Tailoring Continuous Word Representations for Dependency Parsing

Mohit Bansal, Kevin Gimpel, and Karen Livescu

TTI-Chicago

Questions We Want to Answer

- What kind of embeddings will help dependency parsing (in-domain and out-of-domain)?
- How can we convert embeddings to parsing features?
- Are there good intrinsic measures of embedding quality?

Representation Models

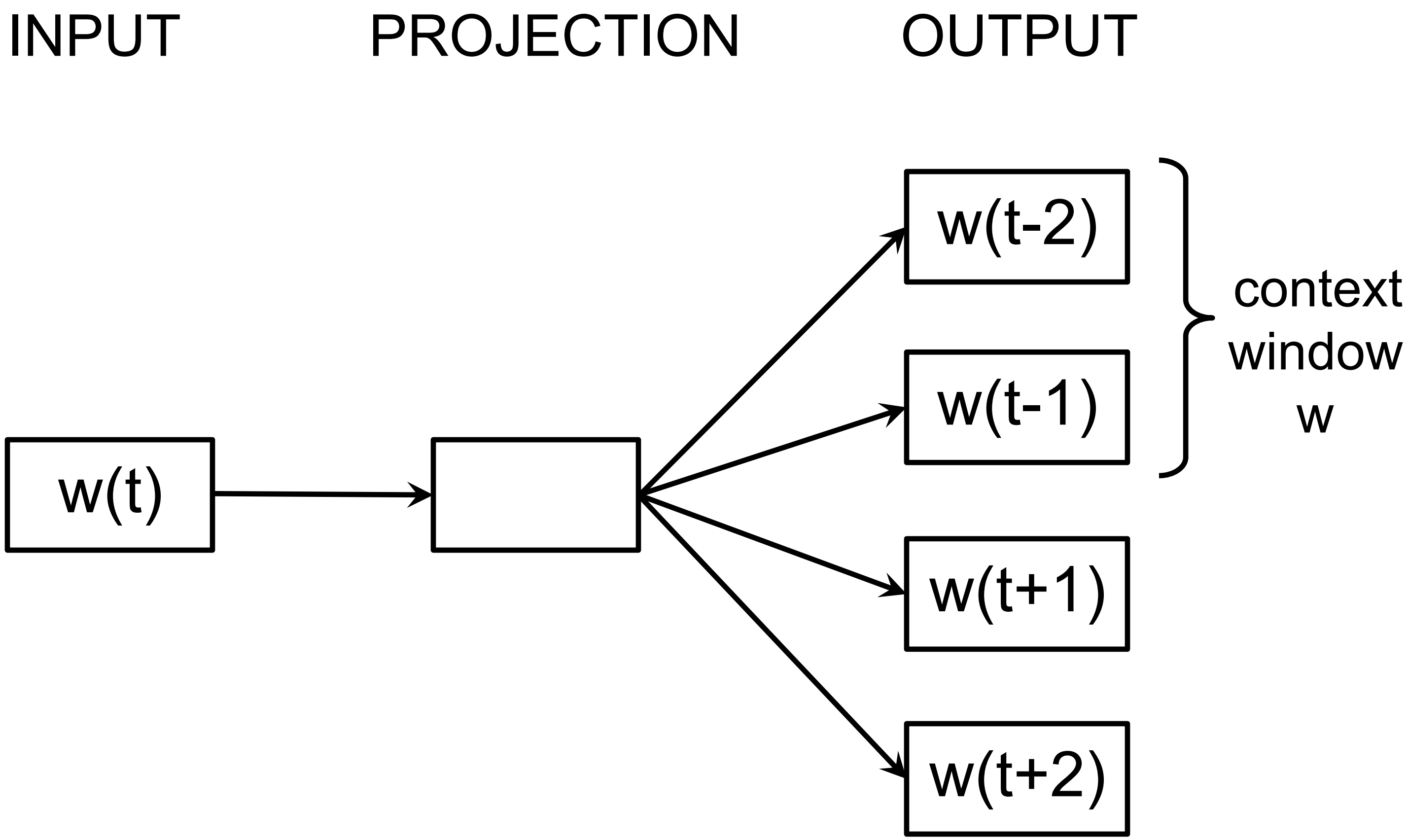
- BROWN (Brown et al., 1992)
- SENNA (Collobert et al., 2011, 2008)
- TURIAN (Turian et al., 2010)
- HUANG (Huang et al., 2012)
- SKIP (Mikolov et al., 2013)

discrete

continuous

SKIP

Mikolov et al., 2013



Few mins. vs. days/weeks/months!!

Syntactically Tailored Embeddings

- Context window size (SKIP)
  - Smaller window → syntactic/functional similarity
  - Larger window → topical similarity

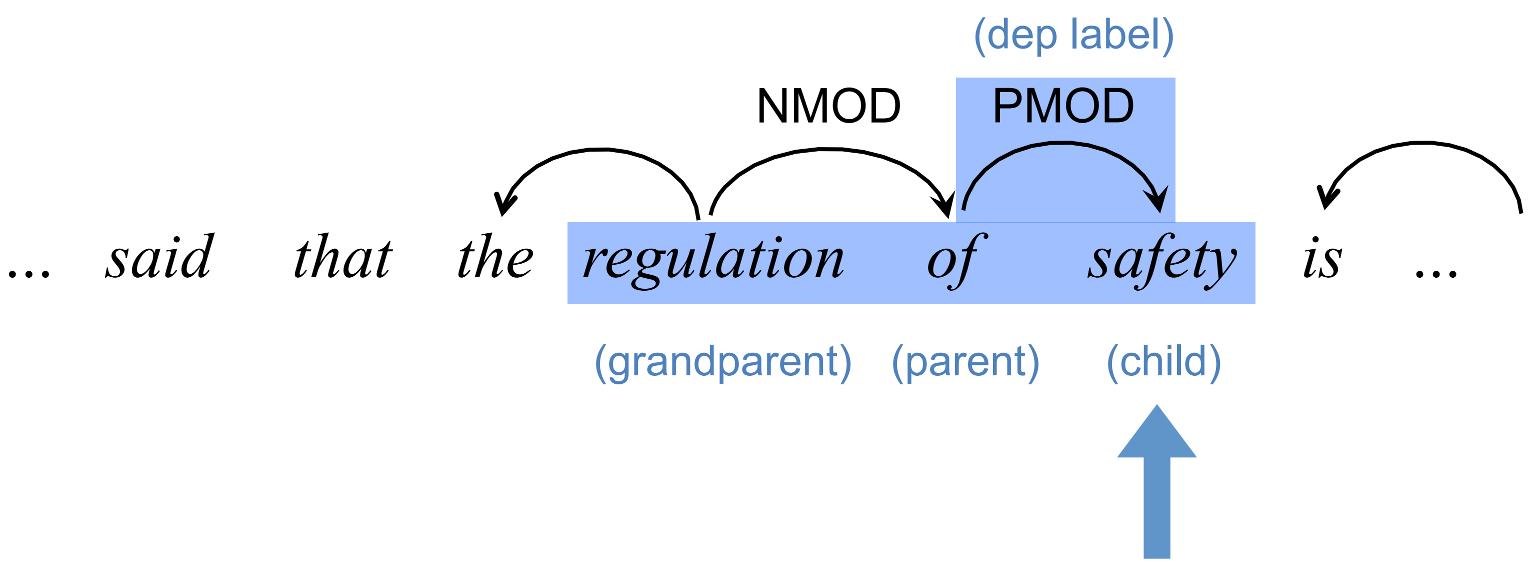
The morning ***flight*** *at the JFK airport was delayed*

context window

- Similar effect in distributional representations (Lin and Wu, 2009)

Syntactically Tailored Embeddings

- Syntactic context (SKIP$_{DEP}$)
  - Condition on dependency context instead of linear
  - First parse a large corpus with baseline parser:



Syntactically Tailored Embeddings

- Syntactic context ($\mathrm{SKIP}_{\mathrm{DEP}}$)
  - Condition on dependency context instead of linear
  - Then convert each dependency to a tuple:

dep label | grandparent | parent | child | dep label

$$\left[PMOD_{<L>} \quad regulation_{<G>} \quad of \quad safety \quad PMOD_{<L>} \right]$$

context windows

- Syntactic information in clustering, topic, semantic space models

(Sagae and Gordon, 2009; Haffari et al., 2011; Grave et al., 2013; Boyd-Graber and Blei, 2008; Pado and Lapata, 2007)

Cluster Examples

- SKIP, w = 10:

  [*attendant, takeoff, airport, carry-on, airplane, flown, landings, flew, fly, cabins,* ...]

  [*maternity, childbirth, clinic, physician, doctor, medical, health-care, day-care,* ...]

  [*transactions, equity, investors, capital, financing, stock, fund, purchases,* ...]

Cluster Examples

- SKIP, w = 1

[*Mr., Mrs., Ms., Prof., III, Jr., Dr.*]

[*Jeffrey, William, Dan, Robert, Stephen, Peter, John, Richard, ...*]

[*Portugal, Iran, Cuba, Ecuador, Greece, Thailand, Indonesia, ...*]

[*his, your, her, its, their, my, our*]

[*Your, Our, Its, My, His, Their, Her*]

[*truly, wildly, politically, financially, completely, potentially, ...*]

Intrinsic Evaluation

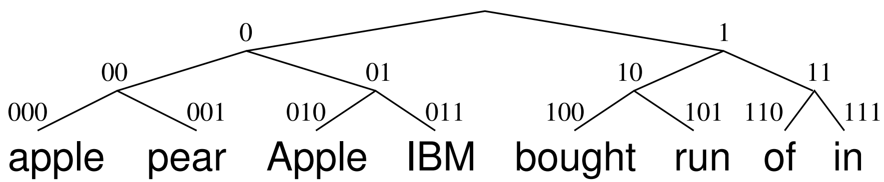
(Finkelstein et al., 2002)

Representation	SIM	TAG
BROWN	–	**89.3**
SENNA	49.8	85.2
HUANG	**62.6**	78.1
SKIP, $w = 10$	44.6	71.5
SKIP, $w = 5$	44.4	81.1
SKIP, $w = 1$	37.8	86.6
SKIP_{DEP}	34.6	88.3

Topical (SIM) Syntactic/ Functional (TAG)

Dependency Parsing Features

- Brown Cluster Features (Koo et al., 2008):



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

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

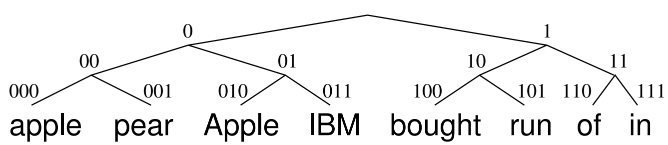
Dependency Parsing Features

- Continuous Representation Features:
  - 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 → $\underbrace{\underbrace{0001}_{\text{prefix4}}01}_{\text{prefix6}}00010$

Parsing Experiments

- Setup: MSTParser (2^{nd} order) w/ standard processing
- Per-dim bucket << Hierarchical clustering features:

System	Test
Baseline	92.0
SENNA (Buckets)	92.0
SENNA (Hier. Clustering)	92.3
HUANG (Buckets)	91.9
HUANG (Hier. Clustering)	92.4

Parsing Experiments

- Main WSJ results:

System	Test
Baseline	91.9
BROWN	**92.7**
SENNA	92.3
TURIAN	92.3
HUANG	92.4
SKIP	92.3
SKIP_{DEP}	**92.7**
Ensemble Results	
ALL – BROWN	92.9
ALL	93.0

(faster)

(complementary)

Parsing Experiments

- Main Web results:

System	Test Avg (5 domains)
Baseline	83.5
BROWN	84.2
SENNA	**84.3**
TURIAN	83.9
HUANG	84.1
SKIP	83.7
SKIP_{DEP}	84.1 (faster)
Ensemble Results	
ALL–BROWN	84.7
ALL	84.9

(complementary)

Correlation w/ Intrinsic Metrics

- Correlation only for variations of a single model

Representation	SIM	TAG	Parsing F1
SKIP, $w = 10$	44.6	71.5	92.70
SKIP, $w = 5$	44.4	81.1	92.86
SKIP, $w = 1$	37.8	86.6	92.94
SKIP_{DEP}	34.6	88.3	93.33

Topical

Syntactic/
Functional

Conclusion

- Improvements ~ Brown but with faster training
- Hierarchical clustering >> bucket (per-dim) features
- Syntactic context helps
- Intrinsic metrics ~correlate with parsing accuracy

Thank you!

Data (dependency embeddings and features) at:

ttic.uchicago.edu/~mbansal