Semi-Supervised Large-Scale Learning for NLP

Pavel Kuksa
Rutgers University

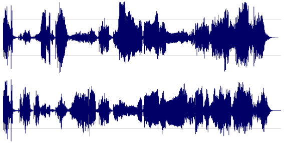
Joint work with:
Yanjun Qi, Bing Bai, NEC Labs
Vladimir Pavlovic, Rutgers University

What are we after?

Understanding of text or audio/music/image corpora

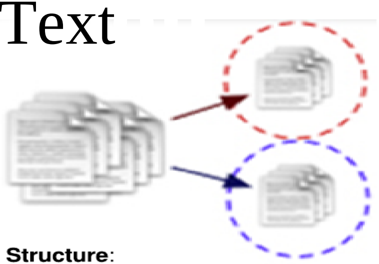
- Large-scale machine learning for matching, annotation, information extraction

Audio/Music



Music Genre
Artist
etc

Text



Bio-informatics

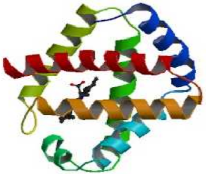
Sequence
VDAAVAKVCGSEAIKANLRRSWGVLSADIEA
TGLMLMSNLFTLRPDTKTYFTRLGDVQKGK
ANSKLRGHAITLTYALNNFVDSLDDPSRLKC
VVEKFAVNHINRKISGDAFGAIVEPMKETLKA
RMGNYYSDDVAGAWAALVGVVQAAL

predict

Class:
Globin-like

Function:
Oxygen transport

3D Structure:



This talk: annotation, information extraction for NLP

Natural Language Processing Tasks

Classical tasks:

- Part-of-speech (POS) tagging: noun, verb, adverb,…
- Chunking: noun phrase, verb phrase,…
- Named Entity Recognition (NER): person, company, location,…
- Semantic role labeling: object, subject, action, …

Practical Information Extraction tasks:

relationship extraction, text summarization, supporting/evidence sentence extraction, etc

Focus on practical tasks of understanding bio-medical texts (normal, e.g., Wiki-English, is a prior work)

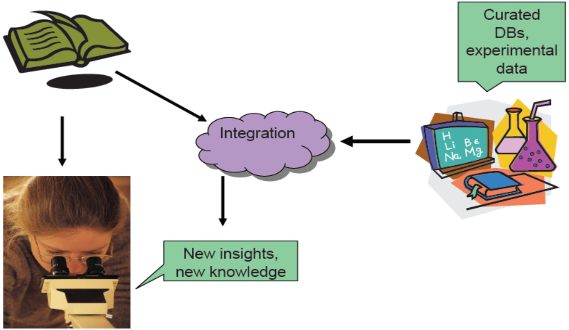
Why need natural language processing (NLP) for bio-medical literature (BioNLP)

MEDLINE: 70 million queries monthly, > 17M articles (wikipedia: 3M articles)

- Impossible to annotate manually

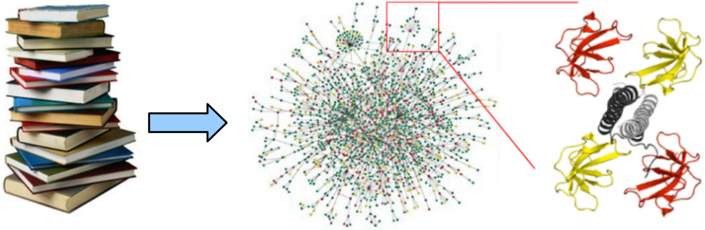
Linking text to databases

- Human curators struggle to process scientific literature
- Efficient access to discoveries/ facts/events crucial in sciences



Goal & Challenges

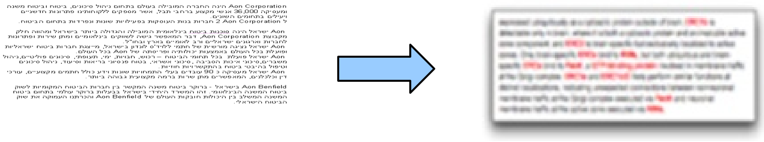
- Goal: Automatic annotation and information extraction from bio-medical texts
  - Bio-Entity Recognition,
  - Relationship Extraction from biomedical texts



- Challenges:
  - annotated data is scarce
  - millions of unannotated articles (e.g., MEDLINE)
  - Learn from unlabeled data with very limited prior knowledge

Three Tasks: Practical information extraction /retrieval problems

- Bio-Entity tagging (genes, proteins, etc)

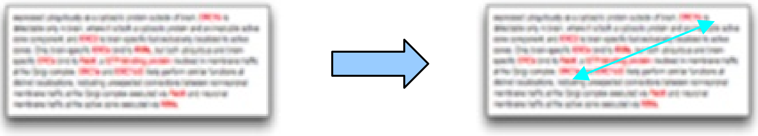


Aon Corporation הינה החברה המובילה בעולם בתחום ניהול סיכונים, ביטוח וביטוח משנה
ומעסיקה 36,000 אנשי מקצוע ברחבי תבל, אשר מספקים ללקוחותינו פתרונות חדשניים
ויעילים בתחומים השונים.
ל Aon Corporation 2 חברות בנות העוסקות בפעילויות שונות ונפרדות בתחום הביטוח.

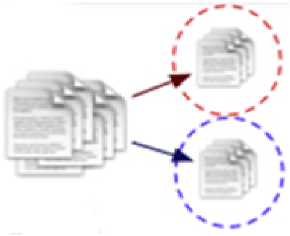
Aon ישראל הינה סוכנות ביטוח בינלאומית המובילה והגדולה ביותר בישראל ומהווה חלק
מקבוצת Aon Corporation, דבר המאפשר גישה לשווקים בינלאומיים ומתן שירות ופתרונות
לחברות וארגונים ישראליים ורב לאומיים בארץ ובחו"ל.
Aon ישראל נציגה מורשית של חתמי ללויד'ס לונדון בישראל, מייצגת חברות ביטוח ישראליות
ופועלת בכל העולם באמצעות יכולותיה ופריסתה של Aon בכל העולם.
Aon ישראל פועלת בכל תחומי הביטוח – רכוש, חבויות, ימי, תעופתי, סיכונים פוליטיים,ניהול
משברים,סיכוני איכות הסביבה ,סיכוני אשראי, בטוח פנסיוני בריאות וסיעוד, ניהול סיכונים
וטיפול בהיבטי ביטוח בהתקשרויות חוזיות.
Aon ישראל מעסיקה כ 90 עובדים בעלי התמחויות שונות וידע כולל חתמים מקצועיים, עורכי
דין וכלכלנים, המאפשרים מתן שרות ברמה מקצועית גבוהה ביותר.

Aon Benfield בישראל - ברוקר ביטוח משנה המקשר בין חברות הביטוח המקומיות לשוק
ביטוח המשנה הבינלאומי. זהו המשרד היחידי בישראל בבעלות ברוקר עולמי בתחום ביטוח
המשנה המשלב בין היכולות חובקות העולם של Aon Benfield והכרתנו העמוקה את שוק
הביטוח הישראלי.

- Protein-Protein Interaction (PPI) extraction

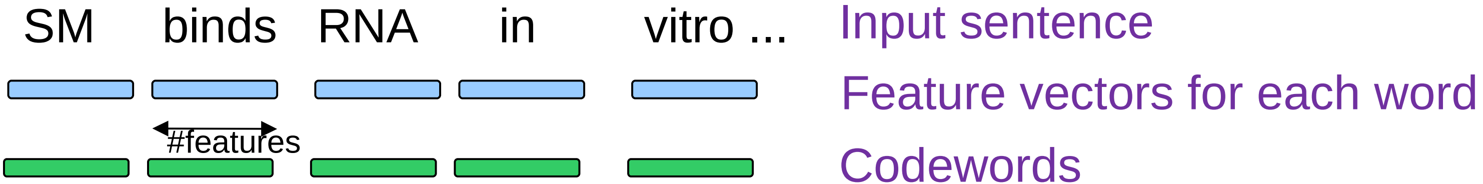


- PPI Article retrieval from abstracts (relevant article detection)



Preview of Results

- State-of-the-art learning systems for three BioNLP tasks
  - Step I: Use semi-supervised and unsupervised methods for learning word-level representations (feature vectors)
    - (1) Word-Class distribution (WCD) patterns
    - (2) Word Co-occurrence patterns
    - (3) Language Model derived word embedding
  - Step II: Use word codebooks (exemplar words) for word embedding



SM binds RNA in vitro ...

Input sentence

Feature vectors for each word

#features

Codewords

Step I (3) Unsupervised: Language Model

- **Language Model**: train low dimensional embedding for words (semantically similar words have close embeddings)
  - Positive examples: Text window extracted from unlabeled corpus (PubMed abstracts 95-present, 1.3G words)
    - trio and Abl **cooperate** in regulating axon
  - Negative examples: Text window with substitution of the middle word by a random word
    - trio and Abl **~~cooperate~~** in regulating axon
      random

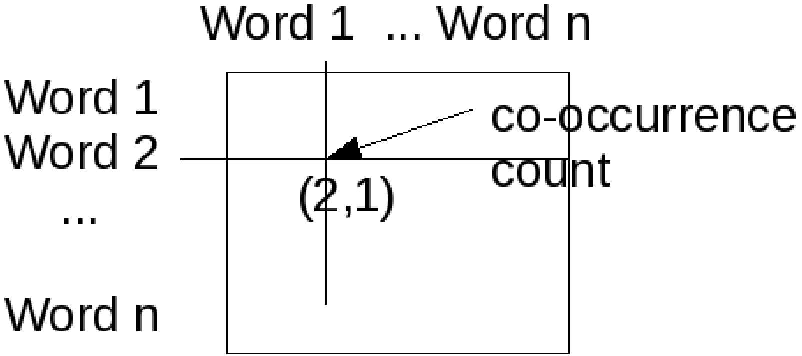
Collobert & Weston, ICML2008, A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning
Collobert & Weston & Kuksa, journal article (in submission)

Step I (2) Unsupervised: Word-Cooccurrence

- **Word Co-Occurrence**: group together words with similar (co)occurrence patterns (e.g., protein, kinase, pkc)

➔ Pairwise co-occurrence matrix (text window-based!)

[w_{-k} … w_{0} … w_{+k}]



Word 1 ... Word n

Word 1

Word 2

...

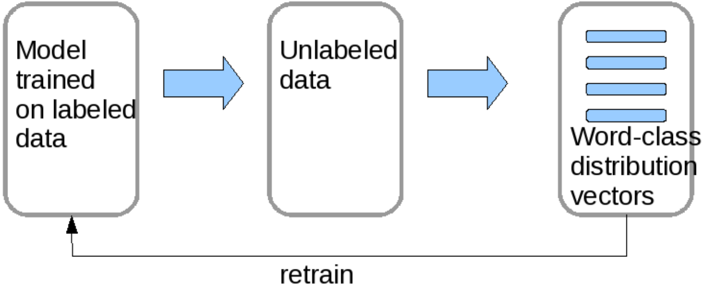
Word n

(2,1)

co-occurrence count

Step I (1) Semi-Supervised: Word-Class Distribution Learning (WCDL)

- **WCDL:** Simple and scalable semi-supervised feature learning
  - Use model trained on labeled examples to estimate (predicted) word-class distribution (WCD) patterns on unlabeled data


Model trained on labeled data
Unlabeled data
Word-class distribution vectors
retrain

  - Add WCD features to the feature set and retrain

Qi, Kuksa, Collobert, Sadamasa, Kavukcuoglu & Weston, ICDM 2009
"Semi-Supervised Sequence Labeling with Self-Learned Features"

Basic WCDL (Word Class Distribution Learning)

- Basic word-class distribution feature (estimated on unlabeled data)

  wcd(word) = [P(class$_1$|word) ... P(class$_n$|word)]

  (for *n*-class classification problem)

  P(class$_i$|word) = #times word is observed in class-i / total #times word is observed in the data

- Example: Using **IOBES** (inside, outside, begin, end, single) representation for the sequence labeling problem

  wcd(word) = [P(**I**|word), P(**O**|word), ... , P(**S**|word)]

- **wcd** features from neighboring words are highly informative for the word to be labeled ➔

Extended WCDL

- Estimate likelihoods for words to be around (i.e. before, after) the named entities
  - Targets unknown name recognition problem

- Extended WCD feature:

  extWCD(word) =

  [P(c_i|word), P(before c_i|word), P(after c_i|word)]

- Improves recognition on previously unseen words/names
  - Effective improvements under transductive setting as well

General WCDL

- Estimate likelihoods of label sequences for words
  - Captures word context better
  - Targets unknown name recognition problem
- General WCD feature:

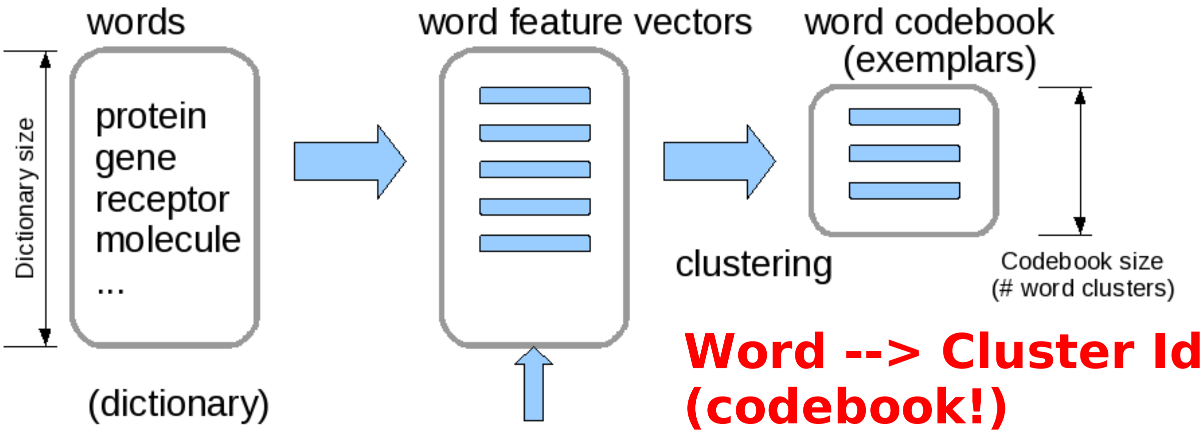
generalWCD(word) =

$$[P(c\_\underbrace{\{-k\}, c\_\{-k+1\},\ldots, c\_\{-1\}, c\_\{0\},c\_\{1\},\ldots,c}_{\text{label sequence (n-gram)}}\_\{k\}|\text{word})]$$

- Improves recognition on previously unseen words/names

Word Codebook Learning (WCL)

- Codebook learning



words

protein
gene
receptor
molecule
...

Dictionary size

(dictionary)

word feature vectors

word codebook
(exemplars)

clustering

Codebook size
(# word clusters)

Word --> Cluster Id (codebook!)

train by back-propagation (Language Model)
co-occurrence counts (Co-occurrence model)
word-class counts (Word-Class Distribution Learning)

Kuksa & Qi, SDM 2010 "Semi-Supervised Bio-Named Entity Recognition with Word-Codebook Learning"

Exemplar Word Embedding

- Query: protein (65)

Co-occurrence	WCDL	LM	SSI
protein	protein	protein	protein
kinase	family	receptor	expression
pkc	mutant	ligand	gene
ampk	antibody	molecule	cell
tyrosine	mutants	polypeptide	pNUMBER

- SSI (supervised semantic indexing),

Bing et al, CIKM 2009; Kuksa et al, ACL 2010

Experiment I: **Gene Mention Prediction (bioNER)**

- Find gene names in text
  - Input text: Phenotypic analysis demonstrates that **trio** and **Abl** cooperate in regulating axon outgrowth…
  - Output gene names: trio, Abl
- Data set:
  - BioCreative II competition
    - Train: 15K sentences from Medline abstracts
    - Test: 5K sentences
  - Unlabeled: 60M sentences (~1.3G words) from Pubmed (compare: Wikipedia 0.6G words)
- Evaluation:
  - precision, recall, F1 for gene names (phrases)

Gene Mention Prediction:
(1) Co-Occurrence

- Compute Dice scores between words from Co-occurrence matrix
- Cluster words using with affinity propagation (AP) method (Frey et al, 2007)
- Baseline (CRF):

Precision	Recall	F1
87.84	76.92	82.02

- +Co-occurrence:

Precision	Recall	F1	Improvement
88.52	79.42	83.72	+1.7 (2.1 %)

Gene Mention Prediction: (2) Basic WCDL

- Estimate WCD features for words on Medline abstracts using pre-trained supervised model
- Cluster WCD features with Vector Quantization (256 clusters)
- Baseline:

Precision	Recall	F1
87.84	76.92	82.02

- +Basic WCDL:

Precision	Recall	F1	Improvement
87.55	80.76	84.01	+1.99 (2.4 %)

- (Compare with co-occurrence):

Precision	Recall	F1	Improvement
88.52	79.42	83.72	+1.7 (2.1 %)

Gene Mention Prediction: (2) Extended WCDL

- Estimate extended WCD features on Medline abstracts using pre-trained supervised model (same model as in basic WCDL case)
- +Basic WCDL:

Precision	Recall	F1	Improvement
87.55	80.76	84.01	+1.99 (2.4 %)

- +Extended WCDL:

Precision	Recall	F1	Improvement
88.88	81.89	85.24	+3.22 (3.9 %)

- +General WCDL:

Precision	Recall	F1	Improvement
89.58	82.93	86.12	+4.08 (5 %)

Gene Mention Prediction: (3) Language Model

- Train Language Model on Medline abstracts
  - 1.3G words (60M sentences)
- Use ~40K dictionary
- Cluster with VQ (1024 clusters)
- +Language Model:

(2 months for 100K words on a single CPU)

Precision	Recall	F1	Improvement
89.19	82.89	85.93	+3.91 (4.8 %)

- Compare with General WCDL:

(hours)

Precision	Recall	F1	Improvement
89.58	82.93	86.12	+4.08 (5 %)

Gene Mention Prediction: multiple WCL (Language Model + WCDL)

- Use both Language Model and extended WCDL
- LM + extended WCDL:

Precision	Recall	F1	Improvement
90.12	84.39	87.16	+5.14 (6.3 %)

- Extended WCDL alone:

Precision	Recall	F1	Improvement
88.88	81.89	85.24	+3.22 (3.9 %)

- LM alone:

Precision	Recall	F1	Improvement
89.19	82.89	85.93	+3.91 (4.8 %)

Gene Mention Prediction: Language Model + WCDL + (word features)

- Use word features (prefix, suffix) with Language Model and extended WCDL
- LM + extended WCDL with word features:

Precision	Recall	F1	Improvement
90.7	85.19	**87.86**	+5.84 (7.1 %)

- LM + extended WCDL (no extra word features):

Precision	Recall	F1	Improvement
89.71	83.34	86.41	+4.39 (5.4 %)

- Previous best system with (many more) word features + POS, etc: F1 86.3

Gene Mention Prediction: adding domain knowledge

- Use NCBI human gene list (0.5M names)
- Use UNIPROT gene/protein names (1M names)
- LM + extended WCDL + gene names:

Precision	Recall	F1	Improvement
90.74	85.74	**88.17**	+6.15 (7.5 %)

- Compare with LM + extended WCDL

Precision	Recall	F1	Improvement
90.7	85.12	87.86	+5.84 (7.1 %)

Gene Mention Prediction: Transductive setting

- Estimate WCD features on test set using model trained on a train set (fast, ~minutes)
- Extended WCDL (transductive):

Precision	Recall	F1	Improvement
88.77	81.01	84.73	+2.71 (3.3 %)

- Compare with extended WCD (Pubmed):

Precision	Recall	F1	Improvement
88.88	81.89	85.24	+3.22 (3.9 %)

- Compare with LM (Pubmed):

Precision	Recall	F1	Improvement
89.19	82.89	85.93	+3.91 (4.8 %)

Gene Mention Prediction Results

Model	Precision	Recall	F1	Improvement
Baseline (Supervised)	87.84	76.92	82.02	
Co-occurrence	88.52	79.42	83.72	+1.7 (2.07 %)
Word-Class-Distributions	88.88	81.89	85.24	+3.22 (3.9 %)
Language Model	89.19	82.89	85.93	+3.91 (4.8 %)
Language Model + Word-Class-Distribution	89.71	83.34	86.41	+4.39 (5.4 %)
*Language Model	90.31	84.54	87.33	+5.31 (6.5 %)
*Language Model+Word-Class-Distribution	90.57	84.93	87.66	+5.64 (6.9 %)
*Language Model+Word-Class+Gene Names	**90.74**	**85.74**	**88.17**	+6.15 (7.5 %)

State-of-the-art gene name recognition performance

Previous best system: 87.21 F1 (complex combination of many classifiers with many more features, dictionaries, etc)

Methods Comparison

- WCDL: single pass over data + (re)training
  - task-focused
  - Time: ~few hours on Pubmed (1.3G words), ~minutes in transductive setting
- Co-occurrence: single pass over data (~ few hours on Pubmed)
  - Task-independent
  - Domain-sensitive
- LM: multiple passes over data (~ 2 month on Pubmed)
  - Task-independent
  - Domain-sensitive: Wiki-English vs Biomedical

Gene Mention Prediction (bioNER): Summary of Results

- State-of-the performance with word features only (87.86 F1 score),
  - 30% reduction in FN, 15% reduction in FP
- Single classifier (as opposed to complex combinations/cascades used by top systems)
- Complex WCDL can be combined with simple models (online prediction)
- System performance can be further improved with better unknown name detection

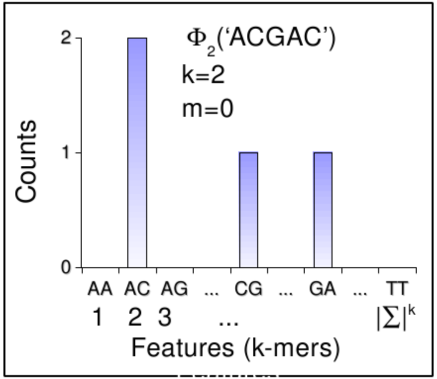
Experiment II & III: Protein-Protein Interaction (PPI) Recognition

- Interaction Article Retrieval: Identify relevant articles about PPI from *abstracts*

- PPI relation recognition: extract pairs of interacting proteins from sentences
  - Example: The protein product of **c-cbl** proto-oncogene is known to interact with several proteins, including **Grb2**, **Crk**, and **PI3 kinase**, and is though to regulate signaling …
    - Interacting pairs: (c-cbl, Grb2), (c-cbl, Crk), etc.

String Kernel for PPI recognition (relationship extraction)

- Use fixed-length feature vectors to represent arbitrary long strings X
- Examples:
  - Word kernel: dot-product of individual word counts
  - Spectrum kernel: dot-product of k-word counts
  - Mismatch kernel: dot-product of k-word counts with inexact matching of k-words
  - Gapped kernel: dot-product of (non-contiguous) k-word counts with gaps allowed between words



Kuksa, Huang, and Pavlovic, NIPS2008, "Scalable Algorithms for String Kernels with Inexact Matching"

String Kernels with Unlabeled data (open research)

- Use word index (codeword) from the codebook to represent the word
  - word -> cluster id
- Evaluate string kernel over codebook representation of a sentence instead of using words directly
  - <word*1* cluster id> <word*2* cluster id> … <word*n* cluster id>

Experiment II: PPI relation extraction from sentences

- Data: AIMed dataset
  - 951 positive examples, 3075 negative
  - 10-fold cross-validation
  - Relation data generation: for a sentence with *n* entities create C(n,2) copies with only 2 entities
- Task: extract all interacting pairs from given sentence (assuming protein/gene name labeling is known)
- Evaluation: F1 score

Experiment II : PPI relation extraction from sentences

Results:

Method	**Precision**	**Recall**	**F1**
WCL (LM)	**61.18**	67.92	64.33
WCL (SSI)	60.68	**69.08**	**64.54**
Baseline 1: (best) Multiple kernel, multiple parser combination (Makoto et al, 2008)	57.8	66.11	61.4
Baseline 2: Dependency and deep parsers (Miyao et al, 2008)	54.9	65.5	59.5

Experiment III : Interaction Article retrieval (relevant article detection)

- Data: BioCreative II competition
  - Train: 3536 negative, 1959 positive abstracts
  - Test: 338 positive, 339 negative
- Binary classification: identify abstracts for articles with experimental evidence for *protein-protein* interaction (not just any interaction)
- Evaluation: F1 score, Accuracy

Experiment III: Interaction Article retrieval (relevant article detection)

Results

Method	**Precision**	**Recall**	F1
WCL (LM)	76.06	84.62	**80.11**
WCL (SSI)	73.59	84.91	78.85
Baseline 1:(best) BioCreative II	70.31	87.57	78.00

- Current best system: F1 78.00 (many more hand-crafted & syntactic features)

Conclusions & Future Work

State-of-the-art learning systems for entity tagging and relationship extraction

- Learning word representations from unlabeled data improves prediction/extraction performance
- Words only, NO syntactic or other complex features
- End-to-end systems (robust, no cascades)
- Future extensions:
  - learning from weakly labeled data (citation graphs, keyword annotations, etc)
  - other tasks (article summarization, evidence sentence retrieval)

Conclusions & Future Work

Search, information retrieval, multi-modal data analysis

- Efficient means (algorithms, models) for analysis and modeling of complex data (text, image, multimedia)
- Large scale matching, annotation, information extraction
- High-dimensional data indexing, embedding

BACKUP SLIDES

Related Work (BioNLP)

- Gene mention recognition
  - Dictionary based
  - Rule based
  - Machine learning systems (CRF)
- PPI article retrieval: SVM on features
  - bag-of-words + bag-of-NLPs (chunk; phrase; pos; protein mention; non-proteins; title phrase, et al.)
- PPI event detection
  - Computational linguistic-based methods (e.g. SRL type)
  - Rule/pattern based methods
  - Machine learning based methods (e.g. co-occurrence)

Related Work (Word Codebook Learning from Unlabeled Data)

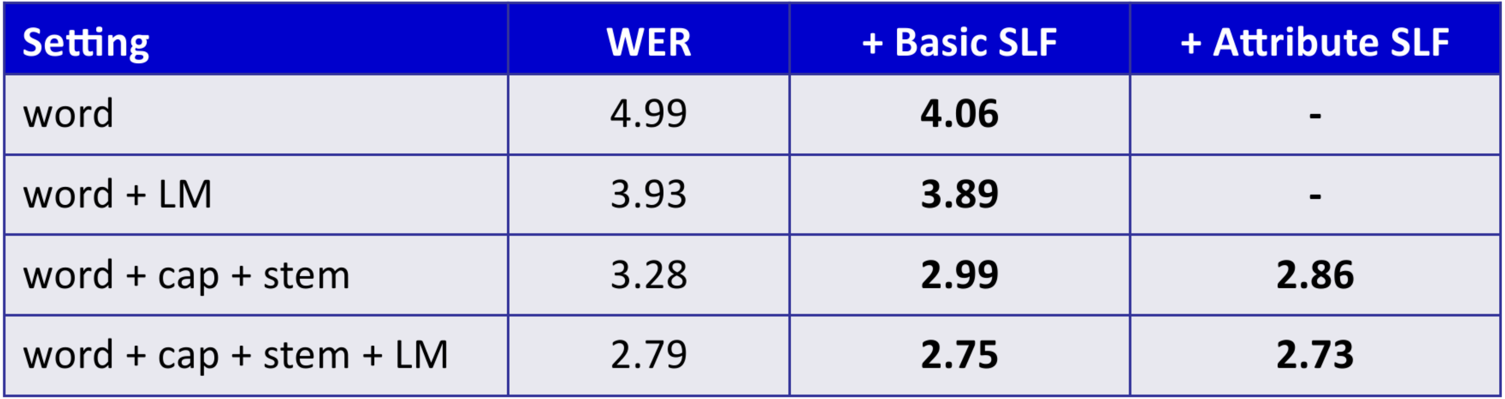
- Word clusters from large unannotated corpus
  - Parser-based hierarchical clustering (Miller et al, ACL 2004)
  - Distributional Similarity methods: (Lee and Pereira, ACL'99), (McCallum et al, SIGIR'98)
  - Mostlly unsupervised, co-occurrence based and no-training (similar to our step(1): co-occurrence)
  - Our methods provide semi-supervised strategy (WCD) and unsupervised model with auxiliary task (LM)

Challenges for bioNLP

- Substantially more difficult:
  - Constantly changing vocabulary, millions of gene names
  - Complex orthographic patterns, variations: expands active vocabulary, complicates building dictionaries
  - Ambiguity. Same name may refer to a range of biological objects and terms.

NLP: Part-of-speech tagging

- 48 classes, 1M words

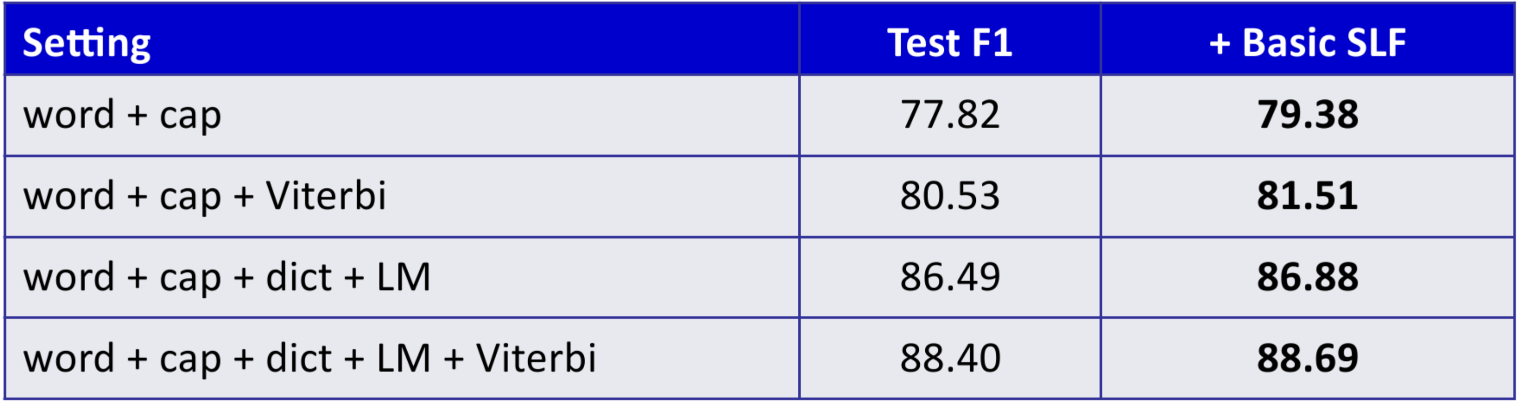


Setting	WER	+ Basic SLF	+ Attribute SLF
word	4.99	**4.06**	-
word + LM	3.93	**3.89**	-
word + cap + stem	3.28	**2.99**	**2.86**
word + cap + stem + LM	2.79	**2.75**	**2.73**

- Best known POS system: 2.76% (Toutanova, 2003), many more complex features

NLP: English NER

- 17 classes, 200K words

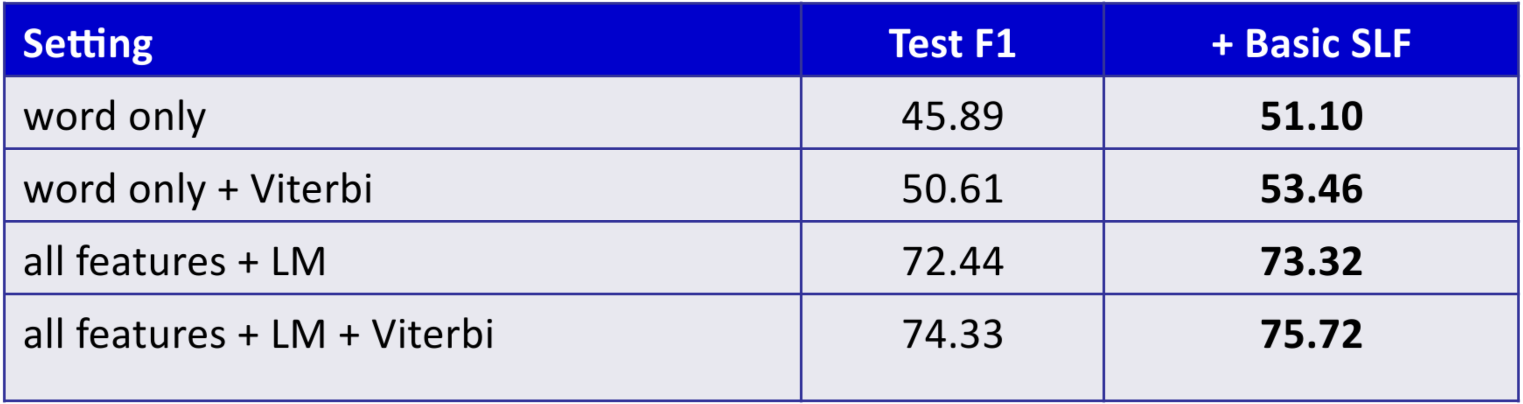


Setting	Test F1	+ Basic SLF
word + cap	77.82	**79.38**
word + cap + Viterbi	80.53	**81.51**
word + cap + dict + LM	86.49	**86.88**
word + cap + dict + LM + Viterbi	88.40	**88.69**

- Best system: 89.31% (extensive use of dictionaries)

NLP: German NER

- 17 classes, 200K words

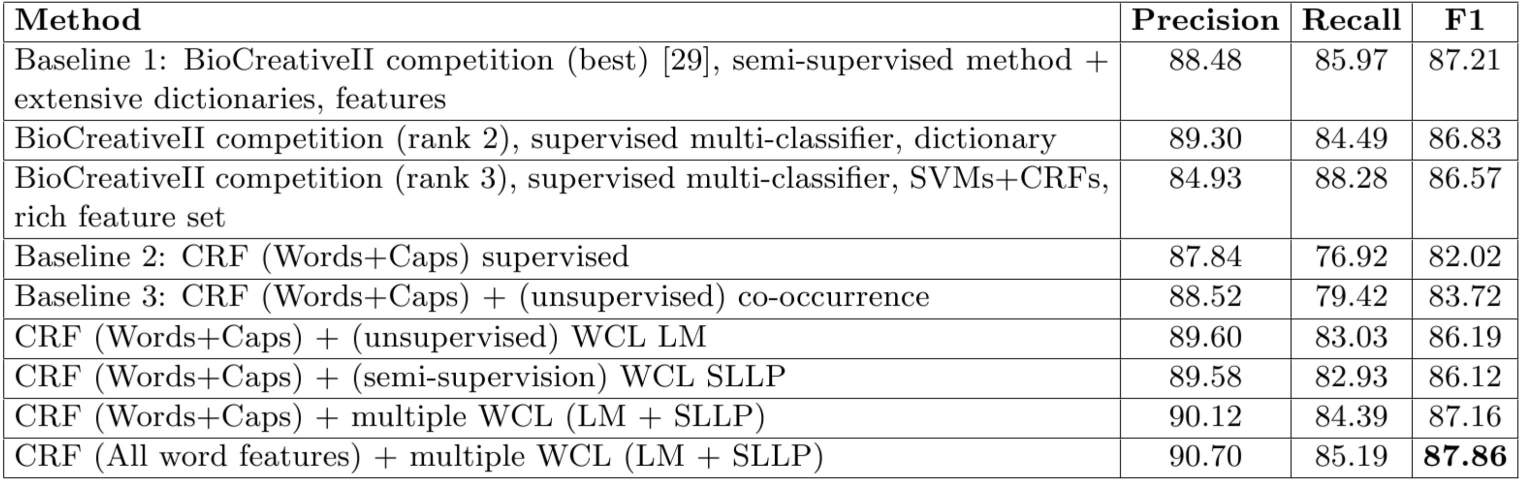


Setting	Test F1	+ Basic SLF
word only	45.89	**51.10**
word only + Viterbi	50.61	**53.46**
all features + LM	72.44	**73.32**
all features + LM + Viterbi	74.33	**75.72**

- Best system: 74.17%

bioNER: Comparison with top systems

- BioCreative II

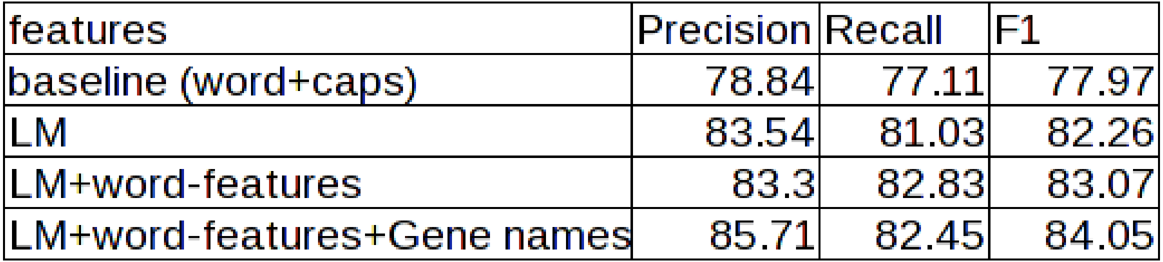


Method	Precision	Recall	F1
Baseline 1: BioCreativeII competition (best) [29], semi-supervised method + extensive dictionaries, features	88.48	85.97	87.21
BioCreativeII competition (rank 2), supervised multi-classifier, dictionary	89.30	84.49	86.83
BioCreativeII competition (rank 3), supervised multi-classifier, SVMs+CRFs, rich feature set	84.93	88.28	86.57
Baseline 2: CRF (Words+Caps) supervised	87.84	76.92	82.02
Baseline 3: CRF (Words+Caps) + (unsupervised) co-occurrence	88.52	79.42	83.72
CRF (Words+Caps) + (unsupervised) WCL LM	89.60	83.03	86.19
CRF (Words+Caps) + (semi-supervision) WCL SLLP	89.58	82.93	86.12
CRF (Words+Caps) + multiple WCL (LM + SLLP)	90.12	84.39	87.16
CRF (All word features) + multiple WCL (LM + SLLP)	90.70	85.19	**87.86**

- Best system: 87.21% (many more features, parsing, dictionaries)

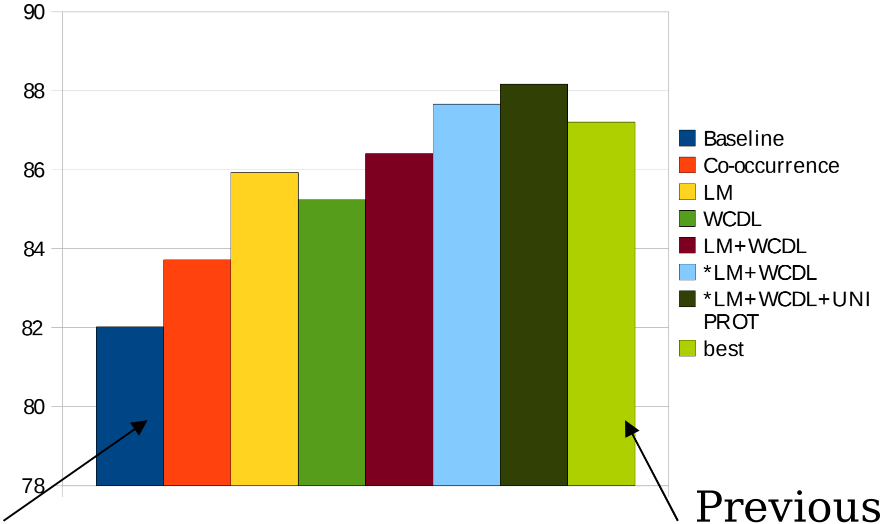
Gene Mention Prediction using NN

- Results



features	Precision	Recall	F1
baseline (word+caps)	78.84	77.11	77.97
LM	83.54	81.03	82.26
LM+word-features	83.3	82.83	83.07
LM+word-features+Gene names	85.71	82.45	84.05

Gene Mention Prediction: system comparison



Baseline

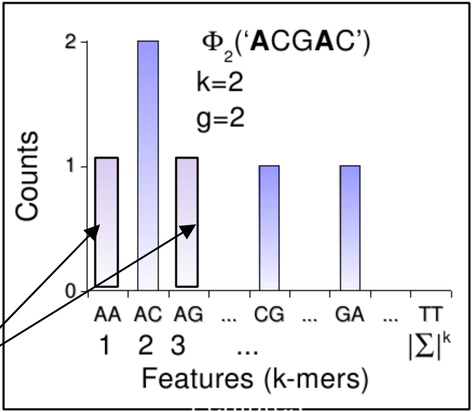
Previous best system

Gapped kernels

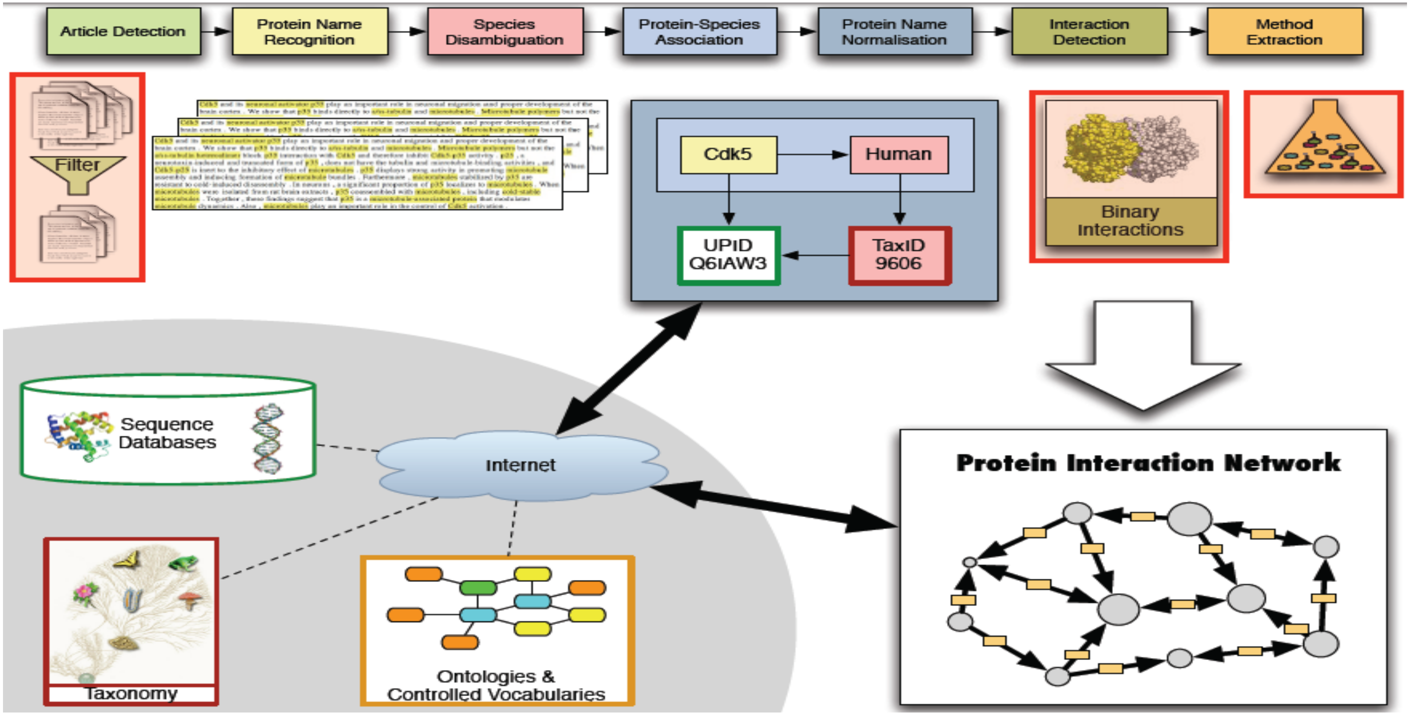
- Count #non-contiguous subsequences of length k and up to g gaps
- Compute kernel between X and Y as a dot-product of two feature vectors:

$$K(X,Y)=\sum_{k-\mathit{word}\ \boldsymbol{a}} C(a|X)\,C(a|Y)$$

$$C(a|X)=\#\,\mathit{subsequences}\ \mathit{matching}\ \boldsymbol{a}\ \mathit{with}\ \mathit{up}\ \text{to}\ \boldsymbol{g}\ \mathit{gaps}$$



Gapped instances



Cdk5 and its neuronal activator p35 play an important role in neuronal migration and proper development of the brain cortex . We show that p35 binds directly to α/β-tubulin and microtubules . Microtubule polymers but not the α/β-tubulin heterodimers block p35 interaction with Cdk5 and therefore inhibit Cdk5-p35 activity . p25 , a neurotoxin-induced and truncated form of p35 , does not have the tubulin and microtubule-binding activities , and Cdk5-p25 is inert to the inhibitory effect of microtubules . p35 displays strong activity in promoting microtubule assembly and inducing formation of microtubule bundles . Furthermore , microtubules stabilized by p35 are resistant to cold-induced disassembly . In neurons , a significant proportion of p35 localizes to microtubules . When microtubules were isolated from rat brain extracts , p35 coassembled with microtubules , including cold-stable microtubules . Together , these findings suggest that p35 is a microtubule-associated protein that modulates microtubule dynamics . Also , microtubules play an important role in the control of Cdk5 activation .