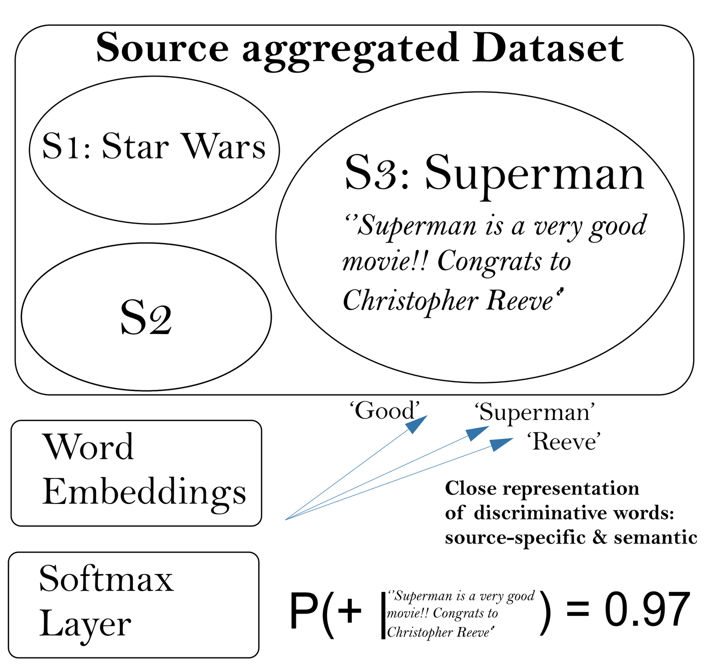
# Adversarial Word Embeddings to Improve Text Classifiers Generalization Power

Victor Bouvier, Céline Hudelot, Philippe Very

Sidetrade and MICS - CentraleSupélec, vbouvier@sidetrade.com

## The multi-sources learning issue

- **Aim**: Increasing the diversity and generality of learned representations.
- **How**: By aggregating various data sources.
- **But**: Learned representations tend to be source-specific rather than multi-source.




### Supervised Continuous Bag of Words [1]

- A strong baseline accuracy-wise.
- Does not discriminate well between source-specific words and words whose meaning is the same regardless of the source

| Word | Similarity |
|---|---|
| 'Obama' | 0.998 |
| 'McConaughey' | 0.997 |
| 'Perfect' | 0.995 |
| 'Shrek' | 0.995 |
| 'Wonderful' | 0.991 |
| 'Good' | 0.887 |

Table 1: Nearest neighbors of the word 'Superman' when training on the K-set of AM.

- **Claim**: Source-specific information deteriorates true generalization power
- **Contribution**: An adversarial approach for building source-independent word embeddings.

## Targeted datasets

- **4 Datasets**: Amazon: Books (AB), Movies and TV (AM), Electronics (AE), Yelp challenge dataset (Yelp).
- **3 targeted datasets** [2]:
  1. Kept set (**K-set**): leakage between the source and the polarity,
  2. Rejected set (**R-set**): inverse leakage (w.r.t K-set) between the source and the polarity.
  3. Unseen set (**U-set**): sources not present in K-set and R-set.
- **Procedure**:
  1. Training on a train set of K-set,
  2. Testing on a test set of K-set,
  3. Evaluating on R-Set and U-set.

| Dataset | Samples | Sources |
|---|---|---|
| AM | 24660 | 137 |
| AB | 43380 | 241 |
| AE | 33300 | 185 |
| Yelp | 28660 | 127 |

Table 2: Statistics on the K-set

| Dataset | R-Set | U-set |
|---|---|---|
| AM | 137 | 787 |
| AB | 241 | 719 |
| AE | 185 | 811 |
| Yelp | 127 | 726 |

Table 3: Number of sources per dataset

## Defining SCBow

- Embeds word sequence $x = (w_1, ..., w_T)$ in $\mathbf{v}_x = \frac{1}{T}\Sigma_{t=1}^{T}\mathbf{v}_{w_t}$ where $\mathbf{v}$ are word embeddings [1],
- $p(y|\mathbf{v}_x)$ is a softmax layer,
- $\mathbf{v}$ and the softmax weights ($\theta_{\text{SCb}}$) are learned with SGD minimizing

$$\mathcal{L}_{\text{SCb}} = \mathbf{E}_{(x,y)}[-\log p(y|\mathbf{v}_x; \theta_{\text{SCb}})]$$

- Best results are obtained with bigrams.

## Inflexion while activating adversarial word embeddings

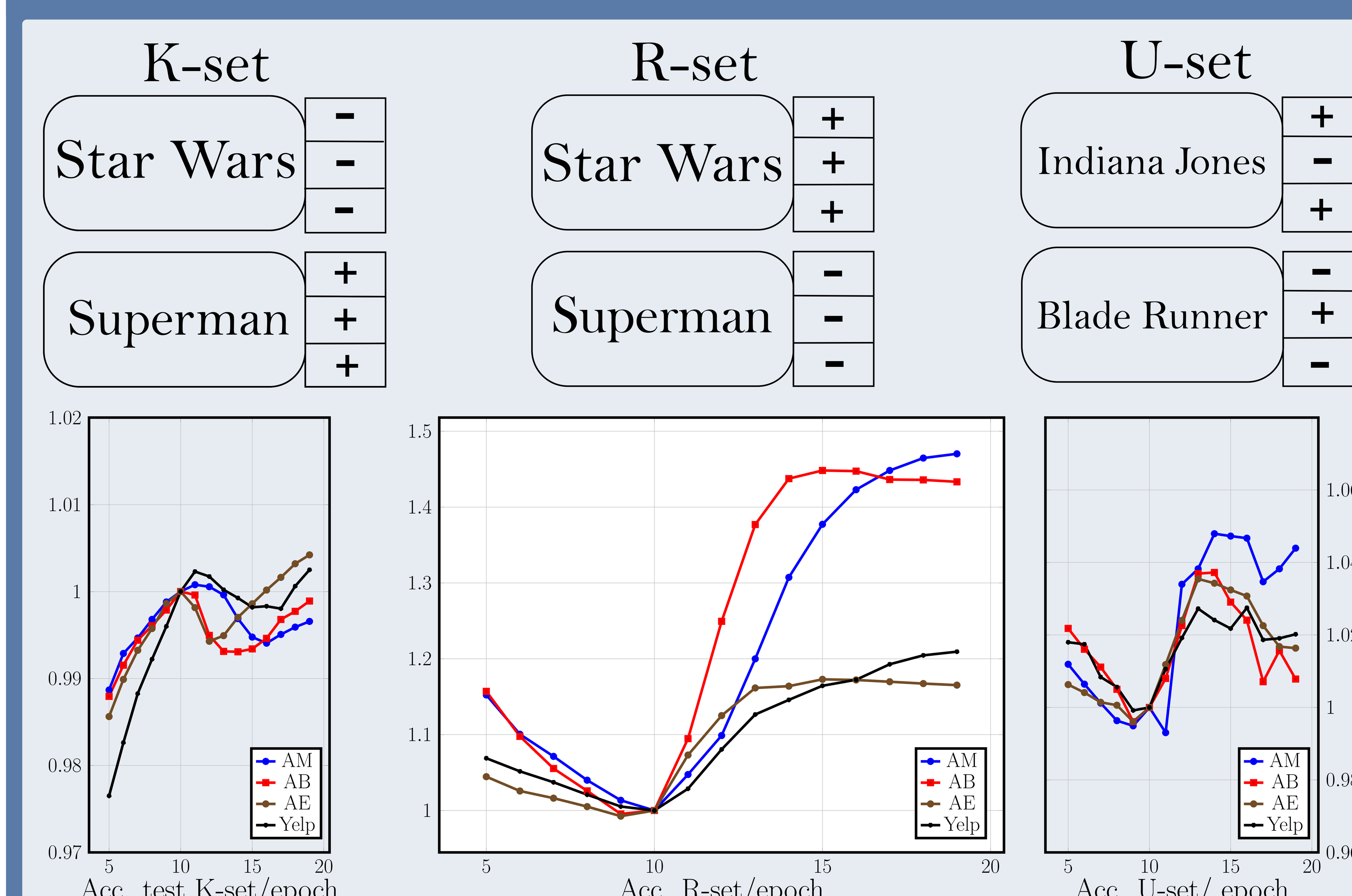



Figure 1: We first train the regularized SCBow with $\lambda = 0$ during 10 epochs then we train with $\lambda = 1$ during 10 epochs. We report the relative accuracy with respect to the performance on the 10th epochs.

$$\mathcal{L}(\mathbf{v}, \theta_{\text{SCb}}) = \mathcal{L}_{\text{SCb}}(\mathbf{v}, \theta_{\text{SCb}}) + \lambda \left| \inf_{\delta\mathbf{v}, \theta_{\text{id}}} \mathcal{L}(\mathbf{v}, \delta\mathbf{v}, \theta_{\text{id}}) - \mathcal{L}_{\text{rand}} \right|$$

## Regularizing SCBow with adversarial embeddings

- $\mathcal{L} = \mathcal{L}_{\text{SCb}} + \lambda\mathcal{L}_{\text{id}}$
- $\mathcal{L}_{\text{id}}$ quantifies the source identifiability of hidden representations $\mathbf{v}_x = (\mathbf{v}_1, ..., \mathbf{v}_T)$

Following the works [3, 4], we suggest an adversarial framework where we learn two embeddings of $x$, $\mathbf{v}_x$ for the classification task and $\tilde{\mathbf{v}}_x$ for the source identification task:

$$\mathcal{L}_{\text{id}} = \left| \inf_{\tilde{\mathbf{v}}, \theta_{\text{id}}} \mathcal{L}(\tilde{\mathbf{v}}, \theta_{\text{id}}) - \mathcal{L}_{\text{rand}} \right|$$

$$\mathcal{L}(\tilde{\mathbf{v}}, \theta_{\text{id}}) = \mathbf{E}_{(x,s)}[-\log p(s|\tilde{\mathbf{v}}_x; \theta_{\text{id}})]$$

$$\tilde{\mathbf{v}} = \delta\mathbf{v} \odot \mathbf{v}$$

1. Coupling $\mathbf{v}$ and $\tilde{\mathbf{v}}$ allows $\tilde{\mathbf{v}}$ to disentangle source hidden information in $\mathbf{v}$.
2. SCb tends to embed discriminative words at the same place, setting $\tilde{\mathbf{v}} = \mathbf{v}$ makes it hard for a neural network $p(\cdot|\tilde{\mathbf{v}}; \theta_{\text{id}})$ to disentangle sources.
3. We build $\tilde{\mathbf{v}}$ as a gated non-linear perturbation of $\mathbf{v}$: $\tilde{\mathbf{v}} = \delta\mathbf{v} \odot \mathbf{v}$


Figure 2: Learning curves on AM with $\lambda = 1$.

## Training details

$\alpha = 0.001$, $\beta = 0.05$

- Pretrain $\mathbf{v}$, $\theta_{\text{SCb}}$ during 10 epochs.
- During 10 epochs do:
  - For a given mini-batch of data $\mathcal{B}$

$$\begin{aligned}\mathbf{v}, \theta_{\text{SCb}} \leftarrow\ & \mathbf{v}, \theta_{\text{SCb}} \\ & - \alpha[\mathbf{E}_{\mathcal{B}}\nabla_{\mathbf{v},\theta_{\text{SCb}}}\mathcal{L}_{\text{SCb}}(\mathbf{v}, \theta_{\text{SCb}}) \\ & - \lambda\mathbf{E}_{\mathcal{B}}\nabla_{\mathbf{v},\theta_{\text{SCb}}}|\mathcal{L}(\mathbf{v}, \delta\mathbf{v}, \theta_{\text{id}}) - \mathcal{L}_{\text{rand}}|]\end{aligned}$$

$$\delta\mathbf{v}, \theta_{\text{id}} \leftarrow \delta\mathbf{v}, \theta_{\text{id}} - \beta\mathbf{E}_{\mathcal{B}}\nabla_{\delta\mathbf{v},\theta_{\text{id}}}\mathcal{L}(\mathbf{v}, \delta\mathbf{v}, \theta_{\text{id}})$$

  - Update $\alpha$ and $\beta$ with Adam.

| $\lambda$ | R-set | U-set |
|---|---|---|
| 0.01 | $< 1.0$ | $< 1.0$ |
| 1.0 | $\times$ 1.33 | $\times$ 1.04 |
| 10.0 | $\times$ 1.64 | $\times$ 1.04 |
| 100.0 | $\times$ 1.23 | $< 1.0$ |

Table 4: Ablation study on $\lambda$ on aggregated performance gain on the 4 datasets

## Remaining challenges

- Making a fast implementation competitive with `fastText` [1].
- Defining a stopping criterion.
- Studying the encoder architecture with respect to source disentanglement in hidden representation.

## References

[1] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov.
Bag of tricks for efficient text classification.
*arXiv preprint arXiv:1607.01759*, 2016.

[2] Victor Bouvier, Céline Hudelot, and Philippe Very.
Proposing datasets to design nlp models robust to distributional shift.
*under review at EMNLP2018*, 2018.

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky.
Domain-adversarial training of neural networks.
*The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[4] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel.
Learning anonymized representations with adversarial neural networks.
*arXiv preprint arXiv:1802.09386*, 2018.

## Acknowledgements

We thank Sidetrade for funding and for providing challenging and insightful datasets. We thank Pablo Piantanida and Jean-Cyril Schutterlé for fruitful discussions.

SIDETRADE

CentraleSupélec