Parse Thicket representations for paragraphs of text

Boris A. Galitsky\textsuperscript{1,2}, Dmitry Ilvovsky\textsuperscript{2}, Sergei O. Kuznetsov\textsuperscript{2} and Fedor Strok\textsuperscript{2}

\textsuperscript{1} eBay Inc San Jose CA USA
\textsuperscript{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

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

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.

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
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

- No answer includes ‘pulmonologist’
What happens if we match keywords not phrases
If we specify “string phrases” a search engine is doing a little better.
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
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">2</int>
  </lst>
  <lst name="params">
    <str name="q">name:lady name:gaga name:sings</str>
    <str name="fl">name</str>
  </lst>
</response>

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

```xml
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">36</int>
  </lst>
  <lst name="params">
    <str name="qt">dismax</str>
    <str name="q">name:lady name:gaga name:sings</str>
    <str name="fl">name</str>
  </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>
    </doc>
  </result>
</response>
```

Phrase (natural language) search
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]
Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons.
UN passes a resolution condemning the work of Iran on nuclear weapons, in spite of Iran's claim that its nuclear research is for peaceful purposes.

Sentence 2: Iran proceeds with the dispute over its nuclear program and develops

Sentence 3: Iran confirms that the evidence of its nuclear weapons program was fabricated by the US and proceeds with the second uranium enrichment process.
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.
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
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:

```java
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