Size (and Domain) Matters: Evaluating Semantic Word Space Representations for Biomedical Text

Pontus Stenetorp*, Hubert Soyer, Sampo Pyysalo
Sophia Ananiadou and Takashi Chikayama

http://pontus.stenetorp.se

<pontus@stenetorp.se>

Aizawa Laboratory
University of Tokyo

4th of September 2012

Keyquote

Keyquote

"A word is characterized by the company it keeps"
– J.R. Firth (1957)

Keyquote

"A word is characterized by the company it keeps"

– J.R. Firth (1957)

Aspects

Keyquote

"A word is characterized by the company it keeps"

– J.R. Firth (1957)

Aspects

- Captures syntax

Keyquote

"A word is characterized by the company it keeps"
– J.R. Firth (1957)

Aspects

- Captures syntax
- But also aspects of semantics

Word Representations: What Are They?

Word Representations: What Are They?

The Good

Word Representations: What Are They?

The Good

- Easy to apply

Word Representations: What Are They?

The Good

- Easy to apply
- Performance boost:

Word Representations: What Are They?

The Good

- Easy to apply
- Performance boost:
  - Named Entity Recognition: $+7.0$ F_1 (Turian et al. 2010)

Word Representations: What Are They?

The Good

- Easy to apply
- Performance boost:
  - Named Entity Recognition: $+7.0$ F_1 (Turian et al. 2010)
  - Parsing: $+1.4\%$ F_1 (Koo and Collins 2008)

Word Representations: What Are They?

The Good

- Easy to apply
- Performance boost:
  - Named Entity Recognition: +7.0 F_1 (Turian et al. 2010)
  - Parsing: +1.4% F_1 (Koo and Collins 2008)

The Bad

Word Representations: What Are They?

The Good

- Easy to apply
- Performance boost:
  - Named Entity Recognition: $+7.0\ F_1$ (Turian et al. 2010)
  - Parsing: $+1.4\%\ F_1$ (Koo and Collins 2008)

The Bad

- No given "intuitive" interpretation

Word Representations: What Are They?

The Good

- Easy to apply
- Performance boost:
  - Named Entity Recognition: $+7.0\ F_1$ (Turian et al. 2010)
  - Parsing: $+1.4\%\ F_1$ (Koo and Collins 2008)

The Bad

- No given "intuitive" interpretation
- Many possible variations

Word Representations: What Are They?

The Good

- Easy to apply
- Performance boost:
  - Named Entity Recognition: $+7.0$ F_1 (Turian et al. 2010)
  - Parsing: $+1.4\%$ F_1 (Koo and Collins 2008)

The Bad

- No given "intuitive" interpretation
- Many possible variations
- Generation computationally costly

Word Representations: How Do They Look?

Word Representations: How Do They Look?

Format

Word Representations: How Do They Look?

Format

- Vector-based

Word Representations: How Do They Look?

Format

- Vector-based
- But details vary...

Word Representations: How Do They Look?

Format

- Vector-based
- But details vary...

For the Protein “fibrin”

Word Representations: How Do They Look?

Format

- Vector-based
- But details vary...

For the Protein "fibrin"

Hard (One-hot) $\begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$

Word Representations: How Do They Look?

Format

- Vector-based
- But details vary...

For the Protein "fibrin"

Hard (One-hot)	$\begin{bmatrix} 0 & ... & 0 & 1 & 0 & ... & 0 \end{bmatrix}$
Hard (Dense)	$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}$

Word Representations: How Do They Look?

Format

- Vector-based
- But details vary...

For the Protein "fibrin"

Hard (One-hot)	$\begin{bmatrix} 0 & ... & 0 & 1 & 0 & ... & 0 \end{bmatrix}$
Hard (Dense)	$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}$
Distance	$\begin{bmatrix} -0.19 & -0.49 & 0.40 & 0.39 & 0.01 & ... \end{bmatrix}$

Word Representations: How Do They Look?

Format

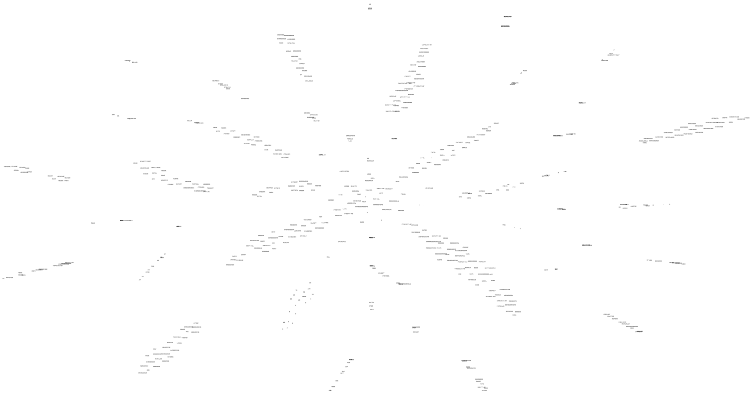
- Vector-based
- But details vary...

For the Protein "fibrin"

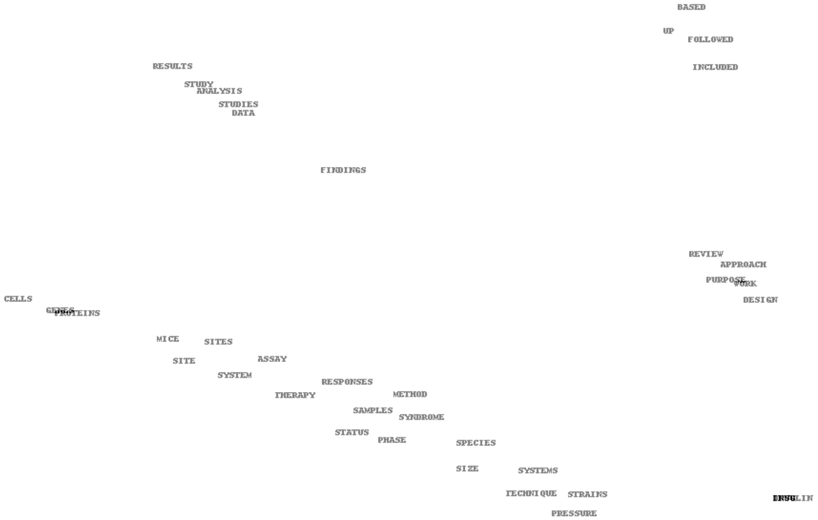
Hard (One-hot)	$\begin{bmatrix} 0 & ... & 0 & 1 & 0 & ... & 0 \end{bmatrix}$
Hard (Dense)	$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}$
Distance	$\begin{bmatrix} -0.19 & -0.49 & 0.40 & 0.39 & 0.01 & ... \end{bmatrix}$
Distance (Sparse)	$\begin{bmatrix} ... & 0 & 0.15 & 0 & ... & 0 & 0.22 & 0.28 & 0 \end{bmatrix}$

Word Representations: How Do They Look?

Word Representations: How Do They Look?



Word Representations: How Do They Look?



Open Questions?

Open Questions?

Newswire

Open Questions?

Newswire

- Helps for a great variety of tasks

Open Questions?

Newswire

- Helps for a great variety of tasks
- "Off-the-rack" representations available

Open Questions?

Newswire

- Helps for a great variety of tasks
- "Off-the-rack" representations available

BioNLP

Open Questions?

Newswire

- Helps for a great variety of tasks
- "Off-the-rack" representations available

BioNLP

- Can word representations boost performance?

Open Questions?

Newswire

- Helps for a great variety of tasks
- “Off-the-rack” representations available

BioNLP

- Can word representations boost performance?
- Are in-domain word representations necessary?

Word Representations Used

Name	Method	Domain	Src.	Dim.	Publication
Brown-news-100	Brown	news	63M	100	
Brown-news-320	Brown	news	63M	320	
Brown-news-1000	Brown	news	63M	1,000	
Brown-news-3200	Brown	news	63M	3,200	Turian et al.
HLBL-news	HLBL	news	63M	100	
C&W-news-200d-0.1	C&W	news	63M	200	
C&W-news-50d-0.3	C&W	news	63M	50	
Google	K-means	web	10^{12}	1,000	Lin et al.
ClarkNE-bio	Clark-NE	bio	31M	45	McClosky et al.
Brown-bio-100	Brown	bio	13M	100	
Brown-bio-320	Brown	bio	13M	320	This study
Brown-bio-1000	Brown	bio	13M	1,000	

Task: Named Entity Recognition (NER)

Task: Named Entity Recognition (NER)

System Details

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)
- Single-class NER

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)
- Single-class NER
- Strong baseline NER model

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)
- Single-class NER
- Strong baseline NER model
- Add word representations and their combinations

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)
- Single-class NER
- Strong baseline NER model
- Add word representations and their combinations

Corpora

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)
- Single-class NER
- Strong baseline NER model
- Add word representations and their combinations

Corpora

- Anatomical Entity Mention (AnEM) (Ohta et al., 2012)

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)
- Single-class NER
- Strong baseline NER model
- Add word representations and their combinations

Corpora

- Anatomical Entity Mention (AnEM) (Ohta et al., 2012)
- BCII Gene Mention (BC2GM) (Smith et al., 2008)

Task: Named Entity Recognition (NER)

System Details

- System by Ratinov and Roth (2009)
- Single-class NER
- Strong baseline NER model
- Add word representations and their combinations

Corpora

- Anatomical Entity Mention (AnEM) (Ohta et al., 2012)
- BCII Gene Mention (BC2GM) (Smith et al., 2008)
- NCBI disease (NCBID) (Islamaj Dogan and Lu, 2012)

Task: Semantic Category Disambiguation (SCD)

Task: Semantic Category Disambiguation (SCD)

The BvrR / BvrS system is essential for Brucella abortus virulence.

Task: Semantic Category Disambiguation (SCD)

The BvrR / BvrS system is essential for Brucella abortus virulence .

2-comp-sys

Pro Pro +Regulation Organism Process

The BvrR / BvrS system is essential for Brucella abortus virulence .

Task: Semantic Category Disambiguation (SCD)

Task: Semantic Category Disambiguation (SCD)

Definition

Task: Semantic Category Disambiguation (SCD)

Definition

- Given a span, in a context, assign a semantic category

Task: Semantic Category Disambiguation (SCD)

Definition

- Given a span, in a context, assign a semantic category

Data

Task: Semantic Category Disambiguation (SCD)

Definition

- Given a span, in a context, assign a semantic category

Data

- Induce contexts using unambigous seed-words

Task: Semantic Category Disambiguation (SCD)

Definition

- Given a span, in a context, assign a semantic category

Data

- Induce contexts using unambigous seed-words
- Seed-words for 10 categories (McIntosh and Curran, 2009)

Task: Semantic Category Disambiguation (SCD)

Definition

- Given a span, in a context, assign a semantic category

Data

- Induce contexts using unambigous seed-words
- Seed-words for 10 categories (McIntosh and Curran, 2009)
- Blind the focus word

Task: Semantic Category Disambiguation (SCD)

Definition

- Given a span, in a context, assign a semantic category

Data

- Induce contexts using unambigous seed-words
- Seed-words for 10 categories (McIntosh and Curran, 2009)
- Blind the focus word
- Stratify semantic categories

Task: Semantic Category Disambiguation (SCD)

Definition

- Given a span, in a context, assign a semantic category

Data

- Induce contexts using unambigous seed-words
- Seed-words for 10 categories (McIntosh and Curran, 2009)
- Blind the focus word
- Stratify semantic categories
- Enables evaluation on large data

Task: Semantic Category Disambiguation (SCD)

Task: Semantic Category Disambiguation (SCD)

The effects of electric fields on and PANC1 cells .

Task: Semantic Category Disambiguation (SCD)

The effects of electric fields on $\underbrace{\qquad}$ and PANC1 cells .

The effects of electric fields on $\overbrace{\qquad}^{\text{Cell line}}$ and PANC1 cells .

Named Entity Recognition: Results

Model	Dataset			
	AnEM	BC2GM	NCBID	μ
Baseline	56.19	78.07	68.02	67.43

Named Entity Recognition: Results

Model	Dataset			
	AnEM	BC2GM	NCBID	μ
Baseline	56.19	78.07	68.02	67.43
Brown-news-100	55.73	78.77	69.30	67.93
Brown-news-320	54.56	78.19	69.10	67.29
Brown-news-1000	56.70	78.43	68.99	68.04
Brown-news-3200	55.31	78.94	69.08	67.78

Named Entity Recognition: Results

Model	Dataset			
	AnEM	BC2GM	NCBID	μ
Baseline	56.19	78.07	68.02	67.43
Brown-news-100	55.73	78.77	69.30	67.93
Brown-news-320	54.56	78.19	69.10	67.29
Brown-news-1000	56.70	78.43	68.99	68.04
Brown-news-3200	55.31	78.94	69.08	67.78
HLBL-news	58.52	79.46	69.06	69.02

Named Entity Recognition: Results

Model	Dataset			
	AnEM	BC2GM	NCBID	μ
Baseline	56.19	78.07	68.02	67.43
Brown-news-100	55.73	78.77	69.30	67.93
Brown-news-320	54.56	78.19	69.10	67.29
Brown-news-1000	56.70	78.43	68.99	68.04
Brown-news-3200	55.31	78.94	69.08	67.78
HLBL-news	58.52	79.46	69.06	69.02
Google	61.66	79.43	69.68	70.26

Named Entity Recognition: Results

Model	Dataset			
	AnEM	BC2GM	NCBID	μ
Baseline	56.19	78.07	68.02	67.43
Brown-news-100	55.73	78.77	69.30	67.93
Brown-news-320	54.56	78.19	69.10	67.29
Brown-news-1000	56.70	78.43	68.99	68.04
Brown-news-3200	55.31	78.94	69.08	67.78
HLBL-news	58.52	79.46	69.06	69.02
Google	61.66	79.43	69.68	70.26
ClarkNE-bio	52.81	78.81	67.83	66.48

Named Entity Recognition: Results

Model	Dataset			
	AnEM	BC2GM	NCBID	μ
Baseline	56.19	78.07	68.02	67.43
Brown-news-100	55.73	78.77	69.30	67.93
Brown-news-320	54.56	78.19	69.10	67.29
Brown-news-1000	56.70	78.43	68.99	68.04
Brown-news-3200	55.31	78.94	69.08	67.78
HLBL-news	58.52	79.46	69.06	69.02
Google	61.66	79.43	69.68	70.26
ClarkNE-bio	52.81	78.81	67.83	66.48
Brown-bio-100	57.20	78.79	69.75	68.58
Brown-bio-150	52.97	79.08	69.72	67.26
Brown-bio-320	57.96	79.33	69.50	68.93
Brown-bio-500	62.11	79.81	69.88	70.60
Brown-bio-1000	**62.49**	80.04	**70.52**	71.02

Named Entity Recognition: Results

Model	Dataset AnEM	BC2GM	NCBID	μ
Baseline	56.19	78.07	68.02	67.43
Brown-news-100	55.73	78.77	69.30	67.93
Brown-news-320	54.56	78.19	69.10	67.29
Brown-news-1000	56.70	78.43	68.99	68.04
Brown-news-3200	55.31	78.94	69.08	67.78
HLBL-news	58.52	79.46	69.06	69.02
Google	61.66	79.43	69.68	70.26
ClarkNE-bio	52.81	78.81	67.83	66.48
Brown-bio-100	57.20	78.79	69.75	68.58
Brown-bio-150	52.97	79.08	69.72	67.26
Brown-bio-320	57.96	79.33	69.50	68.93
Brown-bio-500	62.11	79.81	69.88	70.60
Brown-bio-1000	**62.49**	80.04	**70.52**	71.02
HLBL-news+Brown-news-1000	58.91	79.31	69.22	69.15
HLBL-news+Brown-bio-1000	62.10	**80.32**	70.15	**71.16**

Semantic Category Disambiguation: Results

Model	Accuracy
BoW	67.61
Comp	71.59

Semantic Category Disambiguation: Results

Model	Accuracy
BoW	67.61
Comp	71.59
Comp-Brown-news-100	71.54
Comp-Brown-news-320	71.93
Comp-Brown-news-1000	71.45
Comp-Brown-news-3200	71.42

Semantic Category Disambiguation: Results

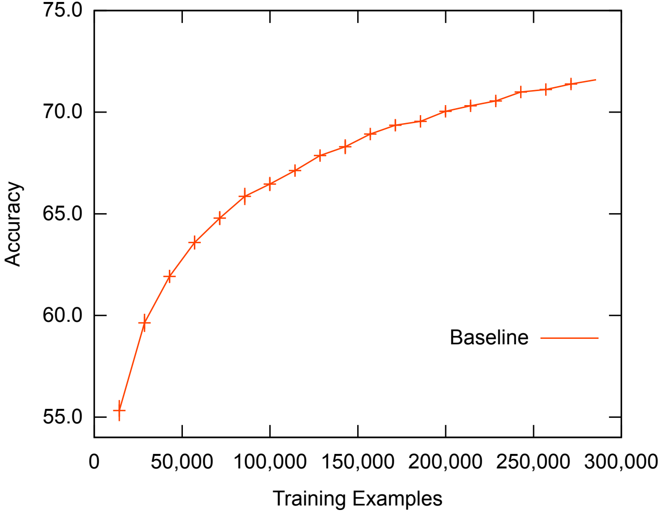
Model	Accuracy
BoW	67.61
Comp	71.59
Comp-Brown-news-100	71.54
Comp-Brown-news-320	71.93
Comp-Brown-news-1000	71.45
Comp-Brown-news-3200	71.42
Comp-ClarkNE-bio	72.05
Comp-Brown-bio-100	71.73
Comp-Brown-bio-320	72.03
Comp-Brown-bio-1000	72.31

Semantic Category Disambiguation: Results

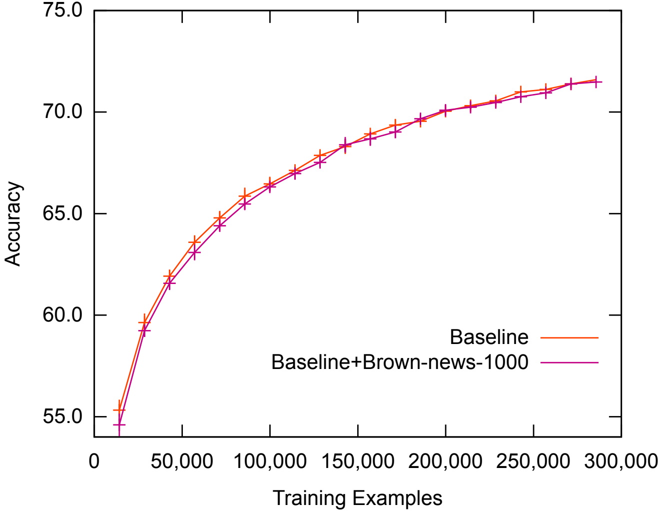
Model	Accuracy
BoW	67.61
Comp	71.59
Comp-Brown-news-100	71.54
Comp-Brown-news-320	71.93
Comp-Brown-news-1000	71.45
Comp-Brown-news-3200	71.42
Comp-ClarkNE-bio	72.05
Comp-Brown-bio-100	71.73
Comp-Brown-bio-320	72.03
Comp-Brown-bio-1000	72.31
Comp-Google	**73.70**

Semantic Category Disambiguation: Learning Curves

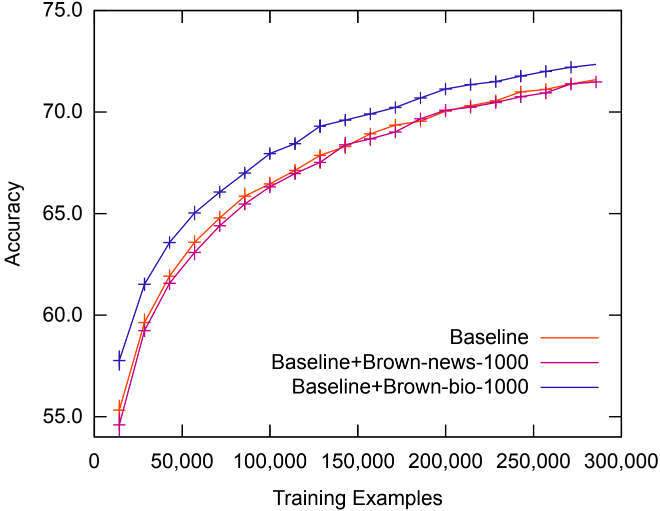
Semantic Category Disambiguation: Learning Curves



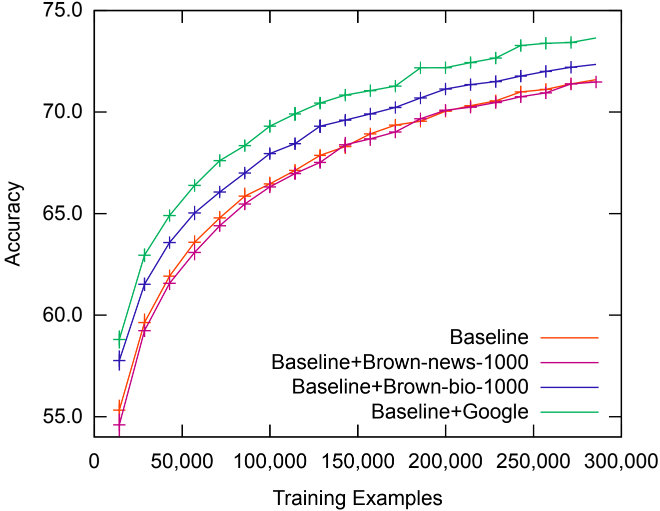
Semantic Category Disambiguation: Learning Curves



Semantic Category Disambiguation: Learning Curves



Semantic Category Disambiguation: Learning Curves



Conclusions and Future Work

Conclusions and Future Work

Conclusions

Conclusions and Future Work

Conclusions

- In-domain word representations outperform out-of-domain

Conclusions and Future Work

Conclusions

- In-domain word representations outperform out-of-domain
- Combinations of the two can potentially be beneficial

Conclusions and Future Work

Conclusions

- In-domain word representations outperform out-of-domain
- Combinations of the two can potentially be beneficial
- Performance benefits does not appear to saturate

Conclusions and Future Work

Conclusions

- In-domain word representations outperform out-of-domain
- Combinations of the two can potentially be beneficial
- Performance benefits does not appear to saturate

Future Work

Conclusions and Future Work

Conclusions

- In-domain word representations outperform out-of-domain
- Combinations of the two can potentially be beneficial
- Performance benefits does not appear to saturate

Future Work

- What is the impact of the size of the data?

Conclusions and Future Work

Conclusions

- In-domain word representations outperform out-of-domain
- Combinations of the two can potentially be beneficial
- Performance benefits does not appear to saturate

Future Work

- What is the impact of the size of the data?
- Is the observation regarding saturation accurate?

Conclusions and Future Work

Conclusions

- In-domain word representations outperform out-of-domain
- Combinations of the two can potentially be beneficial
- Performance benefits does not appear to saturate

Future Work

- What is the impact of the size of the data?
- Is the observation regarding saturation accurate?
- Consider more embedding types?

Thank You for Your Attention

ご清聴ありがとうございました

Tack för er uppmärksamhet

Code and Data: `http://wordreprs.nlplab.org/`
Slides: `http://pontus.stenetorp.se/`

Seed-words Used

Category	Seed words
Antibodies	MAb IgG IgM rituximab infliximab
Cells	RBC HUVEC BAEC VSMC SMC
Cell lines	PC12 CHO HeLa Jurkat COS
Diseases	asthma hepatitis tuberculosis HIV malaria
Drugs	acetylcholine carbachol heparin penicillin tetracycline
Molecular functions	kinase ligase acetyltransferase helicase binding
Mutations and mutants	Leiden C677T C282Y 35delG null
Proteins and genes	p53 actin collagen albumin IL-6
Signs and symptoms	anemia hypertension hyperglycemia fever cough
Tumors	lymphoma sarcoma melanoma neuroblastoma osteosarcoma

NER Corpora Statistics

	Corpus		
	AnEM	BC2GM	NCBID
Words	91,420	450,991	174,062
Sentences	4,548	20,000	7,844
Entities	3,135	24,596	6,900

Bibliography (1/2)

References

- J. Firth. 1957. **A synopsis of linguistic theory 1930–1955**. In *Studies in Linguistic Analysis*.
- J. Turian, L. Ratinov, and Y. Bengio. 2010. **Word representations: a simple and general method for semi-supervised learning**. In *Proceedings of ACL 2010*.
- T. Koo, X. Carreras and M. Collins. 2008. **Simple semi-supervised dependency parsing**. In *Proceedings of ACL 2008*.
- D. Lin and X. Wu. 2009. **Phrase clustering for discriminative learning**. In *Proceedings of ACL-IJCNLP 2009*.
- D. McClosky, M. Surdeanu, and C. Manning. 2011. **Event extraction as dependency parsing for BioNLP 2011**. In *Proceedings of BioNLP Shared Task 2011*.
- L. Ratinov and D. Roth. 2009. **Design challenges and misconceptions in named entity recognition**. In *Proceedings of CoNLL 2009*.

Bibliography (2/2)

References

- T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou. 2012. **Open-domain anatomical entity mention detection**. In *Proceedings of DSSD 2012.*
- L. Smith, L.K. Tanabe, R.J. Ando, C.J. Kuo, I.F. Chung, et al. 2008. **Overview of BioCreative II gene mention recognition**. *Genome Biology, 9(Suppl2):S2.*
- R. Islamaj Dogan and Z. Lu. 2012. **An improved corpus of disease mentions in PubMed citations.** In *Proceedings of BioNLP 2012.*
- T. McIntosh and J.R. Curran. 2009. **Reducing semantic drift with bagging and distributional similarity.** In *Proceedings of ACL 2009.*