LexSemTM: A Semantic Dataset Based on All-words Unsupervised Sense Distribution Learning

Andrew Bennett (*The University of Melbourne*)

Timothy Baldwin (*The University of Melbourne*)
Jey Han Lau (*IBM Research / The University of Melbourne*)
Diana McCarthy (*The University of Cambridge*)
Francis Bond (*Nanyang Technological University*)

Table of Contents

Unsupervised Sense Distribution Learning

- Task is to automatically learn relative frequencies of word senses, from unlabelled text

- Learning is type-level (one sense distribution per lemma type)

- Shown to be useful in WSD, particularly in domain adaptation, plus other applications (Lau et al., 2014)

Example

Example uses of word

“The crane flew north over the marsh”

“The crane is a graceful bird”

“A crane is a type of tall wading birds”

“The crane lifted the beam to the top of the building”

“Before they could construct the building, they needed a new crane”

Sense glosses from dictionary

crane -- (large long-necked wading bird of marshes and plains in many parts of the world)

crane -- (lifts and moves heavy objects; lifting tackle is suspended from a pivoted boom that rotates around a vertical axis)

Example

Example uses of word

“The crane flew north over the marsh”

“The crane is a graceful bird”

“A crane is a type of tall wading birds”

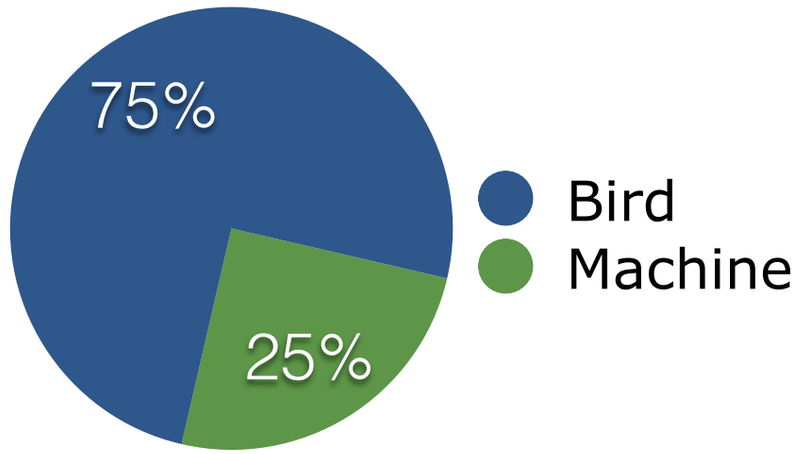
“The crane lifted the beam to the top of the building”

“Before they could construct the building, they needed a new crane”

Sense glosses from dictionary

crane -- (large long-necked wading bird of marshes and plains in many parts of the world)

crane -- (lifts and moves heavy objects; lifting tackle is suspended from a pivoted boom that rotates around a vertical axis)



Existing Sense Frequency Datasets

- Most comprehensive English language sense frequency resource is *SemCor*
- *SemCor* contains major gaps / inconsistencies
- Resources for other languages similarly limited

Our Goals

1. Apply unsupervised sense distribution learning to create resource to replace or supplement *SemCor*

2. Refine existing unsupervised sense distribution learning methodology to facilitate this

Table of Contents

Lau et al., (2014)

HDP-WSI

- Existing state-of-the-art method for unsupervised sense distribution learning
- Main computational component is HDP topic modelling (Teh et al., 2006)
  - HDP is a non-parametric generalisation of LDA (automatically learns “right” number of topics)
- Process performed separately for each lemma

Lau et al., (2014)

HDP-WSI

1. **Run HDP on usages to obtain topics**
2. Assignment of each usage to single topic
3. Soft alignment of each topic to all senses

Example uses of word

“The crane flew north over the marsh”

“The crane is a graceful bird”

“A crane is a type of tall wading birds”

“The crane lifted the beam to the top of the tall building”

“Before they could construct the building, they needed a new crane”

Lau et al., (2014)

HDP-WSI

1. **Run HDP on usages to obtain topics**
2. Assignment of each usage to single topic
3. Soft alignment of each topic to all senses

Example uses of word

“The crane flew north over the marsh”

“The crane is a graceful bird”

“A crane is a type of tall wading birds”

“The crane lifted the beam to the top of the tall building”

“Before they could construct the building, they needed a new crane”

Lau et al., (2014)

HDP-WSI

1. **Run HDP on usages to obtain topics**
2. Assignment of each usage to single topic
3. Soft alignment of each topic to all senses

Example uses of word

"The crane flew north over the marsh"

"The crane is a graceful bird"

"A crane is a type of tall wading birds"

"The crane lifted the beam to the top of the tall building"

"Before they could construct the building, they needed a new crane"

Topic Distributions (bags of words)

t_1

crane flew north marsh crane is graceful bird crane is type of tall wading bird of tall before they they crane

t_2

the over the the a a a the crane lift the beam to the top the building could construct the building need a new

Lau et al., (2014)

HDP-WSI

1. Run HDP on usages to obtain topics
2. **Assignment of each usage to single topic**
3. Soft alignment of each topic to all senses

Example uses of word

"The crane flew north over the marsh"

"The crane is a graceful bird"

"A crane is a type of tall wading birds"

"The crane lifted the beam to the top of the tall building"

"Before they could construct the building, they needed a new crane"

Topic Distributions (bags of words)

t_1

crane flew north marsh crane is graceful bird crane is type of tall wading bird of tall before they they crane

$$P(t_1) = 0.6$$

t_2

the over the the a a a the crane lift the beam to the top the building could construct the building need a new

$$P(t_2) = 0.4$$

Lau et al., (2014)

HDP-WSI

1. Run HDP on usages to obtain topics
2. Assignment of each usage to single topic
3. **Soft alignment of each topic to all senses**

Sense Gloss Distributions (bags of words)

s_1

large long neck wading bird of marsh and plain in many part of the world

s_2

lift and move heavy object lift tackle is suspend from a pivoted boom that rotate around a vertical axis

Topic Distributions (bags of words)

t_1

crane flew north marsh crane is graceful bird crane is type of tall wading bird of tall before they they crane

$$P(t_1) = 0.6$$

t_2

the over the the a a a the crane lift the beam to the top the building could construct the building need a new

$$P(t_2) = 0.4$$

Lau et al., (2014)

HDP-WSI

1. Run HDP on usages to obtain topics
2. Assignment of each usage to single topic
3. **Soft alignment of each topic to all senses**

Sense Gloss Distributions (bags of words)

s_1

large long neck wading bird of marsh and plain in many part of the world

$prev(s_1)$

s_2

lift and move heavy object lift tackle is suspend from a pivoted boom that rotate around a vertical axis

$prev(s_2)$

Topic Distributions (bags of words)

t_1

crane flew north marsh crane is graceful bird crane is type of tall wading bird of tall before they they crane

$P(t_1) = 0.6$

t_2

the over the the a a a the crane lift the beam to the top the building could construct the building need a new

$P(t_2) = 0.4$

$$prev(s_i) = \sum_{j=1}^{T} P(t_j)(1 - JSD(s_i, t_j))$$

JSD = Jensen-Shannon Divergence between corresponding bags-of-words distributions

Lau et al., (2014)

HDP-WSI

1. Run HDP on usages to obtain topics
2. Assignment of each usage to single topic
3. **Soft alignment of each topic to all senses**

Sense Gloss Distributions (bags of words)

s_1

large long neck wading bird of marsh and plain in many part of the world

$prev(s_1)$ 

s_2

lift and move heavy object lift tackle is suspend from a pivoted boom that rotate around a vertical axis

$prev(s_2)$ 

Topic Distributions (bags of words)

t_1

crane flew north marsh crane is graceful bird crane is type of tall wading bird of tall before they they crane

$$P(t_1) = 0.6$$

t_2

the over the the a a a the crane lift the beam to the top the building could construct the building need a new

$$P(t_2) = 0.4$$

$$prev(s_i) = \sum_{j=1}^{T} P(t_j)(1 - JSD(s_i, t_j))$$

JSD = Jensen-Shannon Divergence between corresponding bags-of-words distributions

Lau et al., (2014)

HDP-WSI

1. Run HDP on usages to obtain topics
2. Assignment of each usage to single topic
3. **Soft alignment of each topic to all senses**

Sense Gloss Distributions (bags of words)

s_1

large long neck wading bird of marsh and plain in many part of the world

$prev(s_1)$ 

s_2

lift and move heavy object lift tackle is suspend from a pivoted boom that rotate around a vertical axis

$prev(s_2)$ 

Topic Distributions (bags of words)

t_1

crane flew north marsh crane is graceful bird crane is type of tall wading bird of tall before they they crane

$P(t_1) = 0.6$

t_2

the over the the a a a the crane lift the beam to the top the building could construct the building need a new

$P(t_2) = 0.4$

$$prev(s_i) = \sum_{j=1}^{T} P(t_j)(1 - JSD(s_i, t_j))$$

JSD = Jensen-Shannon Divergence between corresponding bags-of-words distributions

Lau et al., (2014)

HDP-WSI

1. Run HDP on usages to obtain topics
2. Assignment of each usage to single topic
3. **Soft alignment of each topic to all senses**

Sense Gloss Distributions (bags of words)

s_1

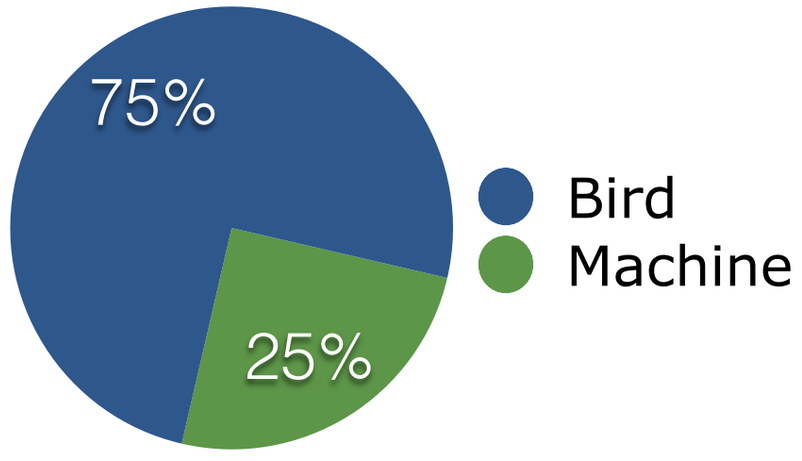
large long neck wading bird of marsh and plain in many part of the world

$prev(s_1)$ 

s_2

lift and move heavy object lift tackle is suspend from a pivoted boom that rotate around a vertical axis

$prev(s_2)$ 



Normalised prevalence scores

$$prev(s_i) = \sum_{j=1}^{T} P(t_j)(1 - JSD(s_i, t_j))$$

JSD = Jensen-Shannon Divergence between corresponding bags-of-words distributions

Why not Use HDP-WSI?

- Too slow!

- ~1 hour per word, doesn't scale well to language-wide computation

Our Method: HCA-WSI

- Follows same procedure as HDP-WSI, except HDP is replaced by HCA (Buntine and Mishra, 2014)

- Differences between HCA and HDP:

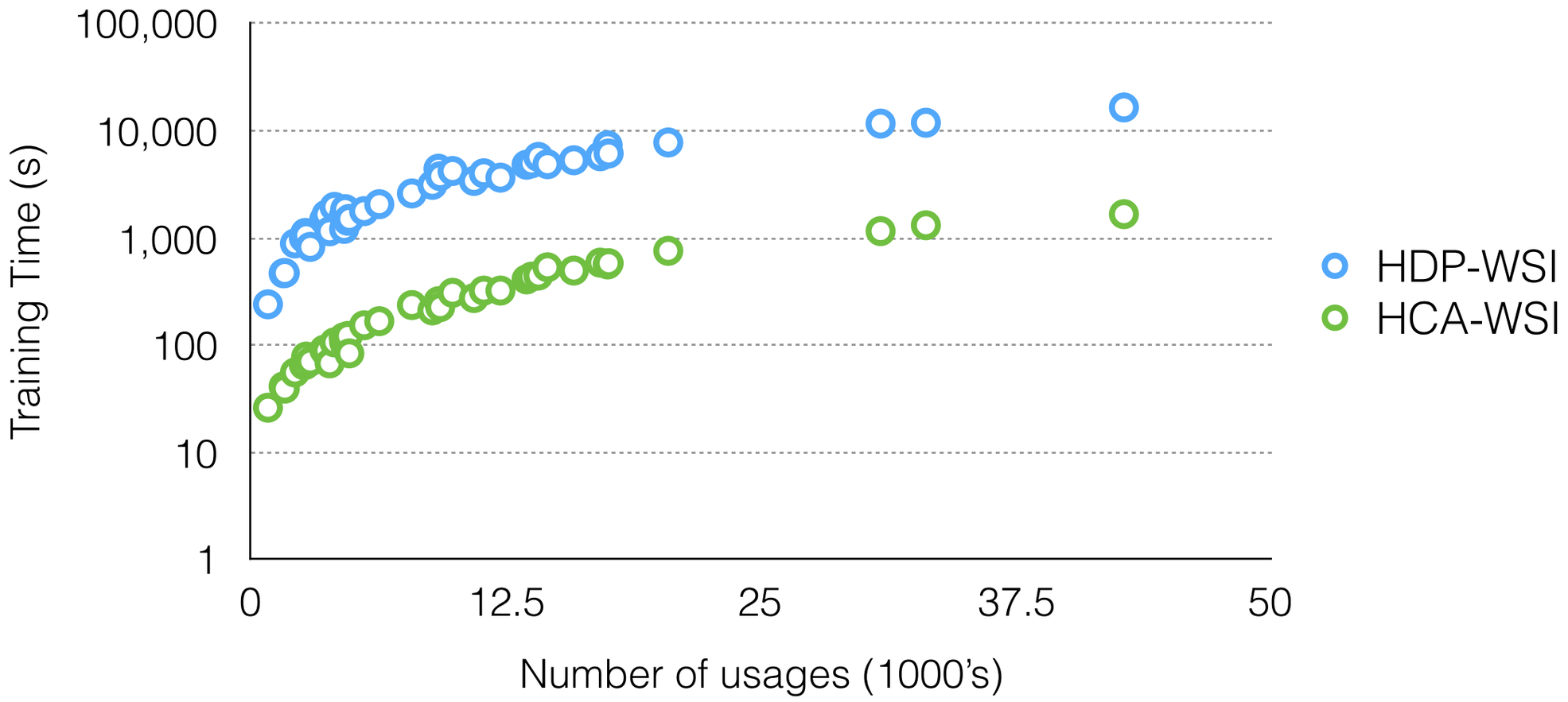
1. **HCA uses more efficient Gibbs sampling inference algorithm than HDP**

2. HCA based on Pitman-Yor processes rather than Dirichlet processes

3. HCA models document-level burstiness

4. HCA requires setting fixed number of topics

HCA-WSI vs HDP-WSI



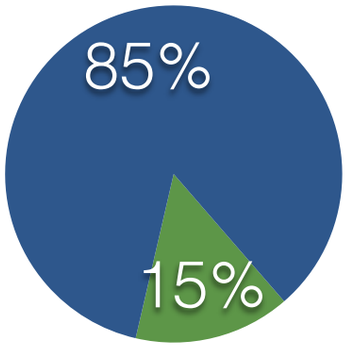
- Impact on sense distribution quality of HCA-WSI versus HDP-WSI not statistically significant ($p > 0.05$)

Lau et al., (2014)

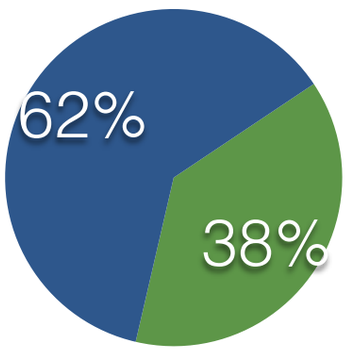
Sense Distribution Quality Metric: JSD

Gold-Standard

Candidate 1



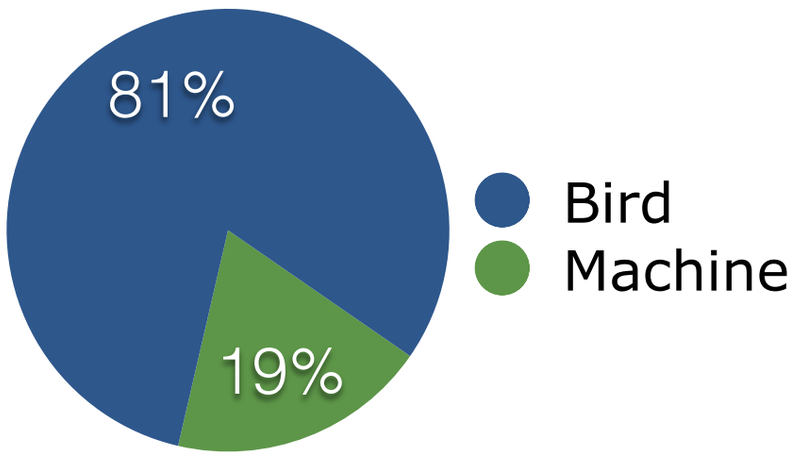
Candidate 2



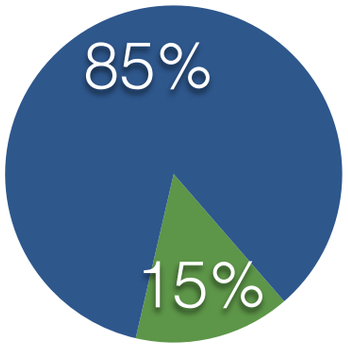
Lau et al., (2014)

Sense Distribution Quality Metric: JSD

Gold-Standard

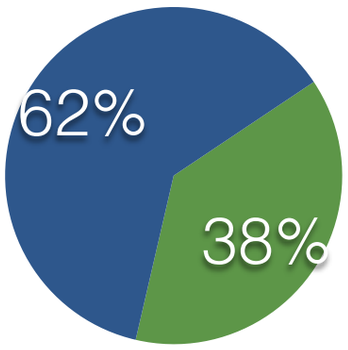


Candidate 1



$JSD = 0.0020$

Candidate 2



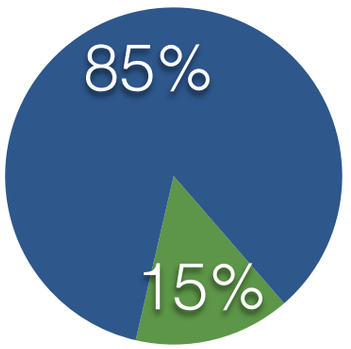
$JSD = 0.0324$

Lau et al., (2014)

Sense Distribution Quality Metric: JSD

Gold-Standard

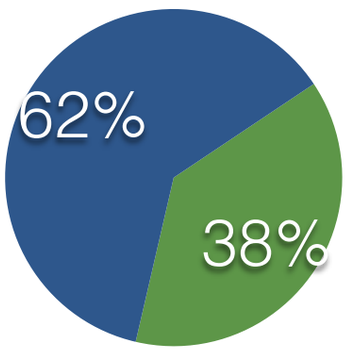
Candidate 1



$JSD = 0.0020$

(lower is better)

Candidate 2



$JSD = 0.0324$

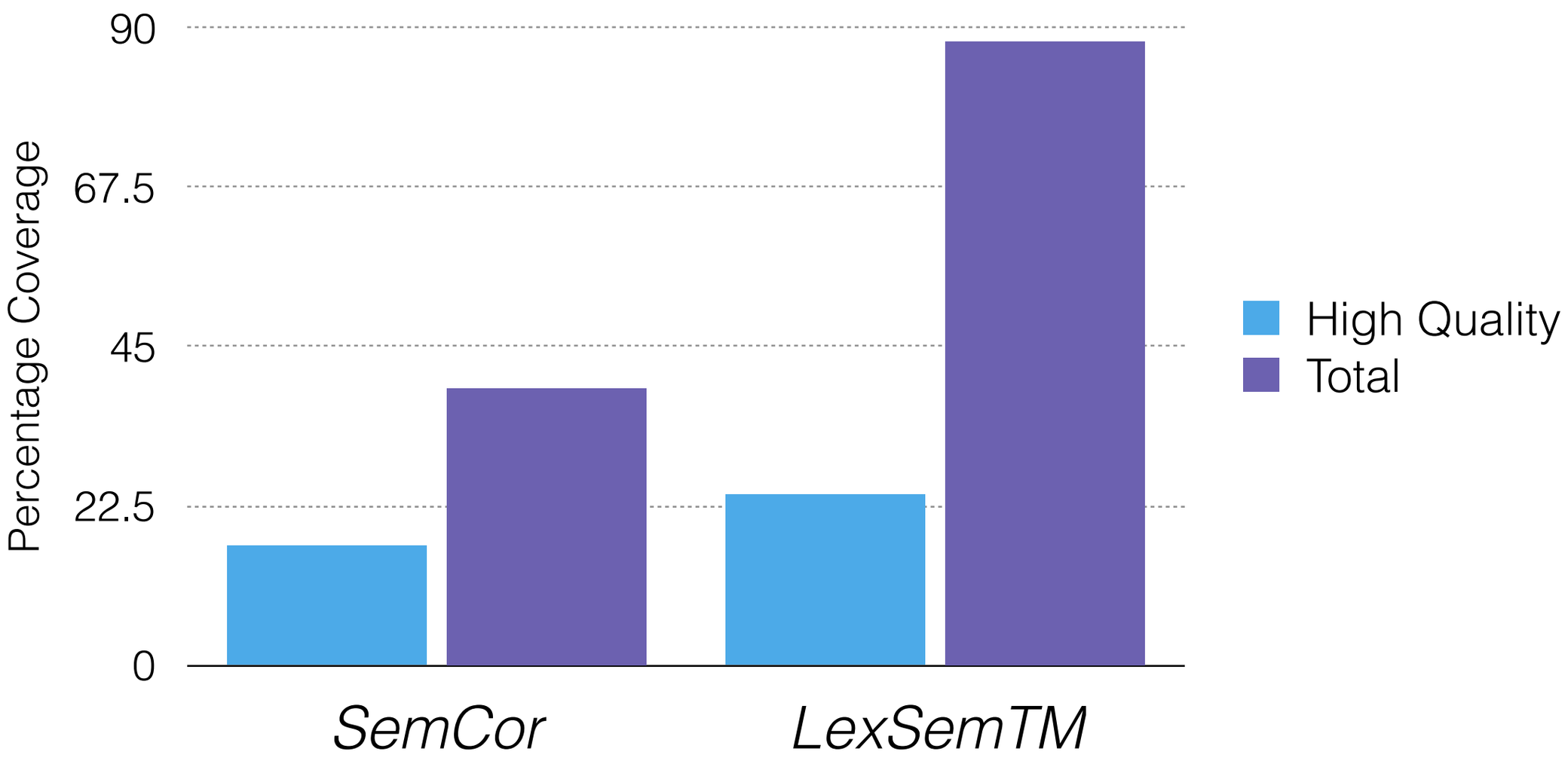
Table of Contents

LexSemTM Dataset

- Created by applying HCA-WSI to usages sampled from English Wikipedia, for all *WordNet* lemmas with at least 20 usages

- Also contains topic HCA topic modelling output for all lemmas (can be realigned to other sense inventories)

Coverage of Polysemous *WordNet* Lemmas



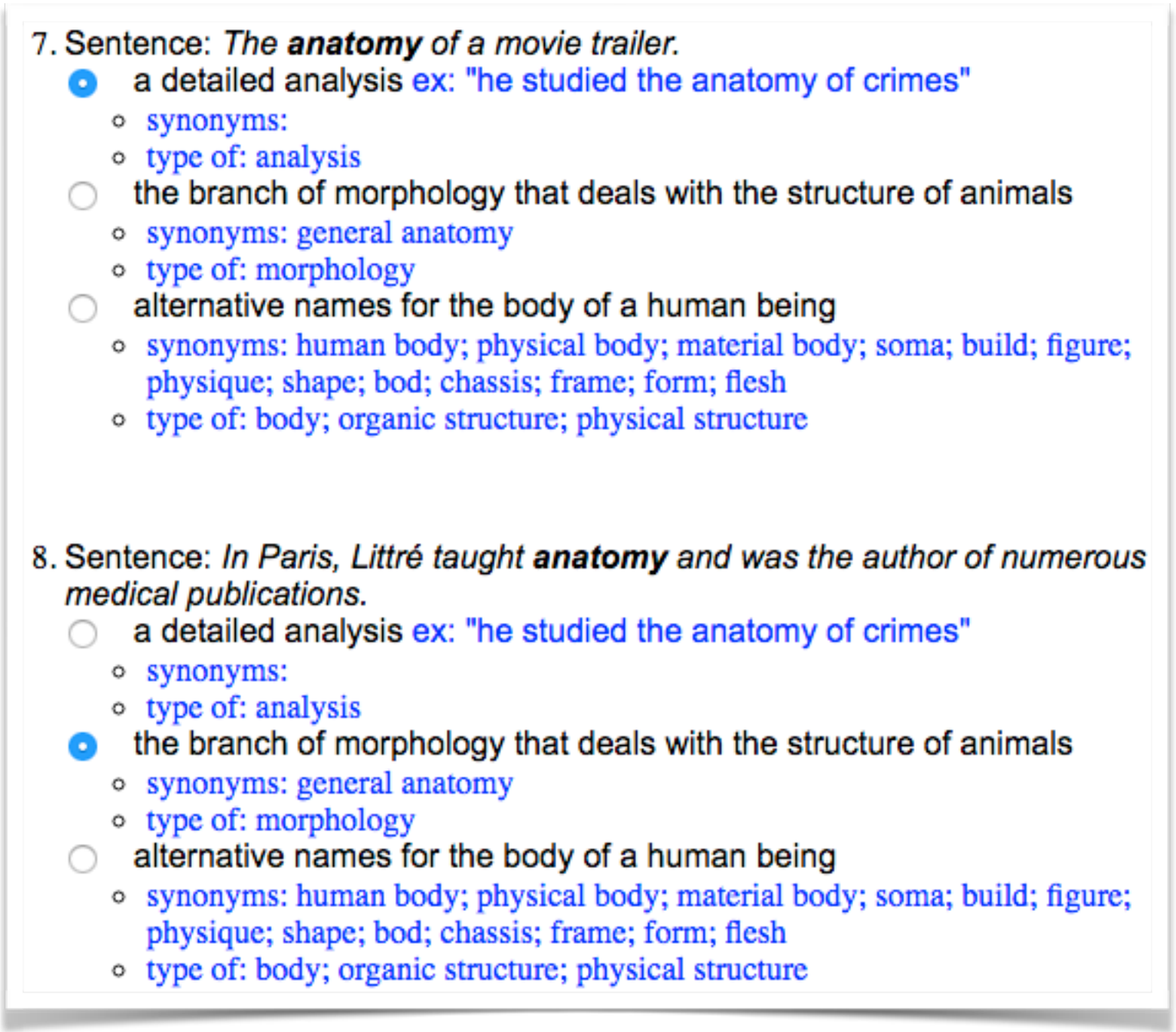
Evaluation

- Evaluating how *LexSemTM* and *SemCor* compare in quality, as a function of *SemCor* frequency

- Want to decide for which lemmas *LexSemTM* can replace/supplement *SemCor*

- Evaluation performed on set of 50 nouns covering wide range of *SemCor* frequencies

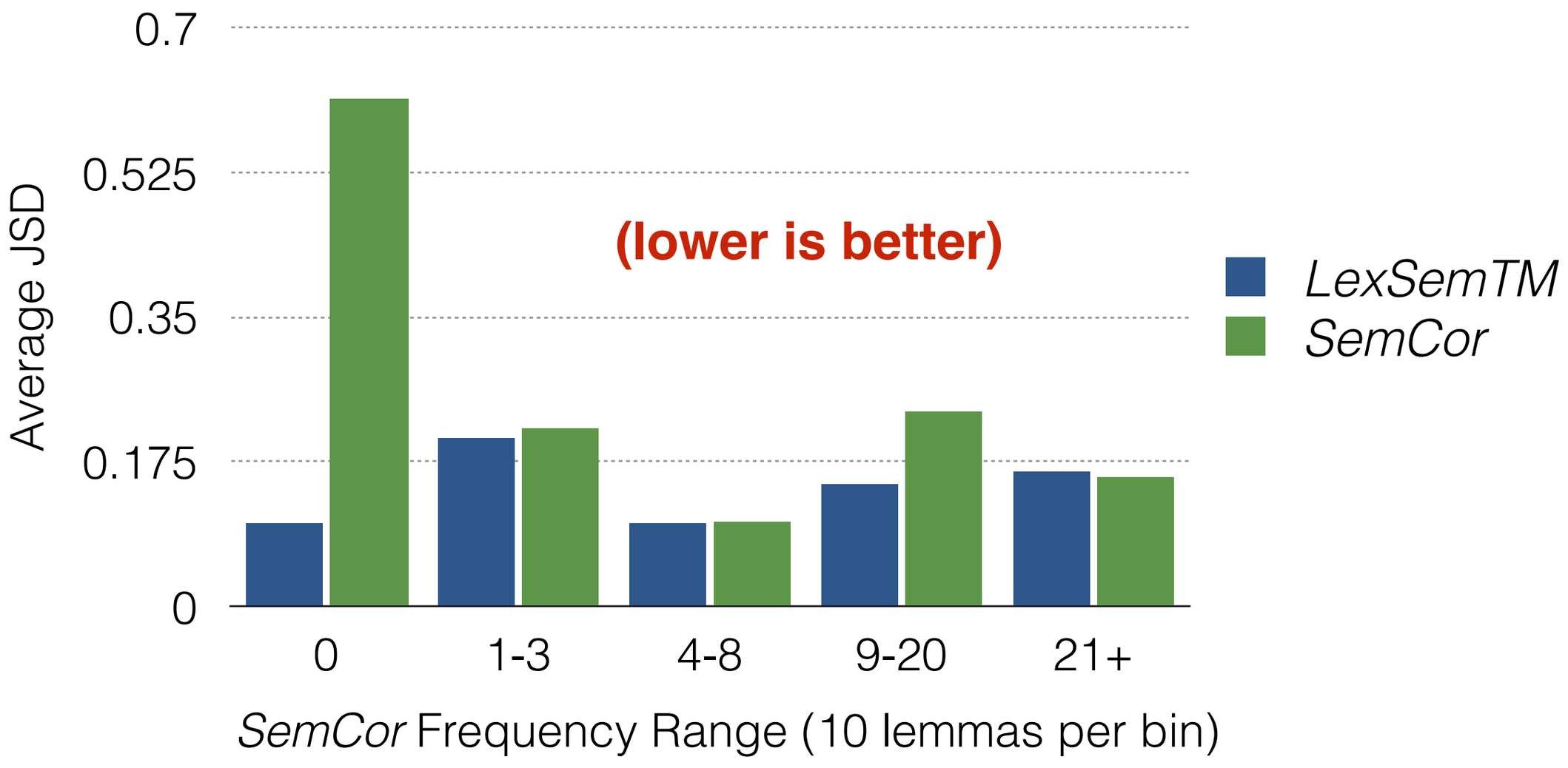
Gold-Standard Annotation



Annotation Interface

- For each of 50 lemmas, 100 sentences randomly sampled to be sense-annotated
- Annotation done using Amazon Mechanical Turk
- Annotation results processed and normalised to give gold-standard distributions

Evaluation Results



Summary

- HCA-WSI is on par with previous state-of-the-art in terms of sense distribution quality, and 10-20 times faster

- *LexSemTM* provides substantially greater coverage of polysemous *WordNet* lemmas than *SemCor*, and appears to be at least as accurate

- *LexSemTM* also contains topic model output that can be re-aligned to other sense inventories

Questions?

References

- Wray L Buntine and Swapnil Mishra. 2014. Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2014)*, pages 881–890, New York City, USA.

- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, USA.

- Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 259–270, Baltimore, USA.

- Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.