Educational Applications of Natural Language Processing

Jan **Wira** Gotama Putra

https://wiragotama.github.io/

Computational Linguistics Laboratory

Tokyo Institute of Technology, Japan

Early Note

Feel free to interrupt my talk when you have questions

"*He who asks a question is a fool for five minutes; he who does not ask a question remains a fool forever*"

– Chinese Proverb

Agenda

- Self introduction
- Introduction to machine learning and deep learning (promotion)
- Representing text as vector (NLP pipeline)
- Essay scoring
- Grammatical Error Correction
- Essay Improvement (structural)

Jan **Wira** Gotama Putra

School	Tokyo Institute of Technology	(Master, PhD—now)
	Institut Teknologi Bandung	(Bachelor)
Research	Natural Language Processing	

Bali Anime Photography Python, C/C++

Introduction to Machine Learning and Deep Learning

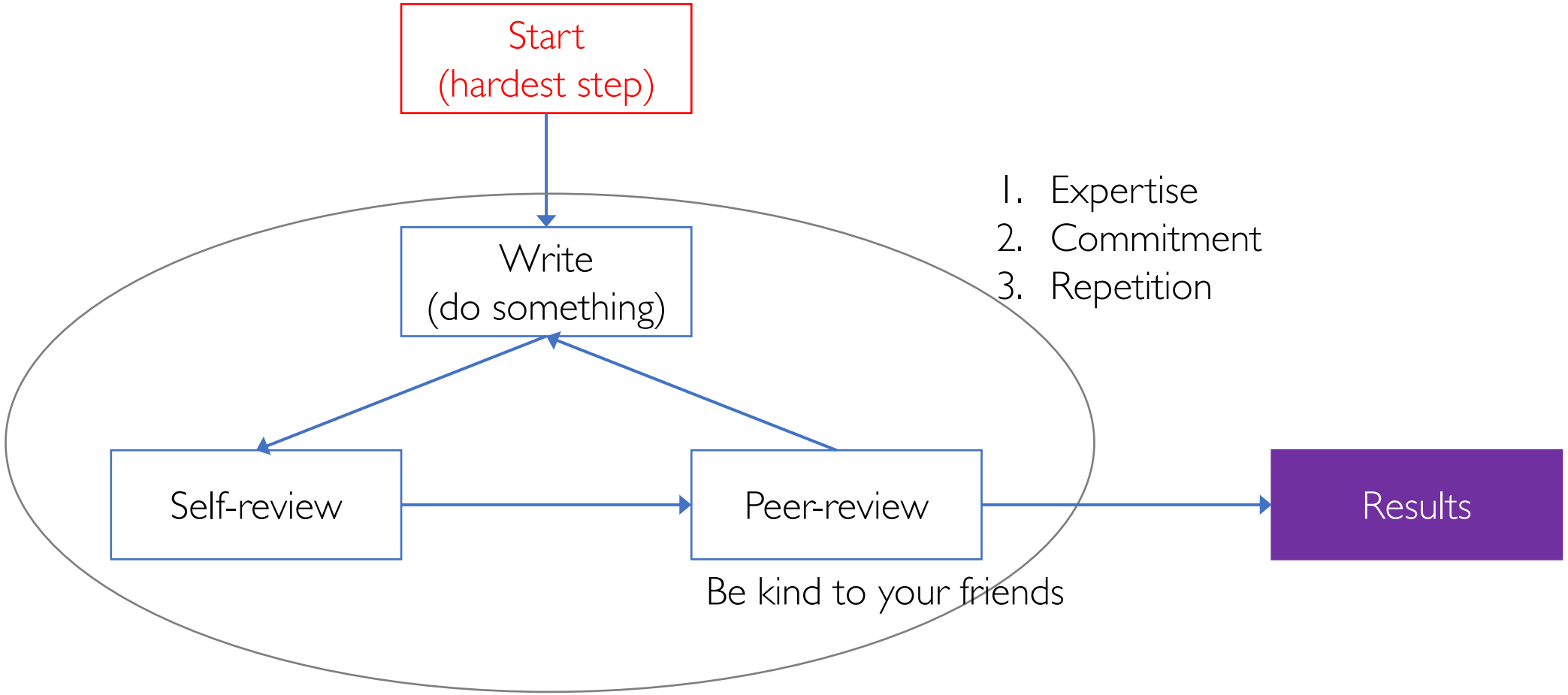
https://wiragotama.github.io/ebook_machine_learning.html

- Problems of Indonesian Students
  - Desperately need high-quality books written in Bahasa Indonesia
  - They like free e-books (no money)
  - Many (if not most) are crazy about discussing "contribution to society" while no action is taken

- Actions taken
  - Writing ML + DL introduction textbook in Bahasa Indonesia (hopefully high in quality)
  - Make it free (contributed to society)

- Why ML + DL
  - ML + DL is (relatively) at more basic-level than NLP and used in many areas of AI
  - Writing exercise

Writing Process



1. Expertise
2. Commitment
3. Repetition

Representing Text

Bag of Words (one-hot-vector) [frequentist]

I would like to request an assistance

Unigram

I	…	would	…	like	…	…	elephant
1	…	1	…	1	…	…	0

Bigram

I would	…	would like	…	request an	…	…	an elephant
1	…	1	…	1	…	…	0

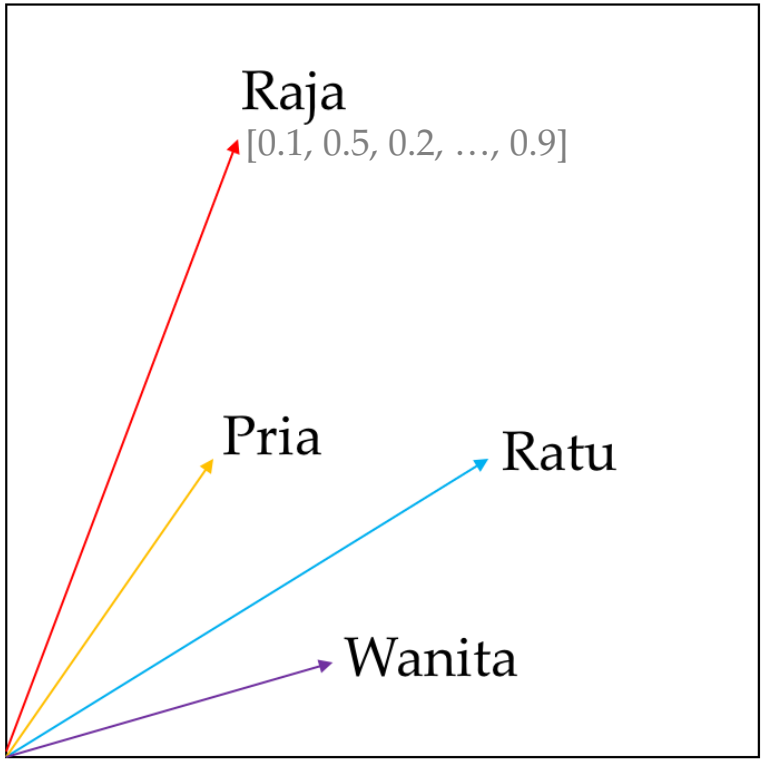
Term Frequency [frequentist]

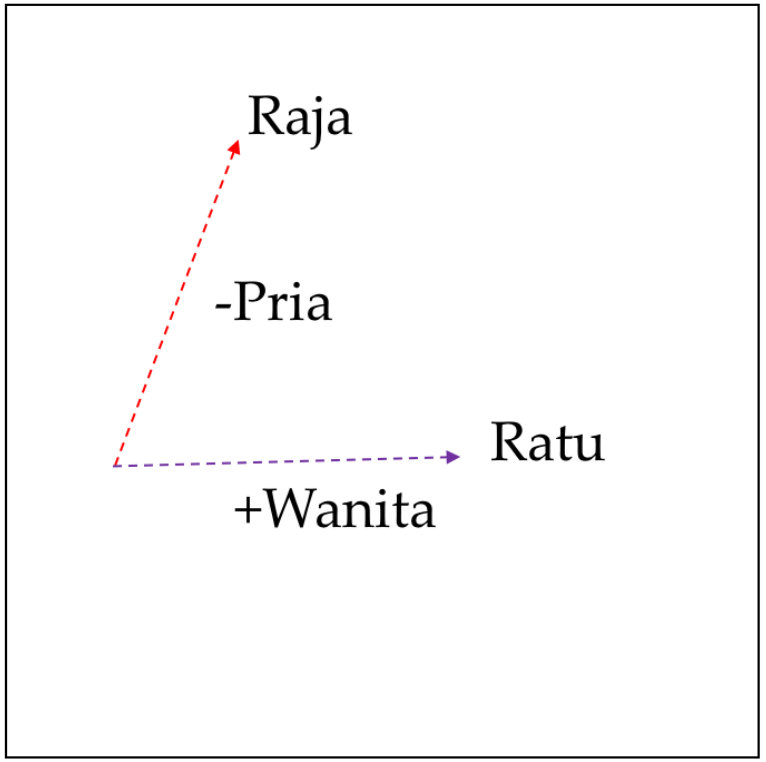
You are what do you think you are

you	…	what	…	think	…	are	elephant
2	…	1	…	1	…	2	0

Representing Meaning [modern]

Mikolov et al. (2013)





Raja - Pria + Wanita = ?

Essay Scoring

Literacy and Scoring

- Literacy: the ability to read and write (Oxford dictionary)

- Essay scoring: predict the quality of writing (using scores)

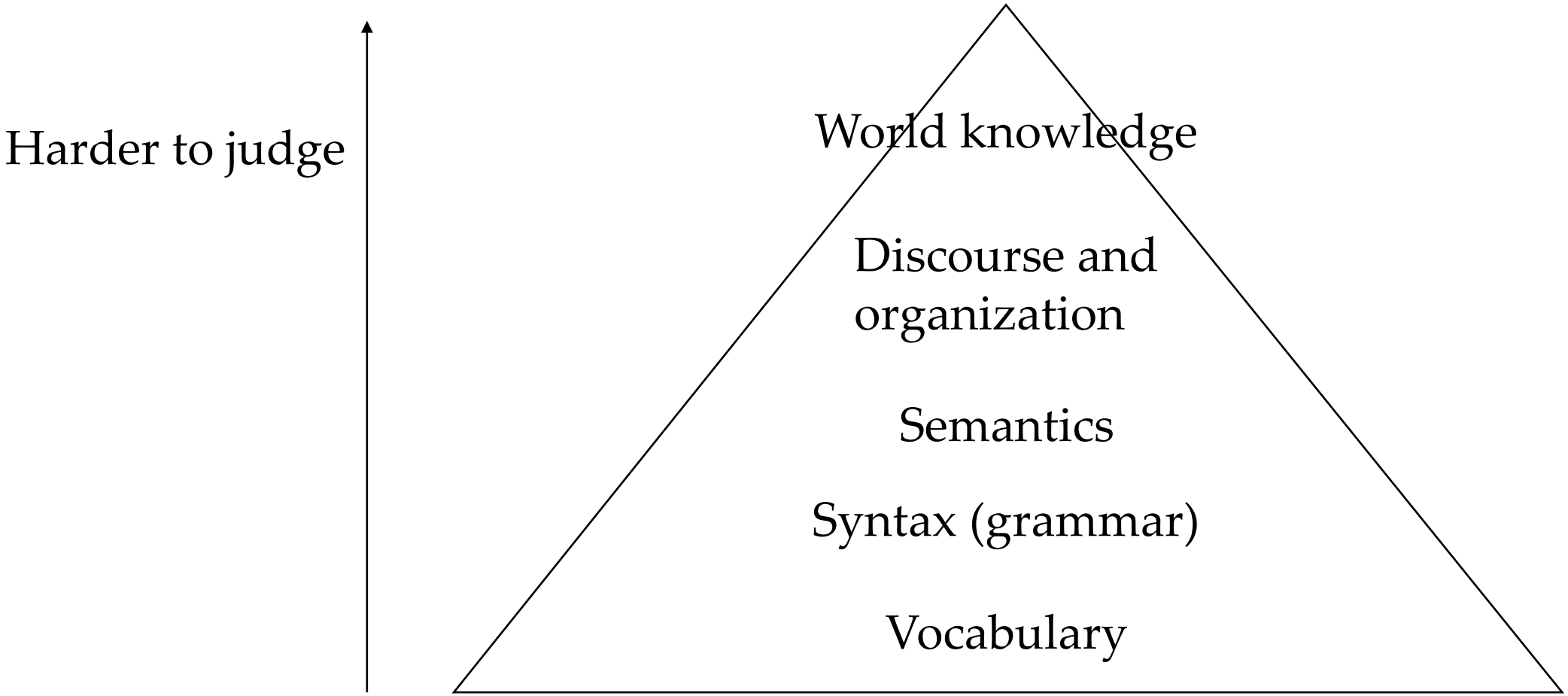
- Broader context
  - Text is a communication tool (source of information)
  - Unclear or low quality writing costs misunderstanding (+time)
  - Standardized testing (TOEFL, IELTS)

Language Education

- The framework of (English) language education is driven by standardized curricula and assessment, e.g., CEFR, Common Core
  - These standards explain comprehensively what language learners have to learn (in an orderly manner) to use language effectively for communication

- There is a strong focus on students' ability to develop an argument and the development of writing assessment and support tools (Burnstein et al., 2013)
  - Teacher vs. students ratio

Judging Text Quality



Judging Text Quality

Aspect	Example
Vocabulary	• I would like to <u>buy</u> a ticket • I would like to <u>purchase</u> a ticket
Grammar	• The man who wears the blue shirt is crazy (grammatical) • The who wears blue shirt man is crazy (ungrammatical)
Semantics	• I would like to eat <u>a boulder</u> • I would like to eat <u>sushi</u>
Discourse and Organization	• Since you said so, let's eat sushi then • Let's eat sushi then, since you said so
World Knowledge	• Obama is the current president of US • Trump is the current president of US

Traditional Readability Formula (Vocabulary)

- Readability: how complex or difficult is the text?

- Flesch Reading Ease (1948)

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllabels}{total\ words}\right)$$

- Flesch-Kincaid (1975)

$$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllabels}{total\ words}\right) - 15.59$$

Types of Essay Scoring

- Holistic Scoring
  - Predict a single score, given a text
  - Somasundaran et al. (2016), etc.
- Aspect based
  - Predict a single score for an aspect (e.g., argumentation, organization), given a text
  - Persing et al. (2010), Persing and Ng (2013), Persing and Ng (2014), etc.
- Ranking
  - Comparing two or more texts, decide which one is the best
  - Barzilay and Lapata (2008), Guinaudeau and Strube (2013), Putra and Tokunaga (2017), etc.

Features (Persing and Ng. 2013)

- N-gram (bag of words) – or word2vec, GloVe, ELMO, BERT, XL-Net, etc.
- Spelling mistakes
- Presence of certain keywords
- POS n-gram (syntactical)
- Semantic roles (think of sentence label)

Why Simple Features Work

Language acquisition

Easier (lower score)	Harder (higher score)
Very good	Extraordinary
Beautiful	Exquisite
Open field	The land is virgin
Complex	Sophisticated

Why Simple Features Work

Syntactical Construct

Easier (lower score)	Harder (higher score)
I need help	I would like to request an assistance
You are beautiful	You have warm eyes like a sunset, sleek skin like a silk

But do essay scoring systems really ``understand" the text?

Evaluating Text Coherence Based on Semantic Similarity Graph

Jan Wira Gotama Putra and **Takenobu Tokunaga**

Presented in TextGraphs-11 (2017)

Evaluating Coherence

Towards Understanding Text Meaning and Structure

- Coherent text is integrated as a whole, rather than a series of independent sentences (Bamberg, 1983; Garing, 2014)

Graph structure

- Every sentence in a coherent text has relation(s) to each other (Halliday and Hasan, 1976; Mann and Thompson, 1988; Grosz et al., 1995; Wolf and Gibson, 2005)

Evaluate coherence through cohesion

- Lexical and semantic (meaning) continuity are indispensable for coherent text (Feng et al., 2014)

Semantic similarity

Coherence (simplified)

Are sentences (or clauses) ordered properly and related to each other (i.e., integrated)?

Idea

- Graph data structure can represent the structure of text and relations among sentences (Wolf and Gibson, 2004; Guinaudeau and Strube, 2013)

- Coherence is achieved through lexical cohesion: **repeated mention of entities.**
  - Disadvantage: cannot capture the relation between related-yet-not identical entities (Li and Hovy, 2014; Petersen et al., 2015)
  - Solution: use distributed representation of words/sentences

Proposed Method (1)

- Formally, text is a graph $G(V, E)$, where
  - V is a set of vertices, v_i represents *i*-th sentence.
  - E is a set of edges, e_{ij} represents relation (cohesion) from *i*-th to *j*-th sentence (weighted & directed).

- Sentences are encoded into their meaning form

  Average of summation of word vectors (distributed representation of words)

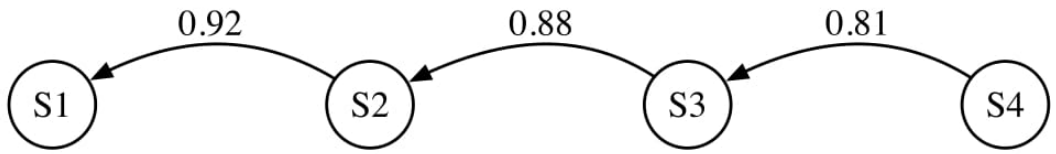
$$\vec{s_i} = \frac{1}{M} \sum_{k=1}^{M} \vec{w_k}$$

- An edge represents cohesion among sentences

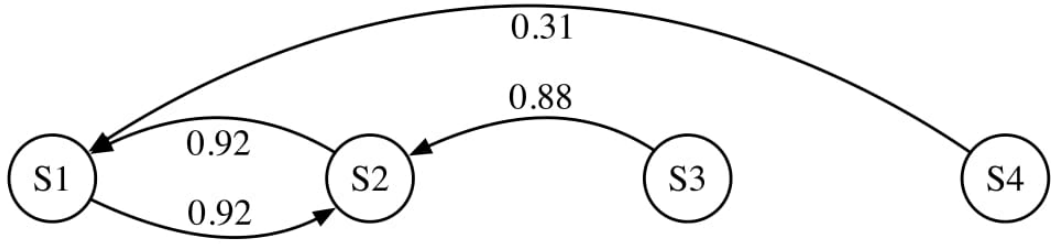
  Establishment of edge is decided as the operation of vectors representation of sentences

Proposed Method (2)

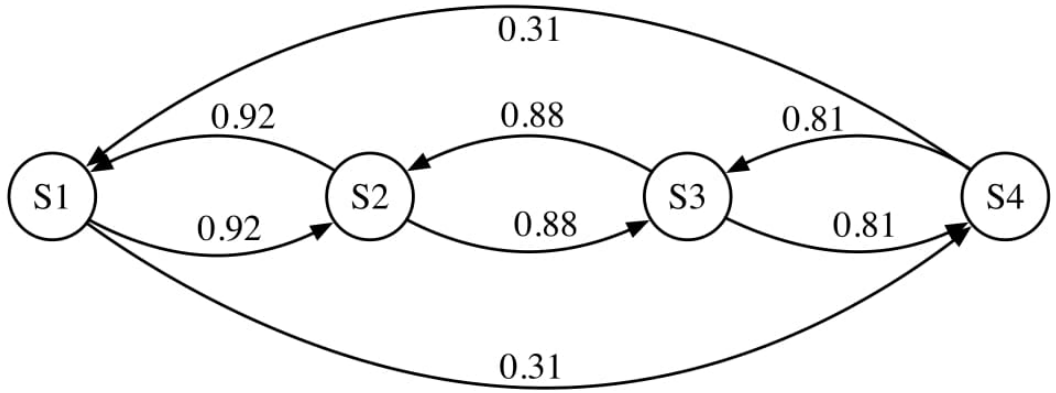
- Preceding Adjacent Vertex (PAV)



- Single Similar Vertex (SSV)



- Multiple Similar Vertex (MSV)



Proposed Method (3)

- An edge is established from the sentence vertex in question to the other vertex with the weight calculated by

$$weight(e_{ij}) = \frac{\text{cos_sim}(\vec{s_i}, \vec{s_j})}{|i - j|}$$

normalization

- Text coherence measure (higher is better) is calculated by averaging the averaged weight of outgoing edges from every vertex in the graph as

$$t_c = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{L_i} \sum_{k=1}^{L_i} weight(e_{ij})$$

vertices

outgoing edges of vertex v_i

Evaluation

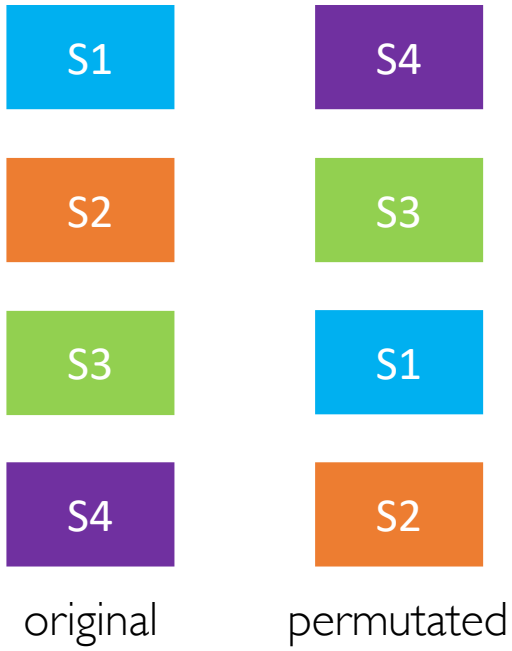
- Task 1: Discrimination (Barzilay and Lapata, 2008)
- Task 2: Insertion (Eisner and Charniak, 2011)

Adversarial samples

- Both tasks evaluate how well the methods in comparing coherence between texts

Task 1: Discrimination Task

- The goal is to compare original vs. permutated text
- Program is considered successful when giving greater score to the more coherent (original) text
- Dataset: 683 WSJ (LDC) texts, 13586 permutations (avg. 24 sentences, 521 tokens)



Task 2: Insertion Task (TOEFL-iBT)

- A text is coherent even without the insertion sentence
- Preservation of coherence is achieved when the question-sentence is inserted in the correct place but disrupt coherence otherwise
- 104 questions from TOEFL-iBT (avg. 7 sentences, 139 tokens)

(A) The raising of livestock is a major economic activity in semiarid lands, where grasses are generally the dominant type of natural vegetation. (B) The consequences of an excessive number of livestock grazing in an area are the reduction of the vegetation cover and trampling and pulverization of the soil. (C) This is usually followed by the drying of the soil and accelerated erosion. (D)

Question: Insert the following sentence into one of (A)-(D)
question sentence = "This economic reliance on livestock in certain regions makes large tracts of land susceptible to overgrazing."

correct answer = B

Result

Method	Accuracy
PAV	**0.774**
SSV	0.676
MSV	0.741
Unsupervised baseline (Guinaudeau and Strube, 2013)	0.725
Supervised baseline (Barzilay and Lapata, 2008)	**0.845**

Discrimination Task

Method	Accuracy
PAV	**0.356**
SSV	0.346
MSV	0.327
Unsupervised baseline (Guinaudeau and Strube, 2013)	0.260
Supervised baseline (Barzilay and Lapata, 2008)	**0.346**

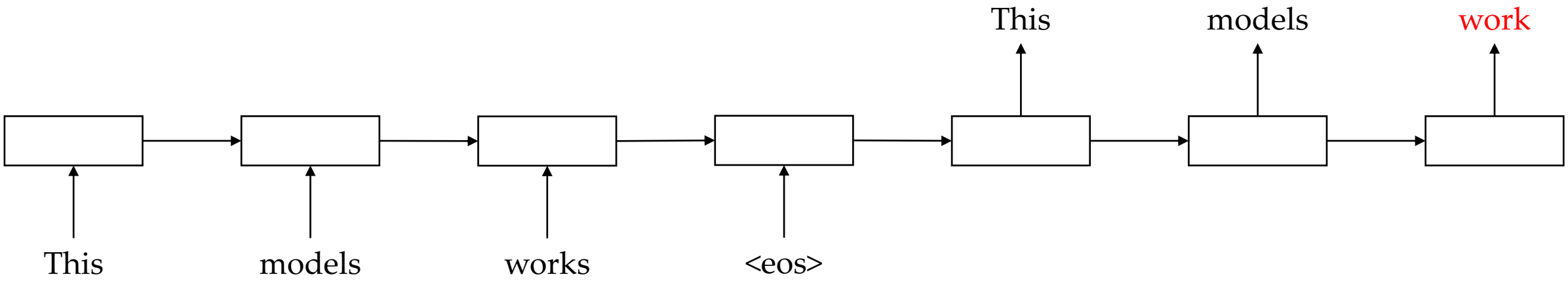
Insertion Task

Limitations

- **Writing assessment tool**
  - Given an essay, output the score for the essay (Persing and Ng, 2014)
  - Automatic essay scoring tools may provide score-associated templated-feedback, e.g., "*essay occasionally wanders off topic*" (score 3 out of 4)
  - Templated-feedback is ineffective for learning since students need to know what their mistakes precisely are
  - *"We can diagnose your diseases with high accuracy, but we cannot cure your diseases"*

- **Writing support tool**
  - Grammarly®, automatic text revision (from grammatical perspective)
  - Tracking revision in writing drafts and automatically predict revision purposes (Zhang and Litman, 2014) → Students can recognize their weaknesses

Grammatical Error Correction



- Schmaltz et al. (2016): sentence-level grammatical error identification and correction
- Cohen et al. (2018): can RNN learn regular grammar?

Discourse-based Essay Improvement

Assessment vs. Improvement

- Assessment
  - Diagnosing the disease

- Improvement
  - Grammatical improvement has been gaining industrial-level performance (e.g., Grammarly), so it is the time to move towards structure-based improvement
  - Let's cure more diseases

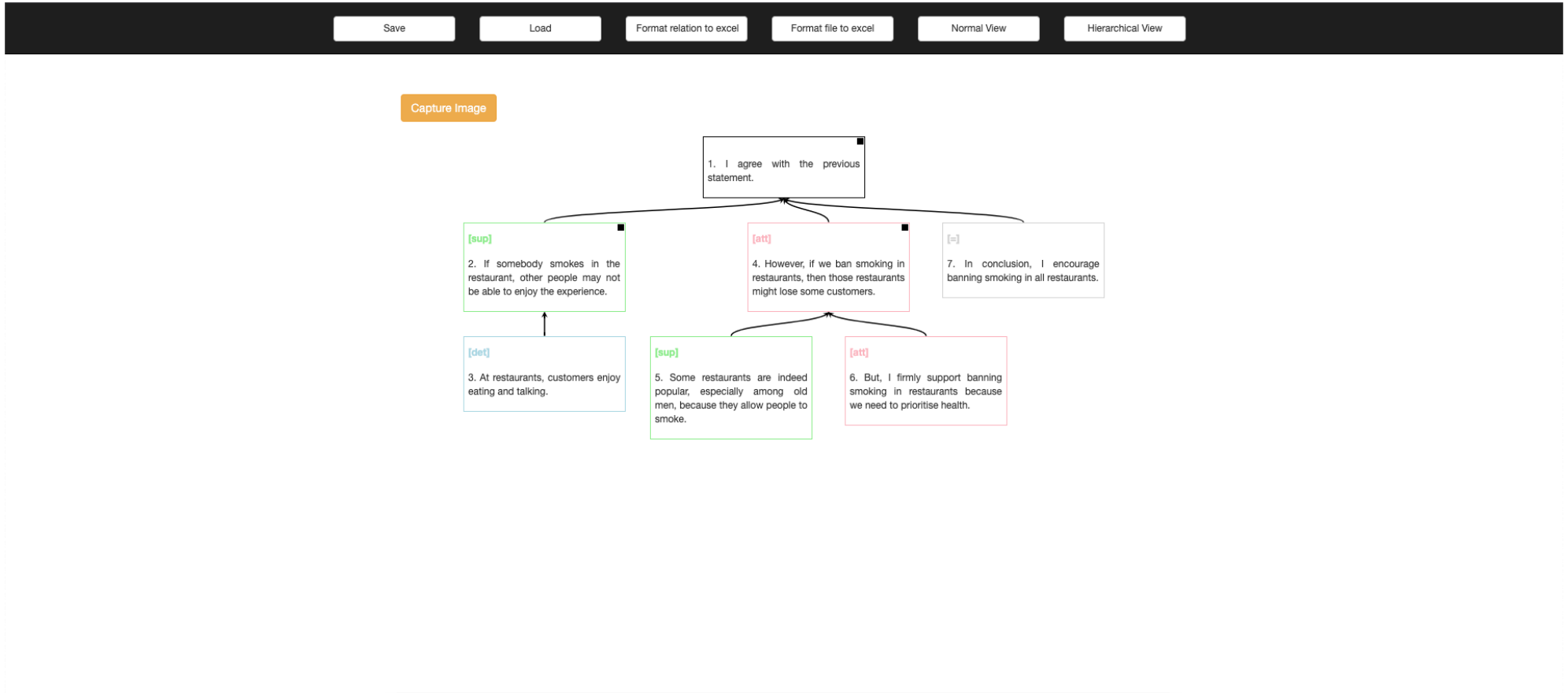
[LIVE ONLY]

[LIVE ONLY]

Annotation Tool TIARA

https://github.com/wiragotama/TIARA-annotationTool

Annotation Tool TIARA



https://github.com/wiragotama/TIARA-annotationTool

Conclusion

- The ability to diagnose a disease is extremely important, but we also need to cure the disease afterwards (current direction)

- Still long way to go (curing the disease)

- Download (the slide): https://wiragotama.github.io/talk.html

References

- Mikolov et al. 2013. Distributed representations of words and phrases and their compositionality. In NIPS.
- Burnstein et al. 2013. Holistic Annotation of Discourse Coherence Quality in Noisy Essay Writing. In Dialogue and Discourse.
- Persing et al. 2010. Modeling Organization in Student Essays. In EMNLP.
- Persing and Ng. 2013. Modeling Thesis Clarity in Student Essay. In ACL.
- Persing and Ng. 2014. Modeling Prompt Adherence in Student Essays. In ACL.
- Somasundaran et al. 2016. Evaluating Argumentative and Narrative Essays Using Graphs. In COLING.
- Zhang and Litman. 2015. Annotation and classification of argumentative writing revisions. In BEA.
- Garing. 2014. Coherence in argumentative essays of first year college of liberal arts students at de la salle university. DLSU Research Congress.
- Grosz et al. 1995. Centering: A framework for modeling the local coherence of discourse. In Computational Linguistics.
- Bamberg. 1983. What makes a text coherent. In College Composition and Communication.
- Guinaudeau and Strube. 2013. Graph-based local coherence modeling. In ACL
- Petersen et al. 2015. Entropy and graph based modelling of document coherence using discourse entities: An application to IR. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval.
- Wolf and Gibson. 2005. Representing discourse coherence: A corpus-based study. Computational Linguistics.

References

- Li and Hovy. 2014. A model of coherence based on distributed sentence representation. In EMNLP
- Eisner and Charniak. 2011. Extending the entity grid with entity-specific features. In ACL-HLT.
- Halliday and Hasan. 1976. Cohesion in English. Longman, Singapore.
- Barzilay and Lapata. 2008. Modeling local coherence: Entity based approach. In Computational Linguistics.
- Feng et al. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In COLING.
- Mann and Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. In Text.
- Schmaltz et al. 2016. Sentence-level grammatical error identification as sequence-to-sequence correction. In BEA.
- Cohen et al. 2018. Inducing regular grammars using recurrent neural networks. In L&R workshop, ICML & IJCAI.

- Free icons: https://stampo.fun/