Text Classification with Heterogeneous Information Network Kernels

AAAI'16 Phoenix, Arizona USA

Chenguang Wang, Yangqiu Song, Haoran Li, Ming Zhang, Jiawei Han

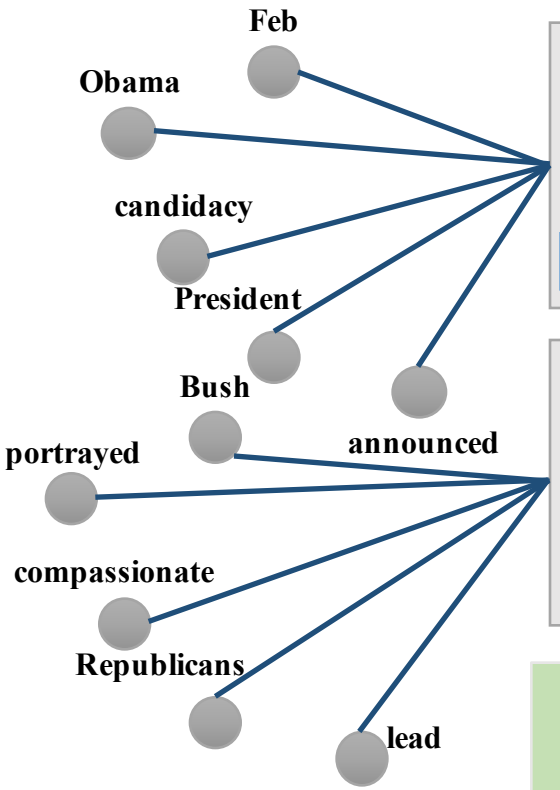
Outline

Motivation	The problem of current classification measures: representation and modelling.
HIN-Kernels	Text represented as network used in kernel for classification.
Experiments	Achieve state-of-art classification results on two benchmark datasets.

Motivation

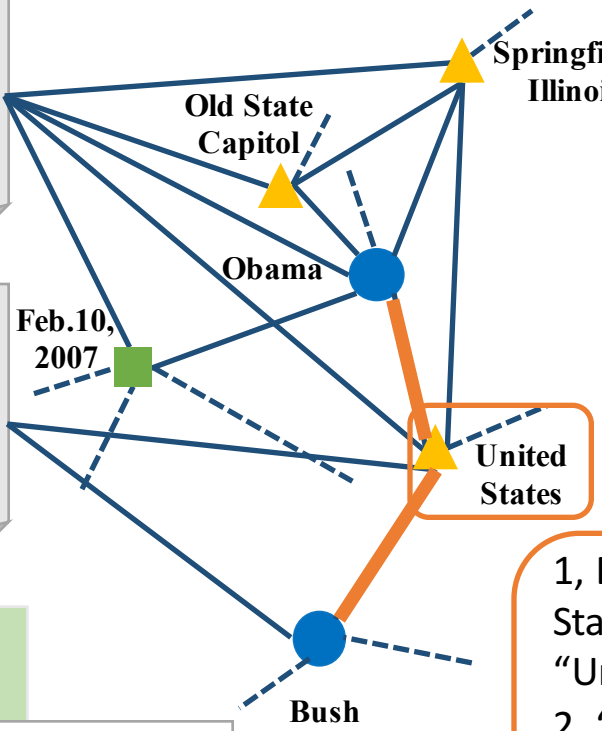
Texts:

KBs:



On Feb.10, 2007 , Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.



Word
Document
Location
Date
Politician

Are the two documents belong to the same class?

"Politics"

1, Entity "United States" not words "United", "States;"
2, "United States" built a **link/meta-path** between "Obama" and "Bush."

Text Classification → Network Classification

Text-Based Heterogeneous Information Network Construction

- Grounding texts to world knowledge framework [Wang et al. KDD'15]

World Knowledge Specification

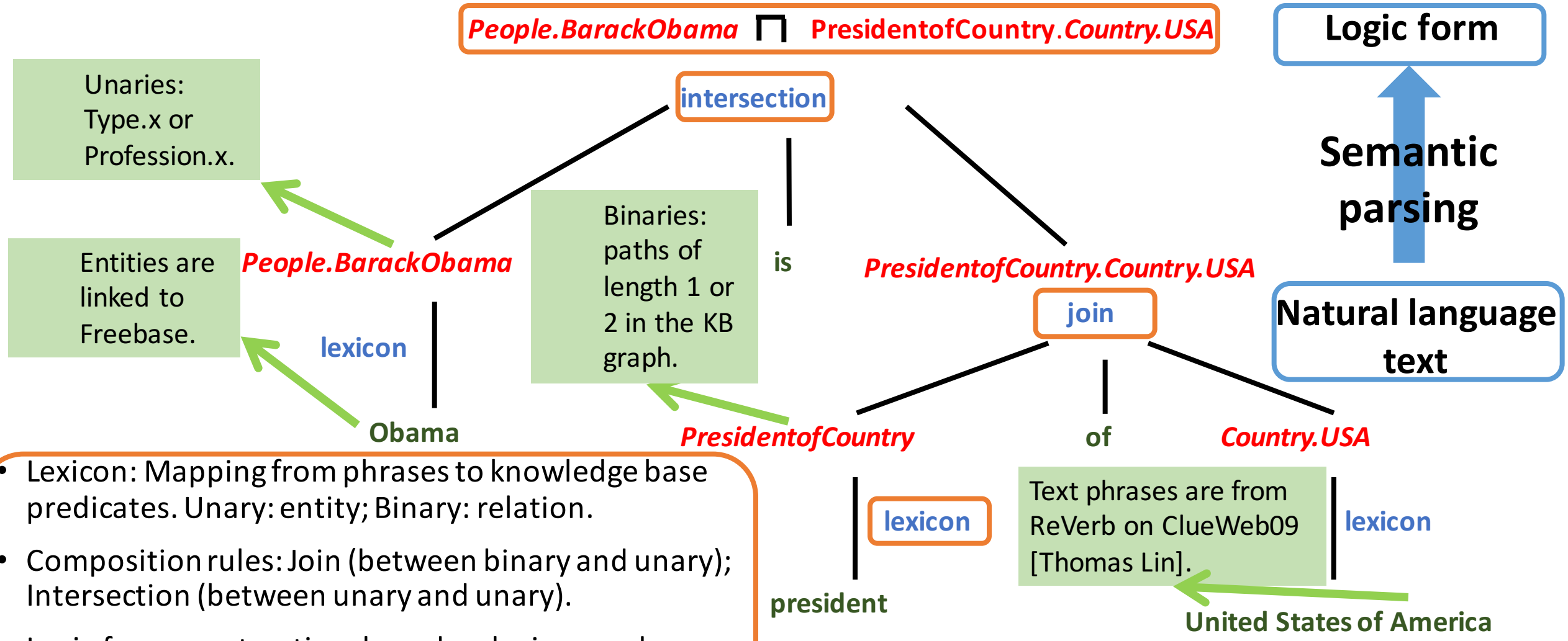
General purpose problem **vs.** Domain specific problem

Specified World Knowledge Representation

Knowledge representation **vs.** Data representation

C. Wang et al. Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD'15

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

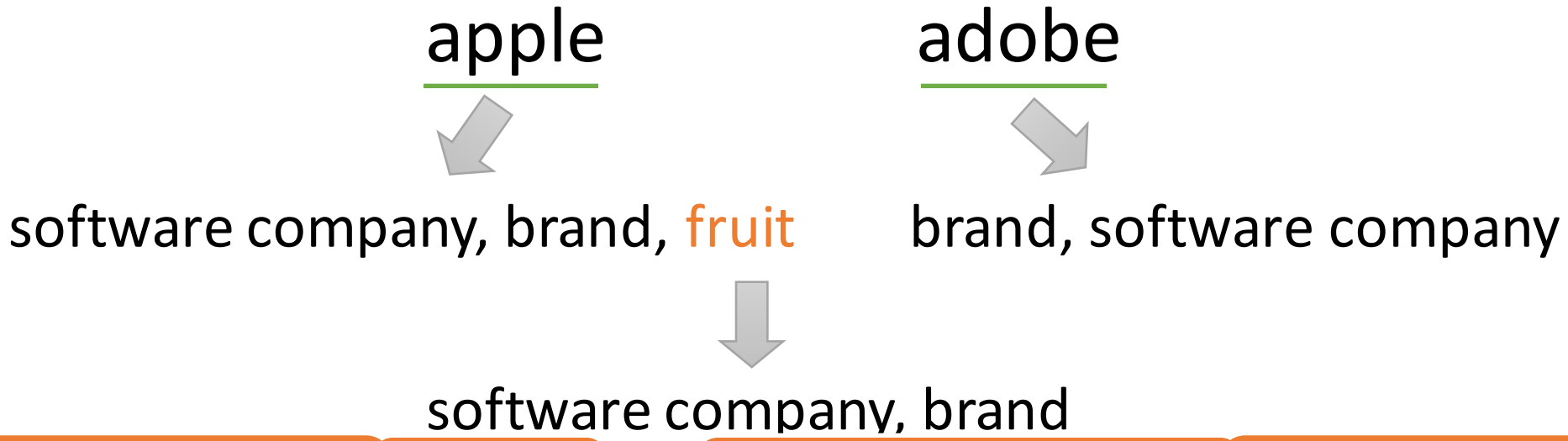


- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.
- Composition rules: Join (between binary and unary); Intersection (between unary and unary).
- Logic form construction: based on lexicon and composition rules recursively.

World Knowledge Specification: Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.



largest probability ones are selected

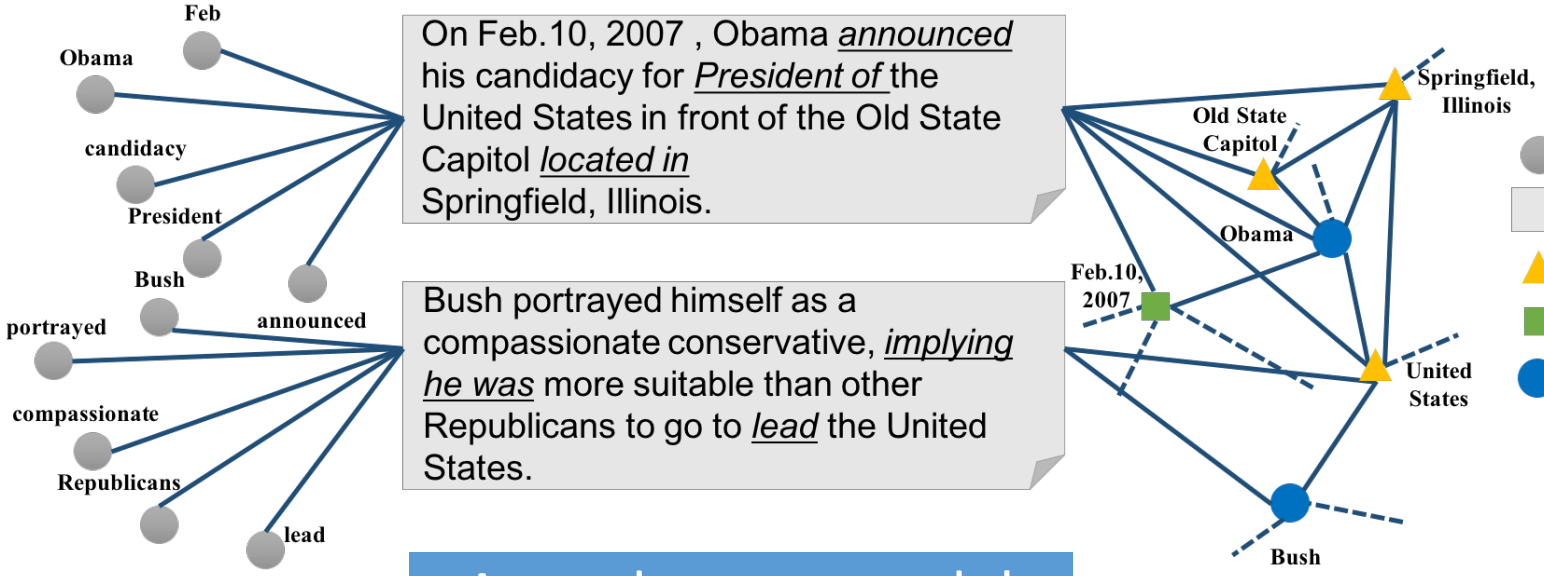
$P(\text{type} \mid \text{related entities})$

A cluster of entities of type features

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI'11.

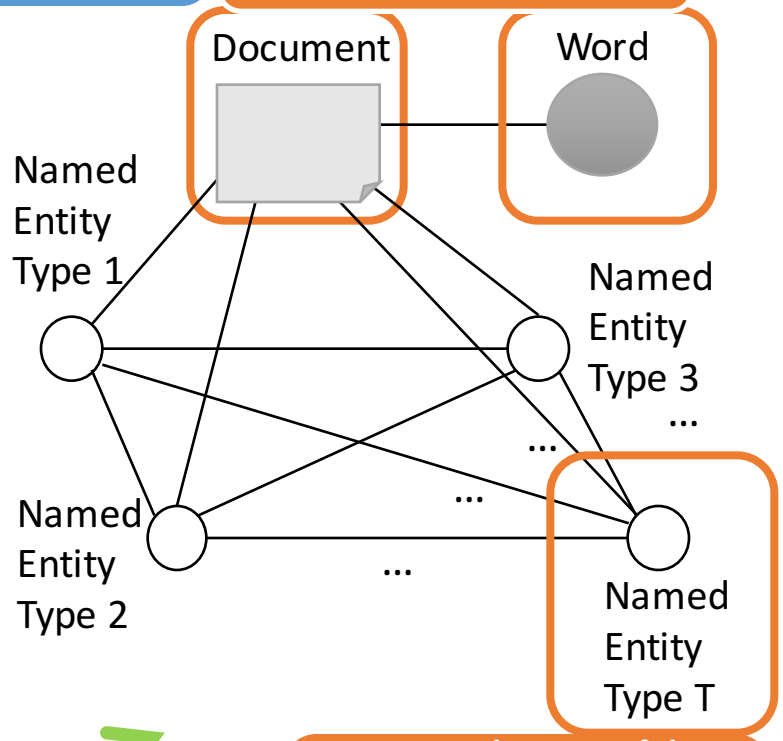
Specified World Knowledge Representation: Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.



A good way to model real world data!

Two entity types in document-based HIN.

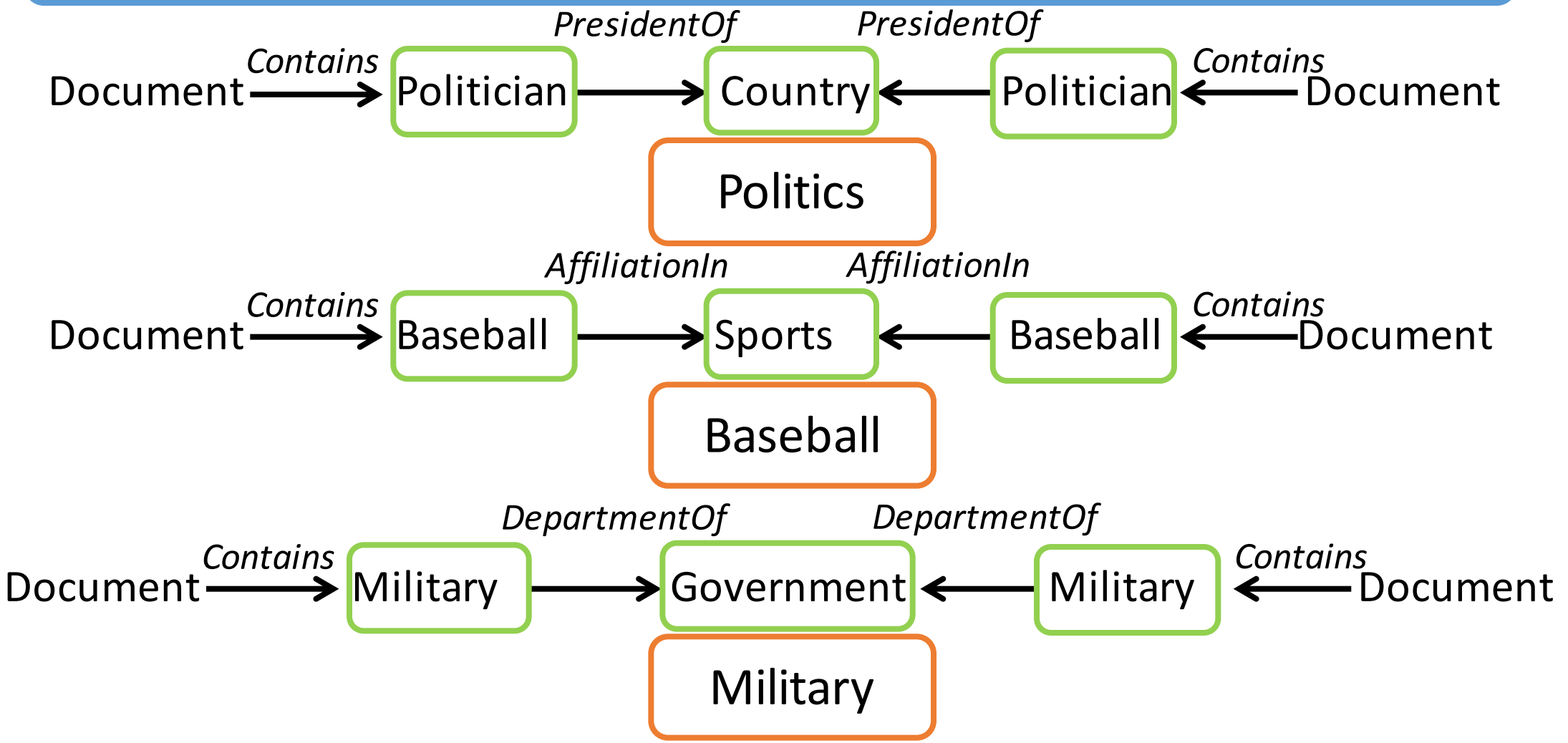


Represent the type of the name in text, e.g, person name.
NOT entity type (node type in HIN).

Network schema: High-level description of a network.

Meta-Path

Meta-path: A **path/link** in the network schema. [Sun et al., 2011]



Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

HIN-Based Kernels for Text Classification

- Link based linear kernel

  Intuition: simply incorporate the links shown via meta-paths into the features.

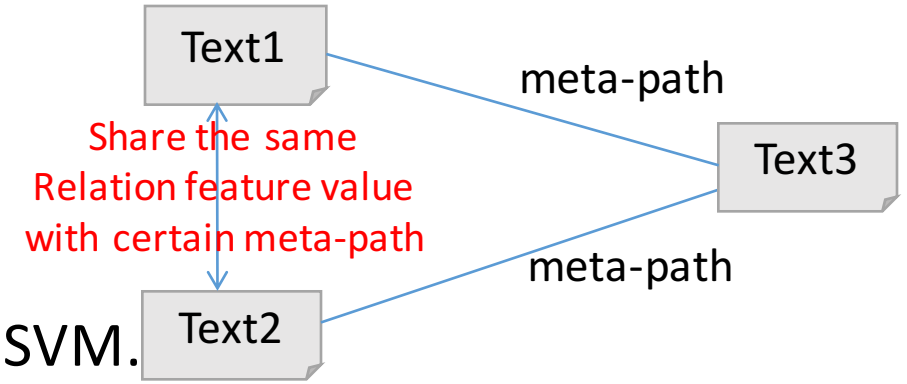
  - Entity features: use bag-of-words with the term frequency weighting mechanism.
  - Relation features: for each meta-path, use number of meta-path instances of the meta-path to documents as feature. [Lu and Getoor 2003']

- Linear kernel framework
  - $\hat{y} = arg\max_{y} f(\boldsymbol{\theta}^T \boldsymbol{x})$

Linear model: linear models are often used for text classification.

Entity features+Relation features

  - Examples: Link based Naïve Bayes and Link based SVM.



Two Simple Link Based Kernels

- Link based Naïve Bayes
  - Intuition: incorporate the links into Naive Bayes model.

$$P(y|\boldsymbol{x}^v) = \frac{P(y)\prod P(\boldsymbol{x}^v|y)}{\sum P(y)\prod P(\boldsymbol{x}^v|y)}$$

Entity features

Combined estimation function

$$P(y|\boldsymbol{x}^e) = \frac{P(y)\prod P(\boldsymbol{x}^e|y)}{\sum P(y)\prod P(\boldsymbol{x}^e|y)}$$

Relation features

$$\hat{y} = arg\max_y p(y)\, P(\boldsymbol{x}^v|y)\, P(\boldsymbol{x}^e|y)$$

- Link based SVM
  - Intuition: provides a simple way to combine the structured information with traditional features.
  - Dual formulation of 1-norm soft margin SVM:

$$\max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Y}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{Y}\boldsymbol{\alpha} \quad \text{s.t. } \boldsymbol{y}^T\boldsymbol{\alpha} = 0,\ 0 \le \boldsymbol{\alpha} \le C\mathbf{1}$$

Entity features+Relation features

  - Optimization: use convex quadratic programming to solve the dual problem.

HIN-Based Kernels for Text Classification

- Indefinite HIN-kernel SVM

  Intuition: disadvantage of the flat features:
  - lost the meaning of meta-paths
  - lost the importance of meta-paths

  Document -> Baseball -> Sports -> Baseball -> Document

  Document -> religion -> Government -> religion -> Document

  For documents talking about sports, the former one is more useful.

Note: we should take the meta-paths as a whole into consideration instead of just treating them as links to the documents.

KnowSim

KnowSim [Wang et al., ICDM'15]:An unstructured data similarity measure defined on structured HIN.

Semantic overlap: the number of meta-paths between two documents.

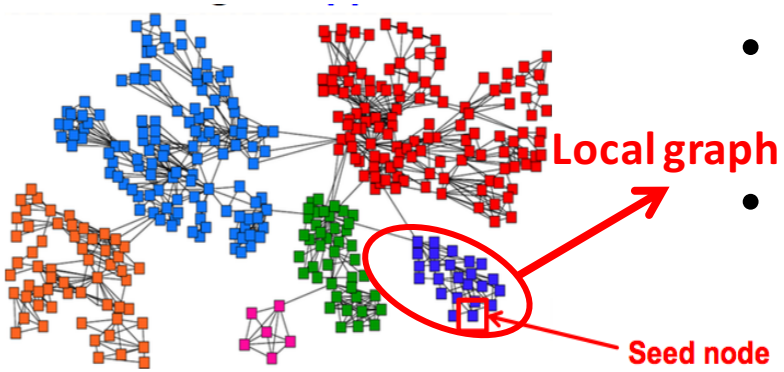
$$KS(d_i, d_j) = \frac{2 \times \sum_m^{M'} w_m \left| \{ p_{i \to j} \in P_m \} \right|}{\sum_m^{M'} w_m \left| \{ p_{i \to i} \in P_m \} \right| + \sum_m^{M'} w_m \left| \{ p_{j \to j} \in P_m \} \right|}$$

Semantic broadness: the number of total meta-paths between themselves.

- Intuition: The larger number of highly weighted meta-paths between two documents, the more similar these documents are, which is further normalized by the semantic broadness.
- KnowSim is computed in nearly linear time.

KnowSim

- To accelerate the meta-path generation
  - Meta-path dependent PageRank-Nibble algorithm [Wang et al., ICDM'15]
  - Intuition: Discovering compact sub-graph based on seed document nodes.



- Compute Personalized PageRank around seed nodes.
- The random walk will get trapped inside the blue sub-graph.

- Algorithm outline
  - Run **PPR** (approximate connectivity to seed nodes) with teleport set = {**S**}
  - Sort the nodes by the decreasing **PPR** score
  - **Sweep** over the nodes and find compact **sub-graph**.
  - Use the sub-graph instead of the whole graph to compute # of meta-paths between nodes.

- To weigh different meta-paths
  - Laplacian score [He, Cai, and Niyogi 2006']
  - Intuition: Laplacian score represents the power of a meta-path in discriminating documents from different clusters.

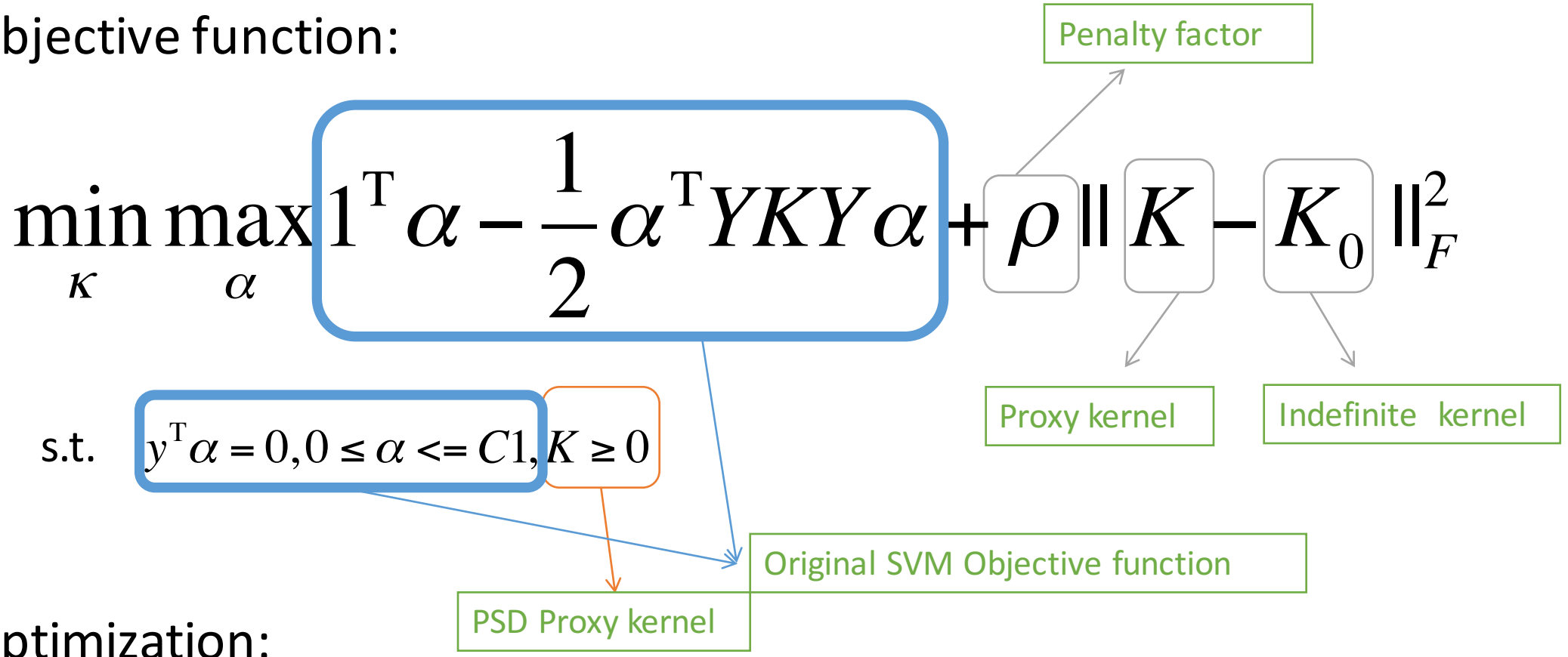
$$L_j = \frac{\widetilde{\boldsymbol{D}}_{.,j}^{T} \boldsymbol{L} \boldsymbol{D}_{.,j}}{\widetilde{\boldsymbol{D}}_{.,j}^{T} \boldsymbol{\Lambda} \boldsymbol{D}_{.,j}}$$

SVM with Indefinite HIN-Kernel

- SVM needs a positive semi-definite(PSD) kernel matrix
- KnowSim matrix K, where $K_{ij} = KS(d_i, d_j)$ is non-PSD.
- Feed the non-PSD KnowSim kernel matrix to SVM [Luss and d'Aspremont 2008']
  - Learn a proxy of non-PSD KnowSim matrix
  - Simultaneously learn a SVM classifier.

SVM with Indefinite HIN-Kernel

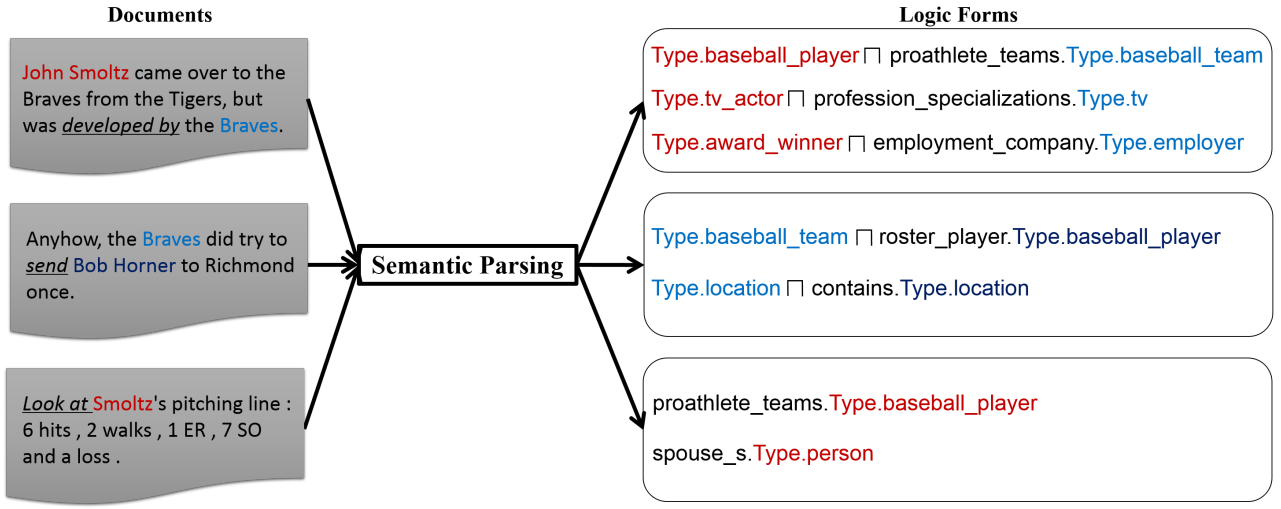
Objective function:



Optimization:

First-order smooth optimization scheme introduced in (Nesterov 2005).
Reason: This scheme has the optimal convergence rate.

Experiments



- Four sub-datasets are constructed

20NewsGroup

RCV1-GCAT

Document datasets					
Sub-datasets	#(Document)	#(word)	#(Entity)	#(Total)	#(Types)
20NG-SIM	3000	22686	5549	31235	1514
20NG-DIF	3000	25910	6344	35254	1601
GCAG-SIM	3596	22577	8118	34227	1678
GCAT-DIF	2700	33345	12707	48752	1523
Each sub-datasets consists of three similar or distinct topics.					

More entities in GCAT

Classification Results

Average accuracy			
Model	SVM		SVM^{HIN}
Settings	BOW	BOW +ENTITY	
20NG-SIM	90.81%	91.11%	91.60%
20NG-DIF	96.66%	96.90%	97.20%
GCAG-SIM	94.15%	94.29	94.82%
GCAT-DIF	88.98%	90.18%	91.19%

The link information actually improve the performance of the SVM.
Same conclusion for Naïve Bayes result.

Classification Results

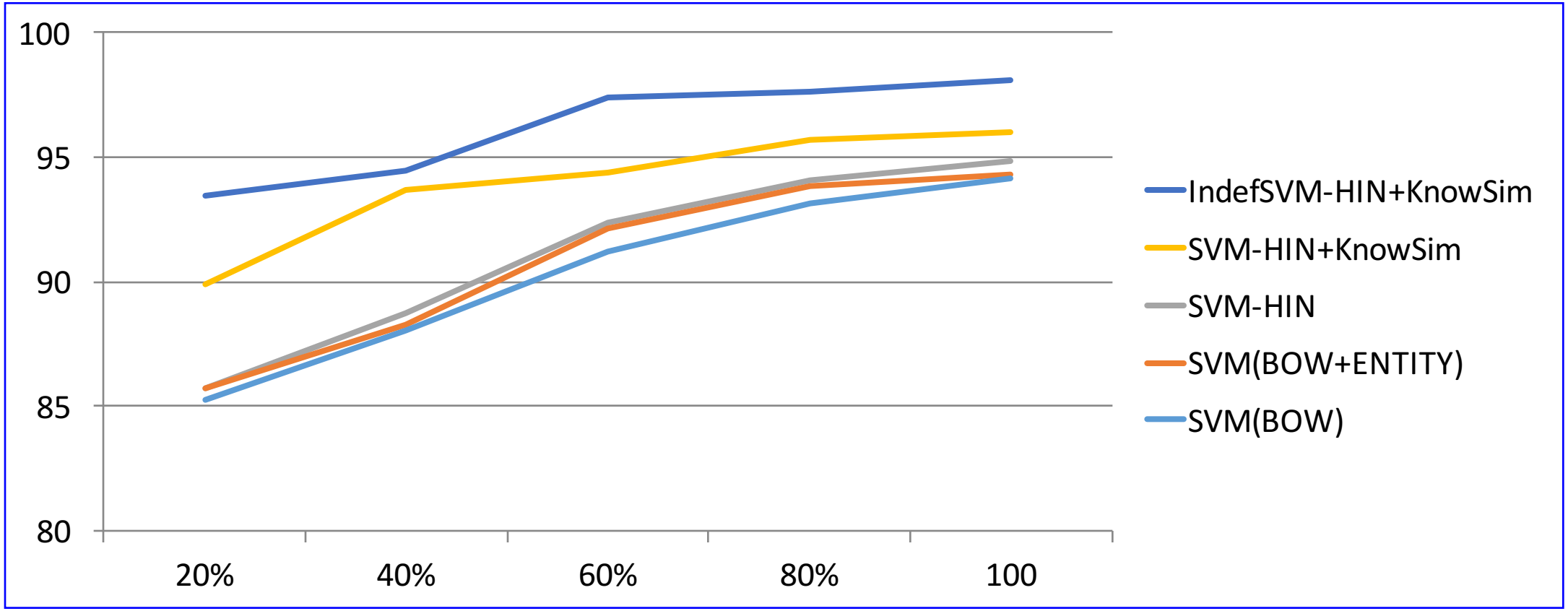
Average accuracy						
Model	SVM^{HIN}	SVM^{HIN}+KnowSim		IndefSVM^{HIN}+KnowSim		**SVM**
Settings		DWD	DWD+ MetaPath	DWD	DWD+ MetaPath	**Word Embedding**
20NG-SIM	91.60%	92.32%	92.68%	92.65%	**93.38%**	91.67%
20NG-DIF	97.20%	97.83%	98.01%	98.13%	**98.45%**	98.27%
GCAG-SIM	94.82%	95.29%	96.04%	95.63%	**98.10%**	96.81%
GCAT-DIF	91.19%	90.70%	91.88%	91.63%	**93.51%**	90.64%

Finding #1: Both kernel methods with DWD+MP outperform SVM^{HIN} .
consider the meta-path information as a whole is better than flat feature.

Finding #2: IndefSVM^{HIN} +KnowSim always works better than SVM^{HIN}+KnowSim
Non-PSD kernel is not suitable for SVM.

Finding #3: IndefSVM^{HIN} +KnowSim always outperforms SVM+WE
Knowledge carries more semantic about the similarities between documents.

Effects of the Size of Training Data



With less training data, the external knowledge can help more on improving the classification accuracy.

Conclusion

Problem	Converting text classification to an HIN classification problem.
Approach	A general classification framework by incorporating typed link information using different types of HIN-kernels.
Results	Indefinite HIN-Kernel SVM performs the best.

Thank You! ☺