Content Generation / Writer Assistance tool as an extension of OpenNLP

Boris Galitsky

Knowledge-Trail Inc. San Jose CA 95127 USA

Motivations

Most web visionaries think that good quality content comes from:

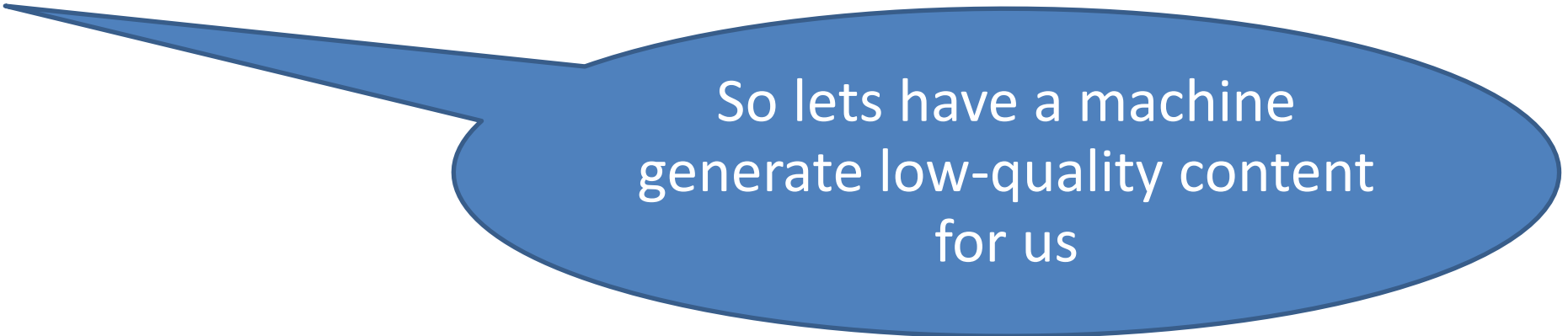
- really passionate fans, or
- from professional journalists, knowing the topic of their writing well.

However, nowadays, the demand for content requires people not as professional and not as passionate to write large amount of text, for a variety of commercial purposes:

- from search engine optimization
- marketing
- self-promotion.

A wide variety of business areas require content creation in some form, including web content,

Expectation of text quality are rather low and deteriorate further, as content is becoming more and more commercial



Objective

To build

- an efficient
- domain-independent
- creative
- interactive

writing tool which produces:

a large volume of content where

- quality and
- effectiveness

are not essential

Plan

- Goals of the project
- Technology
  - Content generation algorithm
  - NLP: Learning parse trees
  - Parse thickets
- Applications
  - Facebook agent
  - Various cases of content generation
- How to use it
  - How to integrate into your application

Innovation

Use of web mining to collect pieces of content
Machine learning of parse trees to filter irrelevant pieces of content from the web
Paragraph-level syntactic structures to form content from pieces
Document-level structure of sections is formed by web mining for attributes of an entity which is a subject of an assay being written

Motivations: why open source

Today search engines are information access gatekeepers
Search engines want their automated agent to analyze / index content created by humans
WE want to automatically build content good enough so that automated agents of search engines do not recognize it as written by humans…
So that search engines would need humans to analyze content created by our machines

Got content from Walmart?

- Today, original content is created by human writers and therefore costly, slowly produced
- Finding a way to automate content creation so that the result is satisfactory for human content consumers and perceived as original by search engine would advance user's web presence and facilitate his professional reputation

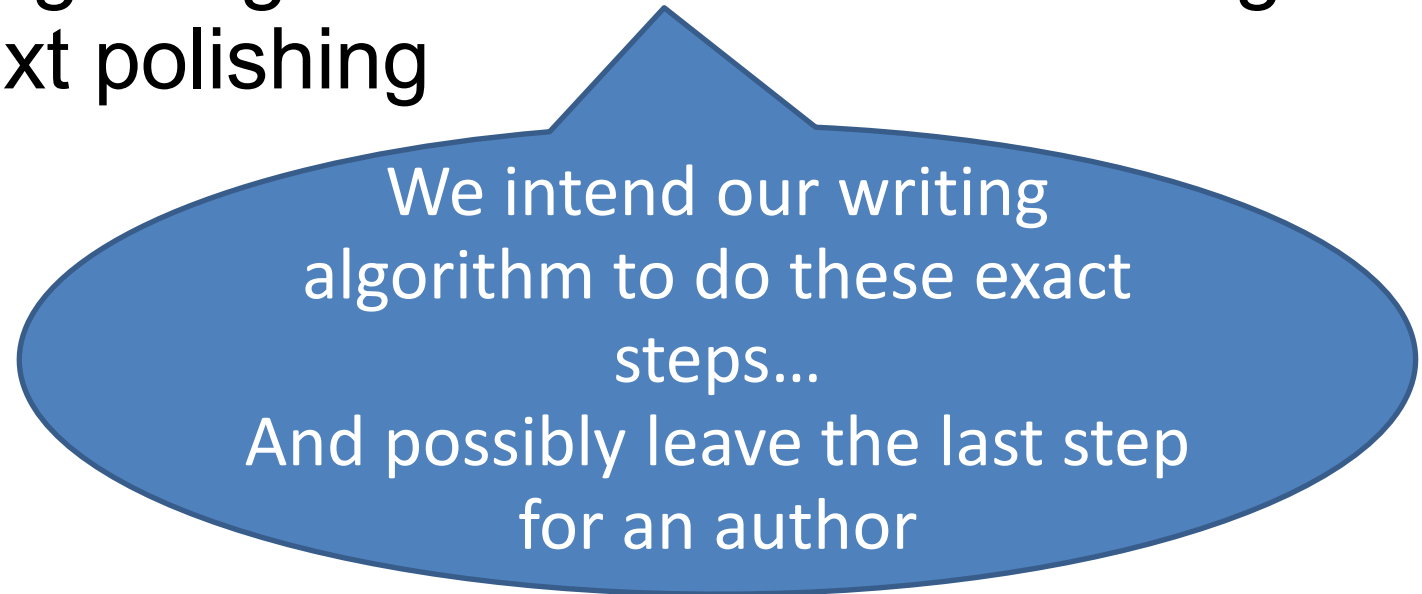
Reformulate a text for the purpose of making it original:

- well known to search engines
- produce uninteresting and semantically non-cohesive

Reproduce the human writing activity

A typical creative writing activity of an average author is:

- searching the web for a subject of writing
- browsing the web for relevant information
- finding candidate pieces
- merging them to form a text
- Inserting images/videos/sound recordings
- final text polishing



Technology

Difficulty = relevance

Eric P.: We will e-mail you a bill for the amount of your money transfer, and you pay using your bank's online bill pay service.

Boris Galitsky: can YOU do this email?

Eric P.: Please clarify and restate your question.

Boris Galitsky: can you please send me email for bill pay information?

Eric P.: Once you have selected WU Pay as the "Pay With" option and clicked the SUBMIT button, the RECEIPT screen will provide instructions on how to fund the Money Transfer.

Boris Galitsky: ok, I will try that

Boris Galitsky: are you an automated agent?

Eric P.: After choosing WU® Pay, you'll receive a bill by email.

Eric P.: Is there anything else I can help you with today?

Boris Galitsky: can you please answer my previous question?

Eric P.: What question Boris?

Boris Galitsky: are you an automated agent?

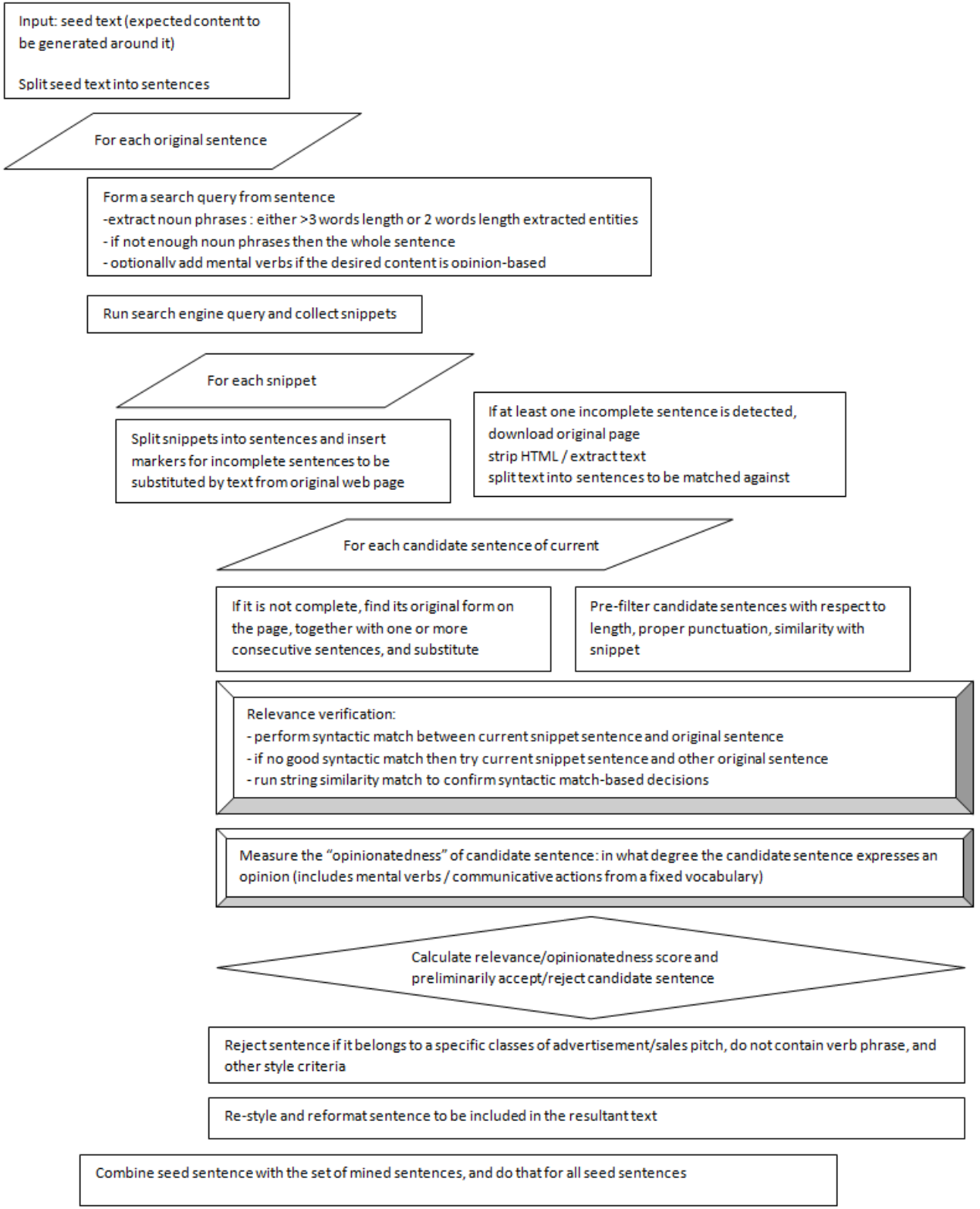
Eric P.: No sir

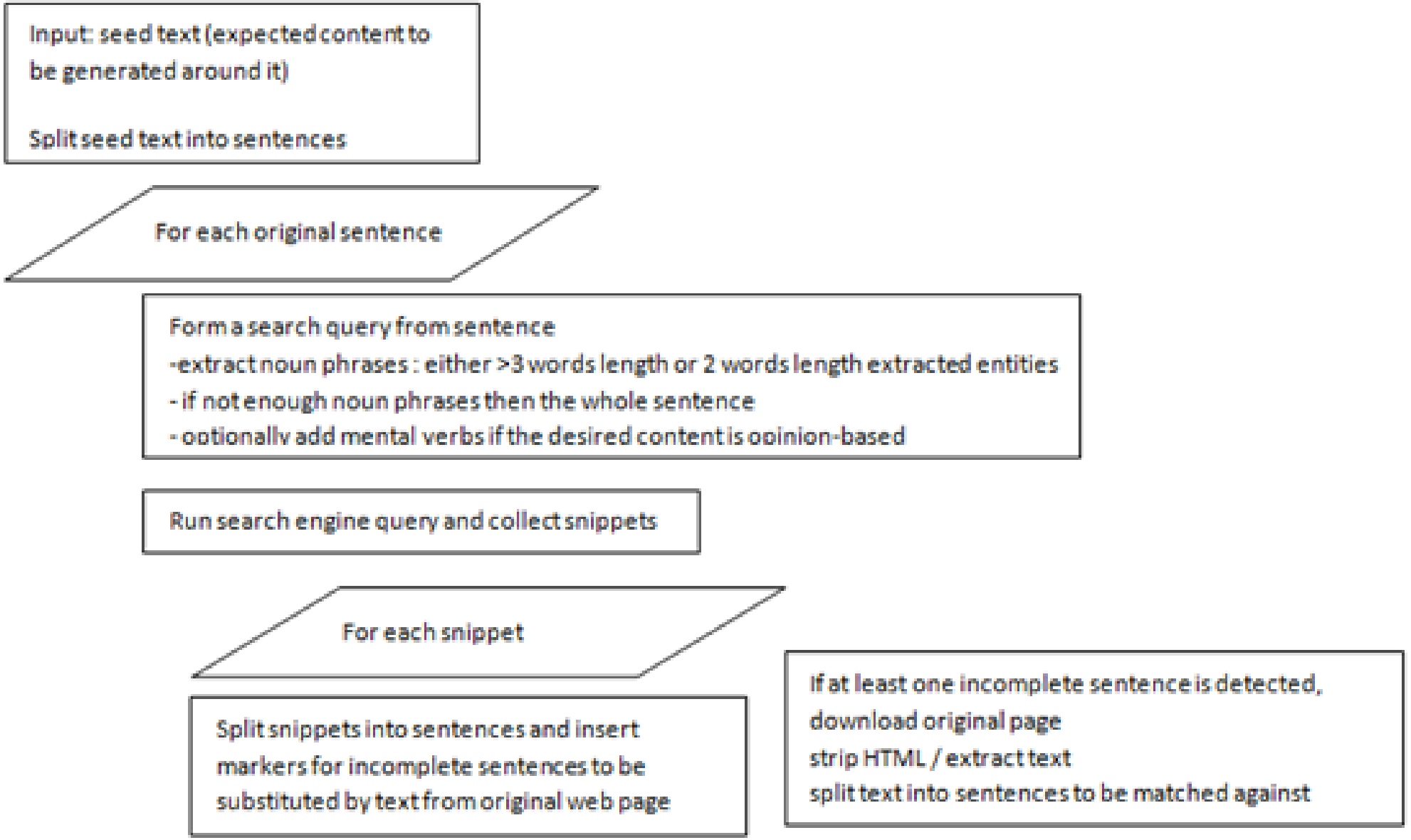
Boris Galitsky: It is my right as a customer to know if I am served by an automated or hur
right?

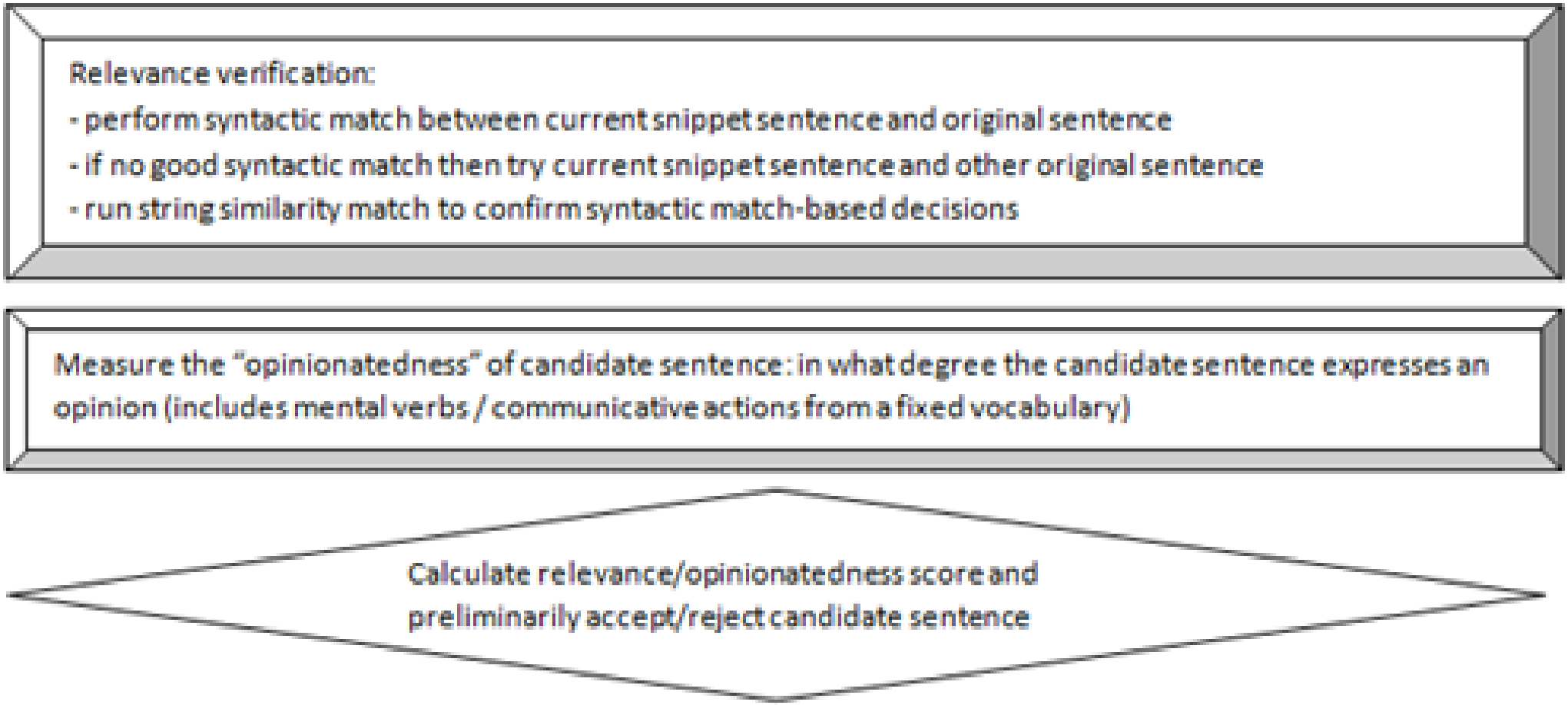
Eric P.: Yes it is

My conversation with WU agent

Content generation algorithm




Input: seed text (expected content to be generated around it)
Split seed text into sentences
For each original sentence
Form a search query from sentence
-extract noun phrases : either >3 words length or 2 words length extracted entities
- if not enough noun phrases then the whole sentence
- optionally add mental verbs if the desired content is opinion-based
Run search engine query and collect snippets
For each snippet
Split snippets into sentences and insert markers for incomplete sentences to be substituted by text from original web page
If at least one incomplete sentence is detected, download original page
strip HTML / extract text
split text into sentences to be matched against


Relevance verification:
- perform syntactic match between current snippet sentence and original sentence
- if no good syntactic match then try current snippet sentence and other original sentence
- run string similarity match to confirm syntactic match-based decisions
Measure the "opinionatedness" of candidate sentence: in what degree the candidate sentence expresses an opinion (includes mental verbs / communicative actions from a fixed vocabulary)
Calculate relevance/opinionatedness score and preliminarily accept/reject candidate sentence

Reject sentence if it belongs to a specific classes of advertisement/sales pitch, do not contain verb phrase, and other style criteria

Re-style and reformat sentence to be included in the resultant text

Combine seed sentence with the set of mined sentences, and do that for all seed sentences

Content generation algorithm: building section structure

We search for a essay topic and obtain main entities associated with it, such as {space, atomic bomb, time machine}

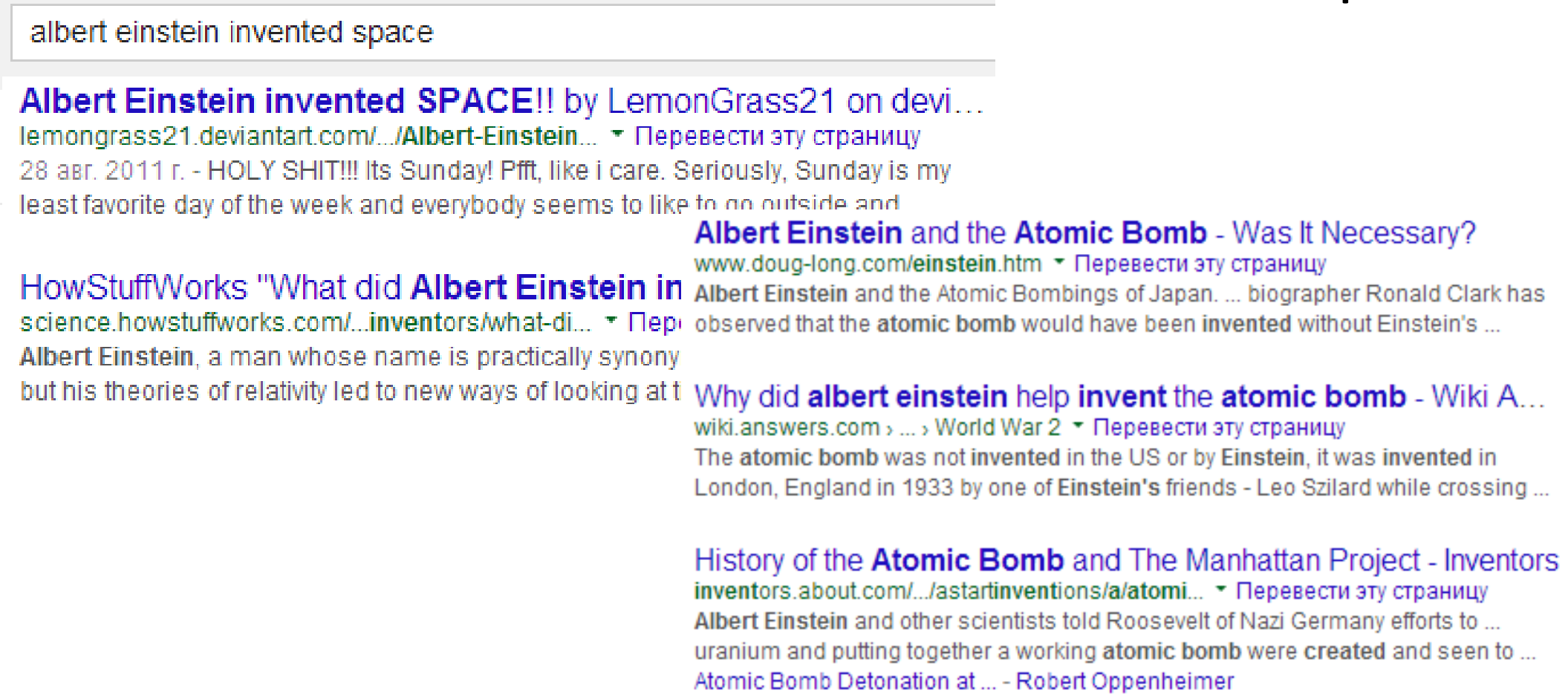
albert einstein invented **space**

albert einstein invented **the atomic bomb**

albert einstein invented **time machine**

Content generation algorithm: building section structure and filling them

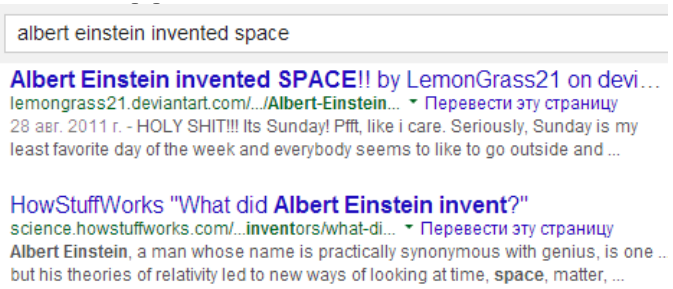
We now search for a essay topic + section topic:



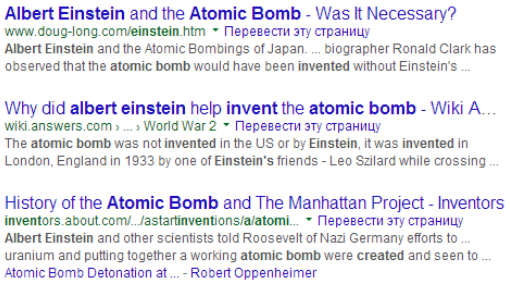
Building section structure and filling them

Now we construct the section structure:

Albert Einstein

1 

nvented Space

2 

ein invented Bomb

Content generation flow 1

For sentence "*Give me a break, there is no reason why you can't retire in ten years if you had been a rational investor and not a crazy trader*"

- We form the query for search engine API: *+rational +investor +crazy +trader*
- From search results we remove duplicates, including *"Derivatives: Implications for Investors | The Rational Walk".*
- From the search results we show syntactic generalization results for the seed and search result:
- Syntactic similarity: np [[IN-in DT-a JJ-*], [DT-a JJ-* JJ-crazy], [JJ-rational NN-*], [DT-a JJ-crazy]], score= 0.9
- Rejected candidate sentence: *Rational opportunities in a crazy silly world.*

Content generation flow 2

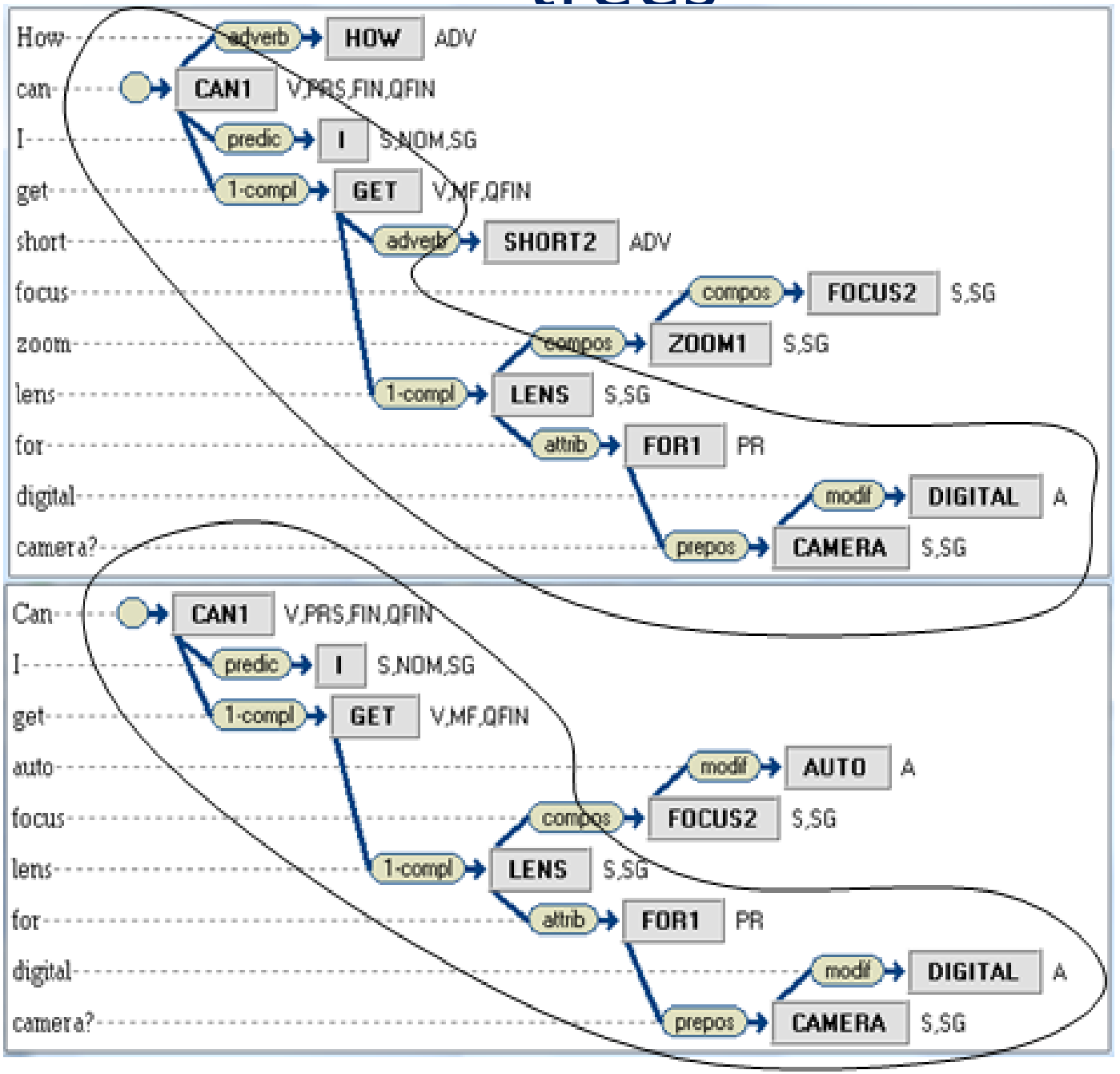
- Syntactic generalization 2: np [[VBN-* DT-a JJ-* JJ-rational NN-investor], [DT-a JJ-* JJ-rational NN-investor]] vp [[DT-a], [VBN-* DT-a JJ-* JJ-rational NN-investor]], score= 2.0
- Accepted sentence: *I have little pretensions about*

.

The latter sentence has significantly stronger semantic commonality with the seed one, compared to the former one, so it is expected to serve as a relevant part of generated content about "*rational investor*" from the seed sentence

NLP: machine learning of parse trees

Similarity = common part among parse trees



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

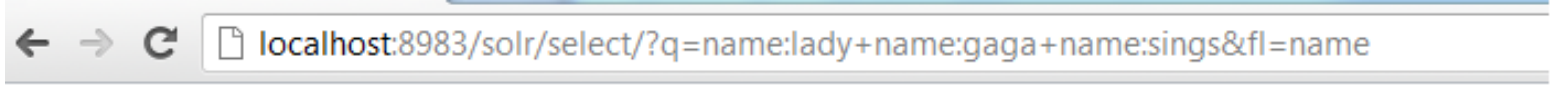
- Extend the operation of least general generalization (unification of logic formula) towards structural representations of paragraph of texts
- Define the operation of *generalization of text paragraphs* to assess similarity between portions of text.
- Use of generalization for similarity assessment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

Now we compare linguistic phrase search and regular SOLR search

SOLR is one of the most popular search frameworks used in industry lucene.apache.org/**solr**/

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



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>
```

Finding similarity between two paragraphs

Original text

"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"

Text, found on the web: is it RELEVANT?

"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]

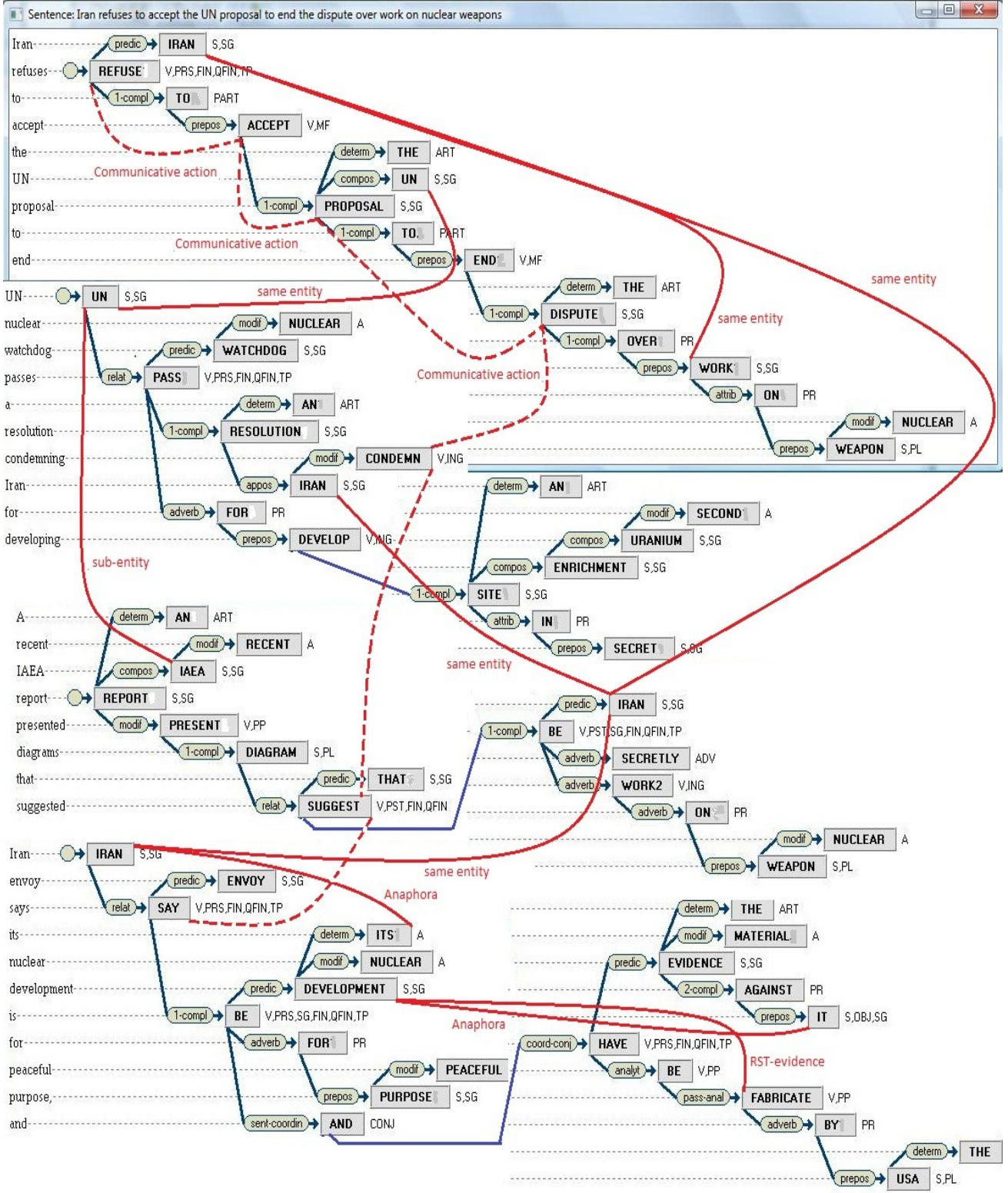
Pair-wise vs. Parse thicket

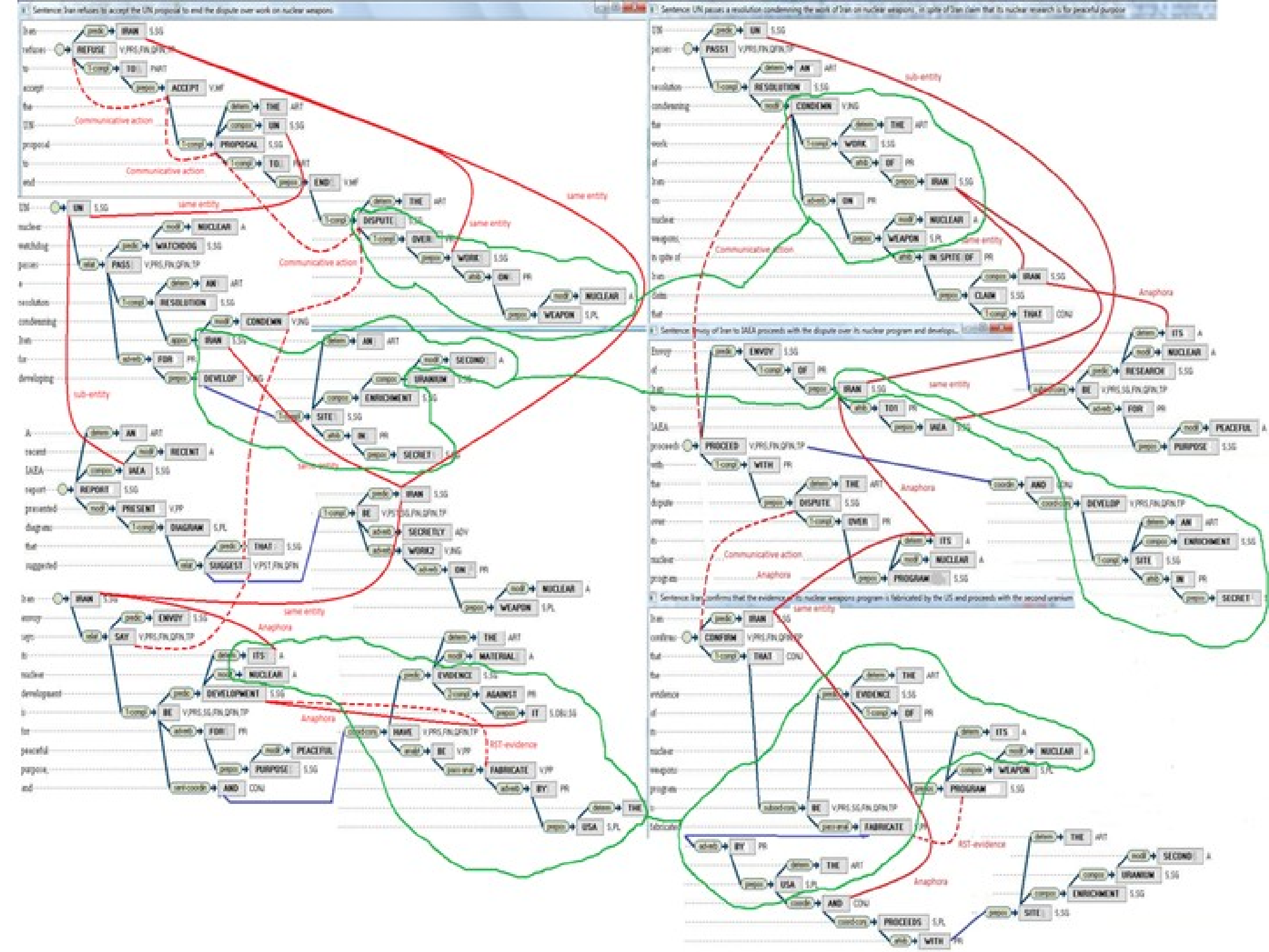
[NN-Iran VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]
[NN-generalization-<UN/nuclear watchdog> * VB-pass NN-resolution VBG condemning NN- Iran]
[NN-generalization-<Iran/envoy of Iran> **Communicative_action** DT-the NN-dispute IN-over JJ-nuclear NNS-*
[**Communicative_action** - NN-work 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
  - 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
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
Communicative action
Communicative action
same entity
same entity
same entity
Communicative action
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
UN
nuclear
watchdog
passes
a
resolution
condemning
Iran
for
developing
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
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
A
recent
IAEA
report
presented
diagrams
that
suggested
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
same entity
Iran
envoy
says
its
nuclear
development
is
for
peaceful
purpose,
and
IRAN S,SG
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
prepos
PURPOSE S,SG
modif
PEACEFUL
sent-coordin
AND CONJ
Anaphora
coord-conj
HAVE V,PRS,FIN,QFIN,TP
analyt
BE V,PP
pass-anal
FABRICATE V,PP
adverb
BY PR
prepos
USA S,PL
determ
THE
RST-evidence
determ
THE ART
modif
MATERIAL A
predic
EVIDENCE S,SG
2-compl
AGAINST PR
prepos
IT S,OBJ,SG


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

Applications

Case Study: Agent acting on behalf of a user posting messages to maintain friendship

Domain of social promotion

- On average, people have 200-300 friends or contacts on social network systems such Facebook and LinkedIn.
- To maintain active relationships with this high number of friends, a few hours per week is required
- In reality, people only maintain relationship with 10-20 most close friends, family and colleagues,
- The rest of friends are being communicated with very rarely. These not so close friends feel that the social network relationship has been abandoned

CASP is about to post a message

Elliott Getaway

Sunset over Lake Tahoe

Like · Comment · Share · 7 hours ago near South Lake Tahoe, CA ·

20 people like this.

Therd Kim Priceless picture ... God's creation at His best.
7 hours ago · Like

Christy Harrisow~amazing purple! 🙂
4 hours ago · Like

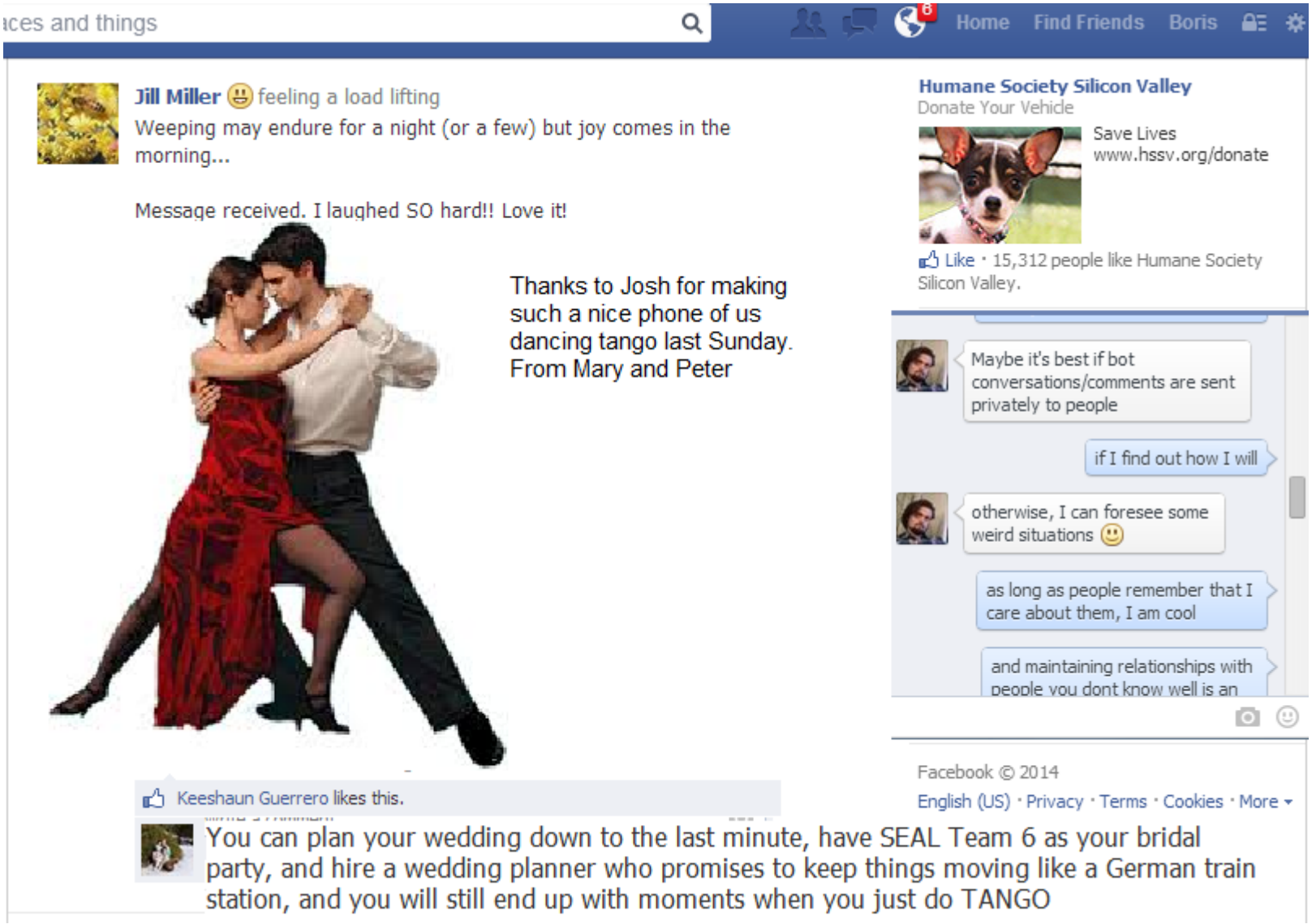
The hike was, round trip, about 8.5 miles and took us just over 6 ... The view from Relay Peak looking south across Lake Tahoe. ... I met Sevski, who had just put in a day of professoring, for a

The hike was, round trip, about 8.5 miles and took us just over 6
... The view from Relay Peak looking south across Lake Tahoe.
... I met Sevski, who had just put in a day of professoring, for a
couple quick runs before sunset.

Case study: what happened

- Friends posted a photo of them dancing tango
- CASP commented on the posting, finding information about “tango at a wedding”
- Friends got upset and communicate that the posting of CASP citing the wedding was irrelevant and they did not like it.
- ….Nevertheless they married 5 months later

Case study


aces and things
Home
Find Friends
Boris
Jill Miller feeling a load lifting
Weeping may endure for a night (or a few) but joy comes in the morning...
Message received. I laughed SO hard!! Love it!
Thanks to Josh for making such a nice phone of us dancing tango last Sunday. From Mary and Peter
Keeshaun Guerrero likes this.
You can plan your wedding down to the last minute, have SEAL Team 6 as your bridal party, and hire a wedding planner who promises to keep things moving like a German train station, and you will still end up with moments when you just do TANGO
Humane Society Silicon Valley
Donate Your Vehicle
Save Lives
www.hssv.org/donate
Like · 15,312 people like Humane Society Silicon Valley.
Maybe it's best if bot conversations/comments are sent privately to people
if I find out how I will
otherwise, I can foresee some weird situations
as long as people remember that I care about them, I am cool
and maintaining relationships with people you dont know well is an
Facebook © 2014
English (US) · Privacy · Terms · Cookies · More

Conversational characters

Simulated human characters and agents are increasingly common components of user interfaces of applications.

They include:

- interactive tutoring environments,
- eldercare systems,
- virtual reality-based systems,
- intelligent assistant systems,
- physician-patient communication training,
- entertainment applications

(Cassell et al., 2000, De Rosis et al., 2003; Dias and Paiva 2005; Lisetti 2008).

Limitations of current conversational characters

- Improvement in the intelligent features, expressiveness, understanding human language, dialog abilities and other aspects.
- Social realism is still far behind
- Don't take into consideration fact that users consistently respond to computers as if they were social actors (Reeves and Nass 1996)
- Poor understanding of user needs in the areas such as shopping, finance, conflict resolution, etc.

Tolerance to dealing with auto agents

- Users / Readers would tolerate:

  - worse than ideal style,

  - discourse and

  - quality of content being communicated,

as long as overall the communication is positive and makes sense.

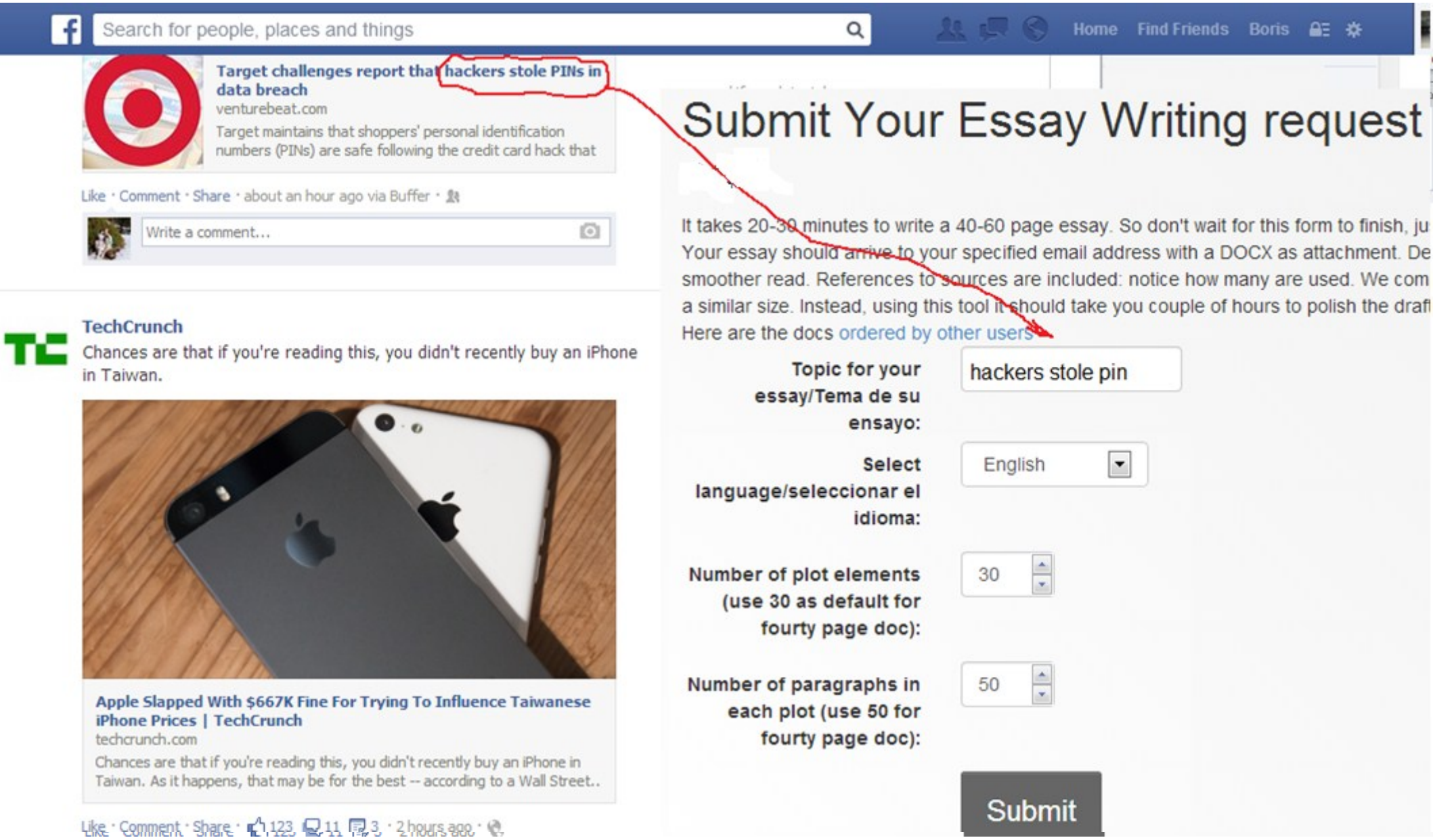
- Messages should at least show social engagement
- Currently available commercial chat bots employed by customer support portals, or packaged as mobile apps, possess too limited

  - NLP

  - text understanding

  - overall robust intelligent capabilities to support conversations to support social promotion.

CASP writes an essay on a topic



Content Generation part of CASP

- The content generation part of CASP can be used at www.facebook.com/RoughDraftEssay
- Given a topic, it first mines the web to auto build a taxonomy of entities which will be used in the future comment or essay
- Then the system searches the web for these entities to find pieces of content to create respective sections
- The resultant document is delivered as DOCX email attachment
- CASP can automatically compile texts from hundreds of sources to write an essay on the topic
- If a user wants to share a comprehensive review, opinion on something, provide a thorough background, then this interactive mode should be used.

Evaluation results: relevance

Query type	Query complexity	Relevance of baseline Bing search, %, averaging over 100 searches	Relevance single-sentence phrase-based generalization search, %, averaging over 100 searches	Relevance of thicket-based phrase generalization search, %, averaging over 100 searches	Relevance of parse thicket-based graph generalization search, %, averaging over 100 searches
Facebook friend agent support search	1compound sent	74.5	83.2	85.3	87.2
	2 sent	72.3	80.9	82.2	83.9
	3 sent	69.7	77	81.5	81.9
	4 sent	70.9	78.3	82.3	82.7

Evaluation of CASP

We measure the failure score of an individual CASP posting as the number of failures in relevance and appropriateness. The postings started with [posted by an agent on behalf of Boris]

After a certain number of failures, friends

- complain
- unfriend
- share negative information about the loss of trust with others
- and even encourage other friends to unfriend an friend who is enabled with CASP.

Trust vs Domain & Complexity

Values are the average numbers of failed CASP postings, for given domain and given complexity, when issues with trust arises

On average, friends complain after 6 failures and want to unfriend CASP's host after 10 failures

Topic of the seed	Complexity of the seed and posted message	A friend complaints to the CASP's host	A friend unfriends the CASP host	A friend shares with other friends that the trust in CASP is lost	A friend encourages other friends to unfriend a friend with CASP
Travel & outdoor	1 sent	6.3	8.5	9.4	12.8
	2 sent	6.0	8.9	9.9	11.4
	3 sent	5.9	7.4	10.0	10.8
	4 sent	5.2	6.8	9.4	10.8
Shopping	1 sent	7.2	8.4	9.9	13.1
	2 sent	6.8	8.7	9.4	12.4
	3 sent	6.0	8.4	10.2	11.6
	4 sent	5.5	7.8	9.1	11.9
Events & entertainment	1 sent	7.3	9.5	10.3	13.8
	2 sent	8.1	10.2	10.0	13.9
	3 sent	8.4	9.8	10.8	13.7
	4 sent	8.7	10.0	11.0	13.8
Job-related	1sent	3.6	4.2	6.1	6.0
	2 sent	3.5	3.9	5.8	6.2
	3 sent	3.7	4.0	6.0	6.4
	4 sent	3.2	3.9	5.8	6.2
Personal life	1 sent	7.1	7.9	8.4	9.0
	2 sent	6.9	7.4	9.0	9.5
	3 sent	5.3	7.6	9.4	9.3
	4 sent	5.9	6.7	7.5	8.9
Average		6.03	7.5	8.87	10.575

Trust vs Relevance

- For less information-critical domains like travel and shopping, tolerance to failed relevance is relatively high.
- Conversely, in the domains taken more seriously, like job related, and with personal flavor, like personal life, users are more sensitive to CASP failures and the loss of trust in its various forms occur faster
- For all domains, tolerance slowly decreases when the complexity of posting increases.

Conclusions for Facebook application

- We proposed a problem domain of social promotion and build a conversational agent CASP to act in this domain
- CASP maintains friendship and professional relationships by automatically posting messages on behalf of its human host
- It was demonstrated that a substantial intelligence in information retrieval, reasoning, and natural language-based relevance assessment is required to retain trust in CASP agents by human users

Content generation agent can be trusted

- We evaluated the scenarios of how trust was lost in Facebook environment.
- We confirmed that it happens rarely enough for CASP agent to improve the social visibility and maintain more friends for a human host than being without CASP.

Although some friends lost trust in CASP, the friendship with most f[illegible] by CASP

Overall CASP's impact on social activity is positive.

Published work

- Galitsky, B., Josep Lluis de la Rosa & Gábor Dobrocsi. Building Integrated Opinion Delivery Environment FLAIRS-24 2011.
- Galitsky, B. and Josep Lluis de la Rosa. Concept-based learning of human behavior for customer relationship management. Special Issue on Information Engineering Applications Based on Lattices. Information Sciences. Volume 181, Issue 10, 15 May 2011, pp 2016-2035, 2011.
- Galitsky, B., Josep-Lluis de la Rosa, and Boris Kovalerchuk. Assessing plausibility of explanation and meta-explanation in inter-human conflict. Engineering Application of AI, V 24 Issue 8, pp 1472-1486, 2011.
- Galitsky, B. Josep Lluis de la Rosa, Gábor Dobrocsi. Inferring the semantic properties of sentences by mining syntactic parse trees. Data & Knowledge Engineering. Available online 28 July 2012.
- Galitsky, B. Machine Learning of Syntactic Parse Trees for Search and Classification of Text. Engineering Application of Artificial Intelligence, 2012 (to appear).
- Galitsky, B. Exhaustive simulation of consecutive mental states of human agents. Knowledge-Based Systems, 2013

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

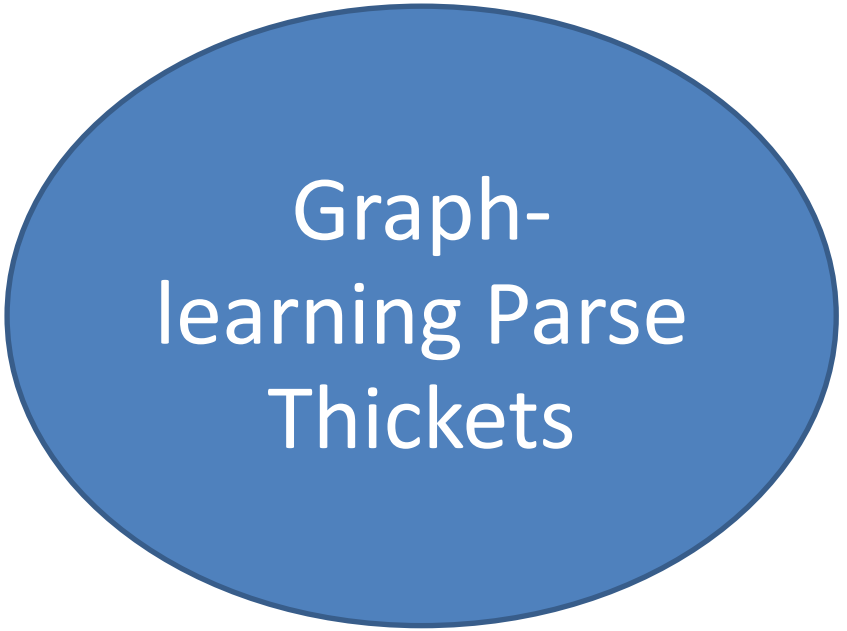
Also available at Google code

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

This project among open source projects

JGraphT

OpenNLP



Stanford NLP

Linguists can now enjoy graph-learning of linguistic structures, and graph community – new rich source of graphs

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(matc
hResult);
double bagOfWordsScore =
parserBOW.assessRelevanceAndGetScore(phrase1, phrase2);
      assertTrue(matchScore+2 < bagOfWordsScore);
      System.out.println("MatchScore is adequate ( =
"+matchScore + ") and bagOfWordsScore =
"+bagOfWordsScore+" is too high");
```

Value of this research

- We believe this is one of the first systems in learning a semantic discourse to solve a content generation relevance problem, based on a set of parse trees.
- Instead of using linguistic information of individual sentences, we can now compute text similarity at the level of paragraphs.

Project @ Google Code

The relevance assessment code is developed as a part of OpenNLP project and is available

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