TECHNISCHE UNIVERSITÄT DRESDEN

BIOTEChnologisches Zentrum

biotec
Biotechnology Center TU Dresden

Relation Discovery between Indirectly Connected Biomedical Concepts

Dirk Weißenborn, Michael Schroeder, George Tsatsaronis

DILS Conference, Lisbon, 2014

Raynaud's syndrome & fish oil

*"Beneficial effect of **fish oil** on **blood viscosity** in peripheral vascular disease"*
[Woodcock et al., 1984]

Blood viscosity

*"…blood was studied in 20 patients with **Raynaud's syndrome**… studies demonstrate increased **blood viscosity** …"*
[Tietjen et al., 1975]

Fish oil

Blood viscosity

Raynaud's disease

Blood viscosity

Hypothesis:
Fish oil **treats** Raynaud's syndrome
[Swanson, 1986]

Confirmation:
clinical study [DiGiacomo *et al*., 1989]

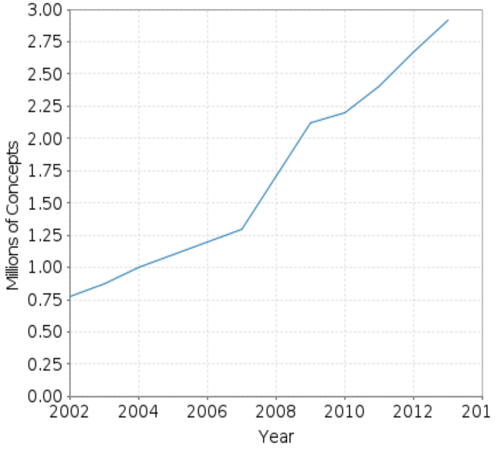
Motivation

- fast growth of knowledge sources → *data mining*

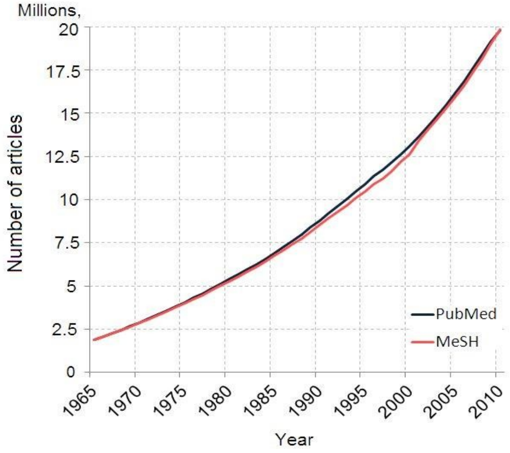
- combination of facts can lead to new knowledge (e.g., Swanson)

- focus of similar work mainly on word statistics neglecting linguistic information

Growth of UMLS



Growth of MEDLINE



Structured vs. Unstructured Knowledge

DrugBank
Open Data Drug & Drug Target Database

Structured knowledge:

- from databases like the *UMLS, DrugBank*
- **fixed** set of **concepts** and **relations**

Unstructured knowledge:

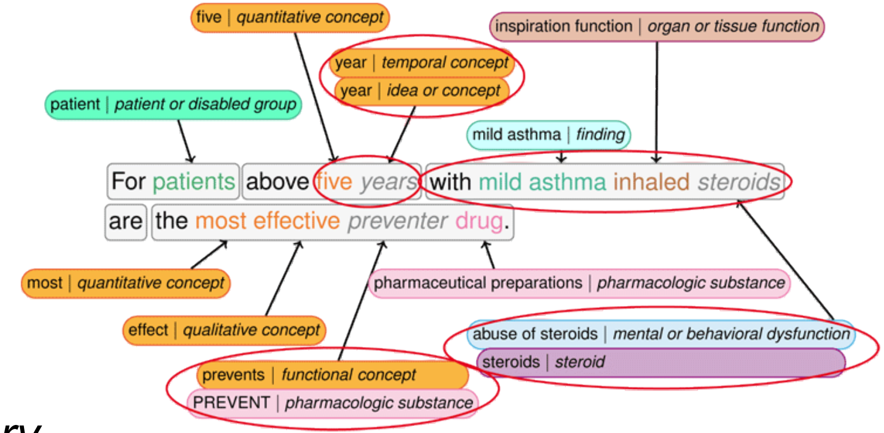
- *MEDLINE* abstracts, annotated by *MetaMap*
- natural language text, expression of **concepts** and their **relations** in **ambiguous** and **synonymous** ways

PubMed

Bridging the Gap

Concepts

- concept annotation of natural language text with *MetaMap*



five | *quantitative concept*

inspiration function | *organ or tissue function*

year | *temporal concept*

year | *idea or concept*

patient | *patient or disabled group*

mild asthma | *finding*

For patients above five years with mild asthma inhaled steroids are the most effective preventer drug.

most | *quantitative concept*

pharmaceutical preparations | *pharmacologic substance*

effect | *qualitative concept*

abuse of steroids | *mental or behavioral dysfunction*

steroids | *steroid*

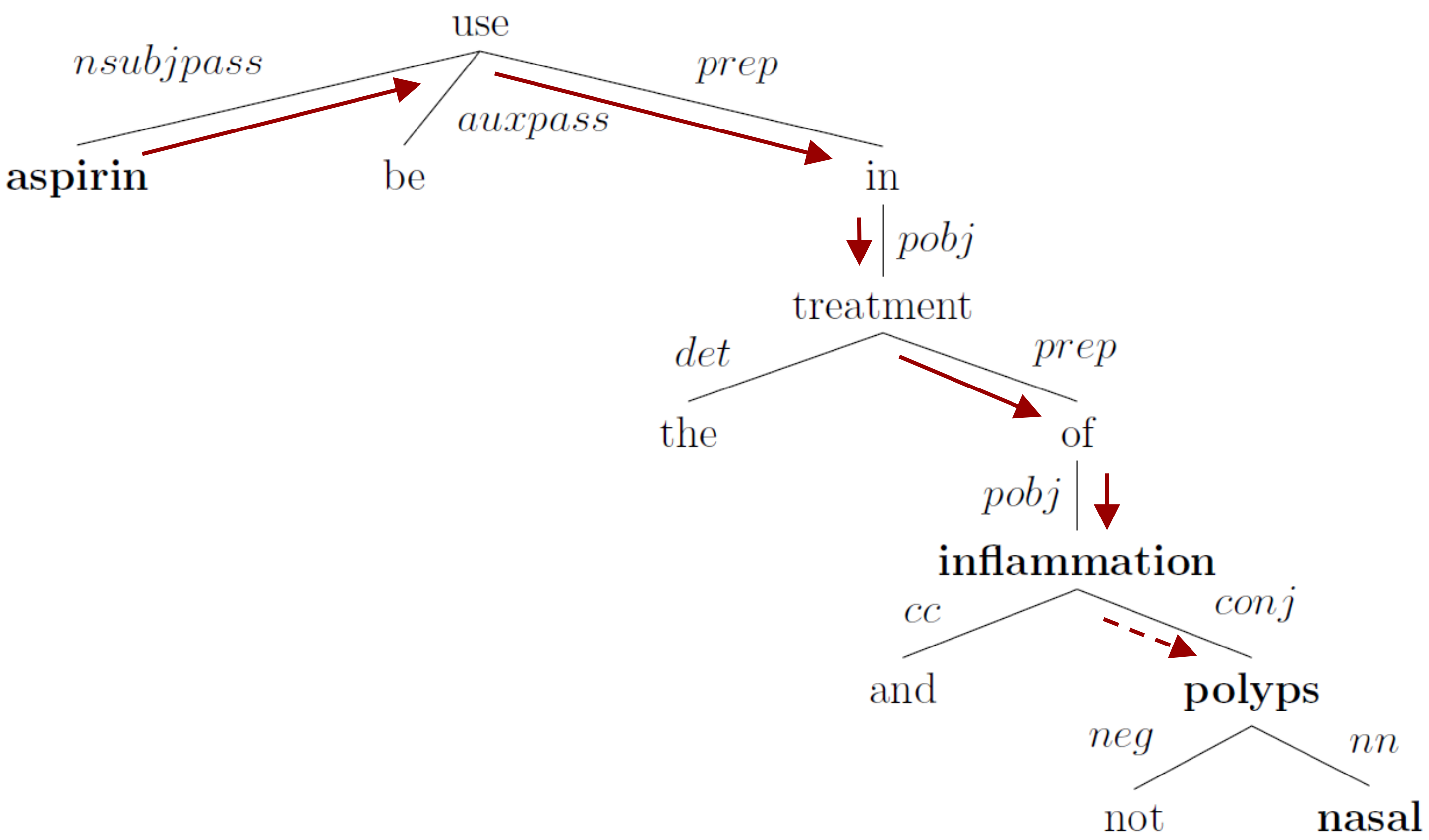
prevents | *functional concept*

PREVENT | *pharmacologic substance*

Relations

- problem: reliable relation annotation *not possible* or *very restricted*
- suggested solution: use plain textual relation between annotated concepts → *dependency paths*

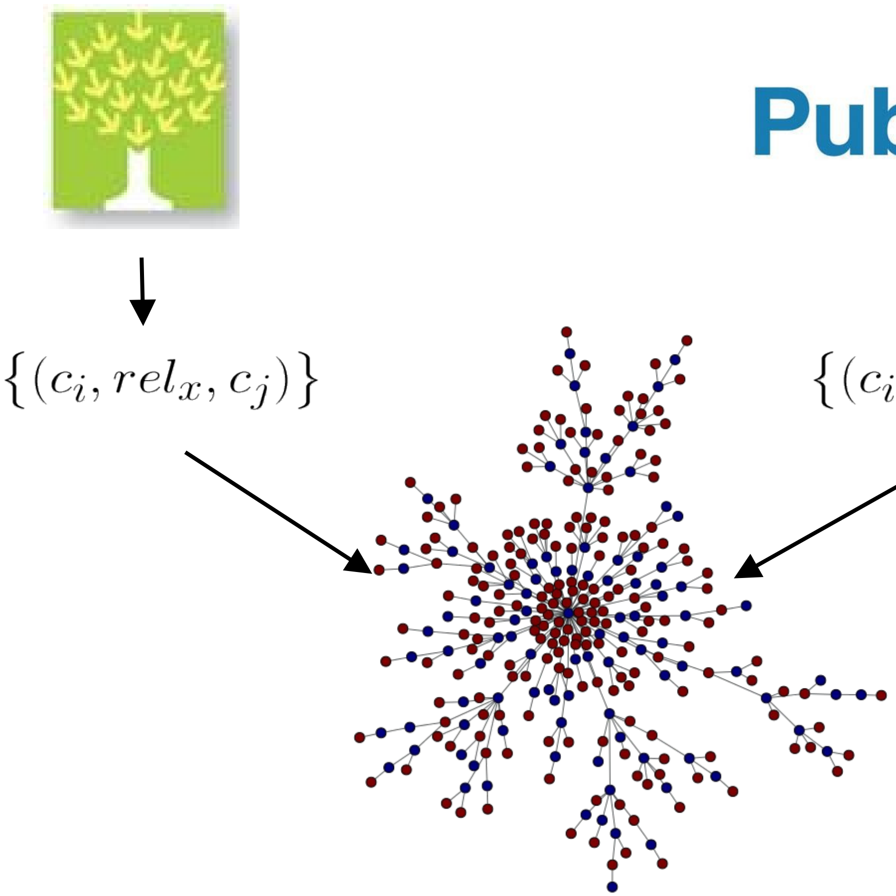
Textual Relations



$$(\mathbf{aspirin}, \overset{nsubjpass}{\rightarrow} use \overset{prep}{\leftarrow} in \overset{pobj}{\leftarrow} treatment \overset{prep}{\leftarrow} of \overset{pobj}{\leftarrow}, \mathbf{inflammation})$$

$$(\mathbf{aspirin}, neg\_ \overset{nsubjpass}{\rightarrow} use \overset{prep}{\leftarrow} in \overset{pobj}{\leftarrow} treatment \overset{prep}{\leftarrow} of \overset{pobj}{\leftarrow}, \mathbf{nasal\ polyps})$$

Combining knowledge



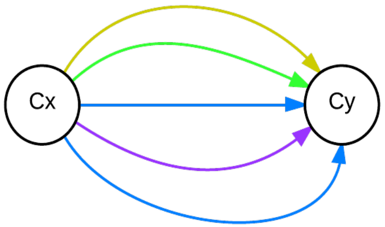
$\{(c_i, rel_x, c_j)\}$

$\{(c_i, rel_x, c_j)\}$

dependency paths
are edges connecting two concepts

paths
are paths in the graph connecting two concepts via intermediate concepts

Representing Relations



- simple model: ***one-in-N*** encoding
  - feature vector of a relation is a vector with only one 1 in the respective dimension
  - feature space is as large as there are relations
- some relations are semantically similar or even synonymous to each other
- simple model assumes all relations to be semantically dissimilar to each other
- need to encode relations semantically
- new model: ***semantic*** encoding
  - apply **LDA** to extract semantic vectors of much lower dimensionality for relations, ensuring semantically similar relations to have similar vectors
  - directly applicable by denoting a pair of concepts as a document with its relations (textual and structured) as words

Does LDA extract semantic vectors?

relation	**most similar relations**
has_target	$\xrightarrow{dobj} form \xleftarrow{prep} with \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} degrade \xleftarrow{agent} by \xleftarrow{pobj}$
	$\xrightarrow{nn} activity \xrightarrow{nsubjpass} inhibit \xleftarrow{agent} by \xleftarrow{pobj}$
	$\xrightarrow{nsubj} inhibit \xleftarrow{dobj} phosphorylation \xleftarrow{prep} of \xleftarrow{pobj}$
	$\xrightarrow{nsubj} show \xleftarrow{dobj} affinity \xleftarrow{prep} for \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} show \xleftarrow{xcomp} interact \xleftarrow{prep} with \xleftarrow{pobj}$
	$\xrightarrow{dep} form \xleftarrow{dobj}$
	$\xrightarrow{nsubj} inhibit \xleftarrow{prep} in \xleftarrow{pobj} presence \xleftarrow{prep} of \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} cross-linked \xleftarrow{prep} to \xleftarrow{pobj}$
	$\xrightarrow{dep} form \xleftarrow{nsubjpass}$
	$\xrightarrow{dobj} inhibit \xleftarrow{prep} with \xleftarrow{pobj}$
	$\xrightarrow{nsubj} potentiate \xleftarrow{dobj} activity \xleftarrow{prep} of \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} prepare \xleftarrow{agent} by \xleftarrow{pobj} reaction \xleftarrow{prep} of \xleftarrow{pobj}$
	$\xrightarrow{nn} substrate \xleftarrow{prep} include \xleftarrow{pobj}$
	$\xrightarrow{nsubj} act \xleftarrow{prep} by \xleftarrow{pobj}$

Does LDA extract semantic vectors?

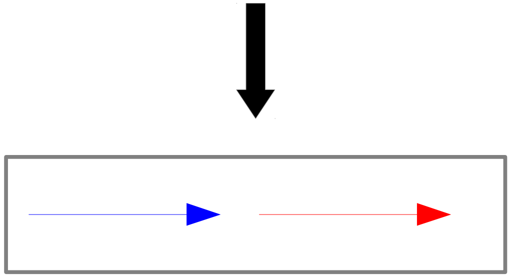
relation	most similar relations
may_treat	$\xrightarrow{pobj} \mathit{with} \xrightarrow{prep} \mathit{patient} \xrightarrow{nsubjpass} \mathit{treat} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$
	$\xrightarrow{nsubj} \mathit{be} \xleftarrow{prep} \mathit{in} \xleftarrow{pobj} \mathit{treatment} \xleftarrow{prep} \mathit{of} \xleftarrow{pobj}$
	$\xrightarrow{nsubj} \mathit{be} \xleftarrow{attr} \mathit{treatment} \xleftarrow{prep} \mathit{for} \xleftarrow{pobj}$
	$\xrightarrow{nn} \mathit{patient} \xleftarrow{partmod} \mathit{treat} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} \mathit{use} \xleftarrow{prep} \mathit{in} \xleftarrow{pobj} \mathit{treatment} \xleftarrow{prep} \mathit{of} \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} \mathit{use} \xleftarrow{prep} \mathit{for} \xleftarrow{pobj} \mathit{treatment} \xleftarrow{prep} \mathit{of} \xleftarrow{pobj}$
	$\xrightarrow{pobj} \mathit{with} \xrightarrow{prep} \mathit{treat} \xleftarrow{prep} \mathit{for} \xleftarrow{pobj}$
	$\xrightarrow{dobj} \mathit{receive} \xleftarrow{prep} \mathit{for} \xleftarrow{pobj}$
	$\xrightarrow{attr} \mathit{be} \xleftarrow{prep} \mathit{in} \xleftarrow{pobj} \mathit{treatment} \xleftarrow{prep} \mathit{of} \xleftarrow{pobj}$
	$\xrightarrow{pobj} \mathit{with} \xrightarrow{prep} \mathit{patient} \xleftarrow{rcmod} \mathit{treat} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} \mathit{administer} \xleftarrow{prep} \mathit{to} \xleftarrow{pobj} \mathit{patient} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$
	$\xrightarrow{nsubjpass} \mathit{use} \xleftarrow{prep} \mathit{in} \xleftarrow{pobj} \mathit{patient} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$
	$\xrightarrow{dobj} \mathit{use} \xleftarrow{prep} \mathit{in} \xleftarrow{pobj} \mathit{patient} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$
	$\xrightarrow{nsubj} \mathit{improve} \xleftarrow{prep} \mathit{in} \xleftarrow{pobj} \mathit{patient} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$
	$\xrightarrow{nsubj} \mathit{have} \xleftarrow{prep} \mathit{in} \xleftarrow{pobj} \mathit{patient} \xleftarrow{prep} \mathit{with} \xleftarrow{pobj}$

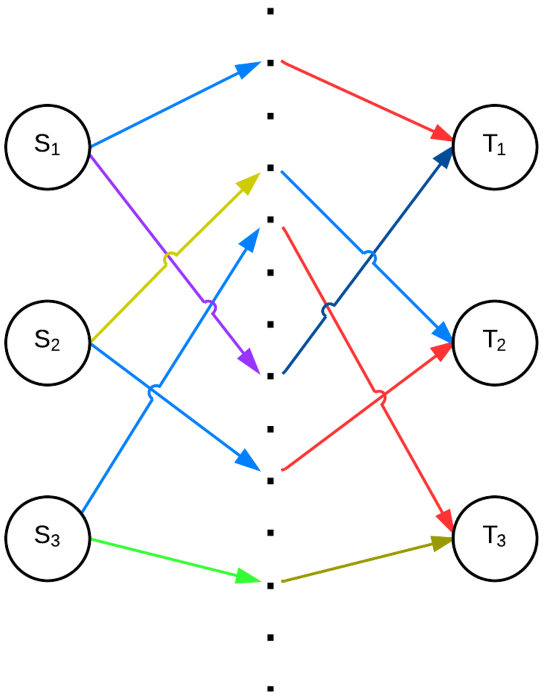
Task

Find characteristic (path-)patterns for relations in the knowledge graph

$$R_l = \{(S1, T1), (S2, T2), (S3, T3)\}$$

Does this relation have a characteristic path pattern?



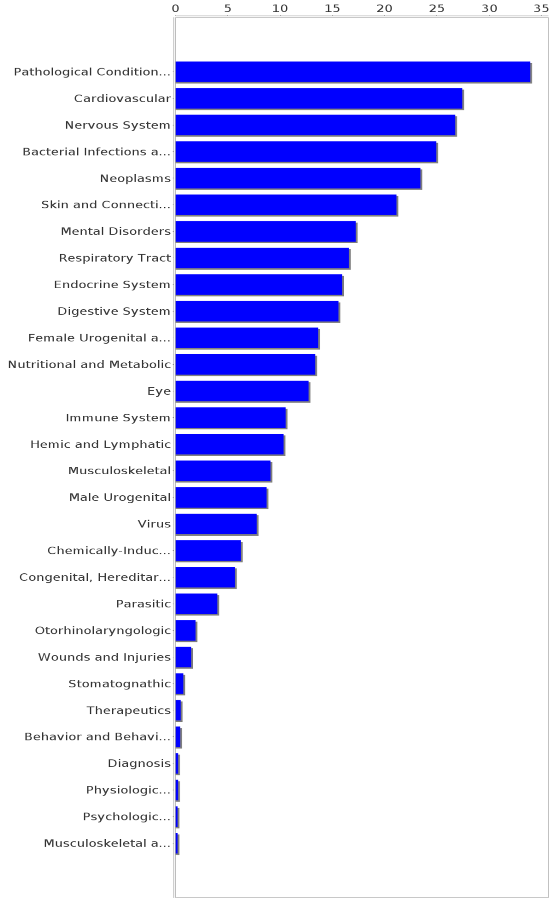


Paths found in knowledge graph

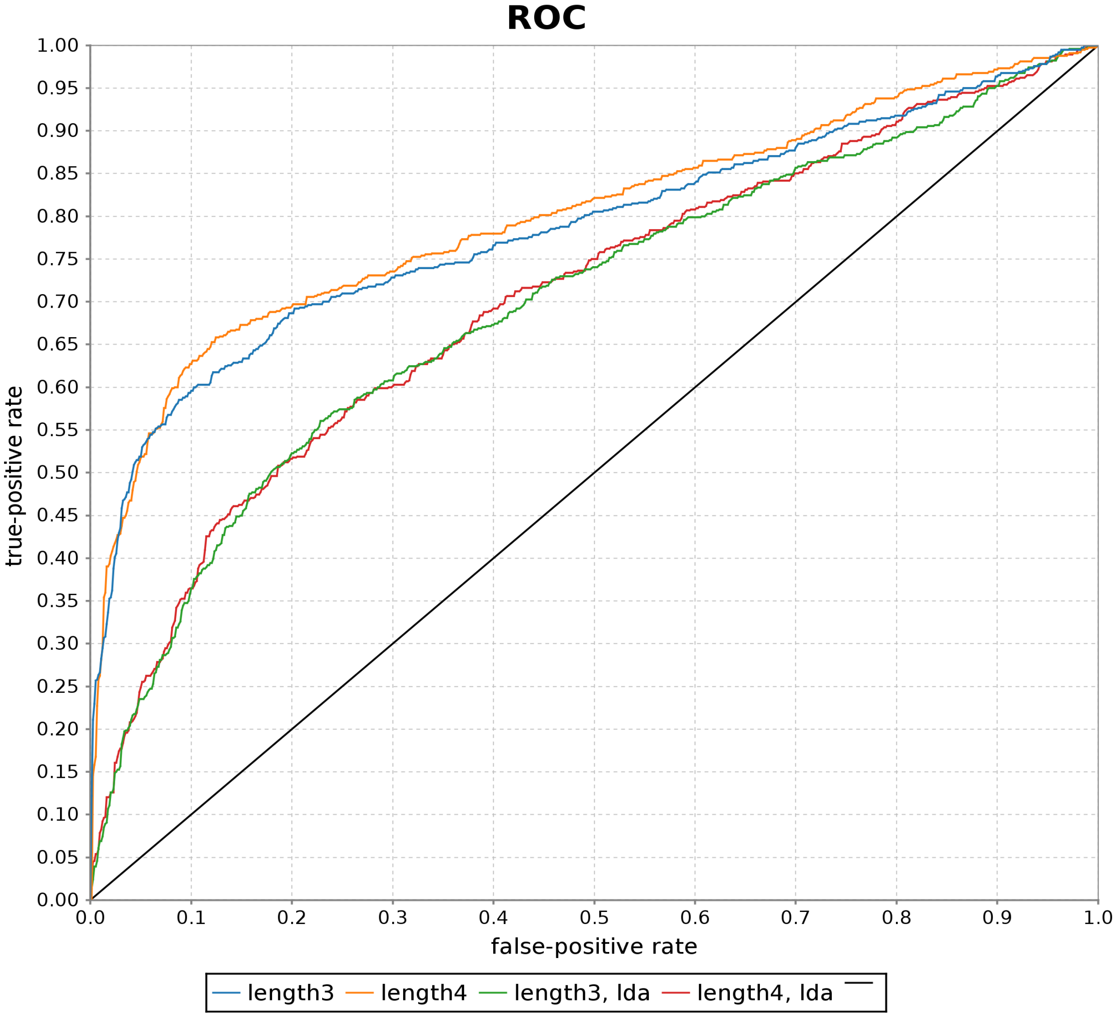
Experiments

- extraction of datasets
  - *may treat* (410 pairs from UMLS)
  - *has target* (740 pairs from DrugBank)
- construction of negative pairs for specific relation
- extraction of paths in knowledge graph for all pairs of the positive and negative examples
- training of logistic regression classifier with both *one-in-N* and *LDA* features
- evaluations focusing on high precision

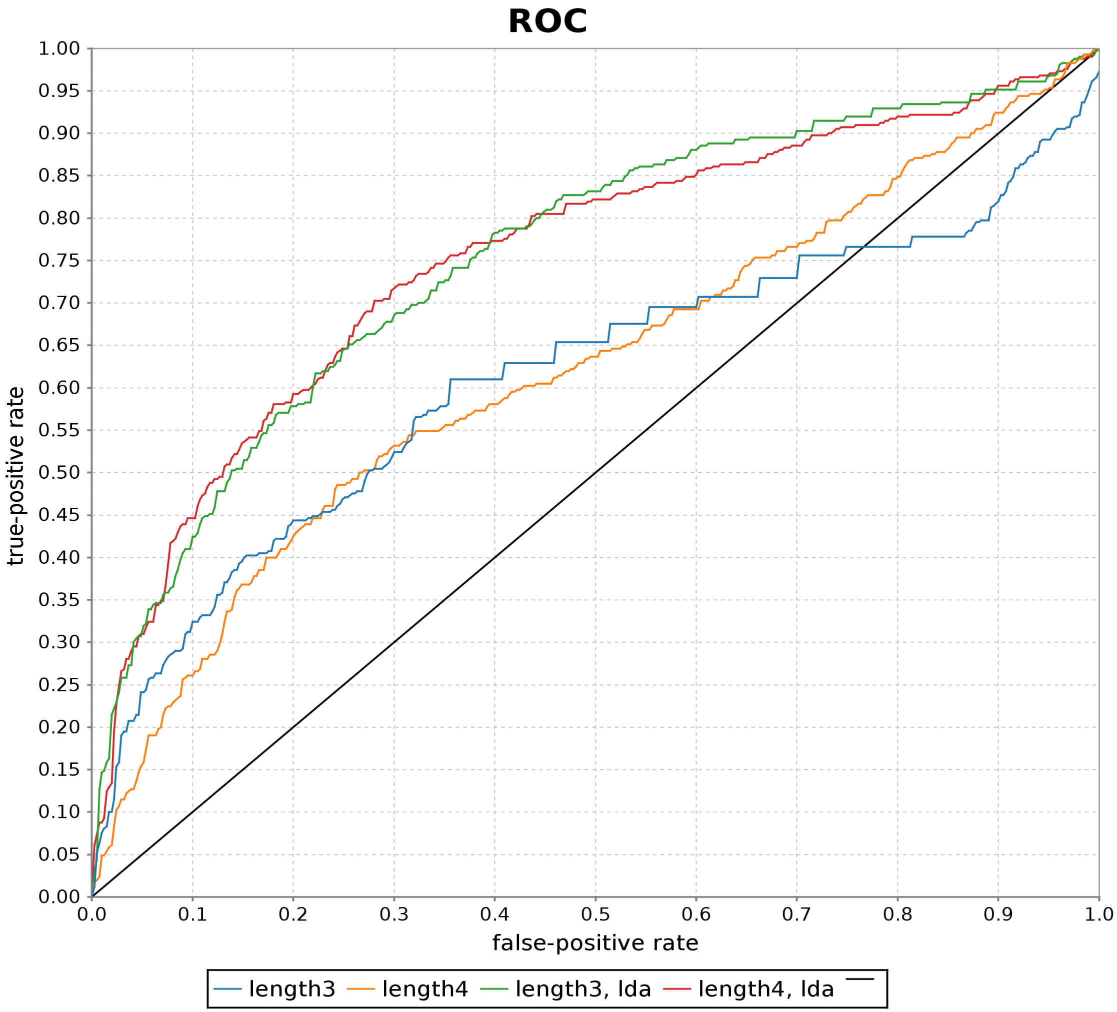
diseases type distribution



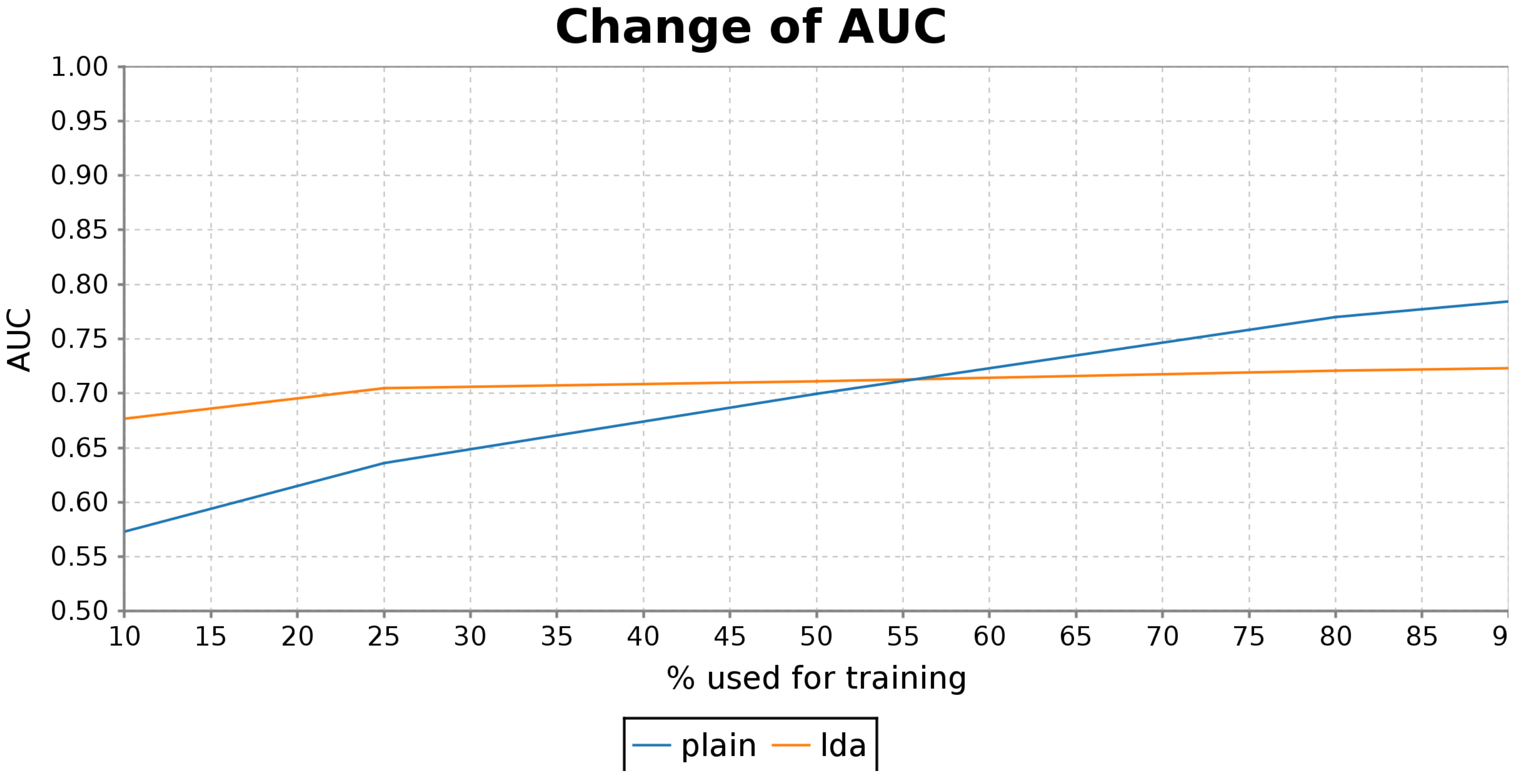
Impact of lengths and feature types, *has target*



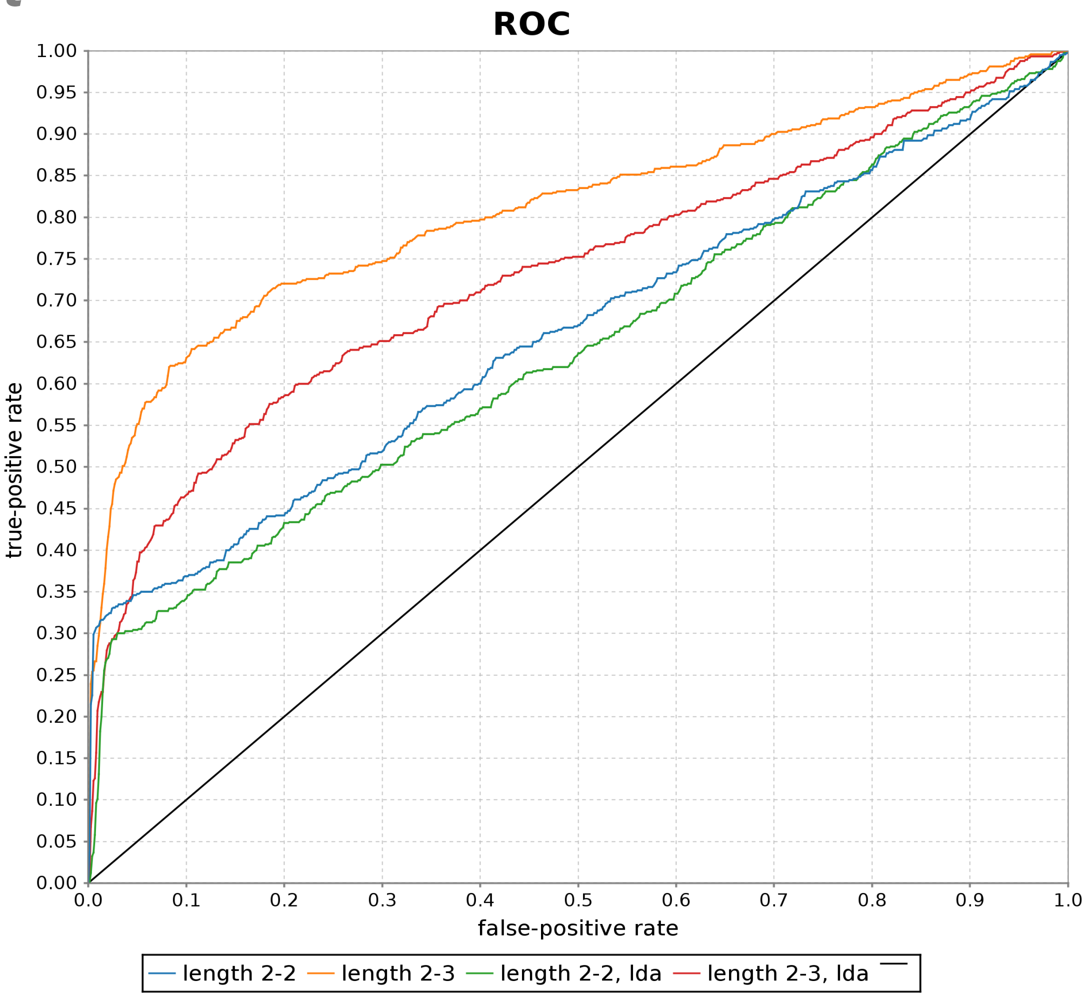
Impact of lengths and feature types, *may treat*



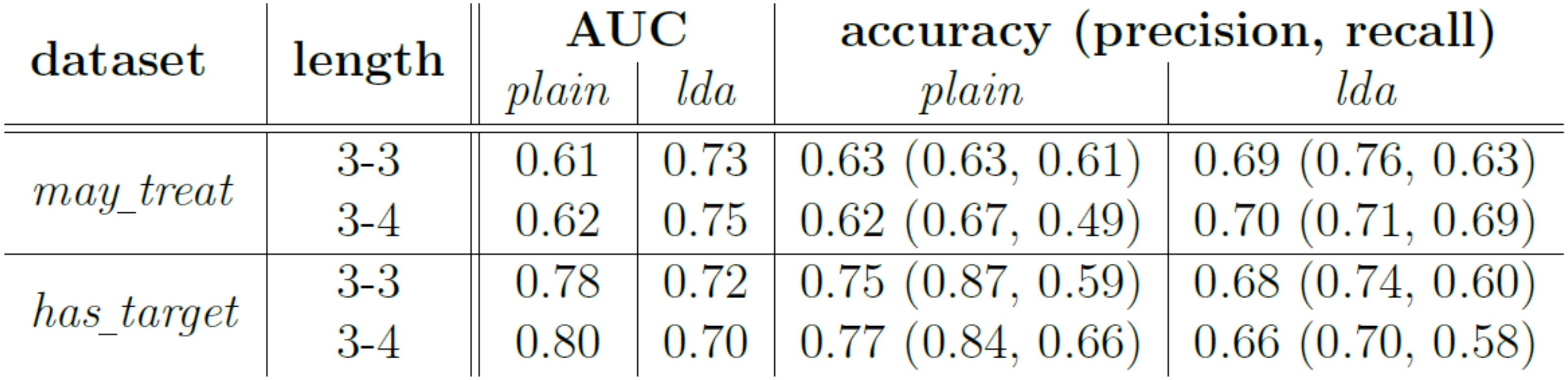
One-in-N vs. *LDA* features, *has target*



Direct vs. Direct + Indirect Connections, *has target*



Summary of Results



dataset	**length**	**AUC**		**accuracy (precision, recall)**	
		plain	*lda*	*plain*	*lda*
may_treat	3-3	0.61	0.73	0.63 (0.63, 0.61)	0.69 (0.76, 0.63)
	3-4	0.62	0.75	0.62 (0.67, 0.49)	0.70 (0.71, 0.69)
has_target	3-3	0.78	0.72	0.75 (0.87, 0.59)	0.68 (0.74, 0.60)
	3-4	0.80	0.70	0.77 (0.84, 0.66)	0.66 (0.70, 0.58)

Example Paths

highly weighted feature	**explanation**
$(\xrightarrow{dep} induce \xleftarrow{prep} in \xleftarrow{pobj}),\ (\xrightarrow{pobj} in \xrightarrow{prep} express \xleftarrow{nsubjpass})$	The substance is induced into something, in which the target (gene/protein) is expressed.
$(\xrightarrow{pobj} by \xrightarrow{agent} suppress \xleftarrow{nsubjpass}),\ (\xrightarrow{nsubj} increase \xleftarrow{prep} at \xleftarrow{pobj})$	The drug suppresses something that is increased by the disease.

Conclusions

- ***automatic discovery*** of relations using only ***indirect knowledge*** is ***possible***

- using not only direct but also ***indirect knowledge*** to ***discover relations*** between concepts is ***very useful***

- ***semantic (LDA)*** encoding ***help***, when data is ***sparse***

Thank you for your attention!
Questions?

Statistics

Textual part of the graph after pruning

Concepts/Vertices: ~ 95,000

Avg. degree: ~ 410.5

Connected Pairs: ~ 9 million

Most common concepts:

cell, rat, mouse, disease, proteins, …

Edges: ~ 39 million

Textual relation labels: ~ 105,000

Avg. occurrence: ~ 371.27

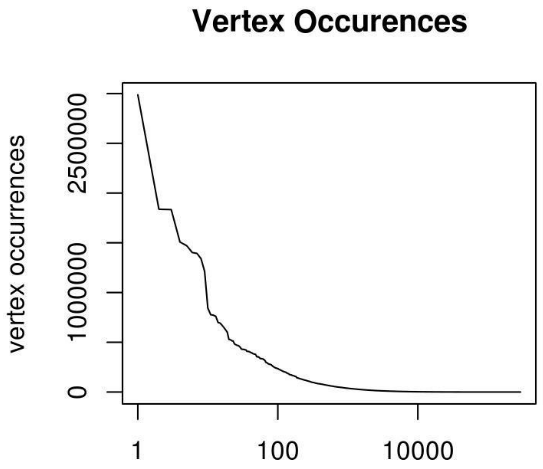
Most frequently occurring textual relation labels:

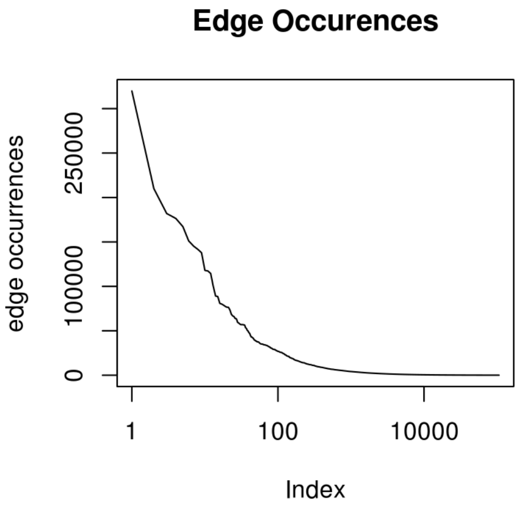
hmod → treat *amod*→

hmod → induce *amod*→

←*prep* include ←*pobj*

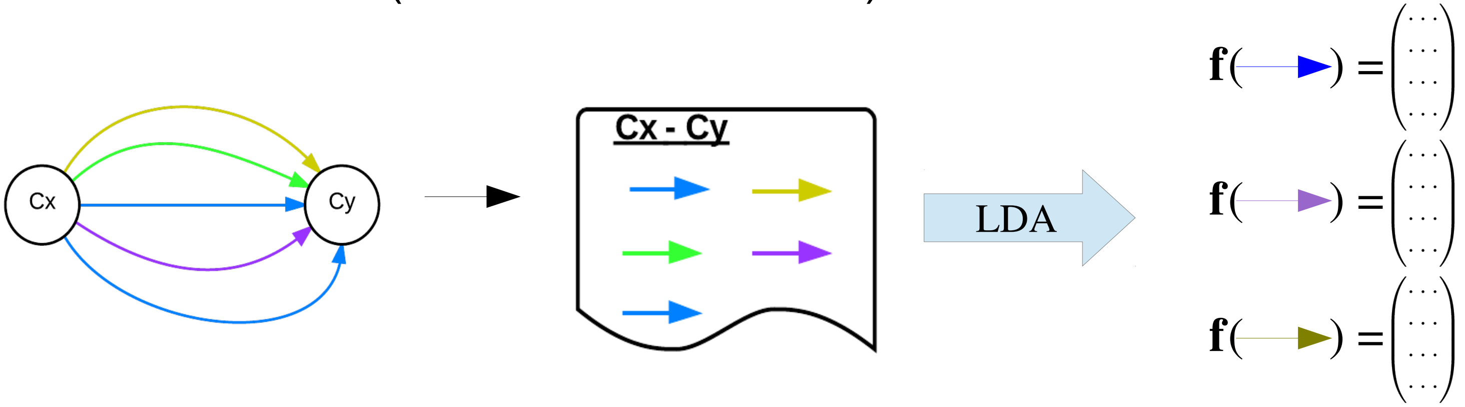
nsubj→ be ←*prep* in ←*pobj*





From Word- to Relation-Spaces

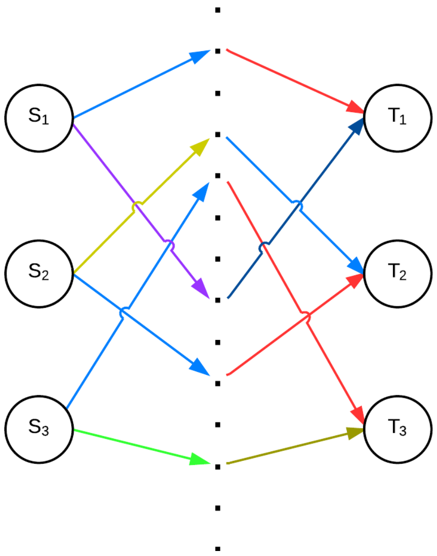
- numerous co-occurrence based algorithms computing semantic vectors for words co-occurring in a set of documents, e.g.,
  - latent semantic analysis (LSA)
  - reflective random indexing (RRI)
  - generalization of principle component analysis (gPCA)
  - **latent dirichlet allocation (LDA)**
- directly applicable by denoting a pair of concepts as a document with its relations (textual and structured) as words



Overview

$$R_l = \{(S_1, T_1), (S_2, T_2), (S_3, T_3), ...\}$$

1. Extract Paths



2. Encode pairs as feature vectors

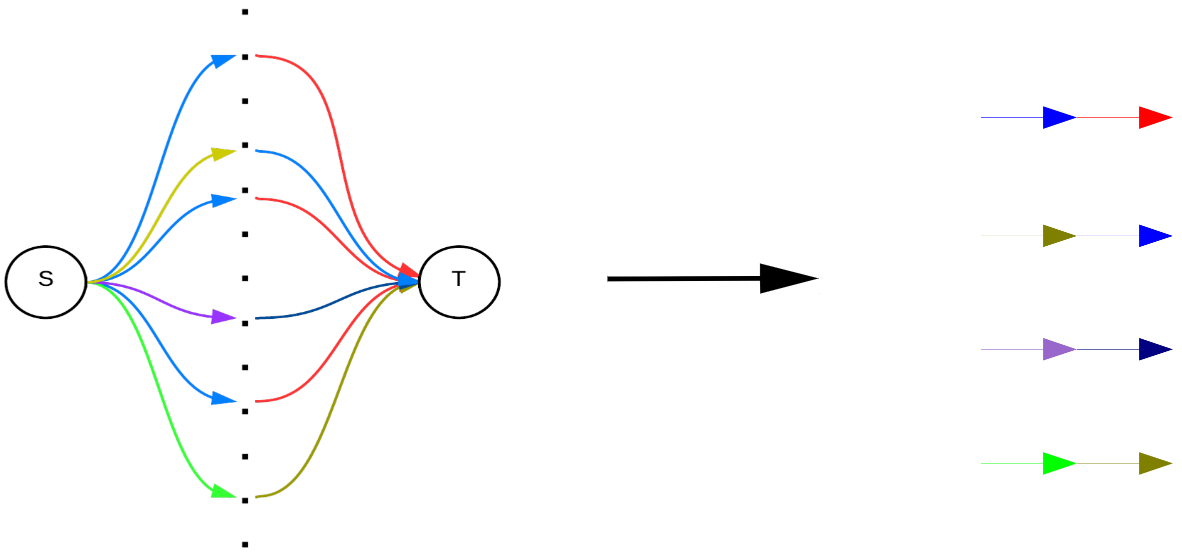
$$\{\boldsymbol{f}_{S_1,T_1}, \boldsymbol{f}_{S_2,T_2}, \boldsymbol{f}_{S_3,T_3}, \cdots\}$$

3. Train Classifier

$$(S, T) \in R_l \quad \Leftrightarrow \quad c_l(\boldsymbol{f}_{S,T}) > \theta$$

Modeling

- How to represent a pair of concepts as a feature vector?
- *concept pair* = ***multiset of paths*** in the knowledge graph
- each path pattern (i.e., sequence of relations) between the source and target concept becomes a feature

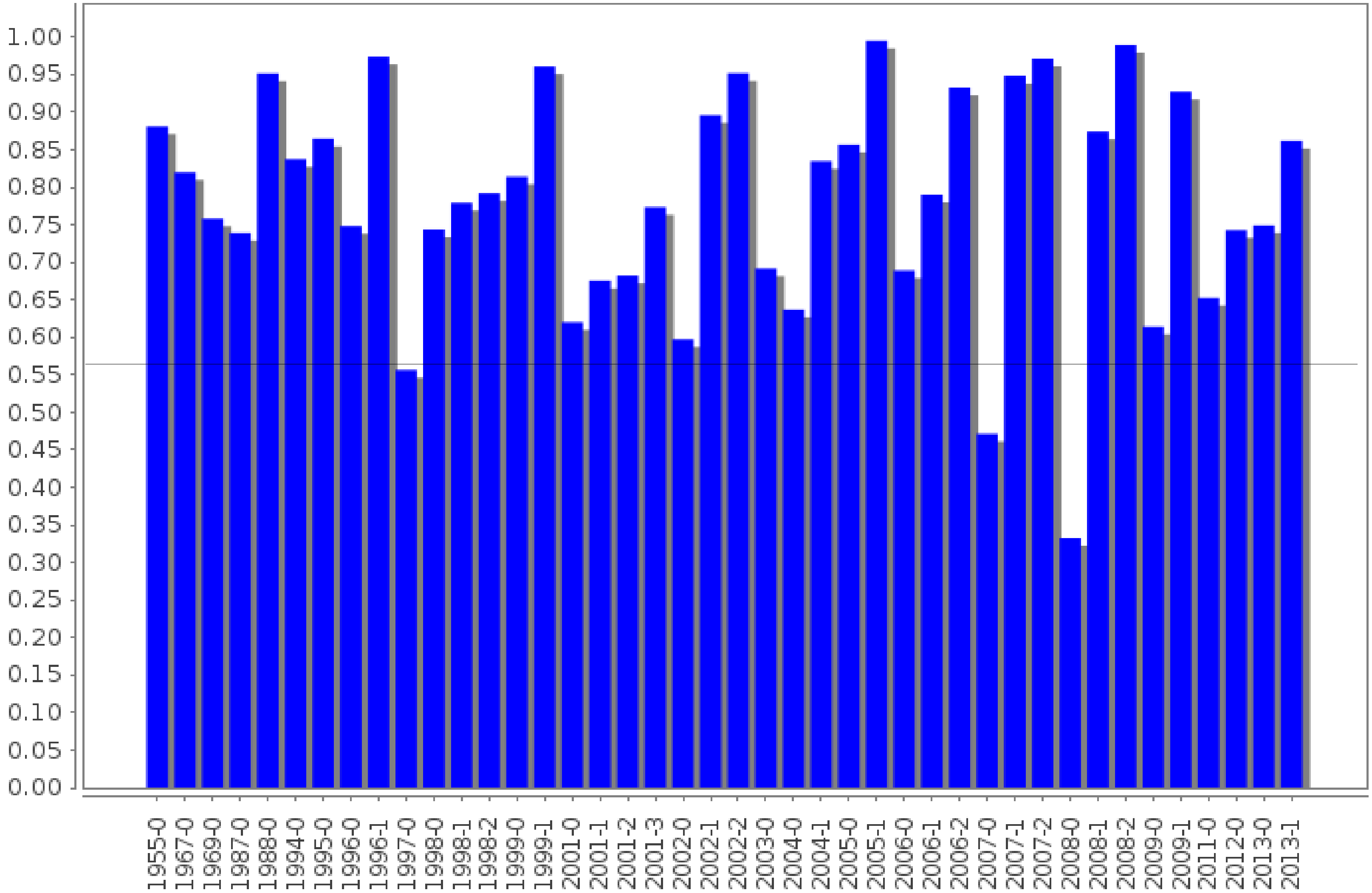


$$\begin{pmatrix}\cdots\\ 3\\ \cdots\\ 1\\ \cdots\\ 1\\ \cdots\\ 1\\ \cdots\end{pmatrix} = \mathbf{f}_{s,t}$$

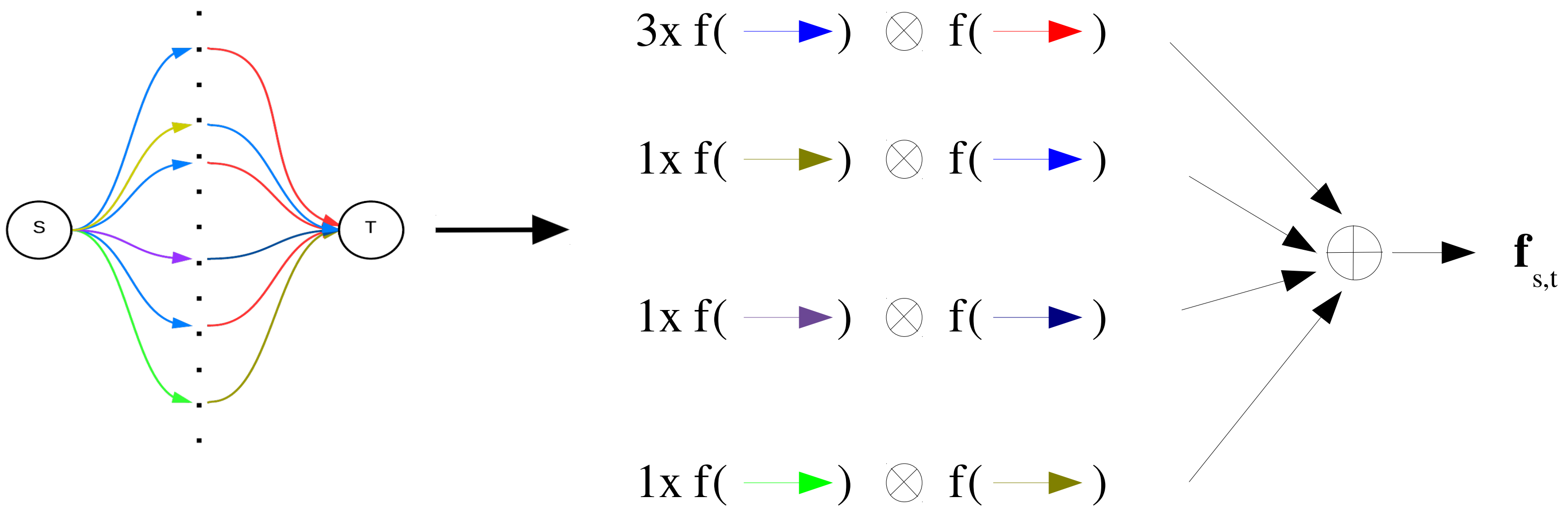
Some problems with this approach

- text mining error at every stage:
  - concept annotation, e.g. the “IMPACT gene”, Retinoic Acid Response Element abbreviated as “RARE”, the “Household gene”, etc.
  - POS tagging + dependency parsing more error prone on scientific text
- dependency paths neglect context of the assertions being made
- attributes of nouns or verbs neglected, e.g., *level* occurs in dependency path, but not the quality of the mentioned *level* (*high*, *low*, …)
- no co-reference resolution → missing knowledge

Classification scores of repositioned drug-disease pairs



General Modeling



3x f(→) ⊗ f(→)

1x f(→) ⊗ f(→)

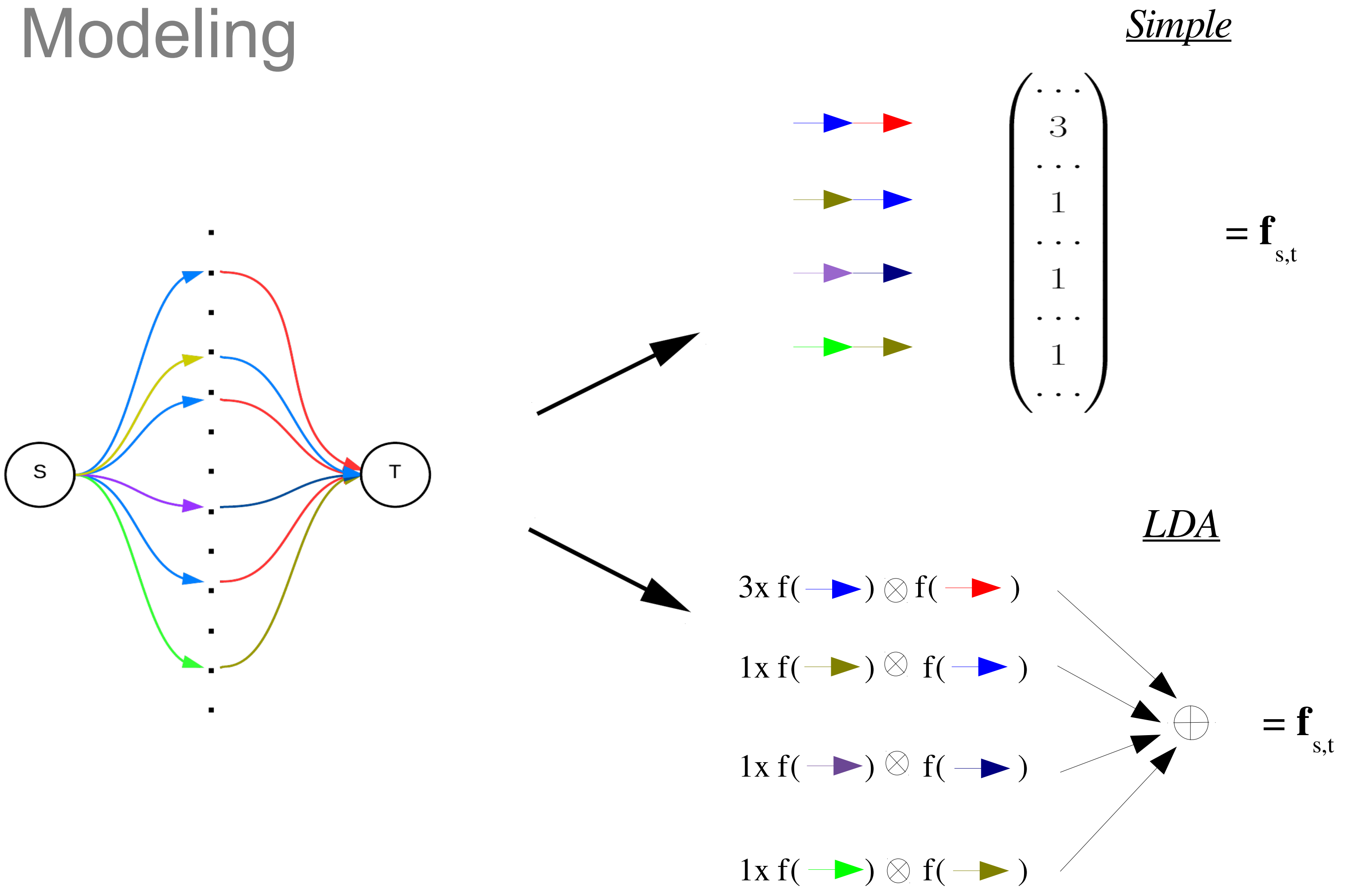
1x f(→) ⊗ f(→)

1x f(→) ⊗ f(→)

$\oplus \rightarrow \mathbf{f}_{s,t}$

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^{\mathrm{T}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \\ u_4v_1 & u_4v_2 & u_4v_3 \end{bmatrix}.$$

Modeling



Simple

→→
→→
→→
→→

$$\begin{pmatrix} \cdots \\ 3 \\ \cdots \\ 1 \\ \cdots \\ 1 \\ \cdots \\ 1 \\ \cdots \end{pmatrix} = \mathbf{f}_{s,t}$$

LDA

3x f(→) ⊗ f(→)

1x f(→) ⊗ f(→)

1x f(→) ⊗ f(→)

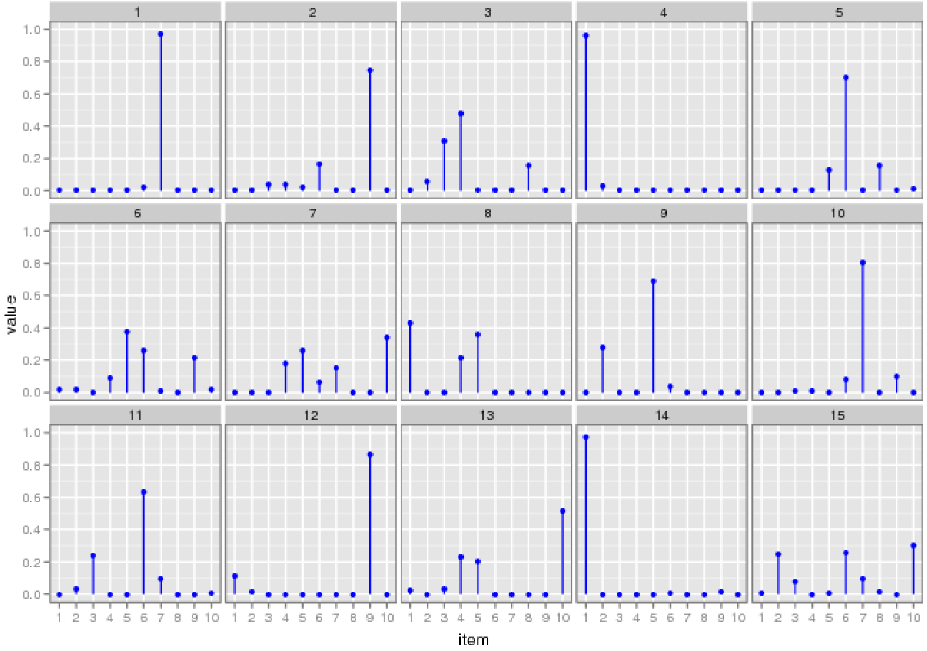
1x f(→) ⊗ f(→)

$\oplus$ $= \mathbf{f}_{s,t}$

LDA modeling relations between concepts

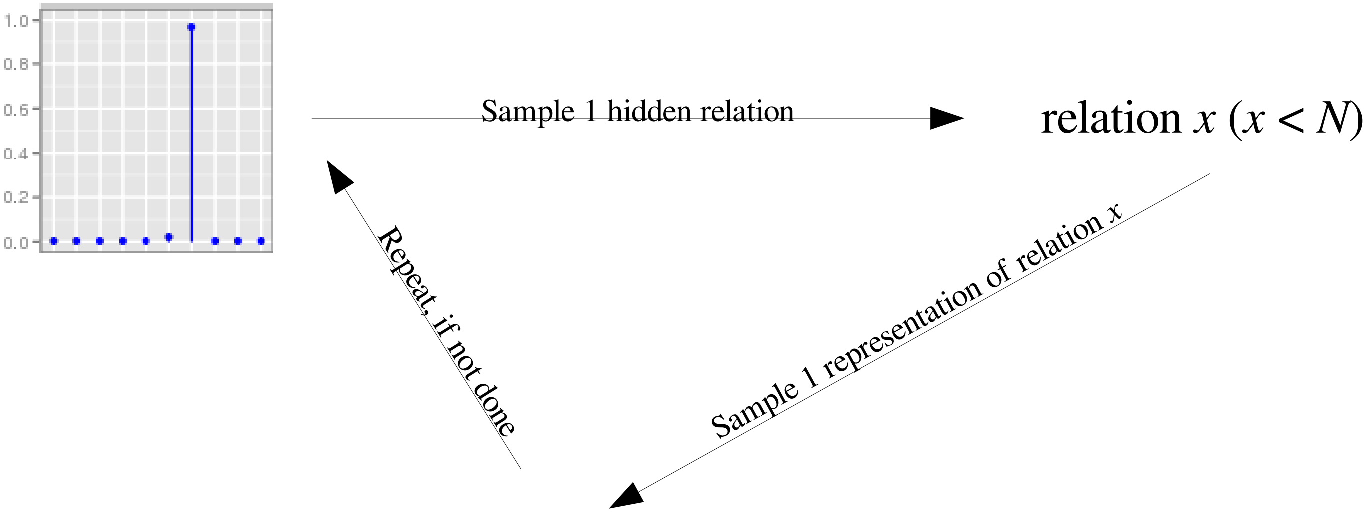
- pair of concepts = probability distribution of *N* real, but hidden relations (topics of LDA)

$\alpha = 0.1$



LDA modeling relations between concepts

- given: 1 pair of concepts = distribution of hidden but real relations



Sample 1 hidden relation

relation *x* ($x < N$)

Sample 1 representation of relation *x*

Repeat, if not done

E.g.:
hmod → treat *amod*→ (from Medline)
hmod → induce *amod*→ (from Medline)
may_prevent (from UMLS)