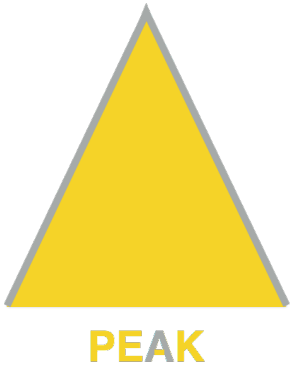


PEAK: Pyramid Evaluation via Automated Knowledge Extraction

Qian Yang, Rebecca J. Passonneau, Gerard de Melo

PhD Candidate, Tsinghua University

Visiting Student, Columbia University

http://www.larayang.com/

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Evaluating Summary Content

- **Human assessors**
  - Judge each summary individually
  - Very time-consuming and does not scale well
- **ROUGE** (Lin 2004)
  - Automatically compares n-grams with model summaries
  - Not reliable enough for individual summaries (Gillick 2011)
- **Pyramid Method** (Nenkova and Passonneau, 2004)
  - Semantic comparison, reliable for individual summaries
  - Has required manual annotation

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Our Contribution

- No need for manually created pyramids
- Also good results on automatic assessment given a pyramid

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

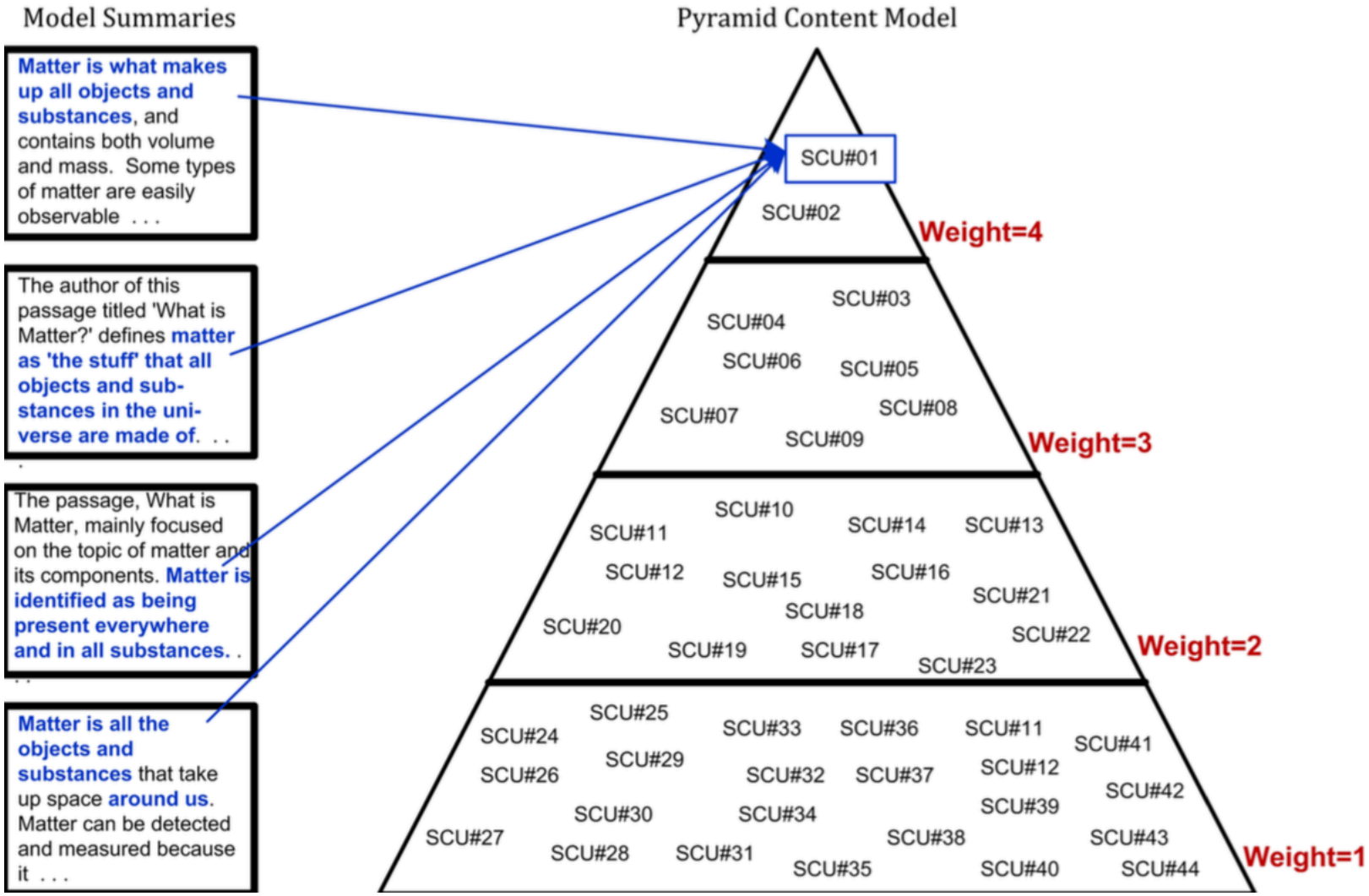
Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Content

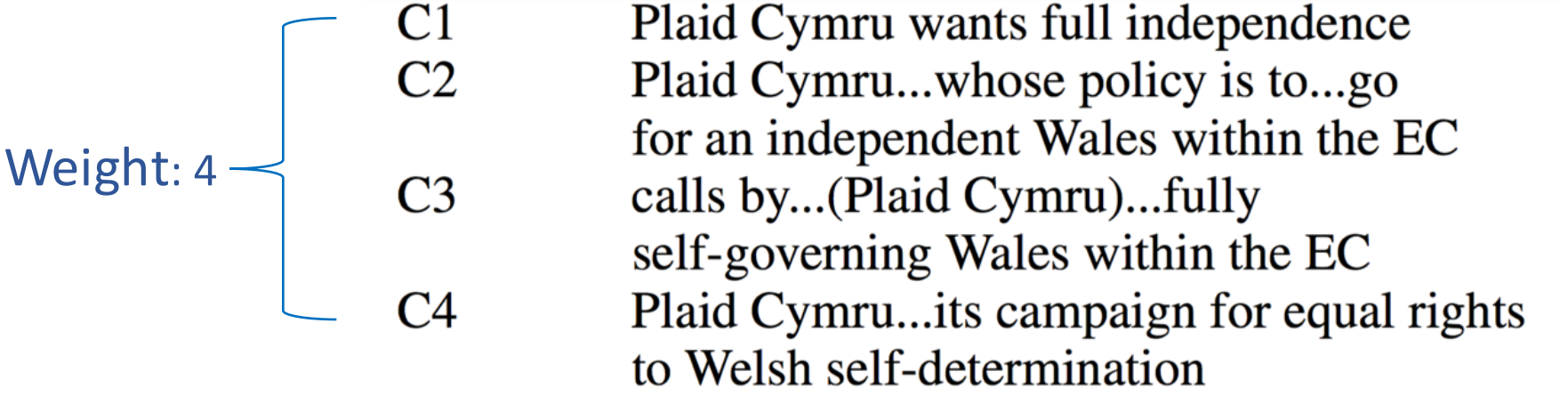
- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Semantic Content Analysis



Source: http://www1.ccls.columbia.edu/~beck/pubs/2458_PassonneauEtAl.pdf

Semantic Content Analysis



SCU 49	Plaid Cymru wants full independence
C1	Plaid Cymru wants full independence
C2	Plaid Cymru...whose policy is to...go for an independent Wales within the EC
C3	calls by...(Plaid Cymru)...fully self-governing Wales within the EC
C4	Plaid Cymru...its campaign for equal rights to Welsh self-determination

Figure 1: Sample SCU from *Pyramid Annotation Guide: DUC 2006*.

Semantic Content Analysis

- "*The law of conservation of energy is the notion that energy can be transferred between objects but cannot be created or destroyed.*"
- Open information extraction (Open IE) methods split them and extract
  <subject,predicate,object>
  triples

Semantic Content Analysis

- "*These characteristics determine the properties of matter*"

  yields the triple

  ⟨*These characteristics*, *determine*, *the properties of matter*⟩
- We use ClausIE (Del Corro and Gemulla 2013)

Semantic Content Analysis

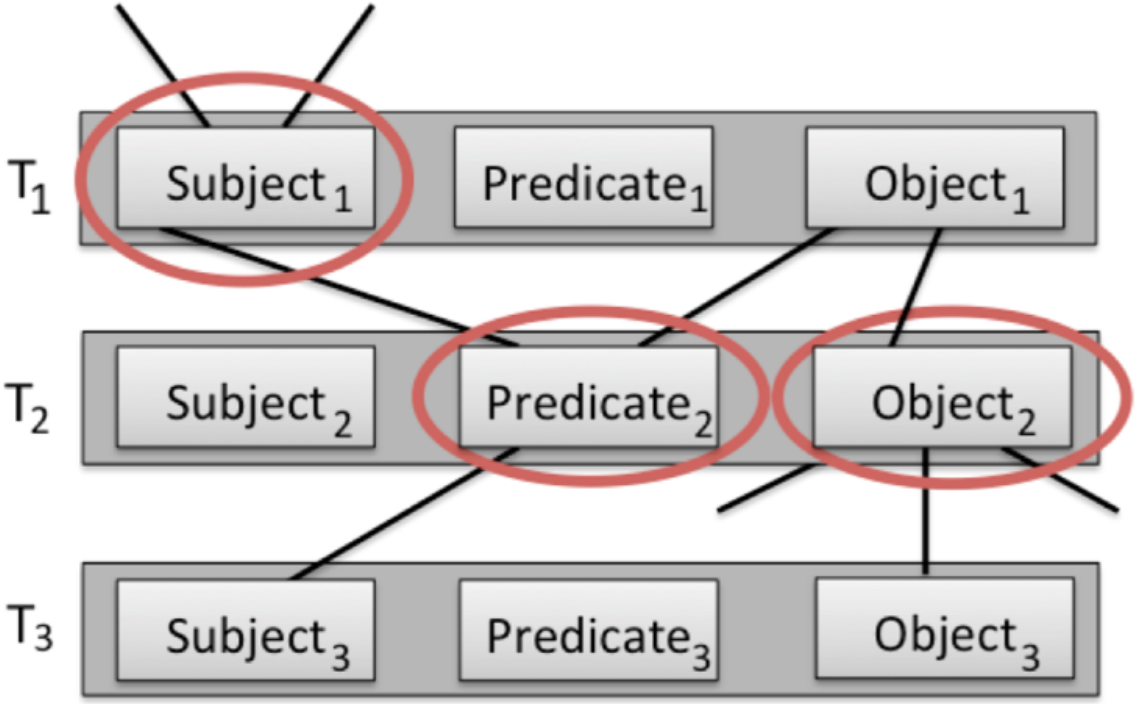


Figure 2: Hypergraph to capture similarites between elements of triples, with salient nodes circled in red

Similarity Score: *Align, Disambiguate and Walk* (ADW) (Pilehvar, Jurgens, and Navigli 2013),

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Pyramid Induction

WEIGHT: 4
ANCHOR: "Matter" "is" "all the objects and substances"

FROM SENTENCE1 of CONTRIBUTOR1: Matter is all the objects and substances that take up space around us.
SALIENT NODES: "Matter" "all the objects and substances"

CONTRIBUTOR1: "Matter" "is" "all the objects and substances"
CONTRIBUTOR2: "Matter" "is identified" "as being present everywhere and in all substances"
CONTRIBUTOR3: "The author of this passage titled What is Matter" "defines" "matter as the stuff that all objects and substances in the universe are made of"
CONTRIBUTOR4: "Matter" "is" "what makes up all objects or substances and contains both volume and mass"

Pyramid Induction

Similarity Class E1 for "Matter"	Similarity Class E2 for "all the objects and substances"

E1 = {
"Matter",
"All matter",
"matter",
"the matter itself",
"a different matter",
"the matter itself systematically", …
}
	E2 = {
	"all the objects and substances",
	"the substance",
	"all objects and substances in the universe",
	"what makes up all objects or substances and contains both volume and mass",
	"as being present everywhere and in all substances", …
	}

Pyramid Induction

Algorithm 1 Merge similar SCUs

```
1:  procedure MERGE(SCU anchors, weights)
2:      set a graph G whose nodes are all SCU anchors
3:      set threshold T_1
4:      for each node anchor_m do
5:          for each node anchor_n do
6:              calculate similarityScore_{m,n}
7:              if similarityScore_{m,n} ≥ T_1 then
8:                  add edge between anchor_m and anchor_n
9:      mergedSCU ← the connected component in G
10:     mergedWeight ← max. weight of connected component
11:     Return mergedAnchor, mergedWeight
```

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Scoring – Pyramid Method

- Score a target summary against a pyramid
  - Annotators mark spans of text in the target summary that express an SCU
  - The SCU weights increment the raw score for the target summary.

- An Example
  - SCU Label: Plaid Cymru wants full independence
  - Target Summary: **Plaid Cymru demands an independent Wales**

Automated Scoring – PEAK

Algorithm 2 Computing scores for target summaries

```
1:  procedure SCORE(target summary sum)
2:      for each sentence s in sum do
3:          T_s ← triples extracted from s
4:      for each triple t ∈ ⋃ T_s do
5:          for each SCU s with weight w do
6:              m ← similarity score between t and s
7:              if m ≥ T then
8:                  W[t][s] ← w                          ▷ store weight
9:      S ← Munkres-Kuhn (Hungarian) Algorithm(W)
10:     Return S
```

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Dataset

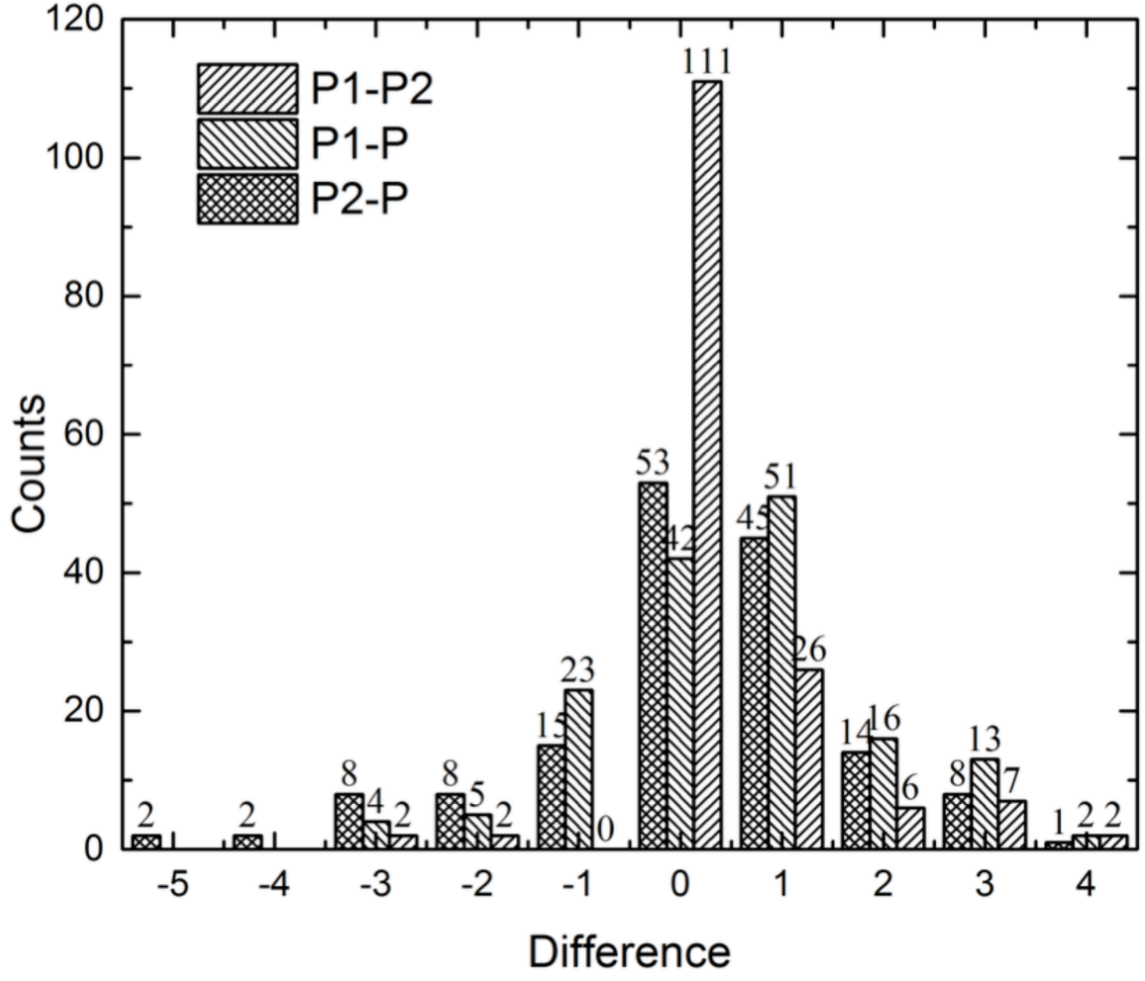
- Student summary dataset from Perin et al. (2013) with 20 **target summaries** written by students
- Passonneau et al. (2013) had produced 5 **reference model summaries**, and 2 **manually created pyramids**

Results

	P1 + M. Scoring	P2 + M. Scoring
P + A.Scoring	0.8263	0.7769
P2 +A. Scoring	0.8538	0.8112
P1 + M. Scoring	1	0.8857

Table 1: Pearson's correlations between scores based on PEAK's pyramid P as well as the two human pyramids P1, P2, with either manual or automatic scoring.

Results


P1-P2
P1-P
P2-P
Counts
Difference
120
100
80
60
40
20
0
-5
-4
-3
-2
-1
0
1
2
3
4
2
2
8
4
2
8
5
2
15
23
0
53
42
111
45
51
26
14
16
6
8
13
7
1
2
2

Result

- Machine-Generated Summaries
  - Dataset: the 2006 Document Understanding Conference (DUC) administered by NIST (“DUC06”)
  - The Pearson’s correlation score between PEAK’s scores and the manual ones is 0.7094.

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Content

- Evaluating Summary Content
- Our Contribution
- How does PEAK work?
  - Semantic Content Analysis
  - Pyramid Induction
  - Automated Scoring
- Our Results
- Conclusion

Conclusion

- The first fully automatic version of the pyramid method
- Not only evaluates target summaries but also generates the pyramids automatically
- Experiments show that
  - Our SCUs are similar to those created by humans
  - The method for assessing target summaries automatically has a high correlation with human assessors

- Overall, our research shows great promise for automated scoring and assessment of manual or automated summaries, opening up the possibility of wide-spread use in the education domain and in information management.

This data and codes are available at

http://www.larayang.com/peak/.

Thank you!