# Semi-Supervised Abstraction-Augmented String Kernel for bio-Relationship Extraction

Pavel P. Kuksa, Rutgers University
Yanjun Qi, Bing Bai, Ronan Collobert, NEC Labs
Jason Weston, Google Research NY
Vladimir Pavlovic, Rutgers University
Ning Xia, University of Minnesota

# ROADMAP

- Bio-Relation Extraction
- Basic String Kernel
- Abstraction-augmented String Kernel (ASK)
- Experimental Results
- Extension to Protein Sequence Classification

# ROADMAP

- Bio-Relation Extraction
- Basic String Kernel
- Abstraction-augmented String Kernel (ASK)
- Experimental Results
- Extension to Protein Sequence Classification

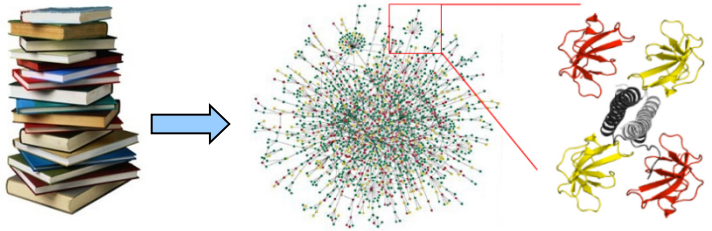
# Relation Extraction for Bio-literature

MEDLINE: 70 million queries monthly, > 17M articles

- Impossible to annotated manually

Linking biomedical text to databases

- Bio-Entity Recognition,
- Bio-Relationship Extraction



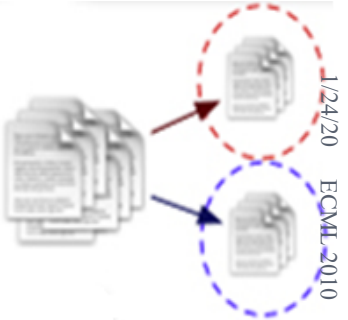
# Goal & Challenges

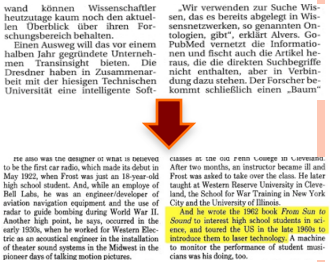
- Challenges:
  - annotated data is scarce
  - millions of unannotated articles (e.g., MEDLINE)
  - learn from unlabeled data with limited prior knowledge

- Bio-Relationships
  - Many important relationships:
  - genetic interaction, disease to phenotype, .....
  - case study here: protein-protein interaction (PPI)

# Three Levels: Bio-Relation Extraction

- (1) Article Level
  - Identify articles about specific relation or not
- (2) Sentence Level
  - Identify sentences about specific relation or not



wand können Wissenschaftler heutzutage kaum noch den aktuellen Überblick über ihren Forschungsbereich behalten.

Einen Ausweg will das vor einem halben Jahr gegründete Unternehmen Transinsight bieten. Die Dresdner haben in Zusammenarbeit mit der hiesigen Technischen Universität eine intelligente Soft-

„Wir verwenden zur Suche Wissen, das es bereits abgelegt in Wissensnetzwerken, so genannten Ontologien, gibt", erklärt Alvers. Go-PubMed vernetzt die Informationen und fischt auch die Artikel heraus, die die direkten Suchbegriffe nicht enthalten, aber in Verbindung dazu stehen. Der Forscher bekommt schließlich einen „Baum"

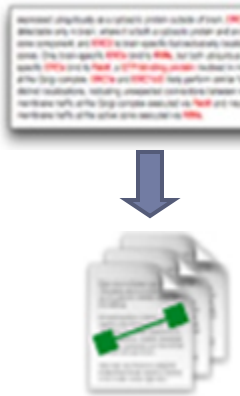
He also was the designer of what is believed to be the first car radio, which made its debut in May 1922, when Frost was just an 18-year-old high school student. And, while an employe of Bell Labs, he was an engineer/developer of aviation navigation equipment and the use of radar to guide bombing during World War II. Another high point, he says, occurred in the early 1930s, when he worked for Western Electric as an acoustical engineer in the installation of theater sound systems in the Midwest in the pioneer days of talking motion pictures.

classes at the old Fenn College in Cleveland. After two months, an instructor became ill and Frost was asked to take over the class. He later taught at Western Reserve University in Cleveland, the School for War Training in New York City and the University of Illinois.

And he wrote the 1962 book *From Sun to Sound* to interest high school students in science, and toured the US in the late 1960s to introduce them to laser technology. A machine to monitor the performance of student musicians was his doing, too.

# Three Levels: Bio-Relation Extraction (Cont.)

- (3) Relation Level (pairwise relation)
  - extract pairs of relating entities (e.g. interacting proteins) from a sentence
  - 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.



# Essentially: Classification of Word Strings

| Task | Data | Labeled Size |
|---|---|---|
| (1) Article Level | BioCreativeII IAS | 5495 train abstracts 677 test abstracts |
| (2) Sentence Level | AIMED Sentence | 1730 sentences |
| (3) Relation Level | AIMED Relation | 4026 relations (built from (2)) |

- annotated data is scarce
- extremely large vocabulary (~2 million words in PubMed)

# ROADMAP

- Bio-Relation Extraction
- String Kernel
- Abstraction-augmented String Kernel (ASK)
- Experimental Results
- Extension to Protein Sequence Classification

# String Kernel

- Use fixed-length feature vectors to represent arbitrary long strings
- 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 m-words
  - Gapped kernel: dot-product of (non-contiguous) k-word counts with g-gaps allowed between words

Kuksa, Huang, and Pavlovic, NIPS2008

# String Kernels


$K(X,Y)=<F(X), F(Y)>$

string kernel

Feature vector F(X)

# e.g. Gapped kernels

- Count #non-contiguous subsequences of length k and up to g gaps

**Example Sentence String: "SM binds RNA in vitro ..."**

| | |
|---|---|
| **Subsequences considered by "Gapped Kernel" With k =3 , g = 1** | **( SM [ ] RNA in ), (binds RNA in [ ] ), (binds [ ] in vitro), ...** |

Kuksa, Huang, and Pavlovic, NIPS2008

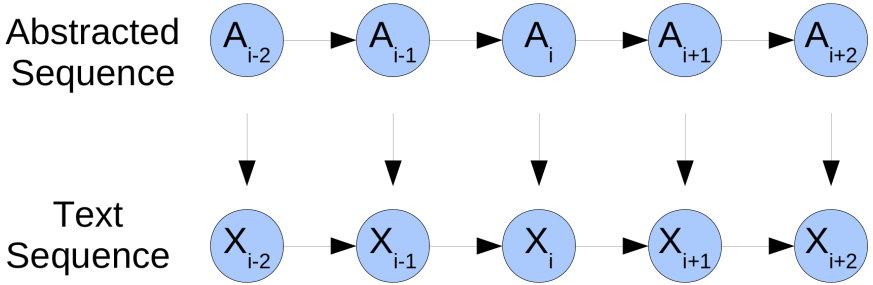
# ROADMAP

- Bio-Relation Extraction
- Basic String Kernel
- Abstraction-augmented String Kernel (ASK)
  - Local ASK
  - Global ASK
- Experimental Results
- Extension to Protein Sequence Classification

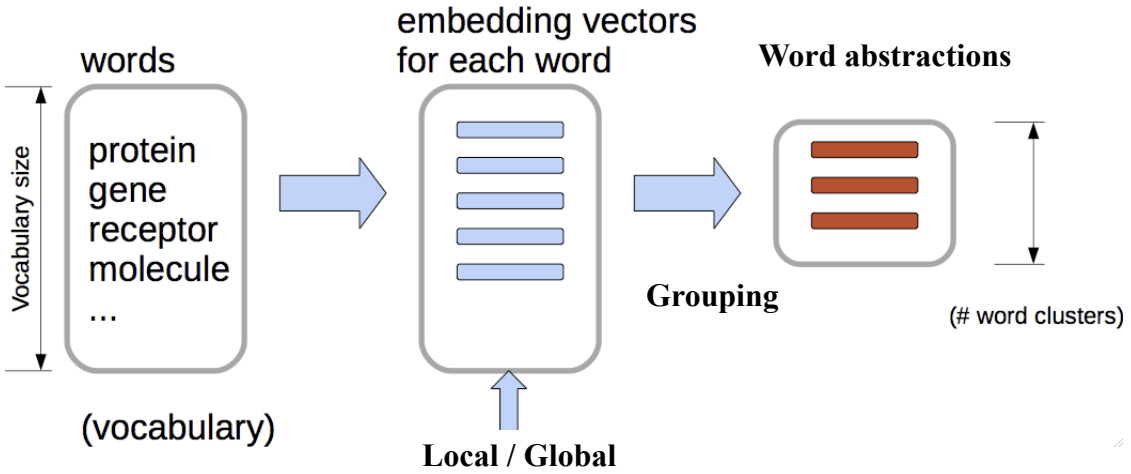
# String Kernels using Unlabeled Data through Word Abstractions

- String kernel over both words and abstracted word representation learned from unlabeled text

$$K(X,Y)=< (F(X),F'(a(X)), (F(Y),F'(a(Y)) >$$



# How to Learn Word Abstractions ?

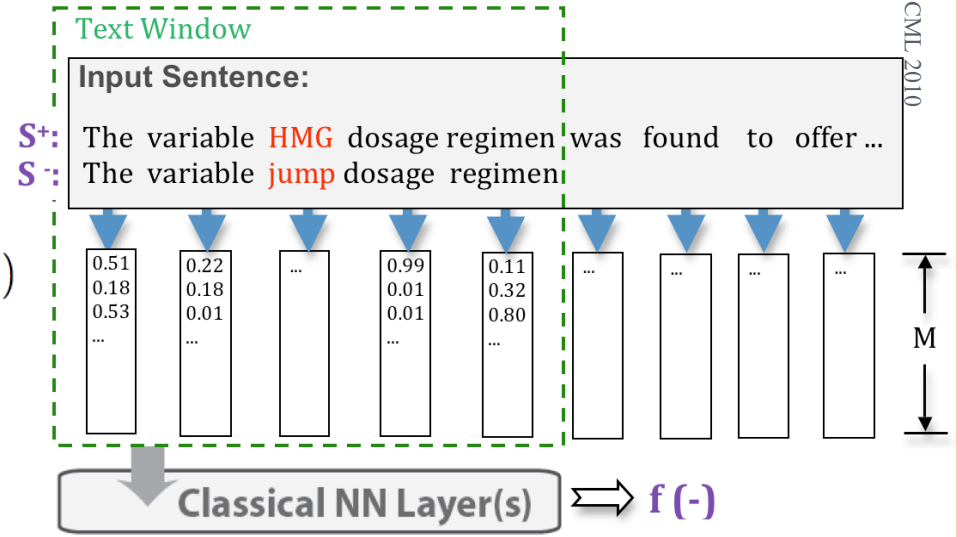



- Step1: train low dimensional embedding for words based on unlabeled data (semantically similar words have close embeddings)
- Step2: group similar words to generate more abstract entities

# (1) Local ASK: Train word embedding from short text window

- Positive examples: Text window extracted from unlabeled corpus (PubMed abstracts 1995-present, 1.3G words)
- Negative examples: Text window with substitution of the middle word by a random word





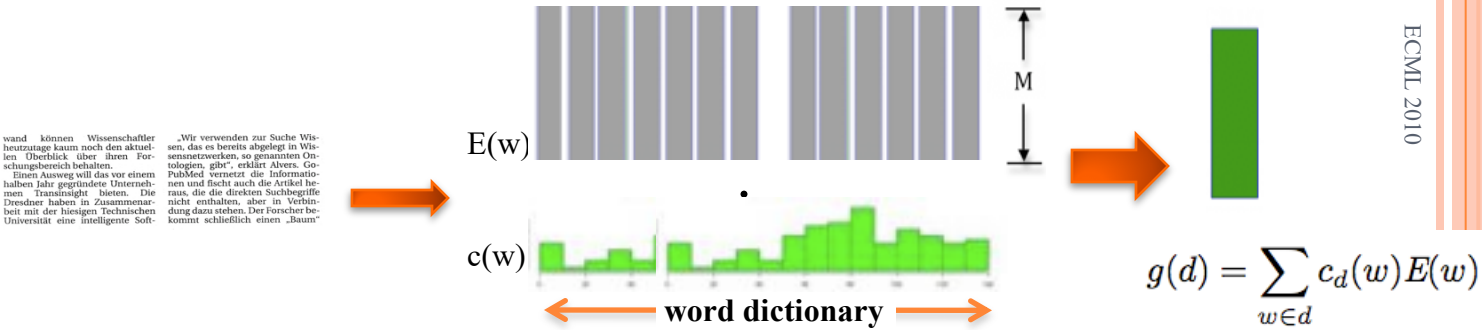

$$\sum \max\left(0, 1 - f(s^+) + f(s^-)\right)$$

Collobert & Weston, ICML2008

# (2) Global ASK: Train word embedding from long text segment

- Consider word distributions globally (not local window structure)
- Model relationships between words in long text segments (e.g. article)






- Force g(-) of two documents with similar meanings to have closer representations, with different meanings to be dissimilar
  - Pseudo supervised signals by splitting each Pubmed abstract into two half segments (~4.5M abstracts)

# Example words with same abstractions as sample query words

| Query | Local ASK | Global ASK |
|---|---|---|
| protein | ligand, subunit, receptor, molecule | proteins, phosphoprotein, isoform, |
| medical | surgical, dental, preventive, reconstructive | hospital, investigated, research, urology |
| interact | cooperate, compete, interfere, react | interacting, interacts, associate, member |
| immunoprecipitation | co-immunoprecipitation, EMSA, autoradiography, RT-PCR | coexpression, two-hybrid, phosphorylated, tbp |

# ROADMAP

- Bio-Relation Extraction
- Basic String Kernel
- Abstraction-augmented String Kernel (ASK)
- Experimental Results
- Extension to Protein Sequence Classification

## Experiment I : Article Level (relevant article detection)

- Binary classification: identify abstracts about
  - protein-protein interaction (not just any relation)

- Data: BioCreative II competition IAS data
  - Train: 3536 negative, 1959 positive abstracts
  - Test: 338 positive, 339 negative

- Evaluation: F1 score, ROC

## Experiment I : Article Level (relevant article detection)

| Method | Precision | Recall | F1 | ROC |
|---|---|---|---|---|
| **Local ASK** | **76.06** | 84.62 | **80.11** | **85.67** |
| **Global ASK** | 73.59 | 84.91 | 78.85 | 84.96 |
| **Mismatch SK** | 69.02 | 83.73 | 75.67 | 81.70 |
| (best) BioCreative II | 70.31 | **87.57** | 78.00 | 81.94 |

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

## Experiment II: Sentence Level (relevant sentence detection)

- Data: AIMed sentences
  - 10-fold cross-validation with 1730 sentences
- Evaluation: F1 score

| Method | Words | Words + Stems |
|---|---|---|
| **Local ASK** | 69.46 | **70.49** |
| **Global ASK** | 67.83 | 67.99 |
| **Spectrum SK** | 61.49 | 65.94 |

## Experiment III : Relation Level (relation extraction from sentences)

- Extract all pairwise interacting pairs from a given sentence
  - assuming protein entities have been labeled already

- Data: AIMed relation dataset
  - 951 positive examples, 3075 negative
  - 10-fold cross-validation
  - Relation data generation:
  - For each regular sentence, with n entities, create C(n,2) copies of the sentence
  - Each copy (example) having only 2 entities replaced with PROT1 and PROT2, all the rest as PROT

## Experiment III : Relation Level (relation extraction from sentences)

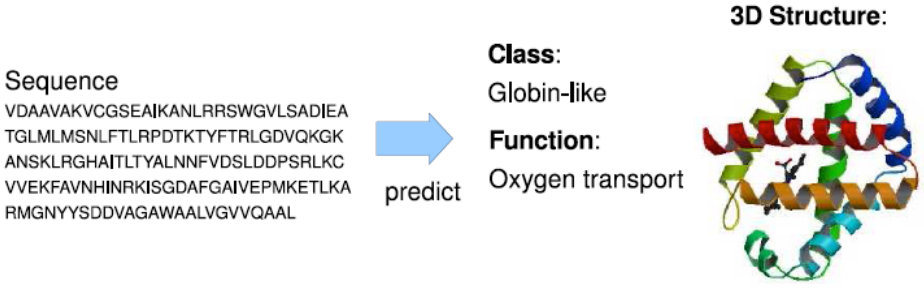
| Method | Precision | Recall | F1 |
|---|---|---|---|
| **Local ASK** | **61.18** | 67.92 | 64.33 |
| **Global ASK** | 60.68 | **69.08** | **64.54** |
| Baseline (EMNLP07) | 59.59 | 60.68 | 59.96 |

- Baseline system: transductive SVM on features from dependency parsing

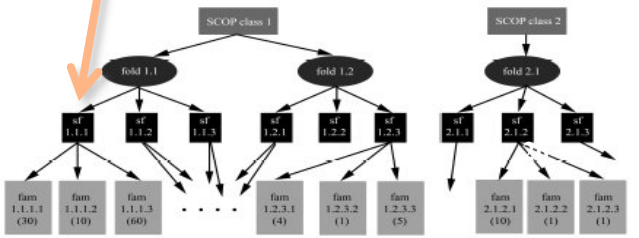
# ROADMAP

- Bio-Relation Extraction
- Basic String Kernel
- Abstraction-augmented String Kernel (ASK)
- Experimental Results
- Extension to Protein Sequence Classification

# Experiment IV: Structural Classification from Protein Sequences

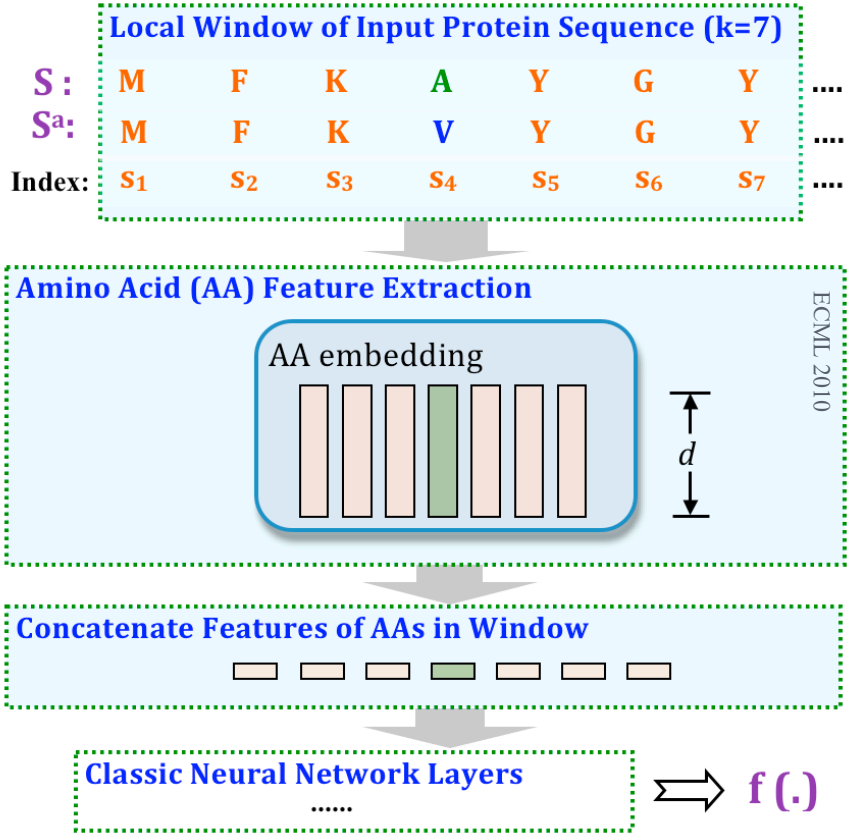



- **Goal**: predict structural/functional class from primary sequence
- Limited labeled data : highly diverse sequences
- Millions of unlabeled sequences



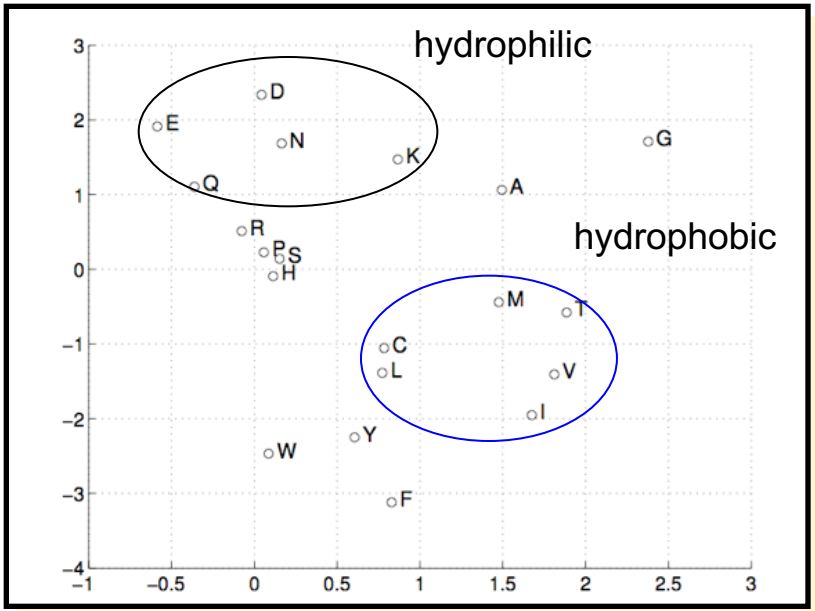
# Experiment IV: Local ASK on Protein Sequence Classification

- Treat each amino acid (AA) as word
- Train embedding representation for each AA on UniprotKB protein sequence dataset



# Experiment IV: Local ASK on Protein Sequence Classification (cont.)

- Learned embedding of amino acid (projected 2D with PCA)




- substitution groups
- structural properties (hydropathy)

# Experiment IV: Local ASK on Protein Sequence Classification (cont.)

- mean ROC50 score

| Method | Baseline | +Local-ASK |
|---|---|---|
| Spectrum | 27.91 | **33.06** |
| Mismatch | 41.92 | **46.68** |
| Spatial sample | 50.12 | **52.75** |
| Semi-supervised Cluster kernel | 67.91 | **70.14** |

- Local ASK improves over both supervised and semi-supervised baselines

# Related Works

- Semi-supervised string kernel
  - Word sequence kernel (soft word match, slow )
  - Cluster kernel just feasible for protein sequences
- Word abstraction based model
  - Mostly unsupervised, co-occurrence based and no-training
  - e.g. Distributional Similarity (Lee and Pereira, ACL'99)
- PPI article retrieval: SVM on features
  - bag-of-words + bag-of-NLPs (chunk; phrase; pos; protein mention; non-proteins; title phrase, et al.)
- PPI relation detection
  - Rule/pattern based methods
  - Graph kernels built on complex parsing trees

# Summary

- Wrapper approach: could apply on both supervised and semi-supervised string kernels
- Efficient: linear cost in the input length
- Provided two models for word-feature learning from unsupervised data: local vs. global
- A unified framework for bRE: at multiple levels where few training examples exist
- Not restricted to biomedical text: workable on general sequence classification tasks
- Can incorporate other types of word similarities : e.g. LSI

# THANKS !