Annotating Argumentative Relations for Mining Coherence Patterns

Jan **Wira** Gotama Putra (D1)

Tokunaga-lab (Computational Linguistics/NLP)

Dept. Computer Science, School of Computing

Tokyo Institute of Technology, Japan

Organisation Example in Student Essay

Prompt: Smoking should be banned at all restaurants in the country

1. Yes, smoking should be completely banned in all restaurants in the country.

Major claim

2. Smoking is dangerous to our health and the government should empower people all over the country to quit
3. Vendors should not have any cigarettes in their stores to avoid smoking
4. When people do not follow the law, the government should give them sanctions

Body: Supporting and opposing arguments

5. Therefore, we should impose NO SMOKING CAMPAIGN

Conclusion

Coherence Example in Student Essay

Prompt: Smoking should be banned at all restaurants in the country

1. Yes, smoking should be completely banned in all restaurants in the country.
2. Smoking is dangerous to our health and the government should empower people all over the country to quit
3. Vendors should not have any cigarettes in their stores to avoid smoking
4. When people do not follow the law, the government should give them sanctions
5. Therefore, we should impose NO SMOKING CAMPAIGN

Is it okay to switch these two sentences?

Illustration: How to Improve Text Coherence

(1) I think smoking should be banned at restaurants.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

(3) At restaurants, customers enjoy eating and talking.

(1) I think smoking should be banned at restaurants.

(3) At restaurants, customers enjoy eating and talking.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy *the experience*.

Why Coherence

- Coherence dictates how to order sentences (information) properly

- Explaining how to order sentences to generate coherent texts is important in many natural language generation applications:
    - Multi-document summarisation of news (Barzilay et al., 2002; Okazaki et al., 2004)
    - Opinion generation in debate (Yanase et al., 2015)
    - Argumentative mirco-texts (Peldszuz and Stede, 2016)
    - Intelligent Language Tutoring Systems (Al-khatib et al., 2016)

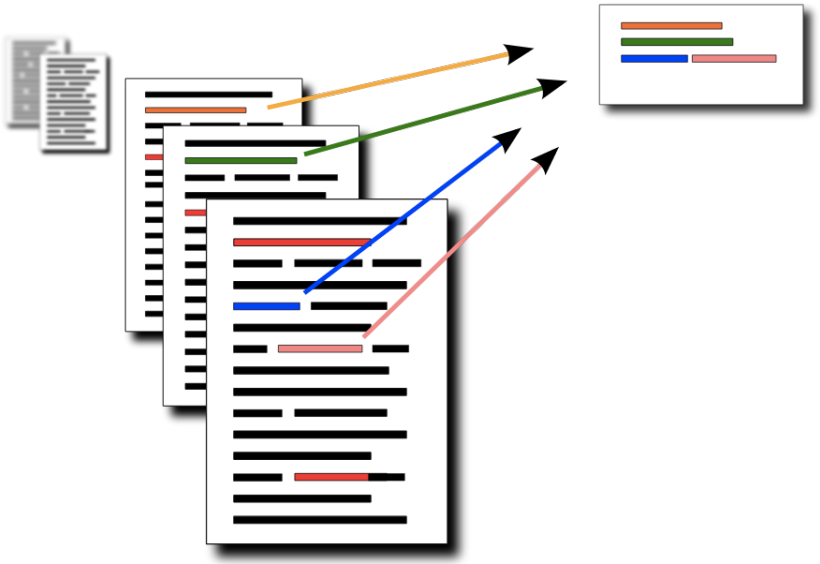


Image courtesy: http://mogren.one/summarization/

Intelligent Language Tutoring Systems

- Automatic Essay Scoring + Score-based Feedback

  Telling students which aspect he/she is lacking

- Example-based Feedback

  Telling students how to order sentences properly by giving a proper ordering example based on his/her own text

Research Goal (Ongoing)

- Our previous study revealed that there is a possibility of difference of relational patterns between coherent and less coherent text (Putra and Tokunaga, 2017)

- Explaining how sentences are related to each other explains text coherence

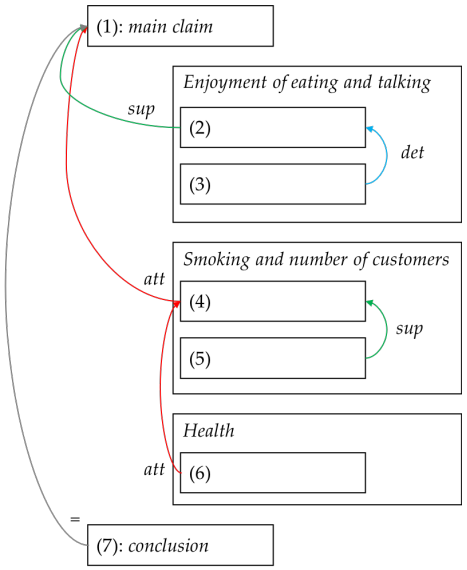
- We aim to explain text coherence by analysing relations between sentences

  Annotating imperfect text, then compare the relational patterns between the imperfect and improved text

Argumentative Structure as a Representation

Incoherent Text

[1,2,3,…]



Argumentative Structure
(Common Representation)

Coherent Text

[1,3,2,…]

How the structure is realised (in sequential ordering) differentiates coherent and incoherent text

Text Collection Criteria

- Imperfect texts (score as proof) → argumentative essays
- Enables us to isolate the study of coherence;
  e.g., do not consider the grammatical issue
- Not to long (to reduce complications), only one paragraph if possible

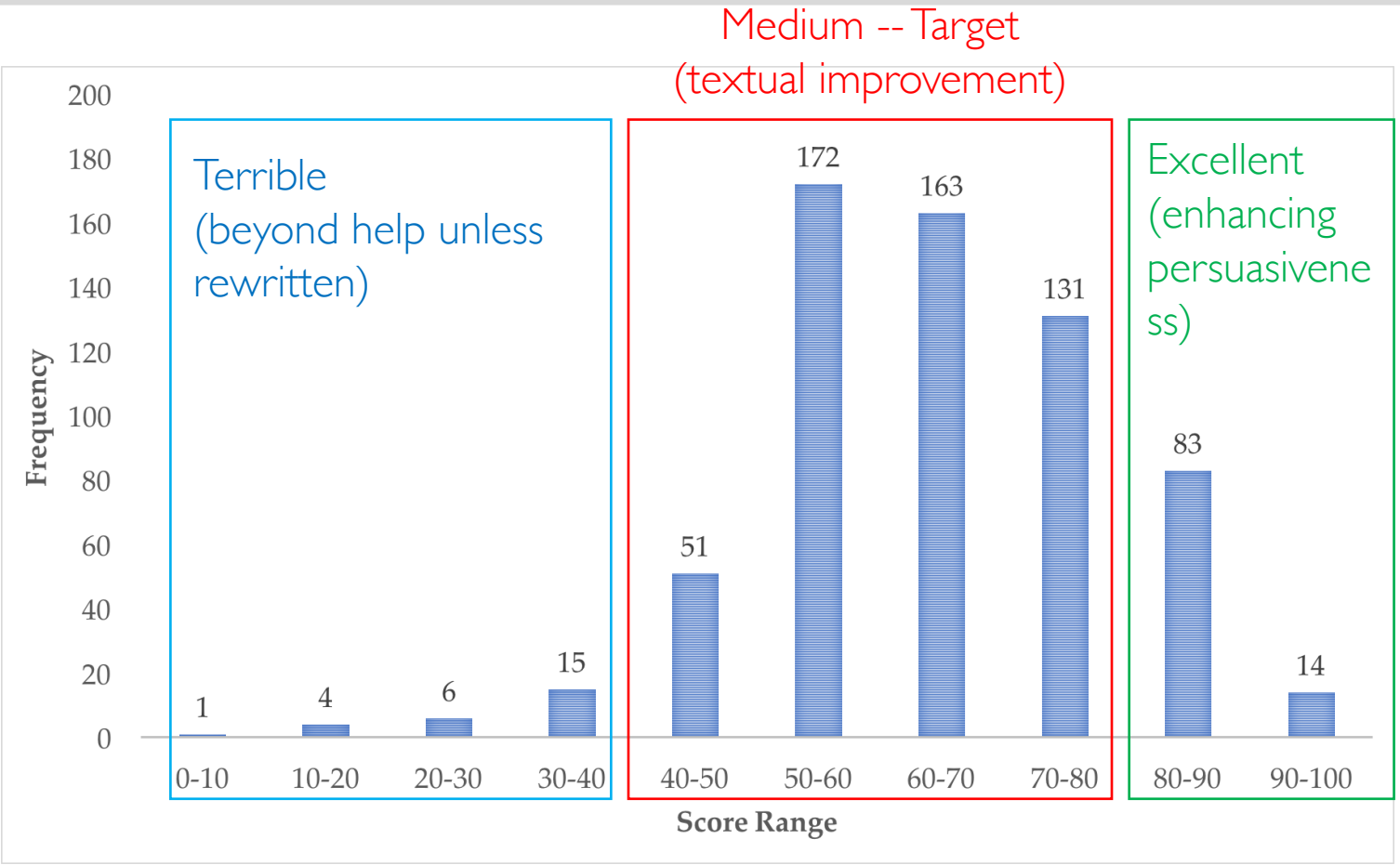
Some Existing Argumentative Corpora

Source	Genre	#Doc	IAA
Stab and Gurev (2014)	Argumentative Essays from online fo (sentence labelling + relations)	90	• Krippendorff α_U = 0.72 (sentence label) • Fleiss K = 0.80~ (relations) • 3 annotators each
Kirschner	Scientific Artic intr	24	• Fleiss K = 0.43 • 3 annotators each
Al-Khatib et al.		300	• Fleiss K = 0.56 • 3 annotators each
Peldszuz and Stede		112	• Fleiss K = 0.83 • #annotators is not clear
Stab and Gu (sentence lab		402	• Krippendorff α_U = 0.767 (sentence label) • Fleiss K = 0.70~ (relations) • 80 essays by 3 annotators for IAA • One expert annotator annotated the rest 300~
Carlile et al. (2018)	Argumentative Essay	102	• Krippendorff α_U = 0.50++ • 2 annotators each

None satisfy our criteria

ICNALE Dataset (Ishikawa, 2013)

- Collection of argumentative writing by Asian students (common writing in education)
- Student essays typically need revisions
- All essays are much or less same at length (no length bias): (200-300 words)
- Grammatically **error-free subset** available (640 essays out of 2800) – April 2018



Relation Annotation

support (sup)

I think banning smoking in all restaurants is necessary

It is essential to protect the citizen's health

detail (det)

I think banning smoking in all restaurants is necessary

Banning means officially forbid them to smoke

attack (att)

I think banning smoking in all restaurants is necessary

But, some restaurants are popular because men are allowed to smoke

Kirschner et al., 2015

restatement (=)

I think smoking should be banned at all restaurants

In conclusion, I believe smoking should be banned at all restaurants

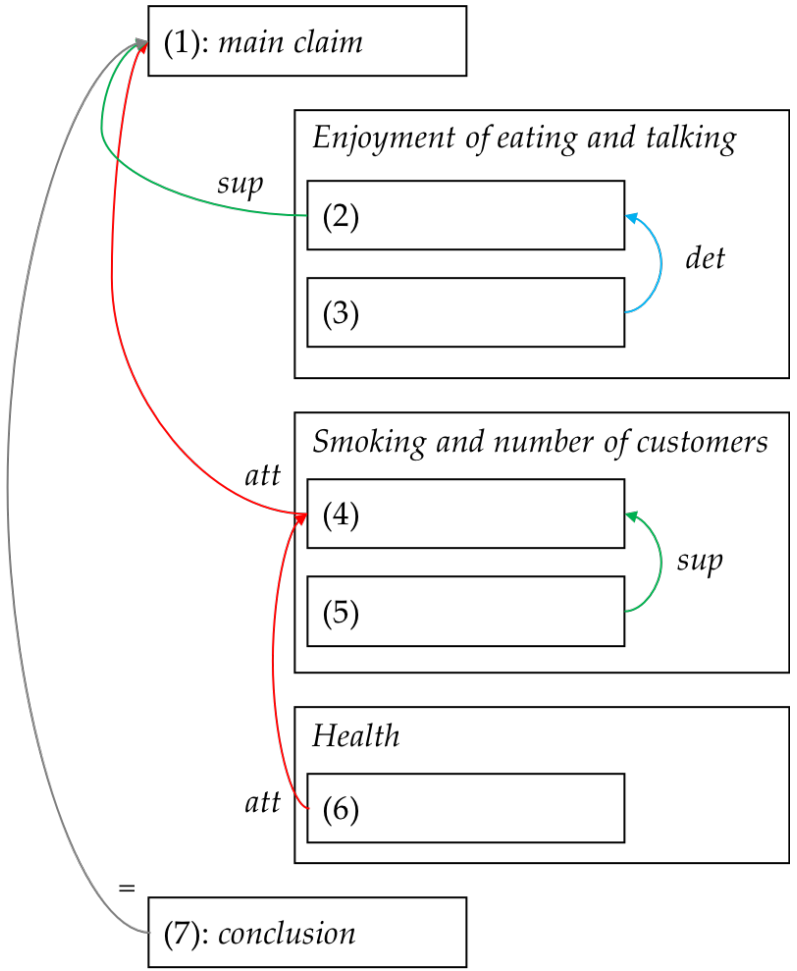
Skeppstedt et al., 2018
two restatement nodes as an equivalence class

Relation Annotation

- Goal:
  Producing structure (tree)

- Objects:
  - argument components,
  - non-argumentative components (dropped)

- Relations:
  {*support, detail, attack, restatement*}
  *from source (*S*) to target (*T*)



Annotation Steps

- Annotating relations (cf. argument mining)
  - Top-down
    - Find a *main claim* (root)
    - Divide sentences into groups (and subgroups at deeper level)
  - Bottom-up
    - Establish relations between sentences in a group
    - Establish relations between groups (higher hierarchical level)
- Reordering sentences (improving coherence)
- Repairing referring and connective expressions

  Reordering may alter how people and things are described or connected

Annotation Illustration

(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

(3) At restaurants, customers enjoy eating and talking.

(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.

(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.

(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.

(7) In conclusion, I encourage banning smoking in all restaurants.

Annotation Illustration: Step 1 (Main Claim)

(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement. → Main Claim

(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

(3) At restaurants, customers enjoy eating and talking.

(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.

(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.

(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.

(7) In conclusion, I encourage banning smoking in all restaurants.

Annotation Illustration: Step 2 (Grouping)

(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

(3) At restaurants, customers enjoy eating and talking.

(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.

(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.

(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.

(7) In conclusion, I encourage banning smoking in all restaurants.

(1): *main claim*

Enjoyment of eating and talking

(2)

(3)

Smoking and number of customers

(4)

(5)

Health

(6)

(7): *conclusion*

Annotation Illustration: Step 3 (Relations)

(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

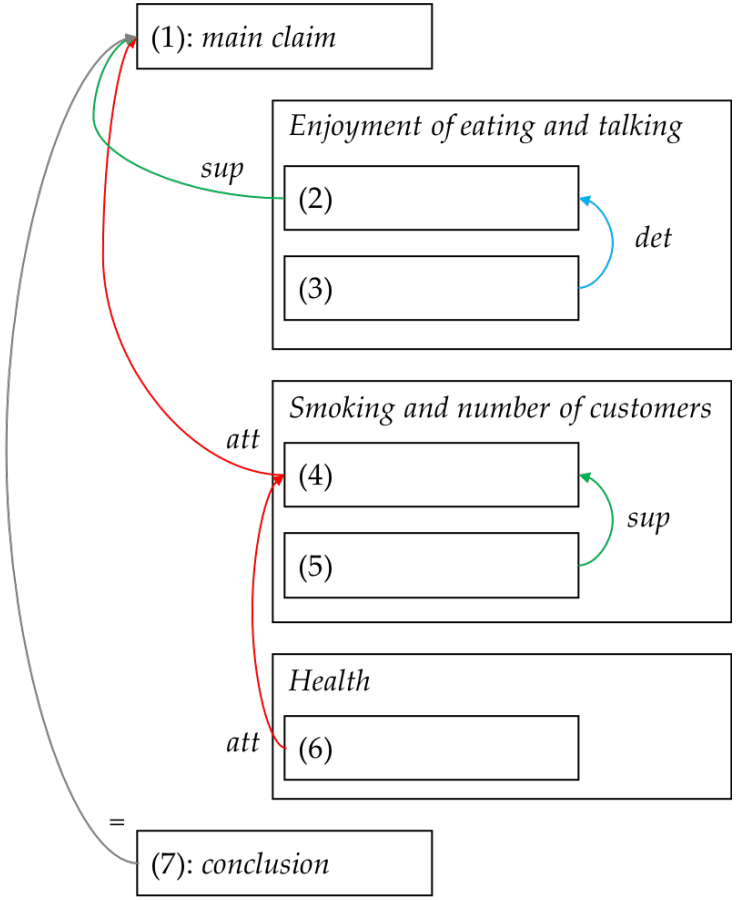
(3) At restaurants, customers enjoy eating and talking.

(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.

(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.

(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.

(7) In conclusion, I encourage banning smoking in all restaurants.



Annotation Illustration: Step 4 (Reordering)

(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement.

(3) At restaurants, customers enjoy eating and talking.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy *the experience*.

(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.

(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.

(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.

(7) In conclusion, I encourage banning smoking in all restaurants.

Annotation Illustration: Step 5 (Text Repair)

(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement.

Implies readers have read the prompt

(3) At restaurants, customers enjoy eating and talking.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.

(5) [...] some restaurants are indeed popular, especially among old men, because they allow people to smoke.

(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.

(7) In conclusion, I encourage banning smoking in all restaurants.

Annotation Illustration: Step 5 (Text Repair)

(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree [~~with the previous statement~~ | that smoking should be banned at all restaurants].

(3) At restaurants, customers enjoy eating and talking.

(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

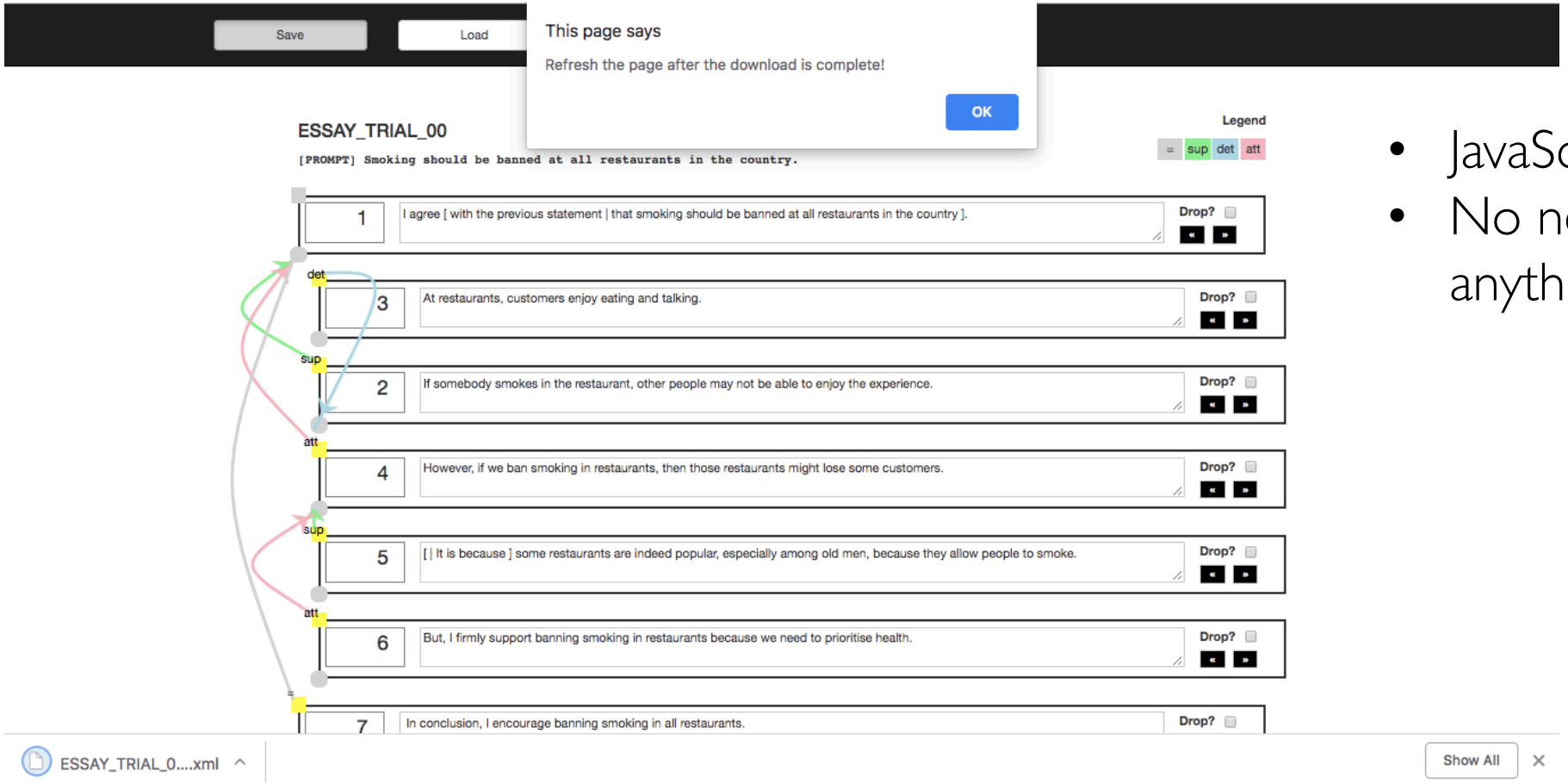
(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.

(5) [| It is because] some restaurants are indeed popular, especially among old men, because they allow people to smoke.

(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.

(7) In conclusion, I encourage banning smoking in all restaurants.

Annotation Tool



- JavaScript
- No need to install anything (web-based)

Measuring Inter-annotator Agreement (IAA)

- Agreement Ratio for Dropping (argumentative vs non-argumentative sentence)

  It is because most sentences are argumentative, so using chance-corrected measure leads to a low score and underestimate the agreement

- Chance-corrected measure + modified agreement ratio for relational agreement

  Calculating how many relations (edges in the graph) having the same label are shared between two annotation

- Modified agreement ratio for *multiplied* relation

  Taking into account restatement resolution when measuring relational agreement

Inter-Annotator Agreement on Dropping

Treat as binary labelling of sentences (drop--TRUE or not drop--FALSE)

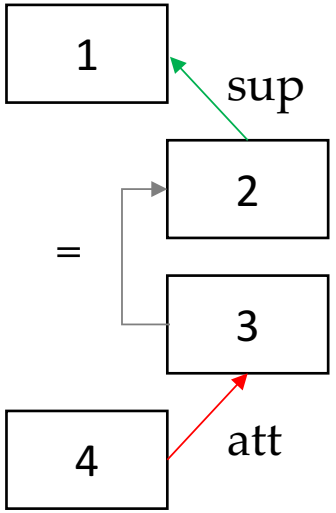
Sentence	Annotator 1	Annotator 2
1	TRUE	FALSE
2	FALSE	TRUE
3	**TRUE**	**TRUE**
4	**FALSE**	**FALSE**
5	**FALSE**	**FALSE**
6	**TRUE**	**TRUE**

4/6 = 0.67

Traditional Inter-Annotator Agreement on Relations

For an essay having N sentences

- $N \times (N-1)$ combinations of source and target
- We measure the relation type + linking at once



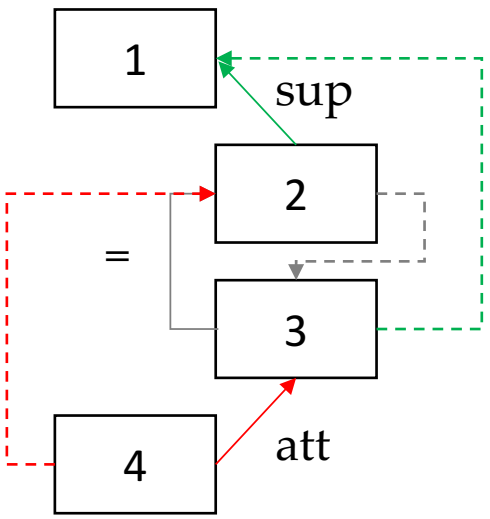
Source	Target	Annotator 1	Annotator 2
1	2	NO	
…			
2	1	sup	
2	3	NO	
…			…
3	2	=	

Cohen's (Fleiss) Kappa
F1-Score
Raw Agreement Ratio
Adapted Percentage (=Ratio) Agreement

Problem: cannot reflect agreement score for restatement and structural similarity

Restatement Resolution: ``Extra Relation" (Multiplication)

We treat two restatement nodes as an equivalence class with respect to incoming and outgoing connections with extra relations



Source	Target	Original Relation	Extra Relation
2	1	sup	sup
2	3		=
3	1		sup
3	2	=	=
4	2	att	att
4	3		att

It is better not to take into account *non-relations* as in traditional agreement ratio, because it is rather meaningless (Kirschner et al., 2015)

Graph Similarity (Structure Similarity)

- Graph Edit Distance -- *inexact matching* (Alberto and King-sun, 1983)
  - Given two graphs G and H, find an *edit path* between G and H, that is a sequence of node or edge *insertion*, *removal* or *substitution* which transforms G to H
  - It is not suitable with our task because we cannot just substitute a vertex with another one

- Maximum Common Edge Subgraph -- *exact matching* (Bokhari, 1981)
  - Given two graphs G and H with the same number of vertices, one has to find a common subgraph of G and H (not necessarily *induced*) with the maximum number of edges
  - The solution to this problem is known as NP-complete and requires vertices matching (graph isomorphism)
  - Straightforward solution: check if a subgraph in G presents in H → $O(2^{|V|})$; $|V|$ = *#vertices*

Modified Agreement Ratio (MAR) with Restatement Resolution: ``Extra Relation" (Multiplication)

To what extent a modified graph subsumes another graph (approximating maximum common edges)

$$G1 = \frac{\text{overlap}(g_2, \text{multiply}(g_1))}{\text{edges}(g_1)}$$

← Number of edges in g_2 exist in the multiplied relations of g_1

← Number of original edges of g_1

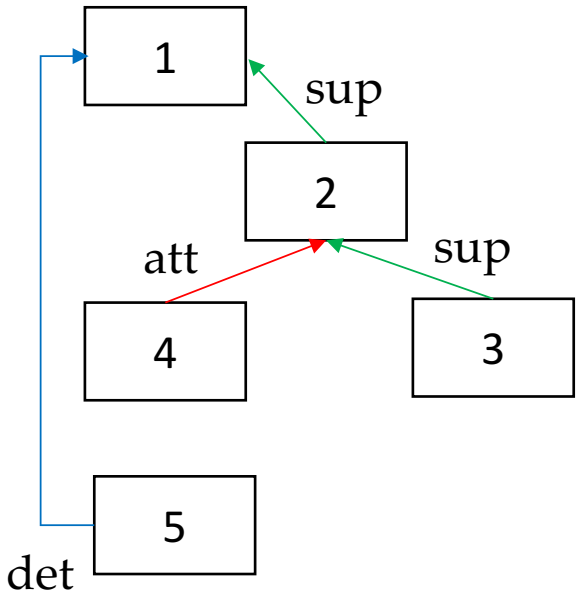
$$G2 = \frac{\text{overlap}(g_1, \text{multiply}(g_2))}{\text{edges}(g_2)}$$

$$\text{MAR} = \frac{G1 + G2}{2}$$

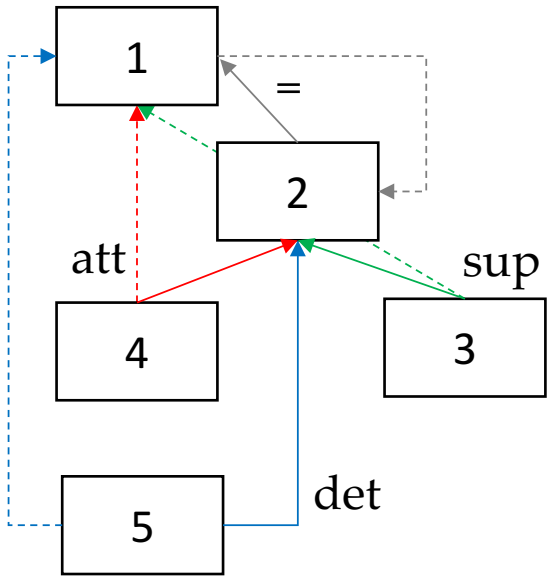
Arithmetic mean
we welcome your suggestions if you think geometric or harmonic mean is better
*) Note: we also have an option of not populating the graph

Illustration: ``Extra Relation" (Multiplication)

Annotator 1 (g_1)



Annotator 2 (g_2)



$$G1 = \frac{\text{overlap}(g_2, g_1)}{\text{edges}(g_1)} = \frac{2}{4} = 0.50$$

$$G2 = \frac{\text{overlap}(g_1, g_2)}{\text{edges}(g_2)} = \frac{2}{4} = 0.50$$

$$\text{MAR} = \frac{G1 + G2}{2} = \frac{0.50 + 0.50}{2} = 0.50$$

$$G1 = \frac{\text{overlap}(g_2, \text{multiply}(g_1))}{\text{edges}(g_1)} = \frac{2}{4} = 0.50$$

$$G2 = \frac{\text{overlap}(g_1, \text{multiply}(g_2))}{\text{edges}(g_2)} = \frac{3}{4} = 0.75$$

$$\text{MAR} = \frac{G1 + G2}{2} = \frac{0.50 + 0.75}{2} = 0.625$$

Our solution is an approximation of maximum common edges (but do not necessarily form a subgraph)

Pilot Study: Dropping

Essay Code	Score	Agreement Ratio			All
		W - D	W - O	D - O	
THA_PTJ0_001_B1_2	45.8	**0.73**	**0.73**	0.60	0.53
JPN_PTJ0_041_B2_0	65.4	0.93	**1.00**	0.93	0.93
SIN_PTJ0_014_B2_0	91.3	**1.00**	0.86	0.86	0.86
Overall Ratio		**0.89**	0.87	0.80	0.78

Overall Fleiss Kappa	**0.56**	0.43	0.35	0.44

Pilot Study: Relational Agreement

Metrics	Annotators		
	W - D	W - O	D - O
Cohen's Kappa	**0.59**	0.50	0.40
Modified Agreement Ratio w/o multiplication	**0.56**	0.46	0.35
Modified Agreement Ratio with multiplication	**0.56**	0.46	0.35
Modified Agreement Ratio with multiplication w/o considering label	**0.64**	0.61	0.48

Fleiss Kappa = 0.50

It happens that in our current pilot study, restatements happen in the exactly same pair of sentences between annotators

Confusion Matrix

W \ D	=	att	det	n	sup
=	2	0	0	0	0
att	0	2	0	5	1
det	0	0	5	8	2
n	0	3	2	584	5
sup	0	0	0	2	11

O \ D	=	att	det	n	sup
=	1	0	0	1	0
att	0	2	0	7	0
det	0	0	4	4	3
n	1	3	2	581	11
sup	0	0	1	6	5

O \ W	=	att	det	n	sup
=	1	0	0	1	0
att	0	6	0	3	0
det	0	0	6	2	3
n	1	2	7	581	7
sup	0	0	2	7	3

- The most confused relations are "det" and "sup" (this is expected)
- However, the difference of resulting hierarchical structures is the most pressing problem

Verdict

- Annotators recognise different structures (different interpretation of texts)
  - Different ordering → different grouping/relations → different structures
  - IAA implicitly measures the difference of dropping and structures

- Higher agreement on higher scored essays
  - I think because of the detailed instruction of recognising groups (organisation).
  - It is easier to recognise groups in highly-scored essays since they are more organised
  - Annotating non-perfect essays is challenging (interpretation problem)

Plan (near future)

- Annotate larger corpus
- Argument quality assessment (organisation score)

Contact Me

URL https://wiragotama.github.io
E-mail gotama.w.aa@m.titech.ac.jp

References (1)

1. Jan Wira Gotama Putra and Takenobu Tokunaga. 2017. Evaluating text coherence based on semantic similarity graph. In TextGraphs-11 ACL, pages 76-85
2. Naoki Okazaki, Y. Matsuo and M. Ishizuka. 2004. Improving choronological sentence ordering by precedence relation. In Proceedings of COLING.
3. Regina Barzilay, Noemie Elhadad, Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. In Journal of Artificial intelligence research 17 (2002), pages 35-55.
4. T. Yanase, T. Miyoshi, K. Yanai, M. Sato, M. Iwayama, Y. Niwa, P. Reisert and K. Inui. 2015. Learning sentence ordering for opinion generation of debate. In Proceedings of Workshop on Argument Mining, ACL.
5. Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In COLING, pages 1501-1510.
6. Christian Kirschner, Judith Eckle-Kohler and Iryna Gurevych. 2015. Linking the thoughts: analysis of argumentation structure in scientific publications. In 2nd Workshop on Argumentation Mining, ACL, pages 1-11.
7. Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In COLING, pages 3433-3443.
8. Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In Argumentation and Reasoned Action: 1st European Conference on Argumentation (ECA 16). College Publications.

References (2)

9. Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays.
10. Marco Lippi and Paolo Torroni. 2016. Argument Mining: State of the art and emerging trends. In ACM Transactions on Internet Technology, Vol. 16, No. 2, Article 10.
11. Winston Carlile, Nishant Gurrapadi, Zixuan Ke, Vincent Ng. 2018. Give me more feedback: annotating argument persuasiveness and related attributes in student essays. In ACL, pages 1-11.
12. Takshak Desai, Parag Dakle, Dan I. Moldovan. 2018. Generating questions for reading comprehension using coherence relations. In BEA, ACL, pages 1-10
13. S. Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of asian learners of english. Learner Corpus Students in Asia and the World 1, 91-118
14. S. Bokhari, On the mapping problem, IEEE Transactions on Computation 30 (3) (1981) 207–214.
15. Sanfeliu, Alberto; Fu, King-Sun. 1983. A distance measure between attributed relational graphs for pattern recognition. IEEE Transactions on Systems, Man and Cybernetics. 13 (3), pages 353–363. doi:10.1109/TSMC.1983.6313167.
16. Maria Skeppstedt, Andreas Peldszus, Manfred Stede. 2018. More or less controlled elicitation of argumentative text: enlarging a microtext corpus via crowdsourcing. In workshop of argument mining, EMNLP, pages 155-163.