

Domain-Weighted Batch Sampling for Neural Dependency Parsing

Jacob Striebel, Daniel Dakota, Sandra Kübler

Indiana University

{jstrieb,ddakota,skuebler}@indiana.edu

1. Summary

• We propose *domain-weighted* batch sampling (DWBS) as a domain adaptation strategy for supervised neural learning.

 We show that DWBS outperforms conventional randomized batch sampling (RBS).

2. Domain-Weighted Batch Sampling (DWBS)

• To perform DWBS, before training begins the training data set is partitioned into disjoint *in*domain and out-of-domain subsets.

Algorithm 1 DomainWeightedBatchGenerator **Require:** training data set *D* partitioned into subsets D_1 and D_2 **Require:** domain-weight parameter $\mu \in [0, 1]$ **Require:** batchSize > 0while validation loss continues to decline do $R_1 \leftarrow \text{shuffle}(\text{copy}(D_1))$ $R_2 \leftarrow \text{shuffle}(\text{copy}(D_2))$ while True do batch \leftarrow for $n \leftarrow 1$ to batchSize do binaryChoice \leftarrow BernoulliSample(μ) if binaryChoice = 0 then if R_1 .hasNext() then $batch.append(R_1.next())$ else

break

break

if R_2 .hasNext() then

if batch.length() < batchSize then

 $batch.append(R_2.next())$

end if

else

end if

break

yield batch

end for

end if

end while

end while

end if

else

4. Results cont.

• In order to evaluate the effectiveness of DWBS, we perform experiments in which we compare a baseline model trained using conventional RBS against domainexpert parsers trained using DWBS.

4.1. Effect on Parsing Accuracy

ТВ	Domain	μ	LAS R	LAS DW	Table 2: Performance in LAS per
EWT	Answers	0.35	86.78	87.56	domain, comparing the baseline parser
	Email	0.35	86.70	88.00	to the highest-LAS-producing domain-
	Newsgr.	0.40	88.64	89.44	ovport parcor LAS Producing domain
	Reviews	0.35	88.27	88.74	expert parser. LAS R. Dasenne parser
	Weblog	0.25	89.52	90.56	trained using RBS; LAS DW: nignest-
GUM	Convers.	0.35	85.41	86.64	LAS-producing domain-expert parser
	Fiction	0.45	89.86	91.23	trained using DWBS; μ : setting resulting
	Interv.	0.50	88.08	89.14	in the highest LAS for the given domain.
	Vlog	0.60	87.74	88.57	Improvements of more than 1.00 LAS
	Whow	0.35	90.46	91.11	are bolded.

- The hyperparameter μ is used to define the probability of choosing the next sample from the in-domain subset.
- So, for example, if $\mu = 0.45$, then there is a 45% chance of drawing the next sample from the in-domain subset and a 55% chance of drawing from the out-of-domain subset.

3. Methodology

3.1. Data

• We use Universal Dependencies treebanks version 2.12 (Nivre et al., 2020; de Marneffe et al., 2021), more specifically the English Web Treebank (EWT; Bies et al., 2012) and the Georgetown University Multilayer Corpus (GUM; Zeldes, 2017).

• From the sixteen domains of EWT and GUM, we select only the ten domains that each have a minimum of 1,000 sentences, which includes all five EWT domains and five of the eleven GUM domains.

85.0

	3.2. Parser	Hyperparameter	Value								
• We use the deep hiaffine attention neural		Optimizer	Adamw	4.2. Effect on Training Duration							
de	ependency parser (Dozat and Manning.	β_1, β_2	0.9, 0.99 False	Treeb.	Domain		BBS NSC	DWBS NSC		Table 2: Training	
20	(17) in the implementation by van der	Learning rate	0.0001	EWT	Answers	0.35	40.40	40.88	0.48	tion por domain	
G	oot et al. (2021).	Weight decay	0.01		Email	0.35	39.84	40.00	0.16	d in number of	
		Gradient normalization	1		Newsgroup	0.40	45.60	45.44	-0.16		
• \//	e modify the parser so that it can be	LR scheduler	Slanted triangular		Reviews	0.35	40.96	41.36	0.40	Sanus of Samples	
CO	nfigured to perform DWRS	Decay factor	0.2		Weblog	0.25	47.04	48.00	0.96		
		Decay factor Discriminative fine tuning	True	GUM	Conversation	0.35	45.52	41.20	-4.32	comparing the ba	
• \//	e use the hypernarameter settings	Gradual unfreezing	True		Fiction	0.45	40.56	42.96	2.40	parser to the high	
nr	ovided by van der Goot et al with the	Batch size	32			0.50	45.20	42.00	-3.20	LAS-producing a	
	by modification being that we specify	Patience batches	200		Vlog	0.60	48.16	42.96	-5.20	expert parser. No	
	rly stopping pations in torms of	Max steps	153,600		Whow	0.35	40.40	40.24	-0.16] number of thousa	
ea bo	toboc rather than encode (coo Table 1)	Embeddings Embeddings dim	bert-base-cased	training	j samples unt	il mode	el converge	ence; RBS N	SC: NS(C for the baseline	
Da	liches fahler than epochs (see fable 1).		700	trained	using RBS; [OWBS	NSC: NSC	; for the highe	est-LAS	-producing domair	
		Table 1: Parser hy	perparameters	parser	trained using	DWBS	3; <i>μ</i> : settinç	yielding the	best (in	i terms of LAS) do	
				expert	parser for the	given	domain.				
	4. Results				• The domains are evenly split on training time reduction with five seeing a re						
	(a) Performance of "EWT	reviews" parsers		and fiv	e experiencin	ng an ir	icrease.			C	
92.5 -	87 07 87 95 88 27 88 47 88 33 88 14 87 97 88 74 88 55 88 36 88 63 8	8 / 8 88 / / 87 97 87 93 88 21 87 88 9	37 93 87 68 86 93 85 50	• The ar	reatest increa	se is e	vnerience l	hy the GLIM 1	iction d	omain which requ	
90.0 -	07.07 07.55 00.27 00.47 00.55 00.14 07.57 00.74 00.55 00.55 0	0.40 00.44 07.57 07.55 00.21 07.00 0	57.55 67.66 66.55 65.56	2.400	more sentenc	ces tha	n the base	line to achiev	/e conve	ergence.	
S				_,						<u> </u>	
87.5 -		 The greatest decrease is experienced by the GUM vlog domain, which show decrease of 5,200 sentences. 									
	0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 (μ	0.55 0.60 0.65 0.70 0.75 0.80	0.85 0.90 0.95 1.00								
(b) Performance of "GUM fiction" parsers					• The average change in training duration is a decrease of 864 sentences.						
92.5 -	87.73 89.52 89.86 90.26 91.02 90.23 90.82 90.93 90.77 91.23 91.06 9	0.50 90.83 90.51 90.48 90.60 90.24 9	90.40 90.19 89.34 88.22	• The hi	gh variability i	in of di	fferences i	n training dur	ation m	av suddest that or	
				domai	n data do not		s have high	intornal con			
90.0 –						aivav			SISIENCY	V. WHICH IS IN IINE M	

• The DWBS-trained parser outperforms the baseline in all ten domains tested, for some setting of μ .

 The average improvement across all ten domains, using each domain's best setting of *µ*, is 0.95 LAS.

• As shown in Table 2, five domains experience gains of more than 1.00 LAS.

• Overall, GUM domains tend to prefer higher values of μ ; in other words, those domains profit more from training examples from the same domain, which indicates that each of those domains is different from all others, either in terms of syntactic structures or annotation.

 We use the deep biaffine attention neural 	Optimizer	Adamw	4.2. Effect on Training Duration									
dependency parser (Dozat and Manning,	ρ_1, ρ_2 Correction bias	False	Treeb.	Domain	μ	RBS NSC	DWBS NSC	Δ NSC	Table 3. Training dura-			
2017) in the implementation by van der	Learning rate	0.0001	EWT	Answers	0.35	40.40	40.88	0.48	tion per domain measur-			
Goot et al. (2021).	Weight decay	0.01		Email	0.35	39.84	40.00	0.16	ad in number of theu			
	Gradient normalization	1		Newsgroup	0.40	45.60	45.44	-0.16	eu in number of thou-			
• We modify the parser so that it can be	LR scheduler	Slanted triangular		Reviews	0.35	40.96	41.36	0.40	sands of samples until			
configured to perform DWRS	Cut fraction	0.2		Weblog	0.25	47.04	48.00	0.96	model convergence,			
configured to perform DVDS.	Decay factor	0.38 Truo	GUM	Conversation	0.35	45.52	41.20	-4.32	comparing the baseline			
· \//~~ the hyperperenter estimate	Gradual unfreezing	True		Fiction	0.45	40.56	42.96	2.40	parser to the highest-			
• we use the hyperparameter settings	Batch size	32		Interview	0.50	45.20	42.00	-3.20	LAS-producing domain-			
provided by van der Goot et al. with the	Patience batches	200		Vlog	0.60	48.16	42.96	-5.20	expert parser. NSC:			
only modification being that we specify	Max steps	153,600		Whow	0.35	40.40	40.24	-0.16	number of thousands of			
early-stopping patience in terms of	Embeddings	bert-base-cased	training	samples unt	il mode	l converae	ence: RBS NS	SC: NSO	C for the baseline parser			
batches rather than epochs (see Table 1).	Embeddings dim	768	trained	using RBS: C	WBS I	NSC: NSC	for the highe	est-LAS	-producing domain-expert			
	parser trained using DWRS: <i>u</i> : setting vielding the hest (in terms of LAS) domain-											
						expert parser for the given domain						
	copert parser for the given domain.											
4. Resu	• The domains are evenly split on training time reduction with five seeing a reduction and five experiencing an increase.											
(a) Performance of "F\//T												
	reviews parsers											
92.5 87.07 87.95 88.27 88.47 88.33 88.14 87.97 88.74 88.55 88.36 88.63	88.48 88.44 87.97 87.93 88.21 87.88 87.97 87.93 88.21 87.88 87.97 87.93 88.21 87.88 87.87 87.88 87.87 87.87 87.87 87.87 87.87 87.87 87.8	7.93 87.68 86.93 85.50	• The ar	eatest increas	se is ex	oerience k	ov the GUM f	iction de	omain, which requires			
92.5 87.07 87.95 88.27 88.47 88.33 88.14 87.97 88.74 88.55 88.36 88.63	88.48 88.44 87.97 87.93 88.21 87.88 87	7.93 87.68 86.93 85.50	• The gr 2 400	eatest increas	se is ex	perience k	by the GUM f	iction do	omain, which requires			
92.5 87.07 87.95 88.27 88.47 88.33 88.14 87.97 88.74 88.55 88.36 88.63	88.48 88.44 87.97 87.93 88.21 87.88 87	7.93 87.68 86.93 85.50	• The gr 2,400	eatest increas nore sentenc	se is ex es thar	perience k the basel	by the GUM f line to achiev	iction do e conve	omain, which requires ergence.			
92.5 92.5 87.07 87.95 88.27 88.47 88.33 88.14 87.97 88.74 88.55 88.36 88.63 90.0 Second Second Levv 1 97.5 87.07 87.95 88.27 88.47 88.33 88.14 87.97 88.74 88.55 88.36 88.63	88.48 88.44 87.97 87.93 88.21 87.88 87	7.93 87.68 86.93 85.50	• The gr 2,400	eatest increas nore sentenc	se is ex es thar	perience k the basel	by the GUM f line to achiev	iction do e conve 4 vlog do	omain, which requires ergence. omain, which shows a			
92.5 90.0 97.5 87.5 85.0	88.48 88.44 87.97 87.93 88.21 87.88 87	7.93 87.68 86.93 85.50	 The gr 2,400 The gr docrose 	eatest increas nore sentenc eatest decrea	se is ex es thar se is e	xperience k n the basel xperience	by the GUM f line to achiev	iction do re conve 1 vlog do	omain, which requires ergence. omain, which shows a			
$\begin{array}{c} 92.5 \\ 90.0 \\ 90.0 \\ 87.5 \\ 85.0 \\ \hline \\ 0.00 \\ 0.05 \\ 0.10 \\ 0.15 \\ 0.20 \\ 0.25 \\ 0.30 \\ 0.35 \\ 0.30 \\ 0.35 \\ 0.40 \\ 0.45 \\ 0.50 \\ \hline \end{array}$	0.55 0.60 0.65 0.70 0.75 0.80 0	7.93 87.68 86.93 85.50	 The gr 2,400 The gr decrea 	eatest increas nore sentenc eatest decrea se of 5,200 s	se is ex es thar se is e entenc	xperience k n the basel xperienced es.	by the GUM f line to achiev	iction do e conve 1 vlog do	omain, which requires ergence. omain, which shows a			
$\begin{array}{c} 92.5\\ 90.0\\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\$	88.48 88.44 87.97 87.93 88.21 87.88 87 0.55 0.60 0.65 0.70 0.75 0.80 0	7.93 87.68 86.93 85.50	 The gr 2,400 The gr decreation 	eatest increas nore sentenc eatest decrea se of 5,200 s	se is ex es than se is e entenc	xperience k n the basel xperienced es.	by the GUM f line to achiev d by the GUN	fiction do re conve 1 vlog do	omain, which requires ergence. omain, which shows a			
92.5 90.0 96.5 87.5 85.0 90.0	A fiction" parsers	7.93 87.68 86.93 85.50	 The gr 2,400 The gr decrea The av 	eatest increas nore sentenc eatest decrea se of 5,200 s rerage change	se is ex es thar ase is e entenc e in trai	xperience k n the basel xperienced es. ning durat	by the GUM f line to achiev d by the GUN	fiction do re conve A vlog do ease of a	omain, which requires ergence. omain, which shows a 864 sentences.			
$\begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} $	A fiction" parsers	7.93 87.68 86.93 85.50	 The gr 2,400 The gr decrea The av 	eatest increas nore sentenc eatest decrea se of 5,200 s verage change	se is ex es than se is e entenc e in trai	xperience k n the basel xperienced es. ning durat	by the GUM f line to achiev d by the GUN ion is a decre	iction do re conve 1 vlog do ease of 2	omain, which requires ergence. omain, which shows a 864 sentences.			
$\begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} $	Tevrevs parsers 88.48 88.44 87.97 87.93 88.21 87.88 87 0.55 0.60 0.65 0.70 0.75 0.80 0 A fiction" parsers 90.50 90.83 90.51 90.48 90.60 90.24 90	7.93 87.68 86.93 85.50 1 1 1 1 <td< th=""><td> The gr 2,400 The gr decrea The av The high </td><td>eatest increas nore sentenc eatest decrea se of 5,200 s verage change</td><td>se is ex es thar ase is e entenc e in trai</td><td>xperience k n the basel xperienced es. ning durat</td><td>by the GUM f line to achiev d by the GUN ion is a decre</td><td>fiction do re conve A vlog do ease of a ation ma</td><td>omain, which requires ergence. omain, which shows a 864 sentences. ay suggest that our target</td></td<>	 The gr 2,400 The gr decrea The av The high 	eatest increas nore sentenc eatest decrea se of 5,200 s verage change	se is ex es thar ase is e entenc e in trai	xperience k n the basel xperienced es. ning durat	by the GUM f line to achiev d by the GUN ion is a decre	fiction do re conve A vlog do ease of a ation ma	omain, which requires ergence. omain, which shows a 864 sentences. ay suggest that our target			
$\begin{array}{c} (u) \ r \ c c c r \ c c r \ c c c c r \ c c c c r \ c c c c c c c c c c$	A fiction" parsers	7.93 87.68 86.93 85.50 1 1 1 1 <td< th=""><td> The gr 2,400 The gr decrea The av The high domain </td><td>eatest increas nore sentenc eatest decrea se of 5,200 s rerage change gh variability i n data do not</td><td>se is ex es than ase is e entenc e in trai n of dif always</td><td>xperience k n the base xperienced es. ning durat ferences in have high</td><td>by the GUM f line to achiev d by the GUN ion is a decre n training dur</td><td>iction do re conve 4 vlog do ease of a ation ma sistency</td><td>omain, which requires ergence. omain, which shows a 864 sentences. ay suggest that our target y, which is in line with</td></td<>	 The gr 2,400 The gr decrea The av The high domain 	eatest increas nore sentenc eatest decrea se of 5,200 s rerage change gh variability i n data do not	se is ex es than ase is e entenc e in trai n of dif always	xperience k n the base xperienced es. ning durat ferences in have high	by the GUM f line to achiev d by the GUN ion is a decre n training dur	iction do re conve 4 vlog do ease of a ation ma sistency	omain, which requires ergence. omain, which shows a 864 sentences. ay suggest that our target y, which is in line with			

0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85

(c) Parser performance averaged over all ten domains



Figure 1: Performance of the DWBS-trained domain-expert parsers on (a) EWT reviews, (b) GUM fiction, and (c) averaged over all ten domains. X-axis: domainweight hyperparameter μ ; y-axis: parser performance in LAS. Because in our experimental setup we use ten domains of equal size, whenever $\mu = 0.10$, DWBS is equivalent to conventional RBS; therefore, in each char we highlight the baseline RBS-trained parser in blue, and we highlight the best performing DWBS-trained parser(s) in green.

cross-domain parsing between EWT and GUM.

5. Conclusion

• Based on the positive results reported above, when the preconditions for performing DWBS are met, it should be preferred over conventional RBS.

7. References

[1] Bies, Ann and Mott, Justin and Warner, Colin and Kulick, Seth. 2012. English Web Treebank LDC2012T13. Linguistic Data Consortium. [2] Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In International Conference on Learning Representations. [3] Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 176–197, Online. Association for Computational Linguistics. [4] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics, 47(2):255–308.

[5] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), pages 4034–4043, Marseille, France.

[6] Amir Zeldes. 2017. Georgetown Multilayer Corpus (GUM). Georgetown University Corpus Linguistics Lab.

[7] Amir Zeldes and Nathan Schneider. 2023. Are UD treebanks getting more consistent? a report card for English UD. In Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 58–64, Washington, D.C.

[8] Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1– 21, Brussels, Belgium.