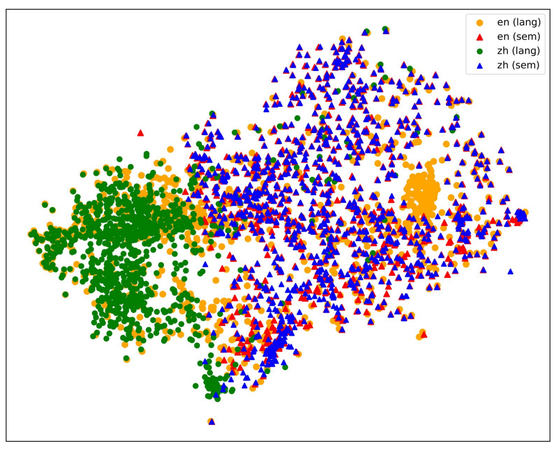
Mitigating Semantic Leakage in Cross-lingual Embeddings via Orthogonality Constraint

Dayeon Ki [1] Cheonbok Park [2] Hyunjoong Kim [2]

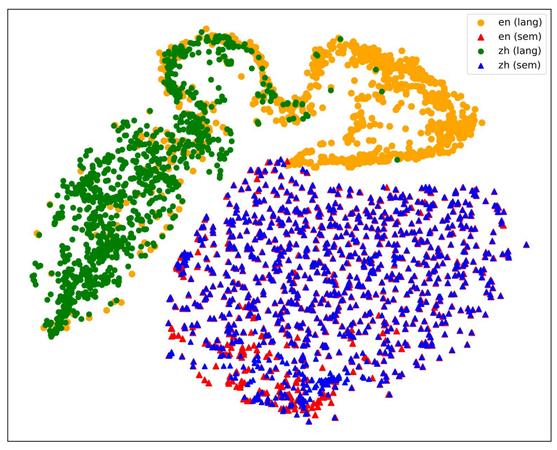
[1] University of Maryland, College Park
[2] NAVER Cloud
dayeonki@umd.edu

Motivation

- Accurately aligning contextual representations in cross-lingual sentence embeddings is key for effective parallel mining.
- One prevalent strategy is disentanglement of semantics and language in sentence embeddings.
- Previous methods suffer from **semantic leakage** - substantial amount of language-specific information is unintentionally leaked into semantic representations



Semantic Leakage



Our Approach (ORACLE)

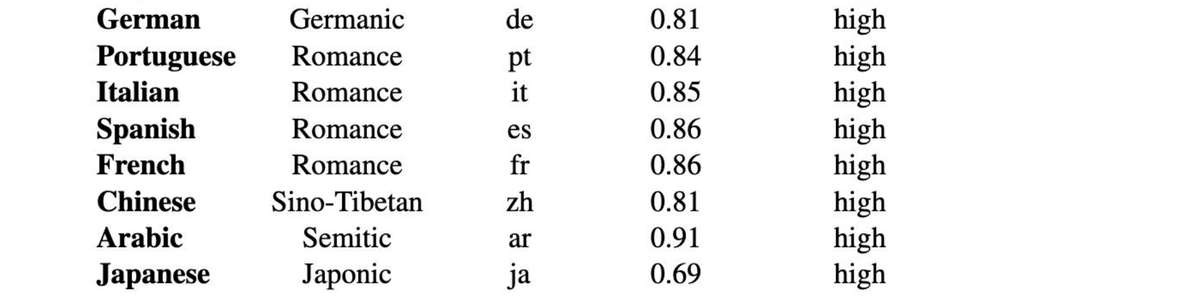
Experiment Setup

- **Dataset -** OPUS / 12 languages in En-XX direction

Language	Family	ISO Code	Similarity	Resource level
English	Germanic	en	-	high
German	Germanic	de	0.81	high
Portuguese	Romance	pt	0.84	high
Italian	Romance	it	0.85	high
Spanish	Romance	es	0.86	high
French	Romance	fr	0.86	high
Chinese	Sino-Tibetan	zh	0.81	high
Arabic	Semitic	ar	0.91	high
Japanese	Japonic	ja	0.69	high
Dutch	Germanic	nl	0.80	medium
Romanian	Romance	ro	0.88	medium
Guaraní	Tupi-Guaraní	gn	0.25	low
Aymara	Andean	ay	0.18	low

- **Baseline**
  - LASER
  - InfoXLM
  - LaBSE

Visualization



(a) DREAM (b) DREAM + ORACLE (c) MEAT (d) MEAT + ORACLE

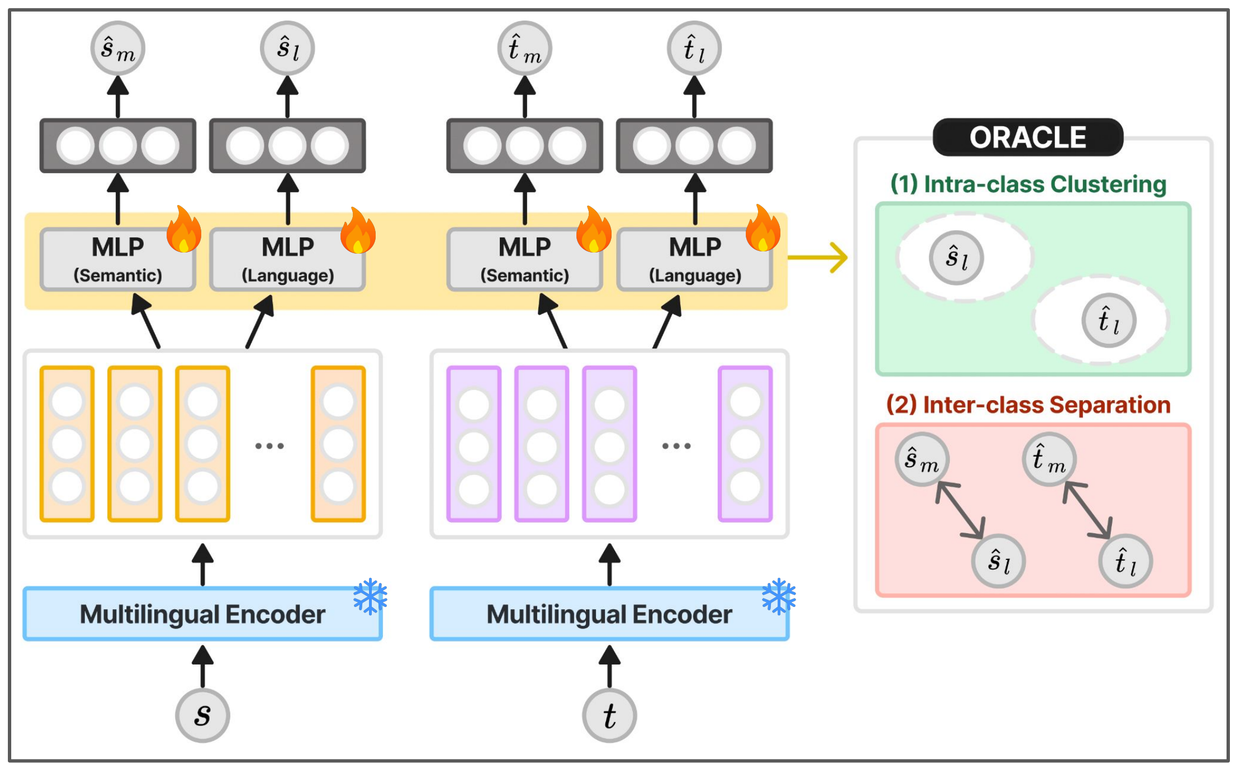
Ablations

Combining both components of ORACLE yields the most balanced performance.

Objective	Tatoeba-14	Tatoeba-36	STS
Semantic Embedding (↑)			
ORACLE	**96.11**	95.53	**74.21**
- $\mathcal{L}_{\text{IC}}$	95.89	95.38	74.13
- $\mathcal{L}_{\text{IS}}$	96.11	**95.54**	72.81
Language Embedding (↓)			
ORACLE	**7.74**	**9.17**	**16.47**
- $\mathcal{L}_{\text{IC}}$	37.78	39.15	30.14
- $\mathcal{L}_{\text{IS}}$	8.07	9.59	18.20

ORACLE

- How well are the semantic representations **aligned**?
- How well are the language-specific representations **separated**?



(1) Intra-class Clustering
: bring related components closer in embedding space

$$\mathcal{L}_{\text{IC}} = \frac{1}{N}\sum_{i=1}^{N}\left(2 - \phi(\hat{\mathbf{s}}_l^i, \hat{\mathbf{s}}_l^j) - \phi(\hat{\mathbf{t}}_l^i, \hat{\mathbf{t}}_l^j)\right)$$

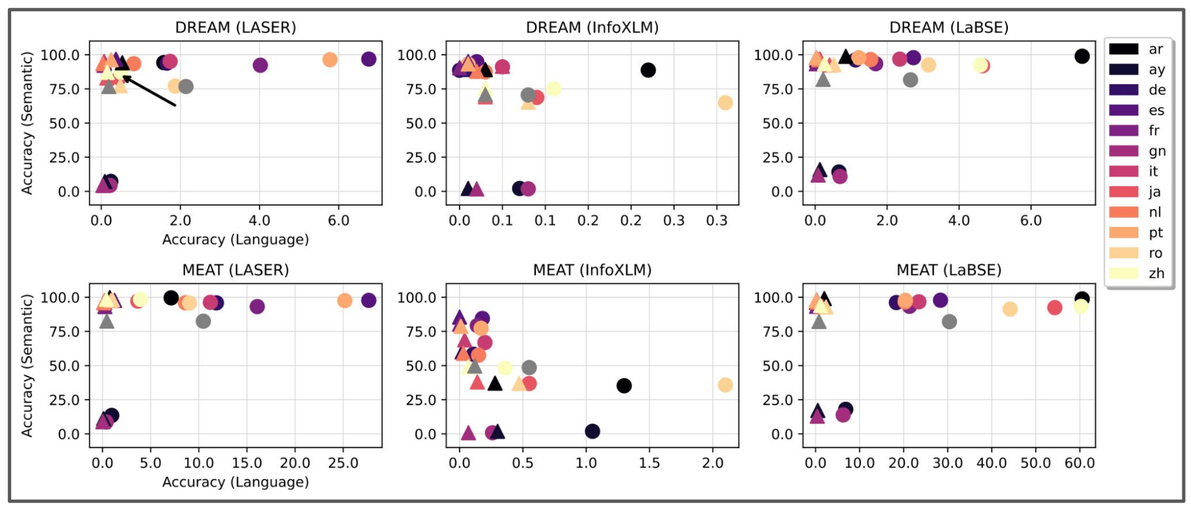
(2) Inter-class Separation
: ensure unrelated components to be distant

$$\mathcal{L}_{\text{IS}} = \frac{1}{N}\sum_{i=1}^{N}\max(0, \phi(\hat{\mathbf{s}}_m^i, \hat{\mathbf{s}}_l^i)) + \max(0, \phi(\hat{\mathbf{t}}_m^i, \hat{\mathbf{t}}_l^i))$$

Results

Cross-lingual Sentence Retrieval

- **Optimal** representation *(towards upper left corner)*
  - Higher semantic retrieval accuracy
  - Lower language retrieval accuracy



Semantic Textual Similarity

Spearman's rank correlation coefficient higher for semantic (●) and lower for language-specific (★) representation with ORACLE.