Entity Linkage with Deep Learning

Shobeir Fakhraei
University of Southern California

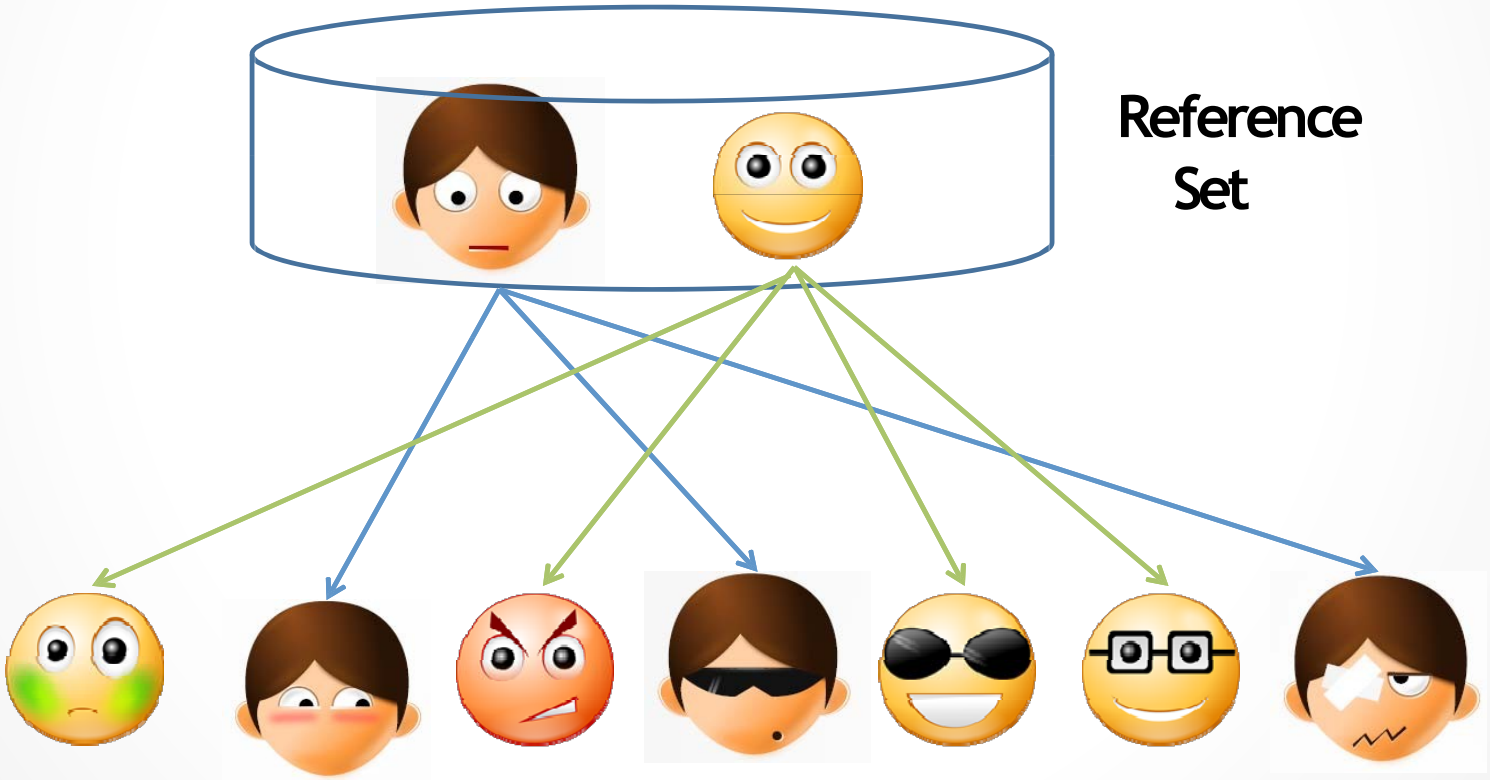
Where are we?

- Information Extraction
  - Semi-structured: Web, HTML
  - Unstructured: Text, Adds, Tweets, …
- Entity Linkage, Data Cleaning, Normalization
- Logical Data Integration
  - Mediators, Query Rewriting
  - Warehouse, Logical Data Exchange
- Automatic Source Modeling/Learning Schema Mappings
- Semantic Web
  - RDF, SPARQL, OWL, Linked Data
- Advanced Topics
  - Geospatial Data Integration, Knowledge Graphs

Deep Entity Normalization …

Reference Matching Problem

- Match noisy records to clean records in a reference table



Biomedical Information Extraction

textual abstract:
summary for human

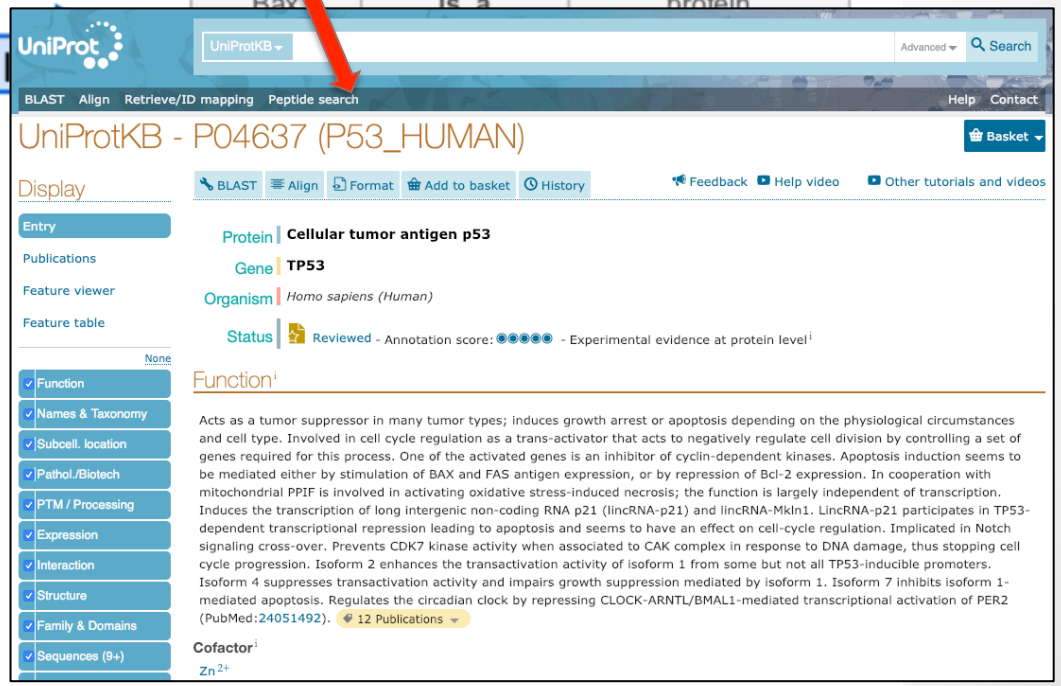
IE

Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

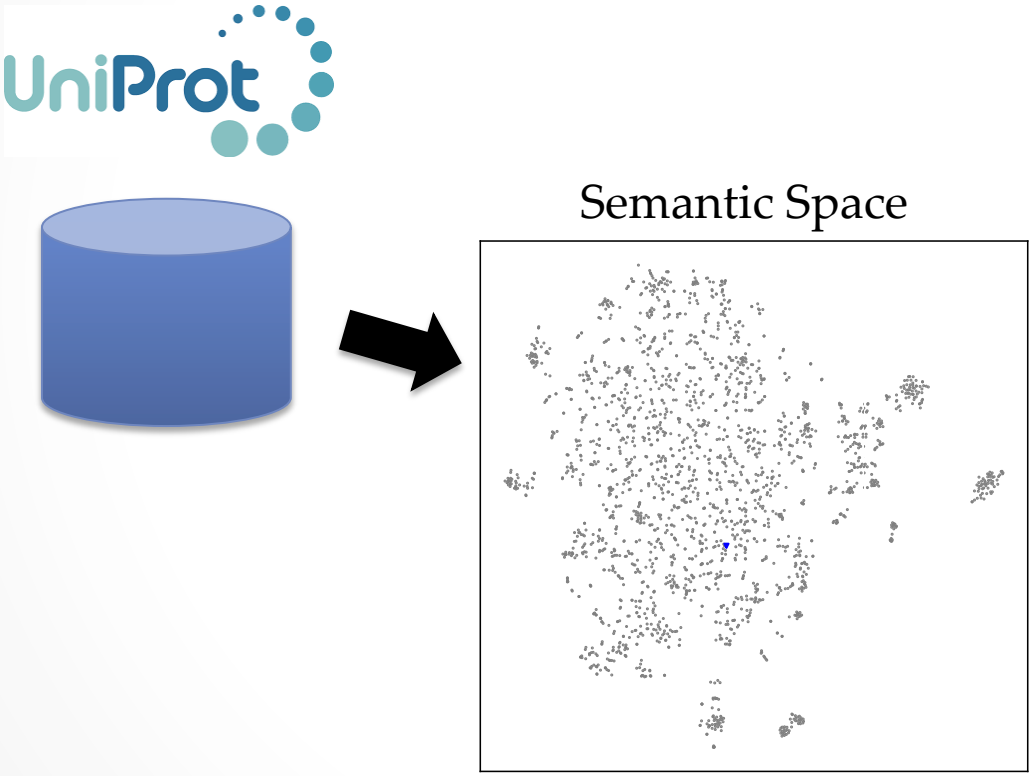
structured knowledge extraction:
summary for machine

Biomedical Information Extraction

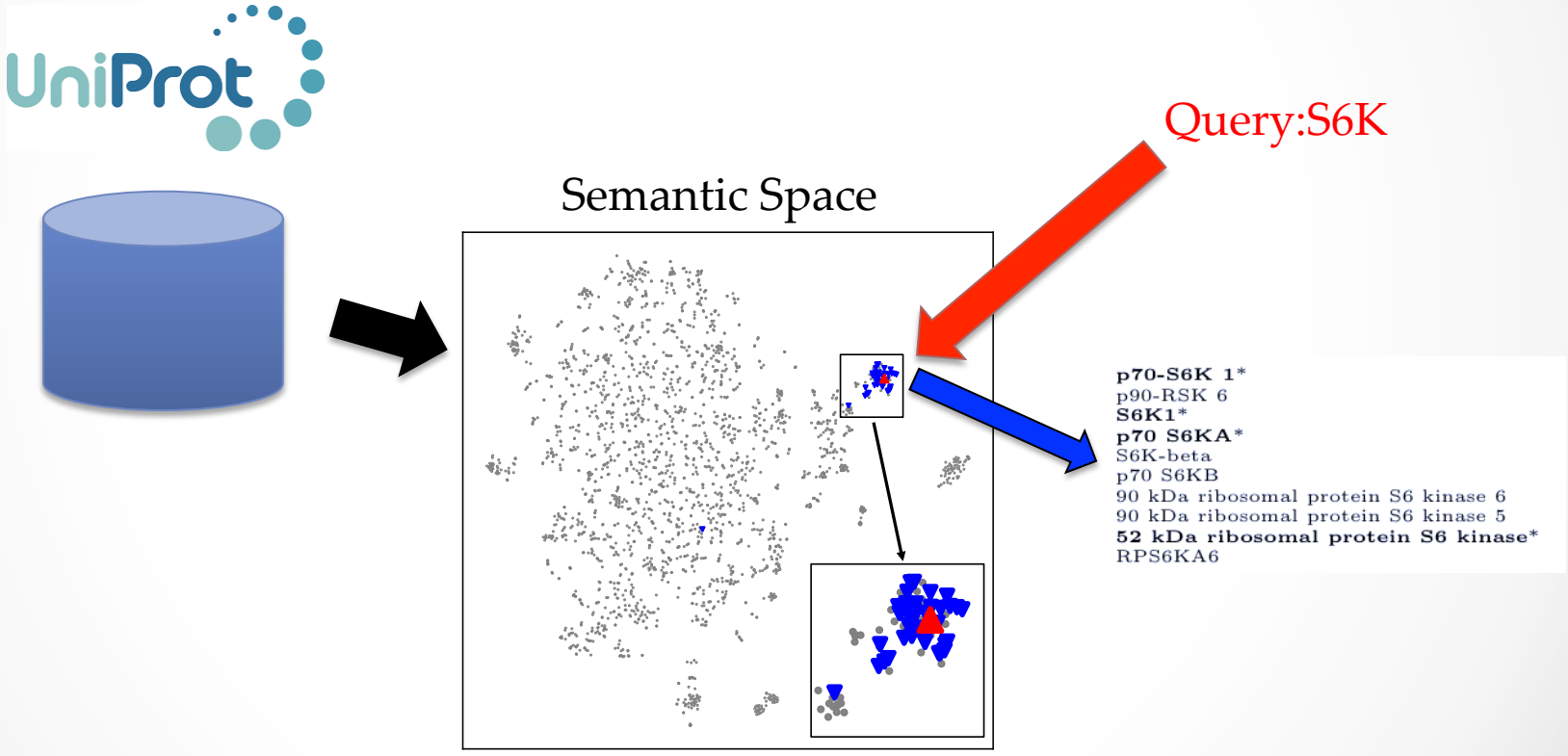
Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein



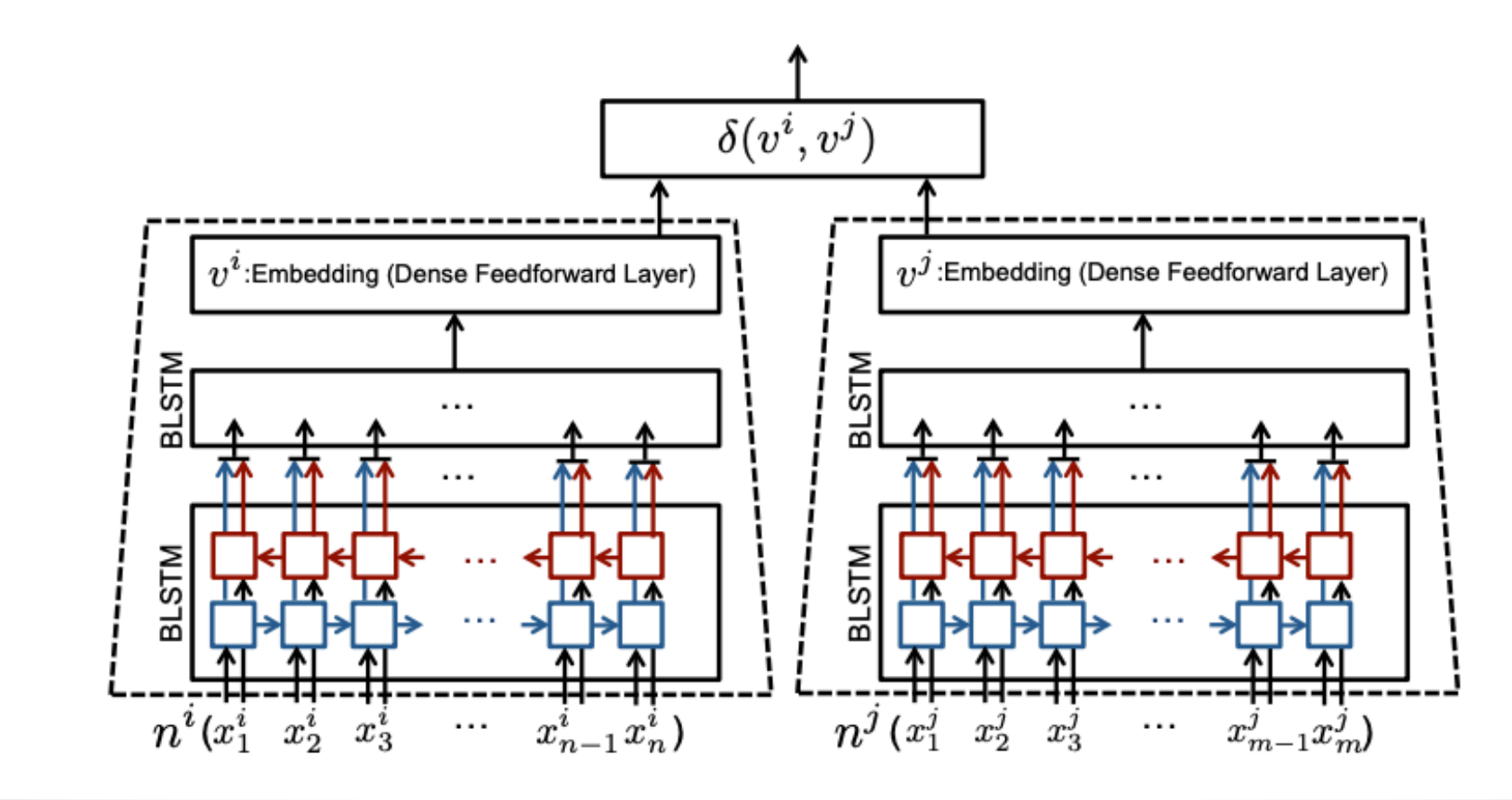
Our Deep Learning-based Solution



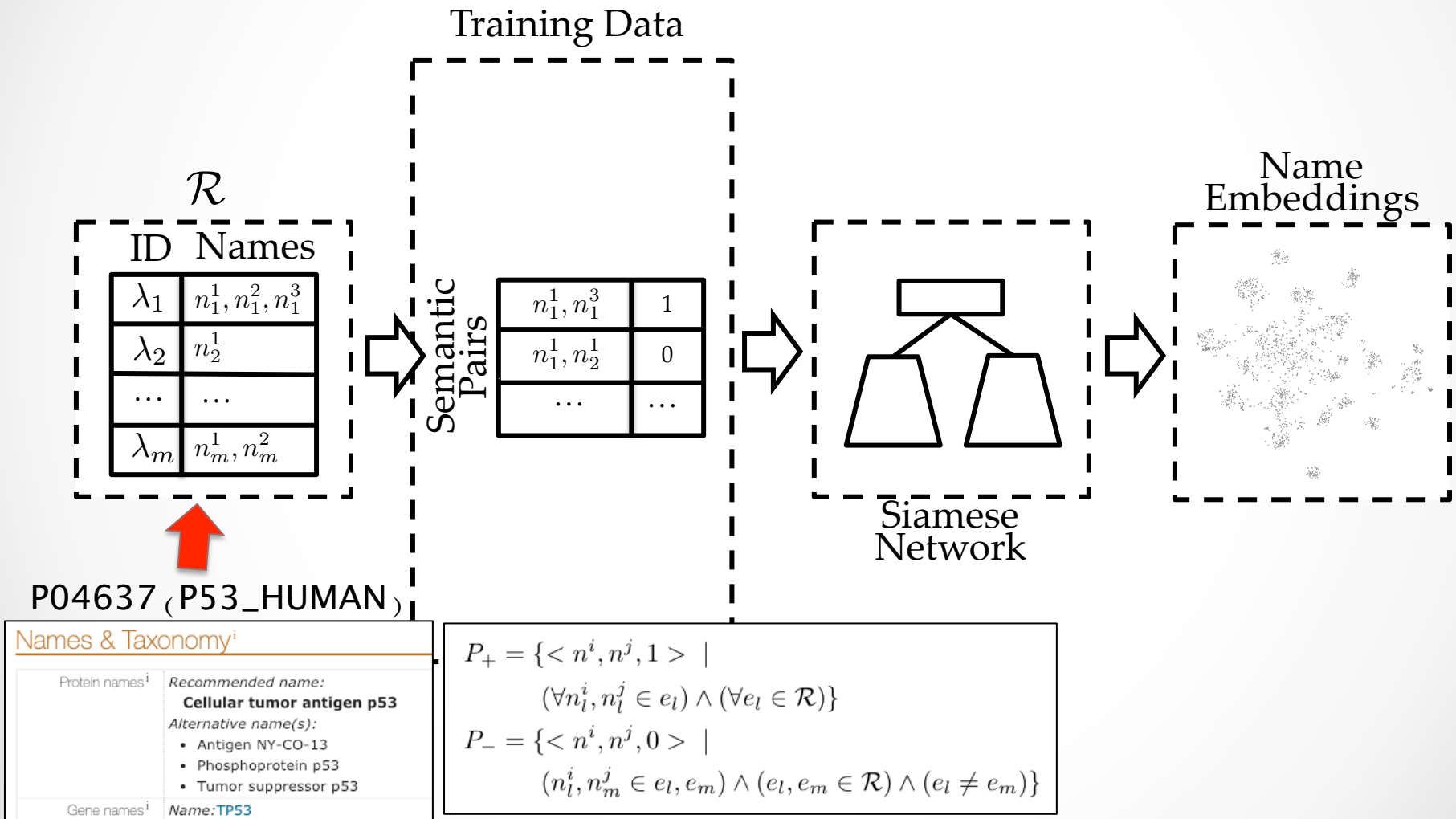
Our Deep Learning-based Solution



Similarity Learning via Siamese Network

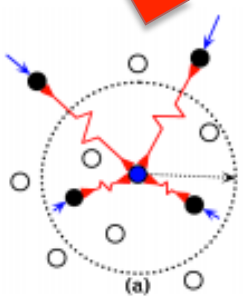


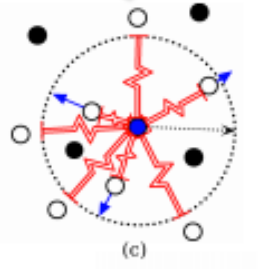
Pair Generation: Semantic



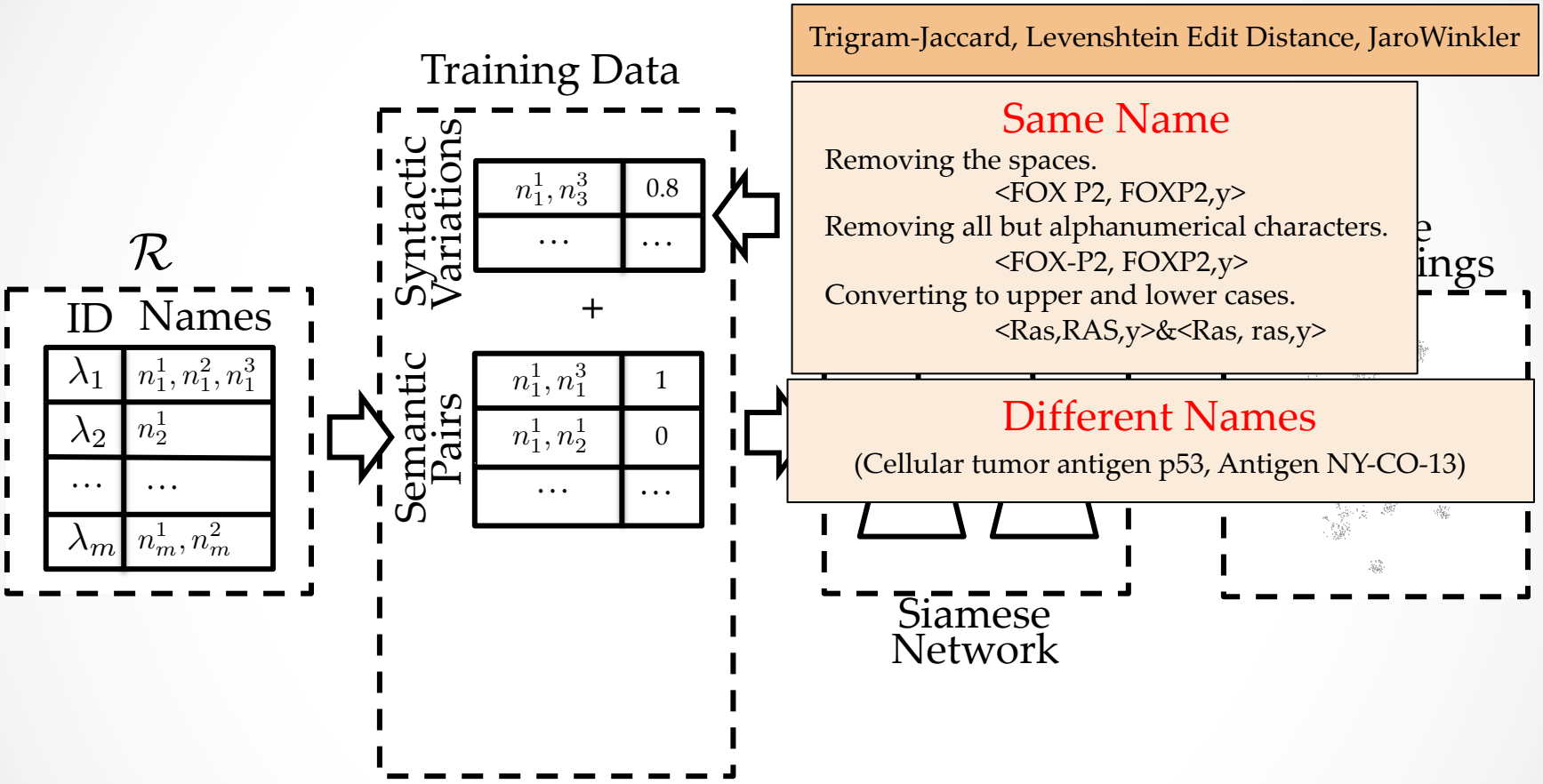
Contrastive Loss

$$\mathcal{L} = \frac{1}{2} y \delta(v_i, v_j)^2 + \frac{1}{2}(1-y)\max(0, m - \delta(v_i, v_j))^2$$



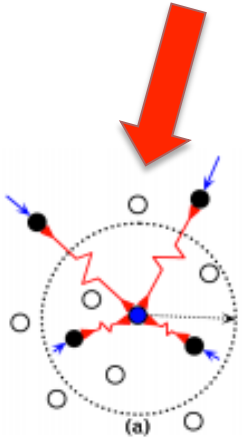


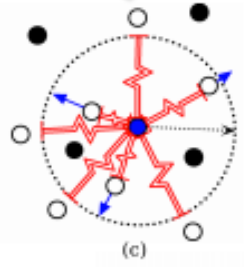
Pair Generation: Syntactic Variations



Contrastive Loss with soft labels

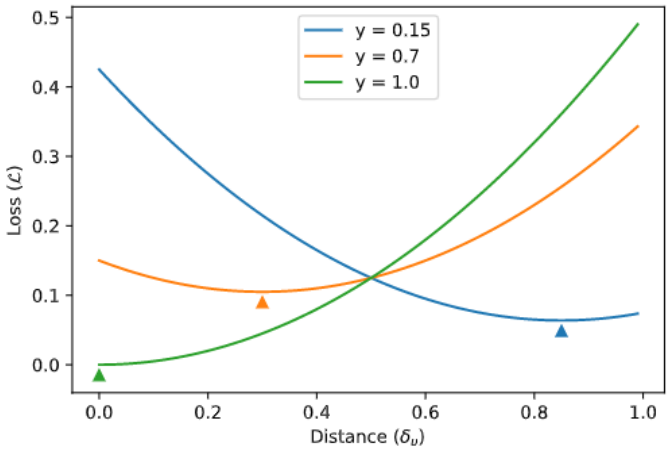
$$\mathcal{L} = \frac{1}{2} y \delta(v_i, v_j)^2 + \frac{1}{2}(1-y)\max(0, m - \delta(v_i, v_j))^2$$



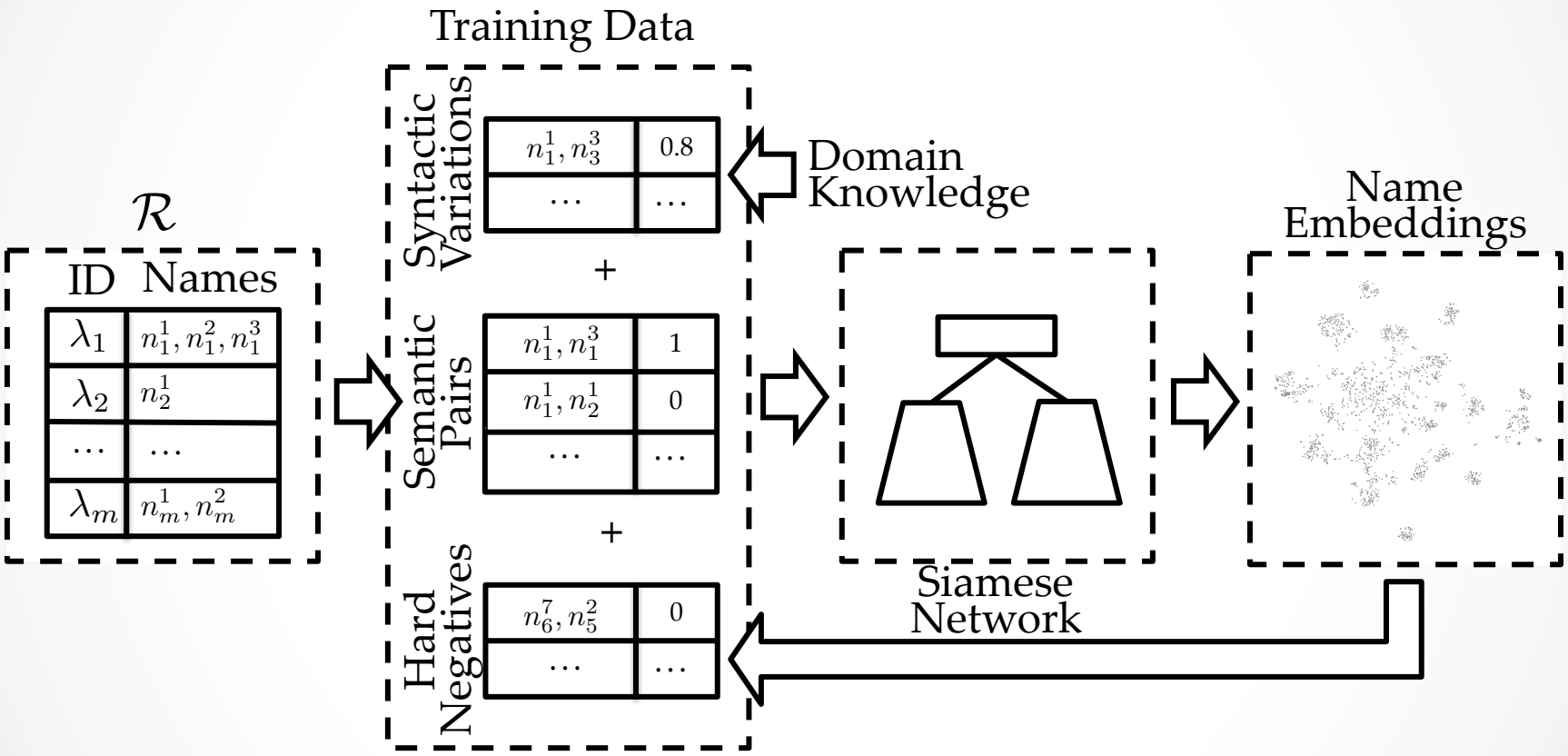


$$\frac{\partial \mathcal{L}}{\partial \delta_v} = y\delta_v - (1-y)(1-\delta_v)$$

$$\arg\min_{\delta_v} \mathcal{L} = \{\delta_v \mid y + \delta_v - 1 = 0\} = 1 - y$$



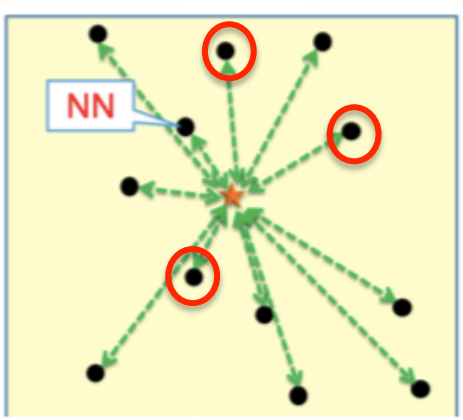
Pair Generation: Hard Negatives

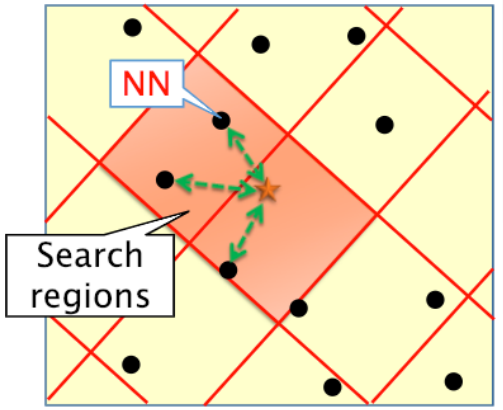


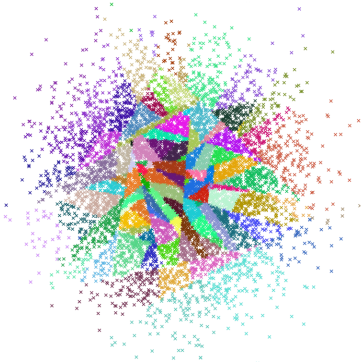
$$P_- = \{< n^i, n^j, 0 > \mid (n_l^i, n_m^j \in e_l, e_m) \wedge (e_l, e_m \in \mathcal{R}) \wedge (e_l \neq e_m)\}$$

Easy Samples

Finding Hard Negatives

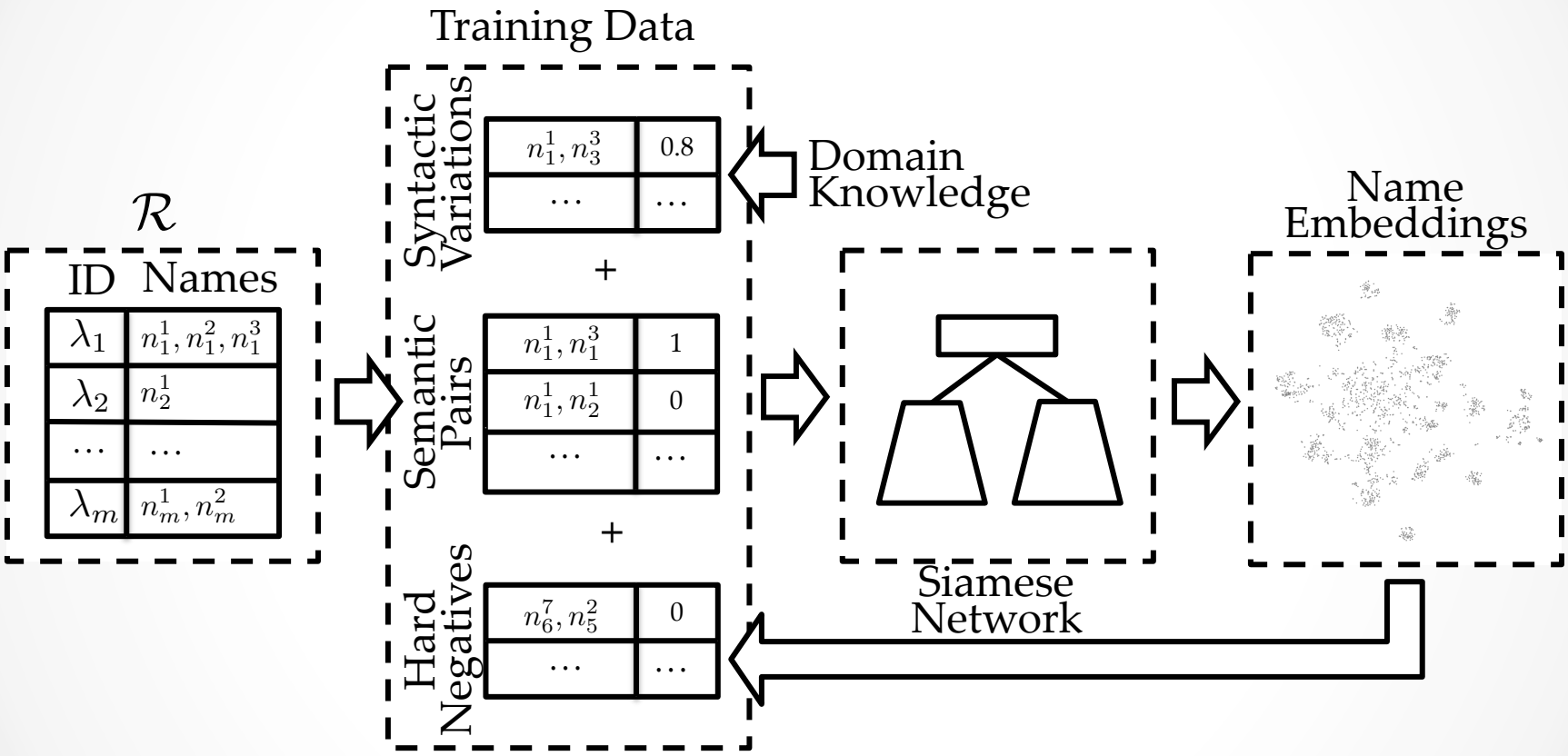






Annoy
(Approximate Nearest
Neighbors Oh Yeah)

Overall Framework (Similarity Learning)



Similarity Learning

Algorithm 1 NSEEN: Similarity Learning

procedure TRAINSIM($\mathcal{R}$, P_d)
    Input: $\mathcal{R}$ reference set
    Input: P_d pairs based on knowledge of syntactic variation in the domain
    Generate pairs based on reference set $\mathcal{R}$ and add them to training data $\mathcal{D}$
    Add P_d pairs to the training data $\mathcal{D}$
    for k times **do**
        Train the model $\mathcal{M}$ (Siamese network) on $\mathcal{D}$
        Embed all the names in $\mathcal{R}$: $n \rightarrow v$
        for all v_l^i **do** ▷ Hard negative mining
            find the k closest v_k^j to v_l^i
            if $k \neq l$ **then**
                add $< n_k^j, n_l^i, 0 >$ to training data $\mathcal{D}$
    return $\mathcal{M}$ ▷ The trained embedding function

Embedding the Database

Algorithm 2 NSEEN: Embedding & Hashing $\mathcal{R}$

1: **procedure** EMBED($\mathcal{R}, \mathcal{M}$)
2: **for all** $n_i \in \mathcal{R}$ **do**
3: $n_i \xrightarrow{\mathcal{M}} v_i$
4: **for all** v_i **do**
5: Hash v_i and store in $\mathcal{H}_{v_i}$
6: **return** $\mathcal{H}_v$ ▷ Hashed embeddings

Retrieval

Algorithm 3 NSEEN: Retrieval

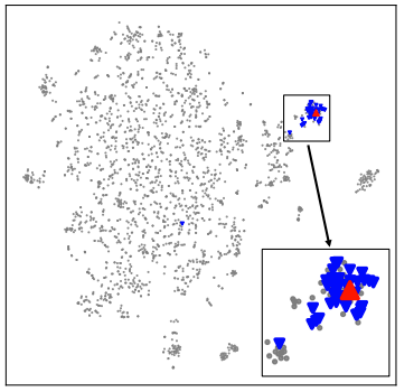
1: **procedure** RETRIEVE($\mathcal{H}_v$, $\mathcal{M}$, n_q)
2: Embed the query name: $n_q \xrightarrow{\mathcal{M}} v_q$
3: Find the closest v_k^j to v_q using approximate nearest neighbor search (Annoy) on $\mathcal{H}_v$
4: **return** λ_k as the ID (i.e., λ_q)

Results

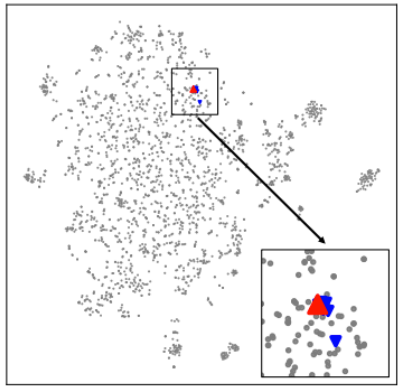
$\mathcal{R}$	DS	Model	H@1	H@3	H@5	H@10
UniProt	BC1	NSEEN	**0.833**	**0.869**	**0.886**	**0.894**
		Baseline	0.814	0.864	0.875	0.885
	BC2	NSEEN	**0.861**	0.888	0.904	**0.930**
		Baseline	0.841	0.888	0.904	0.919
ChEBI	BC1	NSEEN	**0.505**	**0.537**	**0.554**	**0.574**
		Baseline	0.418	0.451	0.460	0.468
	BC2	NSEEN	**0.578**	**0.608**	**0.624**	**0.641**
		Baseline	0.444	0.472	0.480	0.491

Examples

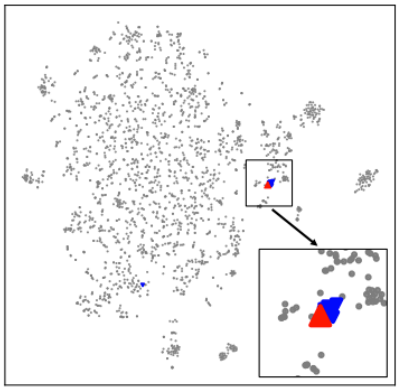
S6K	PLCγ2	IKKϵ	H3
p70-S6K 1*	**PLC-gamma-2***	**IKK-epsilon***	**Histone H3/a***
p90-RSK 6	PLC-gamma-1	**IKKE***	**Histone H3/o***
S6K1*	**PLCG2***	**I-kappa-B kinase epsilon***	**Histone H3/m***
p70 S6KA*	**Phospholipase C-gamma-2***	**IkBKE***	**Histone H3/b***
S6K-beta	Phospholipase C-gamma-1	**IKBKE***	**Histone H3/f***
p70 S6KB	PLC	IKBE	**HIST1H3C***
90 kDa ribosomal protein S6 kinase 6	PLCG1	IK1	**Histone H3/k***
90 kDa ribosomal protein S6 kinase 5	**Phosphoinositide phospholipase C-gamma-2***	IK1	**Histone H3/i***
52 kDa ribosomal protein S6 kinase*	**PLC-IV***	IKKG	**HIST1H3G***
RPS6KA6	PLCB	INKA1	**Histone H3/d***



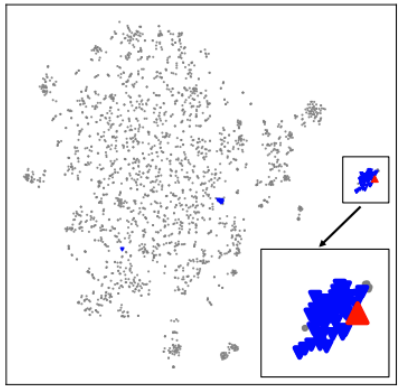
(a) S6K



(b) PLCγ2



(c) IKKϵ

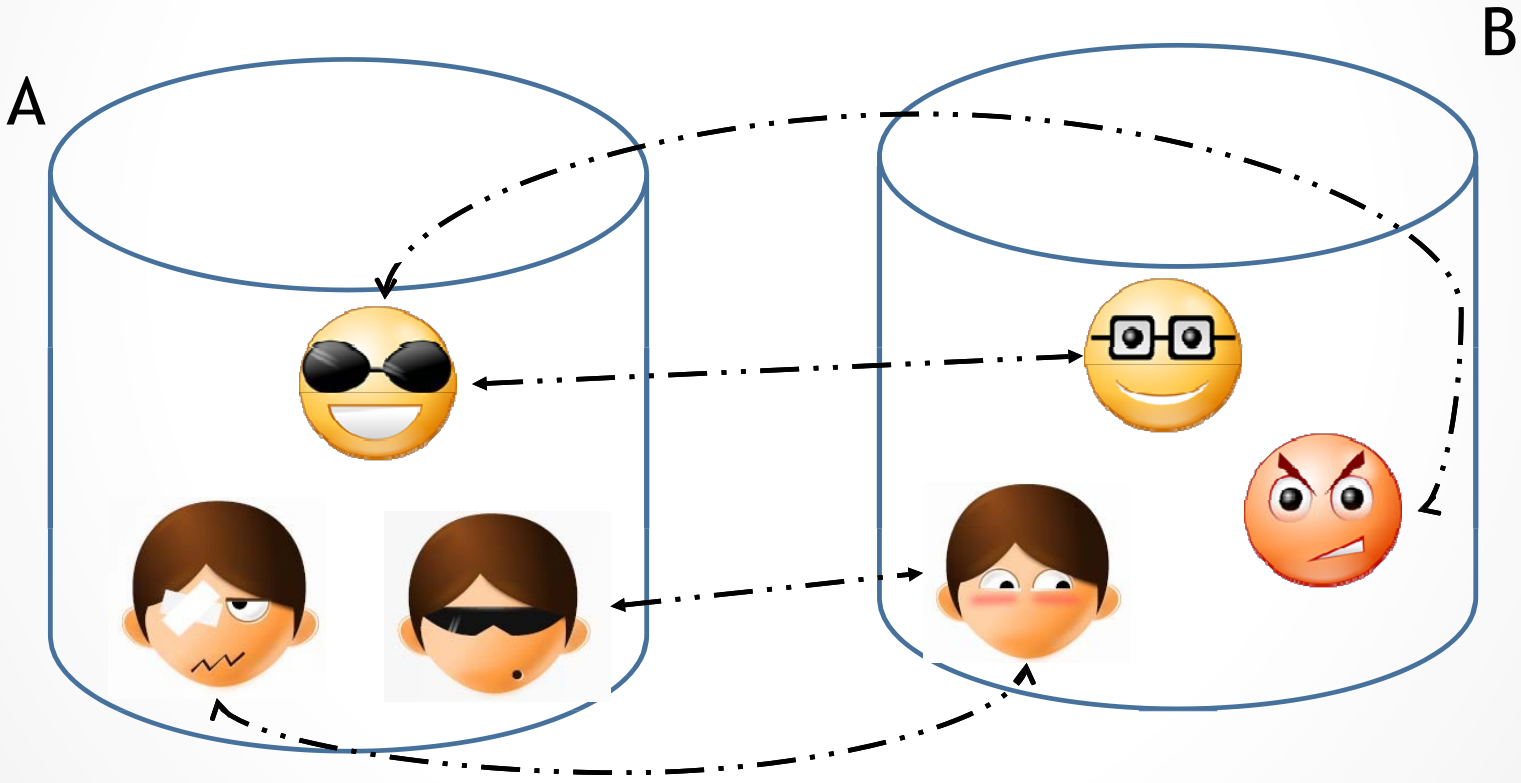


(d) H3

Deep Record Matching

Record Linkage Problem Statement

- Link records that match across databases



Different Kinds of Records

	Name	City	Age
t_1	Dave Smith	New York	18
t_2	David Smith	New York	18

(a) structured

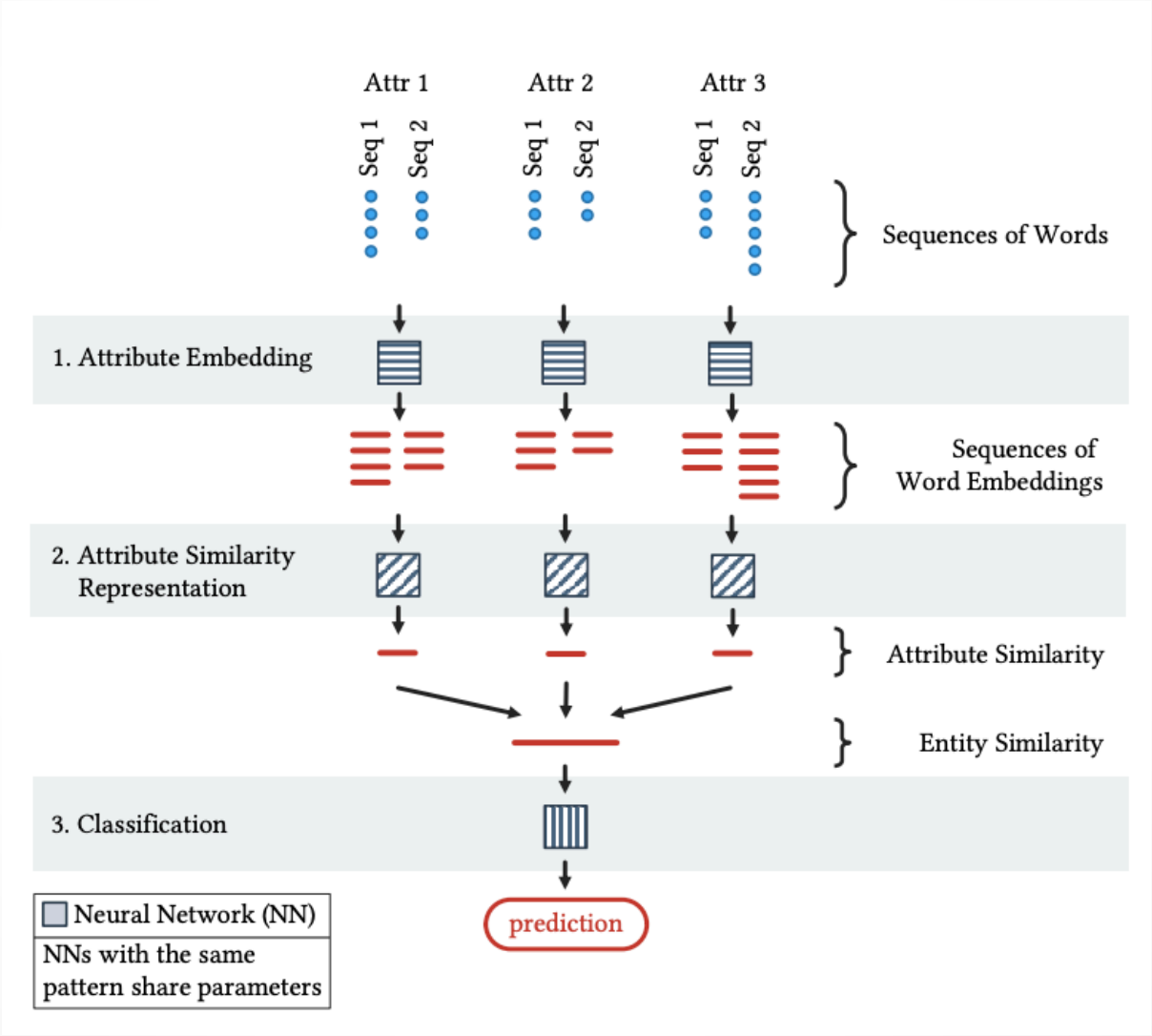
	Description
t_1	Kingston 133x high-speed 4GB compact flash card ts4gcf133, 21.5 MB per sec data transfer rate, dual-channel support, multi-platform compatibility.
t_2	Kingston ts4gcf133 4GB compactflash memory card (133x).

(b) textual

	Name	Brand	Price
t_1	Adobe Acrobat 8		299.99
t_2	Acrobat 8	Adobe	299.99

(c) dirty

Overall Architecture



Options at each stage

Architecture module	Options	
Attribute embedding	*Granularity:* (1) Word-based (2) Character-based	*Training:* (3) Pre-trained (4) Learned

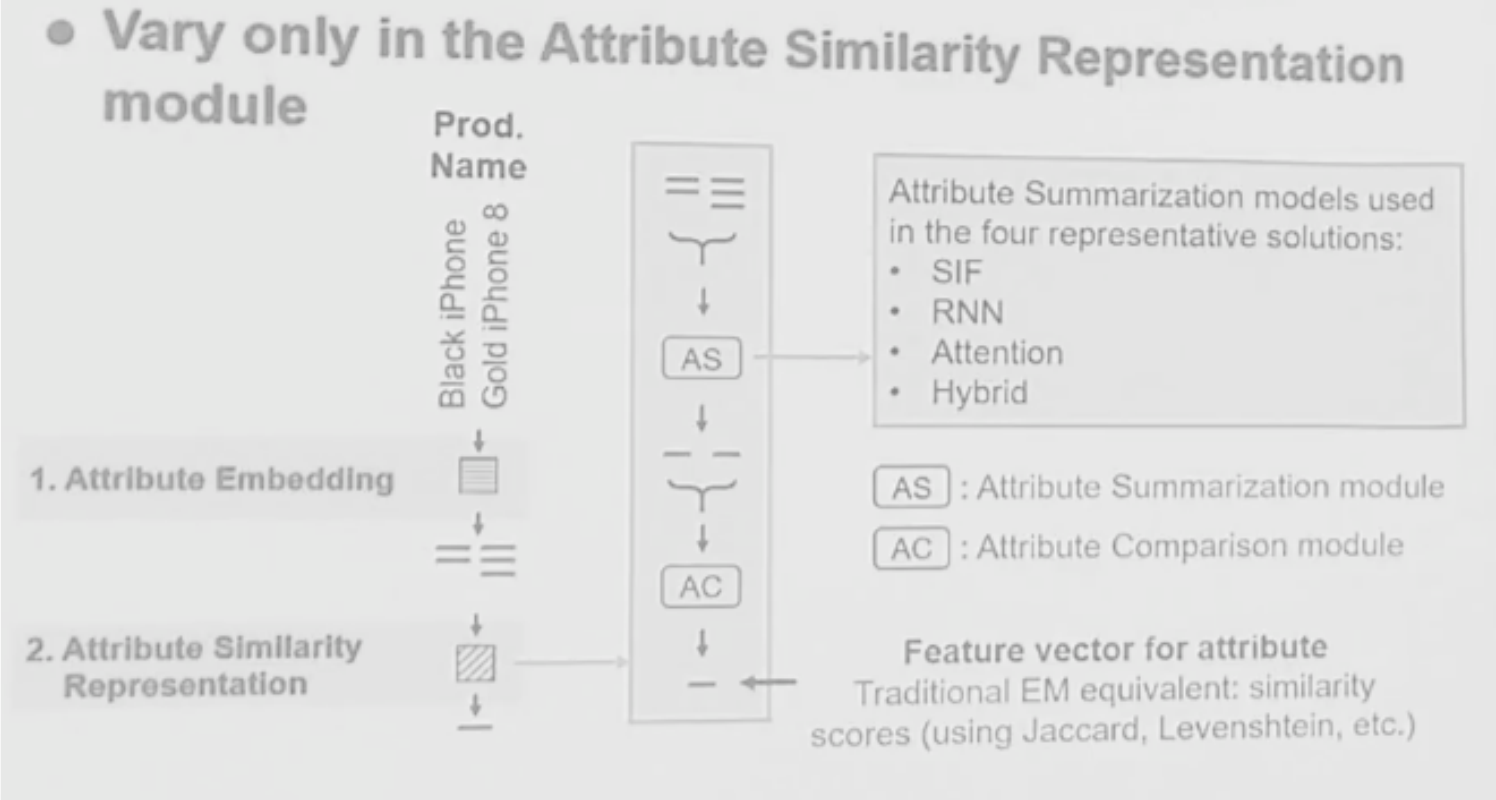
Figure 3: The design space of DL solutions for EM.

Options at each stage

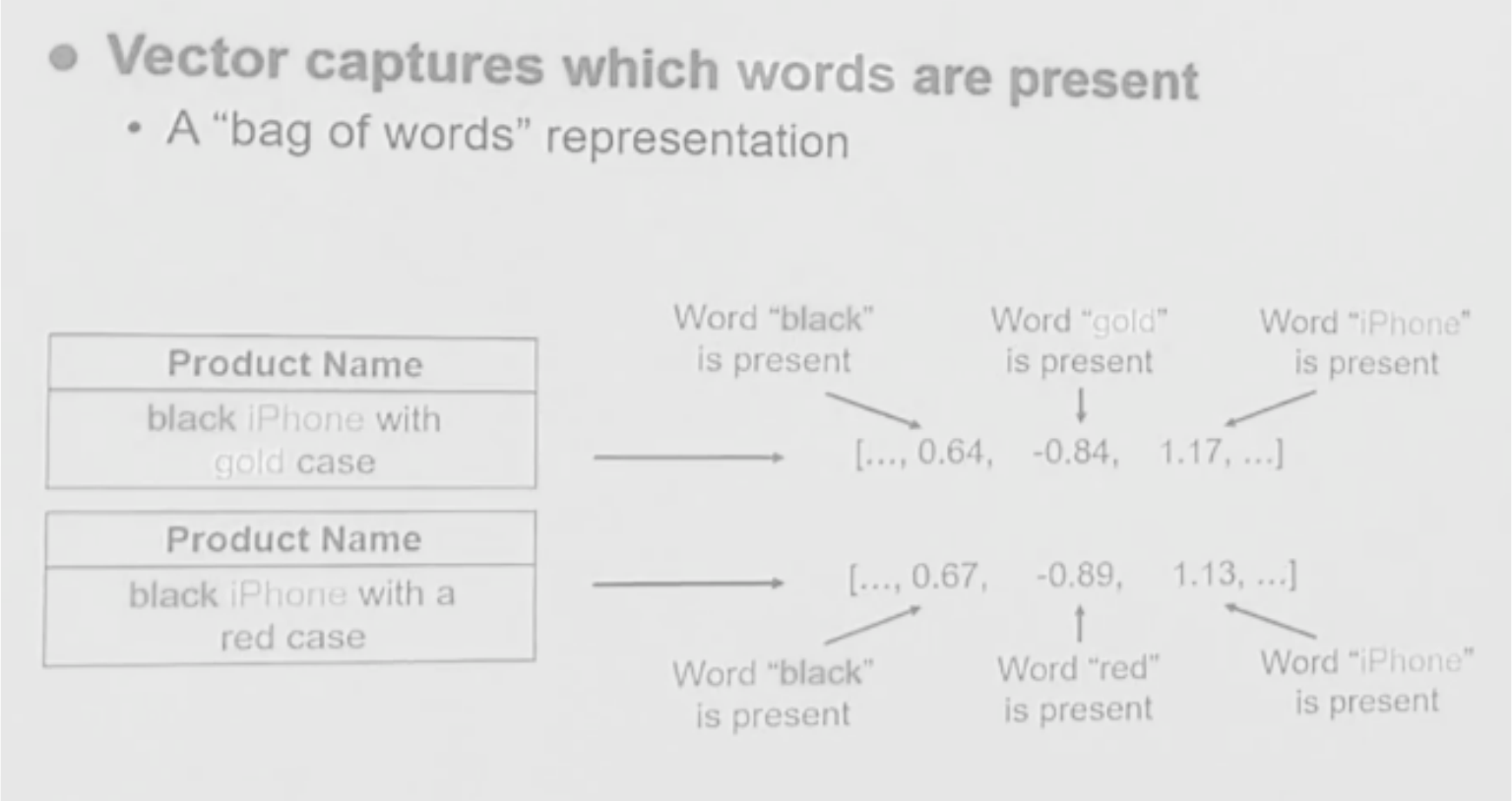
Architecture module		Options	
Attribute embedding		*Granularity:* (1) Word-based (2) Character-based	*Training:* (3) Pre-trained (4) Learned
Attribute similarity representation	(1) Attribute summarization	(1) Heuristic-based (2) RNN-based (3) Attention-based (4) Hybrid	
	(2) Attribute comparison	(1) Fixed distance (cosine, Euclidean) (2) Learnable distance (concatenation, element-wise absolute difference, element-wise multiplication)	

Figure 3: The design space of DL solutions for EM.

Attribute Similarity



Aggregation-based attribute summarization



SIF: Attribute Aggregation

- **Uses smooth inverse frequency**
- **Weighted sum of word embeddings**
  - Weight is inversely proportional to freq. of word

Tuple pair

	Product Name
t_1	Laptop HP191

	Product Name
t_2	Notebook HP191

Word weights

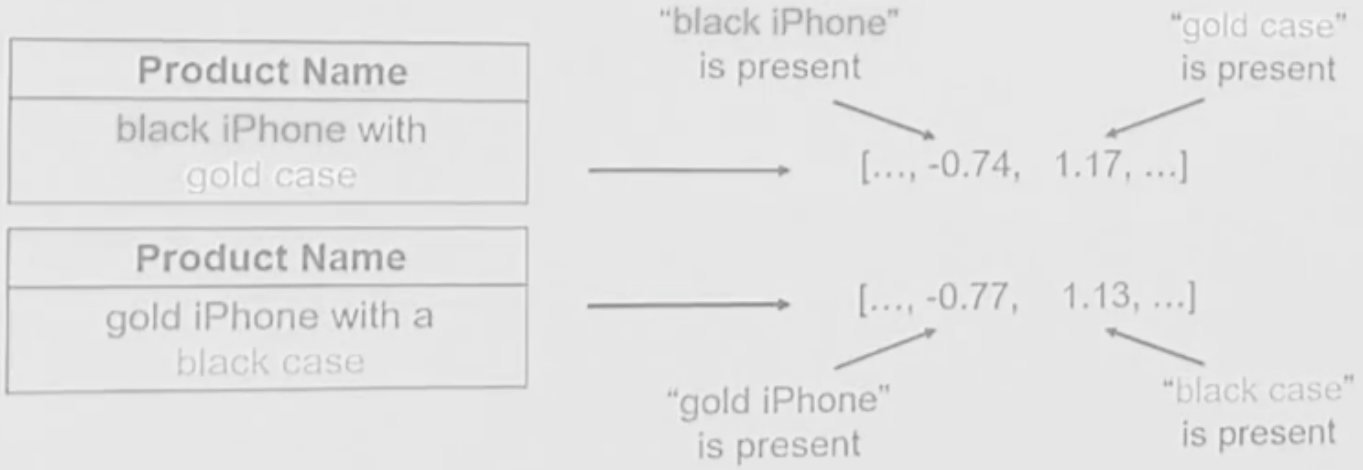
Word	Weight
Laptop	0.05
Notebook	0.07
HP191	0.9

SIF attribute summary computation

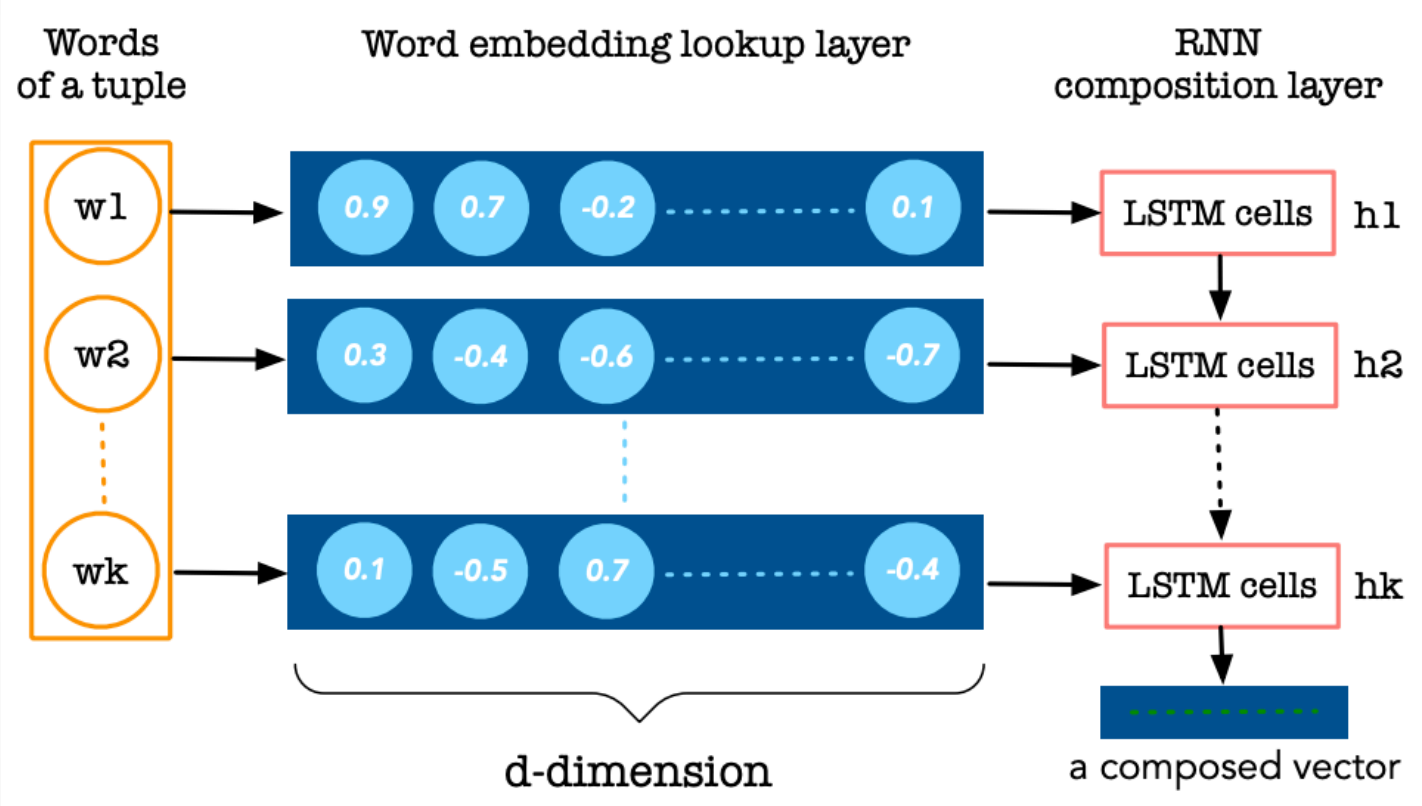
Product Name
(Laptop) (HP191) t_1 0.05 x I + 0.9 x I = I (Attr. Sum. t_1) t_2 0.07 x I + 0.9 x I = I (Attr. Sum. t_2) (Notebook) (HP191)

Sequence-based attribute summarization

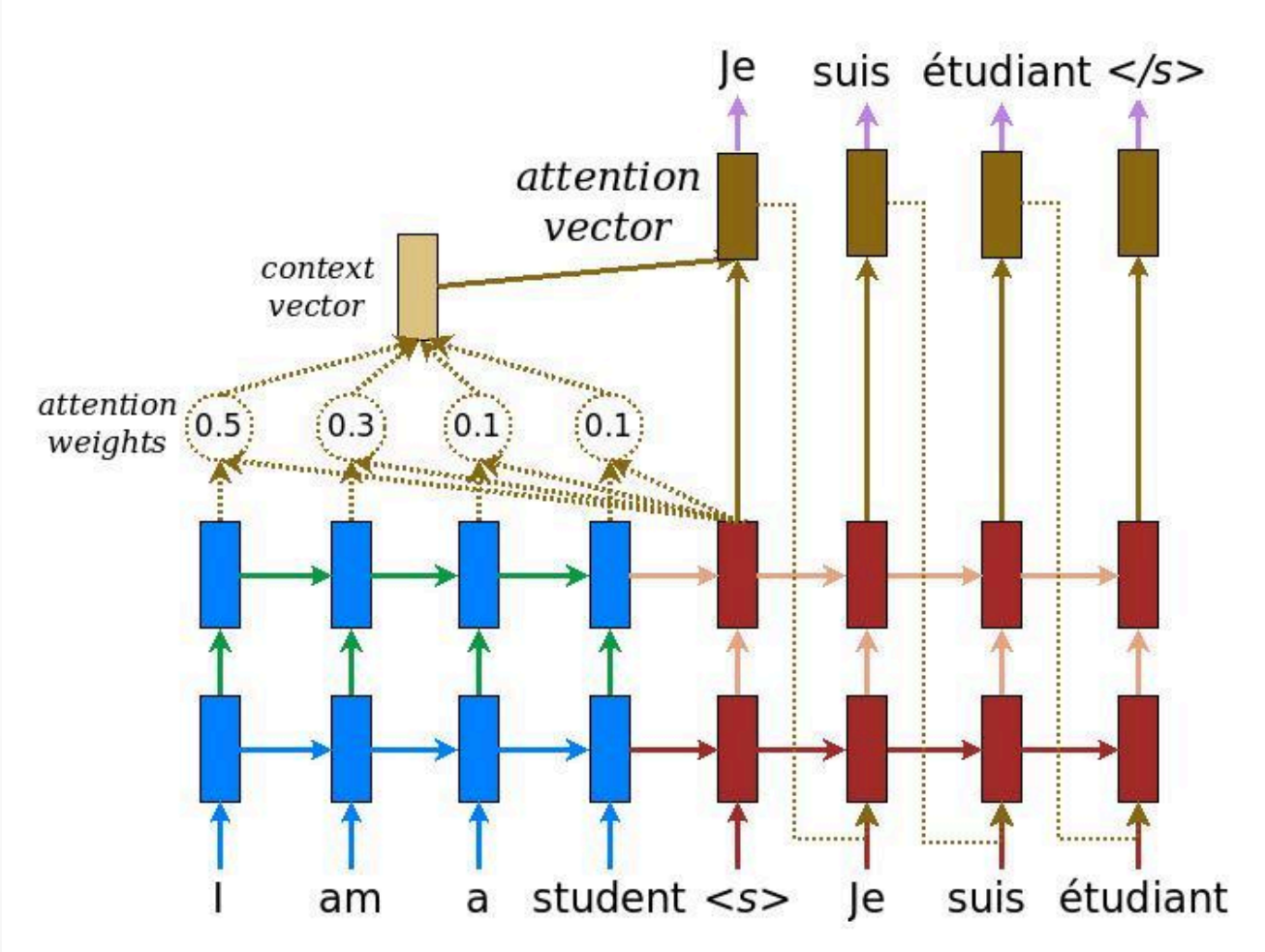
- **Vector captures which word sequences are present**
  - Automatically learns which word sequences are important



RNN: attribute summarization



Attention


Je
suis
étudiant </s>
attention vector
context vector
attention weights
0.5
0.3
0.1
0.1
I
am
a
student <s>
Je
suis
étudiant

Options at each stage

Architecture module		Options	
Attribute embedding		*Granularity:* (1) Word-based (2) Character-based	*Training:* (3) Pre-trained (4) Learned
Attribute similarity representation	(1) Attribute summarization	(1) Heuristic-based (2) RNN-based (3) Attention-based (4) Hybrid	
	(2) Attribute comparison	(1) Fixed distance (cosine, Euclidean) (2) Learnable distance (concatenation, element-wise absolute difference, element-wise multiplication)	
Classifier		NN (multi-layer perceptron)	

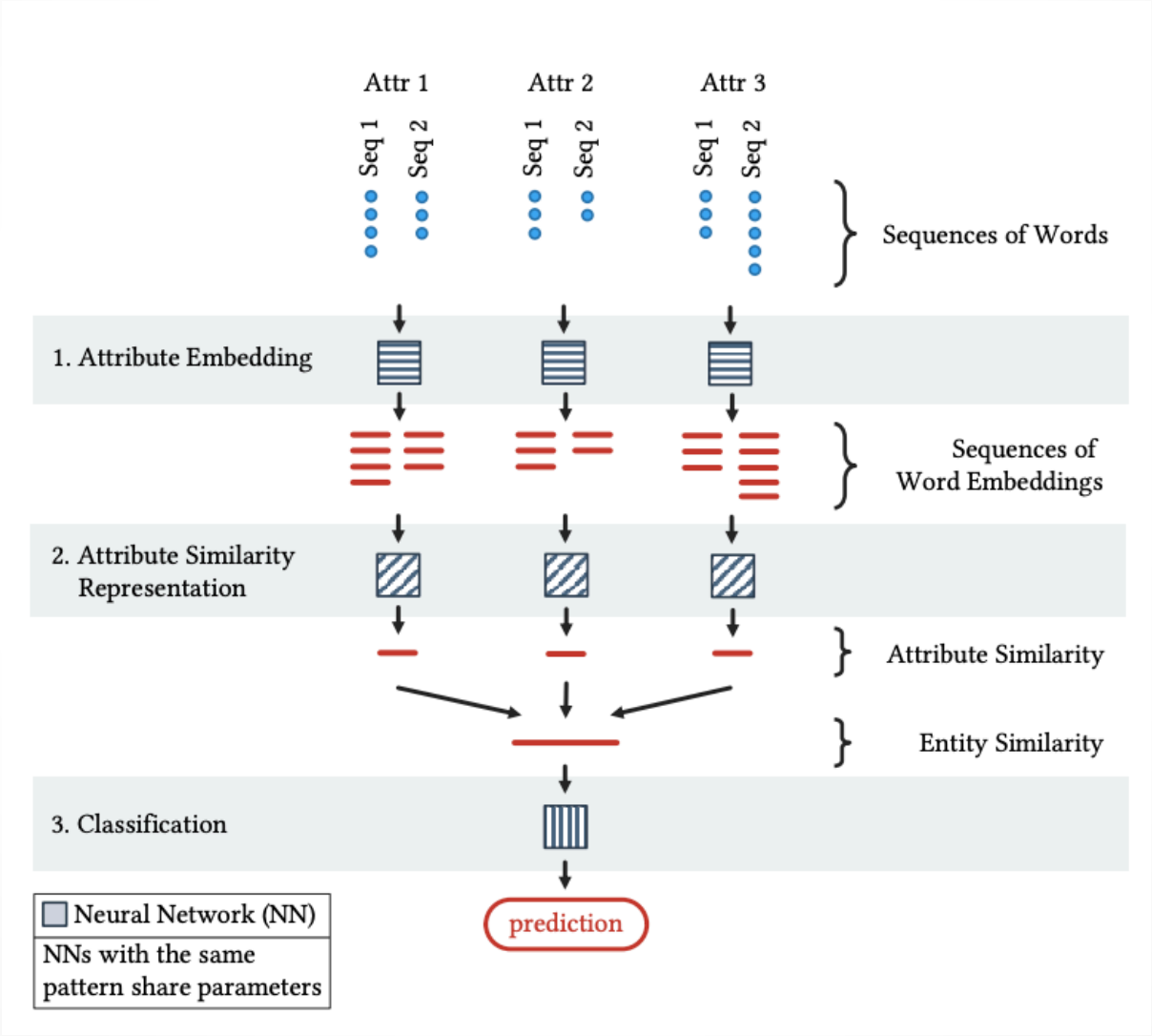
Figure 3: The design space of DL solutions for EM.

Results

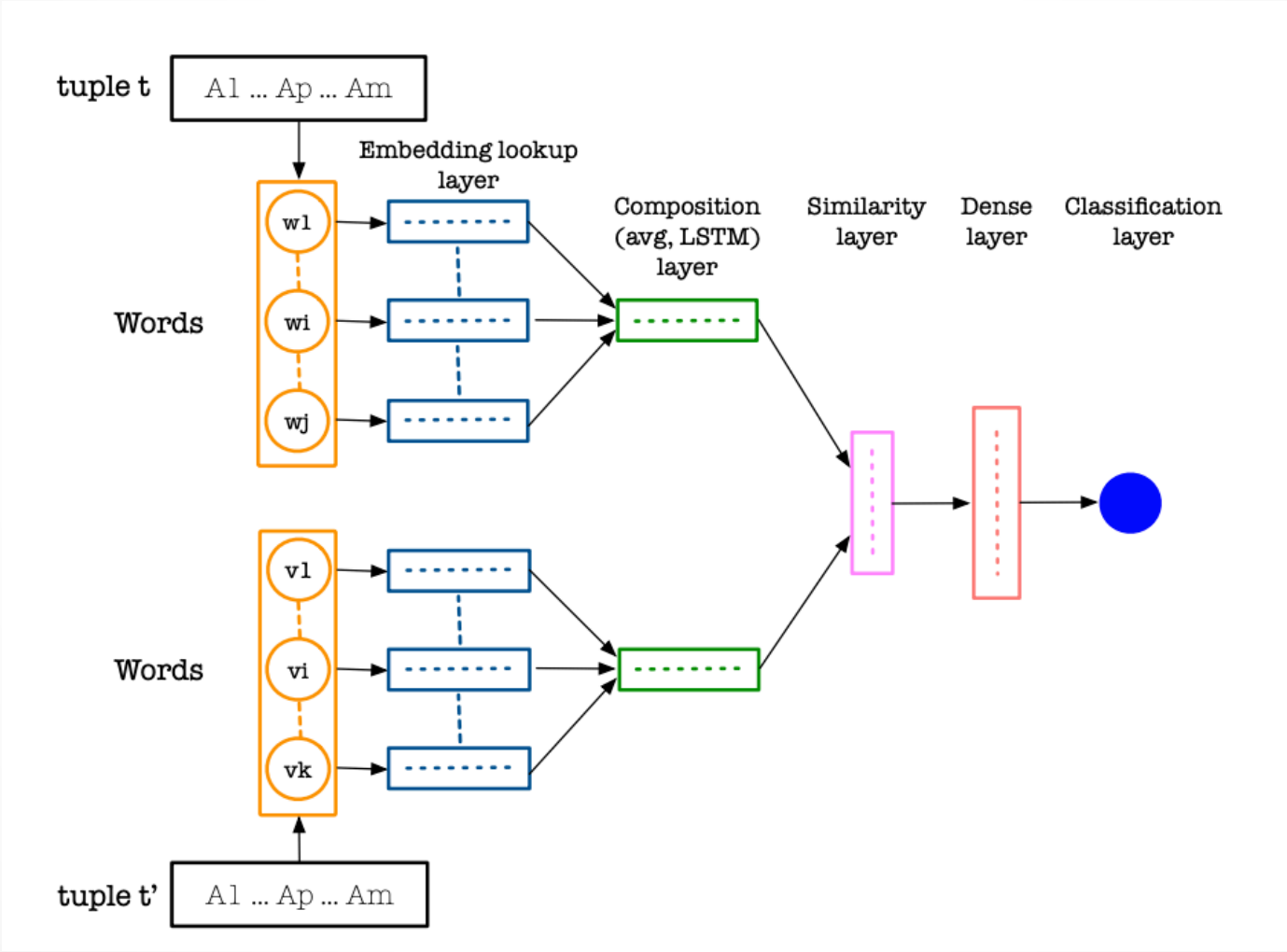
Table 9: F_1-score comparison for different deep learning solutions and Magellan for different dataset types and sizes.

Dataset		SIF	RNN	Attention	Hybrid	Magellan
Structured	small	79.1	83.9	76.2	84.6	86.5
	large	91.0	91.8	91.8	91.5	91.3
Textual	small	35.1	39.4	56.8	62.8	43.6
	large	87.5	88.9	88.8	89.3	83.9
Textual w/o info. attr.	small	32.0	38.5	55.0	47.7	33.0
	large	86.1	86.9	88.4	89.7	80.2
Dirty	small	76.8	86.2	78.2	84.2	64.7
	large	85.6	89.6	90.3	90.0	72.3

Comparing Attributes



Comparing Entities



Results

Dataset	Magellan	DeepER	Published
Prod-WA	82.99	88.06	89.3 [26] (Crowd)
Prod-AG	87.68	96.029	62.2 [33] (ML)
Pub-DA	97.6	98.6	N/A
Pub-DS	98.84	97.67	92.1 [26] (Crowd)
Pub-DC	96.4	99.1	95.2 [17] (Crowd)
Rest-FZ	100	100	96.5 [26] (Crowd)