Parse Thicket representations for paragraphs of text

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

[1] eBay Inc San Jose CA USA

[2] Higher School of Economics, Moscow Russia

{bgalitsky@hotmail.com

; dilv_ru@yahoo.com

; fdr.strok@gmail.com

; skuznetsov@hse.ru

}

Goals

1. Parsing is reliable now. What can we systematically do with parse trees in applications such as search, classification, recommendation?
2. How can search engineers take advantage of linguistic processing (for text similarity assessment) and machine learning in a plug&play mode?
3. Parse trees are for sentences. How can we apply machine learning to paragraphs of text?
4. No keywords please.

Outline

1. Problem and motivations
2. Search Application
3. Introducing Parse Thicket
4. From simple relations between words in sentences to Rhetoric Structures
5. Generalization operation
6. Evaluation

Why Parse Thickets

- To represent a linguistic structure of a *paragraph of text* based on parse trees for each sentence of this paragraph.
- We will refer to the sequence of parse trees extended by a number of arcs for inter-sentence relations between nodes for words as Parse Thicket (PT).
- A PT is a graph which includes parse trees for each sentence, as well as additional arcs for inter-sentence relationship between parse tree nodes for words.

What are we going to do with Parse Thickets

- Extend the operation of least general generalization (unification of logic formula) towards structural representations of paragraph of texts
- Define the operation of *generalization of text paragraphs* to assess similarity between portions of text.
- Use of generalization for similarity assessment is inspired by structured approaches to machine learning versus unstructured, statistical where similarity is measured by a distance in feature space (Moschitti et al)

Applications

- Text analysis application such as search, classification, categorization, and content generation
- Machine Learning of parse trees is used at eBay and other content portals for search and content processing, where text similarity assessment is required

UI for entertainment product search

what Lady Gaga, who wants her g when Jun 20 2012 near Sydney Olympic Park, NSW

Address, Neighborhood, City & State, or ZIP

events movies venues restaurants performers

Sydney Olympic Park, NSW [change my location]

Home > Sydney Olympic Park Events > Sydney Olympic Park Rock Events > Lady Gaga

Tweet Share Like

Lady Gaga

Wednesday, Jun 20, 2012 8:00p

More dates & times (3)

Acer Arena
Sydney Olympic Park, NSW

Buy Tickets

Lady Gaga at Acer Arena.

Call for details and mention you found this information on Zvents.

Price $394.35
Age Suitability None Specified
Category Rock

If you haven't been under a rock the past couple of weeks, you'll likely have heard that Lady Gaga's in town for a couple of concert dates at the brand spankin' new SM Mall of Asia (MOA) Arena. And you'll likely have heard that certain religious groups are up in arms and would dearly love to see her gig(s) cancelled.

These aforementioned groups cite her "immorality" and "lewd behavior and attire," branding her as some sort of demonic pied piper who might lead the youth astray with songs like the controversial "Judas."

While we understand how these groups may feel that the aforementioned song is disrespectful

Searching for events relevant to Facebook posting *'I want to go to Lady Gaga 'Born This Way', and who wants her gigs stopped?'*

Analysis of coreferences is required to index a document for search properly

- Text for indexing1: … Tuberculosis is a lung disease. It is cured by doctors specializing in pulmonology.
- Question: Which specialist doctor should treat my tuberculosis?
- Text for indexing2: … Tuberculosis is a lung disease… Pulmonology specialist Jones was awarded a prize for curing a special form of disease.
- Question: Which specialist doctor should treat my tuberculosis?

In the first case, establishing coreference link *Tuberculosis → disease → is cured by doctors pulmonologists* helps to match these entities with the ones from the question.

In the second case this portion of text does not serve as a relevant answer to the question, although it includes keywords from this question.

Hence at indexing time, keywords should be chained not just by their occurrence in individual sentences, but additionally on the basis of coreferences.

Modern search engines do not do well when a query needs to be matched across multiple sentences

Which specialist doctor should treat my tuberculosis?

About 19,600,000 results (0.51 seconds)

Tuberculosis | Overview -- **FamilyDoctor**.org
family**doctor**.org/family**doctor**/en/diseases.../**tuberculosis**.html
Tuberculosis (say: too-burr-cue-low-sis), also called **TB**, is an infection caused by a bacteria (germ). **Tuberculosis** usually affects the lungs, but it **can** spread to ...

[PDF] **Tuberculosis** Getting Healthy, Staying Healthy
www.niaid.nih.gov/topics/**tuberculosis**/understanding/.../**tb**.pdf
File Format: PDF/Adobe Acrobat - Quick View
What is **TB** infection? 3. **Can** I give **TB** infection to someone else? 4. How **can my doctor** tell if I have **TB** infection? 6. How does **my doctor treat TB** infection? 9 ...

Tuberculose - **Treatment** and advices
www.msss.gouv.qc.ca/sujets/prob_sante/tuberculose/index.php?...
Jump to What happens if I stop taking one or several of **my** medications ...: It **would** then be very difficult to find medication strong enough to cure you. Also, **tuberculosis** is a disease that ... without authorization from your **doctor**.

Uterus **Tuberculosis** (**TB**), **TB** And Pregnancy, **Tuberculosis** And ...
www.parentingnation.in/.../the-uterus-**tuberculosis**-(tb)-during-pregn...
Complete course of **TB treatment should** taken otherwise **TB** bacteria may .... Now **my** 6 months **treatment** has been passed and **my doctor** suggest that i **can** ... me what

- No answer includes ‘pulmonologist’

What happens if we match keywords not phrases

Yandex α
has found
6 million answers

how does us citizen pay tax on income abroad

search within results

Global results Russian results

US Citizens Living **Abroad**, Expatriates, IRS **Income Tax** Explained
However, being temporarily present in the **United States**, or maintaining a dwelling there, **does** not necessarily mean ... Generally, you must **pay** U.S. **tax on** the **income** earned **abroad** in the same way you **pay** the **tax on income** earned in the **United States**.
taxmeless.com › IRS593Publication.htm cached more

Is Foreign **Income** Earned by U.S. **Citizens** Living **Abroad** Subject to Social
The U.S. requires **citizens** to **pay taxes on** all **income** earned **abroad**. **tax** forms image by Chad McDermott from Fotolia.com. ... **How Does** Living in Mexico Effect **Us** Citizenship?
e**how**.com › info_7747879_foreign-subject...**tax**.html cached more

Tax Guide for U.S. **Citizens** and Resident Aliens **Abroad**
View Download 1.5 MB
Your employer **does** not have to withhold U.S. **income taxes** from wages you earn **abroad** if it is reasonable to believe that you will exclude them from **income** under ... **How do** I **pay** my self-employment **tax**? File a Form 1040 with Schedule SE and Form 2555.
irs.gov › pub/irs-pdf/p54.pdf more

U.S. **Citizens** Living **Abroad Taxes**: Frequently Asked Questions
United States citizens living **abroad** are required to file annual U.S. **income tax** returns and report their worldwide **income** if ... **How** can I get a nonworking social security number for her? Each foreign person who **does** not have and cannot obtain a...
groco.com › ...irs_**citizens**living**abroad**.aspx cached more

If we specify “string phrases” a search engine is doing a little better

Yandex α
has found
567 thousand answers

how does us citizen pay tax on "income abroad"

search within results

Global results Russian results

How to Avoid **Income Taxes** | eHow.com
Find a country or locale that levies significantly lower **income tax on** its **citizens** than the **United States**. Some such places are the British Virgin Islands, the Cayman Islands, Switzerland, Lichtenstein, Monaco, Andorra, Singapore and Hong Kong.
ehow.com › **how**_6731095_avoid-**income-taxes**.html cached more

Do I **pay tax on** UK rental **income abroad**? | This is Money
By This Is Money. Last updated at 4:15 PM on 5th January 2011. ... Put your question to our team of independent experts. Click the more button to email **us**.
thisismoney.co.uk › money...article...Do...**pay-tax-on**... more

How many **citizens pay taxes** in the **US**
How many **US citizens pay income taxes**? The "**tax** net" refers to the types of payment that are **taxed**, which included personal earnings (wages), capital gains, and business **income**. ... **Does** India have more practising Christians than the UK?
wiki.answers.com › Q/**How**_many_**citizens_pay_taxes**...US cached more

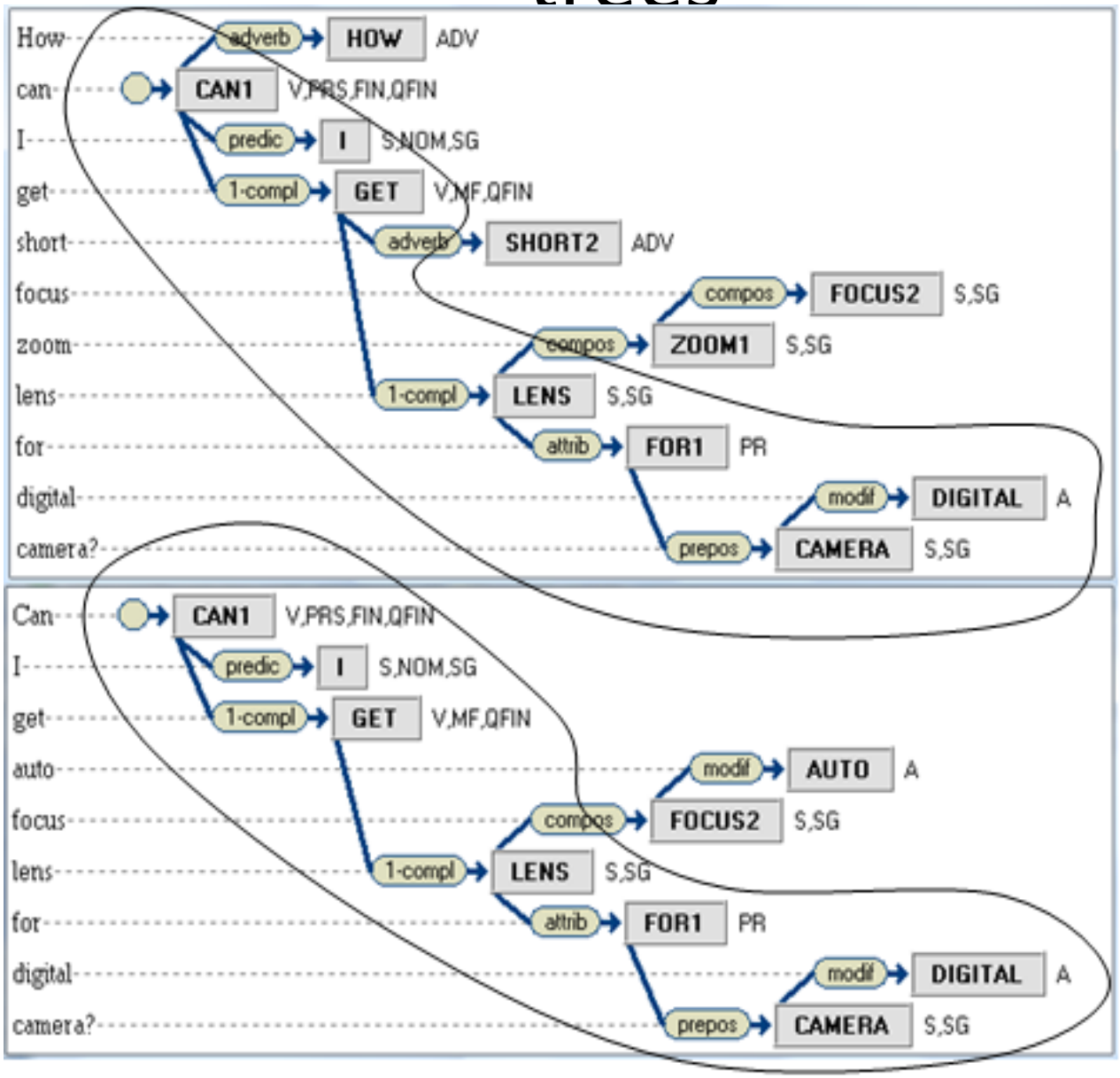
CitizenTax | **Citizen Tax** FAQ
What should I **do**? Contact Customer Support through the CONTACT **US** link located in the "**How** May I Contact You" section in the FAQs above or by selecting the CONTACT **US** link on the toolbar.
https://www.**citizentax**.com › node/4/

Now we compare linguistic phrase search and regular SOLR search

Linguistic phrase is not a string phrase query "…". There are noun phrases, verb phrases, prepositional phrases, and others

SOLR search takes into account term frequency, inverse document frequency, and the number of words between keywords in the document

Similarity = common part among parse trees



localhost:8983/solr/select/?q=name:lady+name:gaga+name:sings&fl=name

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">2</int>
  ▼<lst name="params">
     <str name="q">name:lady name:gaga name:sings</str>
     <str name="fl">name</str>
   </lst>
 </lst>
 ▼<result name="response" numFound="13" start="0">
  ▼<doc>
   ▼<str name="name">
      lady gaga does belly donce but katy perri sings romance
    </str>
   </doc>
  ▼<doc>
   ▼<str name="name">
      lady gaga does belly dance but katy perri sings romance
    </str>
   </doc>
  ▼<doc>
     <str name="name">lady perri sings and katy gaga does dance</str>
   </doc>
  ▼<doc>
     <str name="name">lady gaga and katy perri sings punk rock and dance</str>
   </doc>
  ▼<doc>
     <str name="name">rock lady sings and gaga perri does dance</str>
   </doc>
  ▼<doc>
   ▼<str name="name">
      lady gaga sings with passion and katy perri does an exotic dance
    </str>
   </doc>
  ▼<doc>
   ▼<str name="name">
      lady gaga sings rock but katy perri performs pop and dance
    </str>
   </doc>
```


Keyword search

Phrase (natural language) search

localhost:8983/solr/syntgen/?q=name:lady+name:gaga+name:sings&fl=name&qt=dismax

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<response>
 ▼<lst name="responseHeader">
   <int name="status">0</int>
   <int name="QTime">36</int>
  ▼<lst name="params">
    <str name="qt">dismax</str>
    <str name="q">name:lady name:gaga name:sings</str>
    <str name="fl">name</str>
   </lst>
 </lst>
 ▼<result name="response" numFound="10" start="0">
  ▼<doc>
   ▼<str name="name">
     lady gaga sings pop and katy perri performs punk rock and dance
    </str>
   </doc>
  ▼<doc>
   ▼<str name="name">
     lady gaga sings rock but katy perri performs pop and dance
    </str>
   </doc>
  ▼<doc>
   ▼<str name="name">
     lady gaga sings with passion and katy perri does an exotic dance
    </str>
   </doc>
  ▼<doc>
    <str name="name">lady gaga and katy perri sings punk rock and dance</str>
   </doc>
  ▼<doc>
    <str name="name">rock lady sings and gaga perri does dance</str>
   </doc>
  ▼<doc>
    <str name="name">lady perri sings and katy gaga does dance</str>
   </doc>
  ▼<doc>
   ▼<str name="name">
     lady gaga does dance, sings pop but katy perri performs funk rock
    </str>
```

Similarity between two paragraphs of text

- Baseline: bag-of-words approach, which computes the set of common keywords/n-grams and their frequencies.
- Pair-wise matching: we will apply syntactic generalization to each pair of sentences, and sum up the resultant commonalities. This technique has been developed in our previous work.
- Paragraph-paragraph match.

Finding similarity between two paragraphs

"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"

Keywords: topic with no details

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

Improvement: pair-wise generalization

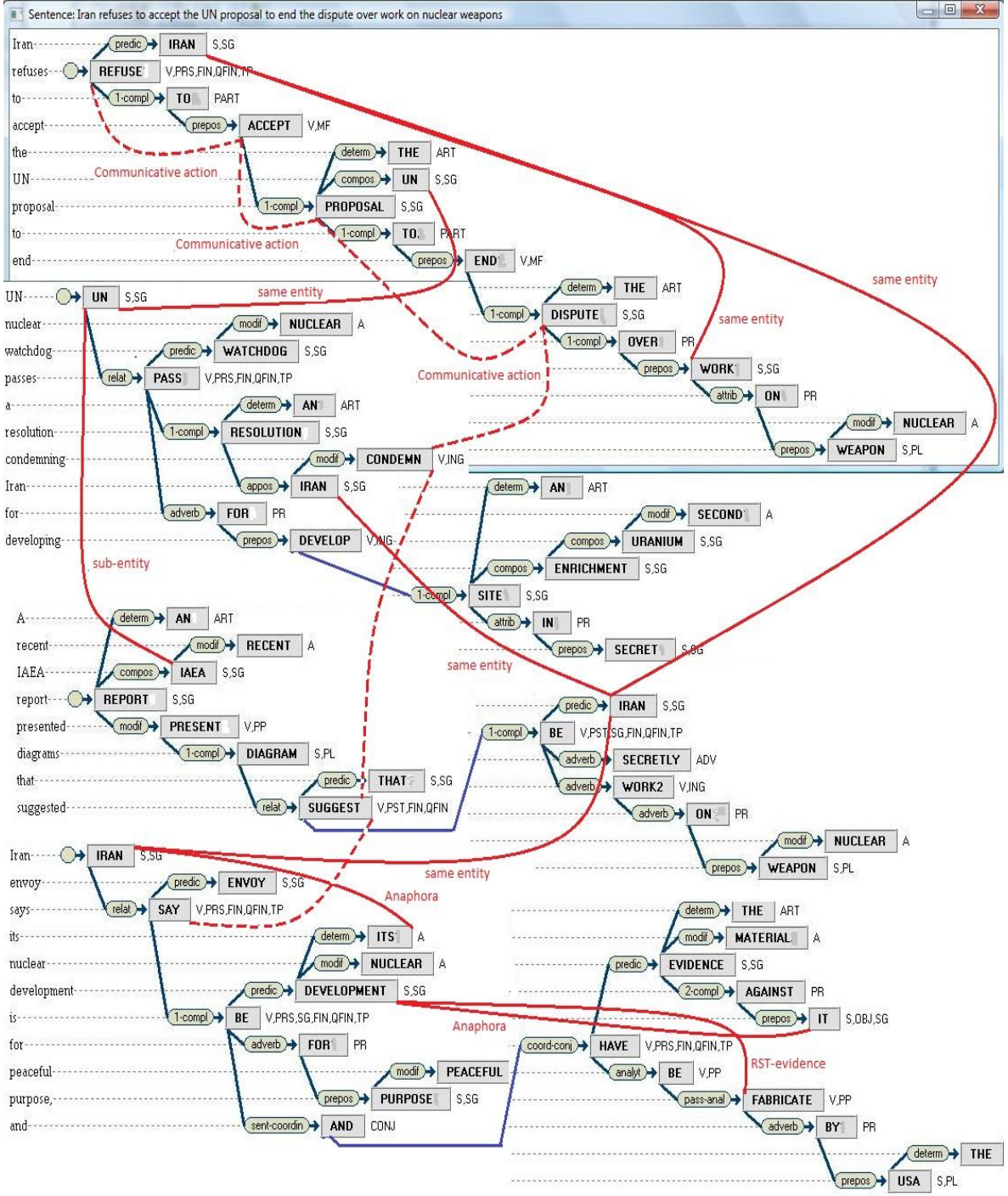
[NN-work IN-* IN-on JJ-nuclear NNS-weapons]**, [DT-the NN-dispute IN-over JJ-nuclear NNS-*]**, [VBZ-passes DT-a NN-resolution],
[VBG-condemning NNP-iran IN-*],
[VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]],
[DT-* JJ-second NN-uranium NN-enrichment NN-site]],
[VBZ-is IN-for JJ-peaceful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

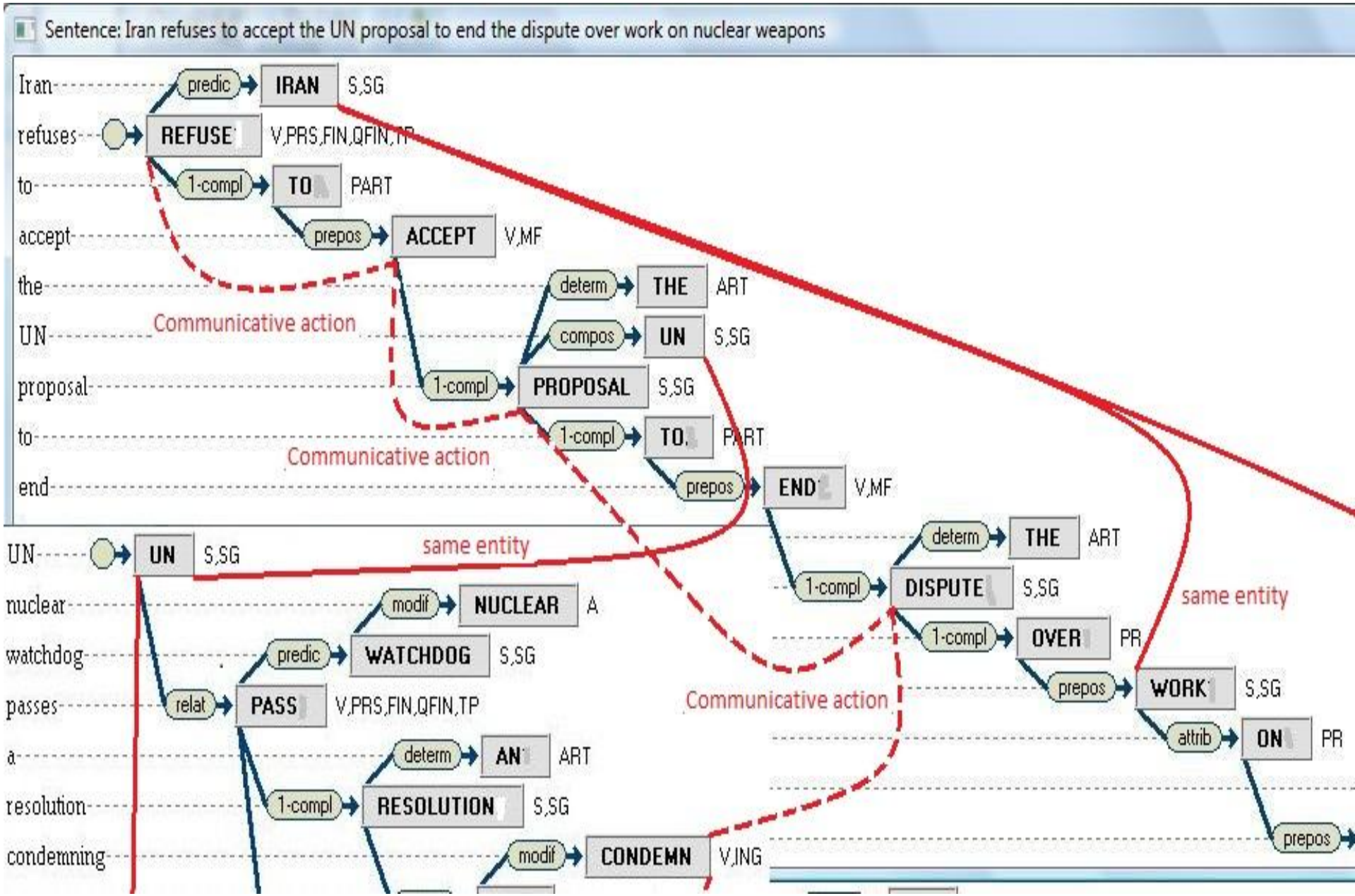
Pair-wise vs. Parse thicket

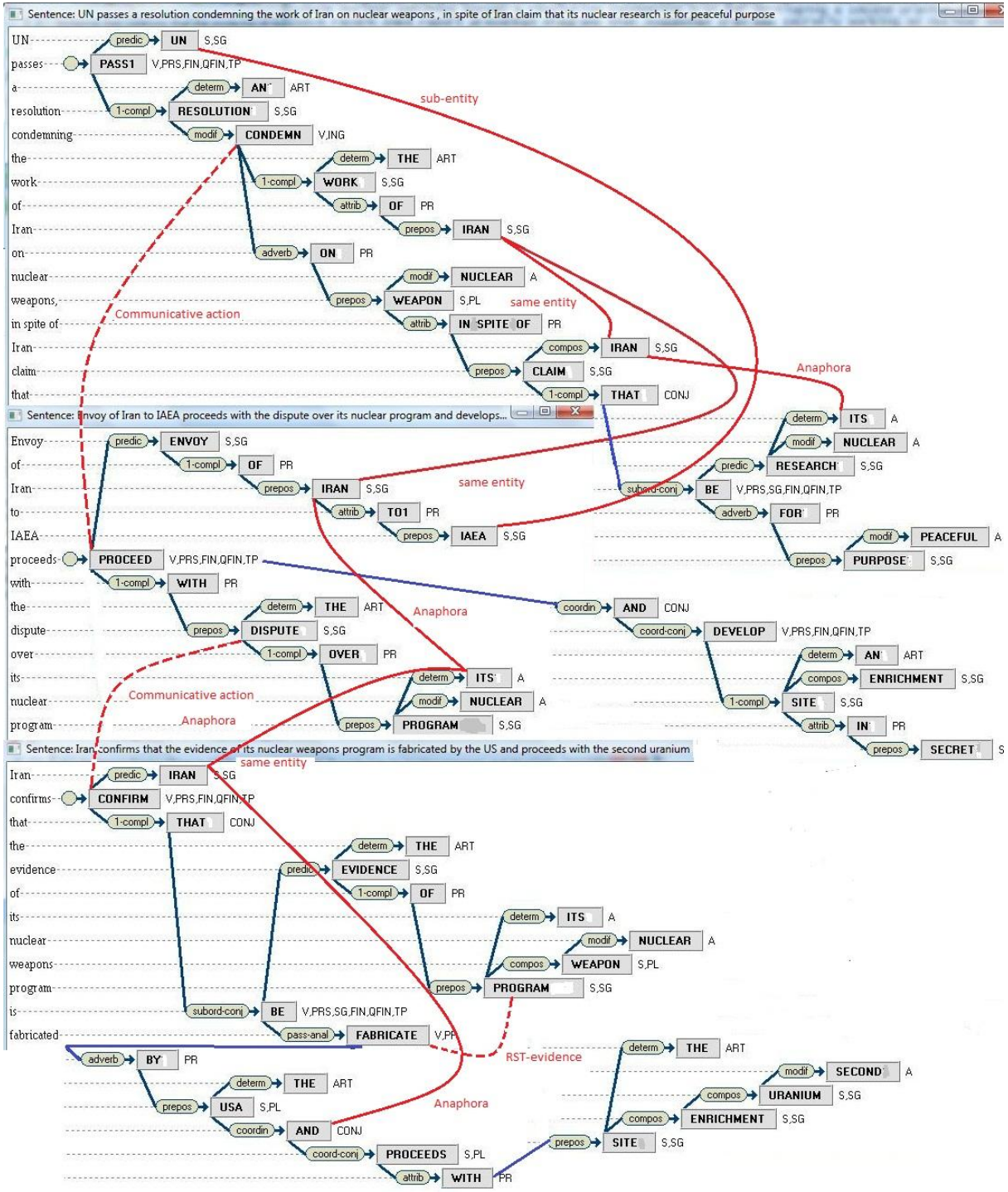
[NN-Iran VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]
[NN-generalization-<UN/nuclear watchdog> * VB-pass NN-resolution VBG condemning NN- Iran]
[NN-generalization-<Iran/envoy of Iran> **Communicative_action** DT-the NN-dispute IN-over JJ-nuclear NNS-*
[**Communicative_action** - NN-work 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*]

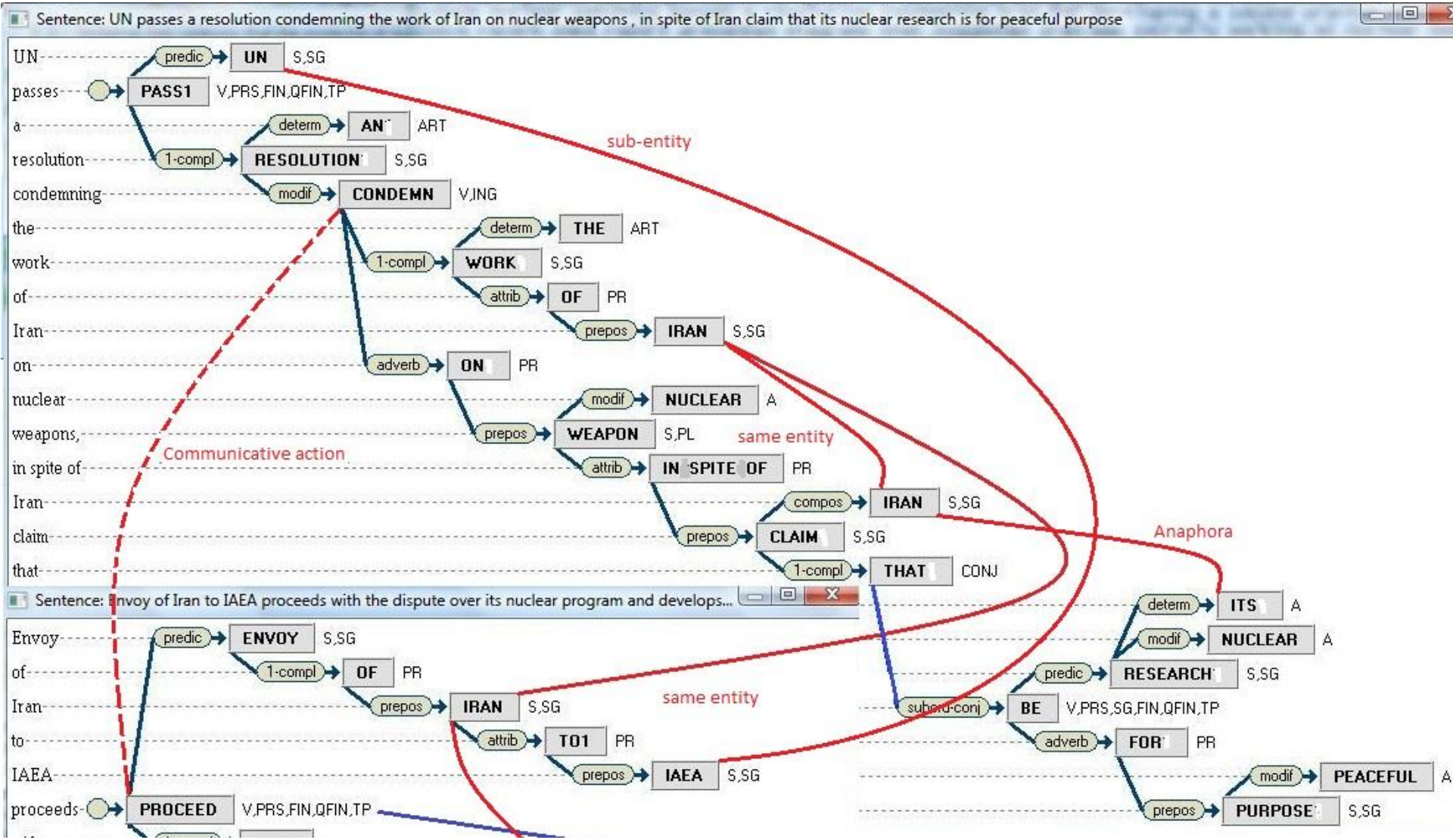
Parse thicket

- Syntactic parse trees
- Links between words from different sentences
  - Anaphora
  - Same entity
  - Hyponym/Hyperonym
  - Rhetoric Structure Theory (RST) [Mann]
  - Speech Act Theory (communicative actions, CA) [Searle]


Sentence: Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons
Iran
refuses
to
accept
the
UN
proposal
to
end
UN
nuclear
watchdog
passes
a
resolution
condemning
Iran
for
developing
A
recent
IAEA
report
presented
diagrams
that
suggested
Iran
envoy
says
its
nuclear
development
is
for
peaceful
purpose,
and
predic
IRAN S,SG
REFUSE V,PRS,FIN,QFIN,TP
1-compl
TO PART
prepos
ACCEPT V,MF
determ
THE ART
compos
UN S,SG
1-compl
PROPOSAL S,SG
1-compl
TO PART
prepos
END V,MF
determ
THE ART
1-compl
DISPUTE S,SG
1-compl
OVER PR
prepos
WORK S,SG
attrib
ON PR
modif
NUCLEAR A
prepos
WEAPON S,PL
Communicative action
Communicative action
same entity
same entity
same entity
UN S,SG
modif
NUCLEAR A
predic
WATCHDOG S,SG
relat
PASS V,PRS,FIN,QFIN,TP
determ
AN ART
1-compl
RESOLUTION S,SG
modif
CONDEMN V,ING
appos
IRAN S,SG
adverb
FOR PR
prepos
DEVELOP V,ING
Communicative action
sub-entity
determ
AN ART
modif
SECOND A
compos
URANIUM S,SG
compos
ENRICHMENT S,SG
1-compl
SITE S,SG
attrib
IN PR
prepos
SECRET S,SG
same entity
determ
AN ART
modif
RECENT A
compos
IAEA S,SG
REPORT S,SG
modif
PRESENT V,PP
1-compl
DIAGRAM S,PL
predic
THAT S,SG
relat
SUGGEST V,PST,FIN,QFIN
predic
IRAN S,SG
1-compl
BE V,PST,SG,FIN,QFIN,TP
adverb
SECRETLY ADV
adverb
WORK2 V,ING
adverb
ON PR
modif
NUCLEAR A
prepos
WEAPON S,PL
IRAN S,SG
same entity
predic
ENVOY S,SG
relat
SAY V,PRS,FIN,QFIN,TP
Anaphora
determ
ITS A
modif
NUCLEAR A
predic
DEVELOPMENT S,SG
1-compl
BE V,PRS,SG,FIN,QFIN,TP
adverb
FOR PR
modif
PEACEFUL
prepos
PURPOSE S,SG
sent-coordin
AND CONJ
Anaphora
determ
THE ART
modif
MATERIAL A
predic
EVIDENCE S,SG
2-compl
AGAINST PR
prepos
IT S,OBJ,SG
coord-conj
HAVE V,PRS,FIN,QFIN,TP
analyt
BE V,PP
pass-anal
FABRICATE V,PP
RST-evidence
adverb
BY PR
determ
THE
prepos
USA S,PL


Sentence: Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons
Iran
refuses
to
accept
the
UN
proposal
to
end
predic
IRAN S,SG
REFUSE V,PRS,FIN,QFIN,TP
1-compl
TO PART
prepos
ACCEPT V,MF
determ
THE ART
compos
UN S,SG
1-compl
PROPOSAL S,SG
1-compl
TO PART
prepos
END V,MF
determ
THE ART
1-compl
DISPUTE S,SG
1-compl
OVER PR
prepos
WORK S,SG
attrib
ON PR
prepos
Communicative action
Communicative action
same entity
same entity
Communicative action
UN
nuclear
watchdog
passes
a
resolution
condemning
UN S,SG
modif
NUCLEAR A
predic
WATCHDOG S,SG
relat
PASS V,PRS,FIN,QFIN,TP
determ
AN ART
1-compl
RESOLUTION S,SG
modif
CONDEMN V,ING


Sentence: UN passes a resolution condemning the work of Iran on nuclear weapons , in spite of Iran claim that its nuclear research is for peaceful purpose
UN
passes
a
resolution
condemning
the
work
of
Iran
on
nuclear
weapons,
in spite of
Iran
claim
that
predic
UN S,SG
PASS1 V,PRS,FIN,QFIN,TP
determ
AN ART
1-compl
RESOLUTION S,SG
modif
CONDEMN V,ING
determ
THE ART
1-compl
WORK S,SG
attrib
OF PR
prepos
IRAN S,SG
adverb
ON PR
modif
NUCLEAR A
prepos
WEAPON S,PL
attrib
IN SPITE OF PR
compos
IRAN S,SG
prepos
CLAIM S,SG
1-compl
THAT CONJ
sub-entity
same entity
Communicative action
Anaphora
Sentence: Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops...
Envoy
of
Iran
to
IAEA
proceeds
with
the
dispute
over
its
nuclear
program
predic
ENVOY S,SG
1-compl
OF PR
prepos
IRAN S,SG
attrib
TO1 PR
prepos
IAEA S,SG
PROCEED V,PRS,FIN,QFIN,TP
1-compl
WITH PR
prepos
DISPUTE S,SG
determ
THE ART
1-compl
OVER PR
prepos
PROGRAM S,SG
modif
NUCLEAR A
determ
ITS A
same entity
Anaphora
Communicative action
Anaphora
coordin
AND CONJ
coord-conj
DEVELOP V,PRS,FIN,QFIN,TP
1-compl
SITE S,SG
determ
AN ART
compos
ENRICHMENT S,SG
attrib
IN PR
prepos
SECRET S
subord-conj
BE V,PRS,SG,FIN,QFIN,TP
adverb
FOR PR
prepos
PURPOSE S,SG
modif
PEACEFUL A
predic
RESEARCH S,SG
modif
NUCLEAR A
determ
ITS A
Sentence: Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium
Iran
confirms
that
the
evidence
of
its
nuclear
weapons
program
is
fabricated
predic
IRAN S,SG
1-compl
CONFIRM V,PRS,FIN,QFIN,TP
THAT CONJ
determ
THE ART
predic
EVIDENCE S,SG
1-compl
OF PR
prepos
PROGRAM S,SG
determ
ITS A
compos
WEAPON S,PL
modif
NUCLEAR A
subord-conj
BE V,PRS,SG,FIN,QFIN,TP
pass-anal
FABRICATE V,PP
adverb
BY PR
prepos
USA S,PL
determ
THE ART
coordin
AND CONJ
coord-conj
PROCEEDS S,PL
attrib
WITH PR
prepos
SITE S,SG
determ
THE ART
compos
ENRICHMENT S,SG
compos
URANIUM S,SG
modif
SECOND A
same entity
RST-evidence
Anaphora


Sentence: UN passes a resolution condemning the work of Iran on nuclear weapons , in spite of Iran claim that its nuclear research is for peaceful purpose
UN
passes
a
resolution
condemning
the
work
of
Iran
on
nuclear
weapons,
in spite of
Iran
claim
that
predic
UN S,SG
PASS1 V,PRS,FIN,QFIN,TP
determ
AN ART
1-compl
RESOLUTION S,SG
modif
CONDEMN V,ING
determ
THE ART
1-compl
WORK S,SG
attrib
OF PR
prepos
IRAN S,SG
adverb
ON PR
modif
NUCLEAR A
prepos
WEAPON S,PL
attrib
IN SPITE OF PR
compos
IRAN S,SG
prepos
CLAIM S,SG
1-compl
THAT CONJ
sub-entity
same entity
Communicative action
Anaphora
Sentence: Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops...
Envoy
of
Iran
to
IAEA
proceeds
predic
ENVOY S,SG
1-compl
OF PR
prepos
IRAN S,SG
attrib
TO1 PR
prepos
IAEA S,SG
PROCEED V,PRS,FIN,QFIN,TP
same entity
determ
ITS A
modif
NUCLEAR A
predic
RESEARCH S,SG
subord-conj
BE V,PRS,SG,FIN,QFIN,TP
adverb
FOR PR
modif
PEACEFUL A
prepos
PURPOSE S,SG

Intro to Rhetoric Structure Theory

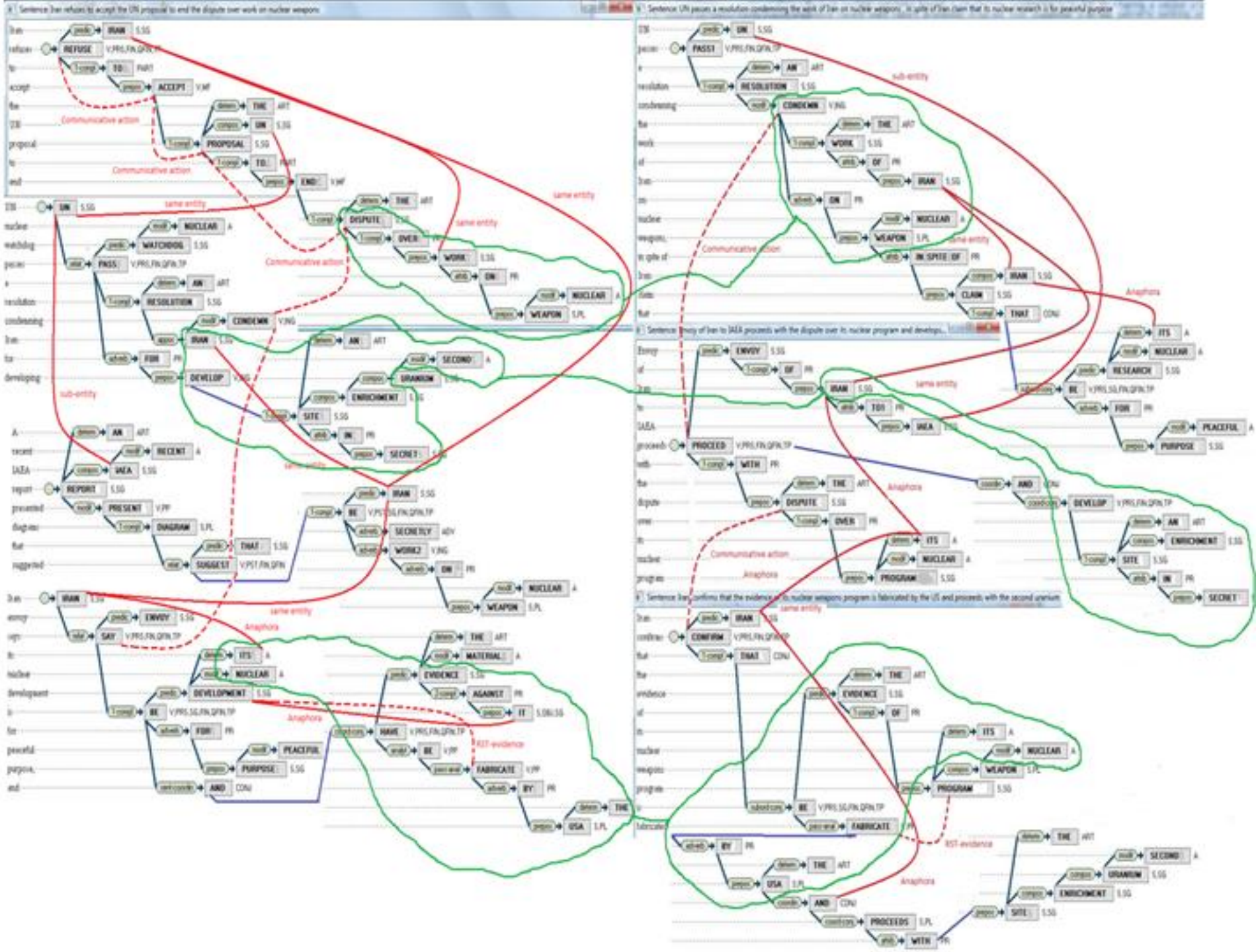
- RST characterizes structure of text in terms of relations that hold between parts of text
- RST helps to discover text patterns such as nucleus/satellite structure.
- RST describes relations between clauses in text which might not be syntactically linked.

Communicative Actions

- Indicates structure of dialogue
- Generalization:
  - one communicative actions from with its subject from T1 against another communicative action with its subject from T2 (communicative action arc is not used)
  - a pair of communicative actions with their subjects from T1 against another pair of communicative actions from T2 (communicative action arcs are used).

Generalization of 2 PT

- Common vertices
- Common arcs
- Arcs with sub-type
- Labeled arcs with intersection


Sentence: Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons
Sentence: UN passes a resolution condemning the work of Iran on nuclear weapons, in spite of Iran claim that its nuclear research is for peaceful purpose
Sentence: Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops...
Sentence: Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium
Communicative action
same entity
sub-entity
Anaphora
RST-evidence

		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 forest generalization based on RST, %, averaging over 20 searches	Relevancy of re-sorting by forest generalization based on SpeechT, %, averaging over 20 searches	Relevancy of re-sorting by hybrid RST+SpeechT forest 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	

We evaluate search relevance improvement, compared to Bing/Yahoo APIs as a baseline, for various complexity of query and answers

Evaluation

For search with multi-sentence match:

- PT-based relevance improves by 6-8% where the query is a single sentence, and then increases to 9-11% as query & answer becomes two-three sentences.
- On average through the cases, PT outperforms single sentence syntactic generalization by 6.7%, whereas RST on its own gives 4.6% and Speech Acts-4.0% improvement respectively.

OpenNLP contribution

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.

It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution.

Syntactic generalization component:

https://svn.apache.org/repos/asf/incubator/opennlp/sandbox/opennlp-similarity

Also available at Google code

https://code.google.com/p/relevance-based-on-parse-trees/

OpenNLP syntactic generalization contribution

Search engineer without knowledge of linguistic / parsing can easily integrate syntactic generalization into an application when relevance is important:

```
SentencePairMatchResult matchRes =
sm.assessRelevance(snapshot, searchQuery);
        List<List<ParseTreeChunk>> match =
matchRes.getMatchResult();
score = parseTreeChunkListScorer.
getParseTreeChunkListScore(match);
        if (score > 1.5) {
           // relevant
        }
```

OpenNLP: how similarity expressions for distinct cases have too high score for bag-of-words

```
String phrase1 = "How to deduct rental expense from income ";
String phrase2 = "How to deduct repair expense from rental income.";
                List<List<ParseTreeChunk>> matchResult  =
parser.assessRelevance(phrase1, phrase2).getMatchResult();
                assertEquals(matchResult.toString(),
                                "[[ [NN-expense IN-from NN-income ],  [JJ-
rental NN-* ],  [NN-income ]], [ [TO-to VB-deduct JJ-rental NN-* ],  [VB-deduct NN-
expense IN-from NN-income ]]]");
double matchScore =
parseTreeChunkListScorer.getParseTreeChunkListScore(matchResult);
double bagOfWordsScore = parserBOW.assessRelevanceAndGetScore(phrase1,
phrase2);
                assertTrue(matchScore+2 < bagOfWordsScore);
                System.out.println("MatchScore is adequate ( = "+matchScore +
") and bagOfWordsScore = "+bagOfWordsScore+" is too high");
```

Conclusions

Parse Thickets is a formalism to systematically learn syntactic representations for paragraphs of text.

Search relevance is improved, when we proceed from keywords to generalizing of parse trees to parse thickets

Thank you