Semi-supervised methods of text processing, and an application to medical concept extraction

Yacine Jernite

Text-as-Data series

September 17. 2015

What do we want from text?

1. Extract information

2. Link to other knowledge sources

3. Use knowledge (Wikipedia, UpToDate,…)

How do we answer those questions?

1. What do people talk about on social media, and how? (Sentiment analysis)

2. What actions are described in a news article? (Semantic parsing)

3. In a medical setting: what symptoms does a patient exhibit?

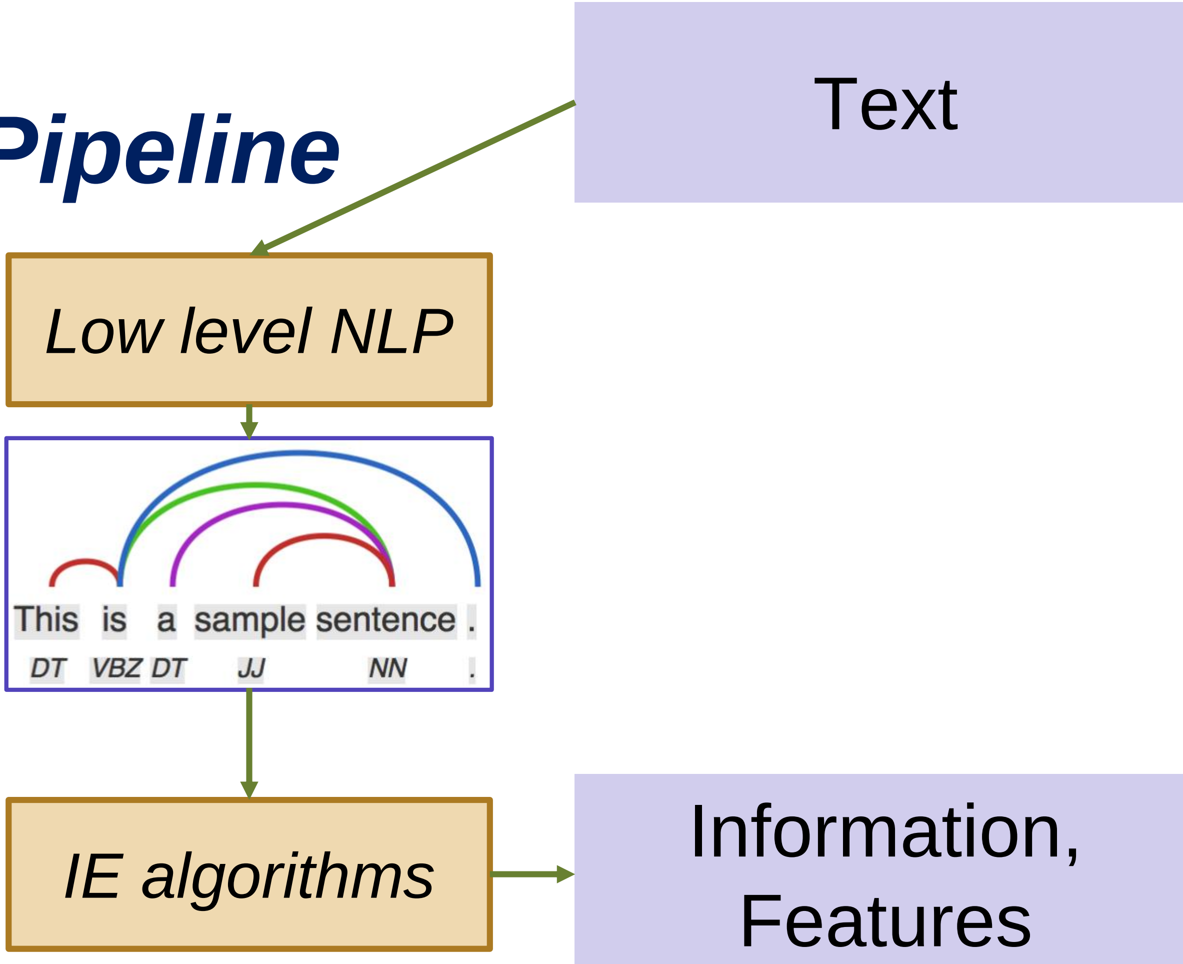
Pipeline

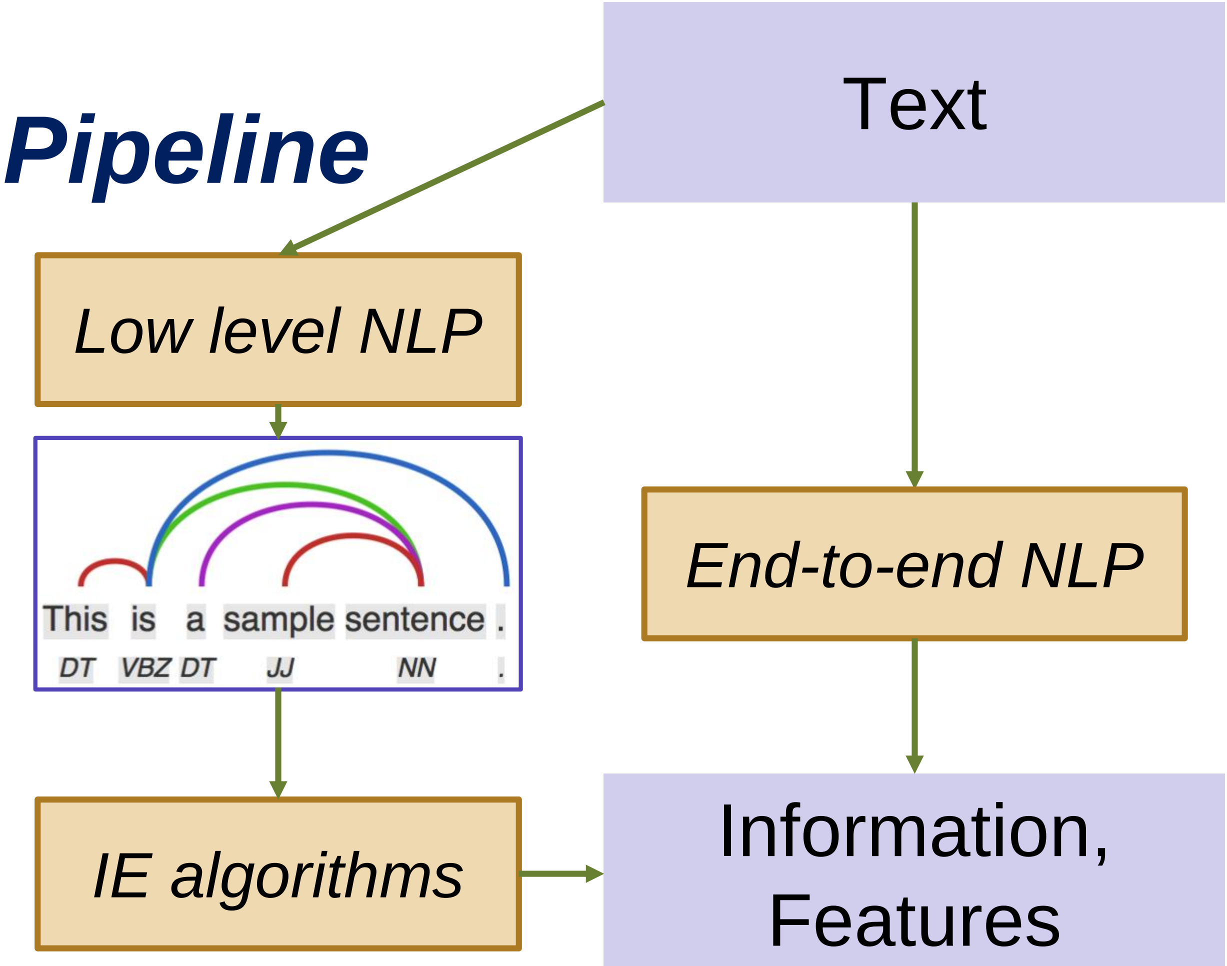
Pipeline

Text

Information, Features

Pipeline


Text
Low level NLP
This is a sample sentence .
DT VBZ DT JJ NN .
IE algorithms
Information,
Features


Pipeline
Text
Low level NLP
This is a sample sentence .
DT VBZ DT JJ NN .
IE algorithms
End-to-end NLP
Information,
Features

Machine learning approach

1. Specify task
2. Specify training algorithm
3. Get data
4. Train

Machine learning approach

1. Specify task

2. Specify training algorithm

3. **Get data**

4. Train

So much text, so few labels

- 5M English Wikipedia articles (3G words)
- 54M Reddit comments
- 1G Words in Gigaword dataset (newswire text)
- 5-grams from 1T words

So much text, so few labels

- 1M words in Penn TreeBank (parsing)
- Machine translation: highly language (and domain) dependent
- A few thousand to few hundred thousand sentence…
- And so many other custom tasks

Presentation outline

1. Literature review on semi-supervised paradigms
    a. Label induction
    b. Feature learning

2. Current work: Semi-Supervised Medical Entity Linking

Overview

Label induction

1. Labeling data is costly
2. Automatically obtain approximate labeling on larger dataset
3. Train using pseudo-labels

Overview

Feature learning

1. Feature quality affects accuracy
2. Learn features using other sources
3. Train with features on small labeled dataset

So much text, so few labels

- Label induction
- Feature learning
- Domain adaptation
- Multi-view learning

Overview

Labels

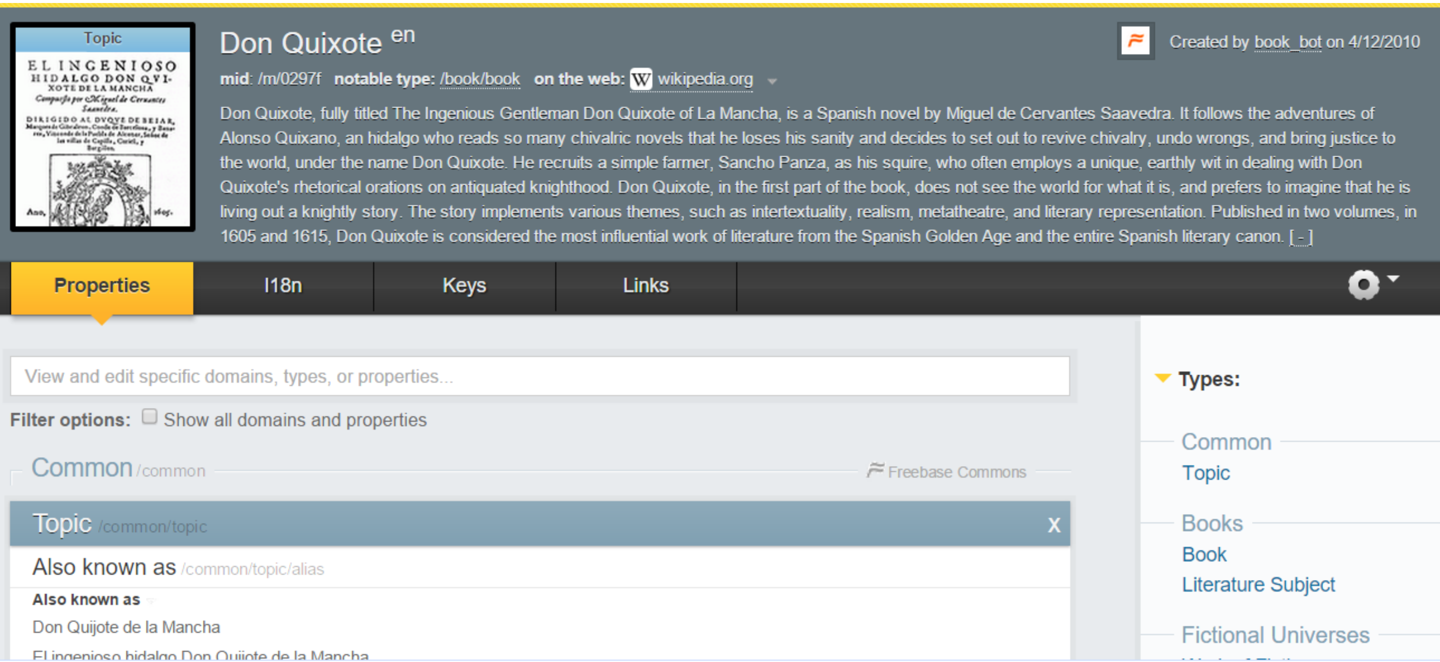
- **Fine Grained Entity Recognition**
  - *Ling and Weld, 2012*
- **Distant Supervision for RE with an incomplete KB**
  - *Min et al., 2013*
- **Co-Training for DA**
  - *Chen et al. 2011*
- **Semi-Supervised FSP for Unknown Predicates**
  - *Das and Smith, 2011*

Fine Grained Entity Recognition

- *Method type*: Automatic labeling
- *Task*: Identify entities in text, and tag them with one of 112 types
- *Labeled data*: Hand-labelled news reports
- *Auxiliary data*: Wikipedia, Freebase

Fine Grained Entity Recognition

Freebase



Fine Grained Entity Recognition

1. Automatically label entity spans in Wikipedia text

Don Quixote

…

Meaning

Harold Bloom says that *Don Quixote* is the writing of radical nihilism and anarchy,…

Fine Grained Entity Recognition

1. Automatically label Wikipedia text
   - Spans are obtained from hyperlinks
   - Types are obtained from Freebase

Don Quixote

…

Meaning

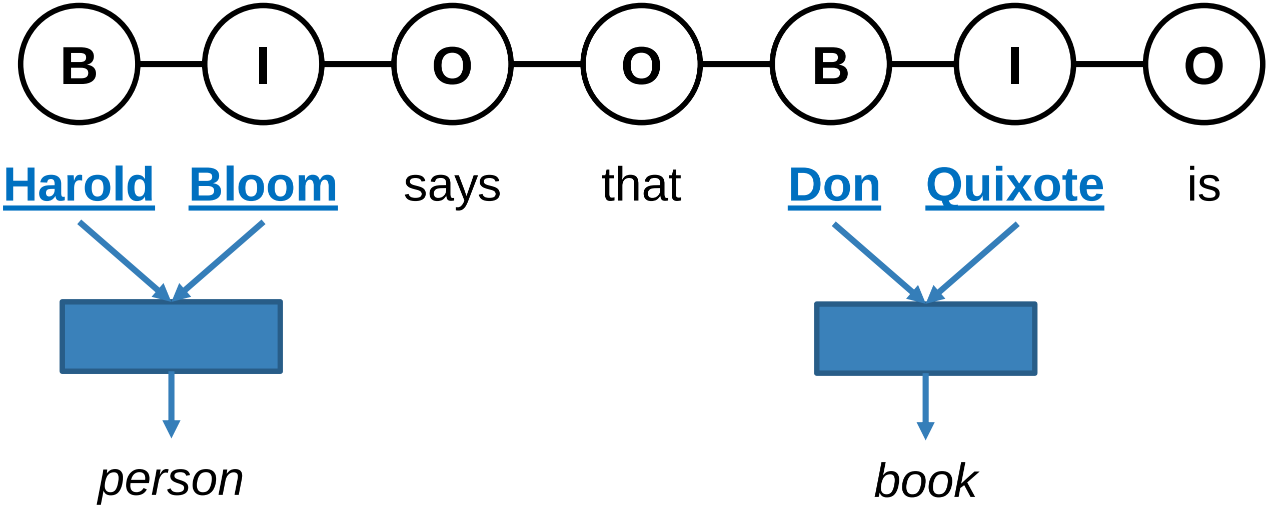
Harold Bloom says that *Don Quixote* is the writing of radical nihilism and anarchy,…

Harold Bloom:
Topic, Academic, Person, Author, Award winner, Influence node

Nihilism:
Topic, Field of study, Literature subject, Religion

Fine Grained Entity Recognition

1. Train CRF and perceptron on pseudo-labeled data



Fine Grained Entity Recognition

- Compares to
  - Stanford NER: 4 most common classes
  - Ratinov et al. Named Entity Linking

- Results:

Measure	Strict	Loose Macro	Loose Micro
NEL	0.220	0.327	0.381
Stanford (CoNLL)	0.425	0.585	0.548
FIGER	0.471	0.617	0.597
FIGER (GOLD)	0.532	0.699	0.693

Distant Supervision for Relation Extraction with an incomplete Knowledge Base

- *Method type*: Automatic labeling, Label inference

- *Task*: Relation extraction

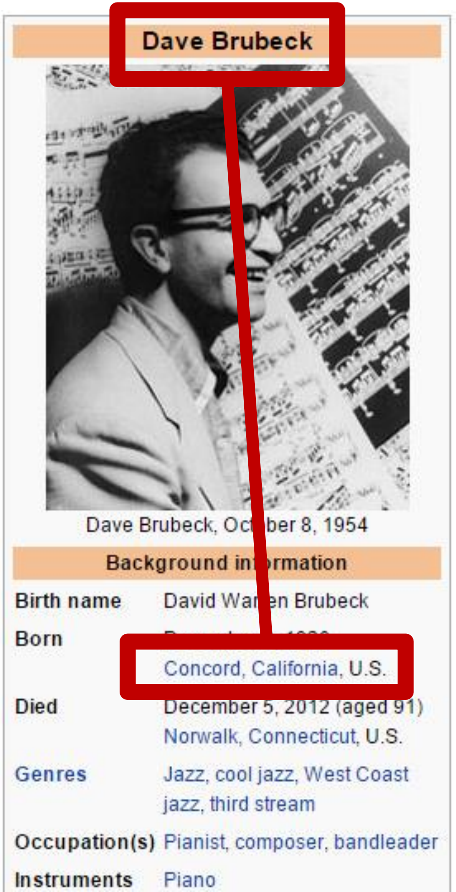
- *Labeled data*: TAC 2011 KBP dataset

- *Auxiliary data*: Wikipedia infoboxes, Freebase

Distant Supervision for Relation Extraction with an incomplete Knowledge Base

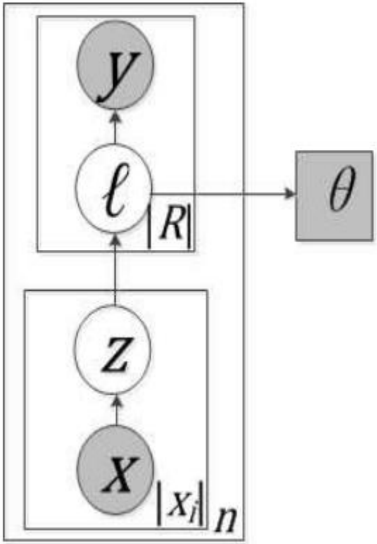
- Entity pairs extracted from Wikipedia infoboxes

- Labeled with FreeBase relations: **origin**



Distant Supervision for Relation Extraction with an incomplete Knowledge Base

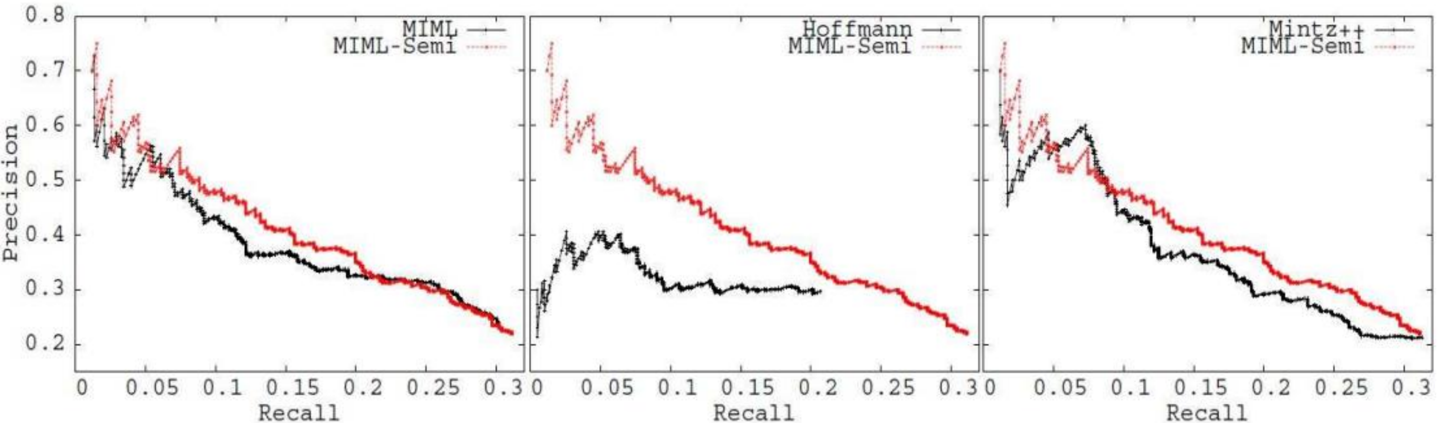
- Latent variable algorithm to learn from positive-only labels



- **X:** entity pair mention
- **Z:** mention level label
- **l:** bag level label
- **Y:** KB entity pair label
- **θ:** Number of positive labels

Distant Supervision for Relation Extraction with an incomplete Knowledge Base

- Learns with EM, compares to *(y = l)*

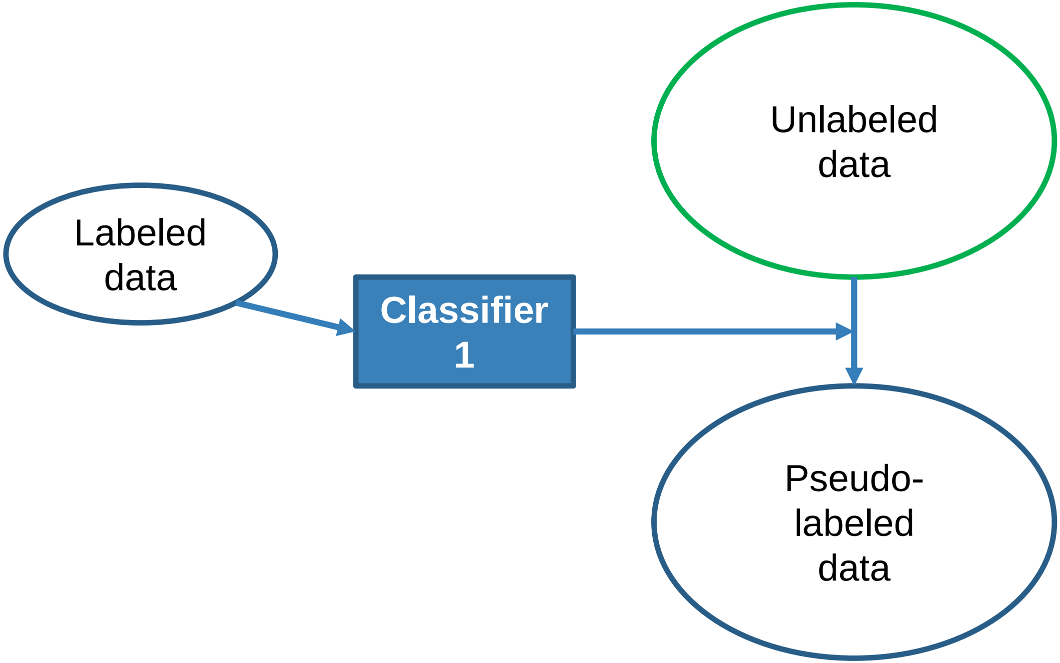


Co-Training for Domain Adaptation

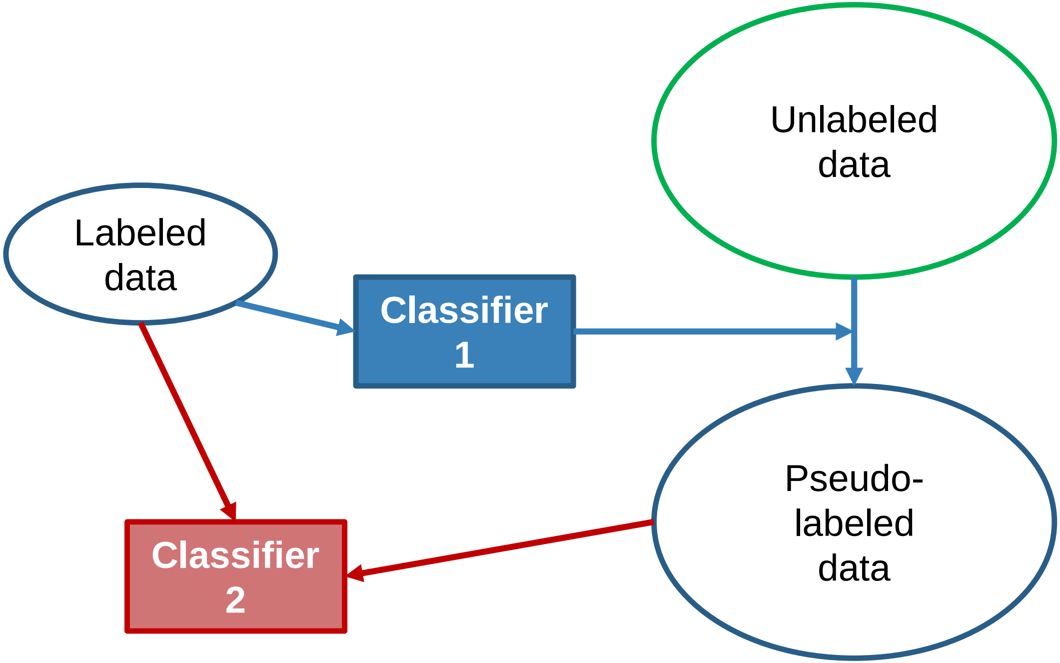
- *Method type*: <u>Automatic labeling</u>, <u>Domain adaptation</u>
- *Task*: Text classification - review polarity
- *Labeled data*: Amazon reviews for books, DVD, electronics, kitchen
- *Auxiliary data*: Cross-domain training

Self-Training

Self-Training



Self-Training

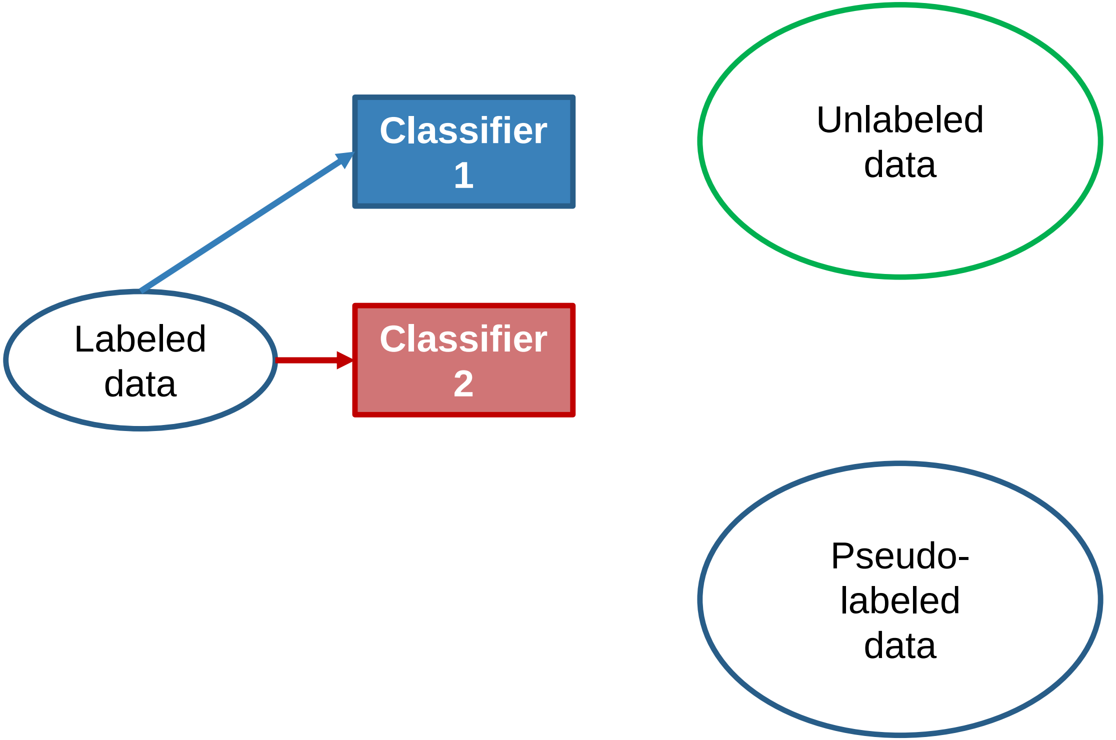


Self-Training

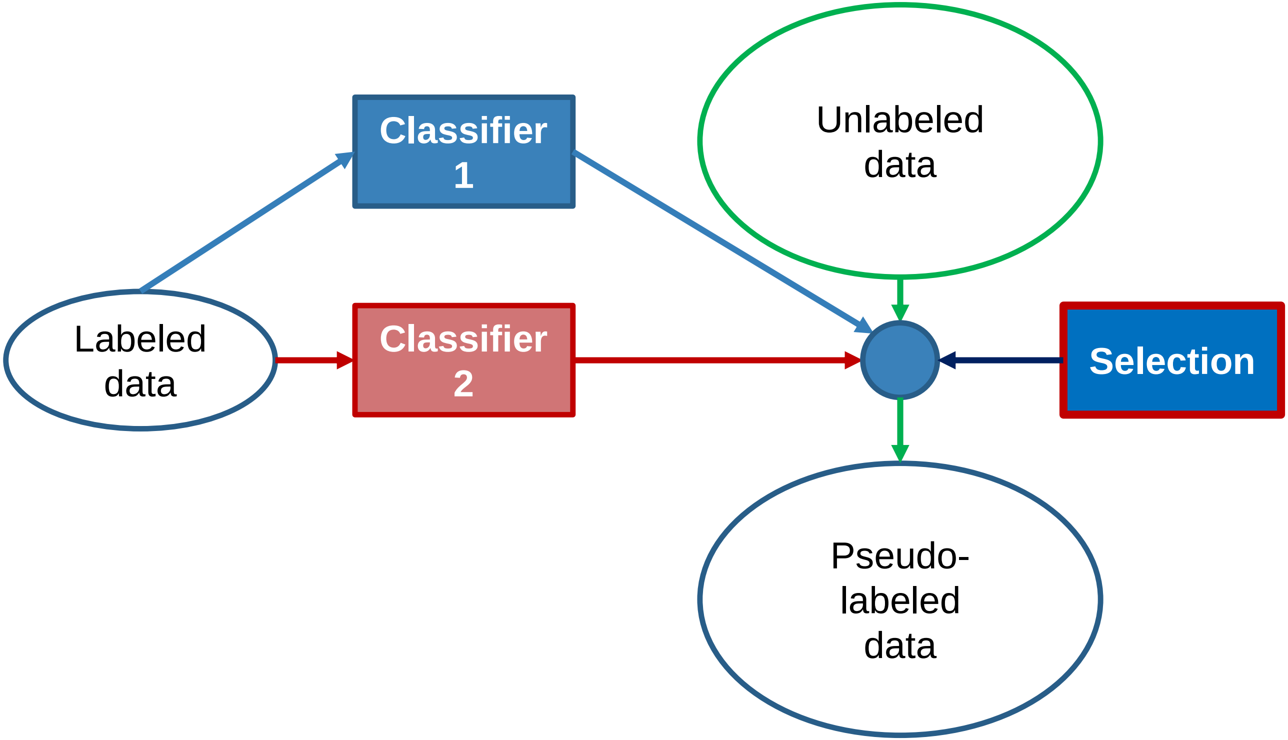
- Algorithm
  - Train System-1 on labeled data
  - Label some data with System-1
  - Train System-2 on combined data
- Not much improvement
  - Less than 1% parsing accuracy
  - Somewhat better “portability”

Co-Training

Co-Training

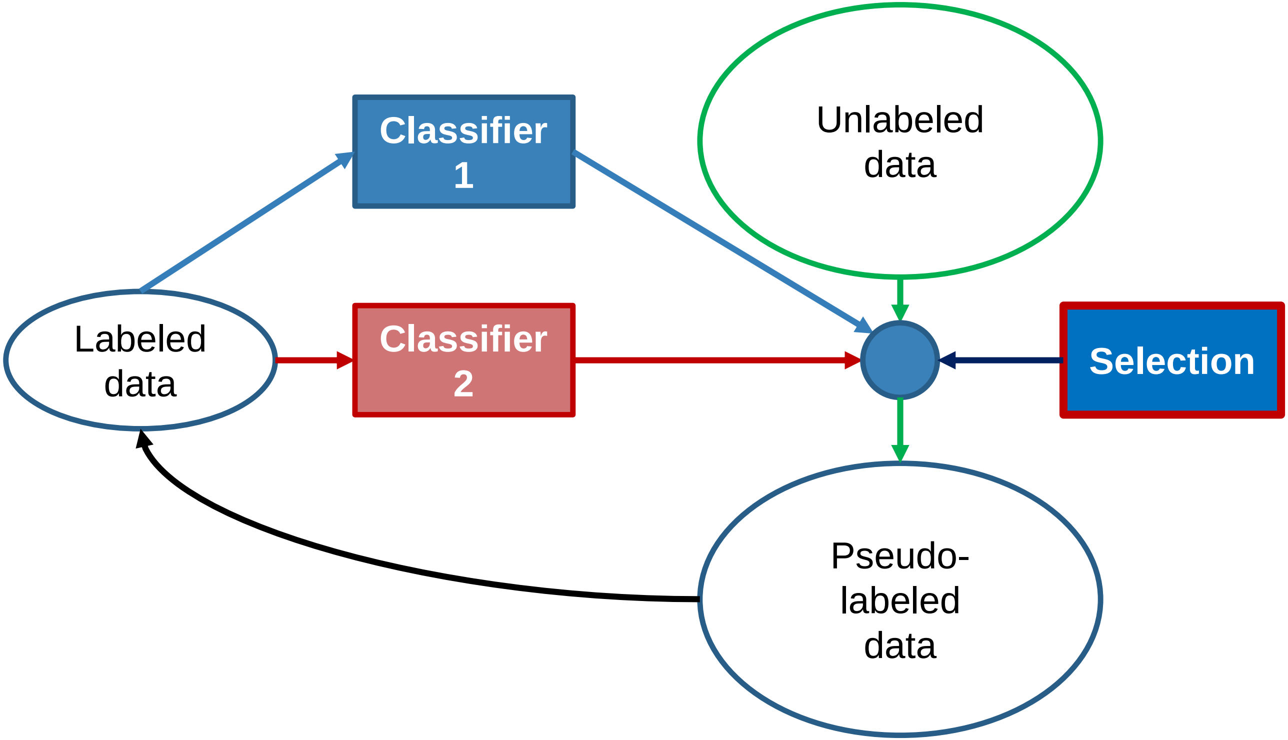

Classifier 1
Unlabeled data
Labeled data
Classifier 2
Pseudo-labeled data

Co-Training



Classifier
1
Unlabeled
data
Labeled
data
Classifier
2
Selection
Pseudo-
labeled
data

Co-Training


Classifier 1
Unlabeled data
Labeled data
Classifier 2
Selection
Pseudo-labeled data

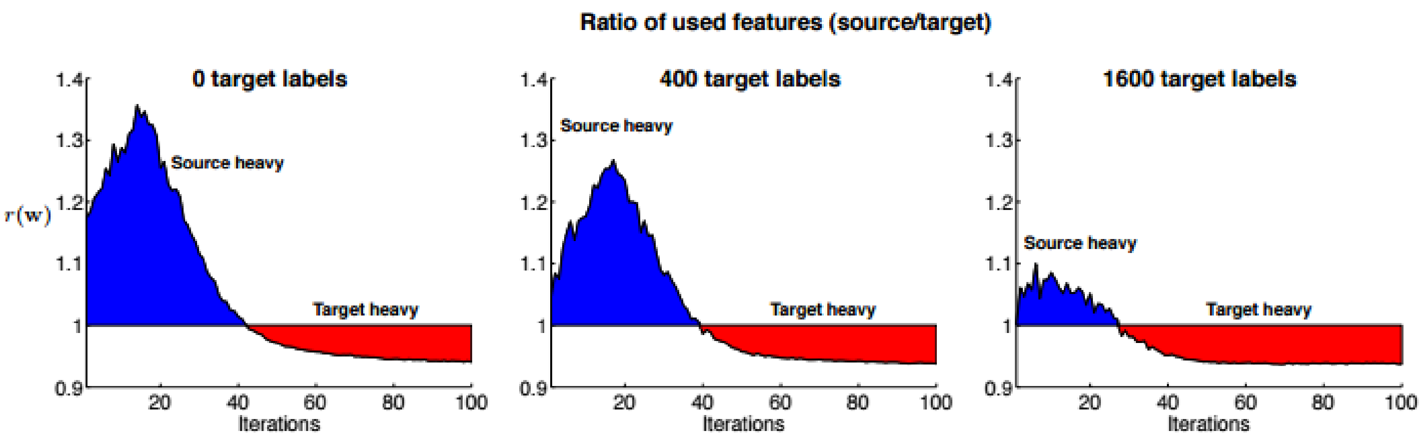
Co-Training

- Algorithm
  - Train System-1 and System-2 on labeled data with disjoint feature sets
  - Add data which is confidently labeled by exactly one system
  - Re-train, iterate

- Theoretical guarantees for "independent" feature sets

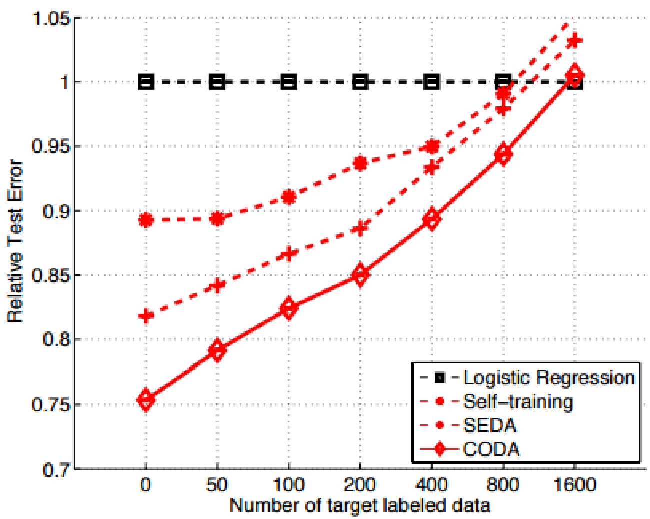
Co-Training for Domain Adaptation

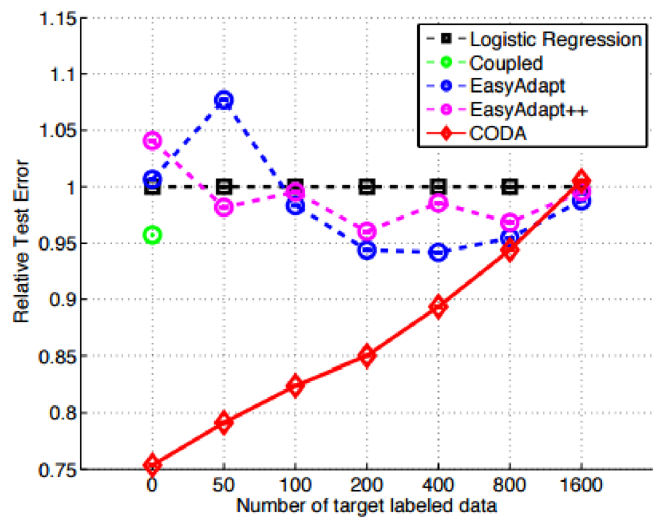
- L1 regularization: starts using more target-domain features



Co-Training for Domain Adaptation

- Best improvement adding a limited number of examples





Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- *Method type*: Label pre-selection
- *Task*: Frame-semantic parsing
- *Labeled data*: SemEval 2007
- *Auxiliary data*: Gigaword corpus, FrameNet

Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

Ted really **tried** to **read Infinite Jest**, but **was discouraged** by the size of the book.

Attempt

Definition:

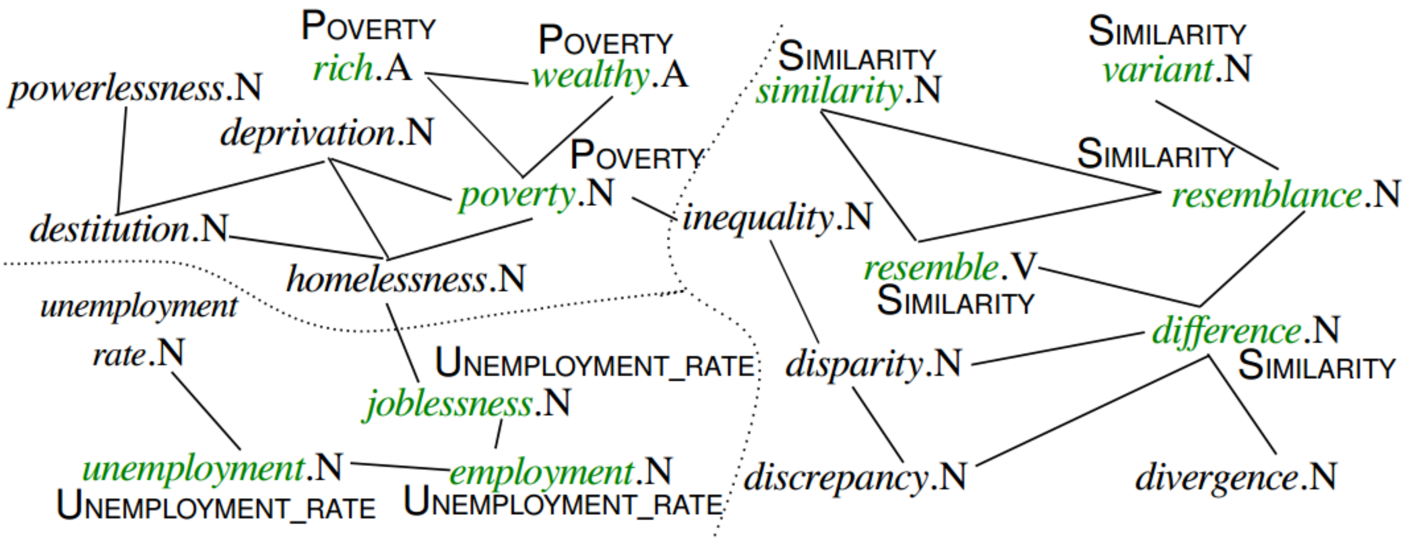
An Agent attempts to achieve a Goal. The Outcome may also be mentioned explicitly.

John ATTEMPTED to climb Mt. Everest.

It was another failed ATTEMPT to climb Mt. Everest.

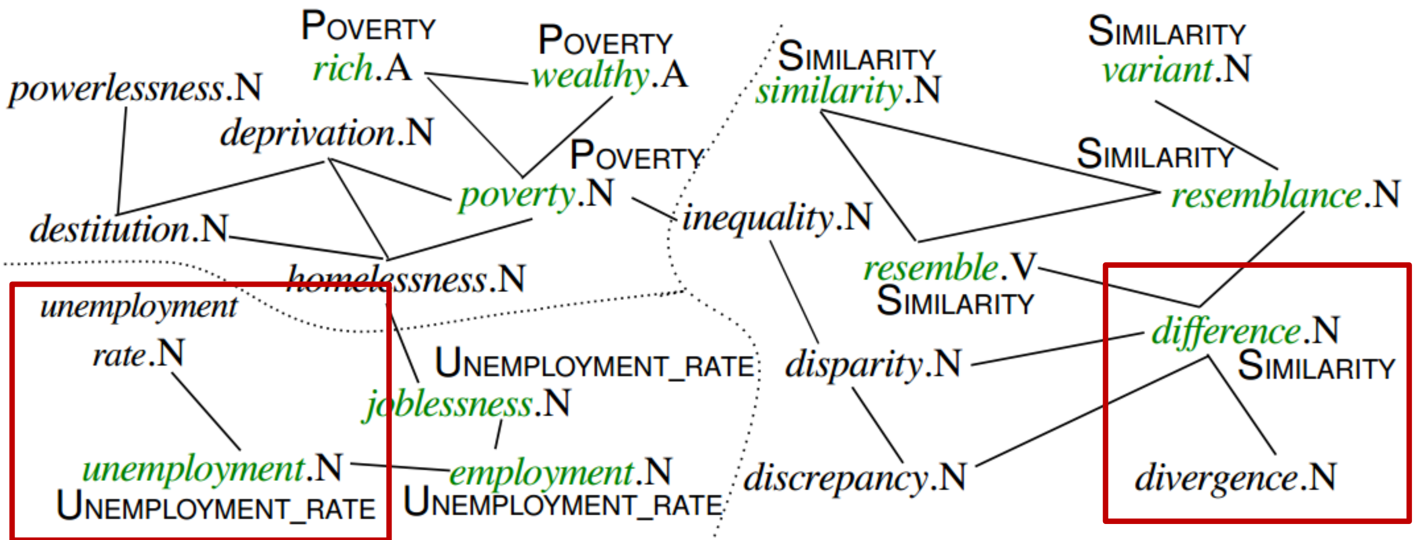
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Extracts possible frame targets from unlabeled data



Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Extracts possible frame targets from unlabeled data



Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Graph construction
  - Distance from dependency parsed text
  - About 60,000 targets (about 10,000 in FrameNet)
  - Convex quadratic optimization problem

Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Learned neighbor frame distribution

$t = \textit{discrepancy}$.N		$t = \textit{contribution}$.N		$t = \textit{print}$.V	
f	$q_t^*(f)$	f	$q_t^*(f)$	f	$q_t^*(f)$
*SIMILARITY	0.076	*GIVING	0.167	*TEXT_CREATION	0.081
NATURAL_FEATURES	0.066	MONEY	0.046	SENDING	0.054
PREVARICATION	0.012	COMMITMENT	0.046	DISPERSAL	0.054
QUARRELING	0.007	ASSISTANCE	0.040	READING	0.042
DUPLICATION	0.007	EARNINGS_AND_LOSSES	0.024	STATEMENT	0.028

Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Parsing results

Model	UNKNOWN TARGETS						ALL TARGETS					
	Exact Match			**Partial Match**			**Exact Match**			**Partial Match**		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
SEMAFOR	19.59	16.48	17.90	33.03	27.80	30.19	66.15	61.64	63.82	70.68	65.86	68.18
Self-training	15.44	13.00	14.11	29.08	24.47	26.58	65.78	61.30	63.46	70.39	65.59	67.90
LinGraph	**29.74**	**24.88**	**27.09**	**44.08**	**36.88**	**40.16**	**66.43**	**61.89**	**64.08**	**70.97**	**66.13**	**68.46**
FullGraph	**35.27***	**28.84***	**31.74***	**48.81***	**39.91***	**43.92***	**66.59***	**62.01***	**64.22***	**71.11***	**66.22***	**68.58***

Overview

Features

- **Prototype-Driven Learning for Sequence Models**
  - *Haghighi and Klein, 2006*
- **DA with Structural Correspondence Learning**
  - *Blitzer et al., 2006*
- **NLP (almost) from scratch**
  - *Collobert et al., 2011*
- **On Using Monolingual Corpora in NMT**
  - *Gulcehere et al., 2015*

Prototype-Driven Learning for Sequence Models

- *Method type*: Feature learning
- *Task*: POS tagging, Classified ads segmentation
- *Labeled data*: PTB/CTB, Classifieds
- *Auxiliary data*: Prototypes

Prototype-Driven Learning for Sequence Models

- Example prototypes:

Label	Prototypes
ROOMATES	roommate respectful drama
RESTRICTIONS	pets smoking dog
UTILITIES	utilities pays electricity
AVAILABLE	immediately begin cheaper
SIZE	2 br sq
PHOTOS	pictures image link
RENT	$ month *number*15*1
CONTACT	*phone* call *time*
FEATURES	kitchen laundry parking
NEIGHBORHOOD	close near shopping
ADDRESS	address carlmont *ordinal*5
BOUNDARY	; . !

Label	Prototype	Label	Prototype
NN	% company year	NNS	years shares companies
JJ	new other last	VBG	including being according
MD	will would could	-LRB-	-LRB- -LCB-
VBP	are 're 've	DT	the a The
RB	n't also not	WP$	whose
-RRB-	-RRB- -RCB-	FW	bono del kanji
WRB	when how where	RP	Up ON
IN	of in for	VBD	said was had
SYM	c b f	$	$ US$ C$
CD	million billion two	#	#
TO	to To na	:	– : ;
VBN	been based compared	NNPS	Philippines Angels Rights
RBR	Earlier duller	“	“ ‘ non-“
VBZ	is has says	VB	be take provide
JJS	least largest biggest	RBS	Worst
NNP	Mr. U.S. Corp.	,	,
POS	'S	CC	and or But
PRP$	its their his	JJR	smaller greater larger
PDT	Quite	WP	who what What
WDT	which Whatever whatever	.	. ? !
EX	There	PRP	it he they
”	”	UH	Oh Well Yeah

Prototype-Driven Learning for Sequence Models

- Gives prototypes of tag-token pairs
- Compute a similarity measure on tokens
- Adds similarity to the prototypes as a feature

Prototype-Driven Learning for Sequence Models

- Results:

	Num Tokens	
Setting	48K	193K
BASE	42.2	41.3
PROTO	61.9	68.8
PROTO+SIM	79.1	80.5

POS tagging

Classifieds segmentation

Setting	Accuracy
BASE	46.4
PROTO	53.7
PROTO+SIM	71.5
PROTO+SIM+BOUND	74.1

Domain Adaptation with Structural Correspondence Learning

- *Method type*: Feature learning, Multi-view learning, Domain adaptation
- *Task* : POS tagging
- *Labeled data*: MEDLINE (target domain)
- *Auxiliary data*: WSJ (source domain)

Domain Adaptation with Structural Correspondence Learning

- Example: pivot features *required*, *from*, *for*

(a) An ambiguous instance

JJ vs. NN			
with	normal	**signal**	transduction

(b) MEDLINE occurrences of signal, together with pivot features

the **signal** *required* to
stimulatory **signal** *from*
essential **signal** *for*

(c) Corresponding WSJ words, together with pivot features

of **investment** *required*
of **buyouts** *from* buyers
to **jail** *for* violating

Domain Adaptation with Structural Correspondence Learning

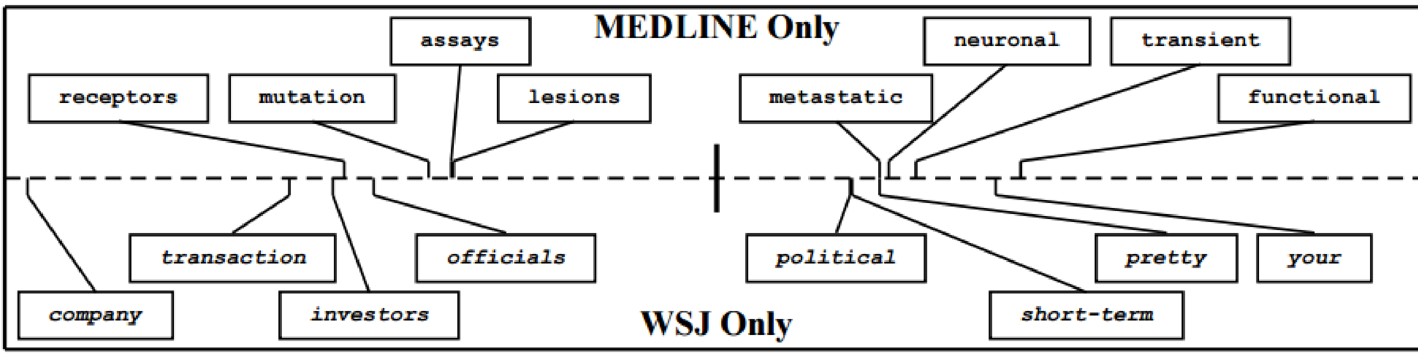
- Defines a set of pivot features, present in both source and target

- Sets up a set of mini-tasks: “predict the presence of pivot feature f”

- Runs SVD on the learned weights w_f

Domain Adaptation with Structural Correspondence Learning

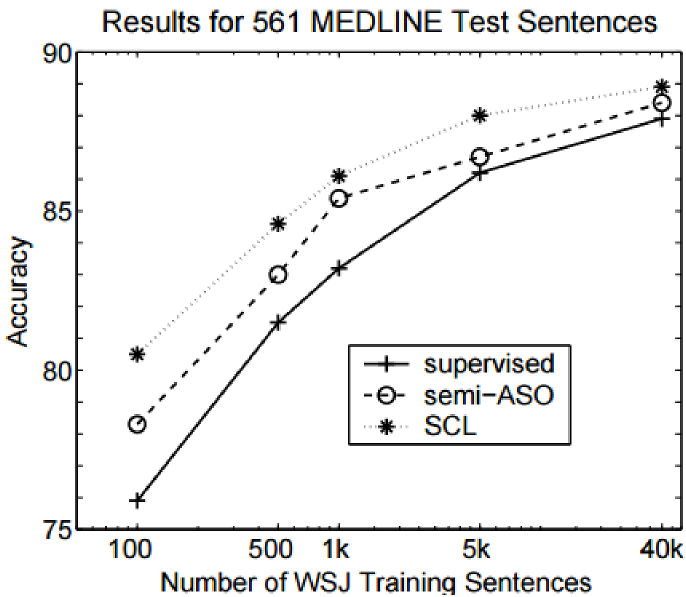
- Projection on first singular vector:



Domain Adaptation with Structural Correspondence Learning

- Results:

(a)



(b) Accuracy on 561-sentence test set

	Words	
Model	All	Unknown
Ratnaparkhi (1996)	87.2	65.2
supervised	87.9	68.4
semi-ASO	88.4	70.9
SCL	**88.9**	**72.0**

(c) Statistical Significance (McNemar's) for all words

Null Hypothesis	**p-value**
semi-ASO vs. super	0.0015
SCL vs. super	2.1×10^{-12}
SCL vs. semi-ASO	0.0003

NLP (almost) from Scratch

- *Method type*: Feature learning, Multi-view learning
- Task : POS, chunking, NER, SRL
- Labeled data: PTB, CoNLL
- Auxiliary data: 852M words from Wikipedia + Reuters

NLP (almost) from Scratch

- Neural network architecture

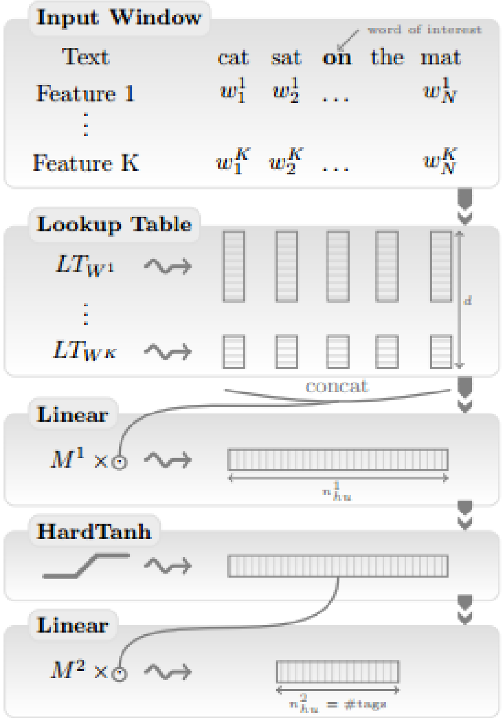


Figure 1: Window approach network.

NLP (almost) from Scratch

- First approach: supervised training of neural networks for tasks

Approach	POS (PWA)	Chunking (F1)	NER (F1)	SRL (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
NN+WLL	96.31	89.13	79.53	55.40
NN+SLL	96.37	90.33	81.47	70.99

NLP (almost) from Scratch

- Second approach: initialize with word representations from LM

Approach	**POS** (PWA)	**CHUNK** (F1)	**NER** (F1)	**SRL** (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
NN+WLL	96.31	89.13	79.53	55.40
NN+SLL	96.37	90.33	81.47	70.99
NN+WLL+LM1	97.05	91.91	85.68	58.18
NN+SLL+LM1	97.10	93.65	87.58	73.84
NN+WLL+LM2	97.14	92.04	86.96	58.34
NN+SLL+LM2	97.20	93.63	88.67	74.15

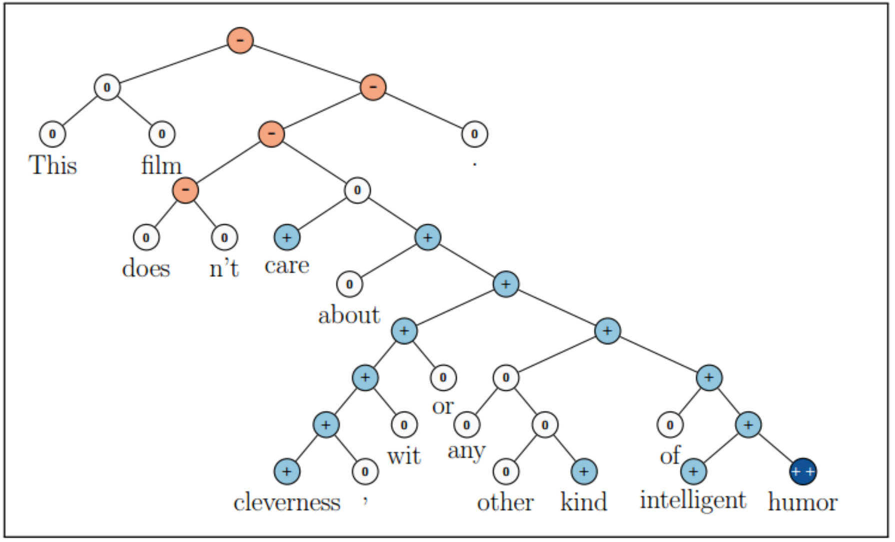
NLP (almost) from Scratch

- Finally: joint training

Approach	**POS** (PWA)	**CHUNK** (F1)	**NER** (F1)	**SRL** (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
	Window Approach			
NN+SLL+LM2	97.20	93.63	88.67	–
NN+SLL+LM2+MTL	97.22	94.10	88.62	–
	Sentence Approach			
NN+SLL+LM2	97.12	93.37	88.78	74.15
NN+SLL+LM2+MTL	97.22	93.75	88.27	74.29

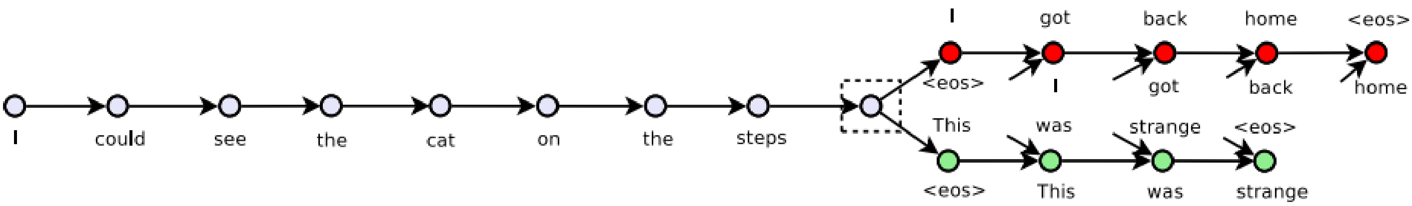
Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

- Sentiment analysis using word embeddings and syntactic parses



Skip-Thoughts Vectors (Kiros et al., NIPS 2015)

- Encodes sentences directly



- Improves sentence-level tasks
  - Classification
  - Paraphrase
  - Image-sentence ranking

On Using Monolingual Corpora in NMT

- *Method type*: <u>Feature learning</u>, <u>Target distribution</u>

- Task : Machine Translation

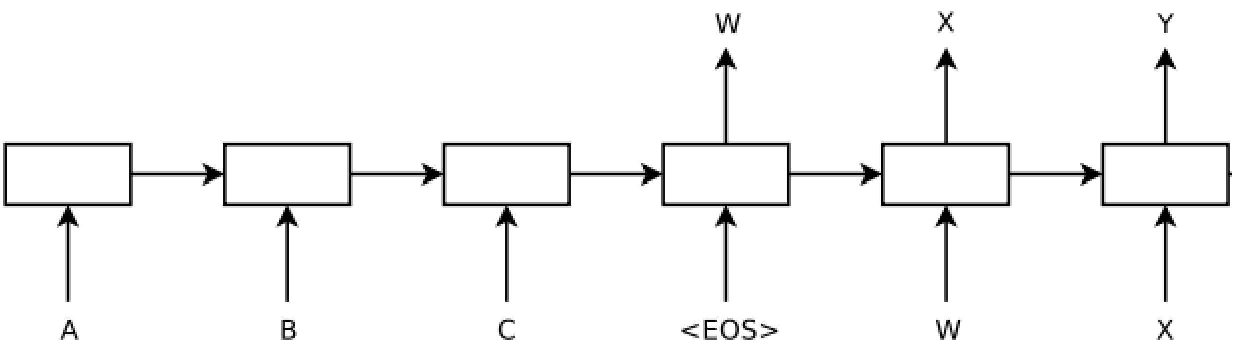
- Labeled data: Aligned text

- Auxiliary data: Monolingual corpora

On Using Monolingual Corpora in NMT

- Neural Machine Translation as sequence to squence modeling

- RNN ancoder and decoder:



On Using Monolingual Corpora in NMT

- Train Neural Machine Translation system
- Train target language model: RNN
- Shallow fusion: beam search on combined scores
- Deep fusion: add language model hidden state as input to decoder (+controller)

On Using Monolingual Corpora in NMT

	Test Set			
	tst2011	**tst2012**	**tst2013**	**Test 2014**
Previous Best **(Single)**	18.77	18.62	18.88	-
Previous Best **(Combination)**	18.83	18.93	18.70	-
NMT	18.40	18.77	19.86	18.64
NMT+LM (Shallow)	18.48	18.80	19.87	18.66
NMT+LM (Deep)	**20.17**	**20.23**	**21.34**	**20.56**

Turkish

On Using Monolingual Corpora in NMT

	SMS/CHAT		CTS	
	Dev	Test	Dev	Test
PB	15.5	14.73	21.94	21.68
+ CSLM	16.02	15.25	23.05	22.79
HPB	15.33	14.71	21.45	21.43
+ CSLM	15.93	15.8	22.61	22.17
NMT	17.32	17.36	23.4	**23.59**
Shallow	16.59	16.42	22.7	22.83
Deep	**17.58**	**17.64**	**23.78**	23.5

Chinese

	De-En		Cs-En	
	Dev	Test	Dev	Test
NMT Baseline	25.51	23.61	21.47	21.89
Shallow Fusion	25.53	23.69	21.95	22.18
Deep Fusion	**25.88**	**24.00**	**22.49**	**22.36**

Semi-Supervised Learning for Entity Linkage using Variational Inference

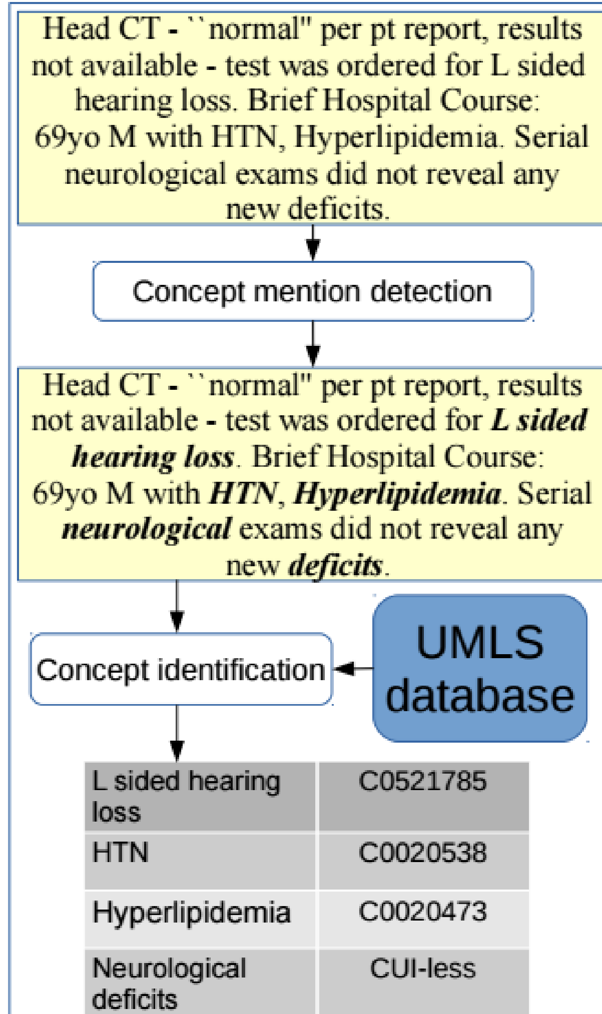
Yacine Jernite, Alexander Rush and David Sontag

Semi-Supervised Learning for Entity Linkage using Variational Inference

- *Method type*: Feature learning, Label inference
- *Task*: Medical concept extraction
- *Labeled data*: Semeval 2015 (annotated medical notes)
- *Auxiliary data*: MIMIC-II (medical text), UMLS

Task description

- We have:
  - Medical text from the MIMIC database
  - Medical knowledge base UMLS with concept descriptions
- We want to identify concepts in the text and link them to UMLS



UMLS samples

- Ambiguous, incomplete

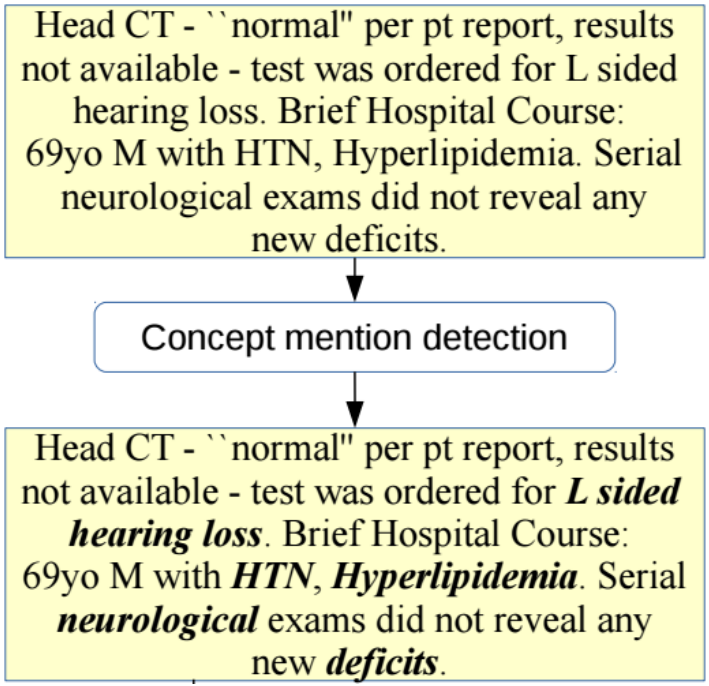
C0027627	C0002895	C0342788
neoplasm metastasis	anemia, sickle cell	renal carnitine transport defect
Neoplastic Process	Disease or Syndrome	Disease or Syndrome
metastases, neoplasm	transient abnormal myelopoiesis	carnitine uptake defect
metastasis	sickle cell anemia	systemic carnitine deficiency
secondaries	hemoglobin ss	scd
metastases	disease sickle-cell	primary carnitine defncy
tumor cell migration	scd	cud

UMLS samples

- Ambiguous, incomplete

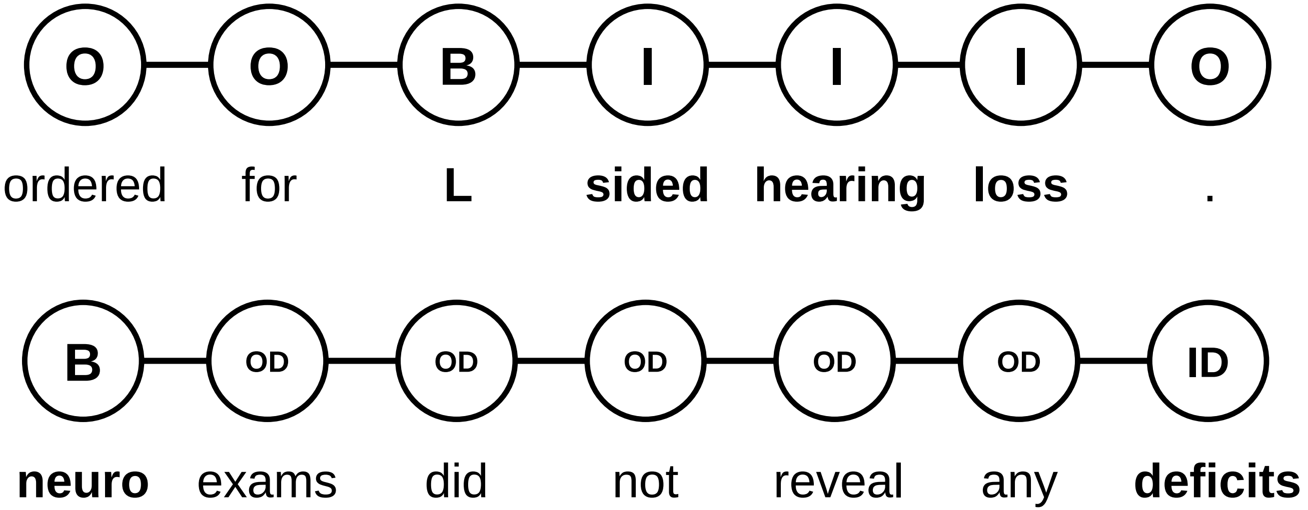
C0027627	C0002895	C0342788
neoplasm metastasis	anemia, sickle cell	renal carnitine transport defect
Neoplastic Process	Disease or Syndrome	Disease or Syndrome
metastases, neoplasm metastasis secondaries metastases tumor cell migration	transient abnormal myelopoiesis sickle cell anemia hemoglobin ss disease sickle-cell scd	carnitine uptake defect systemic carnitine deficiency scd primary carnitine defncy cud

Step 1: Mention Detection



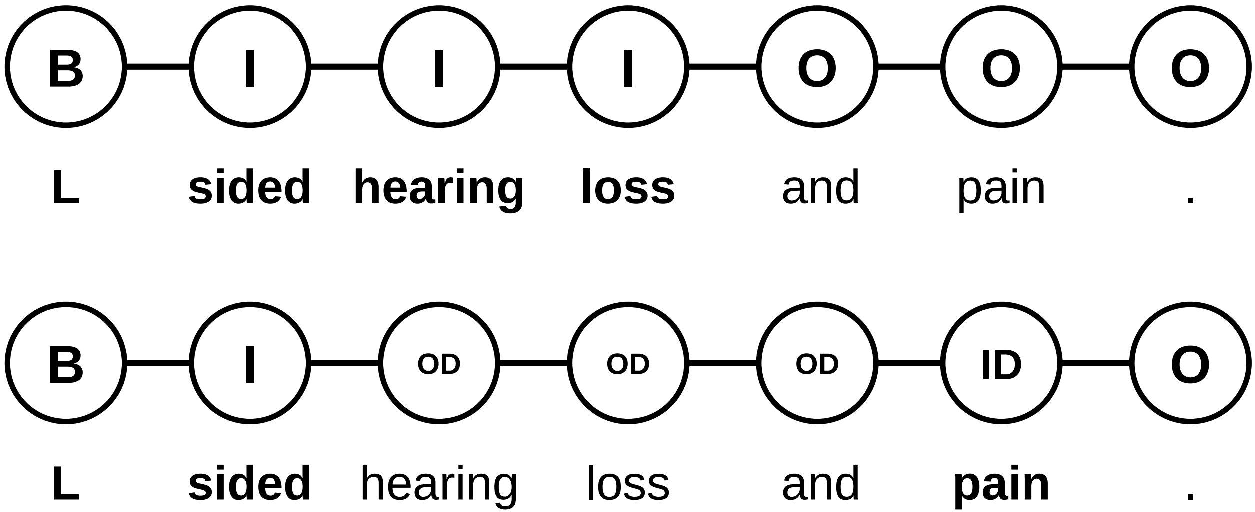
Step 1: Mention Detection

- B, I, O – ID, OD tagging with CRF



Step 1: Mention Detection

- Duplicating incompatible examples

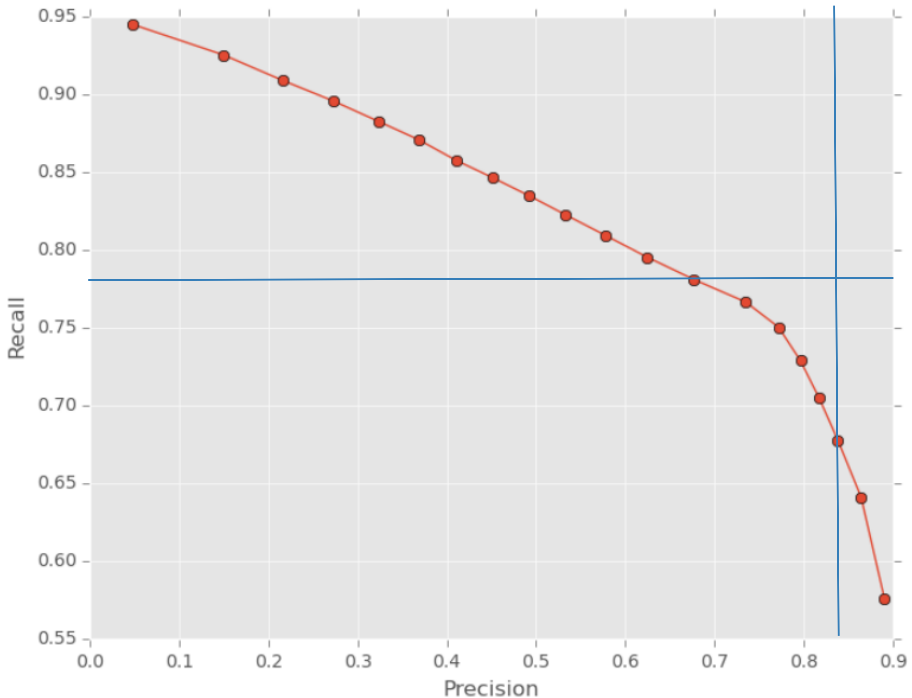


Step 1: Mention Detection

- Run inference on unlabeled and test set
- Approximate marginal probability
- Threshold

Step 1: Mention Detection

- PR curve:

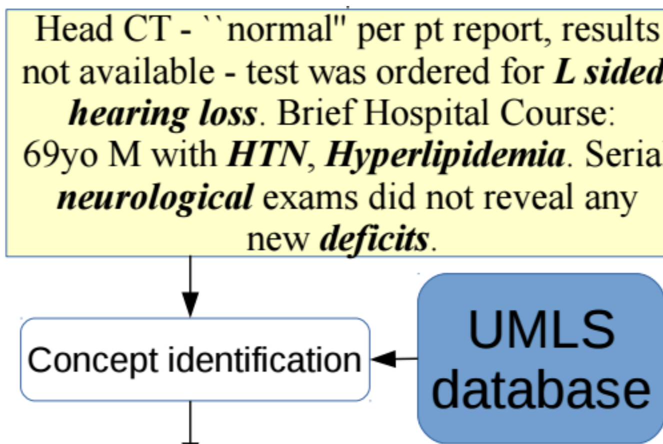


Step 1: Mention Detection

- Other approaches:
  - ezDI: A Supervised NLP System for Clinical Narrative Analysis, Pathak et al., 2015
  - BIO for continuous, SVM to join

  - ULisboa: Recognition and Normalization of Medical Concepts, Leal et al., 2015
  - BIOENS tagging scheme, Brown clusters, domain lexicons

Step 2: Mention Identification



L sided hearing loss	C0521785
HTN	C0020538
Hyperlipidemia	C0020473
Neurological deficits	CUI-less

Step 2: Mention Identification

- Pathak et al.:
  - Simple lookup
  - Semi-automated modified descriptions

CUI	Text	P1	P2	P3
C001 3132	Dribbling from mouth	Dribbling	from	mouth
C001 4591	Bleeding from nose	Bleeding	from	nose
C002 9163	Hemorr- hage from mouth	Hemo- rrhage	from	mouth
C039 2685	Chest pain at rest	Chest pain	at	rest
C026 9678	Fatigue during pregnancy	Fatigue	during	pregn ancy

  - Edit distance

Step 2: Mention Identification

- Leal et al.
  - Abbreviation dictionary
  - UMLS lookup
  - Similarity: Lucene, n-gram and edit distane
  - Lowest Information Content (specificity, using UMLS tree structure)

Step 2: Mention Identification

- A Generative Entity-Mention Model for Linking Entities with KB (Han and Sun, ACL 2011)

- $p(m, e) = p(s, c, e) = p(e)p(s|e)p(c|e)$

- $p(s|e)$: translation model from main description

- $p(c|e)$: unigram language model

Step 2: Mention Identification

- Our model:

- $p(m, e) = p(m|e)p(e)$

  - $p(m|e)$: multinomial with automatically curated support

  - $p(e)$: joint distribution on all entities in the document

Step 2: Mention Identification

- Our model:

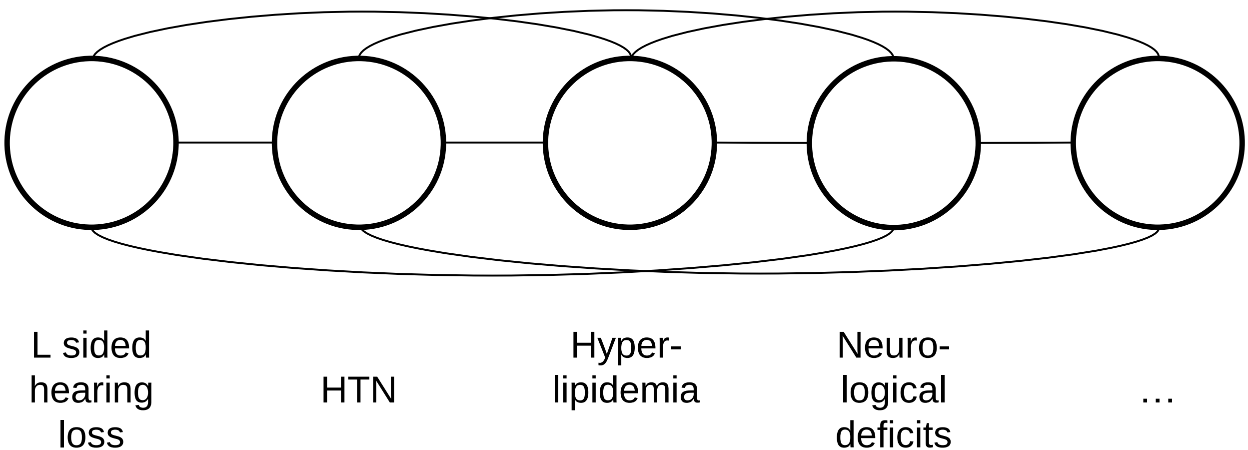
- $p(m, e) = p(m|e)p(e)$

    - $p(m|e)$: multinomial with automatically curated support

    - **$p(e)$: joint distribution on all entities in the document**

Step 2: Mention Identification

- $p(e)$: MRF on CUIs



Step 2: Mention Identification

- Problem: CUIs are latent variables on MIMIC (unlabeled)

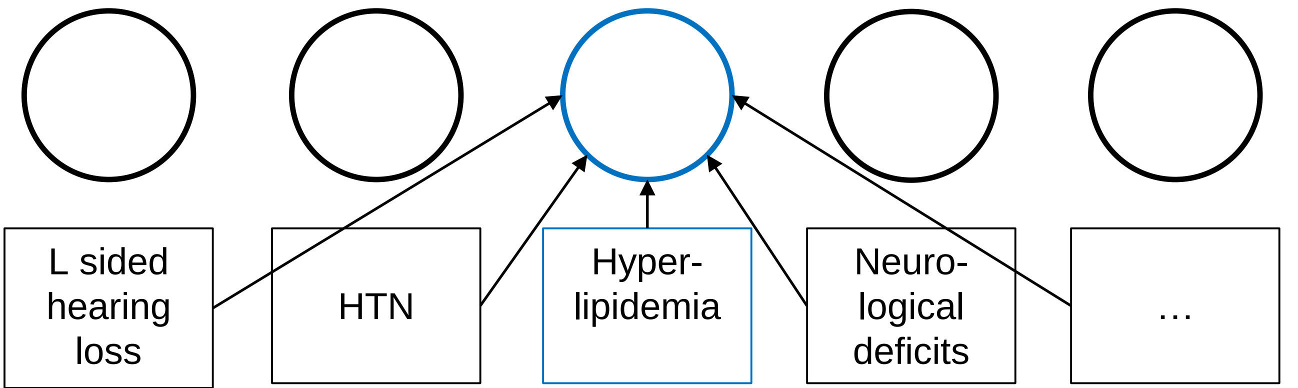
- Variational learning, following:
  - Autoencoding Variational Bayes, Kingma and Welling, ICLR 2014

Step 2: Mention Identification

- Objective:
  - Maximize $\log(\sum_e p(m,e;\theta))$
- Jensen's inequality:
  - $\forall q, \log(\sum_e p(m,e;\theta)) \geq \sum_e q(e|m,\xi)\log(\frac{p(m|e,\theta)}{q(e|m,\xi)})$
- Joint maximization in ξ, θ

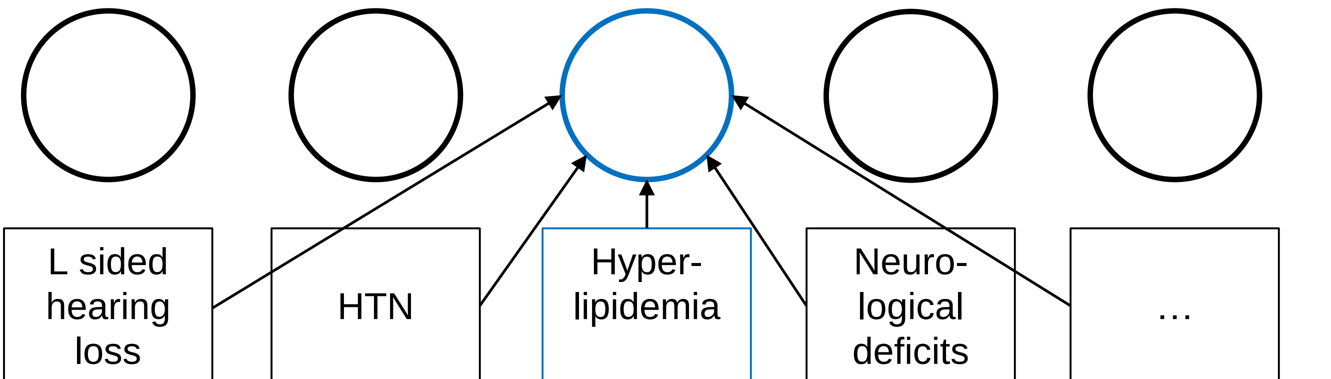
Step 2: Mention Identification

- Factorized q:
  - $q(e|m) = \prod_i q(e_i|m)$



Step 2: Mention Identification

- Considers mention and neighbors:
  - $q(e_i|m) = q(e_i|m_{i-2}, m_{i-1}, m_i, m_{i+1}, m_{i+2})$



Step 2: Mention Identification

- Neural network parameterization
- Semi-automated restricted support
- Supervised training gives 2nd best accuracy on 2014 task

Step 2: Mention Identification

- Next steps:
  - Pre-train parameters
  - Use correlation model
  - Train with variational algorithm

Review of Semi-Supervised methods

- Automatic labeling of data
- Label pre-selection
- Use prototypes
- Use features learned on larger corpus

Review of Semi-Supervised methods

- Domain adaptation: PubMed
- Multi-view learning

Review of Semi-Supervised methods

- Multi-view learning:
  - Other information on the patient: diagnosis codes, procedures, demographics, etc…
  - Jointly learn to predict those

Questions?