Master’s Colloquium

LEARNING DISTRIBUTED EMBEDDINGS FROM KNOWLEDGE BASE WITH FOCUS ON RELATION EXTRACTION

EHSAN KHODDAM MOHAMMADI

UNIVERSITÄT DES SAARLANDES

PROF. DIETRICH KLAKOW

UBIQUITOUS KNOWLEDGE PROCESSING

PROF. IRYNA GUREVYCH

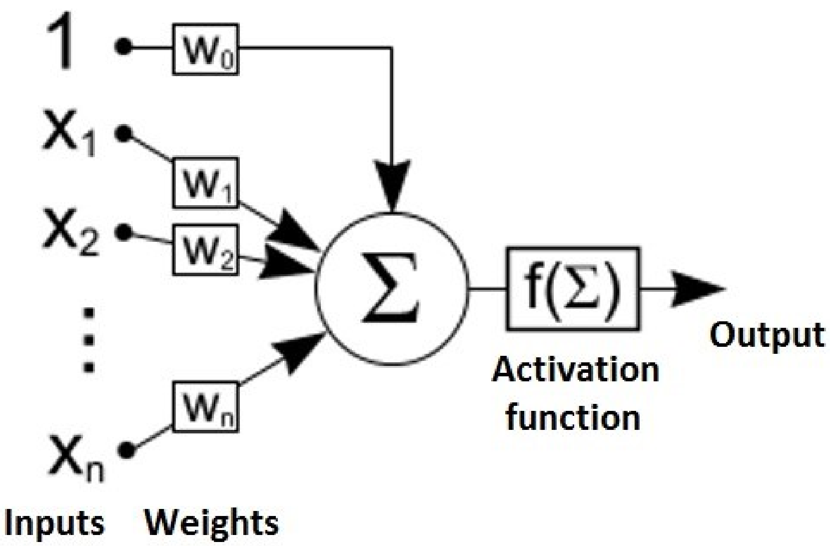
OUTLINE

1. **Artifical Neural Networks:**
   I. ANN architecture
   II. Learning parameters
2. **Representation Learning**
   I. Definition and motivation
   II. Different families of repr. Learning
   III. Learning Representation of a Knowledge Base
3. **Linking Text to a KB**
   I. Problem formulation
   II. Experiments
   III. Evaluation & Analysis
4. **Learning Word Features from Multiple Resources**
   I. Motivation
   II. Experiments
   III. Evaluation & Analysis

SECTION 1

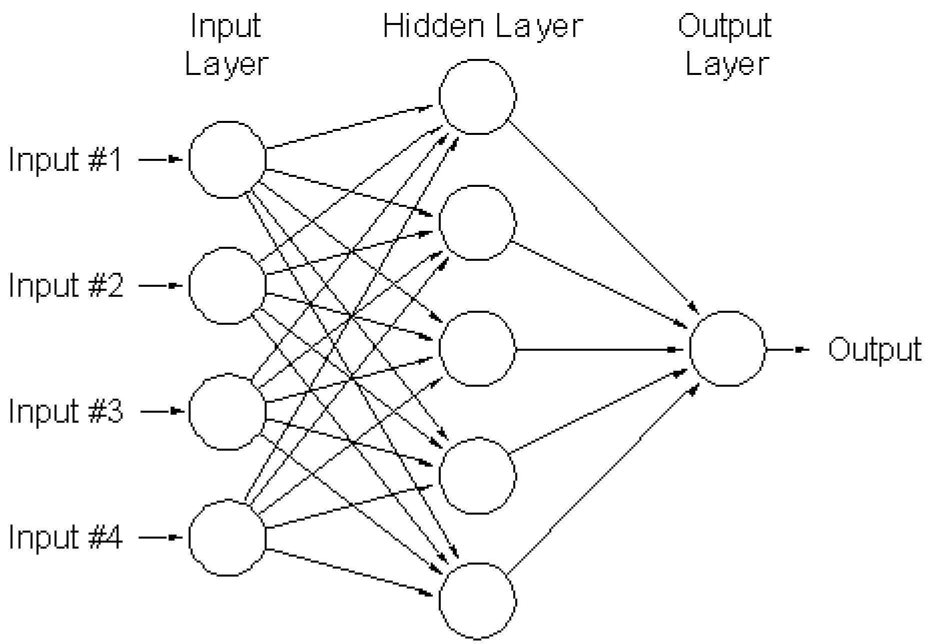
ARTIFICIAL NEURAL NETWORKS

SINGLE NEURON MODEL



- **Linear** $f(x) = ax + b$
- **Step** $f(x) = \begin{cases} 0, & x < \theta \\ 1, & x \geq \theta \end{cases}$
- **Tangent hyperbolic** $f(x) = tanh(x)$
- **Log-sigmoid** $f(x) = \frac{1}{1+e^{-x}}$

A NETWORK



Input
Layer
Hidden Layer
Output
Layer
Input #1
Input #2
Input #3
Input #4
Output

LEARNING ANN PARAMETERS

1. **Optimization Problem**
    - **Error**: difference between output of network and true value
    - Minimizing Error

2. **Single Supervised Layer**
    - Batch Learning
    - Online Learning
    - Stochastic Gradient Descent

3. **Other Layers**
    - Backpropagation: Backward propagation of errors

SECTION 2

REPRESENTATION LEARNING

REPRESENTATION LEARNING

- **Feature Engineering is labor-intensive:**
  - 90% of labor in industrial ML

- **Performance is dependent on feature engineering.**

Learn Features

WORD REPRESENTATION

- **A mathematical object , often a vector**
- **Each dimension corresponds to a feature.**
- **Each feature or subset of features has grammatical or semantical interpretation.**
- **Can be designed by hand or can be learned.**
- **Most of NLP applications can benefit from it:**
  - NER [Turian et al., ACL '10]
  - Chunking [Turian et al., ACL '10]
  - Parsing [Socher et al., ACL '13]
  - SRL [Collobert & Weston, ICML '08]
  - Language Models [Huang et al., ACL '12] [Bengio,JMLR '03]

FAMILY OF REPRESENTATIONS

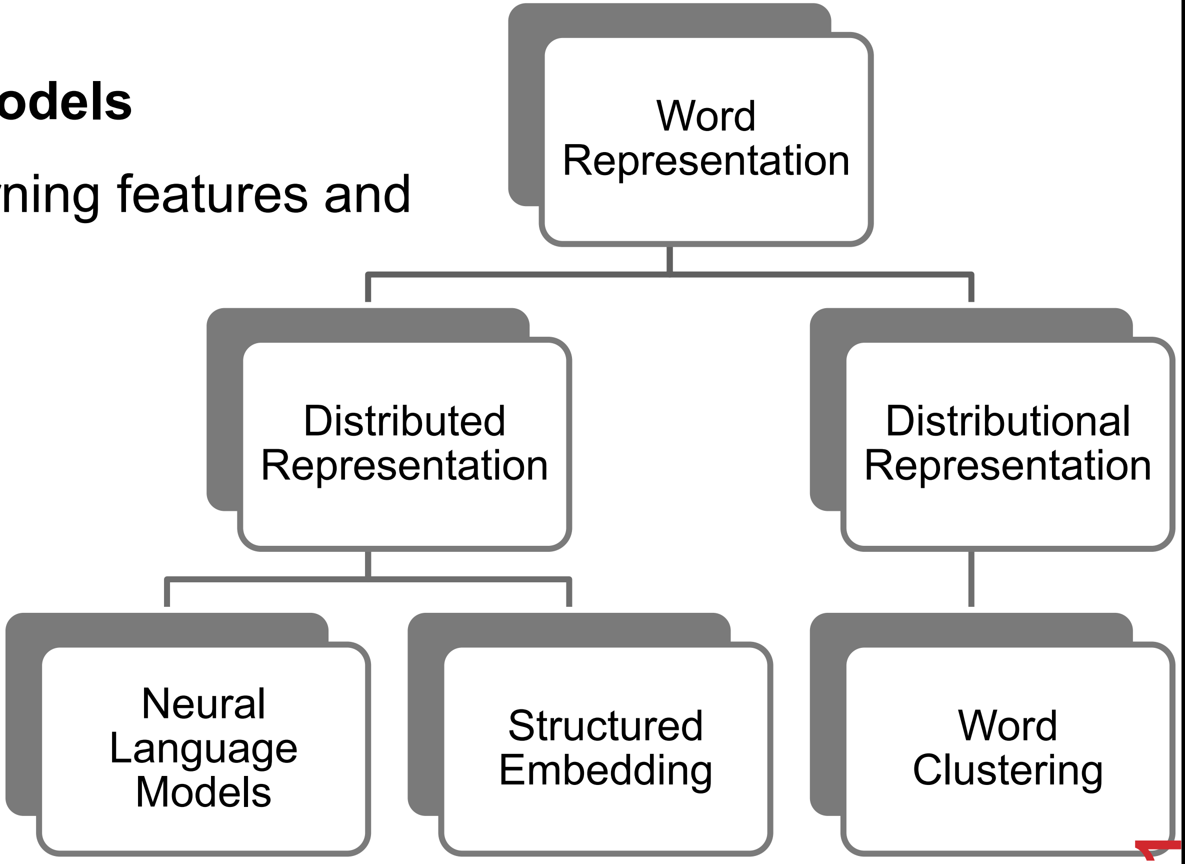
1. **Distributional Representations**
   - Features are windows around a word
   - Usually co-occurrence matrix
   - Dimensionality Reduction: SVD, LDA, PCA,…
   - Word Clustering:
     - Brown clustering

2. **Distributed Representation**
   - High Dimensional but sparse!
   - Real-values (can be binary too)
   - Compact (exponential number of clusters)

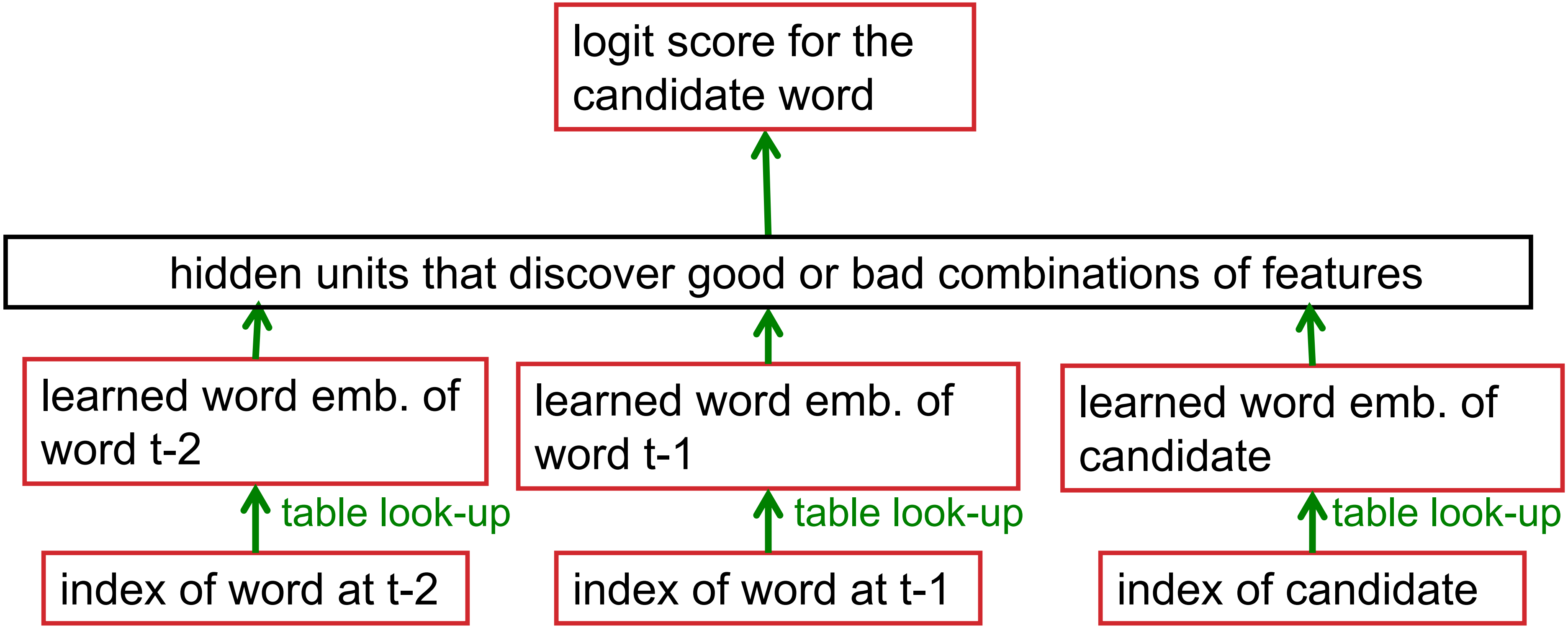
WORD EMBEDDINGS

- **Word representations induced by neural networks.**

- **Neural Language Models**
  - Joint task of learning features and a classification task

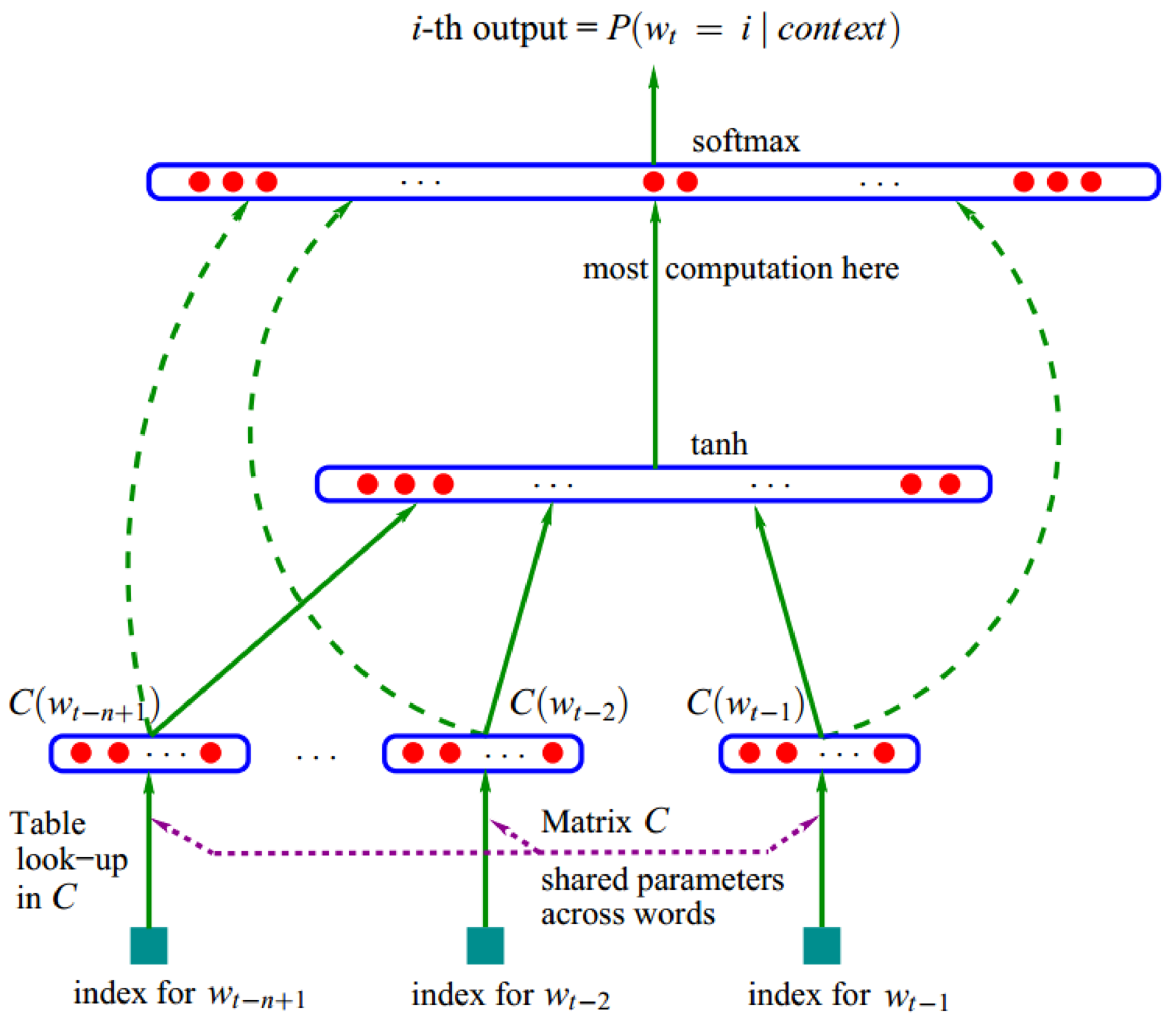


A TEMPLATE OF LEARNING WORD EMBEDDINGS



Courtesy: Geofry Hinton

A NEURAL LANGUAGE MODEL



BUT …

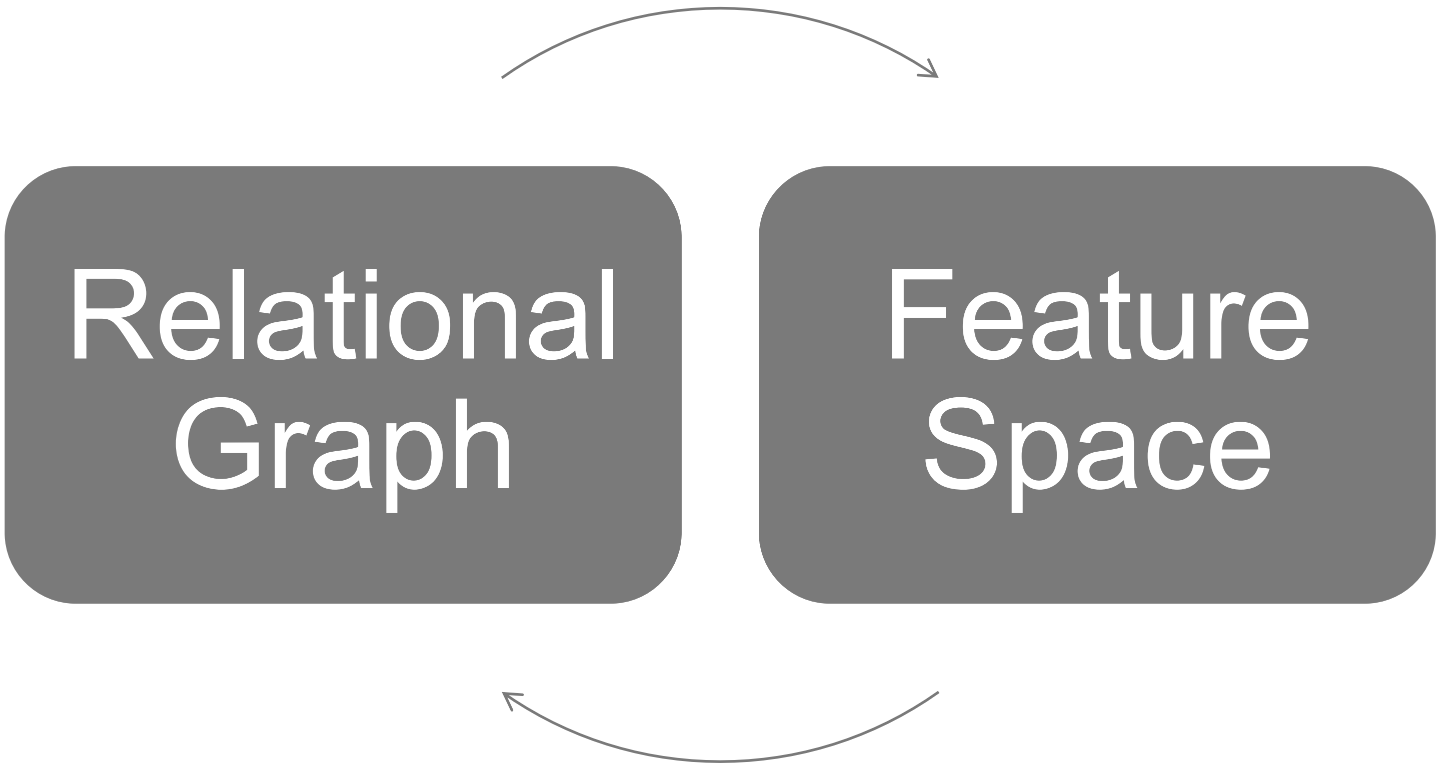
CORPUS IS NOT THE ONLY RESOURCE.

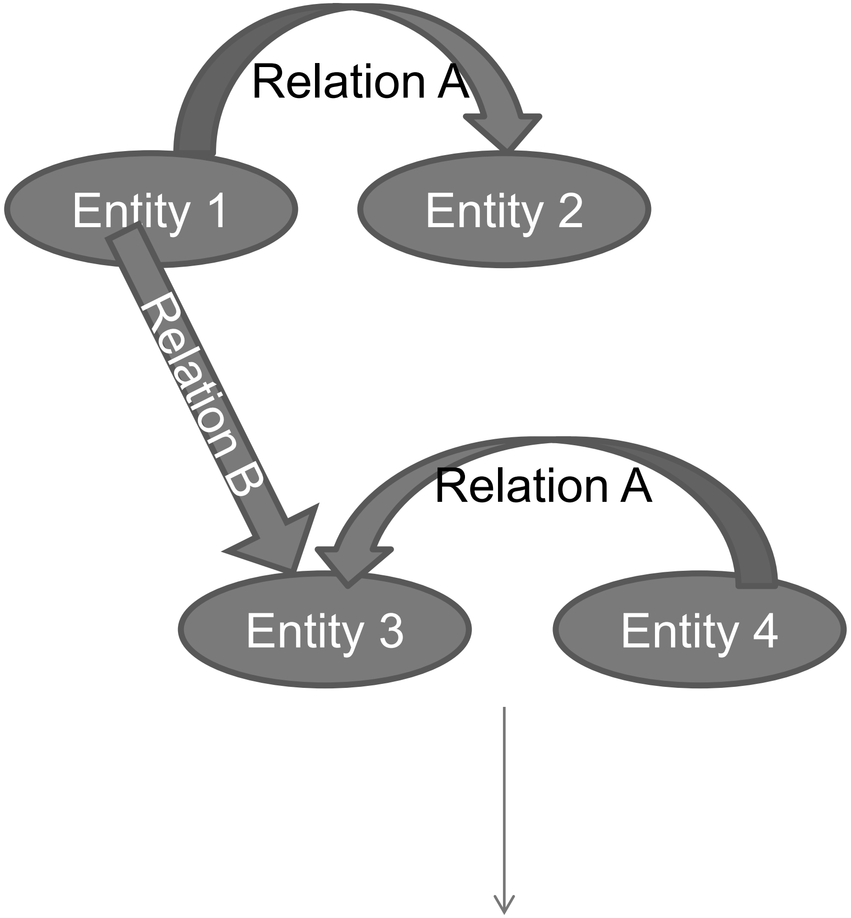
STRUCTURED RESOURCES

1. **Lexical Resources**
   1. WordNet
   2. GermaNet
   3. FrameNet
   4. …

2. **Knowledge Base:**
   1. Freebase
   2. Yago
   3. …

LEARNING REPRESENTATION OF KNOWLEDGE BASES





	f_1	f_2	f_3	f_4	f_5	f_6
Entity 1						
Entity 2						
Entity 3						
Entity 4						

TERMINOLOGY

1. **Word Embeddings:**

$$\boldsymbol{E_i \in \mathbb{R}^d}$$

2. **Relations:**

$$\boldsymbol{R_k = \left(R_k^{lhs}, R_k^{rhs}\right)}$$

3. **Triplet:**

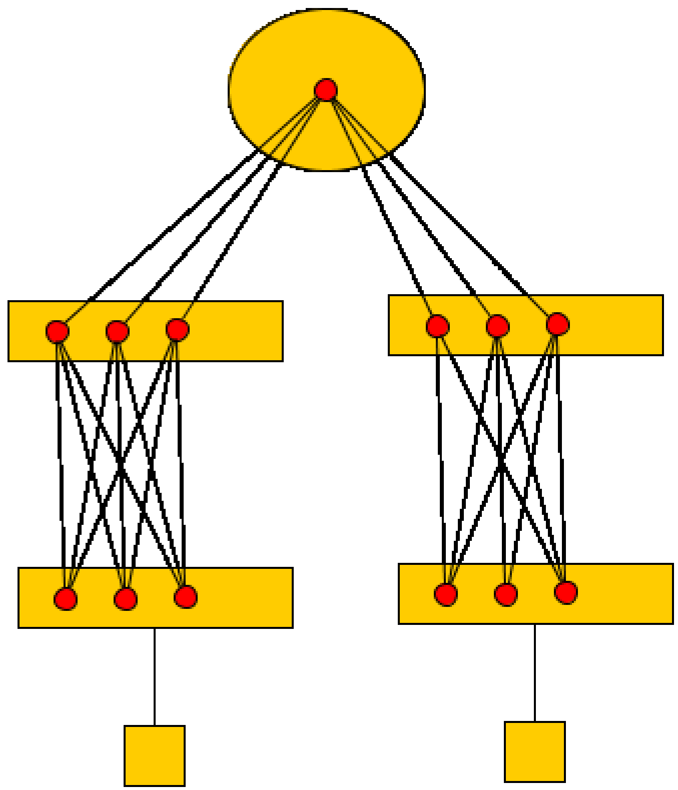
$$\boldsymbol{triplet_{i,j}^k = (E_i, R_k, E_j)}$$

4. **Ranking function**

$$\boldsymbol{f\left(triplet_{i,j}^k\right) = \left\|R_k^{lhs}\, E_i - R_k^{rhs}\, E_j\right\|_1}$$

A DISTRIBUTED MODEL FOR LEARNING STRUCTURED EMBEDDINGS (BORDES ET AL., 2011)

$f(triplet_{i,j}^{k}) - f(triplet_{neg}), 0)$



R_k^{lhs}

R_k^{rhs}

Right Embedding

Left Embedding

Entity Index

Entity Index

TRAINING

1. Randomly initialize embeddings and relation matrices

2. Generate random negative triplets: $\boldsymbol{triplet_{neg}}$

3. Rank training triplets and negative triplets

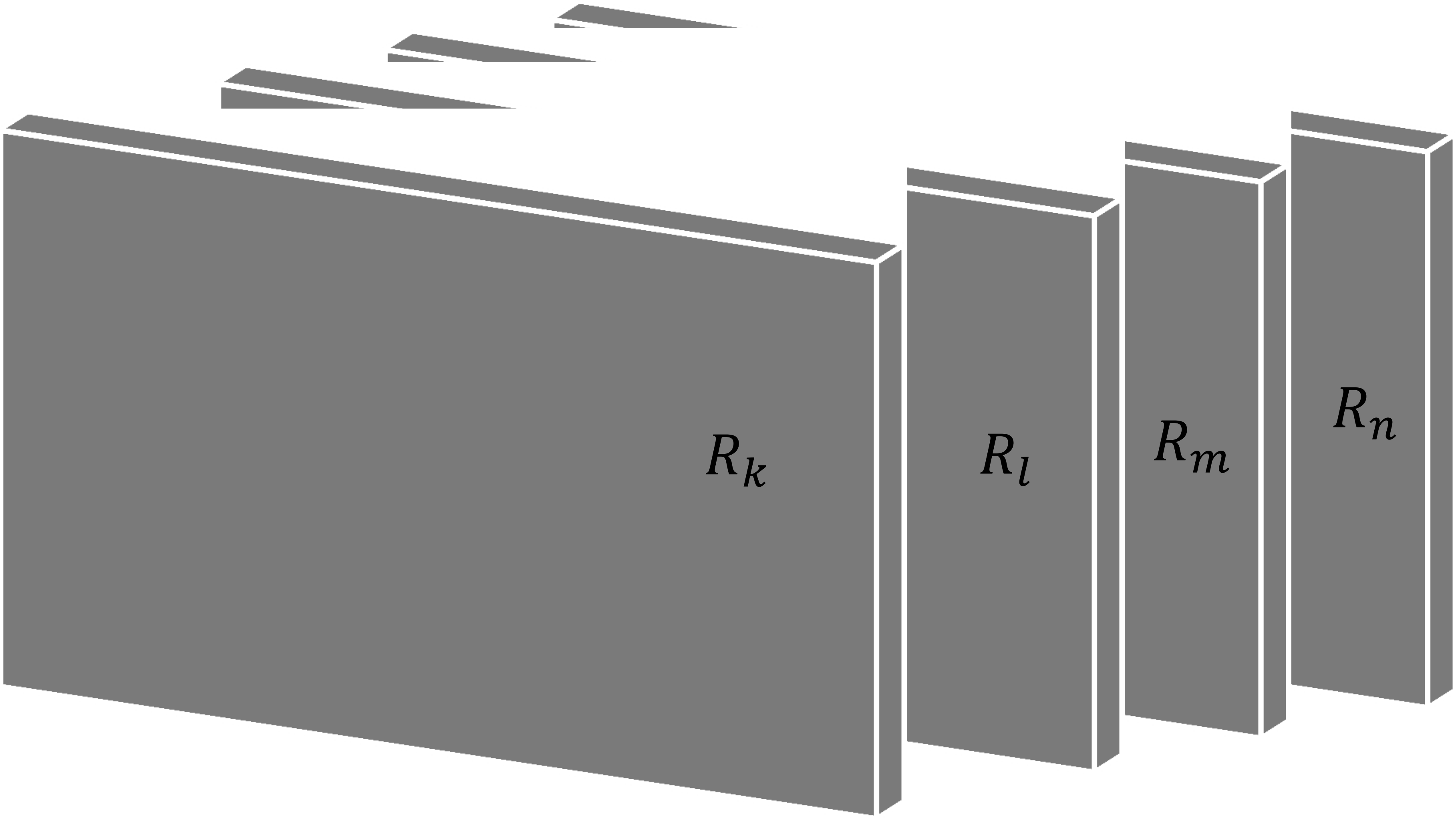
4. Using Stochastic Gradient Descent to tune embeddings and relations with large margin.

$$\boldsymbol{f\left(triplet_{i,j}^{k}\right) < f\left(triplet_{neg}\right) - 1}$$

$$\max(\ \boldsymbol{1} + \boldsymbol{f\left(triplet_{i,j}^{k}\right) - f\left(triplet_{neg}\right)}, \quad \boldsymbol{0})$$

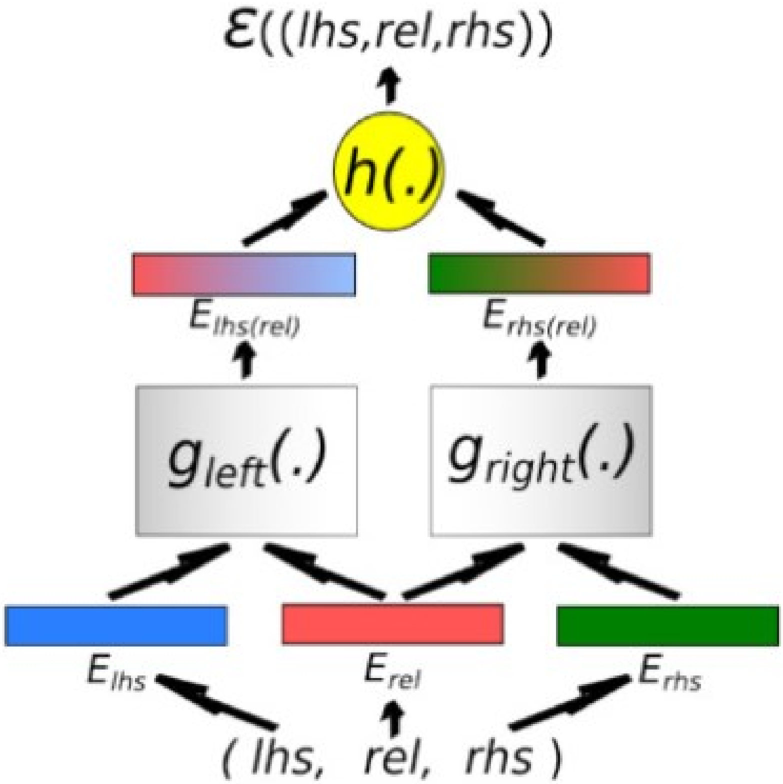
MULTI-TASK LEARNING

Learning each relation is a separate task.



Word Embeddings are shared among each task and transfer information

PARAMETER SHARING FOR RELATION EMBEDDINGS (BORDES ET AL., 2012)



SECTION 3

LINKING TEXT TO KNOWLEDGE BASE FOR RELATION DISCOVERY

RELATION DISCOVERY/EXTRACTION

Relation Extraction is the task of detecting and classifying semantic relationship between **n**amed **e**ntities (NE).

1. **DIRT:** frequent (dependency) surface patterns from text

2. **TextRunner:** bootstraping a classifier on small annotated data set

3. **Distant supervision**: using Freebase to annotate a text and learn a classifier on it

PROBLEM

1. Shallow, hand-crafted features extracted from text
2. Some facts are missed in Freebase
3. Features extracted from Freebase is 0/1

SOLUTION

1. Learning and using word embeddings
2. Learning and using entity/relation embeddings
3. Populating Freebase

TASK DESCRIPTION

Bordes et al.

Given a KB $\boldsymbol{F}$ learn features of entities and relations in F such that it enables us to discover new relations among the entities.

Us

Given a Corpus $\boldsymbol{C}$ and a KB $\boldsymbol{F}$ learn features of entities and relations in $\boldsymbol{F}$ such that it enables us to discover new relations among the entities.

AVAILABLE INFORMATION

Freebase

/GOVERNMENT/POSITION_HELD/PRESIDENT (MIR-HOSSEIN MOUSAVI, IRAN)

Corpus

NYTimes corpus, parsed, POS and NE tagged, Only sentences with two NEs

Problem

Two different formalization, we need to unify this information

UNIFIED FORMALIZATION

- **Represent any desired contextual feature in form of predicate-argument**
- **Introducing auxiliary predicates for different type of features**

- **Contextual features are**
  - Type of NEs: *LOC, PER, ORG,…*
  - Dependency role of a word: *DOBJ, NSUBJ,…*
  - Head of a dependency path: *president, head,…*

Mir-Hossein Mousavi, president of Iran, said . . .

↓

PATH#APPOS|->APPOS->PRESIDENT->PREP->OF->POBJ->|POBJ

↓

HAS_TYPE (Iran, LOC)
HAS_TYPE(Mousavi, PER)

HAS_DEP_ROLE (Iran, POBJ)
HAS_DEP_ROLE (Mousavi, APPOS)

PRESIDENT (Mousavi, Iran)

+

/Government/Position_held/President (Mousavi, Iran)

HAS_TRIGGER (/Government/Position_held/President, PRESIDENT)

EXPERIMENTS

1.	**KB**	only freebase relations, Bordes et al. settings
2.	**KB+Trigger**	freebase relations and surface patterns
3.	**Text+KB\Trigger**	all features except surface patterns
4.	**Text+KB**	all features

EVALUATION

Idea

Use embeddings to predict unseen relations

Given: Two NEs
Task: Rank Relations

EVALUATION

Dataset	Feature Type		Micro	Macro
Experiment 1	KB	mean	71.11	78.39
		median	31.0	72.73
		r@100	67.05	62.79
Experiment 2	KB+Trigger	mean	639.67	534.72
		median	544.0	503.20
		r@100	20.92	19.85
Experiment 3	All \Trigger	mean	59.63	61.59
		median	25.50	56.49
		r@100	73.85	73.88
Experiment 4	All	mean	6.72	57.13
		median	2.0	55.71
		r@100	98.85	76.88

ANALYSIS

Per Relation Evaluation

Relation	Frequency		KB	KB+Trigger	Text+KB\Trigger	Text+KB
		mean	88.65	789.27	83.27	2.55
		median	60.0	797.0	64.00	1.0
NA	2000	r@100	61.55	14.10	62.79	99.95
		mean	49.37	577.79	28.85	3.60
		median	9.0	468.00	5.00	2.00
/location/containedby	688	r@100	78.34	20.34	89.39	99.85
		mean	39.03	531.25	25.09	6.68
		median	12.50	410.5	8.0	4.0
/people/person/place_lived	132	r@100	84.84	25.75	93.18	100.0
		mean	47.00	359.48	16.43	2.83
		median	7.5	105.5	3.0	2.0
/person/company	124	r@100	79.83	50	95.16	100.0
		mean	47.95	433.36	15.60	3.92
		median	15.0	315.5	4.0	2.0
/deceased_person/place_of_death	80	r@100	78.75	36.25	95.00	100.0
		mean	77.71	444.57	34.00	33.14
		median	31.00	413.0	37.0	20.0
/people/person/ethnicity	7	r@100	57.14	28.57	100.0	85.71
		mean	89.25	331.25	51.75	52.50
		median	77.0	295.5	24.5	22.5
/music/composer/compositions	5	r@100	50.00	0.0	75.00	75.00
		mean	46.33	354.00	135.66	71.0
		median	43.00	157.0	142.0	90.0
/book/book_edition/publisher	3	r@100	100.0	33.33	33.33	100.0
		mean	20.33	220.00	4.33	51.33
		median	8.0	179.0	2.0	18.0
/people/person/religion	3	r@100	100.0	33.33	100.0	66.66

CONCLUSION

- **A formalization and process proposed which incorporates information from text to KB to learn entity/relation embeddings from both resources.**

- **Joint *Text+KB* embeddings perform much better than *KB* embeddings in predicting unseen relations among entities.**

- **We can induce new relations: this model is not only able to predict Freebase relations among entities but with using *Triggers* or surface patterns can predict relations that is not mentioned in it.**

SECTION 3

ENTITY LINKING AMONG MULTIPLE LEXICAL RESOURCES

MOTIVATION

1. **In previous section we were predicting semantic relations among NEs by learning representation of NEs and relations.**
2. **We can do the same for any semantic relation among words in general.**
3. **Several advantages:**
   - Induced word features can be used in almost every other NLP task: parsing, tagging, …
   - We can learn word embeddings from multiple resources with different perspective
   - We can link resources together or learn multi-lingual word embeddings: machine translation, word sense disambiguation

TASK DESCRIPTION

Bordes et al.

Given a lexical resource $\boldsymbol{L}$ learn features of entities and relations in $\boldsymbol{L}$ such that it enables us to discover new relations among the entities.

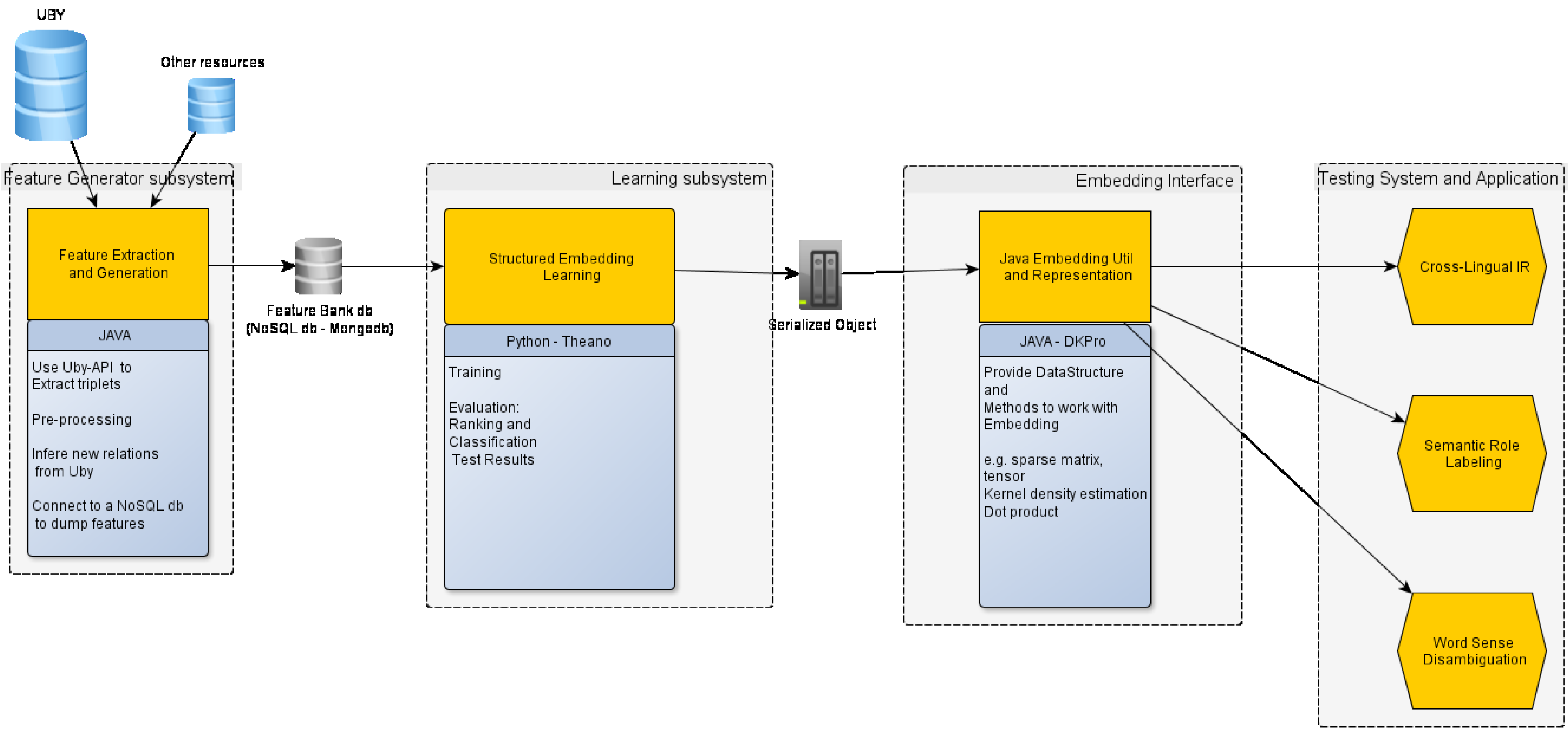
Us

Given a set of lexical resources $\boldsymbol{L_1}, \boldsymbol{L_2}, \dots, \boldsymbol{L_i}, \dots, \boldsymbol{L_j}, \dots, \boldsymbol{L_n}$ and a set of cross-resource relations $\boldsymbol{L_i} \rightarrow \boldsymbol{L_j}$ learn features of entities and relations in $\boldsymbol{L_i}$ and $\boldsymbol{L_j}$ such that it enables us to discover new relations among the entities.

BI-LINGUAL WORD EMBEDDINGS

- **Relations:**
  - Semantic relations: meronymy, holonymy,…
- **Entities:**
  - Word senses
- **Learning WordNet-GermaNet embeddings**
  - WordNet triples
  - GermaNet triples
  - ILI cross-lingual triples

SYSTEM ARCHITECTURE



INTRINSIC EVALUATION

Idea

Use embeddings to predict missing entities

Given: An entity and a relation
Task: Rank the missing entity

MODEL COMPARISON

Intrinsic Evaluation (Ranking Score Performance)

Dataset	#relations	#entities		Micro	Macro
GN SE	16	64025	mean	1003.59	3739.85
			median	5.0	2213.37
			global	84.23	72.49
GN SME-Bil	16	64025	mean	407.90	308.01
			median	10.0	54.18
			global	81.18	69.85
WN SE	23	148976	mean	148.72	623.10
			median	5.0	4.69
			global	92.10	89.86
WN SME-Bil	23	148976	mean	128.82	511.21
			median	10.0	26.63
			global	84.14	75.57
WN-GN SE (WN held out)	32	213002	mean	293.16	1356.30
			median	5.0	5.10
			global	91.19	88.95
WN-GN SME-Bil(WN held out)	32	213002	mean	124.85	331.82
			median	11.0	33.86
			global	82.91	73.55
WN-GN SE (GN held out)	32	213002	mean	3031.44	15470.56
			median	7.0	10080.5
			global	80.87	70.313
WN-GN SME-Bil(GN held out)	32	213002	mean	984.79	1021.37
			median	40.0	428.90
			global	64.16	55.98
WordNet-GermaNet-DD SME-Bil (WN held out)	32	213002	mean	166.18	466.91
			median	18.0	55.41
			global	77.07	65.082
WordNet-GermaNet-DD SME-Bil (GN held out)	32	213002	mean	932.49	719.47
			median	56.0	175.56
			global	59.22	50.84

EXTRINSIC EVALUATION

Idea

Use embeddings to predict semantic similarity of word pairs judged by humans

Given: a pair of words
Task: predict their similarity

WORD-PAIR SIMILARITY

Word-pair Similarity Performance for English

Dataset		WN-SE	WN-GN-SE	WN-SME-Bil	WN-GN-SME-Bil	WN-GN-SME-Bil-DD	HLBL	Turian et al.	Klementiev et a
RG-65	P	0.682	0.666	0.703	0.833	0.725	-0.115	0.233	-0.380
	S	0.769	0.741	0.741	0.811	0.825	-.083	0.118	-0.398
MC-30	P	0.611	0.644	0.601	0.740	0.599	-0.363	0.150	-0.768
	S	0.720	0.648	0.756	0.846	0.954	-.450	-0.198	-0.522
WS-353	P	0.181	0.206	0.239	0.246	0.238	0.233	0.236	0.029
	S	0.093	0.146	0.185	0.224	0.201	0.197	0.210	0.040
YangPowers-130	P	0.482	0.637	0.584	0.627	0.610	-0.130	-0.076	0.154
	S	0.401	0.472	0.406	0.553	0.533	-0.186	-0.116	0.113

Word-pair Similarity Performance for German

Dataset		GN-SE	WN-GN-SE	GN-SME-Bil	WN-GN-SME-Bil	WN-GN-SME-Bil-DD	Klementiev et al.
ZG222	P	-0.010	0.156	0.073	0.130	0.196	0.107
	S	-0.125	0.234	0.152	0.175	0.111	0.152
Gur30	P	0.865	0.984	0.185	0.287	0.301	-0.887
	S	1.0	1.0	-0.500	-0.500	0.500	-1.0
Gur350	P	-0.085	0.063	0.185	0.188	0.127	0.187
	S	-0.157	0.009	0.172	0.194	0.142	0.142
Gur65	P	0.800	0.558	-0.572	0.485	-0.166	0.233
	S	0.800	0.800	-0.8	0.399	0.200	0.200

CONCLUSION

- A framework is proposed here to learn word features from multiple resources.
- The motivation behind the framework has been empirically tested by learning bi-lingual word embeddings from WordNet and GermaNet.
- Bi-lingual structured embeddings have been evaluated on several word-pair similarity datasets and shown significant improvement over mono-lingual and other type of corpus-based embeddings.

FUTURE WORK

1. Our models don’t have probabilistic output which hurts their ability to be used along other NLP models in cascade or for evaluation.

2. More types of features can be used for all of our models to examine their effect.

THANK YOU, LSV!

THANK YOU, UKP!

THANK YOU FOR YOUR ATTENTION.

QUESTION?