Smaller Text Classifiers with Discriminative Cluster Embeddings

Mingda Chen Kevin Gimpel
Toyota Technological Institute at Chicago

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

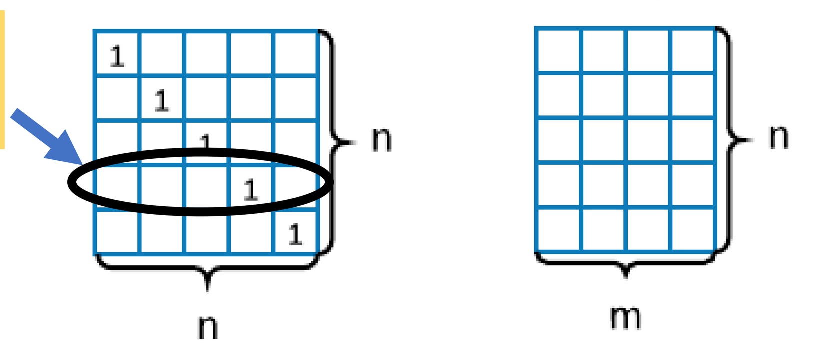
Motivation

❖ Word embedding parameters often **dominate** overall model sizes in neural methods for natural language processing. Can we reduce embedding parameters for text classification tasks?

Discriminative Cluster Embeddings

❖ **Standard Embedding (SE)**: Each *word* has its own word embedding vector.

One-hot vector represents embedding membership



n: Vocabulary size
m: Embedding dimension

❖ **Cluster Embedding (CE)**:

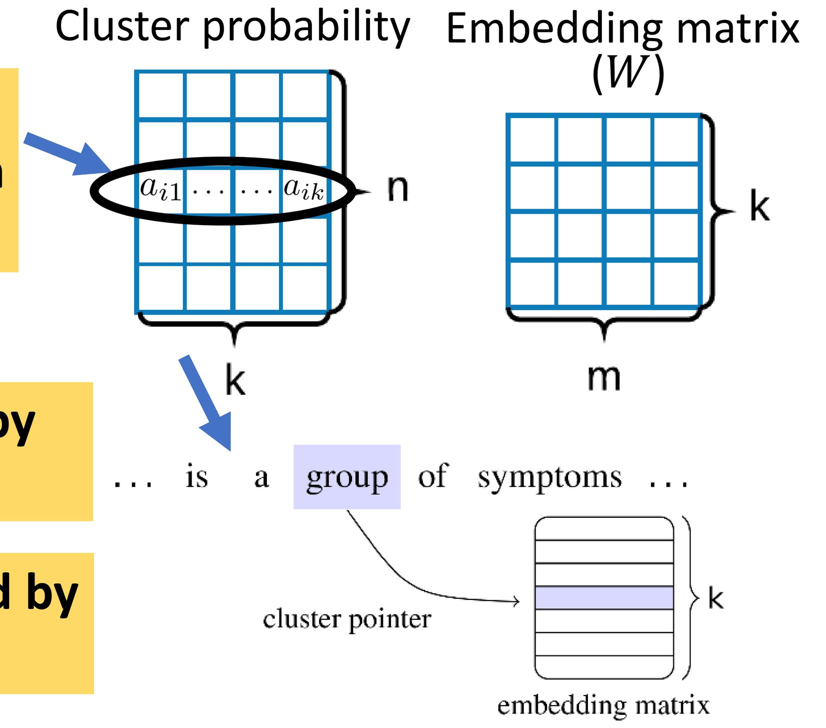
- ❑ Each *cluster* has its own embedding vector.
- ❑ Each word i has a cluster probability vector a_i.
- ❑ End-to-end training with classification loss.
- ❑ Cluster membership h_i is treated as latent variable.
- ❑ Difficulty: non-differentiable argmax $h_i = \arg\max_{1 \le j \le k} a_{ij}$

k: Number of clusters

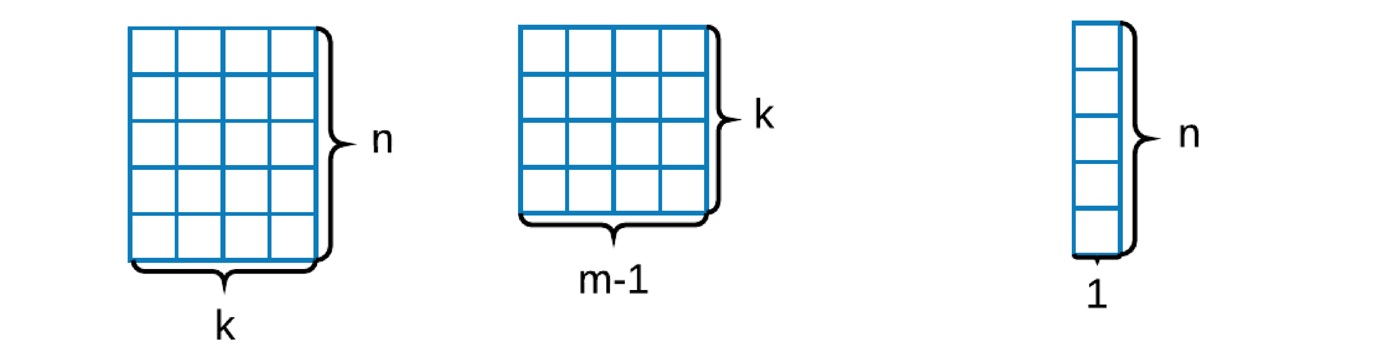
During training, use samples t_i from Gumbel-Softmax(a_i) as an approximation to one_hot(h_i)

Test time, h_i is determined by t_i = one_hot(argmax(a_i))

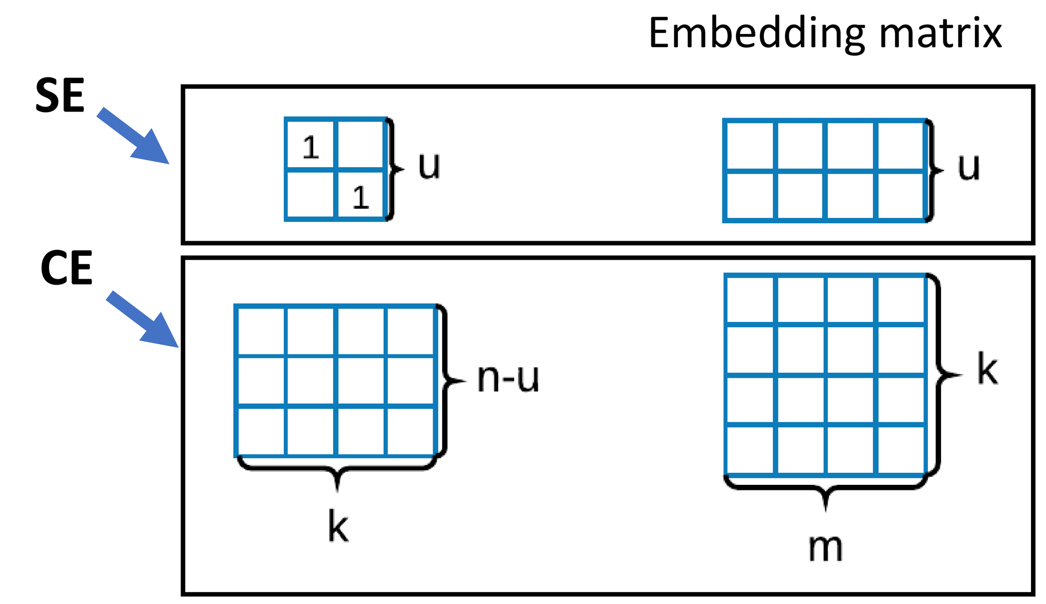
Embedding vector is computed by $e_i = t_i \cdot W$



❖ **Cluster Adjusted Embeddings (CAE)**: Each word has an additional 1-dimensional vector that gets concatenated to the cluster embedding.



❖ **Mixture Embedding (ME)**: Select the most frequent u words to use unique embedding vectors and the remaining words use cluster embedding.

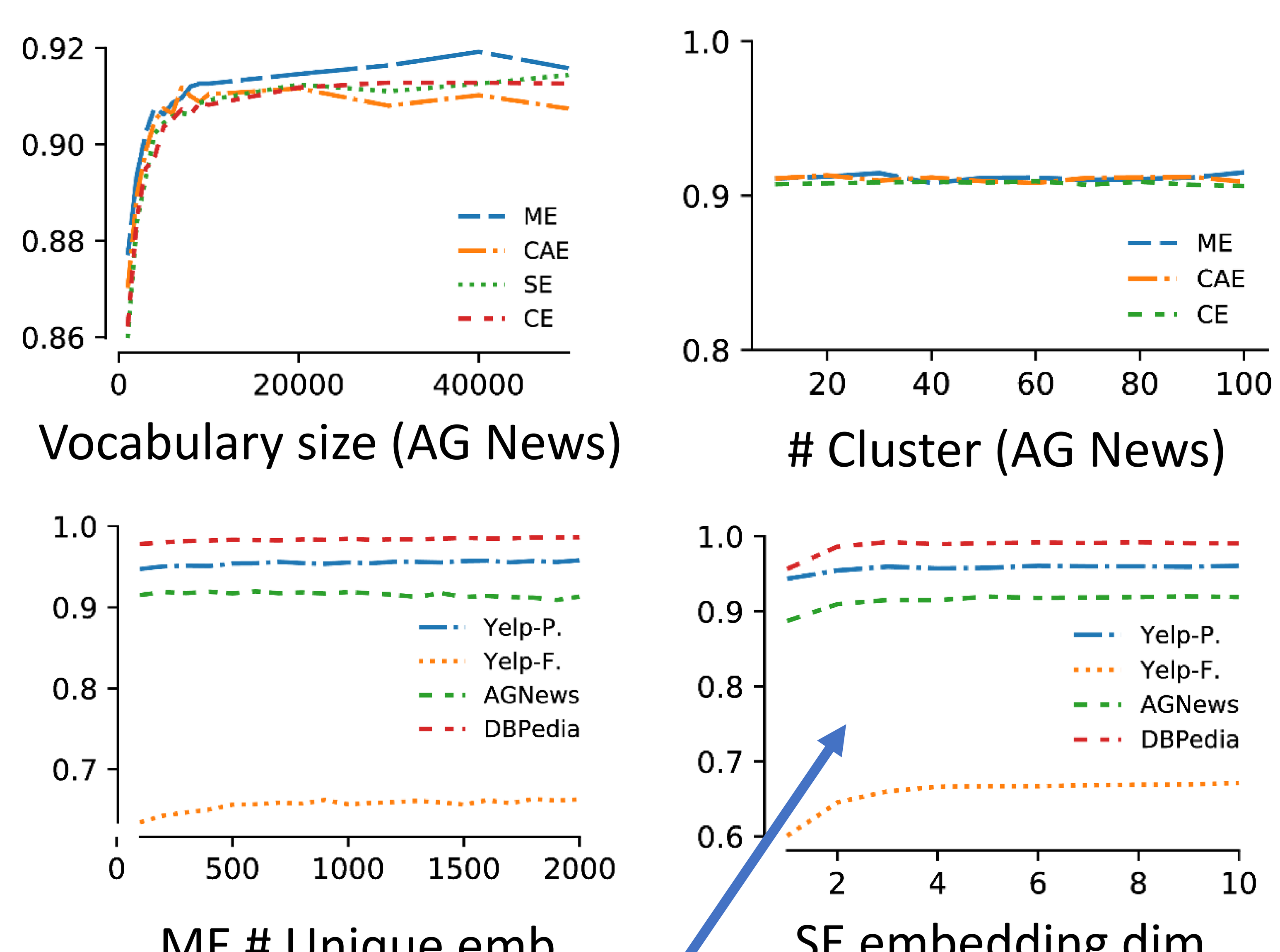


Learned Word Clusters

Data labels: World Business Sci/Tech Sports

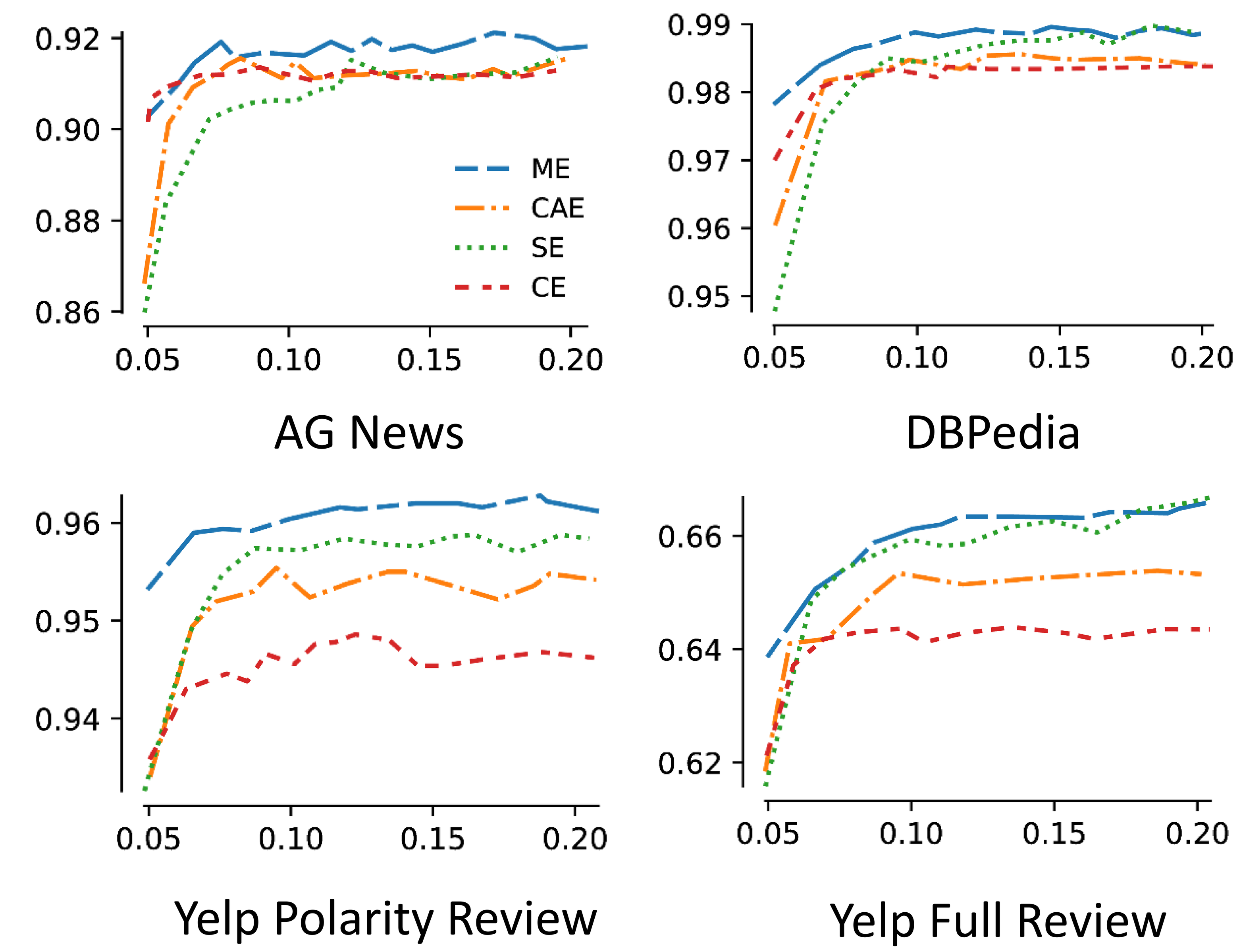
Word clusters learned using CE model on AG News
million week third percent which 000 ago reports once
are after from has another down home than but end
official security china international country court city
Heavyweights operational coordinated healing rewarded
com internet technology ibm google research windows
market quarter sales deals bid growth trade economic
Championship yankees defense player contract football
Troops press attack forces peace iran led army killing

Effect of Hyperparameters



Largest accuracy gains occur when increasing dimensionality from 1 to 2.

Experiments



Development accuracy vs model size (MB) on four datasets.

	AG News		DBPedia		Yelp Full		Yelp Polarity	
Size	0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1
SE	84.8	90.4	95.3	98.1	59.2	62.6	93.4	95.5
CE	89.2	90.7	96.9	97.9	60.3	61.0	93.9	94.4
CAE	86.3	90.7	96.1	98.1	61.2	62.3	93.7	95.3
ME	**90.3**	**91.5**	**97.5**	**98.3**	**61.4**	**63.4**	**95.2**	**95.8**

Test results (%). Model sizes are in MB.

	Embedding size (MB)	Model size (MB)	Test acc (%)
Glove Baseline	85.947	-	87.18
Compositional coding*	2.305	-	88.15
Re-implemented compositional coding	0.245	0.353	83.43
Standard Embedding	0.092	0.137	86.84
Cluster Embedding	**0.004**	0.046	85.58
Cluster Adjusted Embedding	0.016	0.058	86.94
Mixture Embedding	0.009	0.051	**88.22**

IMDB test results.

* Raphael Shu and Hideki Nakayama. Compressing word embeddings via deep compositional code learning. ICLR 2018