Combining Distant and Partial Supervision for Relation Extraction

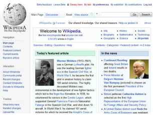
Gabor Angeli, Julie Tibshirani, Jean Y. Wu, Christopher D. Manning

Stanford University

October 28, 2014

Motivation: Knowledge Base Completion

Unstructured Text



$\Longrightarrow$

Structured Knowledge Base

Barack Obama

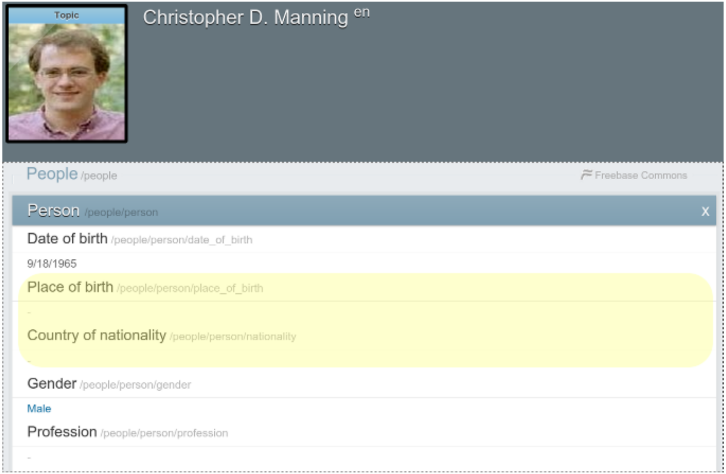
44th President of the United States

Personal details

Born	Barack Hussein Obama II August 4, 1961 (age 52) Honolulu, Hawaii, U.S.
Political party	Democratic
Spouse(s)	Michelle LaVaughn Robinson (m. 1992–present)
Children	Malia Ann Obama (b. 1998) Natasha Obama (b. 2001)

Motivation: Question Answering

Motivation: Question Answering



Motivation: Question Answering

Christopher Manning

Professor of Linguistics and Computer Science

Natural Language Processing Group, Stanford University

Brief Bio

- I'm Australian ("I come from a land of wide open spaces ...")
- BA (Hons) Australian National University 1989 (majors in mathematics, computer science and linguistics)
- PhD Stanford Linguistics 1995
- Asst Professor Carnegie Mellon University Computational Linguistics Program 1994-96
- Lecturer University of Sydney Dept of Linguistics 1996-99
- Asst Professor Stanford University Depts of Computer Science and Linguistics 1999-2006
- Assoc Professor Stanford University Depts of Linguistics and Computer Science 2006-2012
- Professor Stanford University Depts of Linguistics and Computer Science 2012-

Motivation: Question Answering

Australia

Christopher D. Manning, origin

Feedback

Relation Extraction

Input: Sentences containing (entity, slot value).

Output: Relation between entity and slot value.

Relation Extraction

Input: Sentences containing (entity, slot value).
Output: Relation between entity and slot value.

Consider two approaches:

- **Supervised:** Trivial as a supervised classifier.
  Training data: $\{$(sentence, relation)$\}$.
  But...

Relation Extraction

Input: Sentences containing (entity, slot value).
Output: Relation between entity and slot value.

Consider two approaches:

- **Supervised:** Trivial as a supervised classifier.
  Training data: $\{$(sentence, relation)$\}$.
  But... this training data is expensive to produce.

Relation Extraction

Input: Sentences containing (entity, slot value).
Output: Relation between entity and slot value.

Consider two approaches:

- **Supervised:** Trivial as a supervised classifier.
  Training data: $\{$(sentence, relation)$\}$.
  But... this training data is expensive to produce.

- **Distantly Supervised:** Artificially produce "supervised" data.
  Training data: $\{$(entity, relation, slot value)$\}$.
  But...

Relation Extraction

Input: Sentences containing (entity, slot value).
Output: Relation between entity and slot value.

Consider two approaches:

- **Supervised:** Trivial as a supervised classifier.
  Training data: $\{$(sentence, relation)$\}$.
  But... this training data is expensive to produce.

- **Distantly Supervised:** Artificially produce "supervised" data.
  Training data: $\{$(entity, relation, slot value)$\}$.
  But... this training data is much more noisy.

Contribution: Combine Benefits of Both

Adding carefully selected supervision improves distantly supervised relation extraction.

Contribution: Combine Benefits of Both

Adding carefully selected supervision improves distantly supervised relation extraction.

- What is “carefully selected”: Propose new active learning criterion.
- Evaluate a number of questions:

Contribution: Combine Benefits of Both

Adding carefully selected supervision improves distantly supervised relation extraction.

- What is “carefully selected”: Propose new active learning criterion.
- Evaluate a number of questions:
  - Is the proposed criterion better than other methods?

Contribution: Combine Benefits of Both

Adding carefully selected supervision improves distantly supervised relation extraction.

- What is “carefully selected”: Propose new active learning criterion.
- Evaluate a number of questions:
  - Is the proposed criterion better than other methods?
  - Where is the supervision helping?

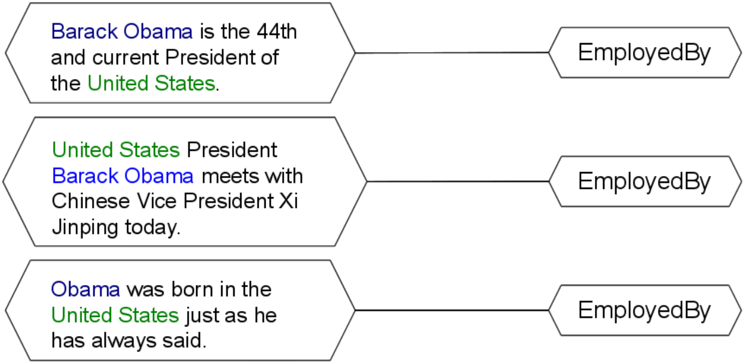
Contribution: Combine Benefits of Both

Adding carefully selected supervision improves distantly supervised relation extraction.

- What is "carefully selected": Propose new active learning criterion.
- Evaluate a number of questions:
  - Is the proposed criterion better than other methods?
  - Where is the supervision helping?
  - How far can we get with a supervised classifier?

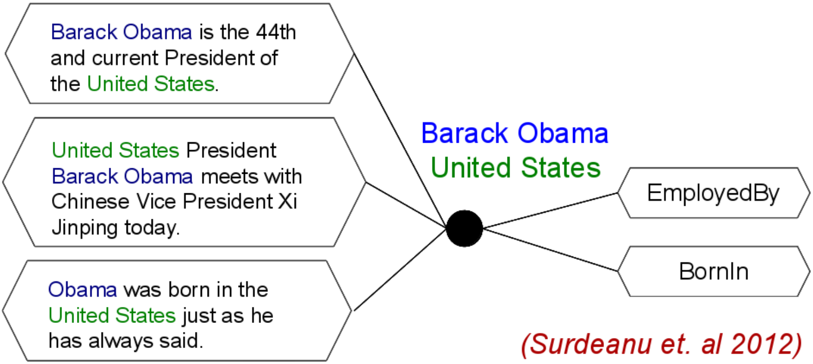
Distant Supervision

≈ (Barack Obama, EmployedBy, United States)



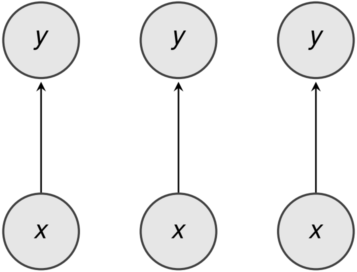
Multiple-Instance Multiple-Label (MIML) Learning

(Barack Obama, EmployedBy, United States)

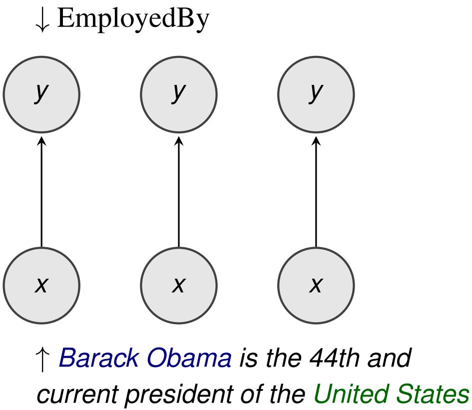


(Surdeanu et. al 2012)

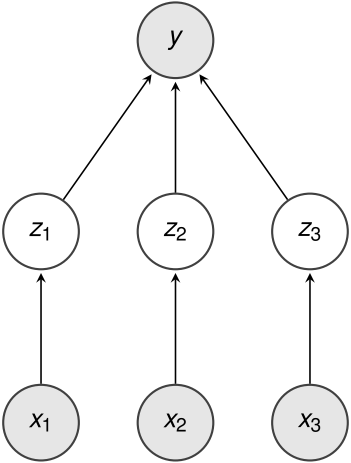
Distant Supervision



Distant Supervision

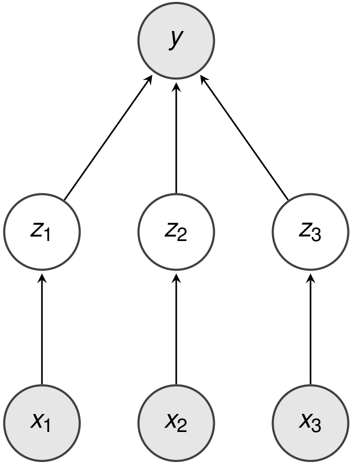


Multiple-Instance

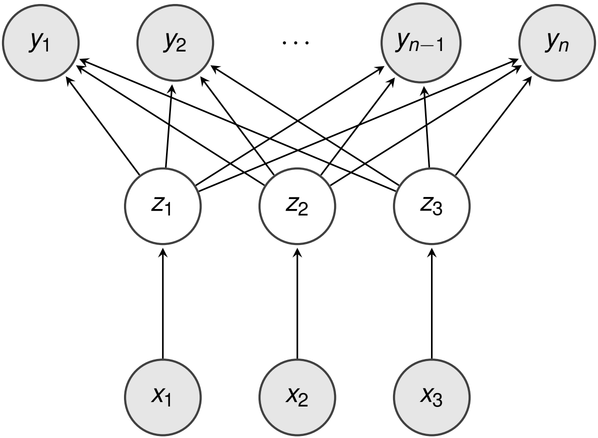


Multiple-Instance

Latent
per-mention relation $\longrightarrow$



Multiple-Instance Multiple-Label (MIML-RE)



Active Learning

Old problem: Supervision is expensive, but very useful.

Old solution: Active learning!

Active Learning

Old problem: Supervision is expensive, but very useful.

Old solution: Active learning!

- Select a subset of latent z to annotate.
- Fix these labels during training.

Active Learning

Old problem: Supervision is expensive, but very useful.

Old solution: Active learning!

- Select a subset of latent z to annotate.
- Fix these labels during training.
- Bonus: this creates a supervised training set.
  - We initialize from a supervised classifier on this training set.

Active Learning

Old problem: Supervision is expensive, but very useful.

Old solution: Active learning!

- Select a subset of latent z to annotate.
- Fix these labels during training.
- Bonus: this creates a supervised training set.
  - We initialize from a supervised classifier on this training set.

Some Statistics

- 1,208,524 latent z which we could annotate.
- \$0.13 per annotation.
- \$160,000 to annotate everything.

Active Learning

Old problem: Supervision is expensive, but very useful.

Old solution: Active learning!

- Select a subset of latent z to annotate.
- Fix these labels during training.
- Bonus: this creates a supervised training set.
  - We initialize from a supervised classifier on this training set.

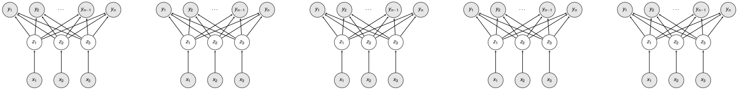
Some Statistics

- 1,208,524 latent z which we could annotate.
- \$0.13 per annotation.
- \$160,000 to annotate everything.

New spin: Have to get it right the first time.

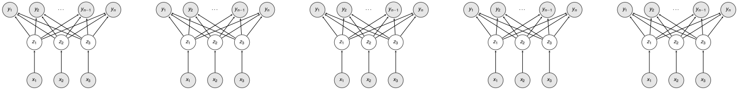
Example Selection Criteria

1. Train *k* MIML-RE models on *k* subsets of the data.



Example Selection Criteria

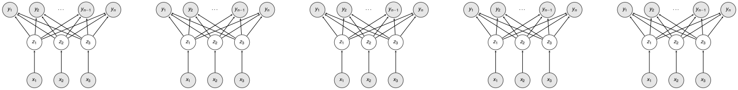
1. Train k MIML-RE models on k subsets of the data.



2. For each latent z, each trained model c predicts a multinomial $P_c(z)$.

Example Selection Criteria

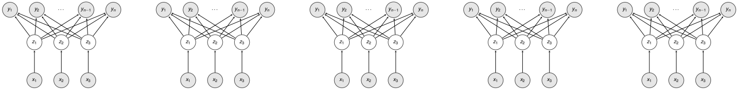
1. Train k MIML-RE models on k subsets of the data.



2. For each latent z, each trained model c predicts a multinomial $P_c(z)$.
3. Calculate Jensen-Shannon divergence: $\frac{1}{k}\sum_{c=1}^{k} \mathrm{KL}(p_c(z)||p_{mean}(z))$.

Example Selection Criteria

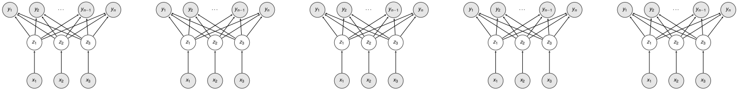
1. Train k MIML-RE models on k subsets of the data.



2. For each latent z, each trained model c predicts a multinomial $P_c(z)$.
3. Calculate Jensen-Shannon divergence: $\frac{1}{k}\sum_{c=1}^{k} \mathrm{KL}(p_c(z) || p_{mean}(z))$.
4. We have measure of *disagreement* for each z.

Example Selection Criteria

1. Train k MIML-RE models on k subsets of the data.



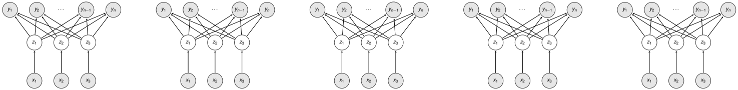
2. For each latent z, each trained model c predicts a multinomial $P_c(z)$.
3. Calculate Jensen-Shannon divergence: $\frac{1}{k}\sum_{c=1}^{k} \mathrm{KL}(p_c(z) || p_{mean}(z))$.
4. We have measure of *disagreement* for each z.

Three selection criteria

- Sample uniformly (*uniform*).

Example Selection Criteria

1. Train k MIML-RE models on k subsets of the data.



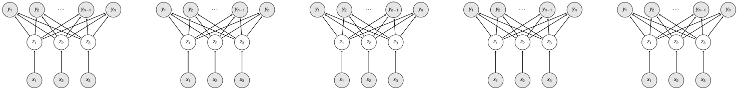
2. For each latent z, each trained model c predicts a multinomial $P_c(z)$.
3. Calculate Jensen-Shannon divergence: $\frac{1}{k}\sum_{c=1}^{k} \mathrm{KL}(p_c(z) || p_{mean}(z))$.
4. We have measure of *disagreement* for each z.

Three selection criteria

- Sample uniformly (*uniform*).
- Take z with highest disagreement (*highJS*).

Example Selection Criteria

1. Train k MIML-RE models on k subsets of the data.

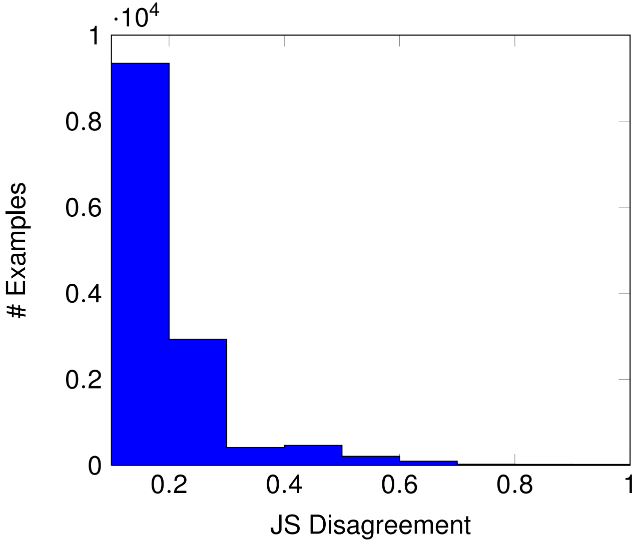


2. For each latent z, each trained model c predicts a multinomial $P_c(z)$.
3. Calculate Jensen-Shannon divergence: $\frac{1}{k}\sum_{c=1}^{k} \mathrm{KL}(p_c(z)||p_{mean}(z))$.
4. We have measure of *disagreement* for each z.

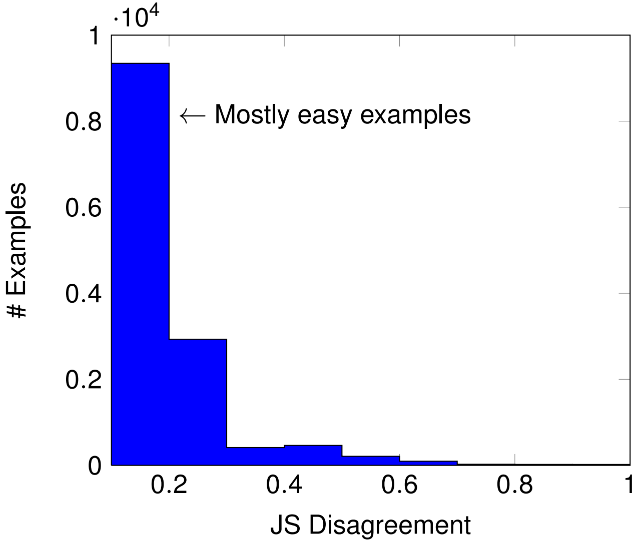
Three selection criteria

- Sample uniformly (*uniform*).
- Take z with highest disagreement (*highJS*).
- **Sample z with highest disagreement (*sampleJS*).**

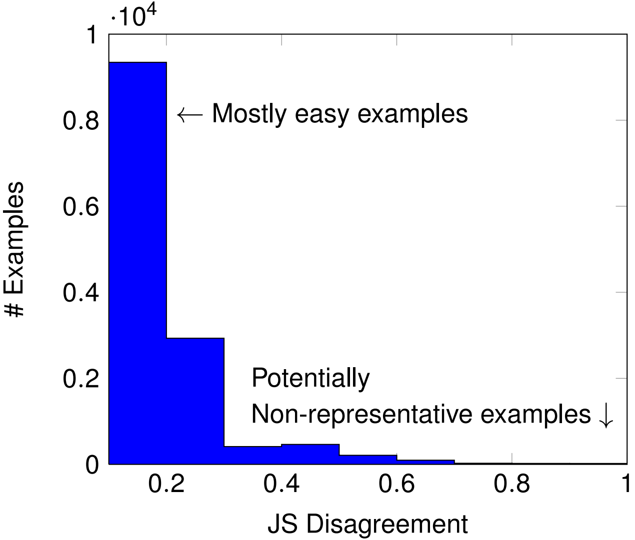
Example Selection Criteria



Example Selection Criteria

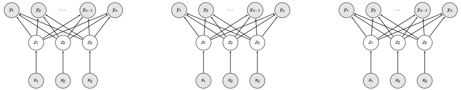


Example Selection Criteria



Example Selection Criteria

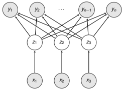
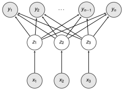
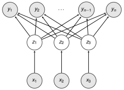
Committee Member Judgments



Sentence	**Member A**	**Member B**	**Member C**
Obama was born in Hawaii	born	born	no relation
Obama grew up in Hawaii	born	lived in	born
Obama Bear visits Hawaii	no relation	born	employee of
President Obama . . .	title	title	title
Obama employed president . . .	employee of	title	employee of

Example Selection Criteria

Committee Member Judgments

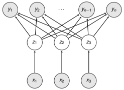
  

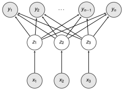
Sentence	**Member A**	**Member B**	**Member C**
Obama was born in Hawaii	born	born	no relation
Obama grew up in Hawaii	born	lived in	born
Obama Bear visits Hawaii	no relation	born	employee of
President Obama …	title	title	title
Obama employed president …	employee of	title	employee of

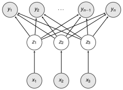
Uniform: Often annotates easy sentences.

Example Selection Criteria

Committee Member Judgments







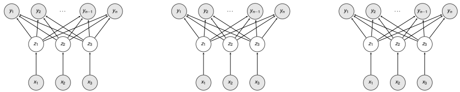
Sentence	**Member A**	**Member B**	**Member C**
Obama was born in Hawaii	born	born	no relation
Obama grew up in Hawaii	born	lived in	born
Obama Bear visits Hawaii	no relation	born	employee of
President Obama …	title	title	title
Obama employed president …	employee of	title	employee of

Uniform: Often annotates easy sentences.

High JS (disagreement): More likely to annotate “rare” sentences.

Example Selection Criteria

Committee Member Judgments



Sentence	**Member A**	**Member B**	**Member C**
Obama was born in Hawaii	born	born	no relation
Obama grew up in Hawaii	born	lived in	born
Obama Bear visits Hawaii	no relation	born	employee of
President Obama . . .	title	title	title
Obama employed president . . .	employee of	title	employee of

Uniform: Often annotates easy sentences.

High JS (disagreement): More likely to annotate “rare” sentences.

Sample JS (disagreement): Mix of hard and representative sentences.

Experiments

Recall our questions:

- Is the proposed criterion better than other methods?
- Where is the supervision helping?
- How far can we get with a supervised classifier?

Experiments

Recall our questions:

- Is the proposed criterion better than other methods?
- Where is the supervision helping?
- How far can we get with a supervised classifier?

Two experimental setups:

- Slot filling evaluation of Surdeanu et al. (2012).
- Stanford's 2013 TAC-KBP slot filling system

Experiments

Recall our questions:

- Is the proposed criterion better than other methods?
- Where is the supervision helping?
- How far can we get with a supervised classifier?

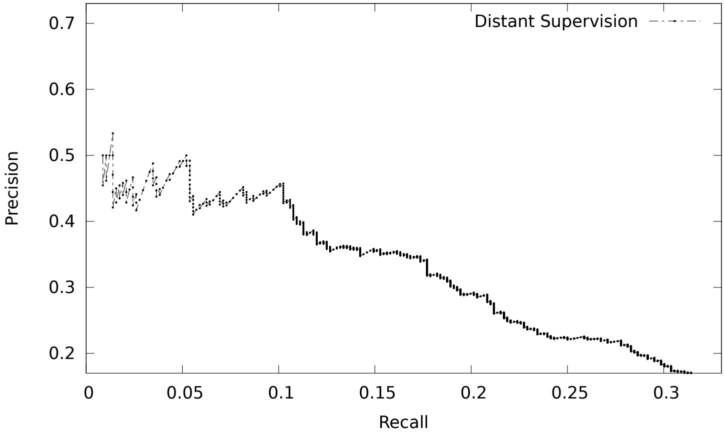
Two experimental setups:

- Slot filling evaluation of Surdeanu et al. (2012).
- Stanford's 2013 TAC-KBP slot filling system

Bonus: 4.4 F_1 improvement on 2014 TAC-KBP competition

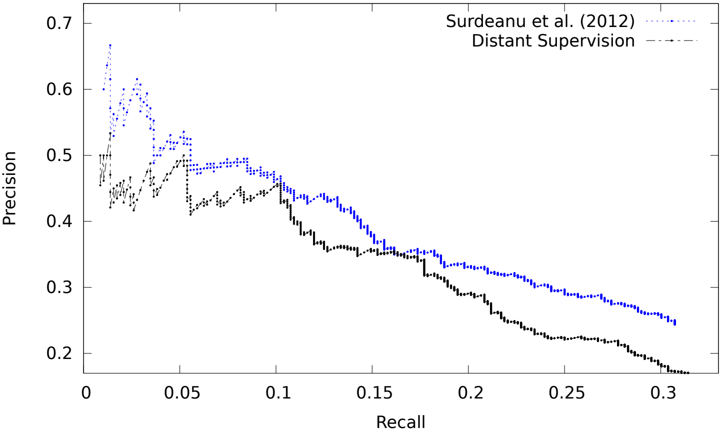
Old News: MIML-RE Works Well

Slot filling evaluation of Surdeanu et al. (2012).



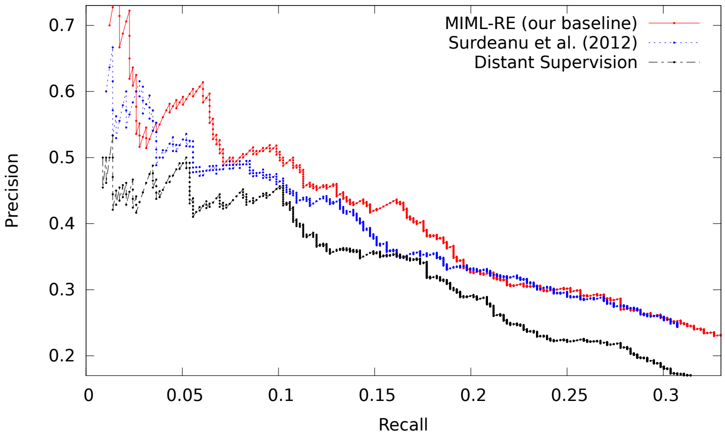
Old News: MIML-RE Works Well

Slot filling evaluation of Surdeanu et al. (2012).



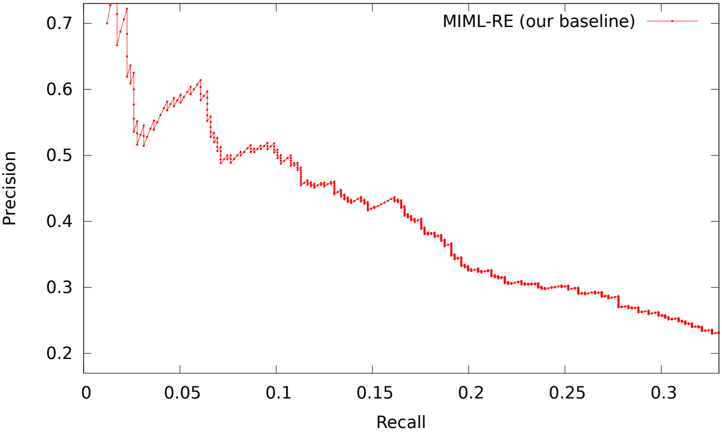
Old News: MIML-RE Works Well

Slot filling evaluation of Surdeanu et al. (2012).



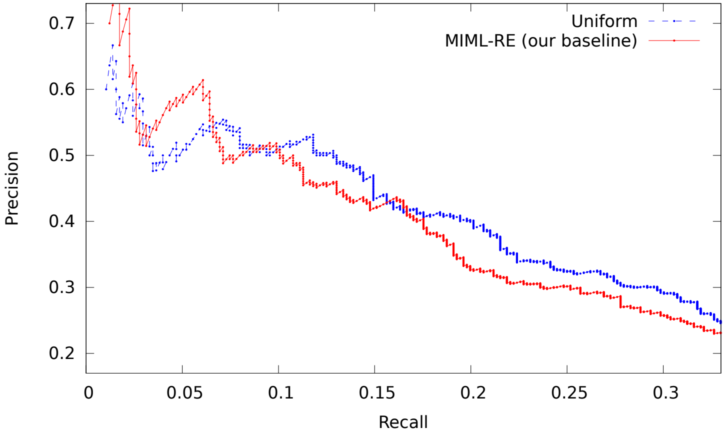
Active learning is important; SampleJS performs well.

Slot filling evaluation of Surdeanu et al. (2012).



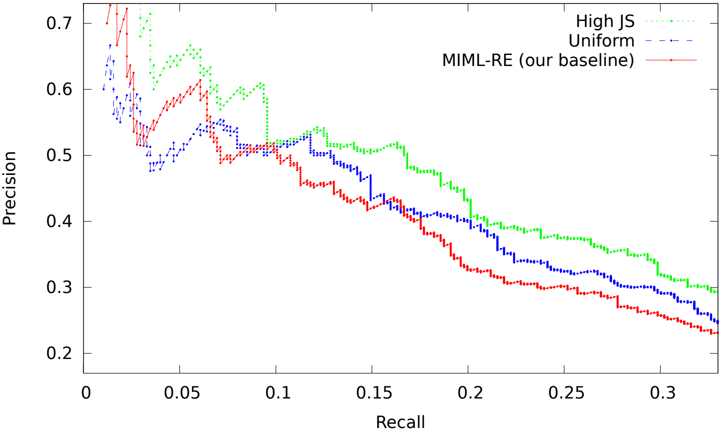
Active learning is important; SampleJS performs well.

Slot filling evaluation of Surdeanu et al. (2012).



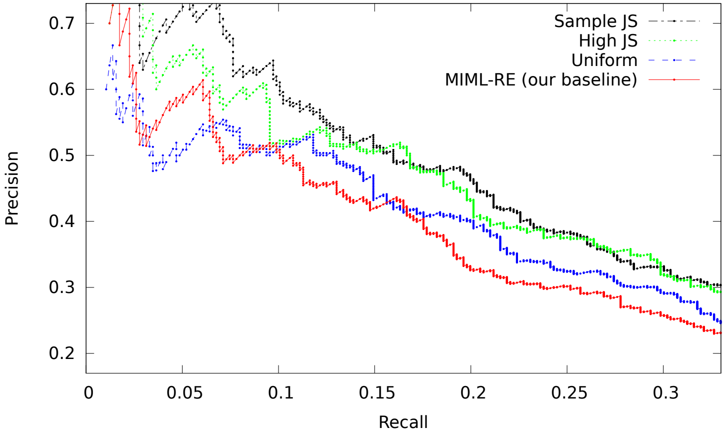
Active learning is important; SampleJS performs well.

Slot filling evaluation of Surdeanu et al. (2012).



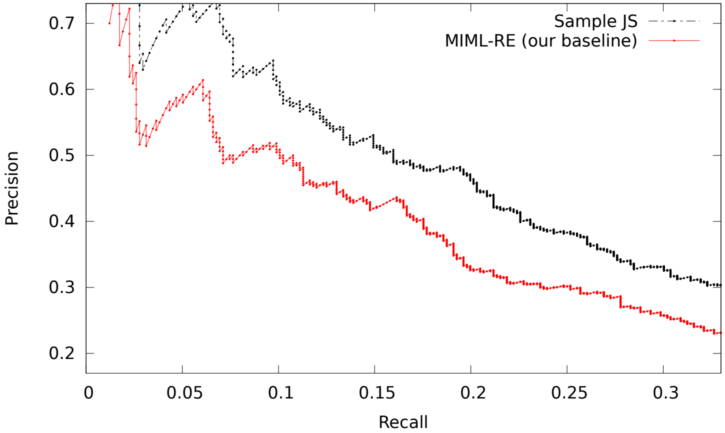
Active learning is important; SampleJS performs well.

Slot filling evaluation of Surdeanu et al. (2012).



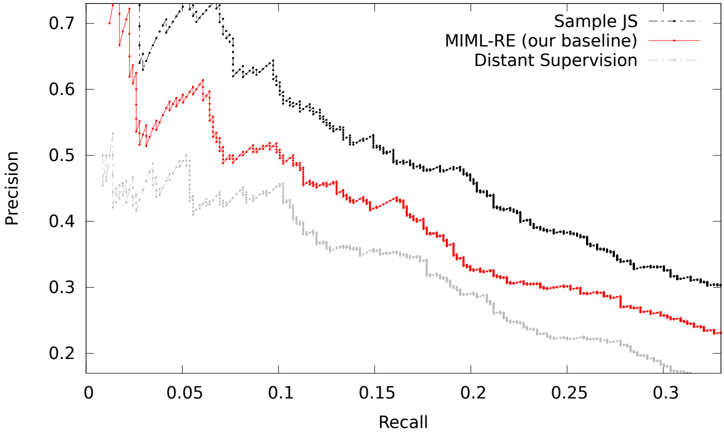
Active learning is important; SampleJS performs well.

Slot filling evaluation of Surdeanu et al. (2012).



Active learning is important; SampleJS performs well.

Slot filling evaluation of Surdeanu et al. (2012).



SampleJS performs best on TAC-KBP challenge.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.

SampleJS performs best on TAC-KBP challenge.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

SampleJS performs best on TAC-KBP challenge.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
No active learning	38.0	30.5	33.8

SampleJS performs best on TAC-KBP challenge.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
No active learning	38.0	30.5	33.8
Sample uniformly	34.4	35.0	34.7

SamplеJS performs best on TAC-KBP challenge.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
No active learning	38.0	30.5	33.8
Sample uniformly	34.4	35.0	34.7
Highest JS disagreement	46.2	30.8	37.0

SampleJS performs best on TAC-KBP challenge.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
No active learning	38.0	30.5	33.8
Sample uniformly	34.4	35.0	34.7
Highest JS disagreement	46.2	30.8	37.0
Sample JS disagreement	39.4	36.2	37.7

SampleJS performs best on TAC-KBP challenge.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

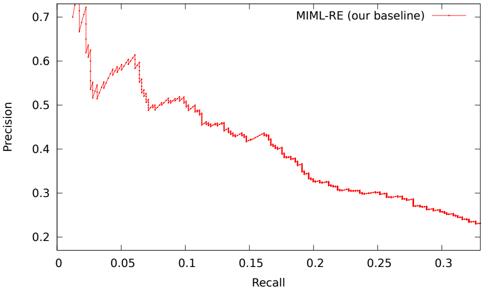
System	P	R	F_1
No active learning	38.0	30.5	33.8
Sample uniformly	34.4	35.0	34.7
Highest JS disagreement	46.2	30.8	37.0
Sample JS disagreement	39.4	36.2	37.7

- 2014 TAC-KBP Slot Filling Challenge: 27.6 $\rightarrow$ 32.0 F_1.

Good initialization is more important than constraining EM.

Is initialization or fixing latent *z*s during EM helping?

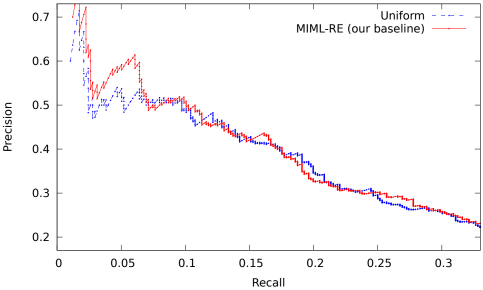
- What if we initialize with distant supervision?



Good initialization is more important than constraining EM.

Is initialization or fixing latent zs during EM helping?

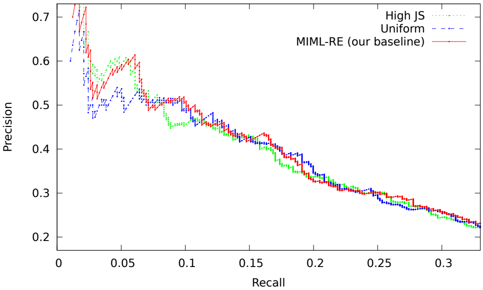
- What if we initialize with distant supervision?



Good initialization is more important than constraining EM.

Is initialization or fixing latent zs during EM helping?

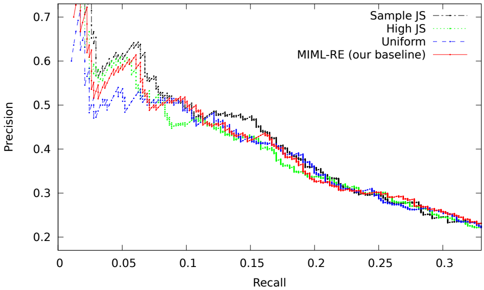
- What if we initialize with distant supervision?



Good initialization is more important than constraining EM.

Is initialization or fixing latent zs during EM helping?

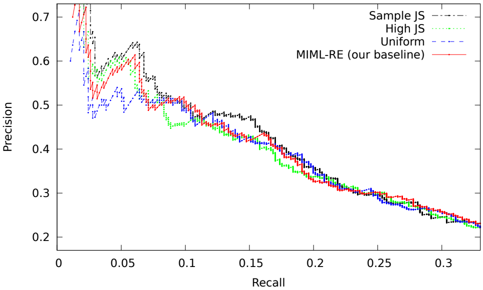
- What if we initialize with distant supervision?



Good initialization is more important than constraining EM.

Is initialization or fixing latent *z*s during EM helping?

- What if we initialize with distant supervision?



Hypothesis: Supervision not only *smooths the objective* but provides *better initialization* for the non-convex objective.

A supervised classifier performs surprisingly well.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
MIML-RE (baseline)	38.0	30.5	33.8

A supervised classifier performs surprisingly well.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
MIML-RE (baseline)	38.0	30.5	33.8
Supervised from SampleJS	33.5	35.0	34.2

A supervised classifier performs surprisingly well.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
MIML-RE (baseline)	38.0	30.5	33.8
Supervised from SampleJS	33.5	35.0	34.2
MIML-RE Supervised init.	35.1	35.6	35.5
SampleJS	39.4	36.2	37.7

A supervised classifier performs surprisingly well.

TAC-KBP 2013 Slot Filling Challenge:

- End-to-end task – includes IR + consistency.
- **Precision:** facts LDC evaluators judged as correct.
  Recall: facts other teams (including LDC annotators) also found.

System	P	R	F_1
MIML-RE (baseline)	38.0	30.5	33.8
Supervised from SampleJS	33.5	35.0	34.2
MIML-RE Supervised init.	35.1	35.6	35.5
SampleJS	39.4	36.2	37.7

- A bit circular: Need MIML-RE to get supervised examples.

A Case for Supervised Classifiers

Stanford's KBP system
(artist rendition)

Supervised Classifier
(150 lines + featurizer)

A Case for Supervised Classifiers

Stanford's KBP system
(artist rendition)

Supervised Classifier
(150 lines + featurizer)

A Case for Supervised Classifiers

Stanford's KBP system
(artist rendition)

Supervised Classifier
(150 lines + featurizer)

Annotating examples: $1330

A Case for Supervised Classifiers

Stanford's KBP system
(artist rendition)

Supervised Classifier
(150 lines + featurizer)

Annotating examples:	$1330
Flight to Qatar:	$1027
Apple 27" Screen:	$999

Conclusions

Things you can use:

- New active learning criterion: *Sampling* disagreement between a committee of classifiers.
- Corpus of supervised examples for TAC-KBP relations.
- 4.4 F_1 improvement on 2014 KBP Slot Filling.

Conclusions

Things you can use:

- New active learning criterion: *Sampling* disagreement between a committee of classifiers.
- Corpus of supervised examples for TAC-KBP relations.
- 4.4 F_1 improvement on 2014 KBP Slot Filling.

Things we've learned:

- Example selection is *very* important for performance.
- MIML-RE is sensitive to initialization.
- Supervised classifiers can perform similarly to distantly supervised methods.

Conclusions

Things you can use:

- New active learning criterion: *Sampling* disagreement between a committee of classifiers.
- Corpus of supervised examples for TAC-KBP relations.
- 4.4 F_1 improvement on 2014 KBP Slot Filling.

Things we've learned:

- Example selection is *very* important for performance.
- MIML-RE is sensitive to initialization.
- Supervised classifiers can perform similarly to distantly supervised methods.

Thank You!

Comparison to Pershina et al. (ACL 2014)

Slot filling evaluation of Surdeanu et al. (2012).