# GILE: A Generalized Input-Label Embedding for Text Classification

Nikolaos Pappas James Henderson
Idiap Research Institute, Martigny, Switzerland

Funded by: SUMMA FNSNF

https://github.com/idiap/gile

## Background

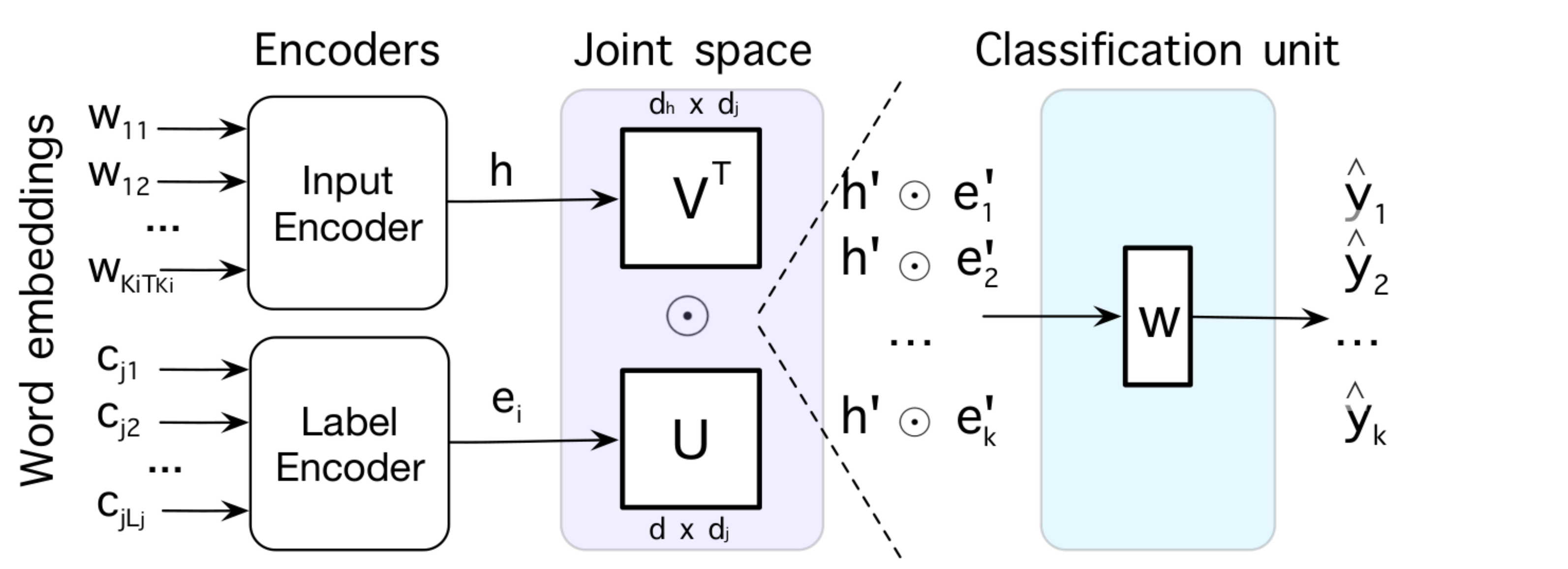
**Problem**: Given $D = \{(x_i, y_i)\}_1^n$ with $y_i \in \{0,1\}^k$ where each input $x_i$ and label $c_j$ are described by word sequences

$$\Phi : X \mapsto \mathcal{Y}_s \cup \mathcal{Y}_u$$

Generalizing well on both seen ($\mathcal{Y}_s$) and unseen ($\mathcal{Y}_u$) labels during training remains a *challenge* because existing models:

- Are tailored for either seen or unseen label prediction
- Have limited expressivity in the output layer

## Proposed Approach: GILE




Given encoded input $h = f_{in}(x_i)$ and encoded label matrix $\mathcal{E} \in \mathbb{R}^{|\mathcal{Y}| \times d}$ with rows $e_j = f_{out}(c_j)$, we have:

$$p(y_i|x_i) \propto \exp\Big(\big(\sigma(\mathcal{E}U + b_u) \odot \sigma(Vh + b_v)\big)w + b\Big)$$

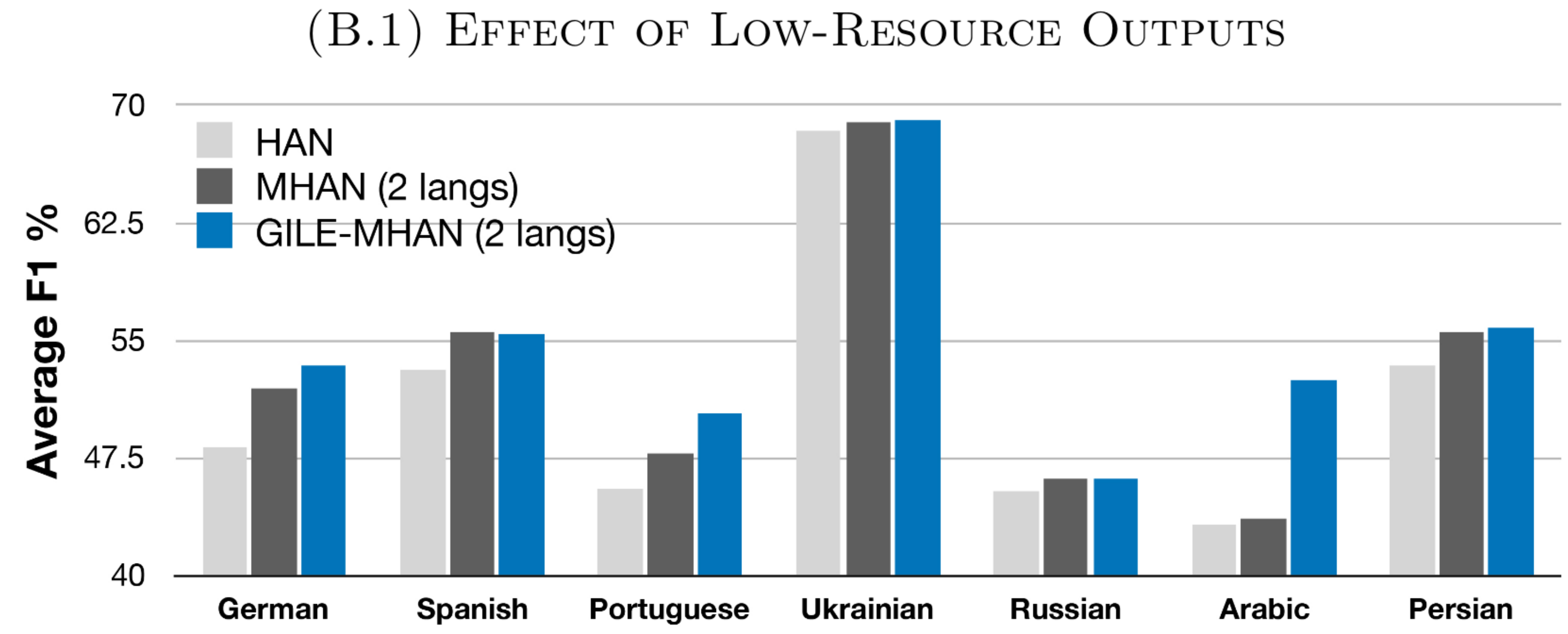
Specifically, based on the joint representation between any input $x_i$ and label $e_j$ and a linear unit $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\hat{y} = p(y_i|x_i) = \frac{1}{1 + e^{-P_{val}^{(i)}}}; \quad P_{val}^{(i)} = \begin{vmatrix} g_{joint}^{(i1)}w + b \\ \dots \\ g_{joint}^{(ik)}w + b \end{vmatrix}$$

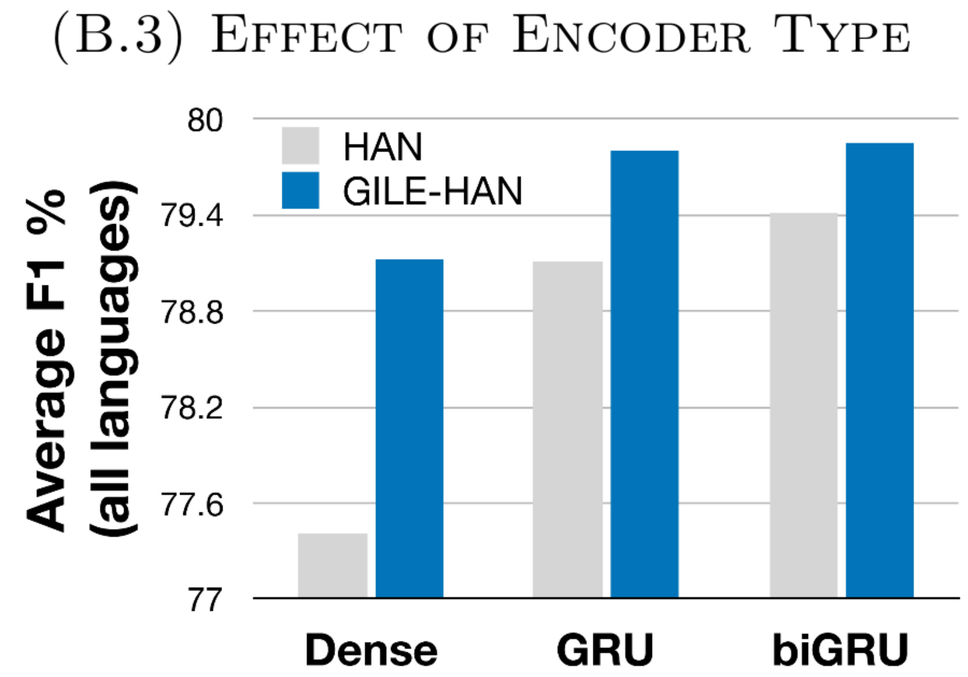
### Novel properties

- Captures nonlinear input and label relationships
- Allows to control the effective capacity of the output layer
- Trained with cross-entropy and is label-set-size independent

### (B.1) Effect of Low-Resource Outputs




### (B.3) Effect of Encoder Type




## References

[YH15] M. Yazdani, J. Henderson. *A Model of Zero-Shot Learning of Spoken Language Understanding*. EMNLP, 2015.
[N16] J. Nam, E. L. Mencía, J. Fürnkranz. *All-In Text: Learning Document, Label, and Word Representations Jointly*, AAAI, 2016.
[Y16] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy. *Hierarchical Attention Networks for Document Classification*, NAACL, 2016.
[P17] N. Pappas, A. Popescu-Belis. *Multilingual Hierarchical Attention Networks for Document Classification*, IJCNLP, 2017.

## Results and Analysis

### (A) Single-task Learning

| Model abbrev. | Layer form Output | Seen labels RL | AvgPr | OneErr | Unseen labels RL | AvgPr | OneErr | Params #count |
|---|---|---|---|---|---|---|---|---|
| AiTextML [N16] | $\mathcal{E}\mathcal{W}h_t$ | 3.54 | 32.78 | 25.99 | 21.62 | 2.66 | 98.61 | 724.4M |
| 1-9 WAN | $W^\top h_t$ | 1.53 | 42.37 | **11.23** | – | – | – | 55.60M |
| BIL-WAN [YH15] | $\sigma(\mathcal{E}\mathcal{W})\mathcal{W}h_t$ | 1.21 | 40.68 | 17.52 | 18.72 | 9.50 | 93.89 | 52.85M |
| BIL-WAN [N16] | $\mathcal{E}\mathcal{W}h_t$ | 1.12 | 41.91 | 16.94 | 16.26 | 10.55 | 93.23 | 52.84M |
| GILE-WAN | $\sigma(\mathcal{E}U)\sigma(Vh_t)$ | **0.78** | **44.39** | 11.60 | **9.06** | **12.95** | **91.90** | 52.93M |
| – constrained $d_j$ | $\sigma(\mathcal{E}\mathcal{W})\sigma(\mathcal{W}h_t)$ | 1.01 | 37.71 | 16.16 | 10.34 | 11.21 | 93.38 | 52.85M |
| – only label | $\sigma(\mathcal{E}\mathcal{W})h_t$ | 1.06 | 40.81 | 13.77 | 9.77 | 14.71 | 90.56 | 52.84M |
| – only input | $\mathcal{E}\sigma(\mathcal{W}h_t)$ | 1.07 | 39.78 | 15.67 | 19.28 | 7.18 | 95.91 | 52.84M |

BioASQ: 10M documents (6.6/0.1/4.9M), 26K labels (23.6 seen/2.4K unseen).

- *AiTextML*: Bilinear input-label embedding trained with a ranking loss.
- *WAN*: Word-level attention network with a sigmoid linear unit.
- *BIL-WAN*: Word-level attention network with a bilinear input-label embedding.

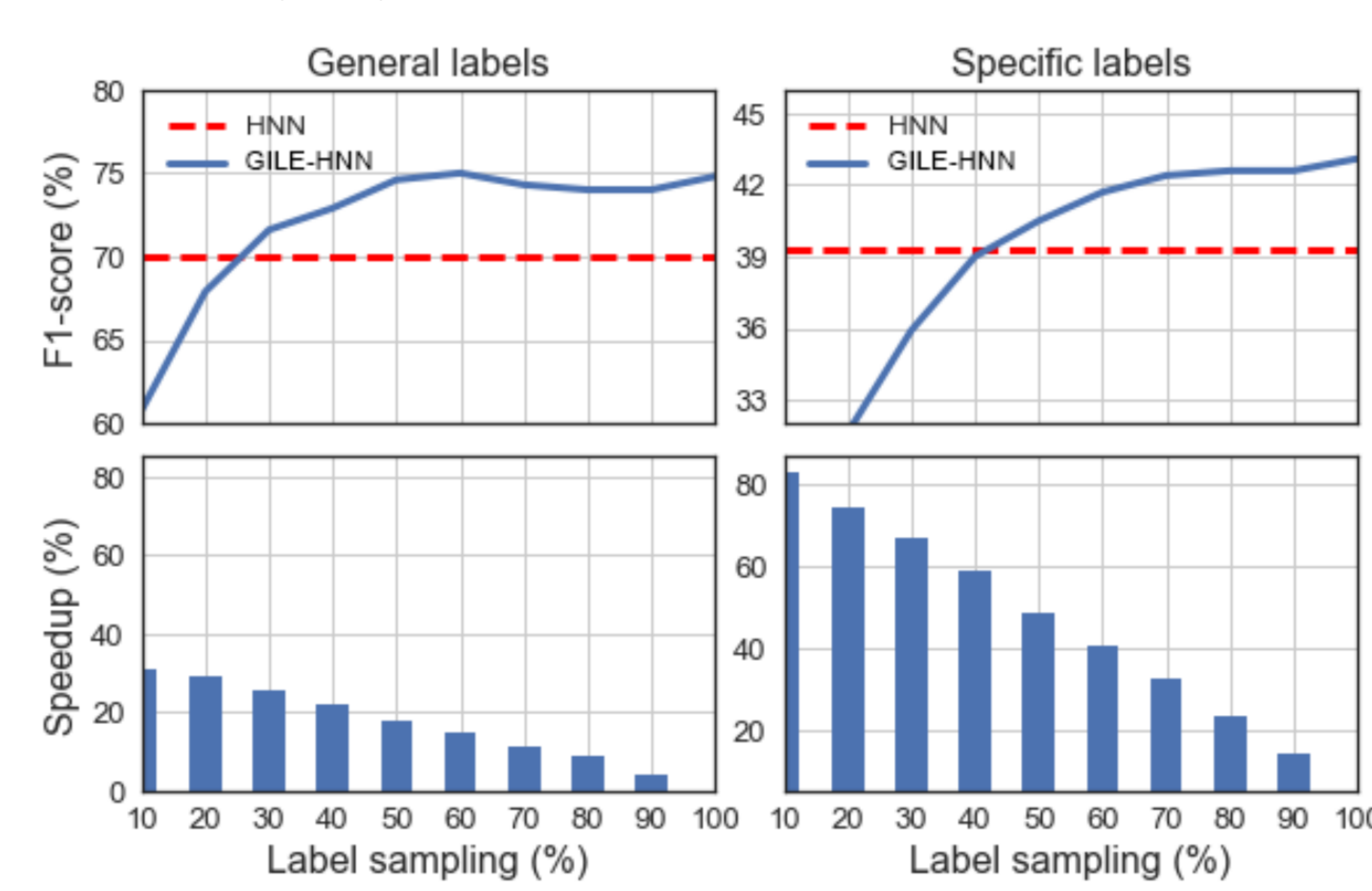
### (B) Multi-task Learning

| Models abbrev. | # lang. | General labels Avg#params | AvgF1 | Specific labels Avg#params | AvgF1 |
|---|---|---|---|---|---|
| HAN [Y16] | 1 | 50K | 77.41 | 90K | 44.90 |
| MHAN [P17] | 2 | 40K | 78.30 | 80K | 45.72 |
| MHAN [P17] | 8 | 32K | 77.91 | 72K | 45.82 |
| GILE-HAN | 1 | 50K | **79.12** | 90K | **45.90** |
| GILE-MHAN | 2 | 40K | **79.68** | 80K | **46.49** |
| GILE-MHAN | 8 | 32K | **79.48** | 72K | **46.32** |

DW dataset: 0.6M documents (80/10/10%), 5K labels (5K seen), 8 languages.

- *HAN*: Hierarchical attention net with no parameter sharing across languages.
- *MHAN*: Multilingual net which shares attention mechanisms across languages.
- *GILE-MHAN*: Multilingual model which shares attention mechanisms and output layer parameters across languages except $w$, $b$.

### (B.2) Effect of Label Sampling




## Conclusion

We proposed an input-label embedding for text classification which:

- Generalizes over previous input-label embedding models.
- Exhibits strong performance on both seen and unseen label prediction.
- Improves multi-task learning models regardless of the encoder type and resource availability of seen labels.

**Future work**: Use more advanced label encoders, perform pretraining on unlabeled data and explore other tasks.