On the Utility of Curricula in Unsupervised Learning of Probabilistic Grammars

Kewei Tu
Department of Computer Science
Iowa State University
Ames, IA 50011, USA
tukw@iastate.edu

Vasant Honavar
Department of Computer Science
Iowa State University
Ames, IA 50011, USA
honavar@iastate.edu

Abstract

We examine the utility of a curriculum (a means of presenting training samples in a meaningful order) in unsupervised learning of probabilistic grammars. We introduce the *incremental construction hypothesis* that explains the benefits of a curriculum in learning grammars and offers some useful insights into the design of curricula as well as learning algorithms. We present results of experiments with (a) carefully crafted synthetic data that provide support for our hypothesis and (b) natural language corpus that demonstrate the utility of curricula in unsupervised learning of probabilistic grammars.

1 Introduction

A *grammar* consists of a set of rules for forming strings over an alphabet. The rules of the grammar collectively describe how to generate sentences that belong to a *language* (i.e., are grammatically valid). The grammar can also be used to recognize whether a given sentence is grammatically valid, and to derive the parse (grammatical structure) of any valid sentence. A stochastic or probabilistic grammar augments the grammar rules with conditional probabilities. The probability of a parse is the product of the probabilities of the rules used in the parse. Examples of probabilistic grammars include hidden Markov models, probabilistic context-free grammars and probabilistic dependency grammars. They have many applications including natural language processing, DNA, RNA and protein sequence analysis, and pattern recognition.

Machine learning offers a potentially powerful approach to learning unknown grammars from data (a training corpus). Because of the high cost of manual labeling of parsed sentences, there is substantial interest in unsupervised grammar learning, which induces a grammar from unannotated sentences. Much of the existing work on unsupervised grammar learning [Lari and Young, 1990; Klein and Manning, 2004; Cohen *et al.*, 2008] starts with all the sentences of a training corpus and tries to learn the whole grammar. In contrast, there is a substantial body of evidence that humans and animals learn much better when the data are not randomly presented but organized into a *curriculum* that helps expose the learner to progressively more complex concepts or grammatical structures. Such a learning strategy has been termed *curriculum learning* by Bengio et al. [2009]. There has been some effect to apply curriculum learning to unsupervised grammar learning. The results of a seminal experimental study by Elman [1993] suggested that grammar induction using recurrent neural networks can benefit from *starting small*, i.e., starting with restrictions on the data or on the capacity of the learner, and gradually relaxing the restrictions. However, the experiments of Rohde and Plaut [1999] called into question the benefits of *starting small* in language acquisition. A more recent study by Spitkovsky et al. [2010] offered evidence that is suggestive of the benefits of curricula in probabilistic grammar induction. To explain the benefits of curricula, Bengio et al. [2009] hypothesized that a well-designed curriculum corresponds to learning starting with a smoothed objective function and gradually reducing the degree of smoothing over successive stages of the curriculum, thus guiding the learning to better local minima of a non-convex objective function. The precise conditions on the curriculum or the learner that lead to improved learning outcomes are far from well-understood.

Against this background, we explore an alternative explanation of the benefits of curricula, especially in the context of unsupervised learning of probabilistic grammars. Our explanation is based on the *incremental construction hypothesis* (ICH) which asserts that when the target of learning is a structure (in our case, a probabilistic grammar) that can be decomposed into a set of sub-structures (in our case, grammar rules), an ideal curriculum gradually emphasizes data samples that help the learner to successively discover new sub-structures. This hypothesis, if true, can help guide the design of curricula as well as learning algorithms. We present results of experiments on synthetic data that provide support for ICH; and we demonstrate the utility of curricula in unsupervised learning of grammars from a real-world natural language corpus.

2 Curriculum Learning

As noted by Bengio et al. [2009], at an abstract level a curriculum can be seen as a sequence of training criteria. Each training criterion in the sequence is associated with a different set of weights on the training samples, or more generally, with a re-weighting of the training distribution. Thus, we can model a curriculum as a sequence of weighting schemes $\langle W_1, W_2, \ldots, W_n \rangle$. The first weighting scheme W_1 assigns larger weights to “easier” samples, and each subsequent weighting scheme increases the weights assigned to "harder" samples, until the last weighting scheme W_n that assigns uniform weights to the training samples. The measure of "hardness" of training samples depends on the learning problem and learning algorithm. Ideally, the information entropy of the weighting schemes increases monotonically, i.e., $\forall i < j, H(W_i) < H(W_j)$. Given a curriculum, learning proceeds in an iterative fashion: at iteration i, the learner is initialized with the model f_{i-1} learned from the previous iteration, and is provided with the training data weighted by the weighting scheme W_i, based on which it generates a new model f_i. The final output of curriculum learning is f_n, the model produced by the last (n-th) iteration.

The baby-step algorithm [Spitkovsky *et al.*, 2010] for unsupervised grammar learning can be seen as an instance of learning with a curriculum. The training data consist of a set of unannotated sentences. The hardness of a sentence is measured by its length (number of words). The i-th weighting scheme W_i assigns a weight of one to each sentence that consists of no more than i words and a weight of zero to any of the other sentences (thus specifying a subset of training sentences). At iteration i of learning, the expectation-maximization algorithm [Lari and Young, 1990] is run to convergence on the subset of training data specified by W_i, and the resulting grammar G_i is then used to initialize iteration $i + 1$. This curriculum introduces increasingly longer sentences into the training data seen by the learner, with the entire training corpus being provided to the learner at the last iteration, which produces the final output grammar.

3 The Incremental Construction Hypothesis of Curriculum Learning

We explore the incremental construction hypothesis (ICH) as a possible explanation of curriculum learning, in the context of learning probabilistic grammars. The hypothesis asserts that an ideal curriculum gradually emphasizes data samples that help the learner to successively discover new substructures (i.e., grammar rules) of the target grammar, which facilitates the learning. Formally, we define an ideal curriculum for grammar learning suggested in ICH as follows.

Definition 1 *A curriculum $\langle W_1, W_2, \ldots, W_n\rangle$ for learning a probabilistic grammar G of a pre-specified class of grammars C is said to satisfy incremental construction if the following three conditions are met.*

1. *for any weighting scheme W_i, the weighted training data corresponds to a sentence distribution defined by a probabilistic grammar $G_i \in C$;*
2. *if R_i and R_j denote the sets of rules of the probabilistic grammars G_i and G_j respectively, then for any i, j s.t. $1 \le i < j \le n$, we have $R_i \subseteq R_j$;*
3. *for any i, j s.t. $1 \le i, j \le n$, and for any two grammar rules r_1, r_2 with the same rule condition (left-hand side) that appear in both G_i and G_j, we have*

$$\frac{P(r_1|G_i)}{P(r_2|G_i)} = \frac{P(r_1|G_j)}{P(r_2|G_j)}$$

In order words, an ideal curriculum that satisfies incremental construction specifies a sequence of intermediate target grammars $\langle G_1, G_2, \ldots, G_n\rangle$, and each intermediate grammar G_i is a sub-grammar of the next intermediate grammar G_{i+1}. Note that curriculum learning requires the last weighting scheme W_n to be uniform, so given enough training data, the last grammar G_n in the sequence should be weakly equivalent to the target grammar G, i.e., they define the same distribution of sentences.

The third of the three conditions in Definition 1 implies that for a grammar rule that appears in two consecutive grammars, its probability either remains unchanged or, if one or more new rules that share the same left-hand side of the rule are introduced in the second grammar, is renormalized to a smaller value that preserves the probability ratios of this rule to other rules that share the same left-hand side. However, since the training data is usually sparse and sometimes noisy in practice, it would be almost impossible to find a curriculum that exactly satisfies the third condition. Therefore we can relax this condition as follows.

3b. *for any i, j s.t. $1 \le i < j \le n$, and for any grammar rule r that appears in both G_i and G_j, we have $P(r|G_i) \ge P(r|G_j)$*

In order to be able to meaningfully assess the benefits of curricula in grammar learning, we need some measures of distance between two probabilistic grammars. There are two commonly used measures. The first is the distance between the parameter vectors (i.e., the vectors of rule probabilities) of the two grammars. For each rule condition p in grammar G_i, the probabilities of the grammar rules with condition p constitute a multinomial vector (in the case that G_i contains no such rule, we add a dummy rule $p \rightarrow \varepsilon$ with probability 1). Let the parameter vector θ_i of a grammar G_i be the concatenation of the multinomial vectors of all the rule conditions. To make the parameter vectors of different grammars comparable, the elements of different parameter vectors are aligned such that a given rule occupies the same position in the parameter vector of each of the grammars $G_1 \ldots G_n$. The second distance measure is the distance between the distributions of grammatical structures (parses) defined by the two grammars. We can use the *total variation distance* of two distributions (defined as one half of the L_1 distance between them) for this purpose.

Now we can express the advantages of an ICH-based ideal curriculum (Definition 1) in the form of the following theorem.

Theorem 1 *If a curriculum $\langle W_1, W_2, \ldots, W_n\rangle$ satisfies incremental construction (with either condition 3 or 3b), then for any i, j, k s.t. $1 \le i < j < k \le n$, we have*

$$\begin{aligned} d_1(\theta_i, \theta_k) &\ge d_1(\theta_j, \theta_k) \\ d_{TV}(G_i, G_k) &\ge d_{TV}(G_j, G_k) \end{aligned}$$

where $d_1(\cdot, \cdot)$ denotes the L_1 distance; $d_{TV}(G_i, G_j)$ represents the total variation distance between the two distributions of grammatical structures defined by G_i and G_j.

The proof of the theorem exploits the fact that both the L_1 norm of the parameter vector and the sum of probabilities over all grammatical structures are constant regardless of the values of i, j and k. We give the detailed proof in [Tu and Honavar, 2011]. This theorem shows that for any $i < j < k$, G_j is a better approximation of G_k than G_i. Therefore, it follows that each stage of curriculum learning tries to induce a grammar that provides a better initialization for the next stage of learning than any of the previous grammars, and the sequence of grammars $\langle G_1, G_2, \ldots, G_n \rangle$ offers a guided sequence of intermediate learning targets culminating in G_n.

In the case of some curricula that have been used in practice (e.g., the length-based curriculum in [Spitkovsky *et al.*, 2010]), condition 3b appears to be still too strong. As will be shown in Section 5, a curriculum may gradually introduce a new grammar rule to the learner across multiple stages. In this case, the probability of the new rule in the sequence of intermediate target grammars does not instantly jump from 0 to its actual value, but instead increases from 0 to its actual value through a series of small changes over several stages. We can prove a theorem similar to Theorem 1 in this setting:

Theorem 2 *If a curriculum $\langle W_1, W_2, \ldots, W_n \rangle$ satisfies the first two conditions in Definition 1 as well as a further relaxed version of the third condition:*

3c. for any grammar rules r, $P(r|G_i)$ first monotonically increases with i and then monotonically decreases with i.

then for any i, j, k s.t. $1 \leq i < j < k \leq n$, we have

$$d_1(\theta_i, \theta_k) \geq d_1(\theta_j, \theta_k)$$

The proof is similar to that of Theorem 1 and is given in [Tu and Honavar, 2011]. However, under condition 3c, the second inequality for the total variation distance of grammars in Theorem 1 no longer holds.

3.1 Guidelines for Curriculum Design and Algorithm Design

ICH offers some guidance on how to design effective curricula. First, an effective curriculum should approximately satisfy the three conditions discussed above. Second, it should effectively break down the target grammar to be learned into as many chunks as possible, so that at each stage of learning the set of new rules introduced by the curriculum can be small and hence easy to learn. Quantitatively, this makes the distance between any two consecutive grammars G_i and G_{i+1} in the sequence $G_1 \ldots G_n$ as small as possible. Third, at each iteration an effective curriculum should introduce the new rule that results in the largest number of new sentences being added into the training data seen by the learner. This ensures that the learner has as many training sentences as possible for learning the new rule. From a theoretical perspective, since each new rule introduced into a grammar leads to some new grammatical structures that were previously invalid (i.e., had zero probabilities in the absence of the new rule), ideally at each iteration the curriculum should introduce the rule that leads to a set of new grammatical structures with the highest sum of probabilities. The third guideline entails two special cases. First, if there are dependencies between rules (i.e., one rule is required for the other rule to be used), then the curriculum should conform to the the partial order defined by the dependencies. Second, among rules that share the same left-hand side, the curriculum should introduce rules in the descending order of their probabilities in the target grammar.

ICH also offers some guidance on designing learning algorithms. Because the learning target at each stage of the curriculum is a partial grammar, it is especially important for the learning algorithm to avoid the over-fitting to this partial grammar that hinders the acquisition of new grammar rules in later stages. Indeed, from our experiments (see the next two sections), we find that if adequate care is not exercised to minimize over-fitting, the results of learning with a curriculum can be worse than the results of learning without curriculum.

4 Experiments on Synthetic Data

To explore the validity of ICH, we designed a set of experiments using synthetic data generated from a known target grammar. With the target grammar known, we were able to construct the ideal curricula suggested by ICH. We used a grammar formalism called the dependency model with valence (DMV) [Klein and Manning, 2004], which has been shown to be amenable to unsupervised learning. We used the dependency treebank grammar of WSJ30 (the set of sentences no longer than 30 in the Wall Street Journal corpus of the Penn Treebank) as our target grammar, and generated a corpus of 500 sentences using this grammar. Expectation-maximization (EM) was used as the base learning algorithm. To deal with the problem of over-fitting mentioned in Section 3.1, we used a dynamic smoothing factor that is set to a large value initially when the effective training set seen by the learner is small; and is decreased as the learner is exposed to more training data. Five-fold cross-validation was used for evaluation: each time 100 sentences were used for training and the rest were used for evaluation. The results reported correspond to averages over the 5 cross-validation runs. Since we knew the correct parses of all the sentences, we used the standard PARSEVAL measures [Manning and Schütze, 1999] to evaluate the learned grammars.

We compared the performance of the learning algorithm when trained with seven different types of curricula as well as without a curriculum. In each curriculum, we used weights of either zero or one in the weighting schemes, which is tantamount to selecting a subset of the training corpus at each stage of the curriculum.

IDEAL CURRICULA that satisfy all the ICH-based guidelines of curriculum design. We construct a curriculum as follows. Given the target grammar and the training set, at each stage of the curriculum we add to the partial grammar the smallest number of new rules of the target grammar that lead to the largest number of new sentences being added to the training set seen by the learner. We assign weight one to each of the training sentences that can be generated by the partial grammar. When there is a tie between two sets of new rules, we randomly select one.

SUB-IDEAL CURRICULA that satisfy the first two guidelines of curriculum design. At each stage, we randomly add a new rule to the partial grammar and assign weight one to each of the sentences in the training corpus that can be generated by the partial grammar.

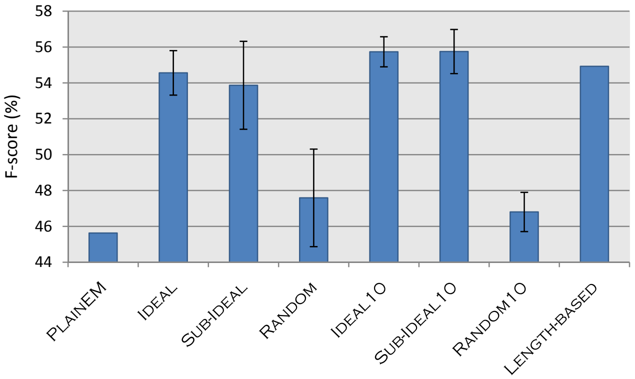


Figure 1: Comparison of the PARSEVAL F-scores of plain EM and learning with seven types of curricula. For each of the six types of curricula that involve nondeterministic construction, ten different curricula were constructed and tested and the mean F-score and standard deviation is shown.

- **Random Curricula** that add new training sentences at random to the training set at each stage of the curricula. We set the number of stages to be the same as that of Ideal Curricula to ensure a fair comparison.
- **Ideal10, Sub-Ideal10 and Random10** curricula that are variants of Ideal, Sub-Ideal and Random curricula respectively except that each stage in the curricula introduces at least 10 new training sentences. Therefore these curricula contain fewer stages.
- **Length-based Curriculum** that introduces new training sentences ordered by their lengths, such that the learner is exposed to shorter sentences before it encounters longer sentences, as described in Section 2.

Figure 1 shows the mean PARSEVAL F-score from cross-validation for each type of curricula as well as learning without curriculum (labeled as PlainEM). The construction procedure of the first six types of curricula is nondeterministic, so we present the mean F-score and standard deviation obtained from experiments with ten curricula of each type.

The results of these experiments show that learning with any of the seven types of curricula, including the random ones, leads to better performance than learning without a curriculum. A possible explanation for the observed gains from the two types of random curricula could be that the target grammar used in this experiment tends to use a rather different set of rules to generate each sentence in the corpus, which would imply that with a small training corpus like ours, even a random partition of the sentences is likely to yield a curriculum that satisfies incremental construction to some extent. The results obtained using the four types of ideal and sub-ideal curricula are significantly better than those obtained using the random curricula. This is consistent with ICH (i.e., the first guideline of curriculum design). Each of the two types of ideal curricula has a slightly better mean F-score and a smaller standard deviation than the corresponding sub-ideal curricula, which suggests that the third guideline of curriculum design also helps facilitate learning.

	Length-based vs Ideal	Sub-Ideal vs Ideal	Random vs Ideal
Kendall	0.7641	0.4125	0.0306
Spearman	0.9055	0.5672	0.0442

Table 1: Average correlations of three types of curricula with the Ideal curricula. Two types of rank correlation, Kendall's and Spearman's correlation, are shown.

However, to the contrary of the second guideline, Ideal and Sub-Ideal have slightly worse performance than Ideal10 and Sub-Ideal10. We speculate that it is because curricula with more stages are more prone to the over-fitting problem discussed in Section 3.1.

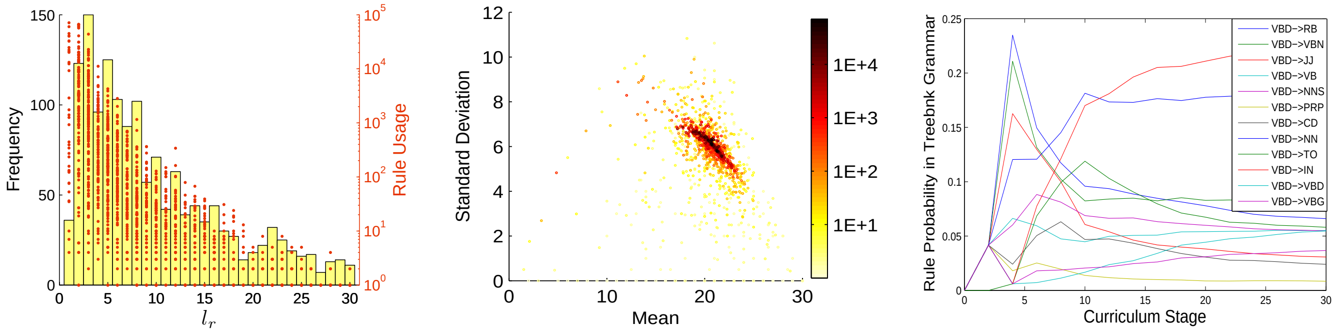
Interestingly, Length-based Curriculum shows performance that is comparable to the four types of ideal and sub-ideal curricula. To explore why this might be the case, we measured how similar the Length-based curriculum is to the Ideal curricula. Since in this set of experiments, each curriculum corresponds to an ordering of the sentences in the training corpus, we can compute the correlation between the orderings to measure the similarity of different curricula. We used two types of rank correlation, Kendall's correlation and Spearman's correlation, for this purpose. Table 1 shows the correlation between Length-based and Ideal, along with the correlations of Sub-Ideal and Random with Ideal for comparison. Because our experiments used ten different Ideal, Sub-Ideal and Random curricula, we report the average values of the correlations between curricula of different types. It can be seen that the Length-based curriculum is very similar to the Ideal curricula in the case of the training corpus and target grammar used in this experiment.

5 Experiments on Real Data

5.1 Analysis of Length-based Curriculum

In practice, since little is known about the target grammar when doing unsupervised learning, it is very difficult, if not impossible, to construct an ideal curriculum suggested by ICH. Hence, curricula that can be constructed without knowledge of the target grammar are preferred. The length-based curriculum offers an example of such curricula. In Section 4, we have shown that on the synthetic data generated from a real-world treebank grammar, the length-based curriculum is a good approximation of an ideal curriculum. In this subsection, we offer some evidence that this may still be true in the case of a real-world natural language corpus.

We use the WSJ30 corpus (the set of sentences no longer than 30 in the Wall Street Journal corpus of the Penn Treebank) to learn a DMV grammar. Since we know the correct parse of each sentence in WSJ30, we can find the grammar rules that are used in generating each sentence. For a grammar rule r, let S_r be the set of sentences in which r is used, and let l_r be the length of the shortest sentence in S_r. Some statistics of grammar rule usage in WSJ30 are shown in Figure 2(a) and 2(b). The histogram in Figure 2(a) in fact shows the distribution of the stages at which the grammar rules are introduced in the length-based curriculum. It can be seen that



(a) The bar graph shows the histogram of l_r (the length of the shortest sentence in the set of sentences that use rule r). Each point in the overlay corresponds a grammar rule r (with x-coordinate being l_r and y-coordinate being the number of times rule r is used in the whole corpus).

(b) Each point in the plot corresponds to a grammar rule r, with its x-coordinate being the mean length of sentences in S_r (the set of sentences in which rule r is used), y-coordinate being the corresponding standard deviation, and color indicating the number of times r is used in the whole corpus (with hotter colors denoting greater frequency of usage).

(c) The change of probabilities of VBD-headed rules with the stages of the length-based curriculum in the treebank grammars (best viewed in color). Rules with probabilities always below 0.025 are omitted.

Figure 2: Analysis of the length-based curriculum in WSJ30

the introduction of grammar rules is spread throughout the entire curriculum, as required by ICH (although more rules are introduced in the early stages). From the overlay plot in Figure 2(a) we can also see that rules that are used more frequently tend to be introduced earlier in the curriculum, which is consistent with the third guideline of curriculum design in Section 3.1. In Figure 2(b), most rules fall within a continuum that ranges from intermediate mean and high standard deviation to high mean and low standard deviation. This suggests that for any grammar rule r, in most cases, the lengths of the sentences in S_r distribute relatively evenly in the interval of $[l_r, 30]$ (where 30 is the length of the longest sentence in WSJ30). So in the length-based curriculum, rules learned in earlier stages can help parse the sentences introduced in later stages of the curriculum, thus facilitating the acquisition of new rules in later stages. This is also consistent with the third guideline of curriculum design.

With the correct parses being known for all the sentences in WSJ30, we can further construct the treebank grammar, in which the rule probabilities are computed from the number of times each rule is used in the parsed corpus. Since each stage of the length-based curriculum specifies a subset of the training sentences, we can construct a sequence of such treebank grammars, one for each stage in the curriculum. Each such grammar is the maximal likelihood grammar of the correct parses of the corresponding sub-corpus, so we can assume that condition 1 in Definition 1 is satisfied. Since each stage of the length-based curriculum adds new sentences to the sub-corpus that is available to the learner, it is easy to see that in this sequence of treebank grammars, once a rule is learned its probability can never drop to zero. This ensures that condition 2 in Definition 1 is also satisfied. How about condition 3? Figure 2(c) shows, for grammar rules that are conditioned on the VBD (past tense verb) head and the right dependency, how the rule probabilities change over the sequence of treebank grammars. We note that most rule probabilities shown in the figure first increase over multiple stages (implying that the rules are being gradually introduced), and then monotonically decrease (due to renormalization of the probabilities as other rules are being introduced). We find that other grammar rules also behave similarly in relation to the sequence of treebank grammars. Therefore, the original condition 3 in Definition 1 is clearly violated, but its relaxed version, condition 3c in Theorem 2, is approximately satisfied. Therefore, the theoretical guarantee of Theorem 2 is likely to hold for the length-based curriculum for the WSJ30 corpus.

Furthermore, from Figure 2(c) we can see that rules are introduced in a specific order. Among the first rules to be introduced are those that produce RB, VBN, JJ and VB (as adverbials, predicatives, etc.); followed by rules that produce NNS, PRP and CD (as objects, etc.); followed by rules that produce NN (as objects) and TO (to head preposition phrases); and ending with rules that produce IN, VBD and VBG (for preposition phrases and clauses). This confirms that rules are introduced incrementally in the length-based curriculum.

5.2 Learning Results

We tested curriculum learning of DMV grammars from the unannotated WSJ30 corpus. Following the standard procedure for evaluating natural language parsers, section 2-21 of WSJ30 were used for training, section 22 was used for development, and section 23 was used for testing. We used expectation-maximization (EM) as the base learning algorithm, with an initialization of the grammar as described in [Klein and Manning, 2004]. To minimize the over-fitting problem discussed in Section 3.1, at each stage of the curriculum we terminated training when the likelihood of the development set stopped increasing. In addition, we set the maximal number of iterations at each stage (except the last stage) of the curriculum to a relatively small value, which further alleviates over-fitting while also speeding up the algorithm.

In addition to plain EM and the length-based curriculum,

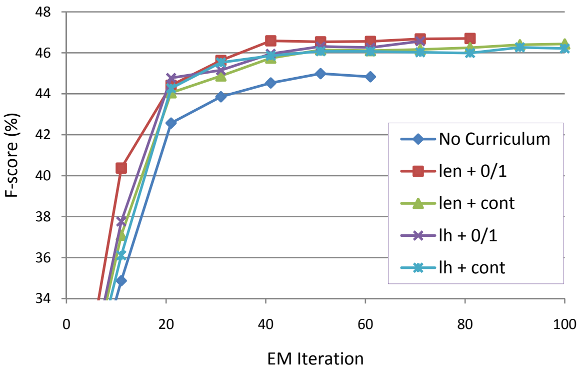


Figure 3: The change of F-scores with the EM iterations. "len" denotes length-based curriculum; "lh" denotes likelihood-based curriculum; "0/1" denotes that weights are set to be either zero or one; "cont" denotes that a continuous-valued weighting function is used in the weighting schemes.

we tested a novel curriculum based on the likelihood of sentences. Because the use of EM as the base learning algorithm guarantees that at any time of the learning we have a complete grammar, we can use the negative log likelihood of a sentence given this grammar as a measure of the relative hardness of the sentence. With this likelihood-based hardness measure, we can construct a new curriculum similar to the length-based curriculum, i.e., sentences with higher likelihood receive larger weights at earlier stages in the curriculum. However, because the grammar used to estimate the hardness of a sentence is continuously updated as a result of learning, so is the hardness measure, making the resulting curriculum an "active" curriculum. We repeated the analysis described in Section 5.1 on this new curriculum, and found the results similar to those reported for the length-based curriculum (data not shown).

In the curricula discussed in Section 4, the weights are set to either zero or one in the weighting schemes, and the set of sentences with weight one expands over successive stages of the curriculum. Here we also tested a different method: a continuous-valued weighting function is used to assign greater weights to easier sentences and less weights to harder sentences, and the weighting function becomes increasingly uniform over successive stages of the curriculum.

We evaluated all the intermediate grammars produced in the course of learning as well as the grammars that was output at the end, using the PARSEVAL metric [Manning and Schütze, 1999]. Figure 3 shows how the F-score changes with the EM iterations when learning with each of four different curricula as well as in the no-curriculum baseline. It can be seen that learning with a curriculum consistently converges to a grammar with a better F-score than the no-curriculum baseline. Also, during the early stages of learning, the use of curricula results in faster improvements in F-score as compared to the no-curriculum baseline. The four curricula behave similarly, with the length-based curriculum using zero/one weights performing slightly better than the others.

We also plotted the change of rule probabilities during learning with a curriculum (data not shown; see [Tu and Honavar, 2011]). The overall trends are very similar to those seen in Figure 2(c): the probability of each rule first rises and then drops, and rules are learned in a specific order.

6 Conclusion

We have provided an explanation of the benefits of curricula in the context of unsupervised learning of probabilistic grammars. Our explanation is based on the *incremental construction hypothesis* which asserts that an ideal curriculum gradually emphasizes data samples that help the learner to successively discover new sub-structures of the target grammar. The hypothesis offers some guidance on the design of curricula as well as learning algorithms. We have presented results of experiments on synthetic data that provide support for the incremental construction hypothesis; we have further demonstrated the utility of curricula in unsupervised learning of grammars from a real-world natural language corpus.

Acknowledgement

The work of Kewei Tu was supported in part by a research assistantship from the Iowa State University Center for Computational Intelligence, Learning, and Discovery. The work of Vasant Honavar was supported by the National Science Foundation, while working at the Foundation. Any opinion, finding, and conclusions contained in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, page 6, 2009.

[Cohen *et al.*, 2008] Shay B. Cohen, Kevin Gimpel, and Noah A. Smith. Logistic normal priors for unsupervised probabilistic grammar induction. In *NIPS*, pages 321–328, 2008.

[Elman, 1993] Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99, 1993.

[Klein and Manning, 2004] Dan Klein and Christopher D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of ACL*, 2004.

[Lari and Young, 1990] K. Lari and S. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–36, 1990.

[Manning and Schütze, 1999] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[Rohde and Plaut, 1999] D. Rohde and D. Plaut. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72:67–109, 1999.

[Spitkovsky *et al.*, 2010] Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *NAACL*, 2010.

[Tu and Honavar, 2011] Kewei Tu and Vasant Honavar. On the utility of curricula in unsupervised learning of probabilistic grammars (supplementary material). Technical report, Computer Science, Iowa State University. Available at `http://archives.cs.iastate.edu/`, 2011.