Dependency Link Embeddings: Continuous Representations of Syntactic Substructures

Mohit Bansal

TTI-Chicago

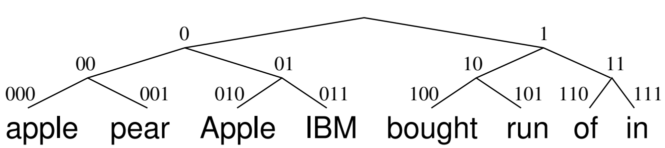
Motivation

(n-ary word cluster/embedding features)

prefix6 → (110010 , 000101)

prefix4 → (1100 , 0001)

ate (parent) → *apple* (child)



apple → 00010100010

prefix4

prefix6

(Lei et al., 2014; Chen and Manning, 2014)

Motivation

(n-ary)

w[i] → (0.6 , -0.9)

ate *apple*
(parent) (child)

ate → [0.6, 0.9, 0.3, -1.0, 0.1, -0.7]

apple → [-0.9, 0.1, -0.3, 0.5, 0.1, 0.6]

Motivation

(n-ary)

w[i]	→	(0.6 , -0.9)
prefix6	→	(110010 , 000101)
prefix4	→	(1100 , 0001)

ate → *apple*

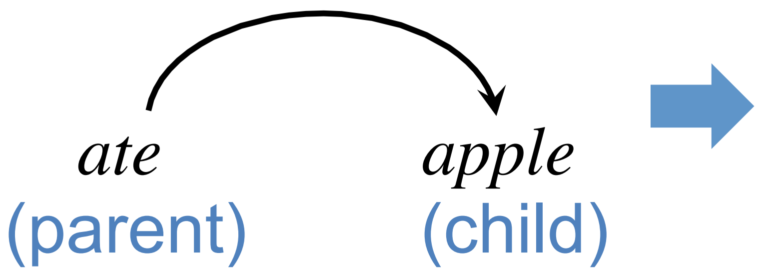
(parent) (child)

Baseline	Cluster-based
ht,mt	hc4,mc4
hw,mw	hc6,mc6
hw,ht,mt	hc$\star$,mc$\star$
hw,ht,mw	hc4,mt
ht,mw,mt	ht,mc4
hw,mw,mt	hc6,mt
hw,ht,mw,mt	ht,mc6
...	hc4,mw
	hw,mc4
	...
ht,mt,st	hc4,mc4,sc4
ht,mt,gt	hc6,mc6,sc6
...	ht,mc4,sc4
	hc4,mc4,gc4
	...

(McDonald et al., 2005; Koo et al., 2008)

Motivation

- Train link embeddings on tons of auto-parsed data (*min-count thresholded to get only the popular links*)



$$\begin{bmatrix} 0.6 \\ -0.2 \\ 0.9 \\ 0.3 \\ -0.4 \\ 0.5 \end{bmatrix}$$

Motivation

(unary)

w[i] → 0.4

ate *apple*

(parent) (child)

ate—apple → [0.4, 0.2, -0.3, 0.4, 0.1, 0.7]

Motivation

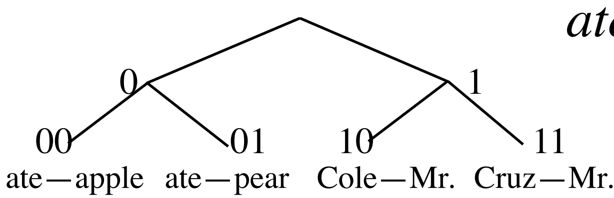
(unary)

prefix6 → 101010

prefix4 → 1101

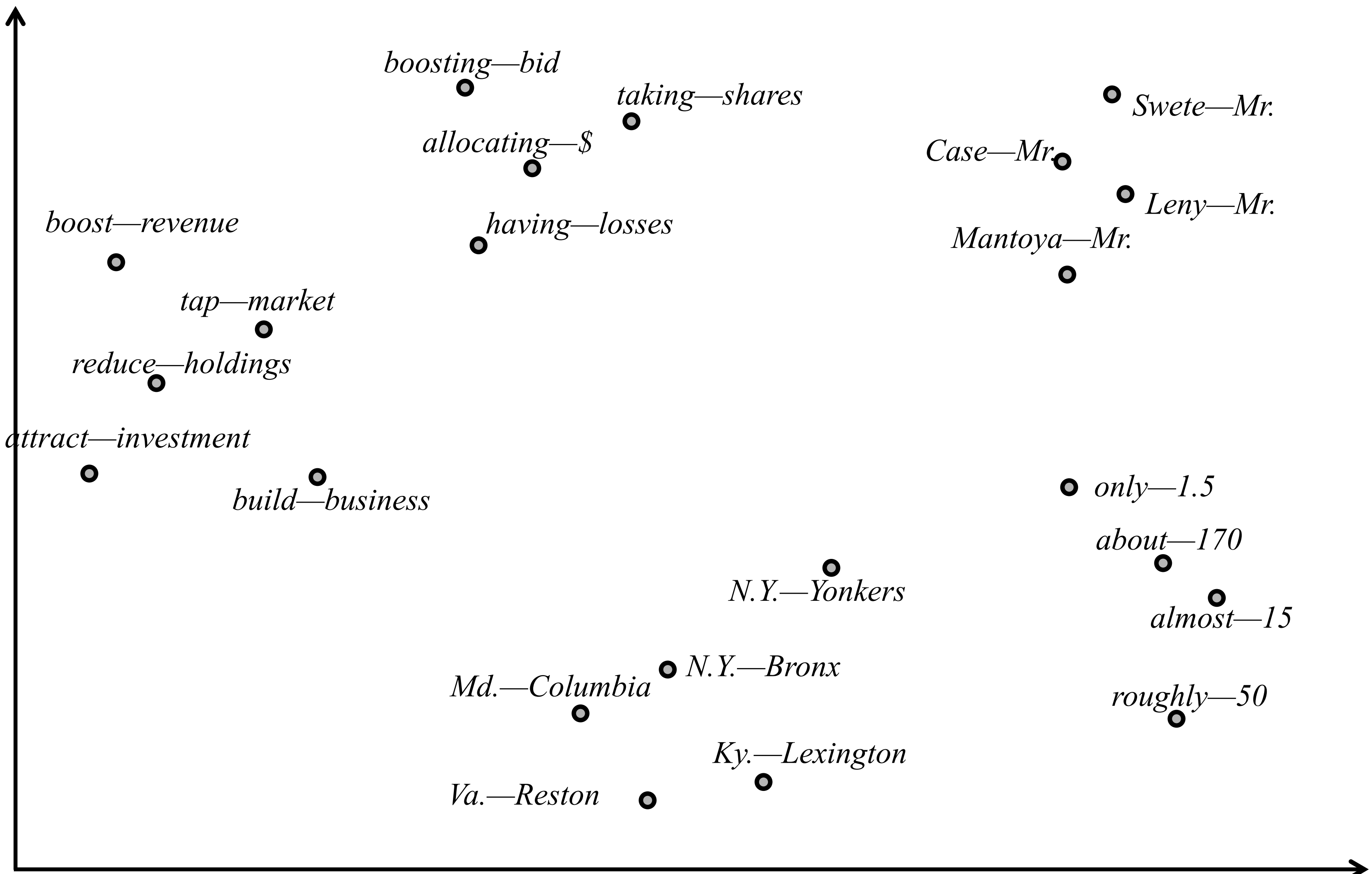
ate (parent)

apple (child)

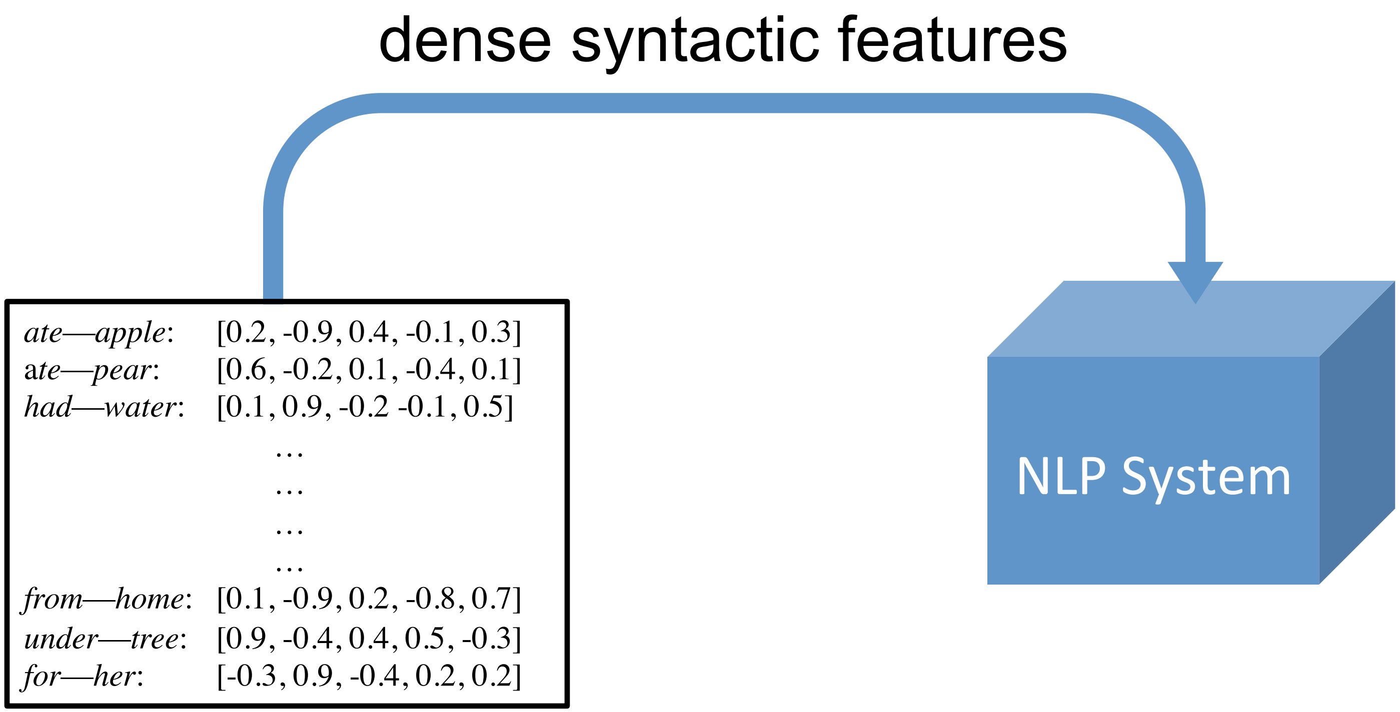


ate — apple → $\underbrace{\underbrace{1101}_{\text{prefix4}}11}_{\text{prefix6}}100010$

Motivation


boosting—bid
taking—shares
allocating—$
having—losses
boost—revenue
tap—market
reduce—holdings
attract—investment
build—business
Swete—Mr.
Case—Mr.
Leny—Mr.
Mantoya—Mr.
only—1.5
about—170
almost—15
roughly—50
N.Y.—Yonkers
N.Y.—Bronx
Md.—Columbia
Ky.—Lexington
Va.—Reston

Motivation

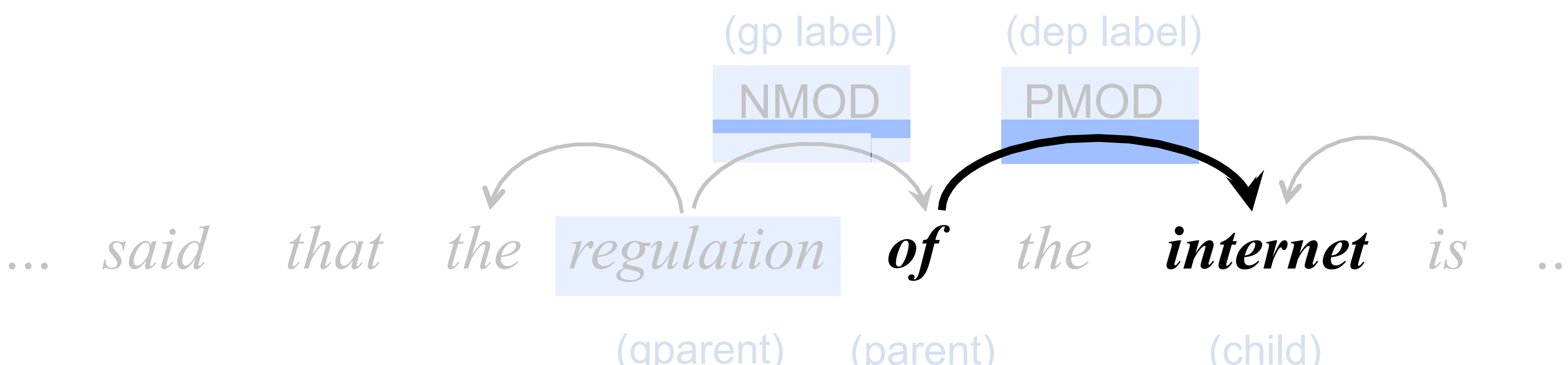

dense syntactic features
ate—apple: [0.2, -0.9, 0.4, -0.1, 0.3]
ate—pear: [0.6, -0.2, 0.1, -0.4, 0.1]
had—water: [0.1, 0.9, -0.2 -0.1, 0.5]
…
…
…
…
from—home: [0.1, -0.9, 0.2, -0.8, 0.7]
under—tree: [0.9, -0.4, 0.4, 0.5, -0.3]
for—her: [-0.3, 0.9, -0.4, 0.2, 0.2]
NLP System

Motivation

- Much fewer, simpler ***unary*** features vs. millions of template-based, ***n-ary*** word-cluster features
- Directly work with higher-order, substructure embeddings that task factors on, and their hidden relationships
- Portable as off-the-shelf, dense, syntactic features (instead of lexicalized or word embedding features)

Training

- Parse a large corpus with baseline parser
- Tuples consist of a dependency link and its context
(Bansal et al., 2014; Levy and Goldberg, 2014)



Training

- Tuples consist of a dependency link and its context

$$d_{<\mathrm{D}>} \quad gl_{<\mathrm{GL}>} \quad p\text{—}c \quad l_{<\mathrm{L}>} \quad d_{<\mathrm{D}>}$$

dist.	gp label	parent—child link	dep label	dist.
[*-2*	$NMOD_{<GL>}$	***of—internet***	$PMOD_{<L>}$	*-2*]

Training

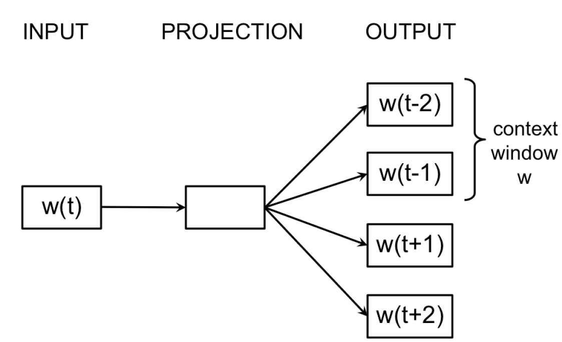
- Tuples consist of a dependency link and its context

$$gl_{\text{<GL>}} \quad gp\text{—}p \quad p\text{—}c \quad d_{\text{<D>}} \quad l_{\text{<L>}}$$

gp label | gp-parent link | parent-child link | dist. | dep label

$$\underbrace{\left[\mathit{NMOD}_{<GL>} \quad \mathit{regulation}\text{—}\mathit{of} \quad \boldsymbol{of\text{—}internet} \quad -2 \quad \mathit{PMOD}_{<L>}\right]}_{\text{context window}}$$

- Run SKIP-gram model to predict context
- **Threshold of 4 to get a vocab of just 92K (and then backoff to word/unk features)**



Training

- Tried other context versions, e.g., **for link to be able to predict the individual words (for backing off to word-level shared info)**

gp label	gp-parent	parent	parent-child	child	dist.	dep label
[$NMOD_{<GL>}$	*regulation—of*	*of*	***of—internet***	*internet*	*-2*	$PMOD_{<L>}$]

context window

Clusters

- Finds useful groups and subtle distinctions at link level

[*N.Y.–Yonkers, Md.–Columbia, N.Y.–Bronx, Va.–Reston, Ky.–Lexington, Mich.–Kalamazoo, Calif.–Calabasas,* ...]

[*boost–revenue, tap–markets, take–losses, launch–fight, reduce–holdings, terminate–contract, identify–bidders,* ...]

[*boosting–bid, meeting–schedules, obtaining–order, having–losses, completing–review, governing–industry,* ...]

[*began–Meanwhile, was–Since, are–Often, would–Now, had–During, were–Over, was–Late, have–Until,* ...]

Clusters

- Finds useful groups and subtle distinctions at link level

[*says–mean, adds–may, explains–have, contend–has, recalls–had, figures–is, asserted–is, notes–would, …*]

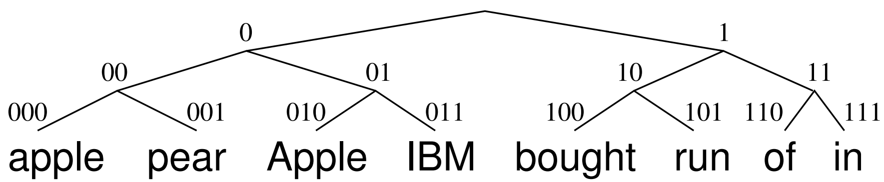
[*would–Based, is–Besides, was–Like, is–From, are–Despite, said–Besides, says–Despite, reported–As, …*]

[*Catsimatidis–Mr., Swete–Mr., Case–Mr., Montoya–Mr., Byerlein–Mr., Heard–Mr., Leny–Mr., Graham–Mrs., …*]

[*only–1.5, about–170, nearly–eight, approximately–10, almost–15, some–80, Only–two, about–23, roughly–50, …*]

Dependency Parsing Features

- Brown cluster **n-ary** features (Koo et al., 2008):



$$\textit{apple} \rightarrow \underbrace{\underbrace{0001}_{\text{prefix4}}01}_{\text{prefix6}}00010$$

prefix6	→	110010	000101
prefix4	→	1100	0001
tag	→	VBD	NN
		ate (parent)	*apple* (child)

Baseline	Cluster-based
ht,mt	hc4,mc4
hw,mw	hc6,mc6
hw,ht,mt	hc*,mc*
hw,ht,mw	hc4,mt
ht,mw,mt	ht,mc4
hw,mw,mt	hc6,mt
hw,ht,mw,mt	ht,mc6
...	hc4,mw
	hw,mc4
	...
ht,mt,st	hc4,mc4,sc4
ht,mt,gt	hc6,mc6,sc6
...	ht,mc4,sc4
	hc4,mc4,gc4
	...

(McDonald et al., 2005; Koo et al., 2008)

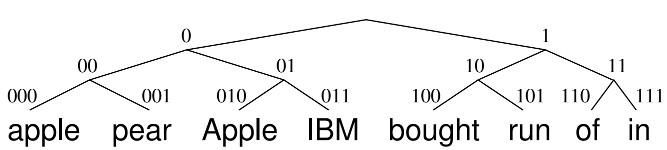
Dependency Parsing Features

- Word embedding **n-ary** features (Bansal et al., 2014):
  - Per-dimension bucket features:

ate → [0.2 0.7 -0.6 0.9]
apple → [0.6 -0.1 0.7 0.2]

  - Hierarchical clustering (bit string) features:

$$linkage(E, \text{‘ward’}, \text{‘euclidean’})$$



apple → $\underbrace{\underbrace{0001}_{\text{prefix4}}01}_{\text{prefix6}}00010$

Baseline	Cluster-based
ht,mt	hc4,mc4
hw,mw	hc6,mc6
hw,ht,mt	hc∗,mc∗
hw,ht,mw	hc4,mt
ht,mw,mt	ht,mc4
hw,mw,mt	hc6,mt
hw,ht,mw,mt	ht,mc6
...	hc4,mw
	hw,mc4
	...
ht,mt,st	hc4,mc4,sc4
ht,mt,gt	hc6,mc6,sc6
...	ht,mc4,sc4
	hc4,mc4,gc4
	...

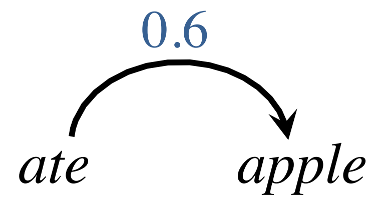
(McDonald et al., 2005; Koo et al., 2008)

Dependency Parsing Features

- Link embedding **unary** features (this work):
  - Per-dimension bucket features:

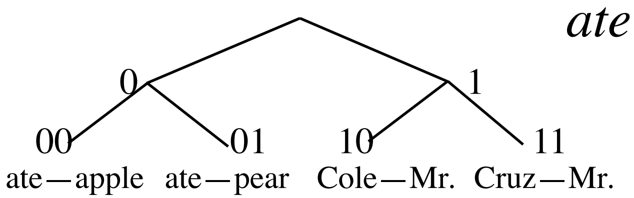
dim=3

ate—apple → [-0.5 0.3 $\boxed{0.6}$ 0.8]



  - Hierarchical clustering (bit string) features:

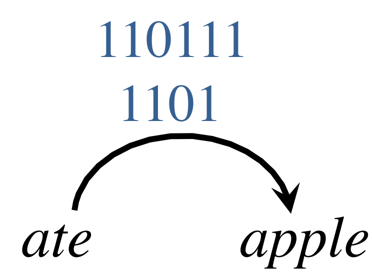
$$linkage(E, \text{‘ward’}, \text{‘euclidean’})$$



ate—apple → 11011100010

prefix4

prefix6



Feature Comparison (Memory, Speed)

- Setup: MSTParser (2^{nd} order), standard data splits, parameters, preprocessing, threshold (Bansal et al., 2014)
- Much fewer features compared to n-ary, word-based
- Quicker to train these SKIP-based link features

System	Number of features	
Baseline	5M	
BROWN	13M	(2.5 days)
Bansal et al. (2014)	30M	
Bucket	**15K**	(15 mins.)
Bit-string	1M	(1 day)

Dependency Parsing Results

- Stat-equal improvements as Brown and stat-significant stacking, at much fewer, simpler, quicker features

System	Test
Baseline	91.9
+ Brown	92.7
+ Bucket	92.3
+ Bit-string	92.6
+ Brown + Bucket	93.0
+ Brown + Bit-string	**93.1**

- Similar improvements and stacking for out-of-domain
- Allows practical, accurate per-dimension features

Off-the-shelf Results

- Portable as simple, dense, syntactic features

- E.g., on constituent reranking, stat-equal improvements as global reranking features and stat-signif. stacking

(Bansal and Klein, 2011)

Parsing Model	Test F1	Test EX
Baseline (1-best)	90.2	37.3
Baseline (log $p(t\|w)$)	89.9	37.3
+ Config	91.1	40.6
+ Bit-string	90.9	40.6
+ Config + Bit-string	**91.4**	**42.0**

Conclusion

- Dense dependency link embeddings allow simpler, fewer (unary) features in dependency parsing
- Get similar improvements to n-ary template-based word cluster/embedding features
- Portable as useful, dense, syntactic features to downstream tasks, e.g., constituent reranking

Current/Future Work

- Export to other extrinsic tasks, e.g., sentence classification or initial units in vector-space composition
- Newer, better representation learning tools, e.g., deeper NNs, GloVe (Pennington et al., 2014)
- Training on larger quantities of automatically-parsed data
- Other back-off approaches for unknown links, e.g., tag-based links (*tag1—tag2*) and one-sided links (*UNK-word or UNK-tag*)
- Compare to approaches like relation matrices on dependency labels, concatenation+SVD of head and argument vectors

Thank you!

Data (link embeddings and features) at:

ttic.uchicago.edu/~mbansal/codedata/linkEmbeddings-skiplink.zip