| Method | Our Approach | | | | SMS | |
| Input Embedding Type | Dense | | Sparse | | Sparse | |
Vocabulary	C	U	C	U	C	U
SensEval2						
BERT Base	67.17	65.95	70.90	69.85	67.79	65.60
BERT Large	66.69	66.03	68.71	66.82	63.49	60.07
SenseBERT Base	–	72.21	–	78.26	–	77.16
SenseBERT Large	–	68.93	–	76.95	–	76.42
DistilBERT Base	66.30	68.40	72.69	73.04	69.32	70.24
RoBERTa Base	62.66	–	67.48	–	63.23	–
RoBERTa Large	67.96	–	72.34	–	68.53	–
XLM-RoBERTa Base	66.30	–	71.16	–	69.69	–
XLM-RoBERTa Large	65.81	–	69.98	–	65.24	–

Table 1: Accuracy of each model on the supersense prediction task using dense and sparse embedding spaces. All of the sparse representations were generated using $\lambda = 0.1$ for the regularization coefficient and $k = 3000$ basis based on the experiments reported. C and U stands for the utilized vocabulary being Cased and Uncased, respectively.

| Method | Our Approach | | | | SMS | |
| Input Embedding Type | Dense | | Sparse | | Sparse | |
Vocabulary	C	U	C	U	C	U
SensEval3						
BERT Base	64.70	62.05	70.32	69.51	64.81	62.05
BERT Large	64.21	62.16	70.21	66.32	63.49	56.81
SenseBERT Base	–	63.56	–	74.75	–	71.45
SenseBERT Large	–	62.91	–	74.91	–	70.86
DistilBERT Base	62.16	64.10	72.59	74.27	66.70	67.45
RoBERTa Base	59.02	–	66.16	–	61.40	–
RoBERTa Large	64.00	–	70.00	–	65.02	–
XLM-RoBERTa Base	62.27	–	70.59	–	67.29	–
XLM-RoBERTa Large	61.67	–	67.67	–	64.75	–

Table 2: Accuracy of each model on the supersense prediction task using dense and sparse embedding spaces. All of the sparse representations were generated using $\lambda = 0.1$ for the regularization coefficient and $k = 3000$ basis based on the experiments reported. C and U stands for the utilized vocabulary being Cased and Uncased, respectively.

Method Input Embedding Type	Vocabulary	Our Approach				SMS	
		Dense		Sparse		Sparse	
		C	U	C	U	C	U
		SemEval2007					
BERT	Base	59.78	58.24	65.93	64.61	62.41	61.53
	Large	58.46	57.14	62.85	61.09	58.24	56.70
SenseBERT	Base	–	61.09	–	71.42	–	67.25
	Large	–	61.75	–	71.42	–	67.25
DistilBERT	Base	58.24	55.38	65.93	68.79	61.31	62.63
RoBERTa	Base	52.96	–	58.90	–	53.18	–
	Large	59.12	–	62.41	–	57.58	–
XLM-RoBERTa	Base	58.68	–	62.19	–	61.31	–
	Large	58.02	–	59.78	–	57.14	–

Table 3: Accuracy of each model on the supersense prediction task using dense and sparse embedding spaces. All of the sparse representations were generated using $\lambda = 0.1$ for the regularization coefficient and $k = 3000$ basis based on the experiments reported. C and U stands for the utilized vocabulary being Cased and Uncased, respectively.

Method Input Embedding Type	Vocabulary	Our Approach				SMS	
		Dense		Sparse		Sparse	
		C	U	C	U	C	U
		SemEval2013					
BERT	Base	63.97	59.73	67.45	67.57	64.05	62.83
	Large	60.27	60.94	67.21	60.58	59.42	53.83
SenseBERT	Base	–	63.38	–	69.46	–	72.68
	Large	–	62.83	–	70.86	–	72.56
DistilBERT	Base	60.94	59.91	67.21	69.03	63.07	65.57
RoBERTa	Base	56.56	–	62.04	–	63.30	–
	Large	61.31	–	67.45	–	63.64	–
XLM-RoBERTa	Base	60.03	–	69.16	–	61.31	–
	Large	58.69	–	66.78	–	63.80	–

Table 4: Accuracy of each model on the supersense prediction task using dense and sparse embedding spaces. All of the sparse representations were generated using $\lambda = 0.1$ for the regularization coefficient and $k = 3000$ basis based on the experiments reported. C and U stands for the utilized vocabulary being Cased and Uncased, respectively.

| Method | | Our Approach | | | | SMS | |
| Input Embedding Type | | Dense | | Sparse | | Sparse | |
Vocabulary		C	U	C	U	C	U
SemEval2015							
BERT	Base	62.72	59.98	68.39	64.67	62.23	59.17
	Large	61.44	58.86	66.43	64.48	59.39	55.77
SenseBERT	Base	–	61.64	–	74.36	–	75.04
	Large	–	60.95	–	74.46	–	74.95
DistilBERT	Base	60.07	63.50	68.98	74.85	64.09	68.10
RoBERTa	Base	57.82	–	64.77	–	63.30	–
	Large	62.32	–	70.64	–	65.08	–
XLM-RoBERTa	Base	63.79	–	68.39	–	65.94	–
	Large	60.07	–	64.72	–	64.38	–

Table 5: Accuracy of each model on the supersense prediction task using dense and sparse embedding spaces. All of the sparse representations were generated using $\lambda = 0.1$ for the regularization coefficient and $k = 3000$ basis based on the experiments reported. C and U stands for the utilized vocabulary being Cased and Uncased, respectively.