EMEA	AUC	MacF
ALLFEATURES	79.60	71.64
–Token:L-PSDBin	77.09	70.50
–Type:RelFreq	78.43	72.19
–Token:G-PSD	**79.66**	72.11
–Type:Context	**79.66**	72.45
–Token:Ctx%	78.91	**73.37**
–Type:TopicSim	78.05	71.33
–Token:CtxCnt	76.90	71.72
–Token:L-PSD	76.03	73.35
–Type:NgramProb	73.32	69.54
–Token:G-PSDBin	74.41	69.76
–Token:NgramProb	69.78	68.89
–Token:PSDRatio	48.38	3.45

Science	AUC	MacF
ALLFEATURES	73.91	47.54
–Token:L-PSDBin	76.26	53.69
–Token:G-PSD	77.04	53.56
–Token:G-PSDBin	77.44	54.54
–Token:L-PSD	77.85	56.05
–Token:PSDRatio	77.92	**57.34**
–Token:CtxCnt	77.85	54.42
–Type:Context	**78.17**	55.45
–Token:Ctx%	78.06	55.04
–Type:TopicSim	77.83	54.57
–Token:NgramProb	76.98	51.02
–Type:RelFreq	74.25	49.57
–Type:NgramProb	50.00	0.00

Subs	AUC	MacF
ALLFEATURES	69.26	52.78
–Type:NgramProb	69.13	53.33
–Token:G-PSDBin	70.23	54.72
–Token:CtxCnt	71.23	58.35
–Token:L-PSDBin	72.07	57.85
–Token:G-PSD	72.17	57.33
–Type:TopicSim	**72.31**	58.41
–Token:Ctx%	72.17	56.17
–Token:NgramProb	71.35	**59.26**
–Token:PSDRatio	70.33	46.88
–Token:L-PSD	69.05	53.31
–Type:RelFreq	65.25	48.22
–Type:Context	50.00	0.00

Table 5: Feature ablation results for all three corpora. Selection criteria is AUC, but Macro-F is presented for completeness. Feature selection is run independently on each of the three datasets. The features toward the *bottom* were the first selected.

	AUC	Macro-F	Micro-F
EMEA			
TYPEONLY	71.43 ± 0.94	52.62 ± 3.41	38.67 ± 1.35
TOKENONLY	**73.75** ± 1.11	**67.77** ± 4.18	45.49 ± 3.96
ALLFEATURES	72.19 ± 4.07	67.26 ± 7.88	**49.29** ± 3.55
XV-ALLFEATURES	79.60 ± 1.20	71.64 ± 1.83	46.83 ± 0.62
Science			
TYPEONLY	**75.19** ± 0.89	**51.53** ± 2.55	37.14 ± 4.41
TOKENONLY	71.24 ± 1.45	47.27 ± 1.11	40.48 ± 1.84
ALLFEATURES	74.14 ± 0.93	48.86 ± 3.94	**43.20** ± 3.16
XV-ALLFEATURES	73.91 ± 0.66	47.54 ± 1.52	40.22 ± 1.03
Subs			
TYPEONLY	60.90 ± 1.47	39.21 ± 14.78	24.77 ± 2.78
TOKENONLY	**62.00** ± 1.16	49.74 ± 6.30	**42.95** ± 3.92
ALLFEATURES	60.12 ± 2.11	**50.16** ± 8.63	38.56 ± 5.20
XV-ALLFEATURES	69.26 ± 0.60	52.78 ± 1.96	43.85 ± 0.90

Table 6: Cross-domain test results on the SENSESPOTTING task. Two standard deviations are shown in small type. Only AUC, Macro-F and Micro-F are shown for brevity.

AUC as the measure on which to ablate. It's quite clear that for Science, all the useful information is in the type-level features, a result that echoes what we saw in the previous section. While for EMEA and Subs, both type- and token-level features play a significant role. Considering the six most useful features in each domain, the ones that pop out as frequently most useful are the global PSD features, the ngram probability features (either type- or token-based), the relative frequency features and the context features.

6.5 Cross-Domain Training

One disadvantage to the previous method for evaluating the SENSESPOTTING task is that it requires parallel data in a new domain. Suppose we have *no* parallel data in the new domain at all, yet still want to attack the SENSESPOTTING task. One option is to train a system on domains for which we *do* have parallel data, and then apply it in a new domain. This is precisely the setting we explore in this section. Now, instead of performing cross-validation in a single domain (for instance, Science), we take the union of *all* of the training data in the other domains (e.g., EMEA and Subs), train a classifier, and then apply it to Science. This classifier will almost certainly be worse than one trained on NEW (Science) but does not require *any* parallel data in that domain. (Hyperparameters are chosen by development data from the OLD union.)

The results of this experiment are shown in Table 6. We include results for TOKENONLY, TYPEONLY and ALLFEATURES; all of these are trained in the cross-domain setting. To ease comparison to the results that do not suffer from domain shift, we also present "XV-ALLFEATURES", which are results copied from Table 4 in which parallel data from NEW is used. Overall, there is a drop of about 7.3% absolute in AUC, moving from XV-ALLFEATURES to ALLFEATURES, including a small improvement in Science (likely because Science is markedly smaller than Subs, and "more difficult" than EMEA with many word types).

6.6 Detecting Most Frequent Sense Changes

We define a second, related task: MOSTFREQSENSECHANGE. In this task, instead of predicting if a given word token has a sense which is brand new with respect to the old domain, we predict whether it is being used with a a sense which is not the one that was observed *most frequently* in the old domain. In our EMEA, Science, and Subtitles data, 68.2%, 48.3%, and 69.6% of word tokens' predominant sense changes.