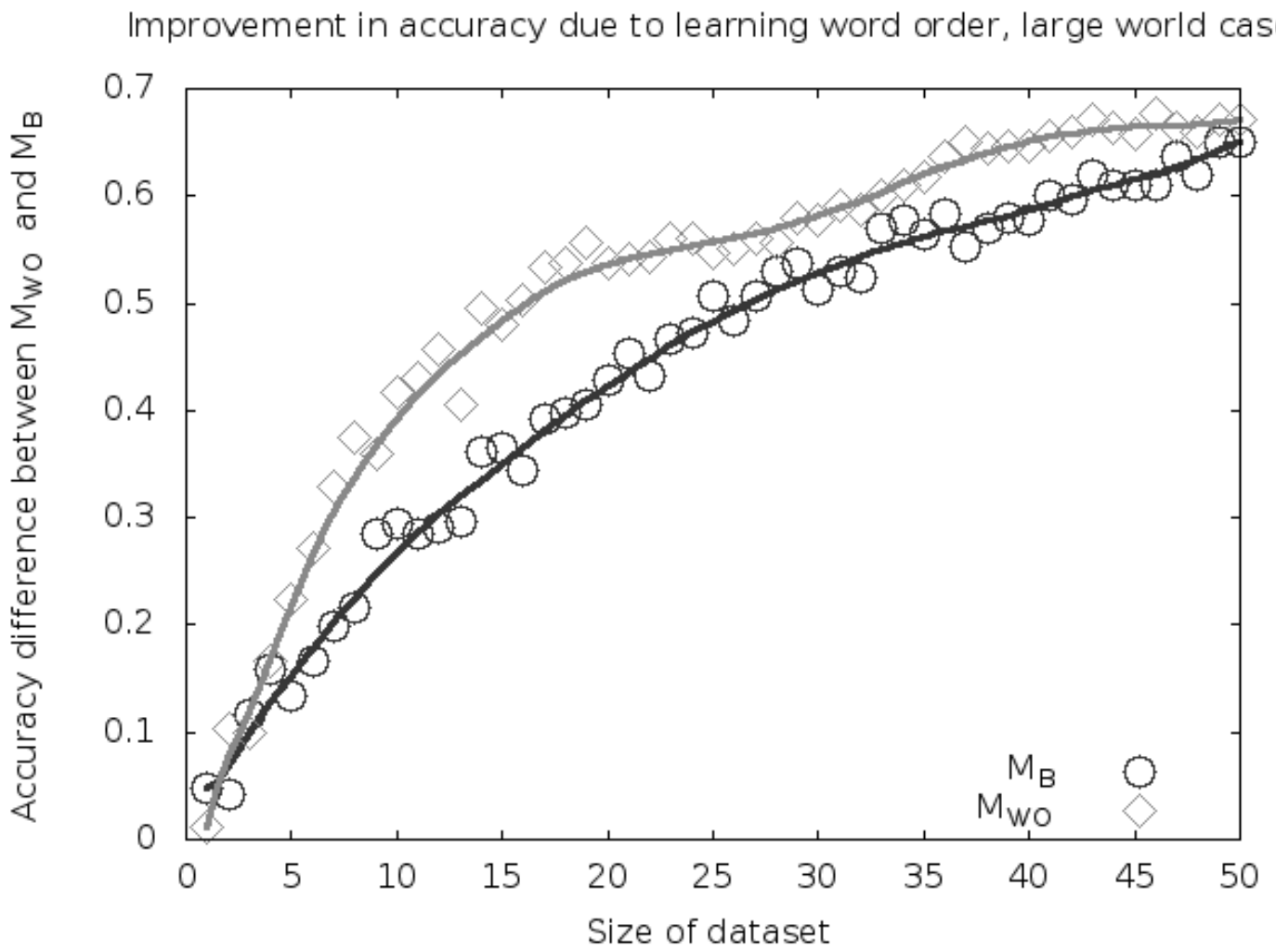
Outline

THE UNIVERSITY OF ADELAIDE AUSTRALIA

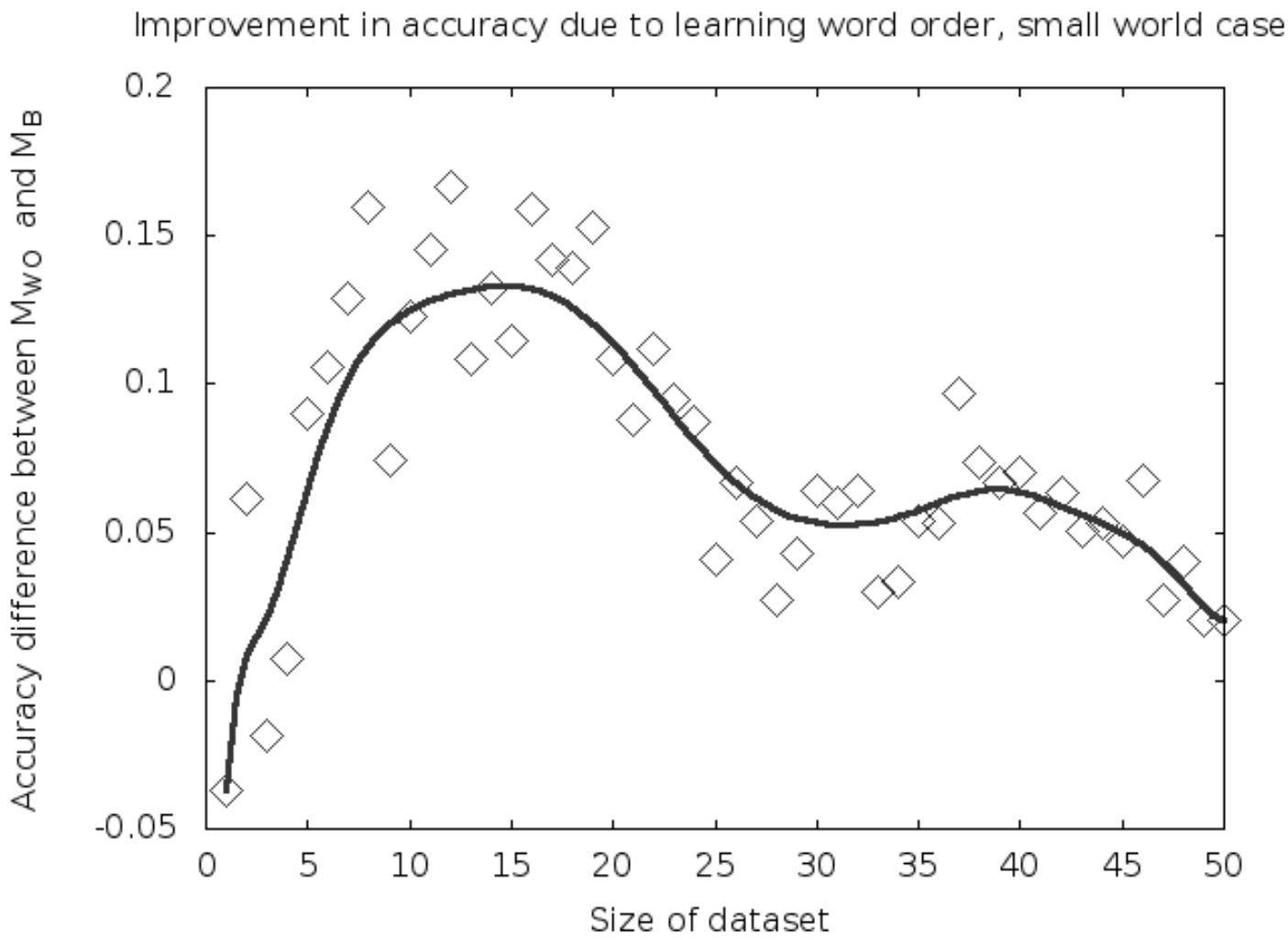
Artificial Data?!

- Results averaged over 10 data sets with random Φ, Λ.
- No, this doesn't necessarily tell us anything at all about "real learners".
- It is a useful litmus test - can the effect we seek exist in principle?
- Experiments on artificial data sets can be useful as a first step to refine the idea and identify characteristics to look for or control for in eventual "real experiments".
- Much less paperwork this way...

Both Models Learn...



...but the Joint Model Learns Better!

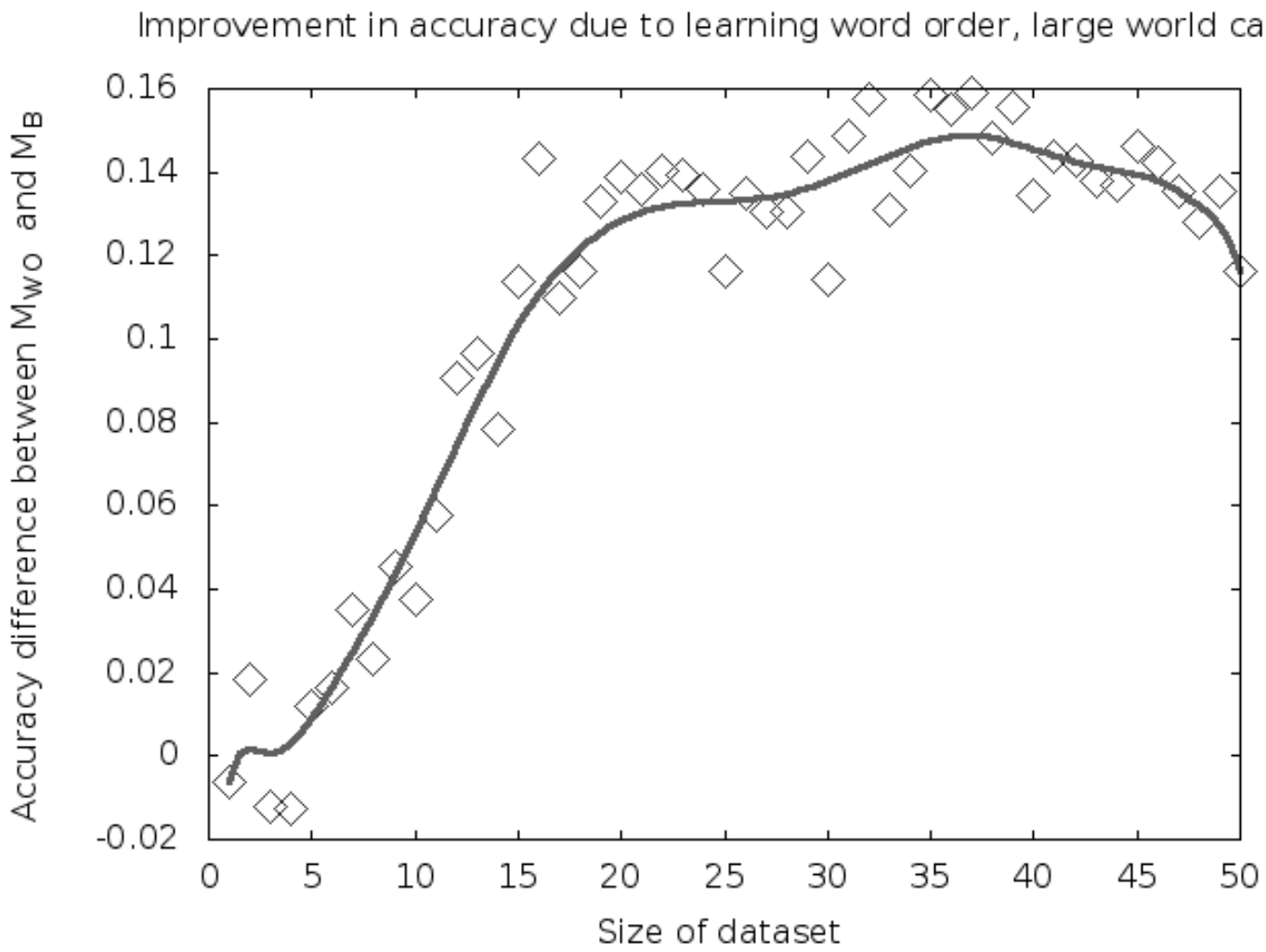


THE UNIVERSITY OF ADELAIDE AUSTRALIA

Is this due to the small world size?

- Maybe, let's see?

The Bigger the World, the Longer the Advantage



THE UNIVERSITY OF ADELAIDE AUSTRALIA

Outline

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Joint learning shows some promise

- Solving both problems simultaneously leads to better performance (at least with regards to word learning).
- The advantage seems to scale up to larger, and more complicated worlds.

Outline

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Leveraging extra data

- We have assumed that the data available for learning is "nice", i.e. every individual utterance is linked unambiguously to exactly one relational observation.
- It seems unlikely that this is actually the case.
- We can extend our model so that learners may make use of "lone" utterances, by estimating Φ from the available relational observations and inferring likely "missing" relations based on Φ and the naming distributions.

Sketch of model extension

- Define three data sets - $\mathcal{D}_{\text{REL}}, \mathcal{D}_{\text{LING}}$ and $\mathcal{D}_{\text{COUP}}$, consisting of lone observations, lone utterances and observation–utterance pairs, respectively.
- To each utterance in $\mathcal{D}_{\text{LING}}$, assign an assumed relation z according to a probability distribution which depends on naming consistency *and* an estimate of Φ.
- To each utterance in $\mathcal{D}_{\text{LING}}$ and $\mathcal{D}_{\text{COUP}}$ assign a word order variable θ just like before.

Benefit of extended model

- Example: A learner who knows the words “cat” and “sit” but who has never heard the word “mat” before and hears the utterance “cat sit mat” (and has learned that SVO word order is most likely) should be able to update their beliefs about the meaning of the word “mat” so that it is not equally likely to mean any object or relationship, but is somewhat more likely to refer to an object a cat is likely to sit on (like a mat, a chair or a person’s lap) and somewhat less likely to refer to an object that a cat is unlikely to sit on (like a chainsaw or fireplace).

Ideal results

- By controlling the relative sizes of $\mathcal{D}_{\text{REL}}$, $\mathcal{D}_{\text{LING}}$ and $\mathcal{D}_{\text{COUP}}$, we may be able to make statements like “to a learner who thinks like the extended model, *m* tightly coupled relational observations and linguistic utterances is ‘as good as’ *n* tightly coupled pairs and *o* individual observations or utterances ”.
- This should improve learning.
- This reduces the implicit explanatory weight on learner’s ability to perform coupling.
- This offers us a glimpse at the possibilities of *inter-domain constraint* - semantic learning constraining language learning instead of language learning constraining intself

Outline

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Inter-domain Mutual Constraint

- Perhaps we can estimate Φ more accurately and quickly by generalising beyond the data.
- We can use a *categorsation* model to investigate this, e.g. *infinite blockmodel*.
- This could provide a nicer example of semantic “world learning” constraining language learning.
- We can also alter our generative process so that sometimes objects are referred to by the name of a category they belong to.
- This could provide an example of language learning constraining semantic “world learning”.

Questions?