台灣科技大學資訊工程系
資訊檢索

基於機器學習之文本可讀性研究

Hou-Chiang Tseng, Yao-Ting Sung, and Berlin Chen

過去的十年...

- Publication
  - Journal Papers：國內外7篇
  - Conference Papers：國內外18篇
    - Best Poster Presentation Award , Weng S. Y., Tseng, H. C., Sung, Y. T. & Chen B., (2019). A Hierarchical Encoding Framework for Text Readability Prediction. The 2019 Conference on Computational Linguistics and Speech Processing (ROCLING 2019), 334-342. Taiwan. (Poster)

- 發明名稱：發散思維測驗自動評分方法及系統。發明人：宋曜廷、張國恩、曾厚強、鄭皓心。(108)智專二(二) 04267 字第10820395260號

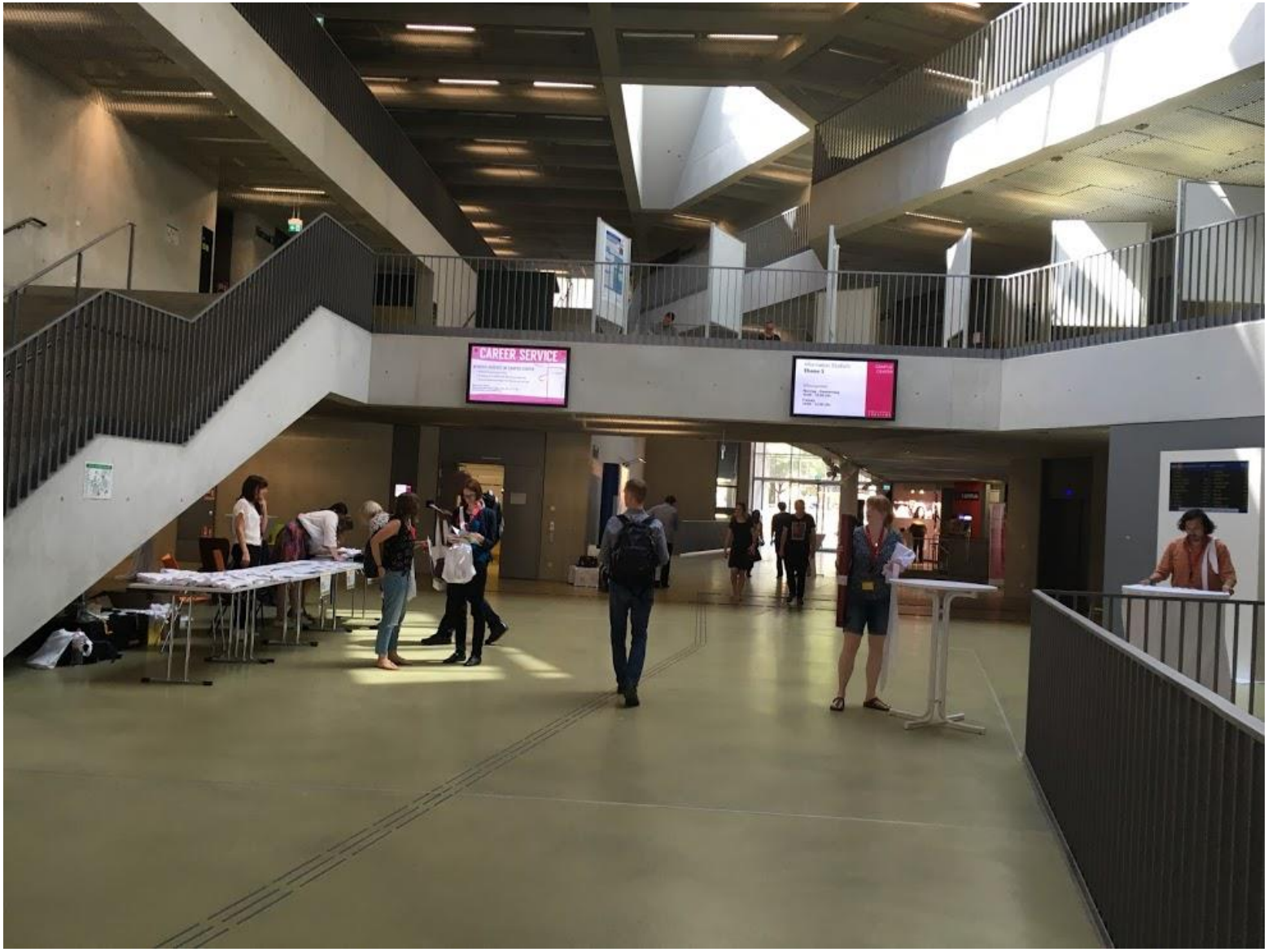
- 中文手寫作文自動辨識、批改與教學回饋系統。108年度財團法人中技社AI創意競賽-AI教育組第一名。成員：宋曜廷、張道行、曾厚強、李少榮

- 科技部補助博士生國外研討會發表 5次

德國 – 卡塞爾


Classification of Text Readability Based
1. Introduction
2. The purposed method
UNIKASSEL

德國 – 卡塞爾



中國 – 華中師範

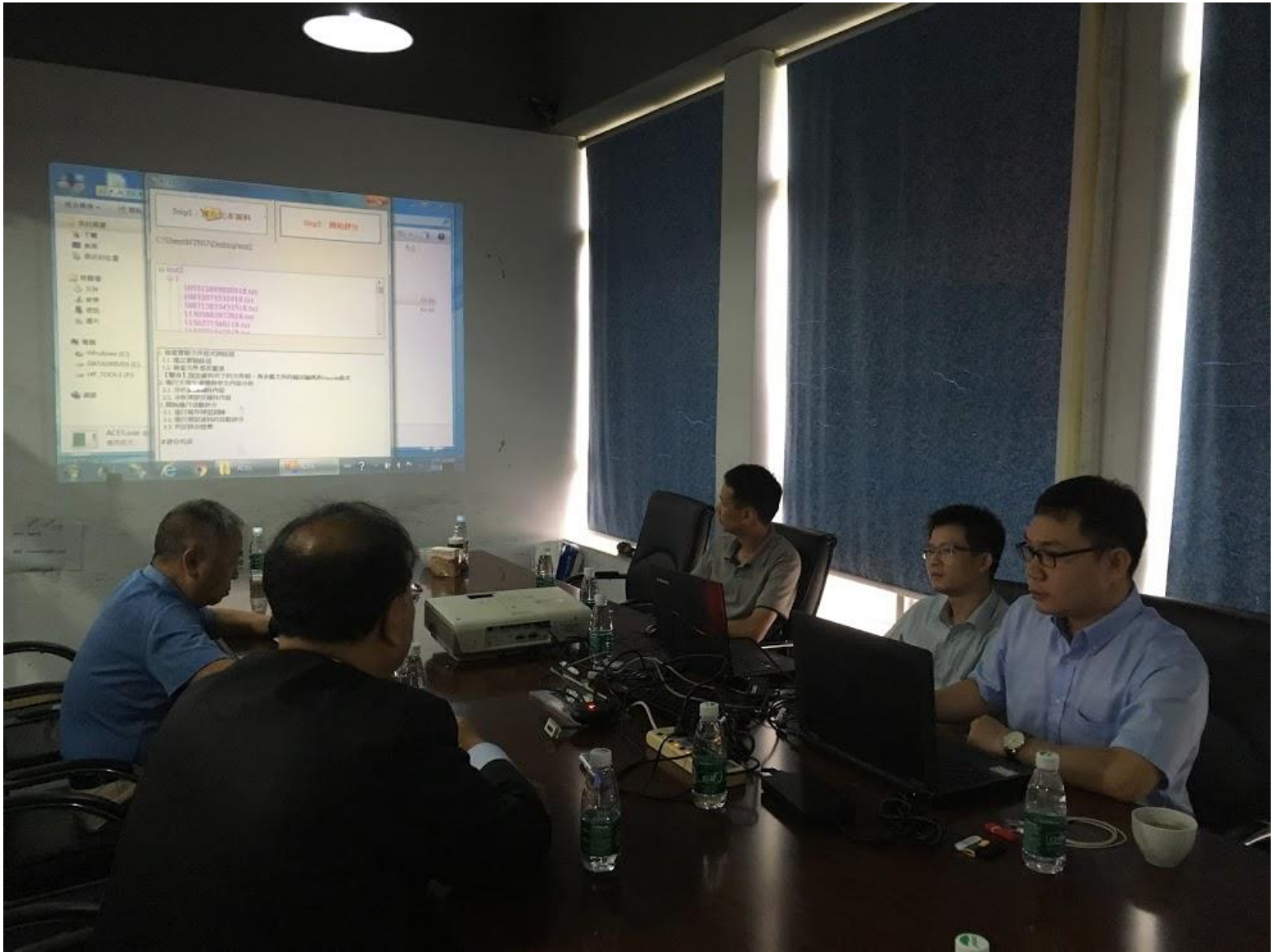

NATIONAL ENGINEERING RESEARCH CENTER FOR E-LEARNING
国家数字化学习工程技术研究中心

美國 – 匹茲堡


6th International Workshop on
Advanced Learning Sciences
IWALS 2018
Perspectives on the Learner:
Cognition, Brain, and Education
June 6-8, 2018
LRDC

美國 – 匹茲堡

中國 – 深圳



挪威

挪威


2019
ICITL
International Conference of
Innovative Technologies and Learning
Dec. 2-5, 2019
Hosted by
Co-Hosted by
Co-Sponsored by
Springer

英國 -倫敦大學學院

英國 – 牛津大學

英國 – BETT

- Bett Show is a global series of education shows.

牛頓 Isaac Newton 說：

如果我看得比其他人遠，是因為我站在巨人的肩膀上。

If I have seen further than others, it is by standing upon the shoulders of giants.

Introduction

- Readability refers to what degree a reading material can be understood by its readers
  - When readers are reading a text with high readability, they will achieve better comprehension and learning retention

(Dale and Chall 1949; Klare 2000)

- In order to facilitate readers to digest and comprehend documents, researchers have long been developing readability models that can automatically and accurately estimate text readability

Introduction & Related Work (cont.)

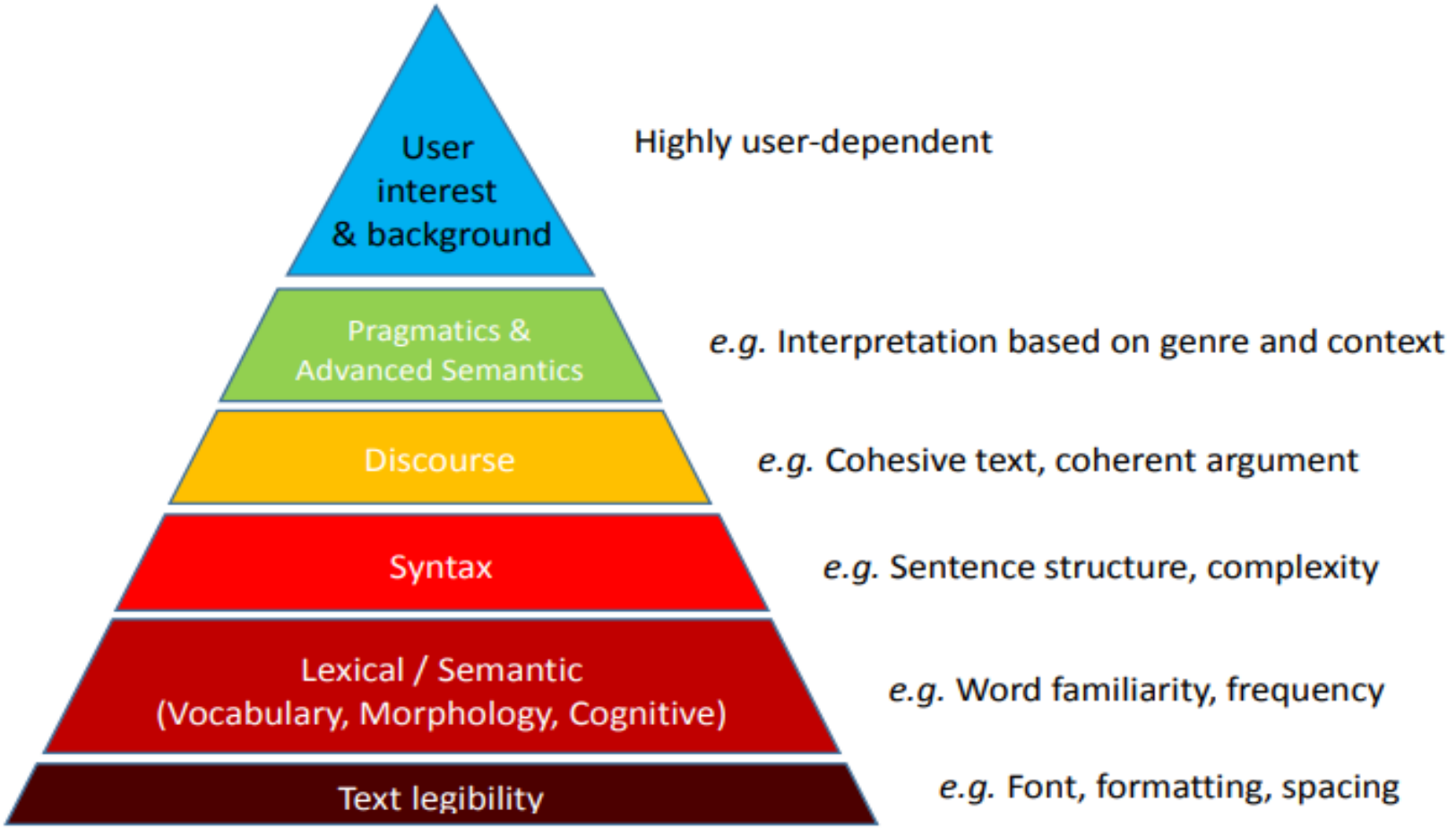
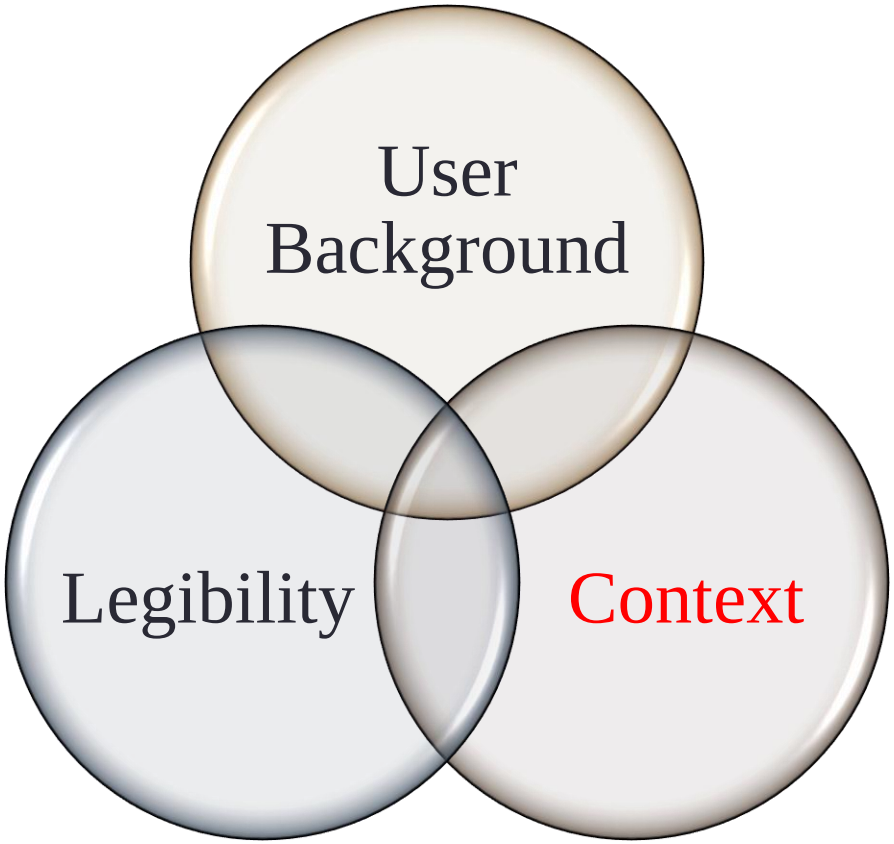


Figure 2: Key aspects of text readability, ordered from lowest level (text legibility) to highest level (user interest and background). These levels are one way to categorize the types of features used by text readability measures for automated assessment.

(Collins-Thompson 2014)

Introduction & Related Work (cont.)



- **Font**
- **Formatting**

- **Linguistics Features**
- **Topic**

Introduction & Related Work (cont.)

- In order to facilitate readers to digest and comprehend documents, researchers have long been developing readability models that can automatically and accurately estimate text readability.

Languages：

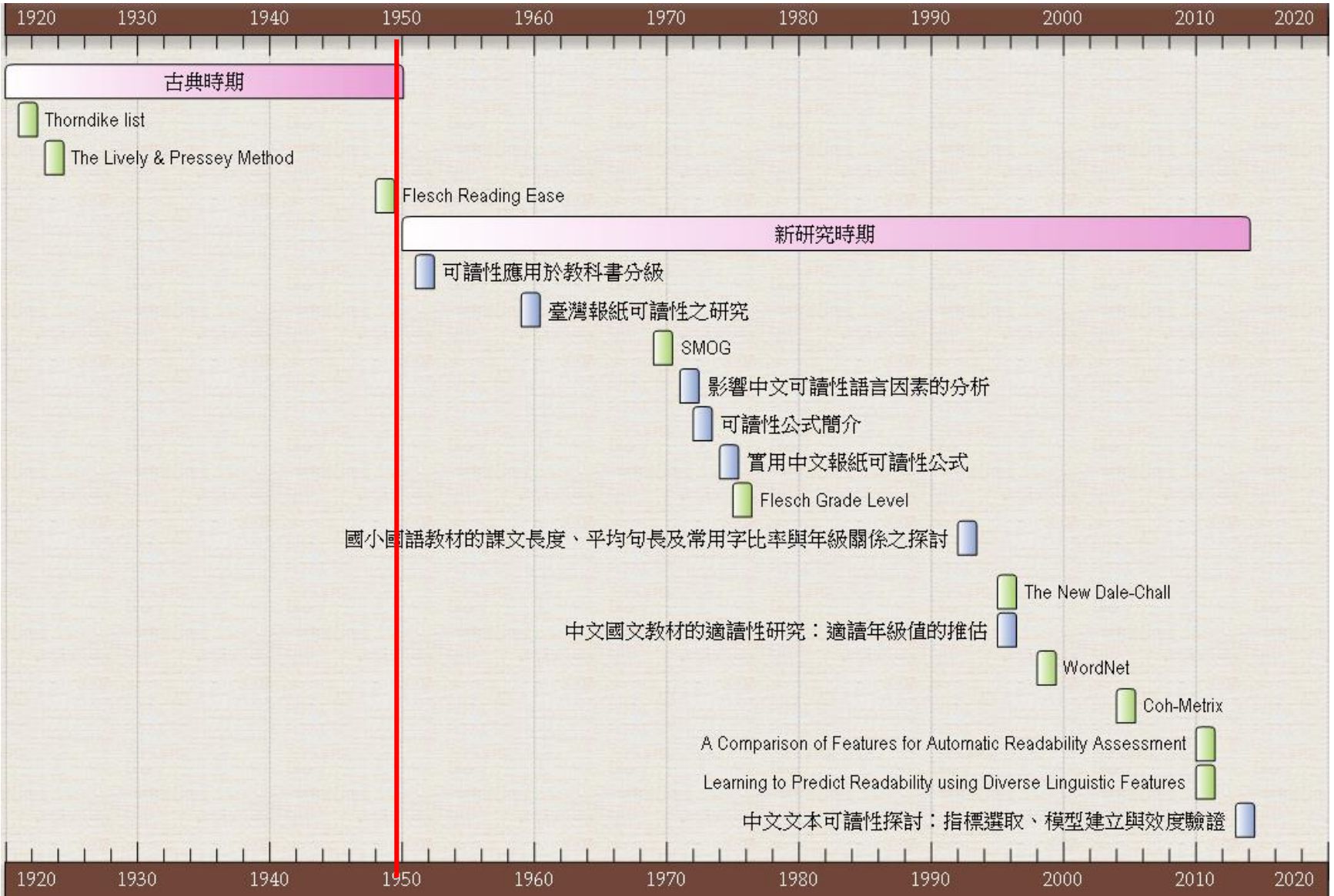
- English (Graesser *et al*., 2004; Pitler and Nenkova 2008)
- French (Todirascu, François, Gala, Fairon, Ligozat and Bernhard 2013; Dascălu 2014)
- Swedish (Pilán, Volodina and Johansson 2014)
- German (Vor Der Brück and Hartrumpf 2007)
- Chinese (Chen, Chen and Cheng 2013; Sung *et al*., 2016)
- Japanese (Sato, Matsuyoshi and Kondoh 2008)

Domains:

- American courts (Dubay 2004)
- Biology textbooks (Belden and Lee 1961)
- Health education messages (Freimuth 1979; Santos, Daar, Badeau and Lei 2017)
- Business communication textbooks (Razek and Cone 1981)
- Economics textbooks (Gallagher and Thompson 1981; McConnell 1982)
- Adult learning materials (Taylor and Wahlstrom 1999)

Related Work - The Development of Readability Studies

Traditional Readability Formulas



Introduction & Related Work (cont.)
Traditional Readability Formulas

Flesch Reading Ease (Flesch, 1948)

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

Dale-Chall (Dale & Chall, 1948)

$$0.1579\left(\frac{difficult\ words}{words} \times 100\right) + 0.0496\left(\frac{words}{sentences}\right)$$

SMOG (McLaughlin, 1969)

$$\text{grade} = 1.0430 \times \sqrt{number\ of\ polysyllables \times \frac{30}{number\ of\ sentences}} + 3.1291$$

- Graesser, Singer and Trabasso (1994) have pointed out that traditional formulas using linguistic features fail to reflect the real process of reading, with semantics and syntax being merely shallow linguistic features, unable to reflect the cohesion of the text.

- Collins-Thompson (2014) also noted the emphasis of traditional readability formulas on the shallow information of a text while overlooking important deeper features.

Related Work - The Development of Readability Studies Traditional Readability Formulas (cont.)

Table 2
Readability Estimates and Corresponding Grade Levels

(Begeny, J. C., and Greene, D. J. 2014.)

Passage	DORF	Dale-Chall	F-K	FOG	Forcast	Fry	Lexile	PSK	SMOG	Spache
A	1st	1.0 (1st)	0.9 (1st)	3.3 (3rd)	7.3 (7th)	1 (1st)	350 (1st)	3.7 (3rd)	4 (4th)	2.0 (2nd)
B	1st	1.0 (1st)	2.2 (2nd)	4.5 (4th)	8 (8th)	2 (2nd)	470 (2nd)	4.1 (4th)	5 (5th)	2.4 (2nd)
C	2nd	3.0 (3rd)	3.2 (3rd)	5.0 (5th)	7.8 (7th)	3 (3rd)	670 (3rd)	4.2 (4th)	4 (4th)	2.8 (2nd)
D	2nd	4.0 (4th)	3.7 (3rd)	5.3 (5th)	8.6 (8th)	4 (4th)	750 (4th)	4.4 (4th)	5 (5th)	2.7 (2nd)
E	3rd	4.0 (4th)	3.8 (3rd)	6.3 (6th)	7.9 (7th)	4 (4th)	750 (4th)	4.4 (4th)	6 (6th)	3.1 (3rd)
F	3rd	4.0 (4th)	5.7 (5th)	7.7 (7th)	9.2 (9th)	7 (7th)	880 (4th)	5.1 (5th)	8 (8th)	3.2 (3rd)
G	4th	5.5 (5th)	5.7 (5th)	8.4 (8th)	9.5 (9th)	7 (7th)	880 (4th)	5.2 (5th)	8 (8th)	3.4 (3rd)
H	4th	4.0 (4th)	6.8 (6th)	8.0 (8th)	9.3 (9th)	7 (7th)	1080 (5th)	5.2 (5th)	7 (7th)	3.8 (3rd)
I	5th	5.5 (5th)	6.3 (6th)	8.8 (8th)	9 (9th)	7 (7th)	900 (4th)	5.1 (5th)	9 (9th)	3.5 (3rd)
J	5th	5.5 (5th)	8.1 (8th)	10.4 (10th)	9.8 (9th)	8 (8th)	1140 (6th)	5.6 (5th)	9 (9th)	3.9 (3rd)
K	6th	5.5 (5th)	8.7 (8th)	10.2 (10th)	9.3 (9th)	8 (8th)	1180 (6th)	5.6 (5th)	9 (9th)	4.4 (4th)
L	6th	5.5 (5th)	8.9 (8th)	9.6 (9th)	11.6 (11th)	10 (10th)	1050 (5th)	6.2 (6th)	10 (10th)	3.9 (3rd)
Accuracy：		41.67%	16.67%	0%	0%	8.3%	25%	25%	0%	33.33%

Overall findings of this study suggested that only one of the readability formulas (Dale-Chall) demonstrated to be a valid and consistent indicator of text difficulty when compared to a commonly used proxy of elementary-aged students' reading ability (i.e., ORF).

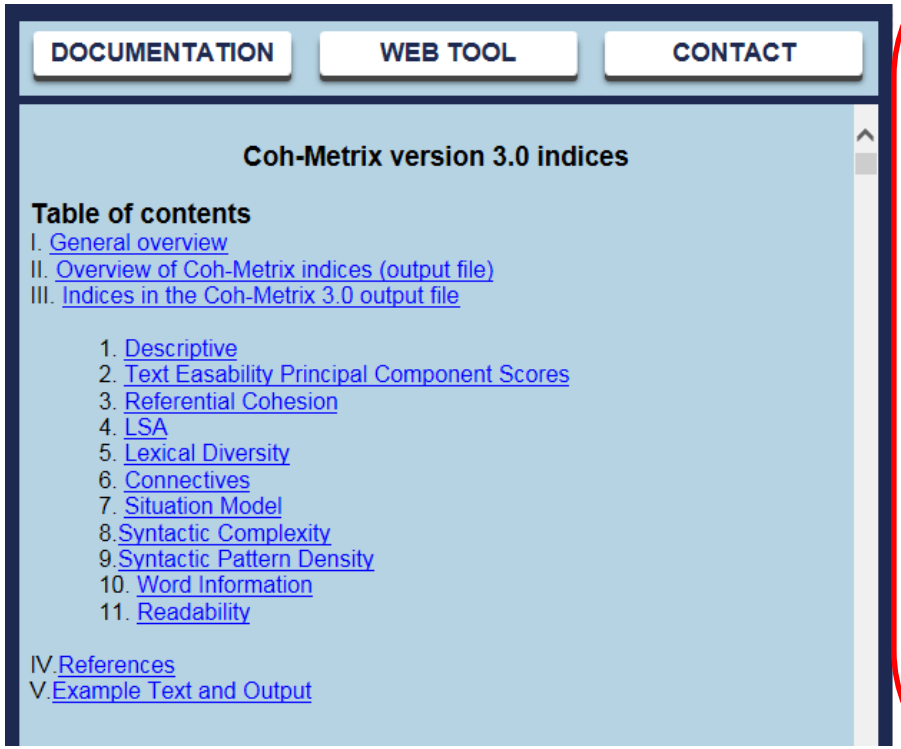
Related Work - The Development of Readability Studies Traditional Readability Formulas (cont.)

The popular readability formulas, namely the Flesch Reading Ease (Flesch, 1943) and the Flesch-Kincaid Grade Level (Kincaid & McDaniel, 1974) formulas rely primarily on word length (number of letters or syllables) and sentence length to assess difficulty level. Consequently, textbook publishers pressured to target textbooks to a certain grade can lower textbooks' grade level estimates by reducing word and sentence lengths. This approach results in short, choppy, sentences with minimal cohesion. However, texts with shorter sentences paradoxically run the risk of being more difficult to comprehend, particularly for readers with low domain knowledge and low reading proficiency. Why? Because there are fewer linguistic cues of cohesion that specify how the sentences should be conceptually related.

McNamara, Louwerse, & Graesser, A. C. (2002).

Research Objectives

English Graesser *et al.* (2004)



Chinese Sung *et al.* (2015)

- Scholars have in recent years begun to conduct research using multiple linguistic features.
- The readability indices that include four main categories: word, syntax, semantics, and cohesion.

Experiment

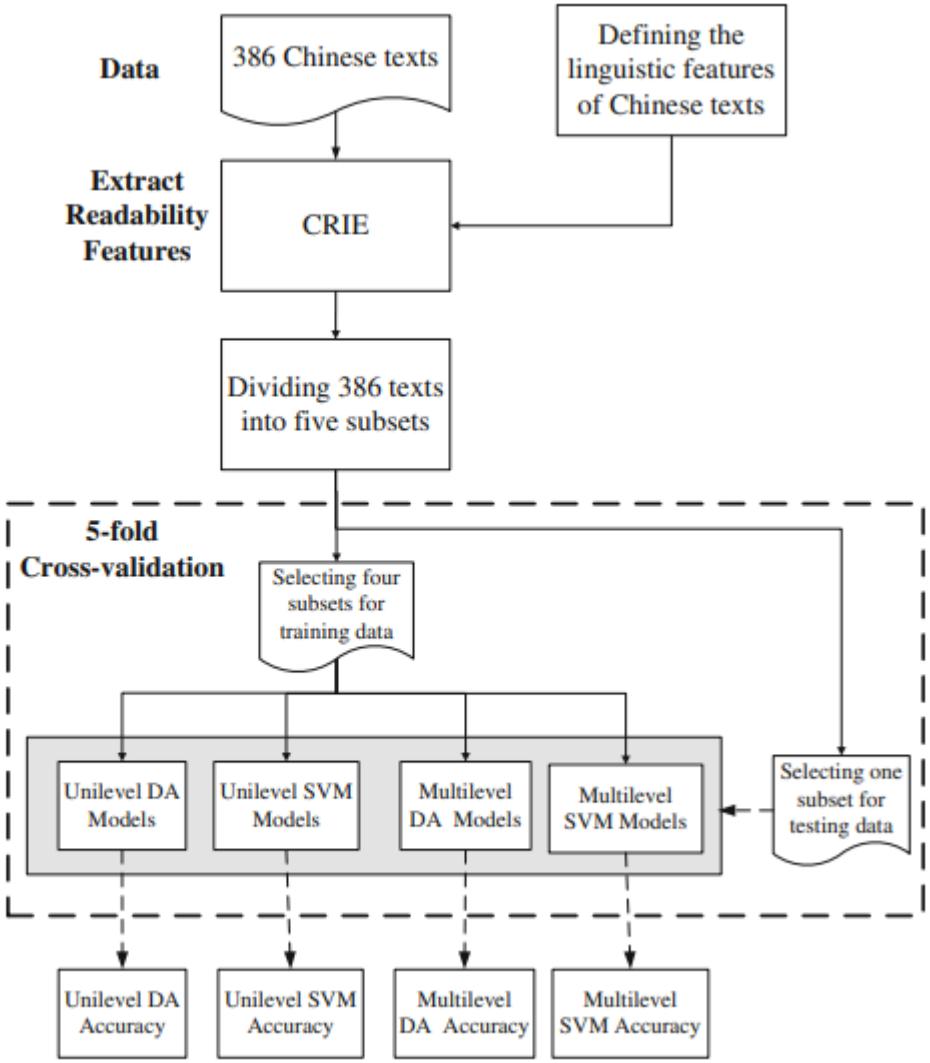


Fig. 3 Procedure for constructing and validating the models

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. Behavior research methods, 47(2), 340-354.

Experiment Results: materials

The experimental materials for this study were adopted from the 1 to 6 grade textbooks published in 2009 by three major publishers in Taiwan, Nan I (Nan I 2009), Han Lin (Han Lin 2009), and Kang Hsuan (Kang Hsuan 2009).

Grade	1	2	3	4	5	6
Chinese	24	67	61	71	69	70

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. Behavior research methods, 47(2), 340-354.

Experiment Results

Table 3 Prediction accuracies (%) of the unilevel and multilevel DA and SVM models

Level	Fold					Average
	1	2	3	4	5	
Word						
DA	61.33	57.33	53.33	62.67	59.30	58.79
SVM	64.00	65.33	62.67	72.00	61.63	65.13
Semantics						
DA	53.33	58.67	49.33	60.00	61.63	56.59
SVM	54.67	64.00	50.67	58.67	61.63	57.93
Syntax						
DA	58.67	50.67	50.67	46.67	52.33	51.80
SVM	56.00	52.00	54.67	53.33	51.16	53.43
Cohesion						
DA	44.00	34.67	45.33	52.00	47.67	44.73
SVM	45.33	37.33	38.67	52.00	46.51	43.97
Multilevel						
DA	56.00	60.00	56.00	58.67	58.14	57.76
SVM	70.67	77.33	65.33	73.33	72.09	71.75

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. Behavior research methods, 47(2), 340-354.

Introduction & Related Work (cont.)
Domain-Specific Features

- Yan and his colleagues (2006) pointed out in no uncertain terms that when calculating using terms included in the Medical Subject Headings (MeSH) database, features used in linguistic feature formulas, such as the number of syllables and length of words, do not correlate with the difficulty of medical terminology.

- "Shock"
  - "Very surprised"
  - "A life-threatening condition that occurs when the body is not getting enough blood flow, which obstructs microcirculation and results in the lack of blood and oxygen in vital organs"

    (Cecconi, De Backer, Antonelli, Beale, Bakker, Hofer, Jaeschke, Mebazaa, Pinsky, Teboul, Vincent and Rhodes 2014: 1796)

?

Introduction & Related Work (cont.)
Domain-Specific Features

- Yan and his colleagues (2006) pointed out in no uncertain terms that when calculating using terms included in the Medical Subject Headings (MeSH) database, features used in linguistic feature formulas, such as the number of syllables and length of words, do not correlate with the difficulty of medical terminology.

- Ontology
  - Yan, X., Song, D., and Li, X. (2006). Concept-based document readability in domain-specific information retrieval, In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, pp. 540-549.
- Latent Semantic Analysis (LSA)
  - Tseng, H. C., Chen, B., Chang, T. H., & Sung, Y. T. (2019). Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering*, 25(3), 331-361.
- Language Model
  - Sato, S., Matsuyoshi, S., and Kondoh, Y. (2008). Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, pp. 654-660.

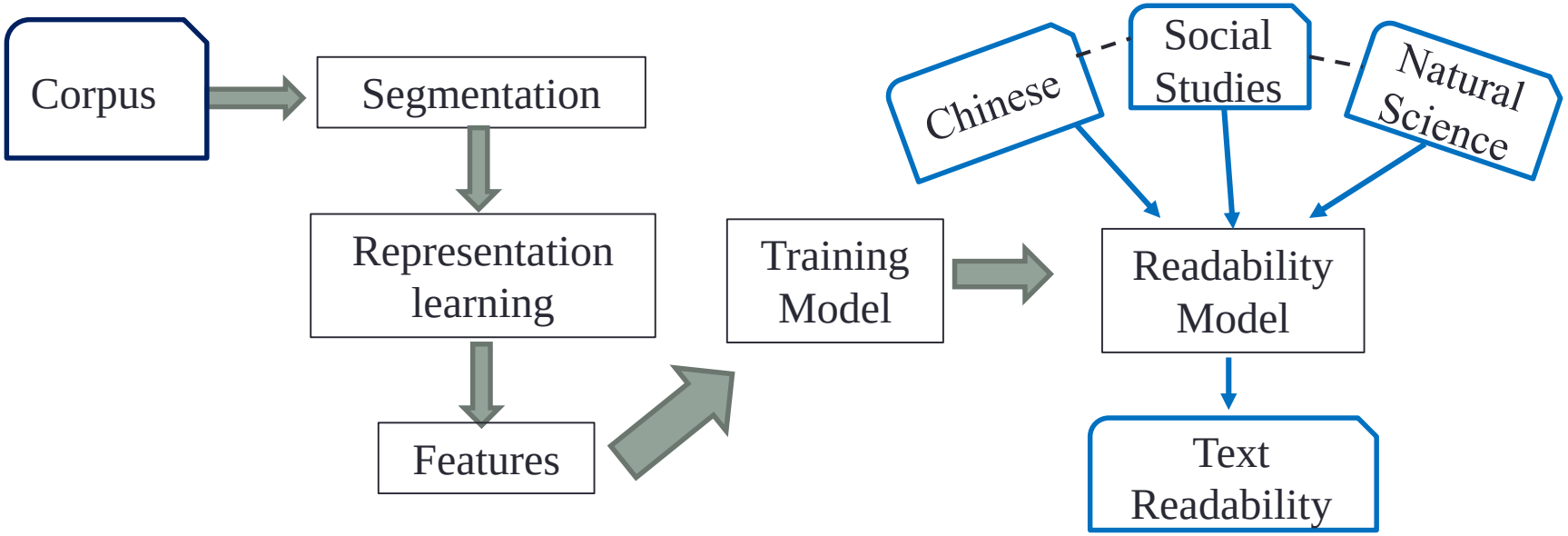
Research Objectives (cont.)

- Yan, Song and Li (2006) found that the number of syllables and word length of a medical term were not related to the term's level of difficulty.
  - Generalization

- Developing the handcrafted features is not only labor-intensive and time-consuming, but also expertise demanding.

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

Research Objectives (cont.)



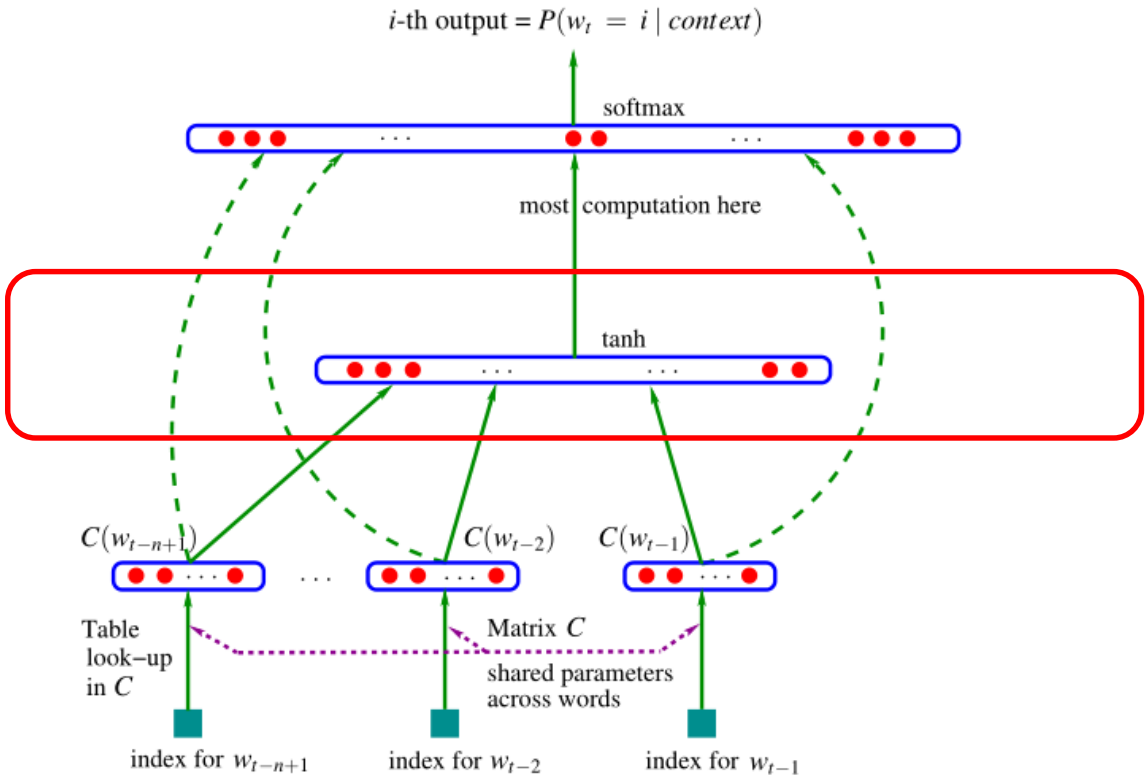
Experiment Results: materials

- This paper used a total of 4,648 texts as the experimental data, which were selected from approved Chinese, social studies and natural science textbooks for grades 1-12 published in 2009 by three major Taiwan publishers, Han-Lin, Kang-Hsuan and Nan-I, all of which were compiled by experts following the curriculum guidelines.

	1	2	3	4	5	6	7	8	9	10	11	12
Social Studies	0	0	80	74	85	81	389	407	325	340	331	270
Natural Science	0	0	72	67	67	62	172	175	157	211	355	295
Chinese	24	67	61	71	69	70	37	34	28	84	41	47

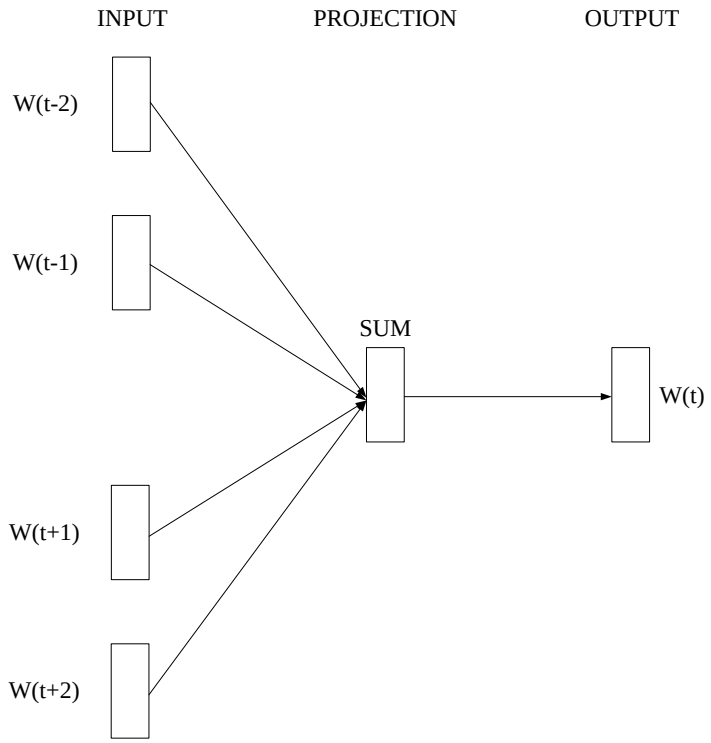
Word2vec

- Word2vec is one of the word vector representation techniques, also known as word representation or word embedding, was first proposed by Hinton in 1986 (Hinton 1986). In 2003, Bengio proposed a training framework called Feed-forward Neural Network Language Model (FFNNLM), which obtained word vector representation through the neighboring relationship of words in a text (Bengio *et al.*, 2003).

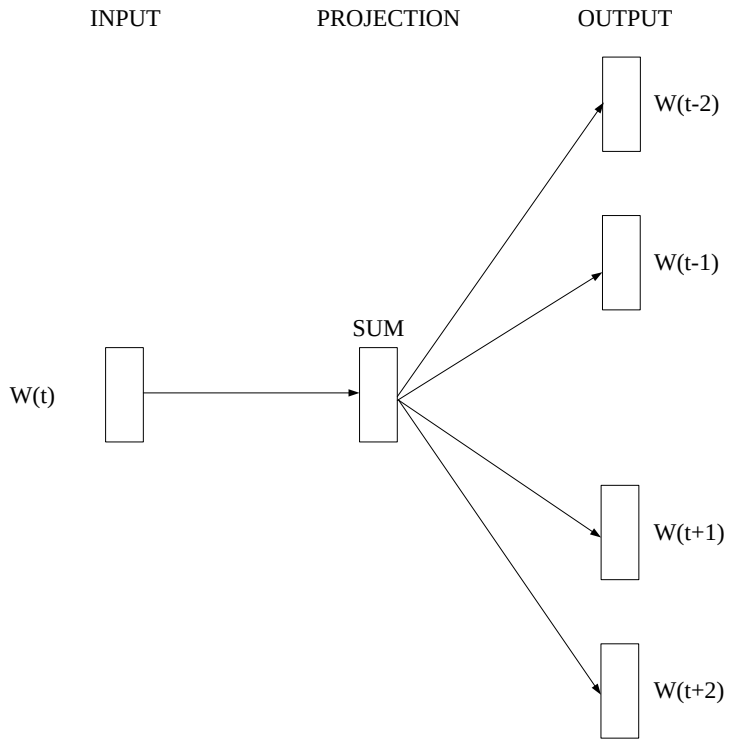


Word2vec (cont.)

- Word2vec (Mikolov *et al.*, 2013), recently released by Google, can be seen as a follow-up method to FFNNLM. However, different from the FFNNLM framework, Word2vec eliminates the nonlinear hidden layer, which is the most time-consuming component in FFNNLM training, retaining only the input layer, projection layer and output layer, streamlining the framework.



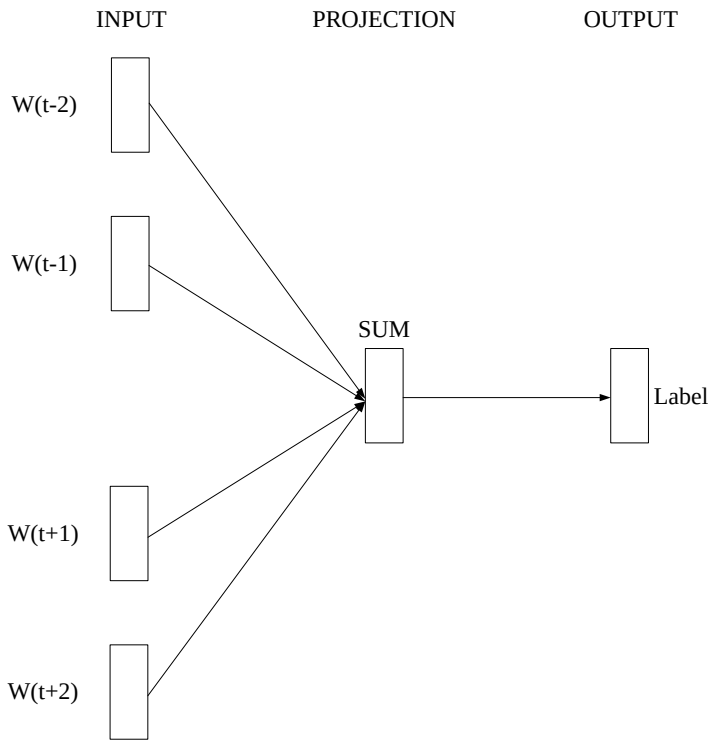
(a) CBOW Training Algorithm.



(b) Skip-gram Training Algorithm.

fastText

- In 2017, Facebook AI research proposed fastText, which is essentially an extension of Word2vec (Mikolov *et al.*).
- FastText architecture is similar to the Word2vec, where the middle word is replaced by a label. (Supervised semantic space)



(a) CBOW Training Algorithm.

Distributed Representation Learning: *StarSpace*

- StarSpace is a general-purpose embedding model brought by Facebook AI Research in 2017 (Wu *et al.*, 2018).
  - Document classification task
  - Text ranking
  - Image recognition

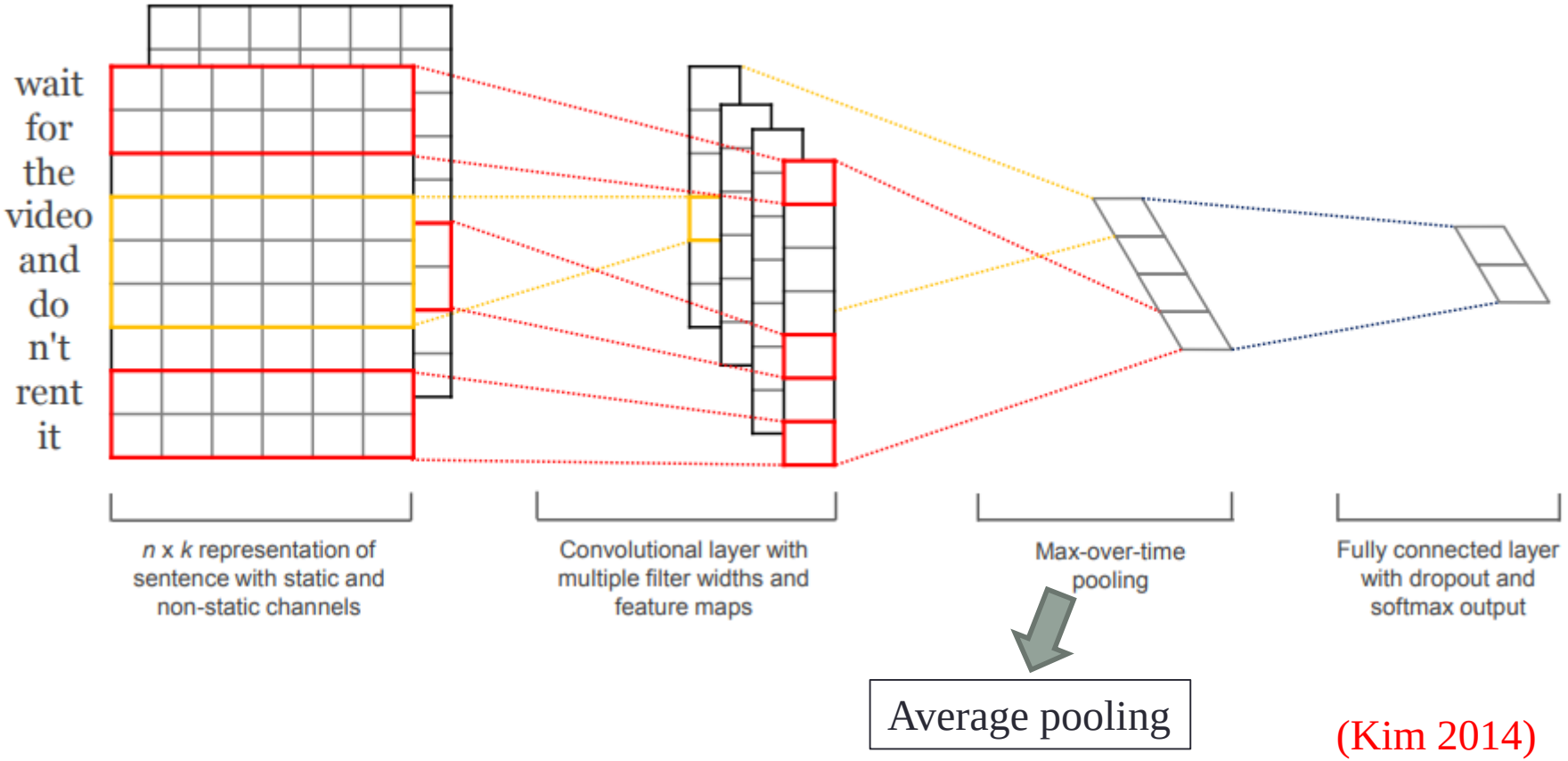
$$\sum_{\substack{(a,b)\in E^{+} \\ (a,b^{-})\in E^{-}}} L^{batch}\left(sim(a,b), sim(a,b_1^{-}), \ldots, sim(a,b_k^{-})\right)$$

Word2vec: negative example

- Different from the design of Word2vec, StarSpace used hinge loss (Rosset, Zhu and Hastie 2003) to compare the relation of positive/negative pairs of loss function.

Distributed Representation Learning: *CNN*

- CNN is a class of deep neural networks, that has successfully been applied to analyzing visual imagery and NLP.
- We utilize Keras to implement this readability model.



(Kim 2014)

Distributed Representation Learning: Based on Distributed Representations to Develop Generalized Readability Model (cont.)

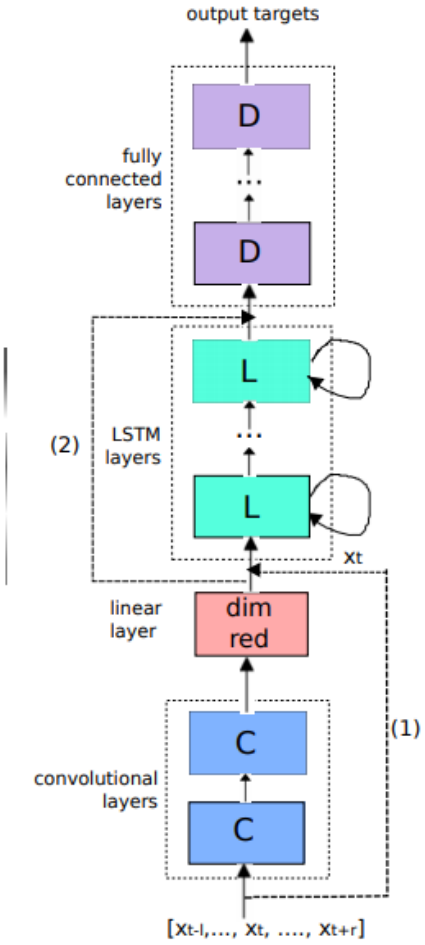
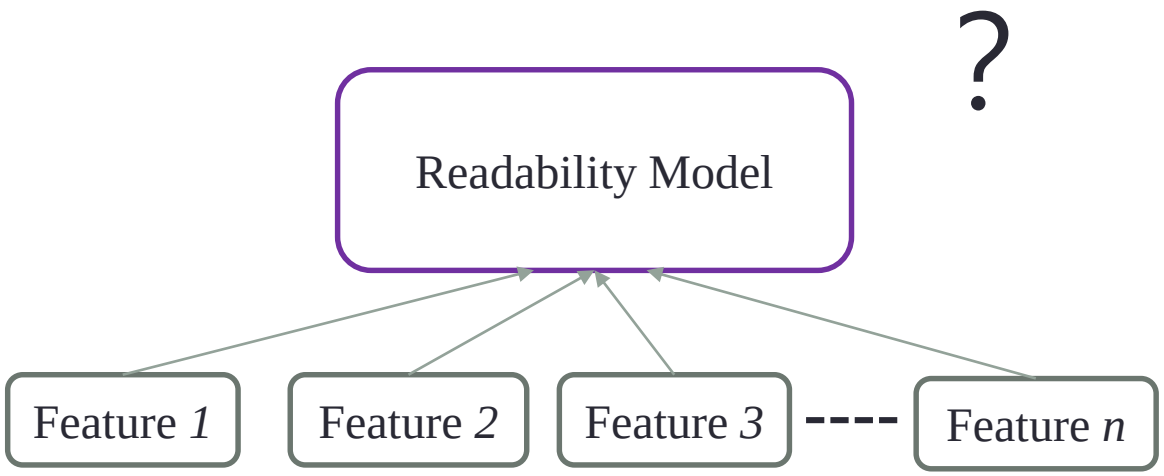
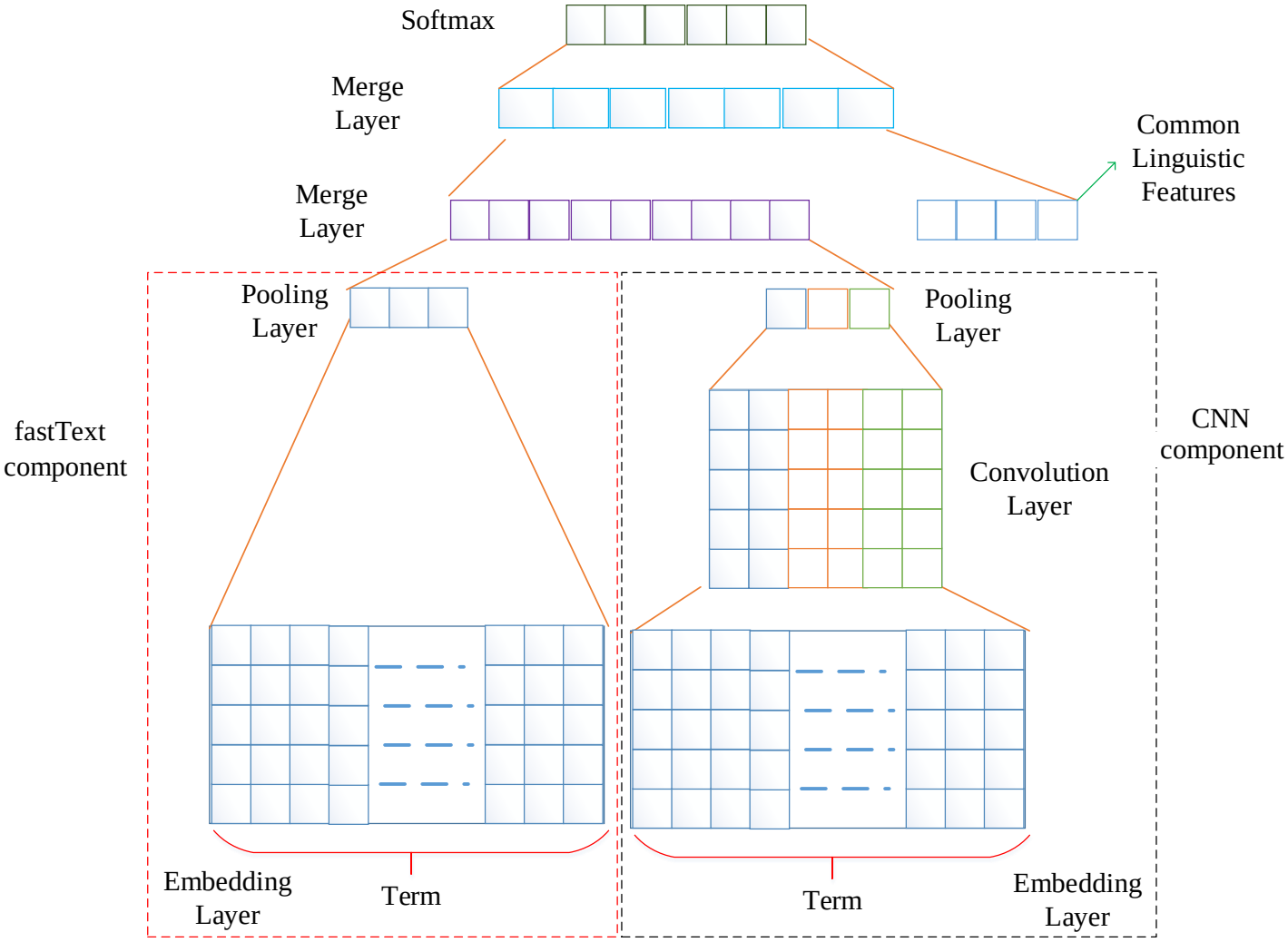


Fig. 1. CLDNN Architecture



➢ Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4580-4584). IEEE.

Distributed Representation Learning: Based on Distributed Representations to Develop Generalized Readability Model (cont.) – *Hybrid*



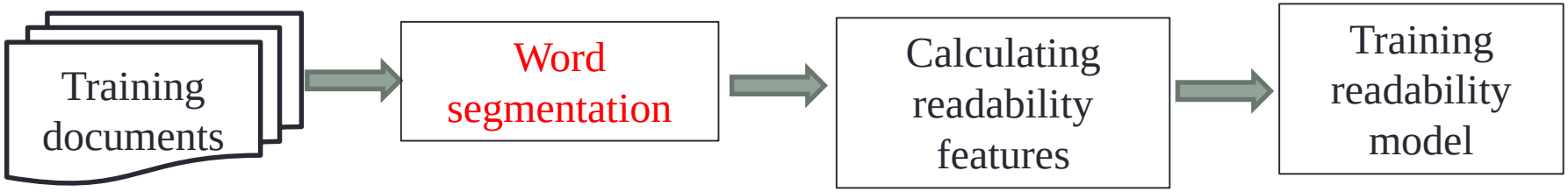
Experiment Results

We used 15 general linguistic features (see Appendix B for details) employed are the same as those used for the readability model developed by Sung *et al*. (2015a)

	Accuracy (%)	Adjacent Accuracy (%)
Common Linguistic Features		
Word2vec Features (CBOW)		
fastText Features (CBOW)		
StarSpace Features		
CNN Features		
Hybrid Features		

Research Objectives

- In conventional Chinese readability models, word segmentation is one of the most basic and important procedures in the pre-processing of texts. However, the ambiguity arising from word segmentation is often a knotty problem that will inevitably happen in the pre-processing of texts.



English

Power outage all over Taiwan.

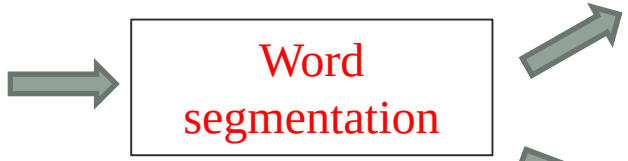
space

Power / outage / all / over / Taiwan.

Chinese

全台大停電

No space



全台 / 大 / 停電

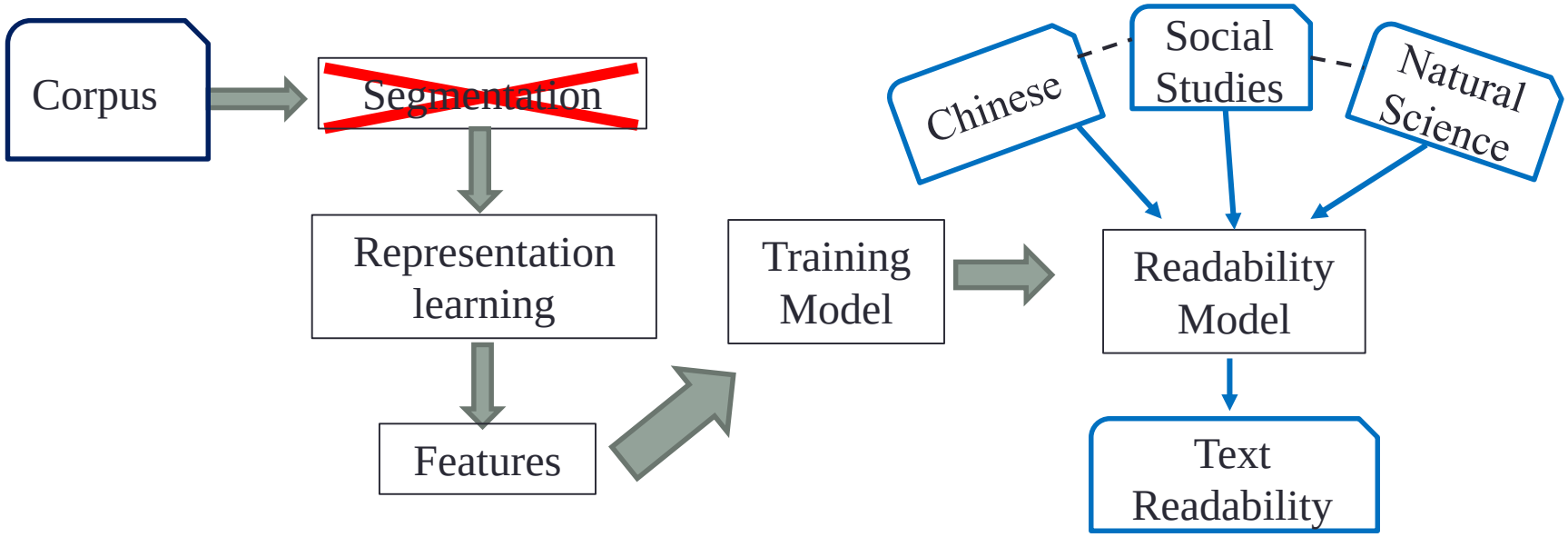
Power outage all over Taiwan.

全 / 台大 / 停電

Power outage in National Taiwan University.

Research Objectives (cont.)

- However, no segmentation model 100% guarantees accuracy.
- Is it possible to train readability model without recourse to word segmentation ?



Experiment Results: materials

- This paper used a total of 4,648 texts as the experimental data, which were selected from approved Chinese, social studies and natural science textbooks for grades 1-12 published in 2009 by three major Taiwan publishers, Han-Lin, Kang-Hsuan and Nan-I, all of which were compiled by experts following the curriculum guidelines.

	1	2	3	4	5	6	7	8	9	10	11	12
Social Studies	0	0	80	74	85	81	389	407	325	340	331	270
Natural Science	0	0	72	67	67	62	172	175	157	211	355	295
Chinese	24	67	61	71	69	70	37	34	28	84	41	47

Training Readability Model - BERT

Bidirectional Encoder Representations from Transformers

- How about use "BERT" ? (proposed by Google On Oct. 2018)
  - BERT is a novel neural language model which makes effective use of bi-directional self-attention (also called the Transformer) to capture both short and long-span contextual interaction between the tokens in the input sequence, usually in the form of words or word pieces.
  - The training of BERT consists of two stages: pre-training and fine-tuning.

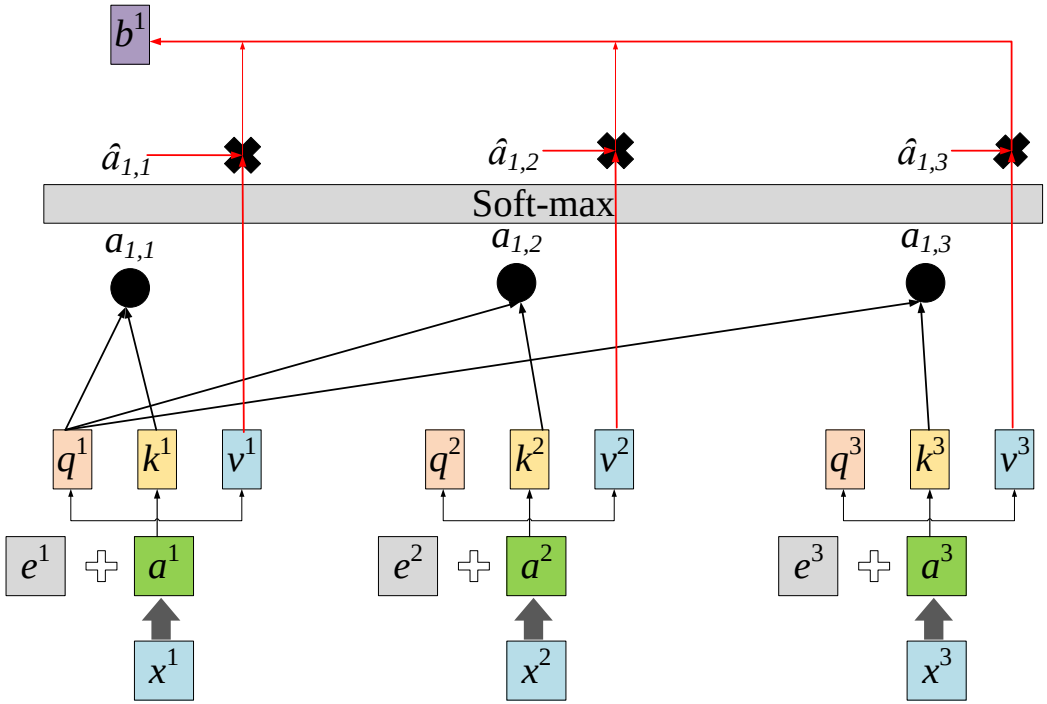
Released by Google

Downstream NLP tasks

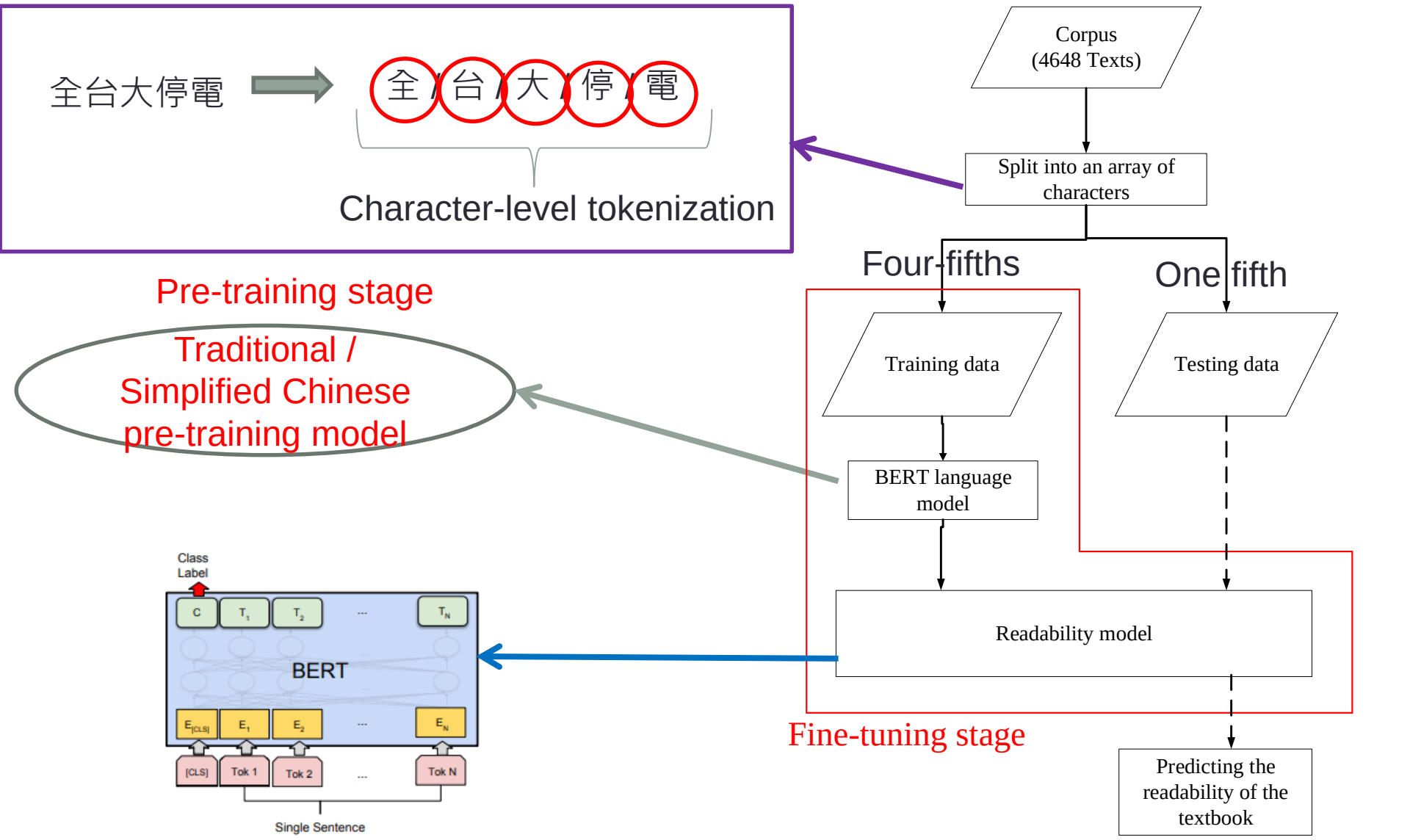
Training Readability Model - BERT

Attention Is All You Need

- Training by 3.3 billion tokens corpus.
- 1.1 billion parameters need to training.
- 21,128 token list.
- Released by Google. ☺

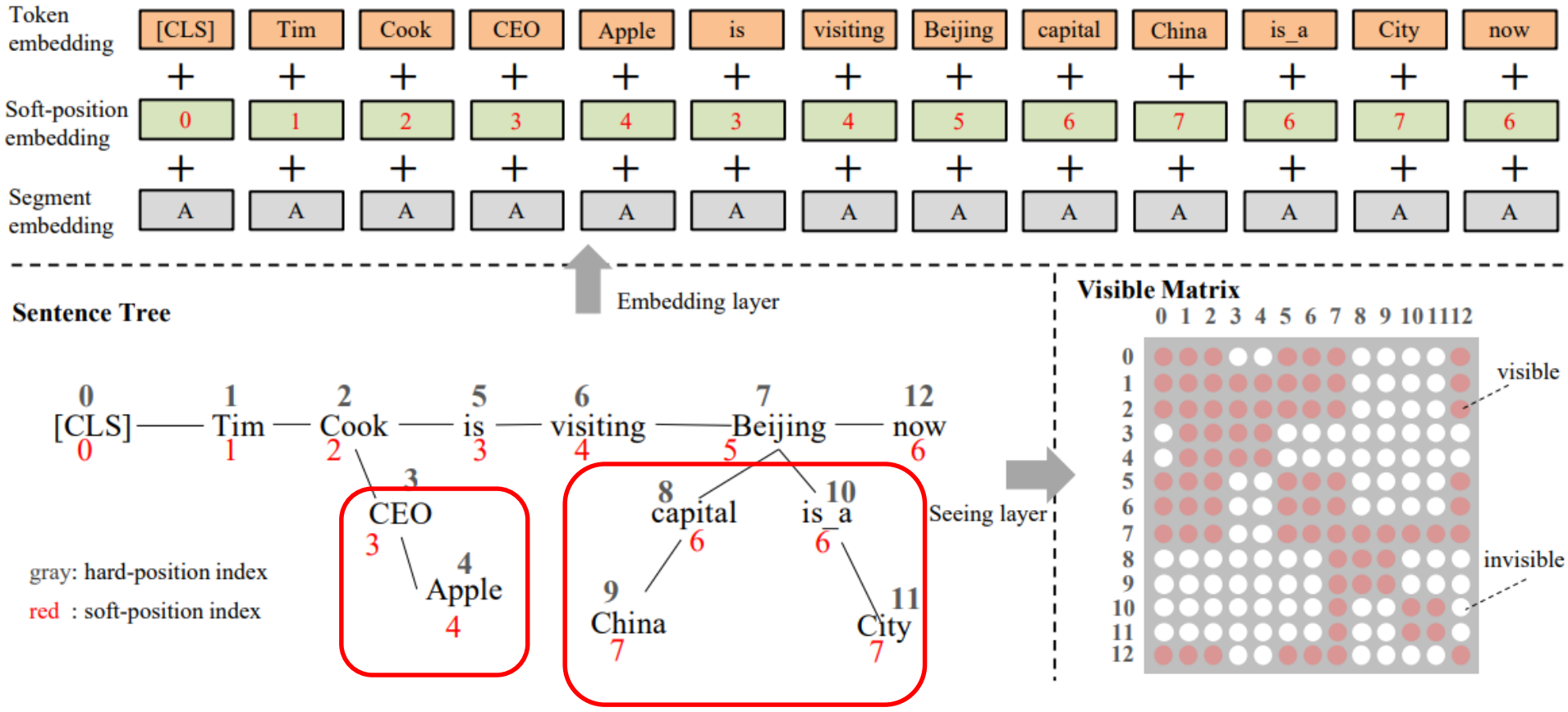


Training Readability Model - BERT

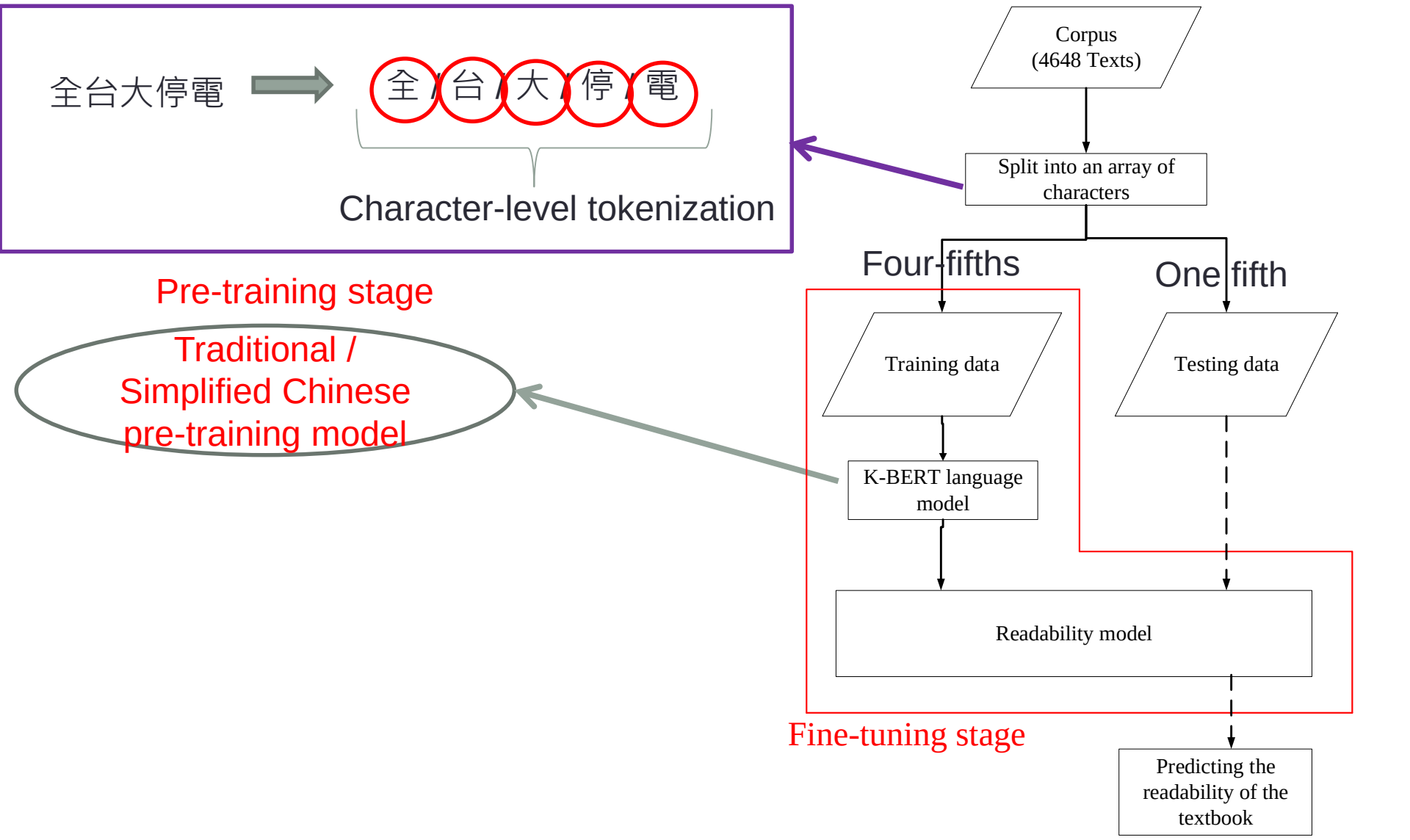


Training Readability Model – K-BERT

- K-BERT：Enabling Language Representation with Knowledge Graph (Proposed by Peking University On 2019)



Training Readability Model – K-BERT



Experiment Results

	Word-level		Character-level	
	Linguistic features + DNN model	**fastText model (proposed by Google on Jul. 2016)**	**BERT model**	**K-BERT model**
Accuracy				
Adjacent accuracy				

Adjacent accuracy: Allowing plus/minus one level error.

SMARTREADING

透過適性閱讀，培養自律自主的閱讀者

Adaptive reading for self-regulated readers

能力檢測


首頁
能力檢測
閱讀計畫
SmartReader金榜
在這裡
你可以瞭解自己的閱讀理解能力程度
能力檢測
或
你可以挑選適合自己能力或喜好的圖書
閱讀計畫
Smart Reading

試題示例

結果報表

閱讀能力

常模參照分數

各分向度能力

評量中文文本的可讀性(CRIE)

Chinese Readability Index Explorer
Smart Reading

- 準確率: %; 鄰近準確率: %
- 目前系統中涵蓋約43,000本(簡體書20,000； 繁體書23,000) 持續增加中

1.從國語、社會、自然及體育和健康教育等四科教科書，
抽取出知識層建構階層式的概念空間

2.為了驗證CRIE這個可讀性模型的效能，
利用CRIE去評量1,565本臺灣所公開販賣的課外讀物

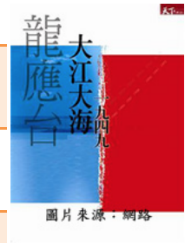
3. 書均由知名的出版社 (如Global Views, Youth Culture, 遠流等)所出版。13大領域包含華文文學、兒童文學、繪本、傳記、藝術、自然科普、社會史地、休閒生活等

現實之中的應用

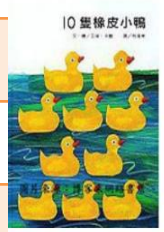

我的閱讀計畫
SmartReading提供符合你能力程度的圖書推薦書單，你可以規劃自己的閱讀時間與進度，透過這些有系統的紀錄，能更加瞭解自己進步的情形。
你的能力難度範圍：393C - 483C
書籍難度範圍設定
目前你累積250分!!
書名
搜尋
繁體書
書籍搜尋功能
圖書難度
書名
類別
出版社
作者
讀後評量：A+ A B
帕瓦娜的旅程
圖書難度：393C
放入我的書庫
讀後評量：A+ A B
魚的臉
圖書難度：393C
放入我的書庫
寓言
讀後評量：A+ A B
大提琴與樹
圖書難度：393C
讀後評量：A+ A B
披風少年
圖書難度：393C
放入我的書庫
加入讀書計畫按鈕

學校年級	SR分級	書籍的分級
G12 高三	SR700↑	蘇東坡詞（下）（SR707）
		蔡明亮與緩慢電影（SR697）
	SR600	蔡康永的說話之道（SR663）
		物理奇遇記：湯普金（SR603）
	SR500	大江大海一九四九（SR554）
		紅樓夢（SR514）
	SR400	老人與海（SR462）
		和小蓮一起遊茉莉花園（SR407）
	SR300	10隻橡皮小鴨（SR342）
		狐狸愛上圖書館（SR311）
G2 小二	SR300↓	爸爸與我（SR183）

書名：蔡康永的說話之道（兔斯基慶功版）
ISBN：9789861363974
作者：蔡康永
出版社：如何
字數：58030
一般類別：哲學/心理
圖書資源：國資圖 北市圖



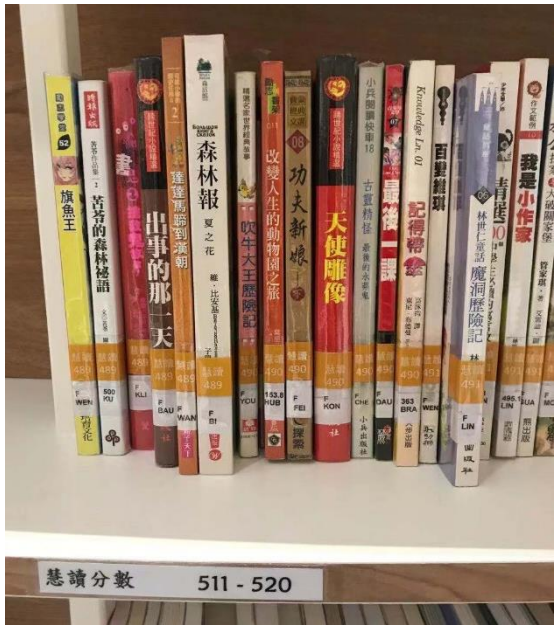
書名：大江大海一九四九
ISBN：9789863870371
作者：龍應台
出版社：印刻文學生活雜誌
字數：214361
一般類別：現代文學
圖書資源：校內圖書館

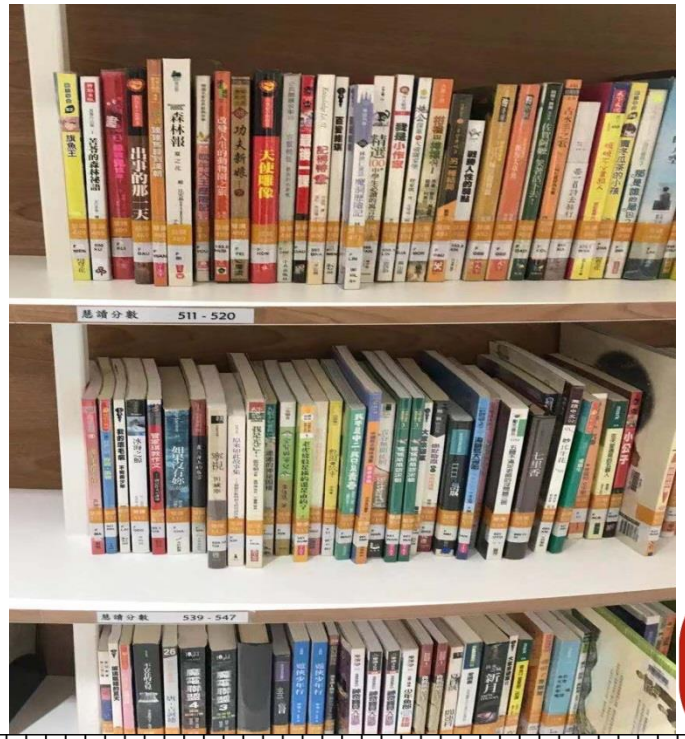


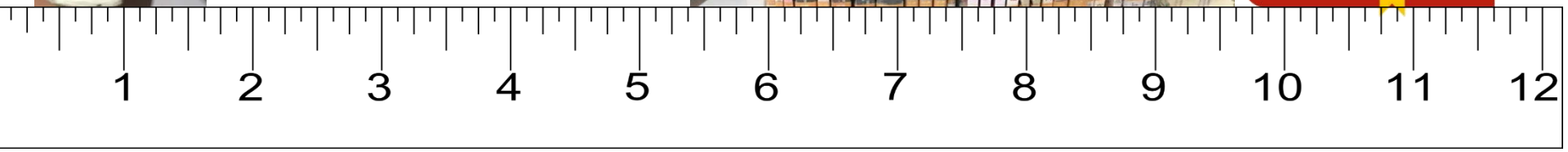
書名：10隻橡皮小鴨
ISBN：9577624146
作者：艾瑞卡爾/文圖；柯倩華/譯
出版社：上誼
字數：1061
一般類別：懸疑推理/冒險/科幻
圖書資源：國資圖 北市圖

SR的讀本分級 應用於海內外學校圖書館

SR的讀本分級 應用於海內外學校圖書館







臺北市立景美女子高級中學
圖書館WEBPAC館藏服務平台

回首頁 | 開館日 | 登入

線上人數：2人
點閱累計：278659人

書籍資料

返回館藏查詢

電學之父

資料類型：印刷本圖書
書目號：0019535

書　名	電學之父：法拉第的故事
作　者	張文亮著
版本項	第一版
出版項	臺北市：文經社，1999[民88]
稽核項	207面：圖, 像；21公分
ISBN / 裝訂 / 價格	9576632463 / 平裝 / NT$160
集叢項	文經文庫.161
適用對象	成人，一般性
分類號	309.941 自然科學類/總論 Pure Sciences/科學史
書籍困難度	SR 571 / 九級(535 ~ 622)
標　題	科學-英國-傳記
著作者	張文亮著

w/library/library.action

讀者服務　跨校查詢 未啟用

讀者留言　線上續借　線上推薦
預約查詢/取消　個人資料　閱讀認證
SR推薦區間

Q & A

Thank you for your attention