Data release note
v.1
October 10, 2016

Beata Beigman Klebanov and **Chee Wee Leong**
Educational Testing Service
{bbeigmanklebanov,cleong,mflor}@ets.org

Abstract

This note presents the data sizes and performance figures obtained by evaluating the metaphor detection system presented in Beigman Klebanov et al. (2015) (henceforth, **original paper**) in a cross-validation scenario (as reported in the original paper) and train-and-test scenario (not reported in the original paper) on the VUA corpus, after additional clean-up of the data. Benchmarking against the feature set UPT+CupDown+DCUpDown from the original paper should be done using the up-to-date figures in this release note. Please refer to the original paper for further details about the features.

1 Data Sizes

Table 1 is copied from the original paper, for the VUA corpus. It reports data sizes in the training data for the VUA corpus; only the training data was used in the paper, in a cross-validation setting.

Data	#Texts	#Instances	%Metaphor
News	49	18,519	18%
Fict.	11	17,836	14%
Acad.	12	29,469	13%
Conv.	18	15,667	7%

Table 1: Summary of the data.

Table 2 reports the data sizes in training and testing partitions in each of the genres in the VUA corpus, as per the current data release, following a substantial data cleaning in view of the release and use of this data for a shared task. Values that changed from the original are shown in blue; since evaluations on the test data were not reported in the original paper, these data sizes are reported here for the first time.

Data	Training			Testing	
	#T	#I	%M	#T	#I
News	49	17,056	20%	14	6,008
Fict.	11	15,892	16%	3	4,810
Acad.	12	27,669	14%	4	6,076
Conv.	18	11,994	10%	6	5,302

Table 2: Summary of the data, corrected and expanded. #T = # of texts; #I = # of instances; %M = percentage of metaphors.

2 Cross-Validation Performance

Table 3 is copied from Table 9 of the original paper for VUA data (Beigman Klebanov et al., 2015). Table 4 shows recomputed figures for the same evaluations as table 3, using the training data in the current release. Cells where performance is different by more than 0.02 are shown in blue. We note that the average F1-scores (rows named Av.) are within 0.01 of the originally reported, for all training regimes.

3 Performance on Test Data

This section provides benchmarks on test data for the UPT+CUpDown+DCUpDown model from Beigman Klebanov et al. (2015), in auto-weighting and optimized-weighting training regimes that were found to consistently outperform the training regime without re-weighting of examples, using Logistic Regression classifier. Please refer to Beigman Klebanov et al. (2015) for details about the weighting regimes.

3.1 Training on genre-specific data only

Table 5 shows the performance of the UPT+CUpDown+DCUpDown model in auto-weighting training regime (top) and optimized weighting regime (bottom). The optimized class weights for the four genres are: 1:4 (Acad.,

Data	UPT			UPT+ CUpDown+ DCUpDown		
	P	R	F	P	R	F
Acad.	.635	.347	.418	.636	.356	.425
Conv.	.506	.240	.316	.487	.236	.309
Fict.	.549	.288	.374	.559	.309	.395
News	.641	.457	.531	.636	.466	.536
Av.	.583	.333	.410	.580	.342	.416

Data	UPT			UPT+ CUpDown+ DCUpDown		
	P	R	F	P	R	F
Acad.	.524	.651	.558	.525	.657	.562
Conv.	.292	.688	.392	.293	.691	.396
Fict.	.400	.600	.476	.411	.607	.486
News	.529	.665	.587	.530	.673	.590
Av.	.436	.651	.503	.440	.657	.509

Data	UPT			UPT+ CUpDown+ DCUpDown		
	P	R	F	P	R	F
Acad.	.521	.671	.565	.531	.655	.564
Conv.	.321	.614	.404	.293	.691	.396
Fict.	.398	.620	.481	.414	.621	.493
News	.506	.711	.586	.513	.709	.590
Av.	.437	.654	.509	.438	.669	.511

Table 3: Performance of a model without any concreteness features (UPT) and the model UPT+CUpDown+DCUpDown, in no-reweighting regime (top), auto-weighting (middle), and optimal weighting (bottom), using cross-validation on training data.

Data	UPT			UPT+ CUpDown+ DCUpDown		
	P	R	F	P	R	F
Acad.	.651	.329	.406	.644	.357	.426
Conv.	.474	.232	.305	.445	.241	.308
Fict.	.567	.272	.364	.559	.325	.408
News	.642	.467	.538	.639	.480	.546
Av.	.584	.325	.403	.572	.351	.422

Data	UPT			UPT+ CUpDown+ DCUpDown		
	P	R	F	P	R	F
Acad.	.486	.707	.556	.516	.692	.571
Conv.	.310	.679	.403	.312	.708	.415
Fict.	.402	.620	.482	.417	.592	.485
News	.505	.698	.582	.525	.672	.586
Av.	.426	.676	.506	.443	.666	.514

Data	UPT			UPT+ CUpDown+ DCUpDown		
	P	R	F	P	R	F
Acad.	.527	.646	.561	.529	.678	.572
Conv.	.321	.617	.403	.327	.657	.418
Fict.	.409	.597	.480	.407	.625	.490
News	.510	.698	.585	.520	.701	.594
Av.	.442	.640	.507	.446	.665	.519

Table 4: Performance of a model without any concreteness features (UPT) and the model UPT+CUpDown+DCUpDown, in no-reweighting regime (top), auto-weighting (middle), and optimal weighting (bottom), using cross-validation on training data. Updated using the datasets in the current data release. Cells where performance is different by more than 0.02 are shown in blue.

Conv.), 1:4.33 (Fict.), 1:3.67 (News).[1]

3.2 Training on all data

Table 6 shows the performance of the UPT+CUpDown+DCUpDown model in auto-weighting training regime (top) and optimized weighting regime (bottom), when the system is trained using training data for all the four genres, and evaluated on test data for each genre separately. Training on data across genres was found to be effective for verbs (Beigman Klebanov et al., 2016). The optimized class weights are: 1:3.67 for all four genres.

[1]These are implemented as class_weight : { "0":1, "1":4.33 }, etc., in the SKLL configuration file.

References

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, Colorado, June. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106, Berlin, Germany, August. Association for Computational Linguistics.

Data	P	R	F
Acad.	.718	.669	.693
Conv.	.327	.746	.455
Fict.	.361	.591	.448
News	.548	.721	.623
Av.	.489	.682	.555

Data	P	R	F
Acad.	.715	.666	.690
Conv.	.385	.625	.477
Fict.	.368	.620	.462
News	.559	.700	.622
Av.	.507	.653	.563

Table 5: Performance of the model UPT+CUpDown+DCUpDown trained on genre-specific data, in auto-weighting regime (top), and optimized weighting regime (bottom), on test data.

Data	P	R	F
Acad.	.697	.718	.708
Conv.	.302	.798	.438
Fict.	.327	.669	.440
News	.575	.643	.607
Av.	.475	.707	.548

Data	P	R	F
Acad.	.723	.667	.694
Conv.	.314	.743	.442
Fict.	.344	.642	.448
News	.612	.618	.615
Av.	.498	.668	.550

Table 6: Performance of the model UPT+CUpDown+DCUpDown trained on data across all genres, in auto-weighting regime (top), and optimized weighting regime (bottom), on test data.