Sensitivity to Input Order: Evaluation of an Incremental and Memory-Limited Bayesian Cross-Situational Word Learning Model

Sepideh Sadeghi
Tufts University
Human-Robot Interaction Laboratory
Medford, MA, USA
sepideh.sadeghi@tufts.edu

Matthias Scheutz
Tufts University
Human-Robot Interaction Laboratory
Medford, MA, USA
matthias.scheutz@tufts.edu

Abstract

We present a variation of the incremental and memory-limited algorithm in (Sadeghi et al., 2017) for Bayesian cross-situational word learning and evaluate the model in terms of its functional performance and its sensitivity to input order. We show that the functional performance of our sub-optimal model on corpus data is close to that of its optimal counterpart (Frank et al., 2009), while only the sub-optimal model is capable of predicting the input order effects reported in experimental studies.

1 Introduction

It has been suggested that early word learning is guided by observing the referent of words across situations (*cross-situational word learning*). Surely the co-occurrence statistics of word-object mappings is noisy (e.g., when the referent of a word is absent in the scene) and only provides ambiguous information (due to referential ambiguity) about word-object mappings, but it is the only source of information used by a true novice who does not have access to social and linguistic cues (Baldwin, 1993; Arunachalam and Waxman, 2010) to narrow down the hypothesis space of words' meanings. In fact, development of some initial meaning interpretations is the prerequisite for the development of useful social and syntactic knowledge (cues) for bootstrapping the process of word learning.

Recent experimental findings (Medina et al., 2011; Trueswell et al., 2013; Zhang and Yu, 2017) indicate that word learning is sensitive to the presentation order of learning trials. More specifically, their results suggest that early wrong interpretations of words' meanings can inhibit the acquisition of correct words' meanings and that performance is best when highly informative learning trials (e.g., with lower referential ambiguity) precede the low informative learning trials, so that the possibility of starting with wrong meaning interpretations is minimized. These results are incompatible with the prediction of the existing well-known word learning models which employ global learning and are equipped with perfect memory of past observations (Yu and Ballard, 2007; Kachergis et al., 2012; Xu and Tenenbaum, 2007; Frank et al., 2009; Alishahi and Fazly, 2010). These models are insensitive to input order due to two major reasons: (1) being purely driven by co-occurrence regularities and assuming full access to co-occurrence statistics which stays the same regardless of input order or (2) processing the data in batch. Word learning is one of a wide range of cognitive tasks which exhibit sensitivity to the presentation style of input. Inductive category learning (Carvalho and Goldstone, 2015), category generalization (Spencer et al., 2011), property projection (Lawson, 2017), etc. all demonstrate important differences under sequential vs. simultaneous presentation of input. Therefore, the root causes of the effects of presentation order on word learning results should be understood as it may generalize to other cognitive tasks and can make up an important aspect of word learning models.

The recent findings regarding the effect of presentation style and order provide evidence for local approaches to learning, while they fail to rule out the possibility of global learning. Exploring the intermix of local and global approaches to learning is important on three levels: (1) replicating human performance, (2) departing from computationally expensive ideal learners that become inefficient as the size of

Proceedings of the 27th International Conference on Computational Linguistics, pages 3170–3180
Santa Fe, New Mexico, USA, August 20-26, 2018.

input data grows, and (3) giving accounts for the real-world constraints faced by word learners (e.g., It is cognitively implausible to assume that infants can remember the details of all past observations). We use a variation of the incremental and memory-limited learning algorithm proposed in (Sadeghi et al., 2017) and provide the first corpus evaluation of this sub-optimal model in terms of its functional performance and the patterns of sensitivity to input order it exhibits which match those reported in the experimental studies (Medina et al., 2011).

2 Model Overview

Similar to previous work (Yu and Ballard, 2007; Alishahi and Fazly, 2010; Frank et al., 2009), we assume that the learner is capable of object and action categorization prior to word learning and the goal is to learn the basic level object category names. The *input* to the model (*observations*) are word learning *situations* (*trials*) each of which consists of an un-ordered set of words as utterance paired with a list of objects as the scene. The notion of *memory* refers to the number of observations (trials) that can be remembered at a time and the current model assumes memory=1. In each situation, the proposed incremental model (1) utilizes a Bayesian approach to infer the correct word-object mappings and (2) integrates the newly inferred word-object mappings (*mini-lexicon*) in the previously learned many-to-many mapping between words and objects (*lexicon*), based on their association strength in the presence of alternative mappings. The model's memory is limited to the word-object mappings stored in its full lexicon and the details of the current situation. The Bayesian approach is constrained on two distinct levels: (1) hypothesis generation (mini-lexicon generation) and (2) hypothesis evaluation (inferring the best mini-lexicon). Both hypothesis generation and hypothesis evaluation are limited to the knowledge stored in the lexicon and the details of the current situation. Therefore in our model, Bayesian approach neither performs global optimization, nor has access to all possible hypotheses (all possibilities if the model had full access to all observations). Instead, the Bayesian approach is a local approach as both hypothesis generation and hypothesis evaluation are limited to the knowledge stored in the lexicon and the details of the current situation.

3 Speaker's Model of Utterance Generation

The learner assumes that in each situation, the speaker uses the generative process illustrated in Fig. 1 to produce an utterance (W_s) corresponding to the current scene (O_s) and using a context appropriate portion (L) of the full lexicon. In each situation, the model has to infer which words are *referential* and to which object they refer to, where a *referential* word is a noun with a concrete object referent in the scene. *Referential intentions* of the speaker (I_s), which refer to the objects that are present in the scene and the speaker is talking about them, determine the space of possible referents for each referential word in the utterance (W_s).

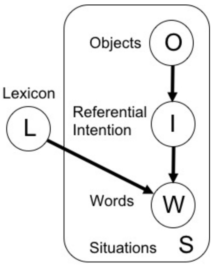

Figure 1: The graphical model describing the probabilistic dependencies between model random variables: utterance (W_s), intention (I_s), lexicon (L), and the objects present in the scene (O_s), where s indexes the situation. The plate indicates multiple copies of the model for different situations (utterance-scene pairs).

In each situation, the speaker uniformly samples a subset from the power-set of all the objects (O_s) present in the scene as the referential intention(s) of the speaker (I_s). For each object in I_s, one word mapping is drawn from the lexicon and added to the utterance. The non-referential words of the utterance are generated randomly with probability $P_{NR} = 1$. In each situation the learner tries to reverse the generative process described in Fig. 1 to infer a context-appropriate portion of the full lexicon (*mini-lexicon*) used by the speaker, where *context* refers to the objects in the current situation. The inferred mini-lexicon then will be integrated in the full lexicon available to the learner (details in Section 4).

Inferring the best mini-lexicon in each situation, requires finding the MAP (maximum a posteriori) mini-lexicon by marginalizing over all possible referential intentions, since I_s is unobserved. The model finds the MAP mini-lexicon ($argmax_L P(L|C)$) according to the Bayes equation and the probability distribution that it defines over unobserved mini-lexica (L) and the corpus of situations (C). The current model only remembers one situation at a time (memory=1) due to which C includes one observed situation plus the context-appropriate situations extracted from the lexicon (see 4.2 for details).

$$P(L|C) \propto P(C|L)P(L) \tag{1}$$

We use $P(L) \propto e^{-\alpha \cdot |L|}$ serving as a soft mutual exclusivity constraint to produce a preference for one-to-one mappings in the mini-lexicon inferred in each situation. Marginalizing over all possible intentions in each situation we can rewrite the likelihood term $P(C|L)$ as:

$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq O_s} P(W_s|I_s, L)P(I_s|O_s) \tag{2}$$

Assuming that $P(I_s|O_s) \propto 1$ and that the words of the utterance are generated independently, we can rewrite the term $P(W_s|I_s, L)$ as:

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1-\gamma)P_{NR}(w|L)] \tag{3}$$

The words of utterance can be generated in two different ways: (1) referentially (with a concrete object referent in the scene) with probability γ and (2) non-referentially with probability $1-\gamma$. The probability of non-referential use (P_{NR}) of a referential word (words in the model lexicon) is set to $\kappa < 1$ (to penalize the non-referential use of referential words), and is set to 1 for non-referential words. The probability of referential use of a referential word in reference to a particular object (P_R) is the probability of the word being chosen uniformly from the set of all words linked to that object in the lexicon.

4 Incremental Learning Algorithm

Our learning algorithm is a variation of the incremental and memory-limited algorithm proposed in (Sadeghi et al., 2017) which accounts for the following real-world constraints: (1) seeing each situation only once with no iteration over data, (2) using only the acquired knowledge and the current observation for hypothesis generation and evaluation. These constraints exclude the use of many proposed incremental algorithms in the literature (Liang et al., 2009; Pearl et al., 2010; Börschinger and Johnson, 2011; Börschinger and Johnson, 2012). In this section we first give a high-level overview of the incremental learning algorithm, while highlighting the differences between our version and the original version proposed in (Sadeghi et al., 2017). We refer the reader to (Sadeghi et al., 2017) for the details shared between the two algorithms. Furthermore, we refer curious readers to (Sadeghi and Scheutz, 2017; Sadeghi and Scheutz, 2018) which demonstrate the application of similar learning algorithms to the augmented versions of the graphical model presented here.

Our algorithm is composed of two major components: (1) inferring the MAP mini-lexicon in each situation, and (2) integrating the new mini-lexicon in the previous lexicon. Inferring the MAP mini-lexicon, subsequently has two components: (1) hypothesis generation (generating mini-lexicon proposals) and (2) hypothesis evaluation (evaluating the mini-lexicon proposals using Eq. 1).

Our algorithm departs from its batch counterpart (Frank et al., 2009) in the notion of C and L. L in (Frank et al., 2009) refers to the entire dictionary used by the speaker while in our model L (mini-lexicon) refers to a part of the dictionary used by the speaker to produce the current utterance in correspondence to the current scene. In each situation, we try to infer a portion of the entire dictionary used by the speaker and we refer to the aggregation of all these mini-lexica as the lexicon. Similarly, C in (Frank et al., 2009) refers to all the observed situations (<*utterance*,*scene*> pairs), assuming full access to data, but in our model C refers to the current situation and a number of relevant situations extracted from the lexicon.

Our algorithm departs from the algorithm in (Sadeghi et al., 2017) in the definition of *context* and subsequently the *context-appropriate mappings* and the *context-appropriate situations* used for hypothesis generation and hypothesis evaluation accordingly in each situation. *Context* in our model refers to the list of objects present in the scene while *context* in (Sadeghi et al., 2017) refers to the objects and words present in the scene and utterance. Given *trial*=<*utterance*,*scene*>, we define *context*=*trial.objects*=$\{o_j|o_j \in \textit{scene}\}$, *trial.words*=$\{w_i|w_i \in \textit{utterance}\}$, and *trial.mappings*=$\{< w_i, o_j > |(o_j \in \textit{scene}) \wedge (w_i \in \textit{utterance})\}$. Then the context appropriate information are defined as: *context-appropriate-mappings*=*trial.mappings* $\cup\{< w_i, o_j > |(< w_i, o_j > \in \textit{lexicon}) \wedge (o_j \in \textit{context})\}$, and C=*context-appropriate-situations*=$\{<$ *utterance*=w_i,*scene*=$o_j > |(< w_i, o_j > \in \textit{lexicon}) \wedge (o_j \in \textit{context})\} \cup$ *trial*.

We focus on objects as the context to extract appropriate information from the lexicon, since the goal of the model is to learn the label of objects while the mappings stored in the lexicon are noisy and can include object mappings for non-referential words. Therefore, we can benefit from filtering out the potential noise (co-occurrence of the non-referential words and objects) and retrieving only the stored mappings of the objects as a proxy for true object labels during hypothesis generation and hypothesis evaluation.

4.1 Generating Mini-Lexicon Proposals

Generating mini-lexicon proposals in each situation, is guided by semi-stochastic search techniques on the context-appropriate mappings based on their association strength (Point-wise mutual information accumulated incrementally and globally) analogous to (Sadeghi et al., 2017). Upon receiving a new input situation, first, we initialize 40 mini-lexica and take the one with the highest un-normalized posterior probability as the best initialized mini-lexicon (*init-lex*). We then explore the neighborhood around *init-lex* by mutating it *nStep* times (we used *nStep*=3). Mutation can be thought of as generating a tree of height=*nStep*+1, with branch factor of 3 (performing 3 operations on mappings of the lexicon at each node: add, remove, swap), the *init-lex* at the root, and its mutated versions on the leaves. We then compare the *init-lex* and its mutated versions against each other and report the one with the highest un-normalized posterior probability as the inferred mini-lexicon.

4.2 Evaluating Mini-Lexicon Proposals

The mini-lexica generated in the previous section are evaluated based on their relative posterior probability computed according to Eq. 1. C includes the context-appropriate situations and is used as evidence for Bayesian inference when evaluating the mini-lexicon proposals. Note that the model does not remember the details of past observations but is able to remember what it learned from them (word-object mappings stored in the lexicon). Hypothesis evaluation in our model departs from hypothesis evaluation in (Sadeghi et al., 2017) by virtue of using a different notion of context and context-appropriate situations used as evidence for Bayesian inference.

4.3 Integrating the MAP Mini-Lexicon in the Lexicon

Integration of the MAP mini-lexicon modifies (adds, deletes or keeps) the word mappings stored in the lexicon, for each item in the context (objects in the current scene). Our algorithm departs from the original version in how it handles conflict resolution when integrating the newly inferred mini-lexicon in the previous lexicon. In (Sadeghi et al., 2017), during conflict resolution, alternative mappings compete with each other and only mappings with the highest co-occurrence statistics are allowed in the output lexicon. This strict mutual exclusivity constraint not only inhibits learning of alternative word-object mappings (e.g., “cat”,“kitten”, and “kittycat” all refer to the same concept “CAT”), but also destabilizes the learning results. Our model on the other hand, modulates this strict mutual exclusivity constraint by allowing the addition of alternative mappings for each object, if their co-occurrence statistics fall within a certain range of the co-occurrence statistics of the best existing mapping for that object. We introduce two parameters: (1) *keep-fraction* and (2) *add-fraction* for the inclusion of alternative mappings in the output lexicon, correspondingly from the previous lexicon and the MAP mini-lexicon. For each object in current context, we allow mappings from the previous lexicon and from the MAP mini-lexicon in the output lexicon if their association strength are correspondingly not less than *keep-fraction* and *add-fraction* of the word mapping with the highest association strength available in memory, for that object. Using small values for *keep-fraction* and *add-fraction* would enforce softer mutual exclusivity constraints and using large values would enforce harder mutual exclusivity constraints.

5 Evaluation Data

5.1 Input Properties

The input properties we study include: (1) the information inherent in learning trials and (2) the serial order of their presentation. We categorize the trials into two categories: *referential* and *non-referential* trials. A trial is *non-referential* if no word in its utterance has a concrete object referent in the scene. Otherwise, the trial is regarded as a *referential trial*. *non-referential* trials are misleading trials as the word-object co-occurrence statistics inherent in them do not guide the model towards learning the correct word-object mappings. Note that *non-referential* trials are generally misleading for any learner which uses object-word co-occurrence statistics to guide word learning. We use *misleading* and *non-referential* interchangeably to refer to the same kind of trials. The referential trials in turn, contain graded levels of useful co-occurrence statistics based on their overall *referential ambiguity*. We use the product of the number of objects in the scene and the number of words in the utterance as a simple measure of trial *ambiguity*.

5.2 Corpus Properties

Table. 1 and Table. 2 report the properties of the two annotated video files used for our corpus evaluations. Note that di06 and me03 differ from each other in the ratio of referential to non-referential trials (1:2.07 for di06 and 1:4.64 for me03) meaning that overall the data in di06 is more informative than the data in me03 for labeling objects. D-di-me and D-me-di differ from each other in the presentation order of the more informative video and D-diRef-meRef differs from D-di-me in that it contains no non-referential (misleading) trials. Our evaluation data is available at `https://tinyurl.com/y9omp52c`.

Table 1: Properties of the annotated child-directed video files used for corpus evaluations. di06 and me03 are taken from the Rollins section of CHILDES (MacWhinney, 1991). We used the annotations and gold lexicon used in (Frank et al., 2009).

Video	Referential trials	Non-referential trials	Average ambiguity	Average ambiguity (referential trials)
me03	56	260	7.3	10.73
di06	100	207	9.35	12.02

Table 2: Different datasets made up of the trials in me03 and di06 for testing the incremental model.

Data	Video files
D-di-me	trials in di06 followed by trials in me03
D-me-di	trials in me03 followed by trials in di06
D-diRef-meRef	referential trials in di06 followed by referential trials in me03

5.3 Target Effects of Input Properties

Recent experimental studies (Medina et al., 2011; Trueswell et al., 2013; Zhang and Yu, 2017) report a huge effect of input order in word learning results, where performance is reported to be sensitive to the presentation order and informativity of learning trials. It has been demonstrated that highly informative trials (referential trials with low ambiguity) facilitate the acquisition of correct meaning interpretations while low informative trials (highly ambiguous referential trials or non-referential trials) facilitate the acquisition of incorrect meaning interpretations. Additionally, it has been reported that early correct interpretations of words' meanings can facilitate the acquisition of correct words' meaning in later encounters while early wrong interpretations of words' meanings can inhibit the acquisition of correct words' meanings in later encounters.

Taking these results together, we expect the model performance: (1) to be sensitive to the presentation order of learning trials, (2) to be best when there is no misleading learning trials in the input data, and (3) to be best when highly informative learning trials precede the low informative learning trials, so that the possibility of starting with wrong meaning interpretations is minimized. We assume that referential trials are more informative than non-referential (misleading) trials and that referential trials with low ambiguity are more informative than referential trials with high ambiguity.

6 Evaluations

In the following sections, for each model, we use precision, recall and fscore of its output lexicon (w.r.t the gold standard lexicon) to measure its functional performance at the end of the simulation and we use *mean word acquisition score* (Alishahi and Fazly, 2010) to evaluate its incremental performance during the simulation.

$$\textbf{mean word acquisition score} = \frac{\sum_{<w,o>\in goldLexicon} P(w|o, lexicon)}{size(goldLexicon)} \tag{4}$$

6.1 Functional Performance

Table. 3 contrasts the functional performance of the our proposed incremental model with that of its batch (optimal) counterpart (Frank et al., 2009) as well as a number of baseline incremental models (please refer to (Frank et al., 2009) for more details about the comparison models). The results reported for the comparison models are taken from (Frank et al., 2009). Analogous to (Frank et al., 2009), we report the precision, recall and fscore of the best lexicon found by the incremental model, where the lexicon with the highest fscore value is considered the best. We performed a grid search in the space of parameters to find the best lexicon for each model and data pair (top three scores are highlighted in each category).

The proposed incremental model is tested on three different datasets (see Table. 2) to evaluate the effect of input properties on model performance. We did not do the same for other models as they are not sensitive to input order.

Table 3: Precision, recall, and f-score of the best lexicon found by each model when run on two annotated video files (di06 and me03) from the Rollins section of (MacWhinney, 1991). The incremental model is the only model which is sensitive to input order and is tested on a number of datasets presented in Table. 2.

Model(data)	Precision	Recall	F-Score
batch model (Frank et al., 2009)	**0.67**	0.47	**0.55**
Incremental model(D-di-me)	**0.259**	**0.617**	**0.365**
Incremental model(D-me-di)	0.204	**0.558**	0.299
Incremental model(D-diRef-meRef)	**0.415**	**0.647**	**0.505**
Association frequency	0.06	0.26	0.1
$P(object\|word)$	0.07	0.21	0.1
$P(word\|object)$	0.07	0.32	0.11
Point-wise MI	0.06	0.47	0.11
Translation Model $P(object\|word)$	0.07	0.32	0.12
Translation Model $P(word\|object)$	0.15	0.38	0.22

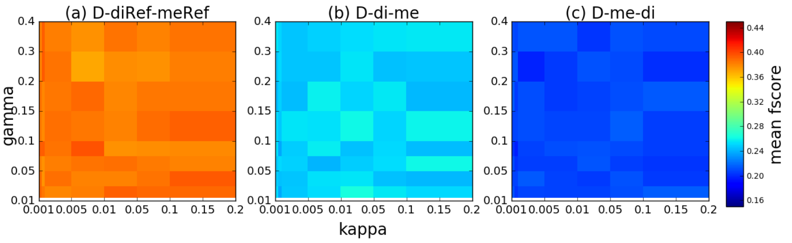

Figure 2: The mean fscore of the lexica found by running the model on (a) D-diRef-meRef, (b) D-di-me, and (c) D-me-di. D-di-me and D-me-di have 623 trials each, while D-diRef-meRef has only 156 trials. In each heatmap, the results are averaged over 50 runs for each set of parameter values using $\alpha = 3$, *keep-fraction* $= 0.5$, *add-fraction* $= 0.5$, and varying κ and γ.

Unsurprisingly, the batch model which has access to all datapoints has the highest precision and fscore on the dataset. Our incremental model has the second best fscore and the highest recall among all models as well as the best fscore among the incremental models. We used *keep-fraction* $=$ *add-fraction* $= 0.5$ (a soft mutual exclusivity constraint) which facilitates the addition of more mappings to the lexicon. This tends to improve the recall and hurt the precision of the model on small datasets such as the ones used here.

Note that the incremental model exhibits its best performance when misleading trials are removed from the corpus data (on D-diRef-meRef) which is inline with our second expectation discussed in Section 5.3 suggesting that misleading trials inhibit the process of learning the correct mappings by facilitating the acquisition of wrong meaning interpretations which in turn can interfere with the acquisition of correct meaning interpretations in later encounters.

As can be seen in Table. 1, the ratio of referential to non-referential trials in di06 is higher than that of me03. Therefore, based on what was discussed in Section 5.3, we expect the incremental model to exhibit better performance when: (1) di06 is presented first compared to when me03 is presented first, and (2) when non-referential trials are removed. Both (1) and (2) help minimize the possibility of early acquisition of wrong meaning interpretations. Our result is compatible with this expectation as the incremental model achieves better results on D-di-me dataset compared to D-me-di (reversed presentation order of video files), and better results on D-diRef-meRef compared to D-di-me (due to removal of misleading trials). Fig. 2 demonstrates that this outperformance pattern is robust against the choice of model parameters and is not an artifact of evaluating the best fscore values.

As can be seen in Fig. 2, regardless of parameter values, mean fscore values on D-diRef-meRef are better than mean fscore values on D-di-me which are in turn better than mean fscore values on D-me-di. Similar outperformance patterns were observed on the heatmaps of mean precision and mean recall scores. Furthermore, additional heatmaps generated using different α values provided further evidence for the robustness of the observed model outperformance on certain datasets regardless of the choice of model parameters and regardless of whether mean or max scores are used for comparison.

6.2 The Presentation Order of Ambiguous Trials

In this section, we seek to examine the effect of the presentation order of ambiguous trials. In order to filter out the effect of misleading trials, we use D-diRef-meRef with no misleading trials (see Table. 2) as the default ordered dataset. We reorded the trials in this dataset and created two new datasets: (1) with ascending trial ambiguity and (2) with descending trial ambiguity.

Fig. 3 demonstrates both the ideal rate of word acquisition (using a perfect model which only learns the correct mappings at the end of each trial) and the word acquisition rate achieved by the incremental model, on each dataset. As can be seen, high referential ambiguity (early trials of the data with descending order of ambiguity as well as the last trials of the data with ascending order of ambiguity) slows the rate of word acquisition as reflected on the slope of the green and red solid lines.

According to recent findings discussed in Section 5.3, performance is best when trials with low ref-

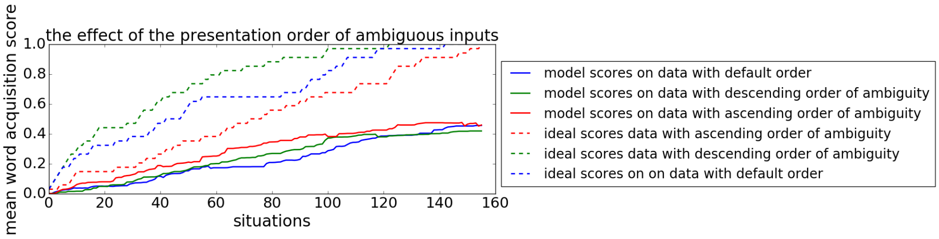

Figure 3: The mean word acquisition score over time as the incremental model (solid lines) is presented with three different datasets: (1) D-diRef-meRef (data with default order), (2) reordered D-diRef-meRef in ascending order of trial ambiguity, and (3) reordered D-diRef-meRef in decreasing order of trial ambiguity. The results are averaged over 50 runs, using the following parameter values: $\kappa = 0.005$, $\gamma = 0.3$, $\alpha = 5$, *keep-fraction* $= 0.5$ and *add-fraction* $= 0.5$. The dashed lines demonstrate the presentation rate of words in the gold standard lexicon in each dataset. The dashed lines also demonstrate the ideal mean word acquisition scores over time, if the model were to learn only the correct mappings at the end of each trial.

erential ambiguity precede the trials with high referential ambiguity. The results from our incremental model is in line with this expectation as the model achieves its best performance on the data with ascending order of ambiguity. Furthermore, the gap between the model performance and the ideal performance is the least when the data is presented in an increasing order of ambiguity. Finally, our results indicate that although input order affects the rate of acquisition, but (1) as more data becomes available, the extra cross-situational information inherent in data can make up for early input order effects resulting in similar word acquisition scores at the end of these experiments, and (2) that it is possible to replicate the input order effects using memory-limited continuous word learning mechanisms that accumulate information across all input trials.

6.3 The Presentation Order of Misleading Trials

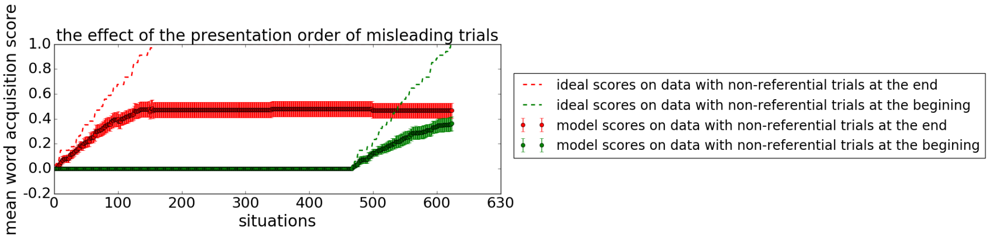

Figure 4: The mean word acquisition score over time as the incremental model (solid lines) is presented with two different datasets: (1) non-referential trials of di06 and me03 added at the end of D-diRef-meRef, (2) non-referential trials of di06 and me03 added at the beginning of D-diRef-meRef. The results are averaged over 50 runs, using the following parameter values: $\kappa = 0.005$, $\gamma = 0.3$, $\alpha = 5$, *keep-fraction* $= 0.5$ and *add-fraction* $= 0.5$. The dashed lines demonstrate the presentation rate of words in the gold standard lexicon in each dataset. The dashed lines also demonstrate the ideal mean word acquisition scores over time, if the model were to learn only the correct mappings at the end of each trial.

in Section 6.1, we demonstrated the inhibitory effect of misleading trials and their early presentation in the input data (as in D-me-di) on final word learning results (fscore, precision and recall). In order to establish what happens during the presentation of misleading trials and how the model is affected by these trials incrementally, we perform two more follow-up experiments using D-diRef-meRef dataset in increasing order of trial ambiguity: (1) putting back the non-referential trials of di06 and me03 at the beginning of D-diRef-meRef, (2) putting back the non-referential trials of di06 and me03 at the end of

D-diRef-meRef. These two follow-up experiments can be used to verify (1) if re-ordering the misleading trials can modulate their negative effect, and (2) how long lasting the effect of misleading trials is on word learning results.

Fig. 4 suggests that presenting the misleading trials after the referential trials seem to have no significant effect on previously learned correct mappings as no significant performance drop is observed on the red line. On the other hand, presenting the misleading trials before the referential trials, significantly affects the model performance as the solid green line remains significantly below the solid red line.

7 Discussion and Conclusion

We presented a variation of the incremental and memory limited word learning model in (Sadeghi et al., 2017) and provided the first corpus evaluations of this type of sub-optimal model. Our corpus evaluations demonstrated superior model performance compared to baseline incremental models. In addition to that the performance of the incremental model was close to the performance of the its batch counterpart (Frank et al., 2009), specially when misleading trials were removed from the corpus but only the incremental model was capable of accounting for input order effects.

We examined the replication of significant input order effects reported in (Medina et al., 2011; Trueswell et al., 2013; Zhang and Yu, 2017) in our model (1) incrementally, (2) on multiple datasets with different size and presentation order of ambiguous and misleading data, and (3) using different parameter values (Fig.2). The predictions of our model regarding the input order effects were in line with the recent empirical findings in (Medina et al., 2011; Trueswell et al., 2013; Zhang and Yu, 2017) as the model exhibited sensitivity to the information content of input trials as well as their presentation order. More specifically the model performance was best when highly informative (with low referential ambiguity) trials preceded the low informative (with high referential ambiguity) trials and when misleading trials were removed. Our simulations with different parameters showed that these outperformance patterns were robust against the choice of model parameters. Furthermore, our simulation results suggest that exposure to more data can neutralize early input order effects and that *memory-limited* continuous word learning mechanisms can replicate the input order effects.

The presented model is more cognitively plausible than its batch counterpart in that it is incremental and memory-limited. In addition to that, it relies on sub-optimal learning approaches (local optimization) as it utilizes a limited memory of past observations for incremental hypothesis generation and hypothesis evaluation performed by Bayesian inference. Applying the Bayesian inference locally and with limited data as evidence, allows the model to avoid becoming inefficient as the size of input data grows.

This work is the first corpus evaluation of a sub-optimal word learner which not only produces comparable functional performance to that of its optimal counterpart (Frank et al., 2009), but also successfully replicates the input order effects reported in experimental studies. Overall, the presented sub-optimal word learner is the first step towards exploring (1) local approaches to learning, (2) which depart from computationally expensive ideal learners, (3) without a huge loss in functional performance, (4) while replicating human performance, and (5) accounting for real-world constraints faced by learners (e.g., an infant or a robot). Furthermore, this work serves as an examination of the loss and gain of departing from optimal learners by performing direct comparisons between a sub-optimal model and its optimal counterpart. Taking (Frank et al., 2009) as a particular optimal model and the proposed model here as its sub-optimal version, we presented the loss (in functional performance) and gain (predicting input order effects) of departing from an optimal model (Frank et al., 2009).

In future work we plan to modulate model's memory of past observations (the number of observations that can be remembered with full details) to establish the effect of memory on replication of input order effects. Furthermore, we plan to compare our model with other sub-optimal models (Stevens et al., 2017) to gain a better understanding of the computational requirements for modeling the input order effects reported in experimental word learning literature.

Acknowledgements

This work was funded in part by Vienna Science and Technology Fund project ICT15-045.

References

Afra Alishahi and Afsaneh Fazly. 2010. Integrating syntactic knowledge into a model of cross-situational word learning. In *Proc. of CogSci*, volume 10.

Sudha Arunachalam and Sandra R Waxman. 2010. Meaning from syntax: Evidence from 2-year-olds. *Cognition*, 114(3):442–446.

Dare A Baldwin. 1993. Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental psychology*, 29(5):832.

Benjamin Börschinger and Mark Johnson. 2011. A particle filter algorithm for bayesian word segmentation. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 10–18. D. Mollá & D. Martinez.

Benjamin Börschinger and Mark Johnson. 2012. Using rejuvenation to improve particle filtering for bayesian word segmentation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 85–89. Association for Computational Linguistics.

Paulo F Carvalho and Robert L Goldstone. 2015. What you learn is more than what you see: what can sequencing effects tell us about inductive category learning? *Frontiers in psychology*, 6:505.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20:578–585.

George Kachergis, Chen Yu, and Richard M Shiffrin. 2012. An associative model of adaptive inference for learning word–referent mappings. *Psychonomic bulletin & review*, 19(2):317–324.

Chris A Lawson. 2017. The influence of task dynamics on inductive generalizations: How sequential and simultaneous presentation of evidence impacts the strength and scope of property projections. *Journal of Cognition and Development*, 18(4):493–513.

Percy Liang, Michael I Jordan, and Dan Klein. 2009. Probabilistic grammars and hierarchical dirichlet processes. *The handbook of applied Bayesian analysis*.

Brian MacWhinney. 1991. *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Inc.

Tamara Nicol Medina, Jesse Snedeker, John C Trueswell, and Lila R Gleitman. 2011. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019.

Lisa Pearl, Sharon Goldwater, and Mark Steyvers. 2010. Online learning mechanisms for bayesian models of word segmentation. *Research on Language & Computation*, 8(2):107–132.

Sepideh Sadeghi and Matthias Scheutz. 2017. Joint acquisition of word order and word referent in a memory-limited and incremental learner. In *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*.

Sepideh Sadeghi and Matthias Scheutz. 2018. Early syntactic bootstrapping in an incremental memory-limited word learner. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Sepideh Sadeghi, Matthias Scheutz, and Evan Krause. 2017. An embodied incremental bayesian model of cross-situational word learning. In *Proceedings of the 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (icdl-epirob)*.

John P Spencer, Sammy Perone, Linda B Smith, and Larissa K Samuelson. 2011. Learning words in space and time: Probing the mechanisms behind the suspicious-coincidence effect. *Psychological science*, 22(8):1049–1057.

Jon S Stevens, Lila R Gleitman, John C Trueswell, and Charles Yang. 2017. The pursuit of word meanings. *Cognitive science*, 41(S4):638–676.

John C Trueswell, Tamara Nicol Medina, Alon Hafri, and Lila R Gleitman. 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1):126–156.

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114:245–272.

Chen Yu and Dana H Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70:2149–2165.

Yayun Zhang and Chen Yu. 2017. How misleading cues influence referential uncertainty in statistical cross-situational learning. In *41st annual Boston University Conference on Language Development*, pages 820–833.