Weakly-Supervised Learning with Cost-Augmented Contrastive Estimation: *Supplementary Material*

Kevin Gimpel **Mohit Bansal**
Toyota Technological Institute at Chicago, IL 60637, USA
{kgimpel,mbansal}@ttic.edu

1 Tag Bigram Costs

We used treebanks for 11 languages from the CoNLL 2006/2007 shared tasks other than those used in our POS tagging experiments. In particular, we used Arabic, Bulgarian, Catalan, Czech, English, Spanish, German, Hungarian, Italian, Japanese, and Turkish. We replicated shorter treebanks a sufficient number of times until they were a similar size as the largest treebank. Then we counted gold POS tag unigrams and bigrams from the concatenation. In Table 1 we show counts and costs for tag bigrams.

tag bigram	count	cost	tag bigram	count	cost	tag bigram	count	cost
CONJ ⟨EOS⟩	5	115.23	CONJ PRT	8494	40.85	ADJ VERB	38685	25.69
DET ⟨EOS⟩	20	101.36	CONJ NUM	8708	40.60	PRON ADP	38982	25.61
ADP ⟨EOS⟩	30	97.31	NUM ADJ	9204	40.05	VERB CONJ	39641	25.44
DET PRT	109	84.41	PRT PRON	9349	39.89	. ADV	40531	25.22
ADV ⟨EOS⟩	201	78.29	ADP CONJ	9362	39.88	ADV .	41062	25.09
X DET	259	75.75	ADJ DET	9385	39.85	ADV ADJ	41244	25.05
PRT ⟨EOS⟩	281	74.94	PRT ADJ	9646	39.58	ADJ CONJ	44536	24.28
PRON ⟨EOS⟩	406	71.26	CONJ CONJ	9939	39.28	CONJ PRON	45286	24.11
ADV X	486	69.46	ADJ PRT	10069	39.15	NOUN DET	46804	23.78
X PRT	506	69.06	ADV NUM	10207	39.01	⟨BOS⟩ CONJ	48265	23.48
DET CONJ	518	68.82	ADV PRT	10230	38.99	ADP VERB	49139	23.30
X ADV	739	65.27	DET DET	10469	38.76	ADJ ADJ	51444	22.84
PRT X	745	65.19	ADV CONJ	10739	38.50	ADP ADJ	54707	22.22
X ⟨EOS⟩	747	65.16	PRON PRT	10873	38.38	ADP PRON	56097	21.97
X PRON	805	64.41	⟨BOS⟩ X	11226	38.06	CONJ DET	57176	21.78
NUM X	1013	62.11	VERB NUM	11281	38.01	. ADP	58998	21.47
VERB ⟨EOS⟩	1023	62.02	PRON CONJ	11922	37.46	VERB ADJ	59035	21.46
CONJ X	1037	61.88	NOUN ⟨EOS⟩	12334	37.12	NOUN ADV	59291	21.42
PRON X	1141	60.92	ADP ADV	12637	36.88	PRT VERB	59408	21.40
DET X	1282	59.76	NOUN X	13247	36.41	. PRON	59523	21.38
NUM ⟨EOS⟩	1475	58.36	DET NUM	14495	35.51	. DET	60663	21.19
X VERB	1490	58.26	DET PRON	14720	35.35	VERB PRON	62171	20.94
NUM ADV	1587	57.63	CONJ .	14921	35.22	⟨BOS⟩ DET	62920	20.82
X NUM	1630	57.36	NUM ADP	15024	35.15	NOUN PRT	70345	19.71
ADJ ⟨EOS⟩	1845	56.12	ADJ PRON	15396	34.90	. .	71475	19.55
PRT CONJ	1936	55.64	ADP .	15595	34.77	VERB ADV	76624	18.85
NUM PRON	1968	55.47	NUM NUM	15807	34.64	NUM NOUN	78804	18.57
PRT ADP	2039	55.12	PRON DET	16134	34.43	ADV VERB	80126	18.41
ADJ X	2477	53.17	⟨BOS⟩ ADJ	18858	32.87	⟨BOS⟩ NOUN	82206	18.15
NUM DET	2564	52.83	⟨BOS⟩.	18939	32.83	. VERB	84367	17.89
ADP X	2595	52.71	X .	18973	32.81	CONJ VERB	85420	17.77
X ADJ	2667	52.43	PRT PRT	21393	31.61	PRON NOUN	91872	17.04
⟨BOS⟩ PRT	2787	51.99	PRON ADV	21569	31.53	DET ADJ	93111	16.91
VERB X	2809	51.92	ADV PRON	23035	30.87	ADJ ADP	104280	15.77
ADP PRT	2885	51.65	. NUM	23833	30.53	NOUN PRON	106705	15.54
PRT DET	3301	50.30	PRT .	24106	30.42	. CONJ	108929	15.34
X CONJ	4259	47.75	ADV NOUN	25692	29.78	ADJ .	109098	15.32
DET ADP	4276	47.71	ADV DET	27526	29.09	CONJ NOUN	113019	14.97
X ADP	4957	46.24	ADV ADV	28654	28.69	. NOUN	118788	14.47
NUM CONJ	5078	45.99	NOUN NUM	28957	28.59	VERB DET	124640	13.99
. PRT	5782	44.70	NUM .	29359	28.45	NOUN CONJ	141525	12.72
X X	5881	44.53	CONJ ADP	29523	28.39	VERB ADP	150002	12.14
NUM PRT	5896	44.50	VERB PRT	29762	28.31	VERB VERB	150999	12.07
PRON NUM	5944	44.42	⟨BOS⟩ VERB	30087	28.20	VERB .	153257	11.92
ADJ NUM	6243	43.93	CONJ ADV	30668	28.01	VERB NOUN	157745	11.63
DET ADV	6501	43.52	ADV ADP	31008	27.90	PRON VERB	164872	11.19
NUM VERB	6834	43.02	PRON ADJ	31259	27.82	ADP DET	193443	9.59
PRT ADV	6841	43.01	CONJ ADJ	32453	27.45	NOUN ADJ	255720	6.80
. X	6844	43.01	PRT NOUN	32560	27.41	NOUN VERB	260172	6.63
ADP ADP	6954	42.85	ADP NUM	32903	27.31	ADJ NOUN	293295	5.43
⟨BOS⟩ NUM	7079	42.67	PRON PRON	33339	27.18	. ⟨EOS⟩	367592	3.17
DET VERB	7390	42.24	⟨BOS⟩ ADP	33470	27.14	NOUN NOUN	409828	2.09
DET .	7413	42.21	⟨BOS⟩ PRON	33486	27.13	NOUN ADP	427409	1.67
PRT NUM	7526	42.06	PRON .	33567	27.11	DET NOUN	454980	1.04
X NOUN	7870	41.61	. ADJ	35337	26.59	ADP NOUN	470575	0.70
ADJ ADV	7932	41.53	⟨BOS⟩ ADV	36636	26.23	NOUN .	504897	0.00

Table 1: Counts and costs for universal tag bigrams based on treebanks for 11 languages not used in experiments. The cost used for unseen bigrams is the maximum of all costs in the table.