Paper:

Automated essay scoring with string kernels and word embeddings

Mădălina Cozma, Andrei M. Butnaru, Radu Tudor Ionescu
Department of Computer Science, University of Bucharest, Romania

Paper:

Highlights

- We propose to combine string kernels (low-level character n-gram features) and word embeddings (high-level semantic features) for automated essay scoring (AED)
- TOK, string kernels have never been used for AED
- TOK, this is the first successful attempt to combine string kernels and word embeddings
- Using a shallow approach, we surpass recent deep learning approaches [Dong et al, EMNLP 2016; Dong et al, CONLL 2017;Tay et al, AAAI 2018]

String kernels

- We use the histogram intersection string kernel (HISK), which is formally defined as follows:

$$k^{\cap}(x, y) = \sum_{v \in \Sigma^n} \min\{\text{num}_v(x), \text{num}_v(y)\},$$

where $x, y \in \Sigma^*$ are two strings over an alphabet Σ, $\text{num}_v(x)$ is the number of occurrences of n-gram v as a substring in x, and n is the length of v.

- We then normalize the kernel as follows:

$$\hat{k}^{\cap}(x, y) = \frac{k^{\cap}(x, y)}{\sqrt{k^{\cap}(x, x) \cdot k^{\cap}(y, y)}}$$

Bag-of-Super-Word-Embeddings

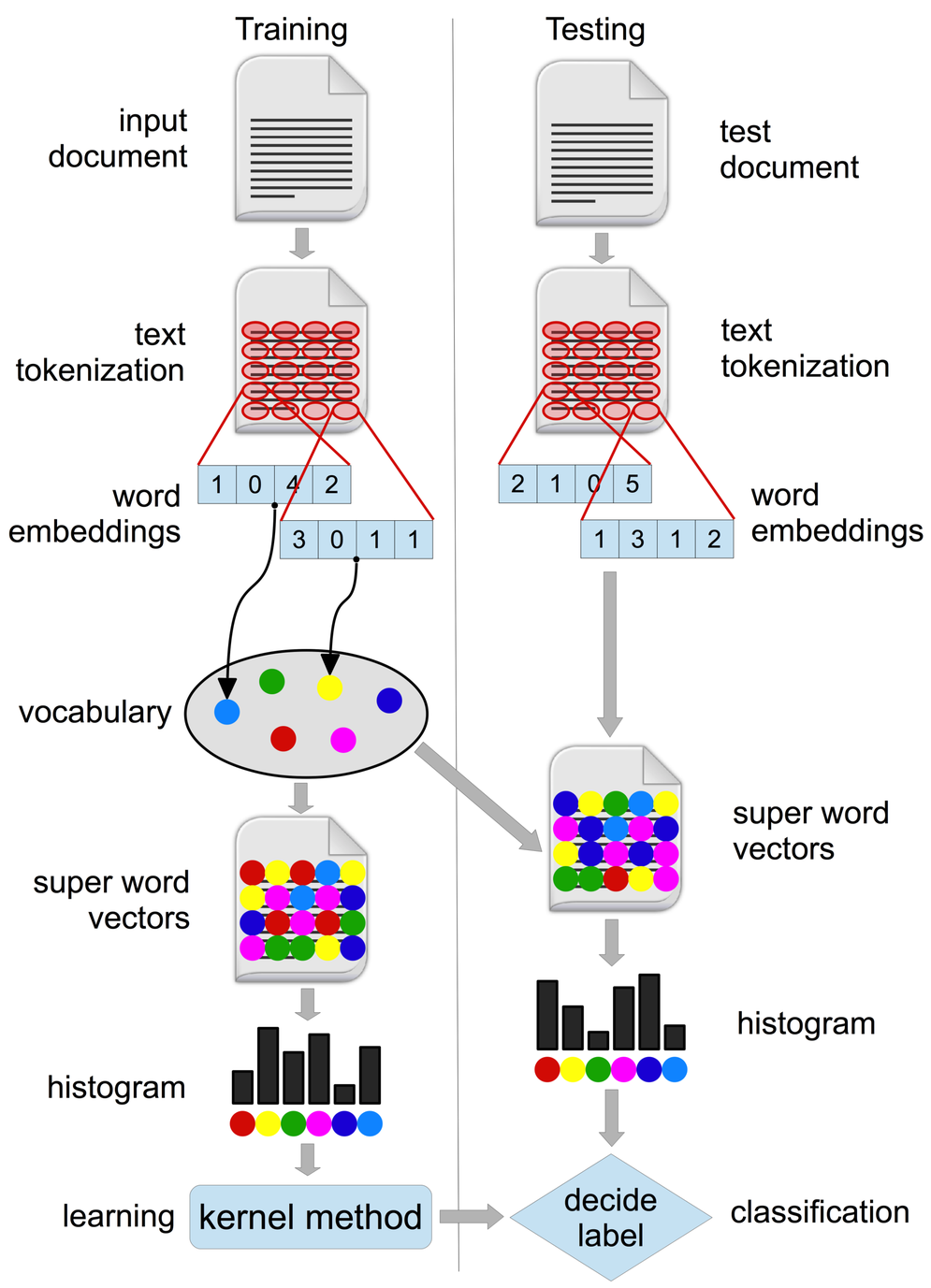

Figure: Each word in the collection of documents is represented as word vector using a pre-trained word embeddings model. The word vectors are then clustered in order to obtain relevant semantic clusters of words. As in the standard bag-of-visual-words model, the clustering is done by k-means. Every embedded word in the collection of documents is then assigned to the nearest cluster centroid (the nearest super word vector). Put together, the super word vectors generate a vocabulary (codebook) that can further be used to represent each document as a *bag-of-super-word-embeddings* (BOSWE). After building the representation, we employ a kernel method to train the model for our task.

Fusion and learning method

- We combine HISK and BOSWE in the dual (kernel) form, by simply summing up the two corresponding kernel matrices
- Note: summing up kernel matrices is equivalent to feature vector concatenation in the primal Hilbert space
- We employ ν-Support Vector Regression (ν-SVR) in order to automatically predict the score for an essay

Data set

Prompt	Number of Essays	Score Range
1	1783	2-12
2	1800	1-6
3	1726	0-3
4	1726	0-3
5	1772	0-4
6	1805	0-4
7	1569	0-30
8	723	0-60

Table: The number of essays and the score ranges for the 8 different prompts in the Automated Student Assessment Prize (ASAP) data set.

In-domain results

Method	1	2	3	4	5	6	7	8	Overall
Human	**0.721**	**0.814**	**0.769**	**0.851**	**0.753**	**0.776**	**0.721**	**0.629**	**0.754**
[Phandi et al, EMNLP 2015]	**0.761**	**0.606**	**0.621**	**0.742**	**0.784**	**0.775**	**0.730**	**0.617**	**0.705**
[Dong et al, EMNLP 2016]	-	-	-	-	-	-	-	-	**0.734**
[Dong et al, CONLL 2017]	**0.822**	**0.682**	**0.672**	**0.814**	**0.803**	**0.811**	**0.801**	**0.705**	**0.764**
[Tay et al, AAAI 2018]	**0.832**	**0.684**	**0.695**	**0.788**	**0.815**	**0.810**	**0.800**	**0.697**	**0.764**
HISK and ν-SVR	**0.836**	**0.724**	**0.677**	**0.821**	**0.830**	**0.828**	**0.801**	**0.726**	**0.780**
BOSWE and ν-SVR	**0.788**	**0.689**	**0.667**	**0.809**	**0.824**	**0.824**	**0.766**	**0.679**	**0.756**
HISK+BOSWE and ν-SVR	**0.845**	**0.729**	**0.684**	**0.829**	**0.833**	**0.830**	**0.804**	**0.729**	**0.785**

Table: In-domain automatic essay scoring results of our approach versus several state-of-the-art methods. Results are reported in terms of the quadratic weighted kappa (QWK) measure, using 5-fold cross-validation. The best QWK score (among the machine learning systems) for each prompt is highlighted with blue.

Cross-domain results

Source→Target	Method	$n_t = 0$	$n_t = 10$	$n_t = 25$	$n_t = 50$	$n_t = 100$
1→2	[Phandi et al, EMNLP 2015]	**0.434**	**0.463**	**0.457**	**0.492**	**0.510**
	[Dong et al, EMNLP 2016]	-	**0.546**	**0.569**	**0.563**	**0.559**
	HISK and ν-SVR	**0.440**	**0.586**	**0.637**	**0.652**	**0.657**
	BOSWE and ν-SVR	**0.398**	**0.474**	**0.478**	**0.492**	**0.506**
	HISK+BOSWE and ν-SVR	**0.542**	**0.584**	**0.632**	**0.657**	**0.661**
3→4	[Phandi et al, EMNLP 2015]	**0.522**	**0.593**	**0.609**	**0.618**	**0.646**
	[Dong et al, EMNLP 2016]	-	**0.628**	**0.656**	**0.659**	**0.662**
	HISK and ν-SVR	**0.703**	**0.716**	**0.724**	**0.742**	**0.751**
	BOSWE and ν-SVR	**0.615**	**0.640**	**0.716**	**0.728**	**0.727**
	HISK+BOSWE and ν-SVR	**0.701**	**0.713**	**0.737**	**0.754**	**0.779**
5→6	[Phandi et al, EMNLP 2015]	**0.187**	**0.539**	**0.662**	**0.680**	**0.713**
	[Dong et al, EMNLP 2016]	-	**0.647**	**0.700**	**0.714**	**0.750**
	HISK and ν-SVR	**0.715**	**0.726**	**0.754**	**0.757**	**0.781**
	BOSWE and ν-SVR	**0.617**	**0.623**	**0.644**	**0.650**	**0.692**
	HISK+BOSWE and ν-SVR	**0.728**	**0.734**	**0.764**	**0.771**	**0.788**
7→8	[Phandi et al, EMNLP 2015]	**0.171**	**0.586**	**0.607**	**0.613**	**0.621**
	[Dong et al, EMNLP 2016]	-	**0.570**	**0.590**	**0.568**	**0.587**
	HISK and ν-SVR	**0.486**	**0.604**	**0.617**	**0.626**	**0.639**
	BOSWE and ν-SVR	**0.419**	**0.526**	**0.577**	**0.582**	**0.591**
	HISK+BOSWE and ν-SVR	**0.522**	**0.606**	**0.637**	**0.638**	**0.649**

Table: Corss-domain automatic essay scoring results of our approach versus two state-of-the-art methods. Results are reported in terms of the quadratic weighted kappa (QWK) measure, using the same evaluation procedure as [Phandi et al, EMNLP 2015; Dong et al, EMNLP 2016]. The best QWK scores for each source→target domain pair are highlighted with blue.

Conclusion

- The in-domain and the cross-domain comparative results indicate that string kernels, both alone and in combination with word embeddings, attain the best performance on the automatic essay scoring task
- Our shallow approach attains better results than recent deep learning methods [Dong et al, EMNLP 2016; Dong et al, CONLL 2017;Tay et al, AAAI 2018]

butnaruandreimadalin@gmail.com

raducu.ionescu@gmail.com