To Annotate or Not?
Predicting Performance Drop under Domain Shift

Hady Elsahar and **Matthias Gallé**
NAVER LABS Europe
{hady.elsahar,matthias.galle}@naverlabs.com

A Appendix

A.1 Dataset and Code Download

All code and created datasets in our work are available to download through the following link: https://github.com/hadyelsahar/domain-shift-prediction. More statistics about the datasets are shown on table 1.

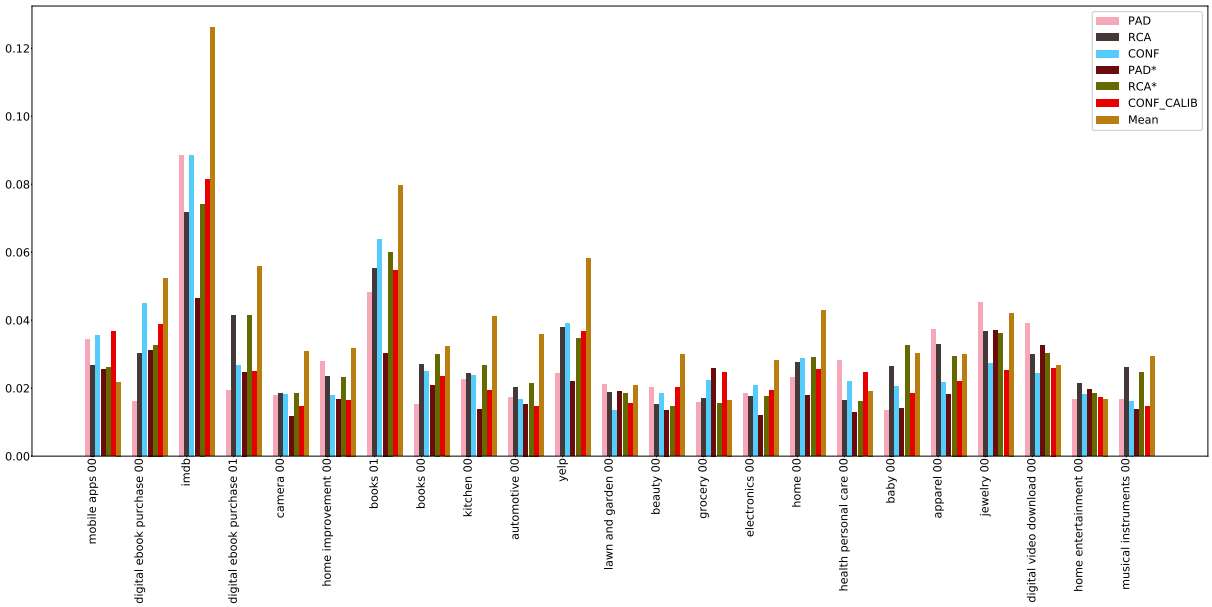

Figure 1: Fig. showing the error in prediction of performance drop for sentiment analysis for each target domain.

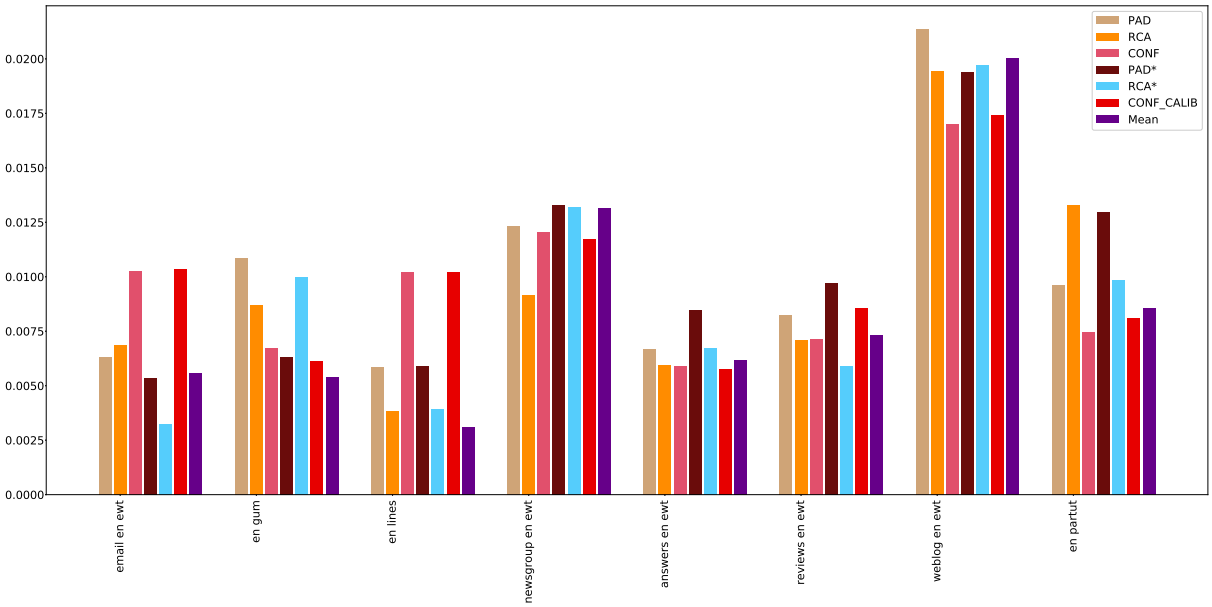

Figure 2: Fig. showing the error in prediction of performance drop for POS tagging for each target domain.

	Sentiment Analysis							
Domain	Train		Dev		Test		Train2	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
apparel	5031	4969	5040	4960	482	518	4970	5030
automotive	5084	4916	4934	5066	497	503	4917	5083
baby	4966	5034	5050	4949	511	489	5069	4931
beauty	4965	5035	5021	4979	507	493	4958	5042
books political	4897	5103	5014	4985	521	479	5043	4957
books fiction	4968	5032	5046	4954	491	509	5000	5000
books historical	5015	4985	5016	4984	490	510	4974	5026
camera	5035	4965	5019	4980	500	500	4955	5045
digital ebook purchase political	5011	4989	5026	4974	496	504	5074	4926
digital ebook purchase historical	5004	4996	4981	5019	529	471	4953	5047
digital video download	5039	4960	5008	4992	503	497	4955	5045
electronics	4969	5031	4985	5015	495	505	4957	5043
grocery	4934	5066	5001	4999	516	484	5045	4955
health personal care	4952	5048	4987	5012	494	506	4991	5009
home	5012	4988	5076	4924	494	506	5085	4915
home entertainment	4996	5004	5046	4954	507	493	4991	5009
home improvement	4961	5039	4990	5010	504	496	4952	5048
imdb	4989	5011	4961	5039	487	513	5019	4981
jewelry	4975	5025	4995	5005	501	499	4957	5043
kitchen	5000	5000	5056	4944	485	515	5018	4982
lawn and garden	5014	4986	5040	4960	513	487	4998	5002
mobile apps	5061	4939	4968	5032	511	489	4972	5028
musical instruments	5028	4972	5028	4972	499	501	4998	5002
yelp	4993	5007	4960	5040	511	489	4920	5080
Total: 2,852,353 Pos. 2,853,603 Neg.								

	POS Tagging					
Domain	Train		Dev		Test	
	Sent.	Tokens.	Sent.	Tokens.	Sent.	Tokens.
en gum	2915	57346	708	14037	779	14266
en lines	2739	50097	913	17108	915	15629
en partut	1782	43544	157	2722	154	3411
answers en ewt	2631	43595	419	5250	438	5402
email en ewt	3770	46296	524	5460	606	6130
newsgroup en ewt	1833	35015	274	4324	284	3807
reviews en ewt	2724	45168	554	5586	535	5565
weblog en ewt	1585	35073	231	4848	214	4509
Total: 27,684 Sentences 474,188 Tokens						

Table 1: Table showing statistics of prepared datasets for pos Sentiment Analysis task and POS tagging. Train2 a is second training dataset from the source domain to calculate the **RCA⋆** measure. In case of POS tagging since the Universal dependency dataset is already split into train/dev/test dataset, to calculate **RCA⋆** we create a second training dataset on the fly by randomly sampling from all other domains $D \setminus \{D_s, D_t\}$

.

	Sentiment Analysis							
Domain	Accuracy%	EMB.	EMB size	$\text{LSTM}_{\text{hidden}}$	$\text{LSTM}_{\text{layers}}$	$\text{Linear}_{\text{layers}}$	Dropout	lr
apparel	91.63	random	180	50	5	1	0.5	0.01
automotive	89.11	glove	100	150	1	1	0.5	0.01
baby	89.16	glove	200	150	1	1	0.5	0.01
beauty	90.41	glove	200	80	1	1	0.5	0.01
books political	89.48	glove	100	150	1	1	0.5	0.01
books fiction	87.6	glove	200	80	1	1	0.5	0.01
camera	90.92	glove	100	80	1	1	0.5	0.01
digital ebook purchase political	90.01	glove	100	150	1	1	0.5	0.01
digital ebook purchase fiction	91.1	glove	100	80	1	1	0.5	0.01
digital video download	90.9	glove	100	80	2	1	0.5	0.01
electronics	89.09	glove	100	150	1	1	0.5	0.01
grocery	88.44	glove	100	80	1	1	0.5	0.01
health personal care	89.56	glove	100	80	2	1	0.5	0.01
home	91.06	glove	200	150	1	1	0.5	0.01
home entertainment	88.67	glove	200	80	1	1	0.5	0.01
home improvement	89.11	random	180	80	2	1	0.5	0.01
imdb	85.9	glove	200	250	1	1	0.5	0.01
jewelry	92.25	glove	100	80	2	1	0.5	0.01
kitchen	90.05	random	180	50	3	1	0.5	0.01
lawn and garden	89.17	random	180	50	3	2	0.5	0.01
mobile apps	90.51	random	180	250	5	1	0.5	0.01
musical instruments	89.17	random	180	50	5	1	0.5	0.01
yelp	86.17	glove	100	80	1	1	0.5	0.01
	POS Tagging							
answers en ewt	97.24	elmo	original	250	2	–	0.5	0.01
email en ewt	98.41	elmo	original	250	2	–	0.5	0.01
en gum	97.71	elmo	original	250	1	–	0.5	0.01
en lines	97.89	elmo	original	80	1	–	0.5	0.01
en partut	98.35	elmo	original	250	1	–	0.5	0.01
newsgroup en ewt	97.53	elmo	original	250	2	–	0.5	0.01
reviews en ewt	97.56	elmo	original	250	1	–	0.5	0.01
weblog en ewt	98.72	elmo	original	250	2	–	0.5	0.01

Table 2: Table showing the best performing hyper-parameters on test dataset for each source domain D_s for both sentiment analysis and POS tagging task.