Pattern Structure Projections for Learning Discourse Structures

Fedor Strok[2], Boris A.Galitsky[1], Dmitry Ilvovsky[2]
Sergei O. Kuznetsov[2]

[1] Knowledge Trail Inc. San Jose CA USA
[2] Higher School of Economics, Moscow Russia

fdr.strok@gmail.com; bgalitsky@hotmail.com; dilvovsky@hse.ru; skuznetsov@hse.ru

Outline

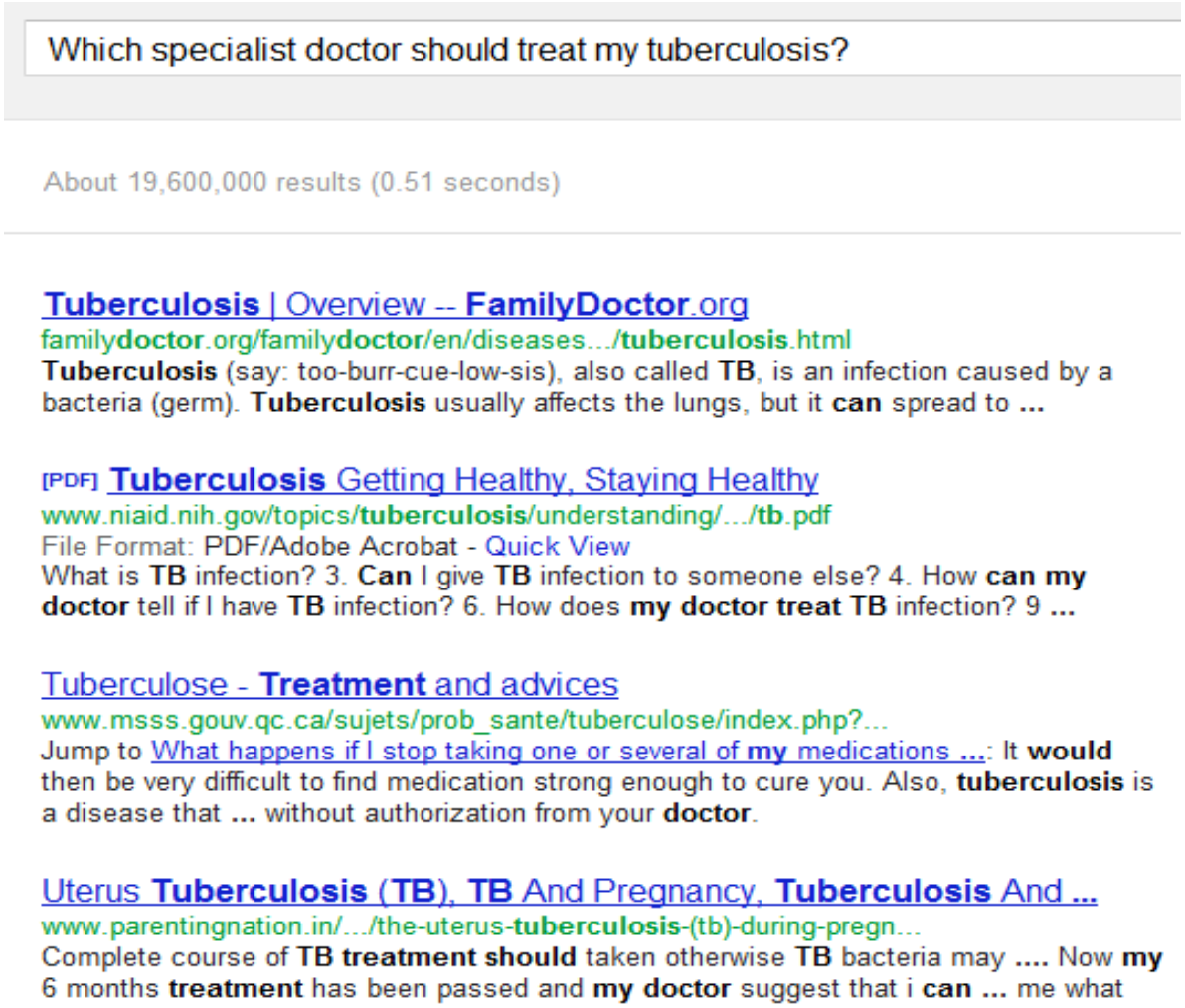
1. NLP search task and motivation
2. Similarity between short texts
3. Parse Thicket model
4. Pattern structures and their projections
5. Clustering of search results

Working with short texts is nice

It sounds unreal but short texts are… useful

- Search and Q/A
- Content analysis:
  - Classification
  - Categorization
- Content generation
- Recommendations
- Advertisement

Answer in multiple sentences?



- No answer includes ‘pulmonologist’

Similarity between question and answer

- **Baseline**: **bag-of-words** approach, which computes the set of common keywords/n-grams and their frequencies.

- **Pair-wise sentence matching**: **syntactic generalization** for each **pair of sentences** and summation of the resultant similarities [Galitsky et al., 2012]

- **Paragraph-paragraph matching**. Similarity as a **generalization** operation

Finding similarity between paragraphs. Small example

(1.1)Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons. (1.2)UN nuclear watchdog passes a resolution condemning Iran for developing a second uranium enrichment site in secret. (1.3)A recent IAEA report presented diagrams that suggested Iran was secretly working on nuclear weapons. (1.4)Iran envoy says its nuclear development is for peaceful purpose, and the material evidence against it has been fabricated by the US.

^

(2.1)UN passes a resolution condemning the work of Iran on nuclear weapons, in spite of Iran claims that its nuclear research is for peaceful purpose. (2.2)Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops an enrichment site in secret. (2.3)Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium enrichment site.

Keywords: common topics with no details

Iran, UN, proposal, dispute, nuclear, weapons, passes, resolution, developing, enrichment, site, secret, condemning, second, uranium

Improvement: pair-wise sentence similarity

[NN-work IN-* IN-on JJ-nuclear NNS-weapons]**, [DT-the NN-dispute IN-over JJ-nuclear NNS-*]**, [VBZ-passes DT-a NN-resolution],
[VBG-condemning NNP-iran IN-*],
[VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]],
[DT-* JJ-second NN-uranium NN-enrichment NN-site]],
[VBZ-is IN-for JJ-peaceful NN-purpose],
[DT-the NN-evidence IN-* PRP-it]**, [VBN-* VBN-fabricated IN-by DT-the NNP-us]**

Improvement: pair-wise sentence similarity

[NN-work IN-* IN-on JJ-nuclear NNS-weapons], **[DT-the NN-dispute IN-over JJ-nuclear NNS-*]**, [VBZ-passes DT-a NN-resolution],
[VBG-condemning NNP-iran IN-*],
[VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]],
[DT-* JJ-second NN-uranium NN-enrichment NN-site]],
[VBZ-is IN-for JJ-peaceful NN-purpose],
[DT-the NN-evidence IN-* PRP-it], **[VBN-* VBN-fabricated IN-by DT-the NNP-us]**

Paragraph-level similarity

[NN-Iran VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]
[NN-generalization-<UN/nuclear watchdog> * VB-pass NN-resolution VBG condemning NN- Iran]
[NN-generalization-<Iran/envoy of Iran> *Communicative_action* DT-the NN-dispute IN-over JJ-nuclear NNS-*]
[Communicative_action - NN-work IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]
[NN-generalization <Iran/envoy to UN> Communicative_action NN-Iran NN-nuclear NN-* VBZ-is IN-for JJ-peaceful NN-purpose],
Communicative_action - NN-generalize <work/develop> IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]*
[NN-generalization <Iran/envoy to UN> Communicative_action NN-evidence IN-against NN Iran NN-nuclear VBN-fabricated IN-by DT-the NNP-us]
***condemn^proceed* [*enrichment site*] <leads to> *suggest^condemn* [*work Iran nuclear weapon*]**

[Iran nuclear NNP-*]<RST-evidence>[fabricated by USA]

Paragraph-paragraph similarity

[NN-Iran VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]
[NN-generalization-<UN/nuclear watchdog> * VB-pass NN-resolution VBG condemning NN- Iran]

[NN-generalization-<Iran/envoy of Iran> *Communicative_action* DT-the NN-dispute IN-over JJ-nuclear NNS-*]

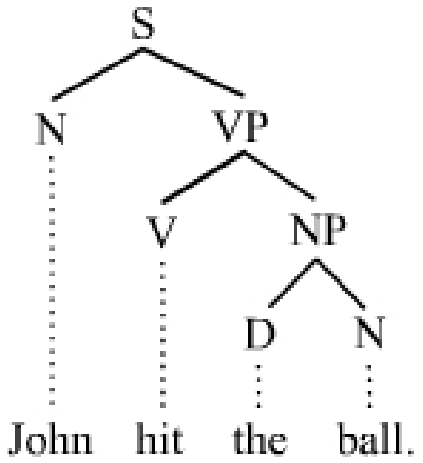
[Communicative_action - NN-work IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]
[NN-generalization <Iran/envoy to UN> Communicative_action NN-Iran NN-nuclear NN-* VBZ-is IN-for JJ-peaceful NN-purpose],
Communicative_action - NN-generalize <work/develop> IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]*
[NN-generalization <Iran/envoy to UN> Communicative_action NN-evidence IN-against NN Iran NN-nuclear VBN-fabricated IN-by DT-the NNP-us]

***condemn^proceed* [*enrichment site*] <leads to> *suggest^condemn* [*work Iran nuclear weapon*]**

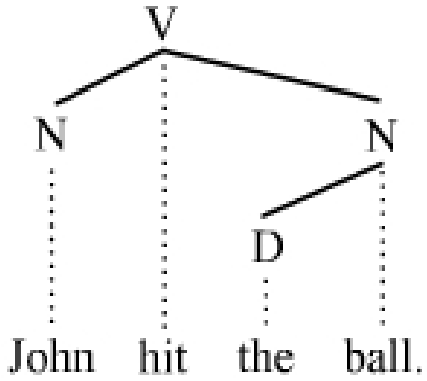
[Iran nuclear NNP-*]<RST-evidence>[fabricated by USA]

Syntactic sentence representation

Syntactic parse trees: constituency and dependency



Constituency-based parse tree



Dependency-based parse tree

Paragraph representation model?

- Sentence parse trees are well-studied. What about using them for paragraphs?
- We are concerned about structures. Can we use structure for treating similarity?
- Search engineers do not want to learn NLP. May we somehow help them?

Introducing Parse Thicket

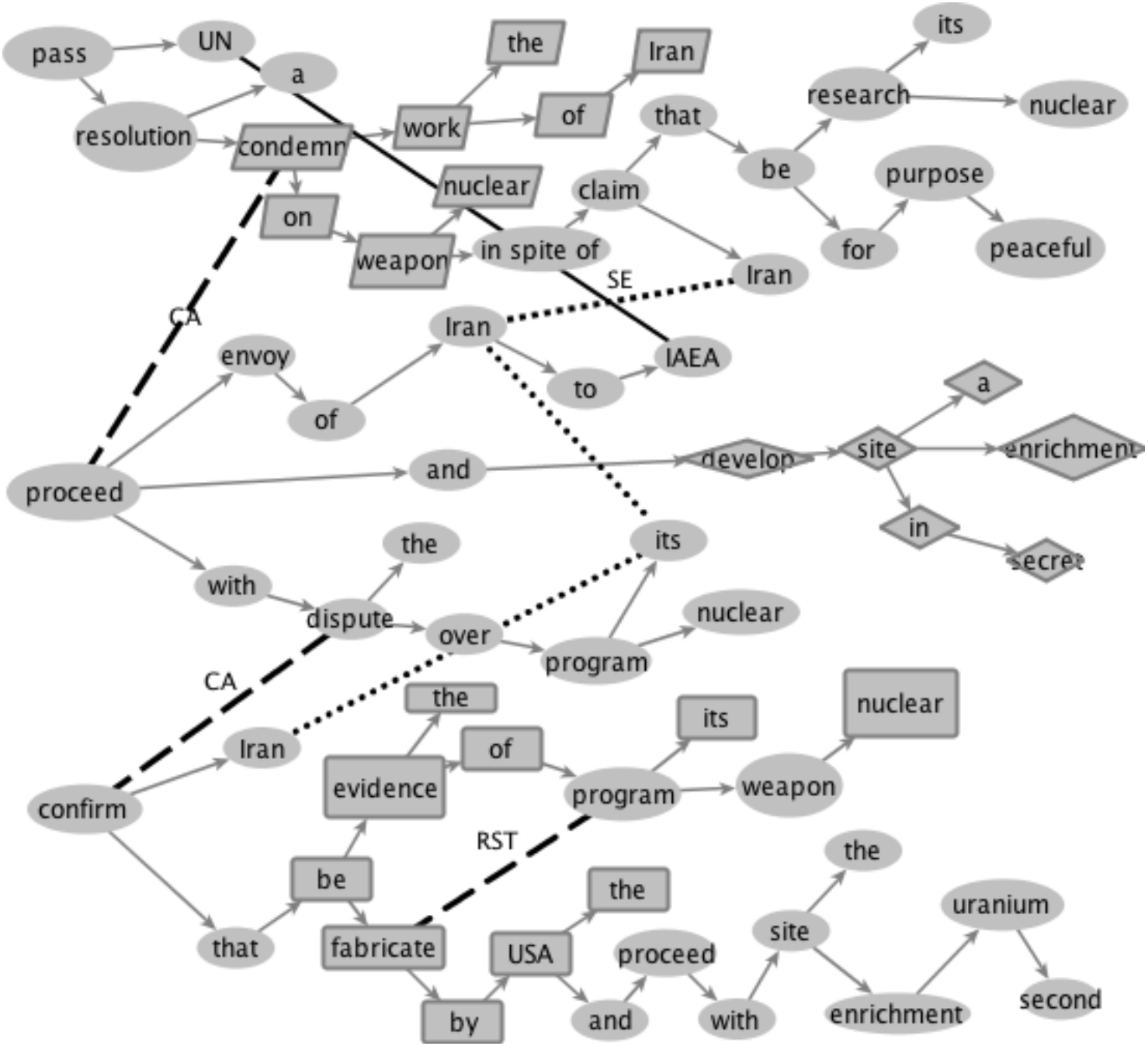
- Representation of a linguistic structure of a **paragraph of text**.
- **Syntactic information** + **discourse**.
- **Graph** structure: **parse trees** + **additional arcs** for inter-sentence relationship **between** parse tree nodes for **words**.

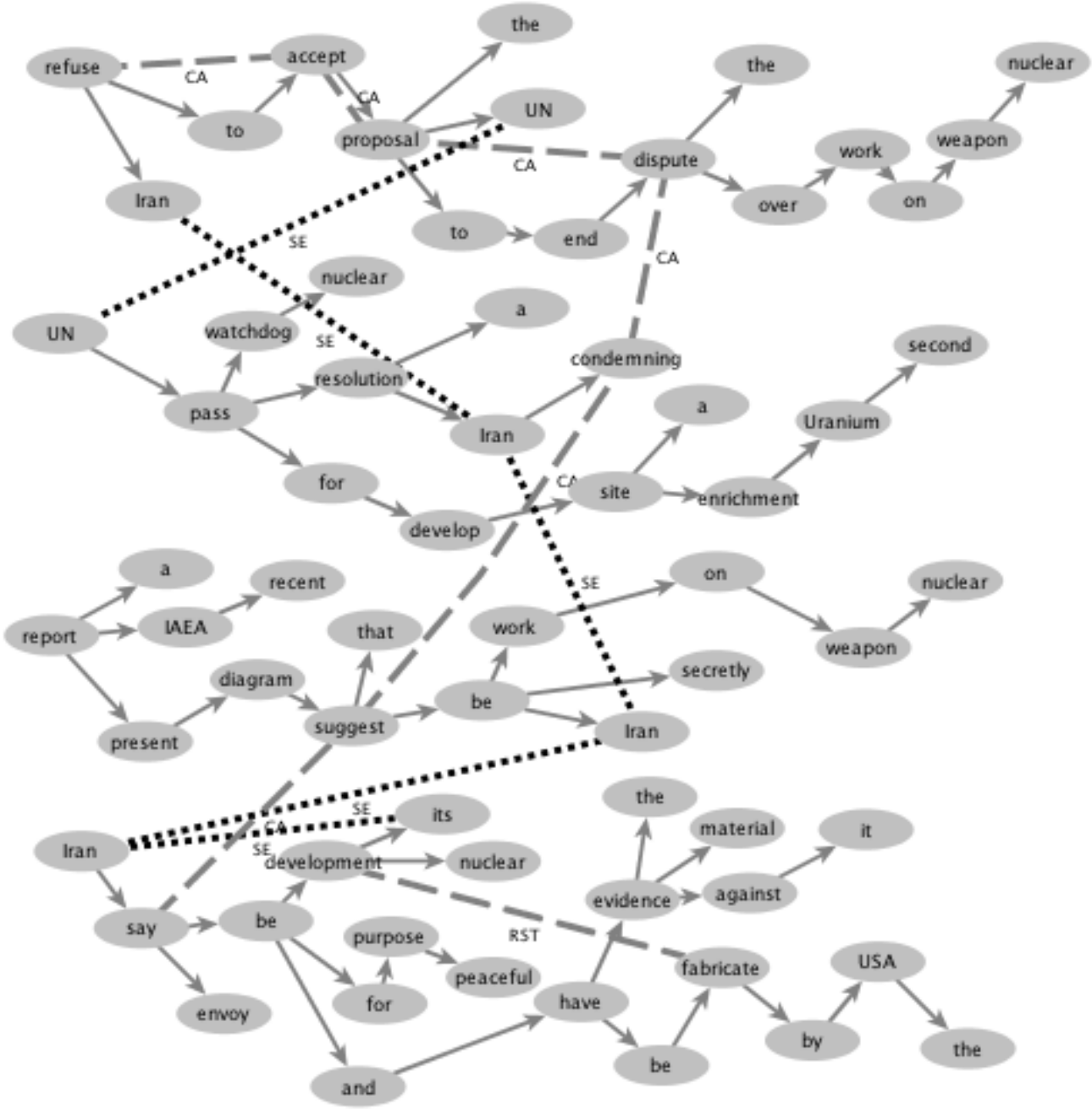
Discourse relations in Parse Thicket

- Taxonomy, coreferences
  - Anaphora
  - Same entity
  - Hyponym/Hyperonym
  - Siblings
  - etc
- Rhetorical Structure Theory (RST) [Mann]
- Speech Act Theory (SpAct) [Searle]

Why Parse Thickets?

- Least general generalization in terms of structural representations of text paragraphs
- Similarity between two texts as a generalization of their PT.
- Exploring machine learning on structures [Moschitti, Sun] at the level of paragraphs


pass
UN
a
resolution
condemn
work
the
of
Iran
that
research
its
nuclear
be
purpose
for
peaceful
on
nuclear
claim
weapon
in spite of
Iran
SE
CA
Iran
envoy
to
IAEA
of
proceed
and
develop
site
a
enrichment
in
secret
with
the
its
dispute
over
program
nuclear
CA
Iran
the
evidence
of
its
program
weapon
nuclear
confirm
RST
be
that
fabricate
the
USA
by
and
proceed
with
the
site
uranium
enrichment
second


refuse
accept
the
CA
to
proposal
UN
CA
dispute
the
nuclear
work
weapon
over
on
Iran
to
end
SE
CA
nuclear
UN
watchdog
SE
a
condemning
second
resolution
pass
Iran
a
Uranium
for
site
enrichment
develop
CA
SE
a
recent
on
nuclear
report
IAEA
that
work
weapon
diagram
secretly
be
present
suggest
Iran
SE
its
the
material
it
Iran
CA
SE
development
nuclear
evidence
against
say
be
purpose
RST
fabricate
USA
peaceful
envoy
for
have
by
the
and
be

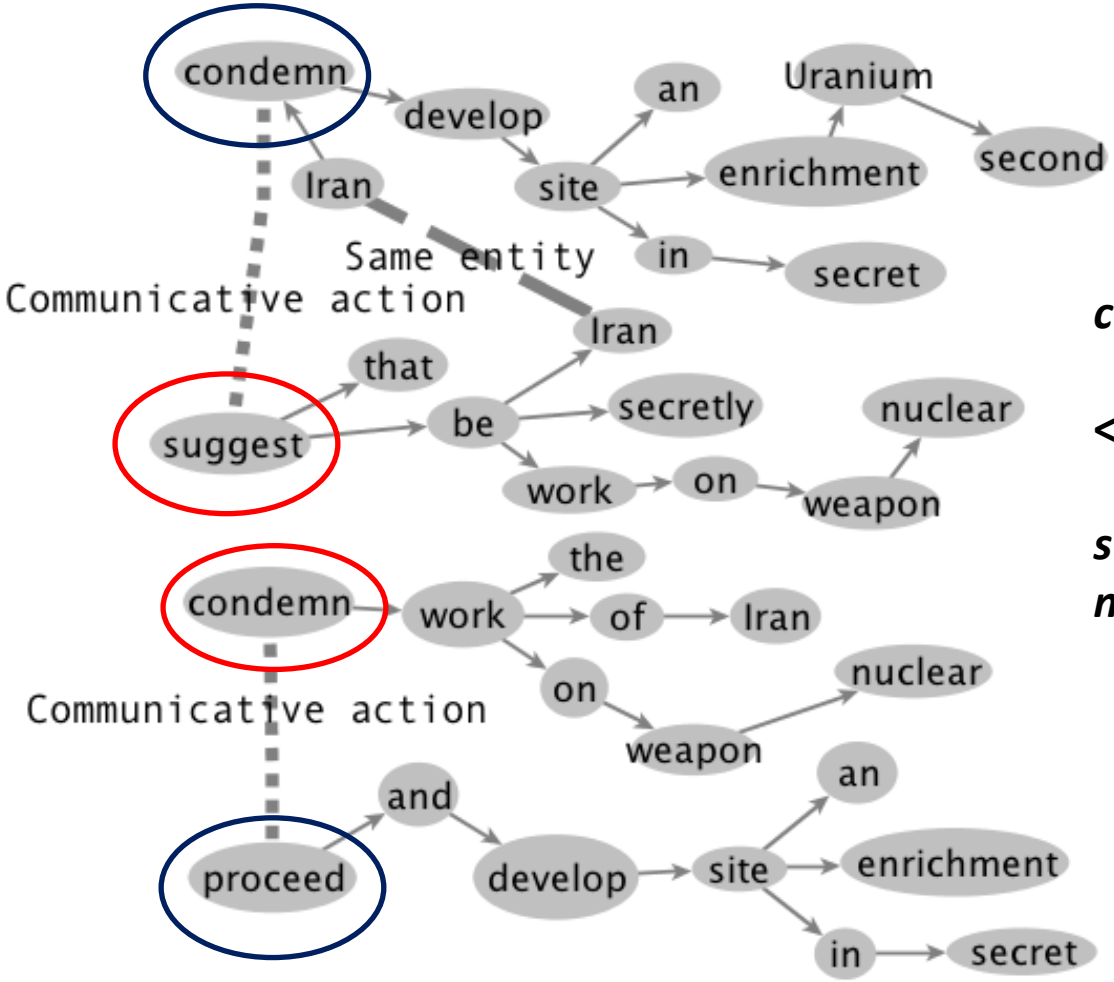
Generalization of 2 PT. Graphs

- PT -> graph
- Generalization is a set of all maximal common subgraphs
- Costs a lot:
  - NP-complete
  - But could be improved using specific of a PT [still in work]
- But explored anyhow [Galitsky et al., GKR 2013]

Generalization of 2 PT. Phrases

- PT -> set of phrases
  - Regular phrases (NP, VP, etc)
  - Thicket phrases (trees in a graph including coreferential and taxonomical arcs)
  - RST-phrases
  - CA-phrases
- Pair-wise generalization of phrases
- Works fast: appr. constant time for modified inverse index

Generalization of 2 PT: CA example

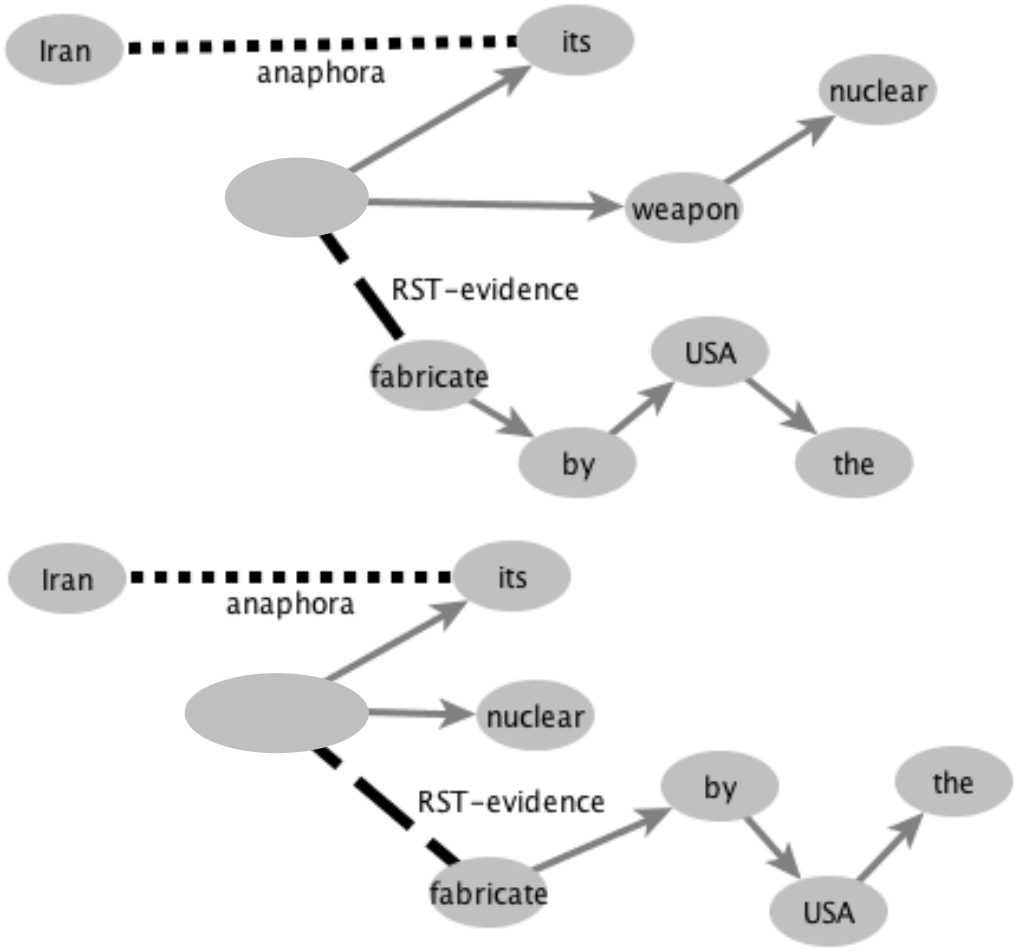


condemn^proceed [enrichment site]

<leads to>

suggest^condemn [work Iran nuclear weapon]

Generalization of 2 PT: RST example



Iran nuclear NN

– RST-evidence –

fabricated by USA

Search results clustering

- Find relevant answers - not the final stage
- Group results somehow – that's the point
- FCA lattices and binary attributes were used ages ago for this task [Romano; Zahiri]
- We have good descriptions, similarity operation. So, maybe…

Pattern structure

Triple $\left(G,\left(D,\mathrm{y}\right),\delta\right)$, where

G – set of objects

$\left(D,\mathrm{y}\right)$ – meet-semilattice of descriptions

$\delta : G \to D$

y - intersection (similarity) of two descriptions

y - associative and commutative

Pattern structure lattice

- Partial order $c \le d \Leftrightarrow c \sqcap d = c$
- Galois operators

$$A^{\square} = \prod_{g \in A} \delta(g), A \subseteq G$$

$$d^{\square} = \{ g \in G \mid d \le \delta(g) \}, d \in D$$

- Closure operator $A^{\square\square} = A, d^{\square\square} = d$
- Pattern (A, d)

$A^{\square} = d, d^{\square} = A$

- All patterns compose ***lattice***

Projection of pattern structure

Function $\psi : D \to D$

- Monotone $x \le y \Rightarrow \psi(x) \le \psi(y)$
- Contractive $\psi(x) \le x$
- Idempotent $\psi(\psi(x)) = \psi(x)$

$$\psi\left(\left(G,\left(D,\mathrm{y}\right),\delta\right)\right) = \left(G,\left(D_\psi,\mathrm{y}_\psi\right),\psi\circ\delta\right)$$

$$D_\psi = \psi(D) = \left\{d \in D \mid \exists d^* \in D : \psi\left(d^*\right) = d\right\}$$

$$\forall x, y \in D,\ x\,\mathrm{y}_\psi\ y = \psi\left(x\,\mathrm{y}\ y\right)$$

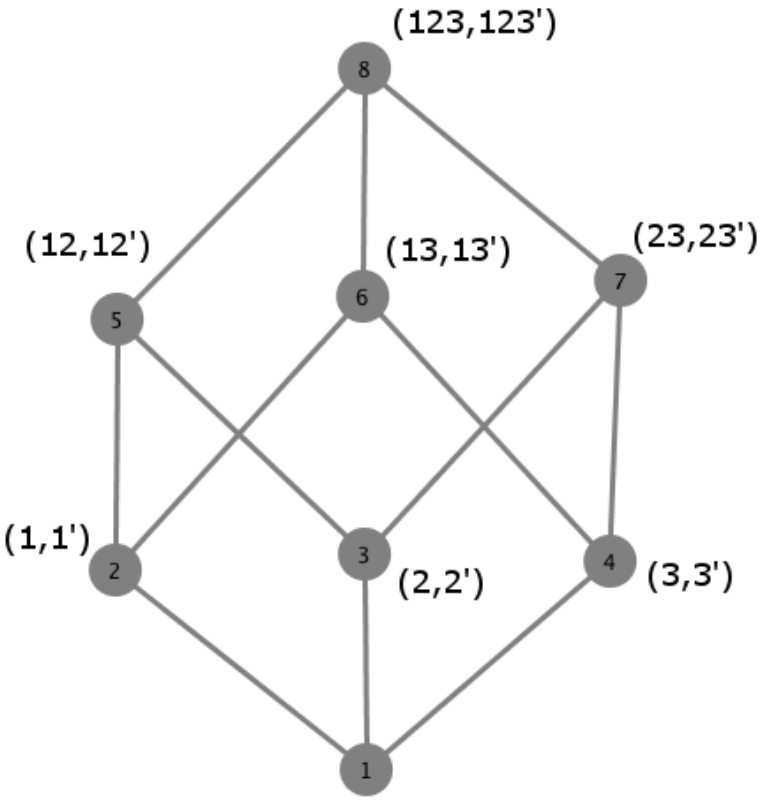
Pattern structure on parse thickets

- *Objects* – paragraphs
- *Descriptions* – parse thicket graph
- *Intersection* – similarity operation
- *Projection* – maximal (on inclusion) thicket phrases

Clustering. Example

1. *At least 9 people were killed and 43 others wounded in shootings and bomb attacks, including four car bombings, in central and western Iraq on Thursday, the police said. A car bomb parked near the entrance of the local government compound in Anbar's provincial capital of Ramadi, some 110 km west of Baghdad, detonated in the morning near a convoy of vehicles carrying the provincial governor Qassim al-Fahdawi, a provincial police source told Xinhua on condition of anonymity.*
2. *Officials say a car bomb in northeast Baghdad killed four people, while another bombing at a market in the central part of the capital killed at least two and wounded many more. Security officials also say at least two policemen were killed by a suicide car bomb attack in the northern city of Mosul. No group has claimed responsibility for the attacks, which occurred in both Sunni and Shi'ite neighborhoods.*
3. *A car bombing in Damascus has killed at least nine security forces, with aid groups urging the evacuation of civilians trapped in the embattled Syrian town of Qusayr. The Syrian Observatory for Human Rights said on Sunday the explosion, in the east of the capital, appeared to have been carried out by the extremist Al-Nusra Front, which is allied to al-Qaeda, although there was no immediate confirmation. In Lebanon, security sources said two rockets fired from Syria landed in a border area, and Israeli war planes could be heard flying low over several parts of the country.*

Clustering. Example



Clustering. Example

$\{1\}'$

[[NP [JJS-least CD-9 NNS-people], NP [CD-43 NNS-others], NP [NNS-shootings CC-and NN-bomb NNS-attacks], NP [NNS-shootings], NP [NN-bomb NNS-attacks], NP [CD-four NN-car NNS-bombings], NP [JJ-central CC-and JJ-western NNP-Iraq], NP [JJ-central], NP [JJ-western NNP-Iraq], NP [NNP-Thursday], NP [DT-the NN-police], NP [DT-A NN-car NN-bomb], NP [DT-the NN-entrance IN-of DT-the JJ-local NN-government NN-compound IN-in NNP-Anbar POS-'s JJ-provincial NN-capital IN-of NNP-Ramadi ,-, DT-some CD-110 NN-km NN-west IN-of NNP-Baghdad], NP [DT-the NN-entrance]

etc.

Clustering. Example

$$\{1,2\}'$$

Same place: *[NN-* NN-* IN-in NNP-**baghdad**]*

Same terms: *[NN-* NN-**bomb** NN-**attack**], [NNS-**attacks**]*

Same info about victims: *[VBD-**wounded**], [VBD-**were** VBN-**killed**], [CD-* NNS-**people**], [CD-**four** NNS-*]*.

etc.

Clustering. Example

$$\{1,2,3\}'$$

Car bombing near capitals:

[DT-a NN-car NN-bombing],

[DT-the NN-capital],

[VBN-killed],

[JJS-least CD- NN-*]*

etc

Search improvement results for PT approach

Query	Answer	Relevancy of baseline Yahoo search, %, averaging over 20 searches	Relevancy of baseline Bing search, %, averaging over 20 searches	Relevancy of re-sorting by pair-wise sentence generalization, %, averaging over 40 searches	Relevancy of re-sorting by thicket generalization based on RST, %, averaging over 20 searches	Relevancy of re-sorting by thicket generalization based on SpActT, %, averaging over 20 searches	Relevancy of re-sorting by hybrid RST-SpActT thicket generalization, %, averaging over 40 searches	Relevancy improvement for parse forest approach, comp. to pair-wise generalization	Standard Deviation for relevancy improvement
3-4 word phrases	1 compound sentence	81.5	82.1	86.7	87.8	87.1	91.4	1.054	0.0091
	2 sentences	79.0	79.7	92.7	86.3	85.1	89.7	0.968	0.0087
	3 sentences	76.2	75.2	78.9	85.3	86.2	88.8	1.125	0.0090
	Average	78.9	79.0	86.1	86.5	86.1	90.0	1.045	
5-10 word phrases	1 compound sentence	78.4	77.5	83.4	87.5	85.0	88.2	1.058	0.0092
	2 sentences	75.9	75.5	80.0	82.6	83.1	88.0	1.100	0.0095
	3 sentences	74.2	74.7	76.9	81.4	80.7	82.6	1.074	0.0082
	Average	76.2	75.9	80.1	83.8	82.9	86.3	1.077	
1 sentence	1 comp. sentence	77.2	76.8	81.3	85.7	86.1	88.6	1.090	0.0079
	2 sentences	73.9	73.7	78.3	82.5	83.0	86.1	1.100	0.0087
	3 sentences	71.3	72.0	76.1	80.9	81.4	83.4	1.096	0.0081
	Average	74.1	74.2	78.6	83.0	83.5	86.0	1.095	
2 sentences	1 comp. sentence	75.8	76.1	82.0	87.1	83.3	83.4	1.017	0.0083
	2 sentences	73.2	71.1	76.5	82.4	81.7	82.0	1.072	0.0086
	3 sentences	69.7	72.1	75.0	79.8	79.7	83.5	1.113	0.0084
	Average	72.9	73.1	77.8	83.1	81.6	83.0	1.066	
3 sentences	1 comp. sentence	73.9	74.1	78.8	85.6	83.2	85.6	1.086	0.0080
	2 sentences	73.9	71.7	76.2	84.4	83.1	84.3	1.106	0.0075
	3 sentences	67.3	69.0	74.8	79.7	81.1	84.3	1.127	0.0071
	Average	71.7	71.6	76.6	83.2	82.5	84.7	1.106	0.0091
Average for all Query and Answer type								1.079	

Conclusions

- Extendable text model
- Effective linguistic projections
- Native clustering technique
- Powerful lattice representation
  - Filtering
  - Indexing
  - Searching
  - Pruning

References

1. Galitsky, B., G. Dobrocsi, J.L. de la Rosa, Kuznetsov, S.O.: From Generalization of Syntactic Parse Trees to Conceptual Graphs, in M. Croitoru, S. Ferré, D. Lukose (Eds.): Conceptual Structures: From Information to Intelligence, 18th International Conference on Conceptual Structures, ICCS 2010, Lecture Notes in Artificial Intelligence, vol. 6208, pp. 185-190.(2010)
2. Galitsky, B., Gabor Dobrocsi, Josep Lluís de la Rosa, Sergei O. Kuznetsov: Using Generalization of Syntactic Parse Trees for Taxonomy Capture on the Web. 19th International Conference on Conceptual Structures, ICCS 2011: 104-117 (2011).
3. Boris Galitsky, Sergei O. Kuznetsov, Daniel Usikov, Parse Thicket Representation for Multi-sentence Search. In: Pfeiffer, H.D.; Ignatov, D.; Poelmans, J.; Nagarjuna, G., Eds., Proc. 20th International Conference on Conceptual Structures(ICCS 2013), Lecture Notes in Artificial Intelligence (Springer), Vol. 7735, pp. 153-172, 2013.
4. Ganter, B, Kuznetsov SO Pattern Structures and Their Projections. In: Conceptual Structures: Broadening the Base. Lecture Notes in Computer Science Volume 2120, 2001, pp 129-142.
5. B. Ganter, P.A. Grigoriev, S.O. Kuznetsov, and M.V. Samokhin, Concept-based Data Mining with Scaled Labeled Graphs. In: K.E. Wolff, H. D. Pfeiffer, H. S. Delugach, Eds., Proc. 12th International Conference on Conceptual Structures (ICCS 2004), Lecture Notes in Artificial Intelligence (Springer), Vol. 3127, pp. 94-108, 2004.