News clustering approach based on discourse text structure

Tatyana Makhalova, Dmitry Ilvovsky, Boris Galitsky

National Research University Higher School of Economics, Moscow, Russia
Knowledge Trail Incorporated, San Jose, USA
t.makhalova@gmail.com, dilvovsky@hse.ru, bgalitsky@hotmail.com

Table of contents

Main Clustering Aspects

- Text preprocessing and representation
- Clustering methods
- Similarity measures

Text Representation Models

Model	**Authors**	**Data structure**	**Words order preserving**	**Embedded semantics**
VSM	Salton et al, 1975	matrix	-	-
GVSM	Wong et al,1985	matrix	-	+
TVSM	Becker and Kuropka, 2003	matrix	-	+
eTVSM	Polyvyanyy and Kuropka, 2007	matrix	-	+
DIG	Hammouda and Kamel, 2004	graph	+	-
"Suffix Tree"	Zamir and Etzioni, 1998	tree	+	-
N-Grams	Schenker et al, 2007	graph	+	-
Parse Thickets	Galitsky, 2013	trees (forest)	+	+

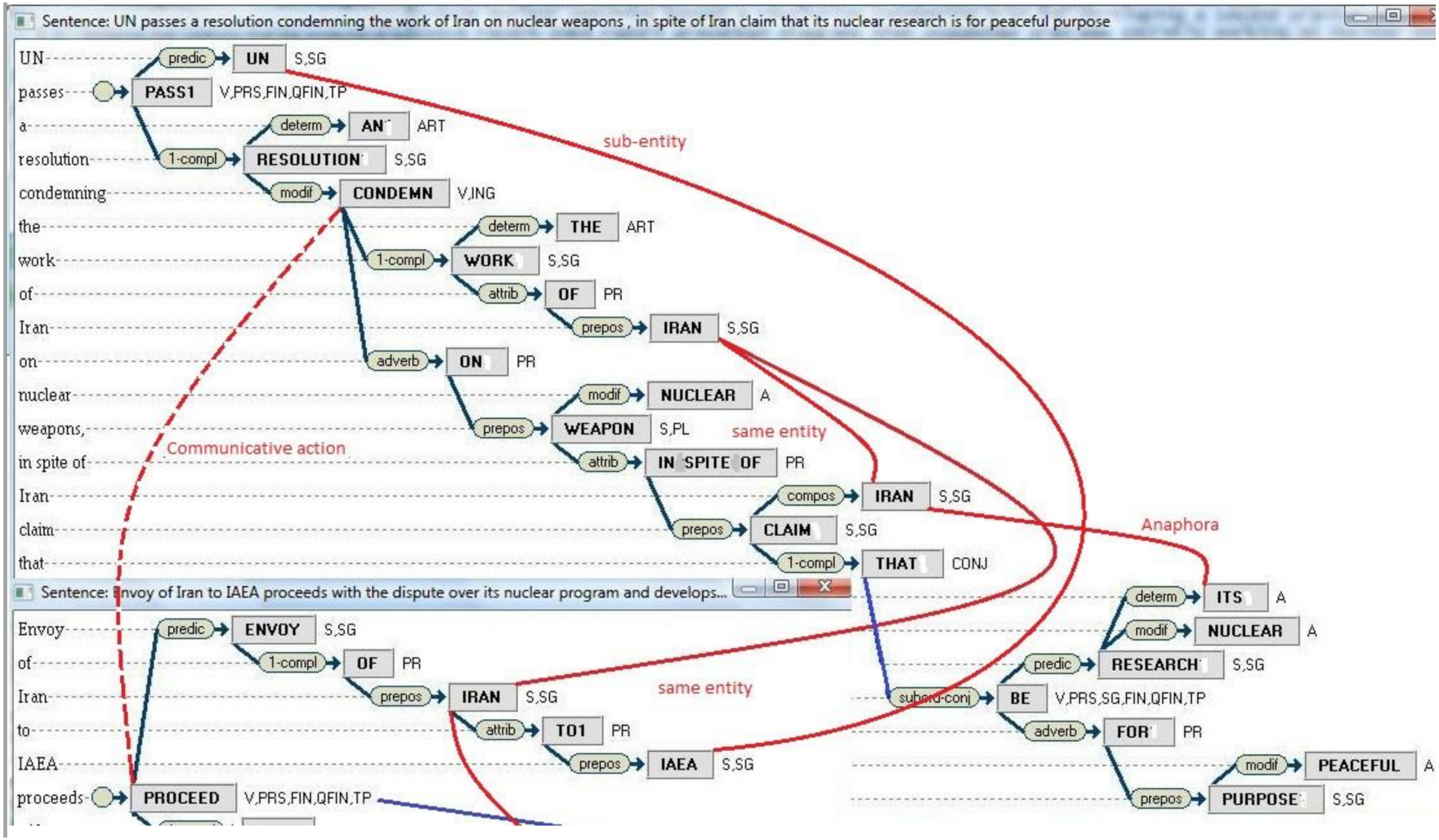
Parse Thickets: basic characteristics

- Preserving a linguistic structure of a text paragraph
- Constructing of parse trees for each sentence within a paragraph
- Adding inter-sentence relations between parse tree nodes

Parse Thickets: types of discourse relations

- Coreferences (Lee et al., 2012)
  - Anaphora
  - Same entity
  - Hyponym/hyperonym
- Rhetoric structure theory (RST) (Mann et al., 1992)
- Communicative Actions (Searle, 1969)

Coreferences: example



Relations based on Rhetoric Structure Theory

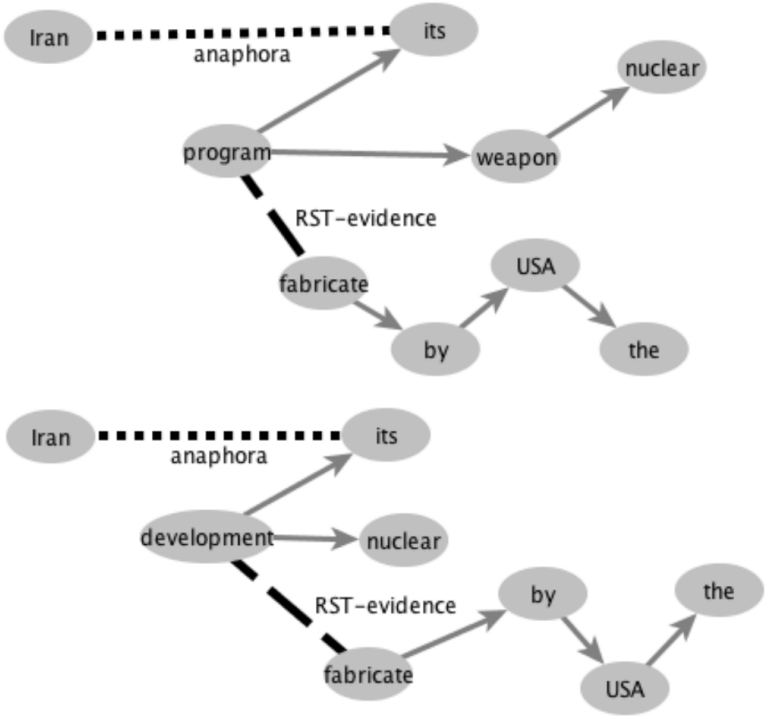
- RST characterizes structure of text in terms of relations that hold between parts of text
- RST describes relations between clauses in text which might not be syntactically linked
- RST helps to discover text patterns such as nucleus/satellite structure with relation such as *evidence*, *justify*, *antithesis*, *concession* and so on.

Parse Thickets: an example

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

"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",
"Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops an enrichment site in secret",
"Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium enrichment site"

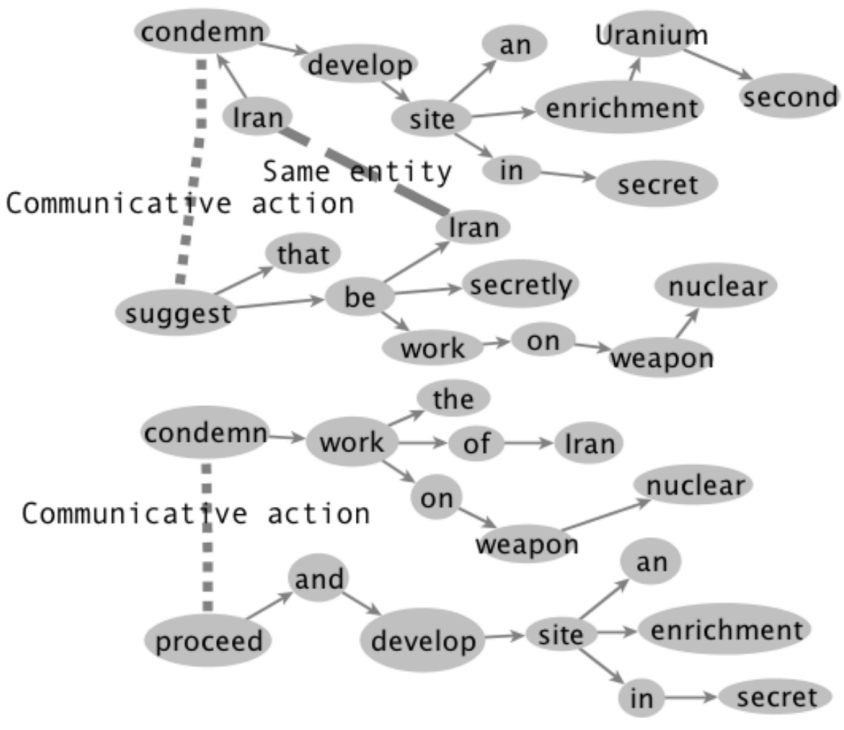
Parse Thickets: discourse relations



"Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium enrichment site"

"Iran envoy says its nuclear development is for peaceful purpose, and the material evidence against it has been fabricated by the US"

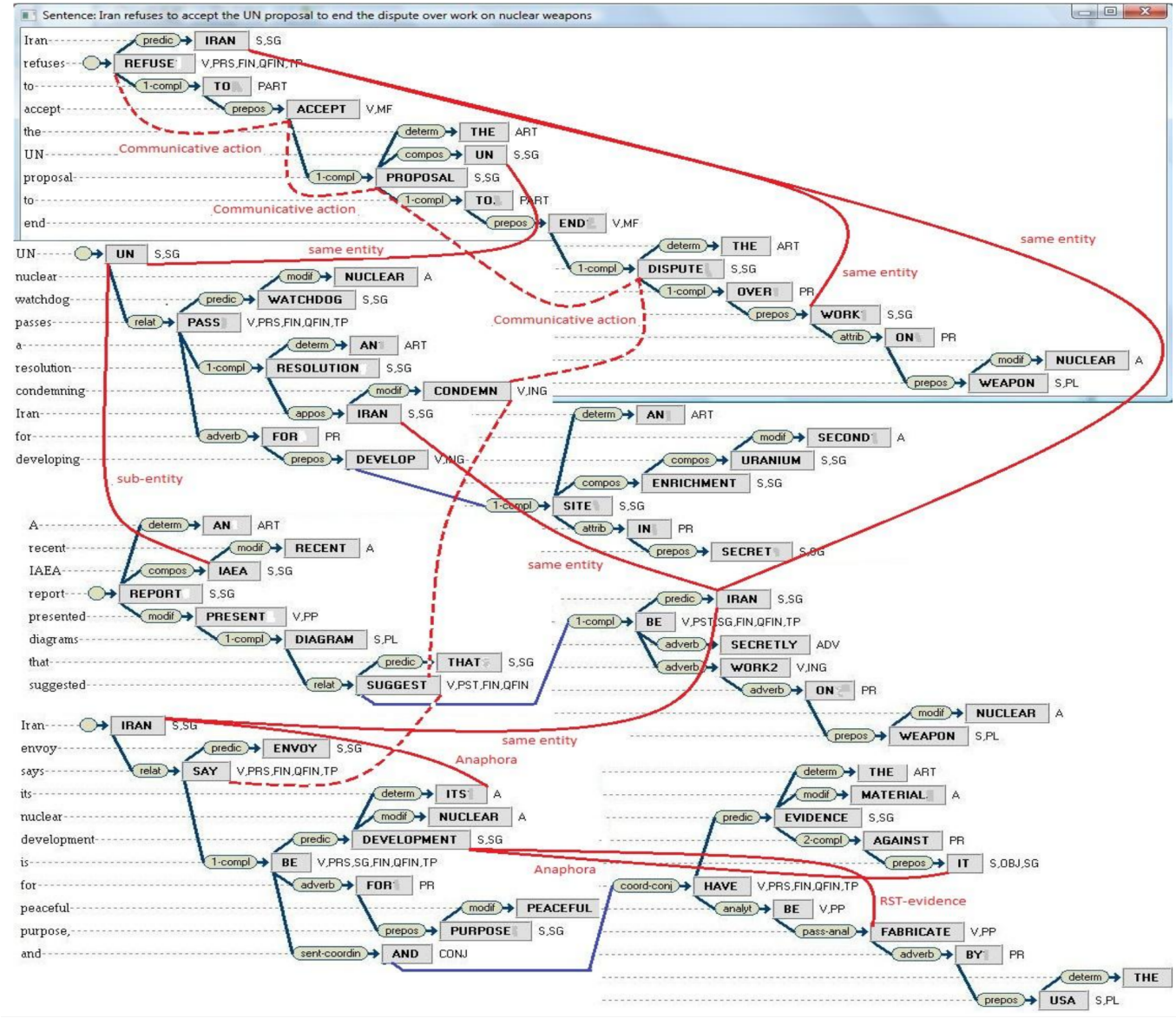
Parse Thickets: discourse relations



"UN nuclear watchdog passes a resolution condemning Iran for developing a second Uranium enrichment site in secret",
"A recent IAEA report presented diagrams that suggested Iran was secretly working on nuclear weapons",

"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",
"Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops an enrichment site in secret"

Parse Thickets: an example



Clustering of Parse Thickets: the main idea

Similarity of parse thickets based on sub-trees matching

- labeled discourse arcs
- unlabeled syntactic arcs
- nodes with part of speech and stem of a word

Clustering of paragraphs: generalisation of syntactic trees

[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 NNsecret],
[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]

Clustering of paragraphs: generalisation of parse thickets

[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-**generalization** <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]
[NN-Iran JJ-nuclear NN-weapon NN-* RST-evidence VBN-fabricated IN-by DT-the NNP-US **condemnproceed** [enrichment site] <leads to> **suggestcondemn** [work Iran nuclear weapon]

Clustering of Parse Thickets: what do we want?

- Adequately represent groups of texts with overlapping content
- Get text clusters with different refinement

Goal: (multi-level) hierarchical structure
Solution: Construction of pattern structures on parse thickets

Clustering of Parse Thickets: the mathematical foundation

Pattern Structure

A triple $(G, (D, \sqcap), \delta)$, where G is a set of objects, $(D, \sqcap)$ is a complete meet-semilattice of descriptions and $\delta : G \to D$ is a mapping an object to a description.

Pattern concept

A pair (A, d) for which $A^{\square} = d$ and $d^{\square} = A$, where $A^{\square}$ and $d^{\square}$ are the Galois connections, defined as follows:

$$A^{\square} := \sqcap_{g \in A} \delta(g) \text{ for } A \subseteq G$$

$$d^{\square} := \{g \in G | d \sqsubseteq \delta(g)\} \text{ for } d \in D$$

Pattern Structures on Parse Thickets

an original paragraph of text	$\rightarrow$	an object $a \in A$
parse thickets constructed from paragraphs	$\rightarrow$	a set of its maximal generalized sub-trees d
a pattern concept	$\rightarrow$	a cluster

Drawback: the exponential growth of the number of clusters by increasing the number of texts (objects)

Reduced pattern structures: meaningfulness estimates of a pattern concept

Average and Maximal Pattern Score

Maximum score among all sub-trees in the cluster

$$Score^{max}\,\langle A, d\rangle := \max_{chunk \in d} Score\,(chunk)$$

Average score of sub-trees in the cluster

$$Score^{avg}\,\langle A, d\rangle := \frac{1}{|d|} \sum_{chunk \in d} Score\,(chunk)$$

where $Score\,(chunk) = \sum_{node \in chunk} w_{node}$

Reduced pattern structures: loss estimates of a cluster with respect to original texts

Average and Minimal Pattern Loss Score

Estimates minimal lost meaning of cluster content w.r.t. original texts in the cluster

$$ScoreLoss^{min}\,\langle A, d\rangle := 1 - \frac{Score^{max}\,\langle A, d\rangle}{\min_{g\in A} Score^{max}\,\langle g, d_g\rangle}$$

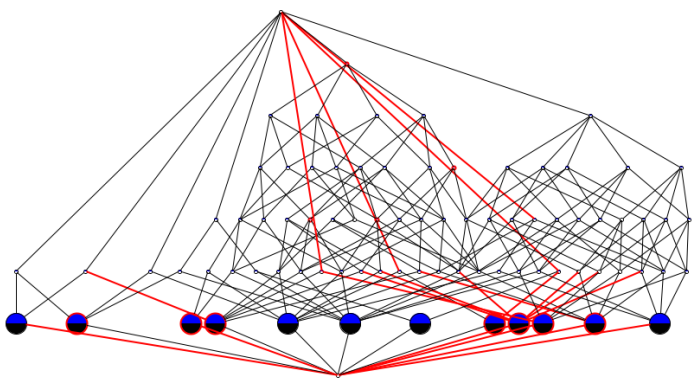
Estimates lost meaning of cluster content on average

$$ScoreLoss^{avg}\,\langle A, d\rangle := 1 - \frac{Score^{avg}\,\langle A, d\rangle}{\frac{1}{|d|}\sum_{g\in A} Score^{max}\,\langle g, d_g\rangle}$$

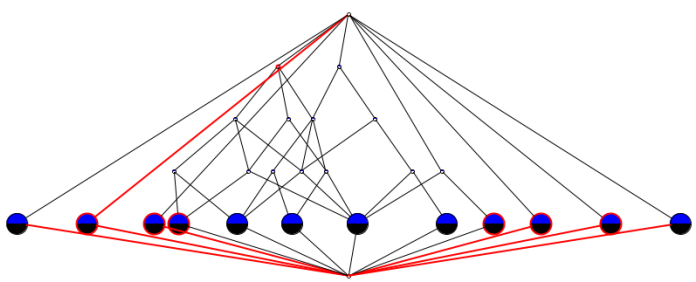
Reduced pattern structures: generalization

Controlling the loss of meaning w.r.t. the original texts

$$ScoreLoss^* \langle A_1 \cup A_2 \ , d_1 \cap d_2 \rangle \leq \theta$$



$\theta = 0,75,\ \mu_1 = 0,1,\ \mu_2 = 0,9$



$\theta = 0,5,\ \mu_1 = 0,1,\ \mu_2 = 0,9$

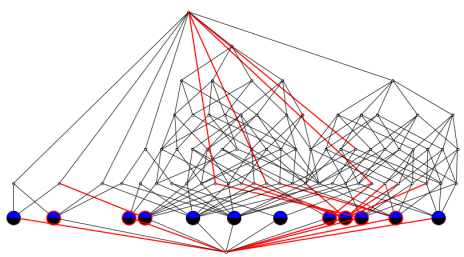
Reduced pattern structures: clusters distinguishability

Controlling the loss of meaning w.r.t. the nearest more meaningfulness neighbors in the cluster hierarchy

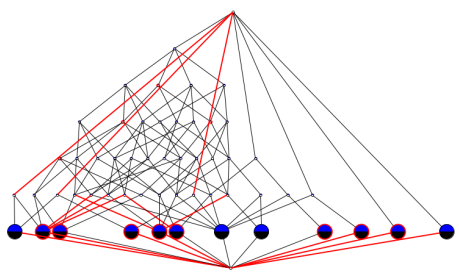
$$Score^* \langle A_1 \cup A_2 \, , d_1 \cap d_2 \rangle \geq \mu_1 min \{Score^* \langle A_1 \, , d_1 \rangle, Score^* \langle A_2 \, , d_2 \rangle\}$$

Controlling the distinguishability w.r.t. the nearest neighbors in the hierarchy of clusters

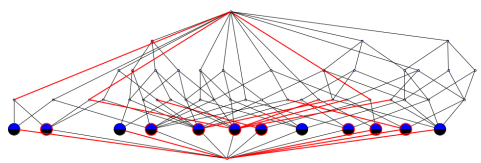
$$Score^* \langle A_1 \cup A_2 \, , d_1 \cap d_2 \rangle \leq \mu_2 max \{Score^* \langle A_1 \, , d_1 \rangle, Score^* \langle A_2 \, , d_2 \rangle\}$$



$\mu_1 = 0,1$, $\mu_2 = 0,9$, $\theta = 0,75$



$\mu_1 = 0,5$, $\mu_2 = 0,9$, $\theta = 0,75$



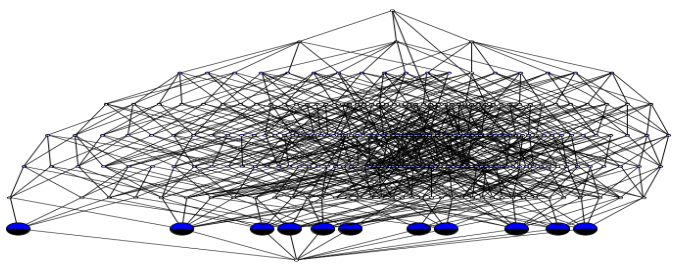
$\mu_1 = 0,1$, $\mu_2 = 0,8$, $\theta = 0,75$

Reduced pattern structures: constraints

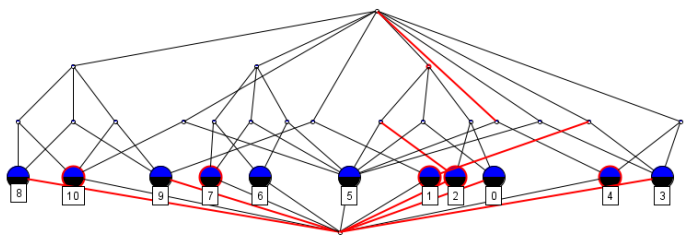
$$ScoreLoss^* \langle A_1 \cup A_2 \, , d_1 \cap d_2 \rangle \leq \theta$$

$$Score^* \langle A_1 \cup A_2 \, , d_1 \cap d_2 \rangle \geq \mu_1 min \{Score^* \langle A_1 \, , d_1 \rangle, Score^* \langle A_2 \, , d_2 \rangle\}$$

$$Score^* \langle A_1 \cup A_2 \, , d_1 \cap d_2 \rangle \leq \mu_2 max \{Score^* \langle A_1 \, , d_1 \rangle, Score^* \langle A_2 \, , d_2 \rangle\}$$



pattern structure
without reduction



reduced pattern structure
with $\theta = 0,75$, $\mu_1 = 0,1$ and $\mu_2 = 0,9$

Implementation

- The Apache OpenNLP library (the most common NLP tasks)
- Bing search API (to obtain news snippets)
- Pattern structure builder: modified by authors version of AddIntent algorithm (van der Merwe et al., 2004)

News Clustering: motivation

- A long list of search results
- Many groups of pages with a similar content
- An overlapping content

User Study: non-overlapping partition

- web snippets on world's most pressing news: "F1 winners", "fighting Ebola with nanoparticles", "2015 ACM awards winners", "read facial expressions through webcam", "turning brown eyes blue"
- inconsistency of human-labeled partitions: low values of a pairwise Adjusted Mutual Information score of human-labeled partitions $0,03 \leq MI_{adj} \leq 0,51$

Example: The Ebola News Set

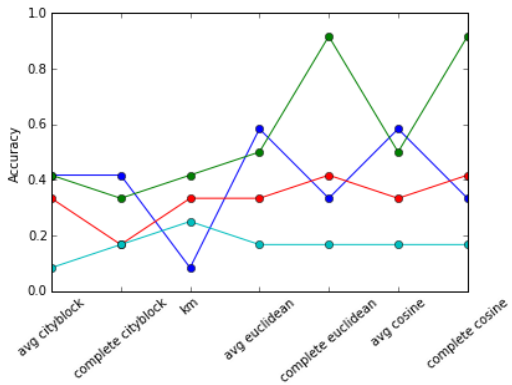
Text ID	# words	# symbols	# sentences	quoted speech	reported speech
1	42	210	3		
2	42	253	3	+	
3	54	287	3	+	
4	75	399	3	+	+
5	31	167	2	+	
6	44	209	2	+	+
7	49	247	2		+
8	61	340	3		+
9	50	242	2	+	
10	62	295	4		+
11	90	526	4	+	+
12	75	370	4		

Accuracy of non-overlapping clustering methods

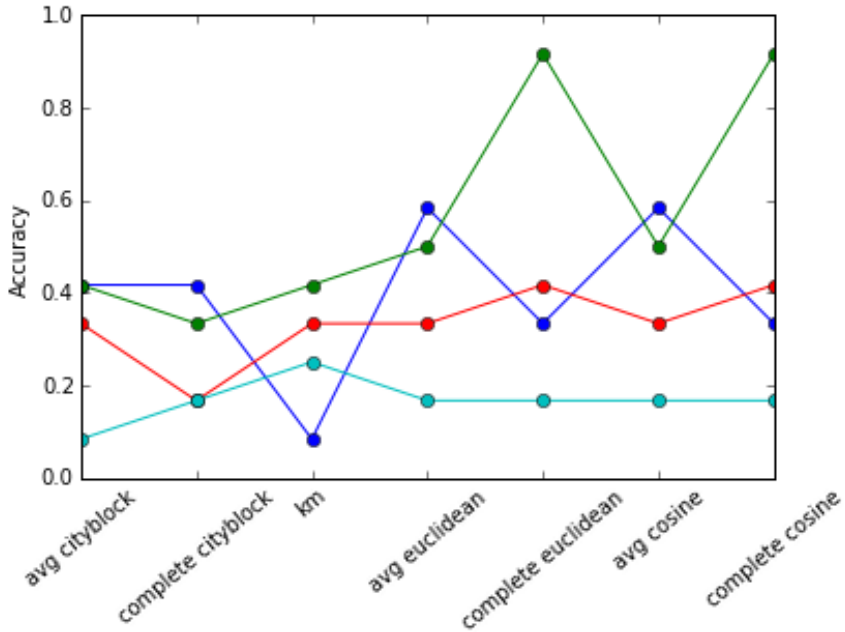
Accuracy of conventional clustering methods in the case of overlapping texts groups

- low (in most cases)
- greatly depends on taken as ground truth a human-labeled partition

Method	Linkage	Distance	A human-labeled partition 1	2	3	4
HAC	average	cityblock	0,42	0,42	0,33	0,08
	complete	cityblock	0,42	0,33	0,17	0,17
	average	euclidean cosine	0,58	0,5	0,33	0,17
	complete	euclidean cosine	0,33	0,92	0,42	0,17
k-means		euclidean	0,08	0,08	0,17	0,25

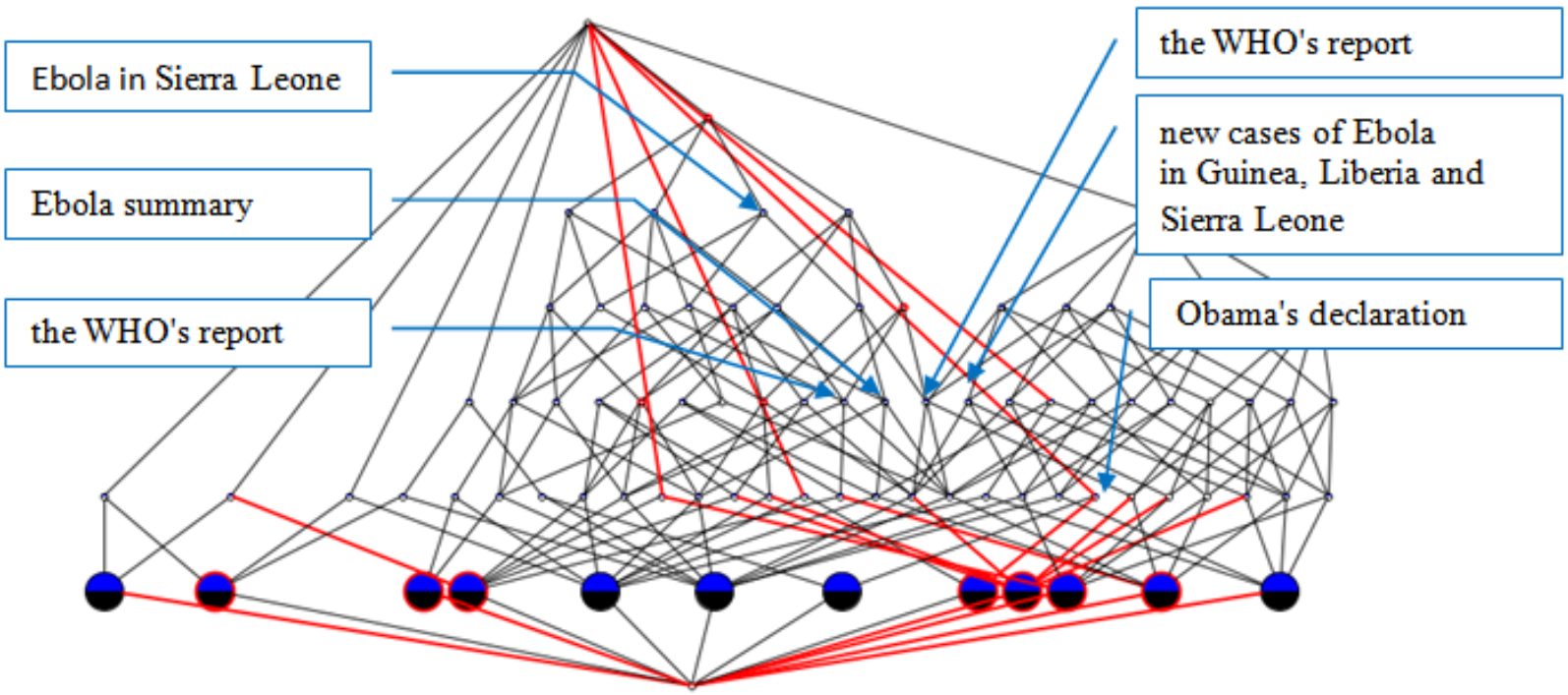


Accuracy of non-overlapping clustering methods



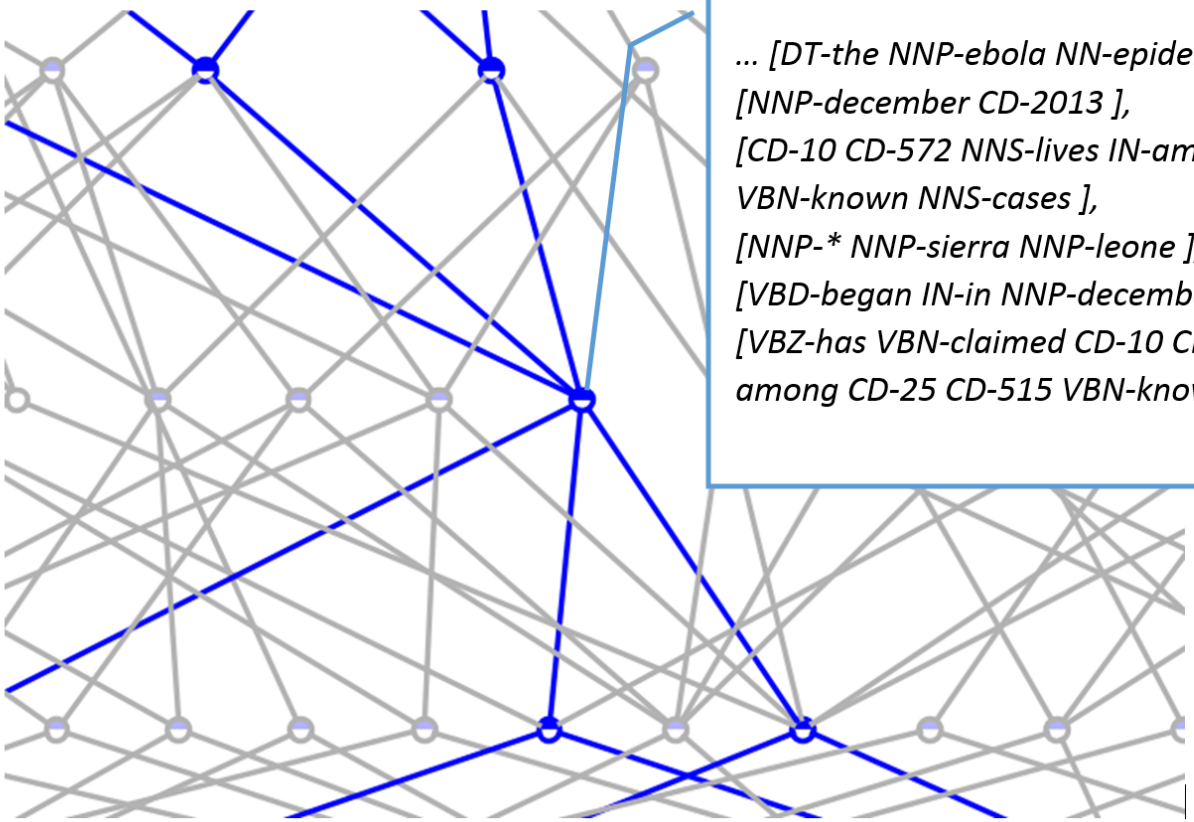
Accuracy of conventional clustering methods
for 4 human-labeled partitions

An example of pattern structures clustering: clusters with maximal score



reduced pattern structure
with $\theta = 0,75$, $\mu_1 = 0,1$ and $\mu_2 = 0,9$

An example of pattern structures clustering: clusters with maximal score



Ebola summary

$\text{Score}^{\max}(\{3, 11, 12\}) = 7{,}8$

… [DT-the NNP-ebola NN-epidemic],
[NNP-december CD-2013],
[CD-10 CD-572 NNS-lives IN-among CD-25 CD-515 VBN-known NNS-cases],
[NNP- NNP-sierra NNP-leone],*
[VBD-began IN-in NNP-december CD-2013],
[VBZ-has VBN-claimed CD-10 CD-572 NNS-lives IN-among CD-25 CD-515 VBN-known NNS-cases]…

Conclusion

- Short text clustering problem
- A failure of the traditional clustering methods
- Parse Thickets as a text model
- Texts similarity based on pattern structures
- Reduced pattern structures with constraints
- *Score* and *ScoreLoss* to improve efficiency and to remove redundant clusters
- Improvement of browsing and navigation through texts set for users