Recent Developments in Machine Learning-based Text Readability Assessment
基於機器學習之文本可讀性評估的研究進展

Berlin Chen (陳柏琳)
Professor, Department of Computer Science & Information Engineering
National Taiwan Normal University
2016/06/29

Outline

- Introduction
- Spectrum of Text Readability Research
- Early Research with Shallow Semantic/Semantic Features
- Machine Learning with Content Features
- Conclusion an Outlook

This talk gives only a partial picture of research on text readability assessment, biased and subject to the presenter's expertise. For more detailed reviews on mainstream developments, please also refer to, among others,

1 K. Collins-Thompson, "Computational assessment of text readability: a survey of current and future research," 2014
2. E. Pitler & A. Nenkova , "Revisiting readability: a unified framework for predicting text quality,", 2008

Introduction

- Research on text readability assessment has witnessed a booming interest in the past decade, partly due to the rapid proliferation of downstream applications and dramatic progress of machine learning technology
  - Early developments in on text readability assessment date back to research efforts conducted in the 40-50's by pioneers such as Dale & Chall (1948); many useful readability formulas have been developed since then
- **Text readability** was formally defined as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the material (Dale & Chall, 1949)
  - Should also be a function of reader's aptitudes

K. Collins-Thompson, "Computational assessment of text readability: a survey of current and future research," 2014

Spectrum of **Text Readability** Research

Modeling

- Human Engineering
- Machine Learning (regression, classification & ranking)

Targets

- Traditional Texts
- Non-traditional Texts (e.g., web/social media)
- Spoken Utterances (E.g., oral proficiency assessment)

Corpora & Evaluation Metrics

- Intrinsic
- Extrinsic

Applications

- Readability Prediction (e.g., Educational Applications)
- Summarization & Simplification
- Information Retrieval
- Producing Instructions and Guides etc.

Features

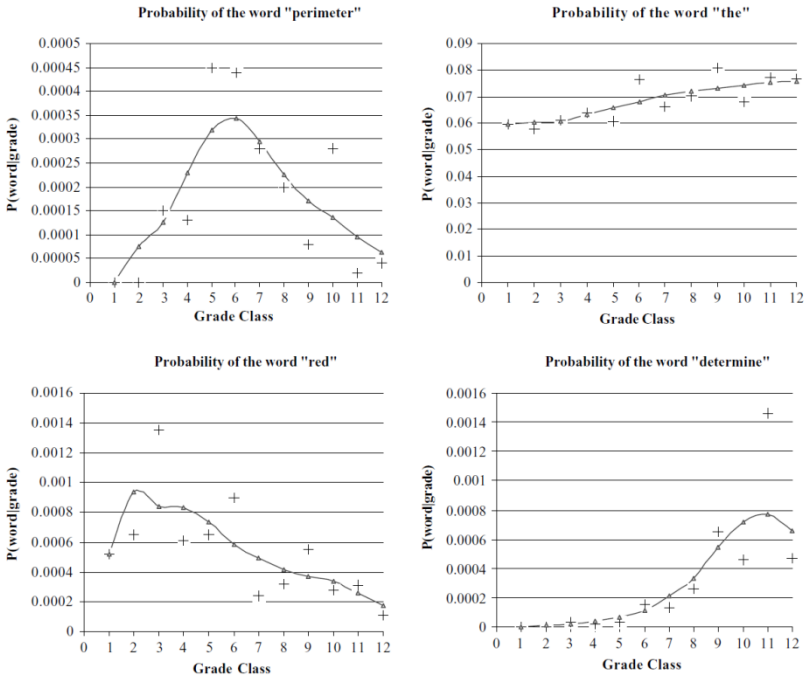
- Lexico-Semantic/Morphological, Syntactic & Content Features
- Discourse: Cohesion & Coherence
- Pragmatic & Genre Features
- Layout and Graphic Illustrations
- Reader's Cognitive Aptitudes

Early Research: Factors for Measures (1/2)

- Most readability measures have focused on two main factors
  - The familiarity of the semantic units (words or phrases) used
  - The syntactic complexity of the sentence structure
- It has also been indicated that (Chall, 1958)
  - **Vocabulary difficulty** is known to account for at least 80% of the total variability explained by readability scores for traditional texts
  - **Sentence structure** giving a small additional amount of predictive power
- Aspects of reading difficulty associated with higher-level linguistic structures in the text, such as its discourse flow or topical dependencies, are largely ignored

Early Research: Factors for Measures (2/2)

- More on vocabulary difficulty and word usage
  - Analysis of word usage across grades revealed that (Chall, 1983)
    - Earlier grade levels tend to use more concrete words like ***red***, whereas later grade levels use more abstract words such as ***determine*** with greater frequency



Examples of four different word usage trends across grades 1–12, as sampled from 400,000-token corpus of English Web documents

K. Collins-Thompson & J. Callan, "Predicting reading difficulty with statistical language models," 2005

Early Research: Some Classic Measures (1/3)

- Flesch-Kincaid Measure (1975)

$$RG_{FK} = 0.39 \cdot \frac{W}{S} + 11.8 \cdot \frac{L}{W} - 15.59$$

Average Word Per Sentence
(sentence length: syntax factor)

Average Syllable Per Word
(word length: semantic factor)

- *W*: total number of word in the text sample
- *S*: total number of sentences in the text sample
- *L*: total number of syllables in the text sample

Early Research: Some Classic Measures (2/3)

- Revised Dale-Chall Measure (1995)

$$RG_{DC} = 3.6365 + 0.1579 \cdot \frac{U}{W} + 0.0496 \cdot \frac{W}{S}$$

  - *U*: total number of unfamiliar words (tokens) in the text sample
  - *W*: total number of words in the text sample
  - *S*: total number of senteces in the text sample

  - A word list consisting of 3,000 words that 80% of tested fourth-grade students were able to read was used
  - A token is labeled *unfamiliar* if the token or simple variants of it do not appear in the 3,000-word list

Early Research: Some Classic Measures (3/3)

- Traditional readability measures are based only on surface characteristics of text, and ignore deeper levels of text processing known to be important factors in readability, such as cohesion, syntactic ambiguity, rhetorical organization, and propositional density

- Readers' cognitive aptitudes are largely ignored
  - Such as the reader's prior knowledge and language skills, which are used while they interact with the text

K. Collins-Thompson, "Computational assessment of text readability: a survey of current and future research," 2014

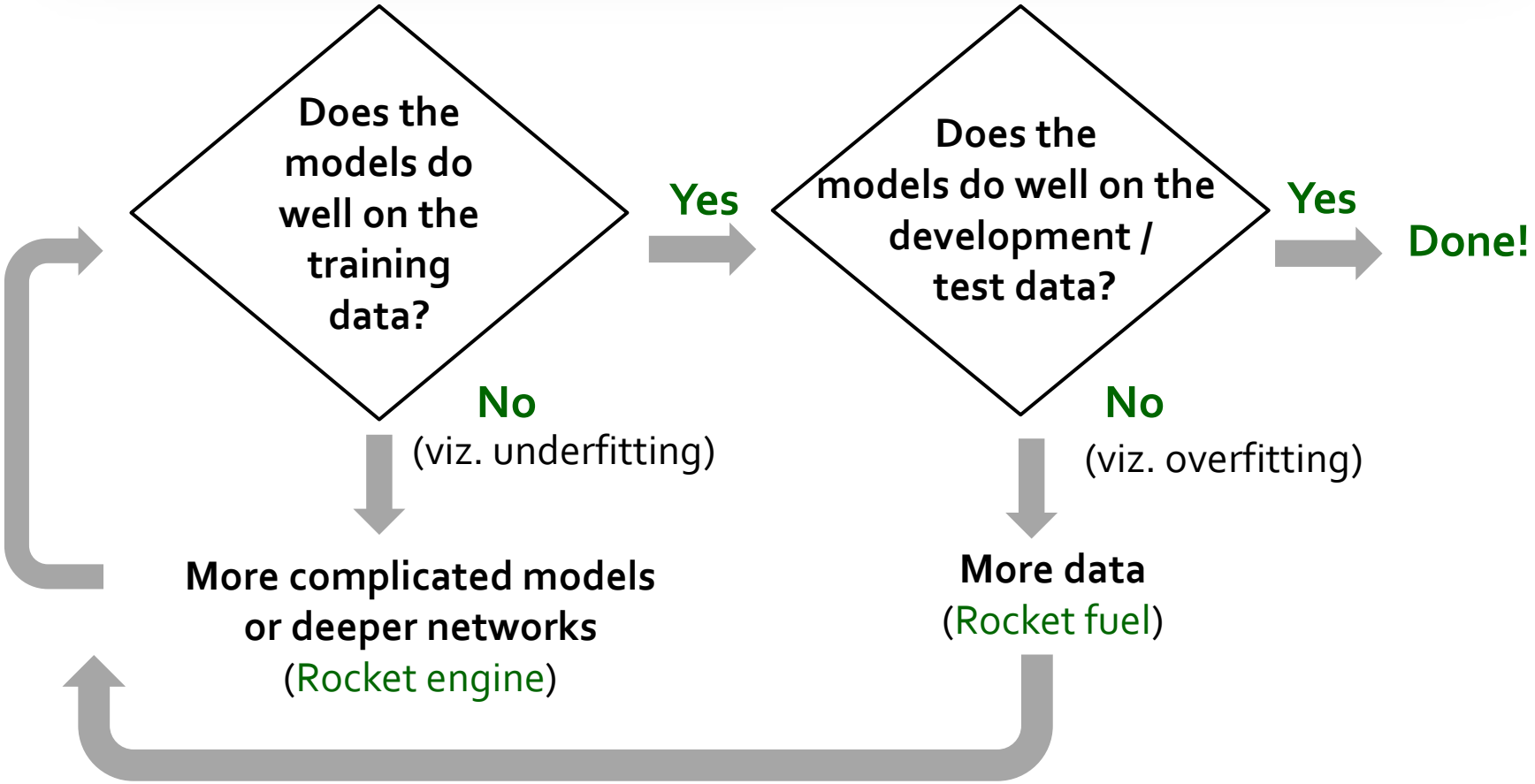
What is Machine Learning?

Machine learning

From Wikipedia, the free encyclopedia

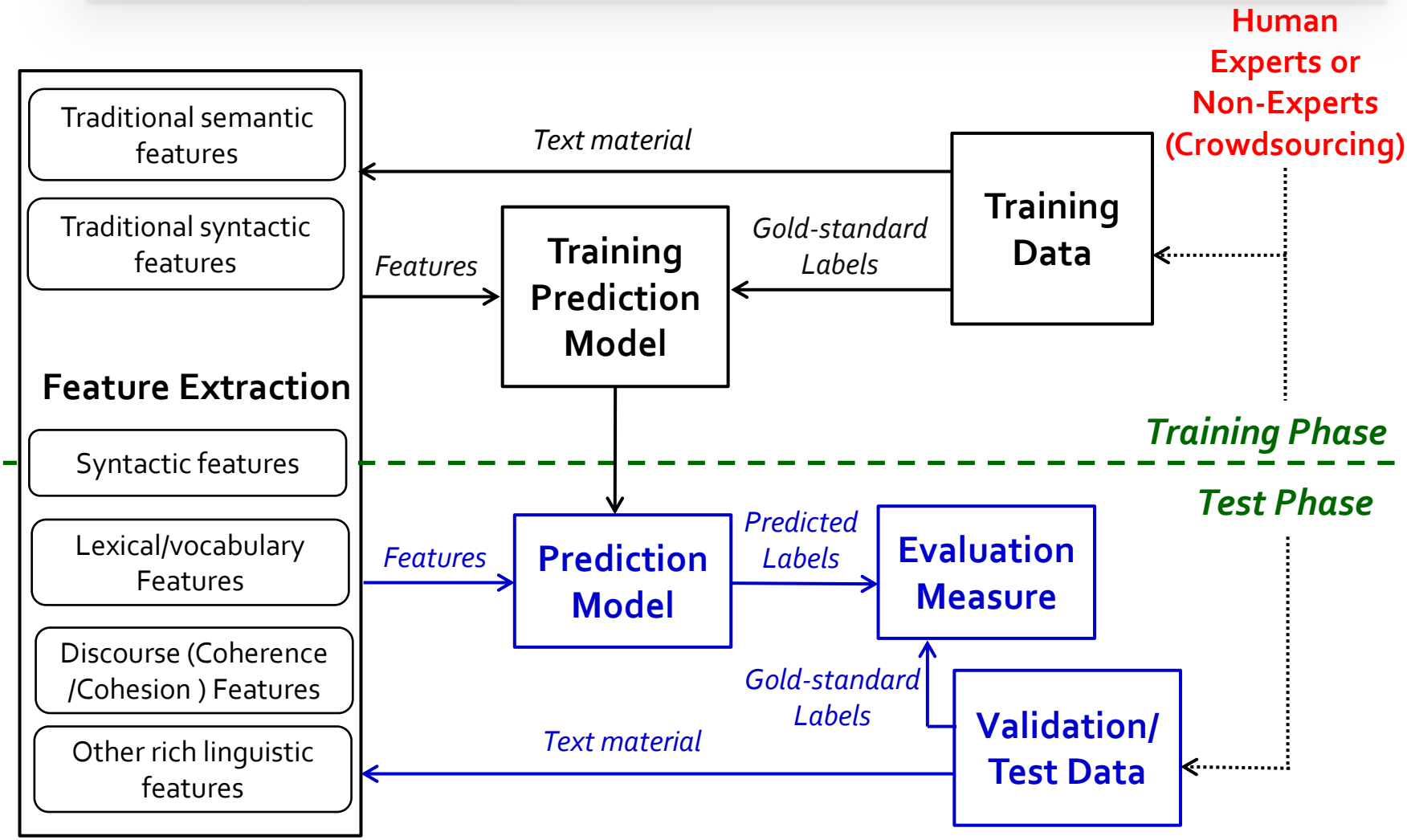
Machine learning is a subfield of computer science[1] that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.[1] In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed".[2] Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.[3] Such algorithms operate by building a model from an example *training set* of input observations in order to make data-driven predictions or decisions expressed as outputs,[4]:2 rather than following strictly static program instructions.

Typical Recipe for **Machine Learning** Research



There is no data like more data!

Machine Learning (ML) for Text Readability



ML: Labeled Corpora (1/2)

- A "gold-standard" training corpus of individual texts is constructed that is representative of the target genre, language, or other aspect of text for which automatic readability assessment is desired
- Each text in the training corpus is assigned a "gold-standard" readability level
  - Typically annotated by human experts (time-consuming and expensive)
  - Or annotated by human non-experts through crowdsourcing platforms
- Some important aspects:
  - Size, language, genre, etc.

ML: Labeled Corpora (2/2)

- Difficulty Levels
  - The standard unit for reading difficulty labels is the school grade level, but other scales of measurement are also used
    - The grade level could be an ordinal value corresponding to discrete ordered difficulty levels, for instance, American grade levels 1 through 12,
    - Or it could be a continuous value within a range, to capture within-level gradations, which are especially important for earlier grade levels (e.g. a text at Grade 5.7)

1. K. Collins-Thompson, "Computational assessment of text readability: a survey of current and future research," 2014
2. E. Pitler and A. Nenkova, "Revisiting readability-a unified framework for predicting text quality," EMNLP 2008

ML: Features (1/3)

- A set of *features* is defined that are to be computed from a text. These features capture semantic, syntactic, and other attributes of the text that are salient to the target readability prediction task.

- Vocabulary-based features
  - Relative frequency of a word
  - Type-token ratio (lexical richness cue)
  - Language models
  - Word maturity measure
  - Word Concreteness (perceivability & imageability)
  - …

ML: Features (2/3)

- Syntactic Features
  - Average number of word per sentence
  - Average parse tree height
  - Average number of noun phrases per sentence
  - Average number of verb phrases per sentence
  - Average number of subordinate clauses per sentence
  - Number of passive sentences
  - …

Having multiple noun phrases (entities) in each sentence requires the reader to remember more items, but may make the article more interesting.
- (Barzilay and Lapata, 2008) found that articles written for adults tended to contain many more entities than articles written for children

While including more verb phrases in each sentence increases the sentence complexity, adults might prefer to have related clauses explicitly grouped together.

ML: Features (3/3)

- Discourse structure
  - Model the semantic/pragmatic connection of sentences in a document, such as elaboration, contrast and background

- Coh-Metrix (Graesser and McNamara, 2004)
  - A computational linguistics tool that has played a prominent role in automated readability assessment, by providing a multi-dimensional set of linguistic and discourse features for text representation
    - Analyze texts on over 200 measures of cohesion, language, and readability

A.C. Graesser, et al., "Coh-Metrix: Analysis of text on cohesion and language," *Behavior Research Methods, Instruments, & Computers*, 2004

ML: Models

- A machine learning model learns how to predict the gold standard label for a text from the text's extracted feature values
  - Language Models (e.g., Unigram, *N*-gram, RNN/LSTM)
  - Topic Models (e.g., LSA, PLSA/LDA)
  - Decision Trees
  - Ensemble Learning (e.g., Adaboost, Bagging, etc.)
  - Support Vector Machines (SVM)
  - Representation Learning (e.g., Deep Neural Networks, Word & Phrase Embeddings)
  - …
- To find a set of model parameters that is likely to generalize well to new texts, during the training phase, models are typically cross-validated against data unseen by the model

ML: Evaluation Metrics

- ***Extrinsic evaluation*** measures the impact of the readability prediction on the performance of downstream applications (tasks)
  - Such as information retrieval, text summarization, essay scoring, among others
  - *Could be time-consuming, expensive and require a considerable amount of careful planning*
- ***Intrinsic evaluation*** examines how well a readability prediction method performs in relation to gold-standard readability levels provided by human experts
  - **Classification Errors**
  - **Root Mean Square Errors (RMSE)**: difference between the predicted grade level and the actual grade level (RMSE weights outliers more heavily
  - Among others

Downstream Applications

- Information Retrieval
  - To Identify documents that not only are relevant to the query but also match the student's reading level
  - To search at a user's preferred level of reading difficulty and have the results reflect that choice
- Text Simplification & Summarization
- Essay Scoring & Second Language Learning
- User Guides, Instructions/ Prescriptions
- …

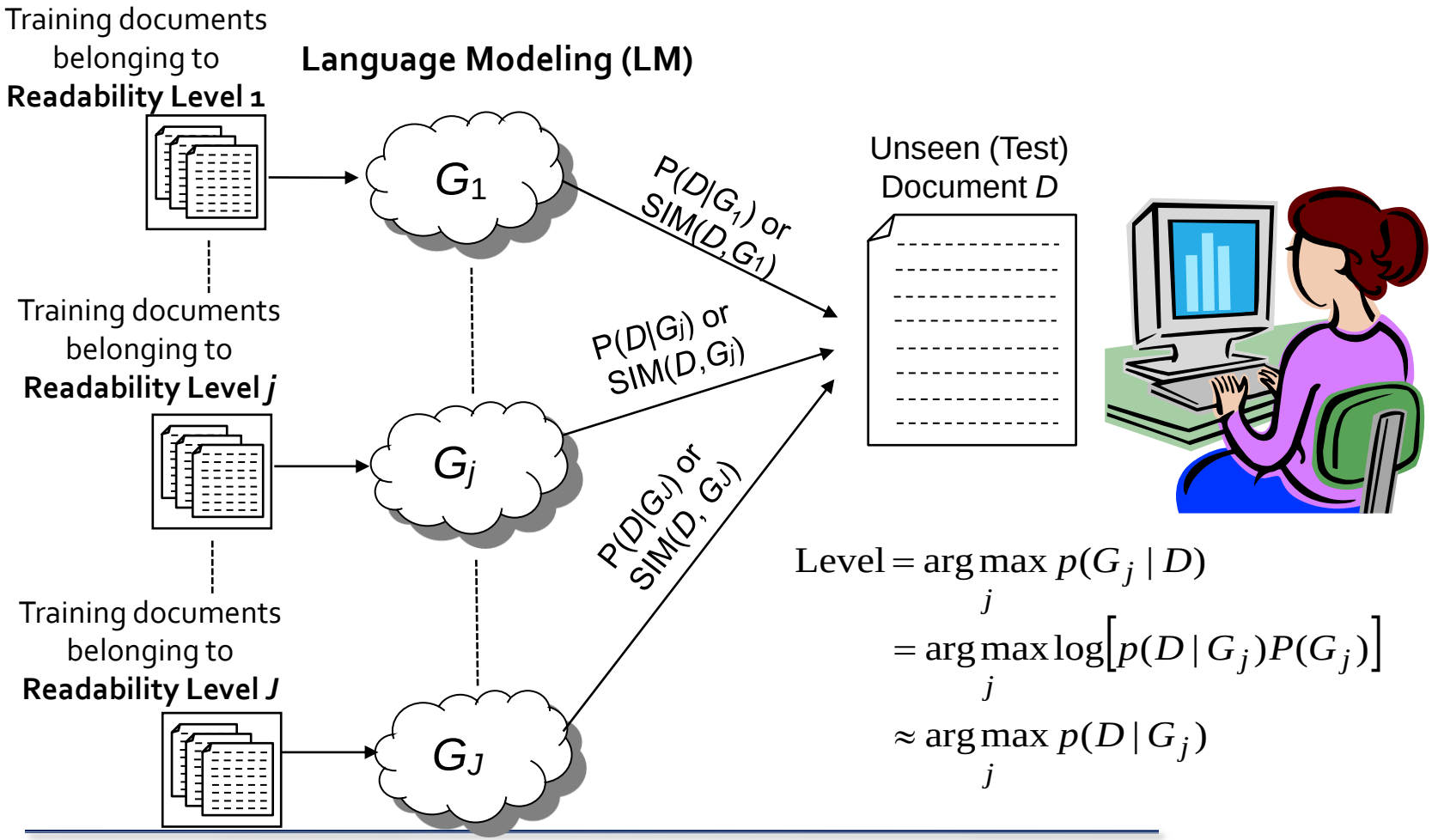
Language Modeling (LM) (1/6)

- LM is a simple and principled approach to using vocabulary information for readability decisions
  - **Formulation:**
    - Recast the well-studied problem of readability in terms of text categorization and use straightforward techniques from NLP, IR and ASR
  - **Working Assumption**
    - There are enough distinctive changes in word usage patterns between grade levels to yield accurate predictions with simple language models, even when the subject domain of the documents is unrestricted

K. Collins-Thompson & J. Callan et al., "Predicting reading difficulty with statistical language models," *Journal of the American Society for Information Science and Technology*, 2005

Language Modeling (2/6)

- Schematic Depiction



$$\text{Level} = \arg\max_j p(G_j \mid D)$$
$$= \arg\max_j \log\left[p(D \mid G_j)P(G_j)\right]$$
$$\approx \arg\max_j p(D \mid G_j)$$

Language Modeling (3/6)

- *N*-grams
  - Unigrams

$$P(D \mid \mathrm{G}_j) = P(D = w_1, w_2, \ldots, w_R \mid \mathrm{G}_j) \approx \prod_{i=1}^{R} P(w_i \mid \mathrm{G}_j)$$

  - Bigrams

$$P(D \mid G_j) = P(w_1 \mid G_j) \prod_{i=2}^{R} P(w_i \mid w_{i-1}, G_j)$$

$$P(w_i \mid G_j) = \frac{C(w_i, G_j)}{\sum_{w' \in \mathbf{V}} C(w', G_j)}$$

$\mathbf{V}$: vocabulary

- Implementation Specifics
  - Word selection (Stop word removal/ Stemming); separation of a document into passages
  - Probability smoothing (e.g., back-off or interpolation) to avoid zero probabilities for OOV words
  - Use of a mixture model of nearby classes might improve accuracy

Language Modeling (4/6)

- A Running Example (Collins-Thompson & J. Callan, 2005)

TABLE 2. Highly simplified unigram language models for grades 1, 5, and 12.

Type w	Grade 1 $P(w)$	Grade 5 $P(w)$	Grade 12 $P(w)$
the	0.06000	0.07000	0.08000
a	0.06000	0.05000	0.06000
red	0.00080	0.00040	0.00020
ball	0.00010	0.00005	0.00001
was	0.01000	0.01000	0.02000
perimeter	0.00005	0.00030	0.00005
optimal	0.000001	0.00001	0.00010

Note. $P(w)$ denotes the probability of type w in the model.

Example 2: Passage T = "the red perimeter"

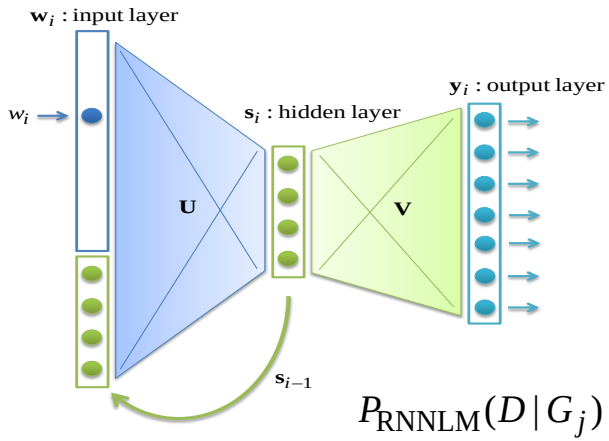
$$
\begin{aligned}
L(T \mid \text{Grade 1}) &= \log 0.06 + \log 0.0008 + \log 0.00001 \\
&= -9.319 \\
L(T \mid \text{Grade 5}) &= \log 0.07 + \log 0.0004 + \log 0.00030 \\
&= \mathbf{-8.076} \\
L(T \mid \text{Grade 12}) &= \log 0.08 + \log 0.0002 + \log 0.00005 \\
&= -9.097
\end{aligned}
$$

Prediction: Grade 5

K. Collins-Thompson & J. Callan et al., "Predicting reading difficulty with statistical language models," *Journal of the American Society for Information Science and Technology*, 2005

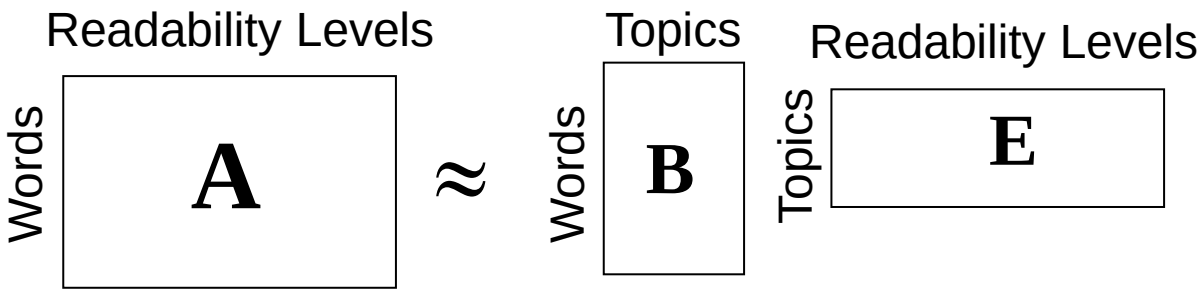
Language Modeling (5/6)

- Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)



$$P_{\text{RNNLM}}(D \mid G_j) = \prod_{i=1}^{L} P_{\text{RNNLM}}(w_i \mid w_1, \ldots, w_{i-1}, G_j)$$

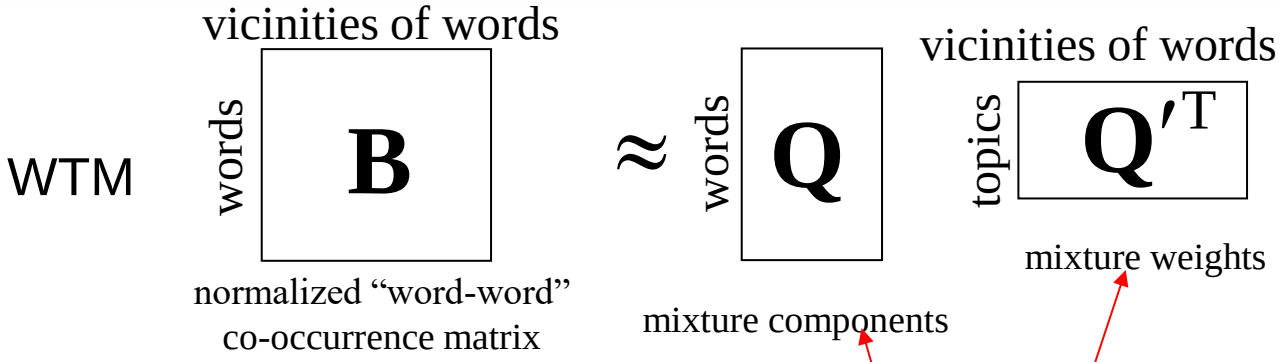
- Topic Models (PLSA/LDA)



$$P_{\text{PLSA/LDA}}(w_i \mid G_j) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid G_j)$$

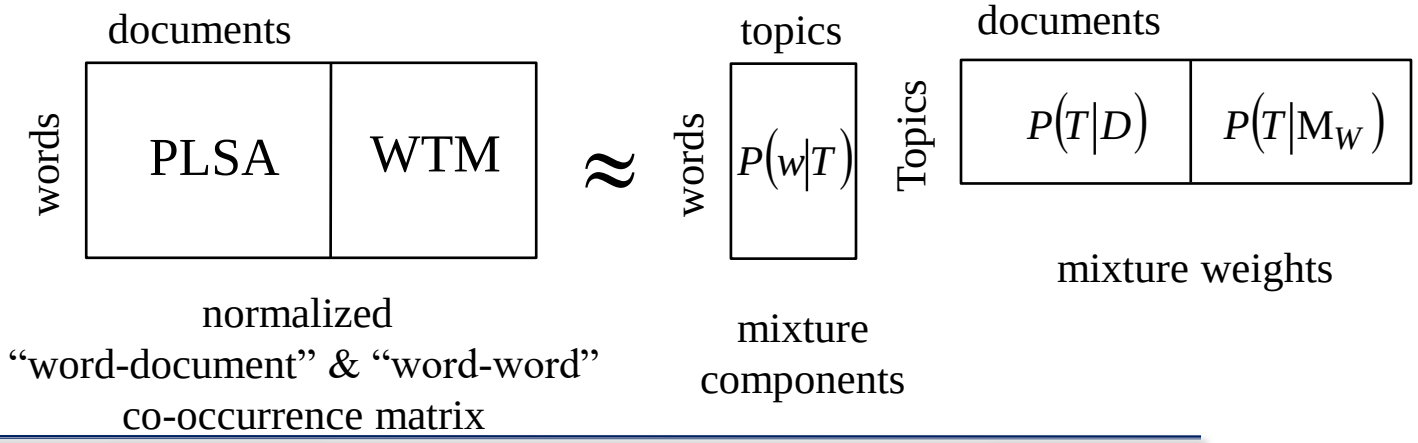
Language Modeling (6/6)

- Word Topic Models (WTM)



$$P_{\mathrm{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right)$$

- Synergy of WTM and PLSA



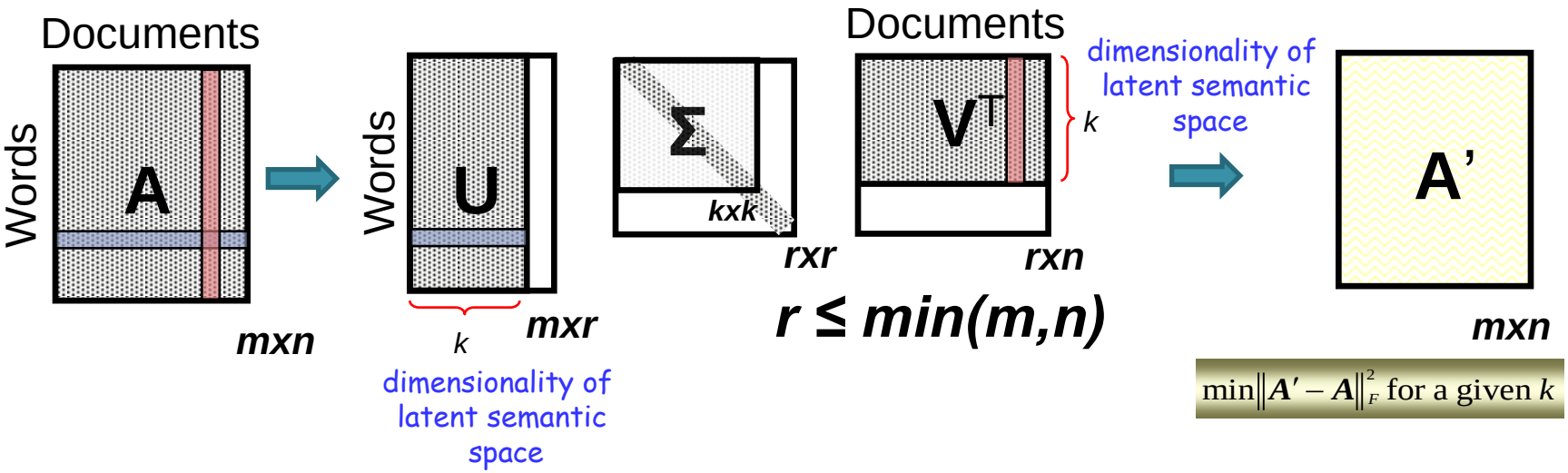
B. Chen, "Word topic models for spoken document retrieval and transcription," *ACM Transactions on Asian Language Information Processing, 8(1), 2009.*

Latent Semantic Analysis (LSA) (1/5)

- Also called Latent Semantic Indexing (LSI), Latent Semantic Mapping (LSM), or Two-Mode Factor Analysis
  - Three important claims made for LSA
    - The **semantic information** can derived from a word-document co-occurrence matrix
    - The **dimension reduction** is an essential part of its derivation
    - **Words and documents can be represented as points** in the Euclidean space
  - LSA exploits the meanings of words by removing "noise" that is present due to the variability in word choice
    - Namely, synonymy and polysemy that are found in documents

Latent Semantic Analysis (2/5)

- Schematic Depiction of Singular Value Decomposition (SVD)

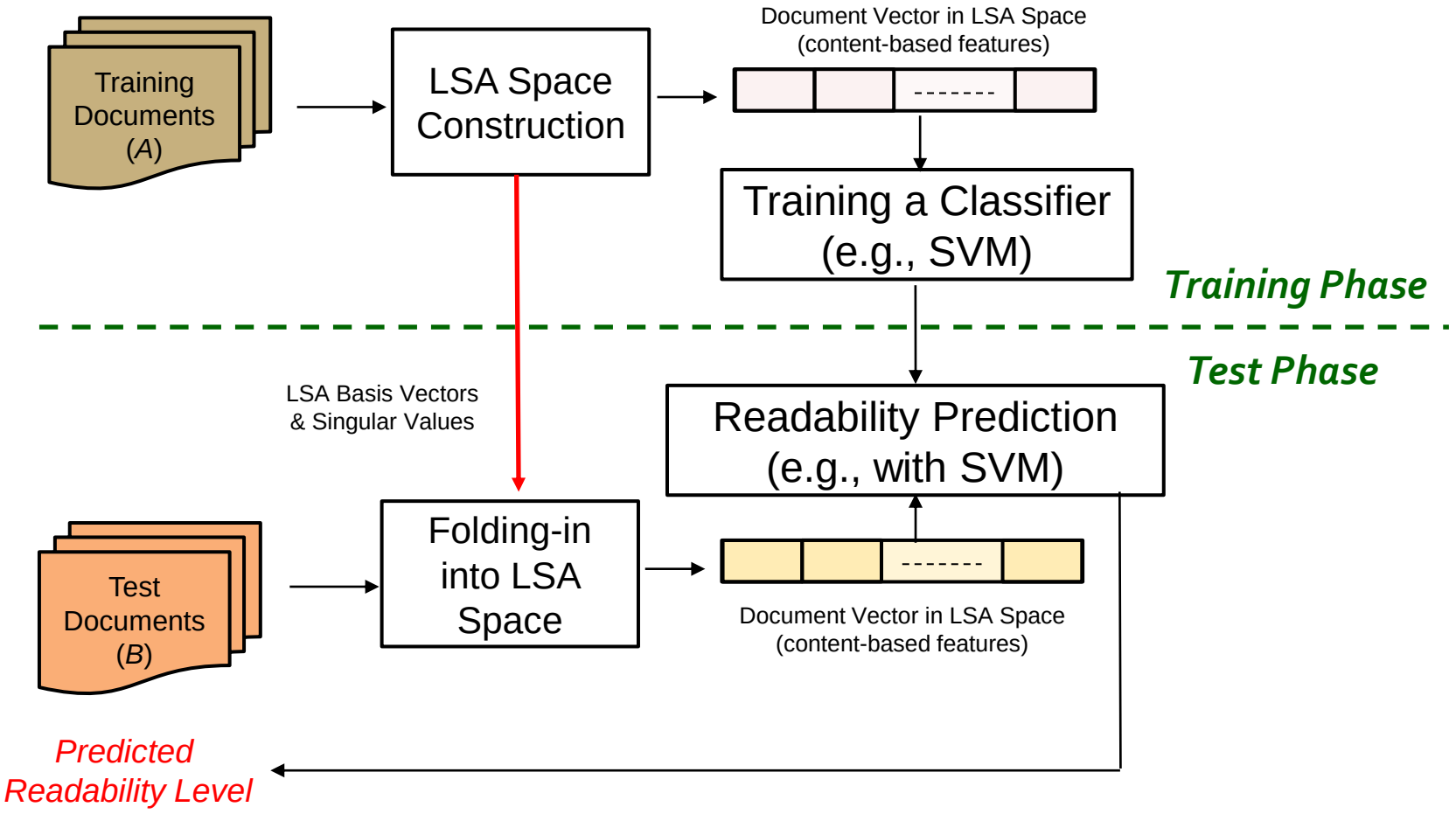


- LSA should balance two opposing effects
  - First, *k* should be large enough to allow fitting all the (semantic) structure in the real data
  - Second, *k* should be small enough to allow filtering out the non-relevant representational details (which are present in the conventional index-term based representation)

G.W. Furnas et al., "Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure," SIGIR1988

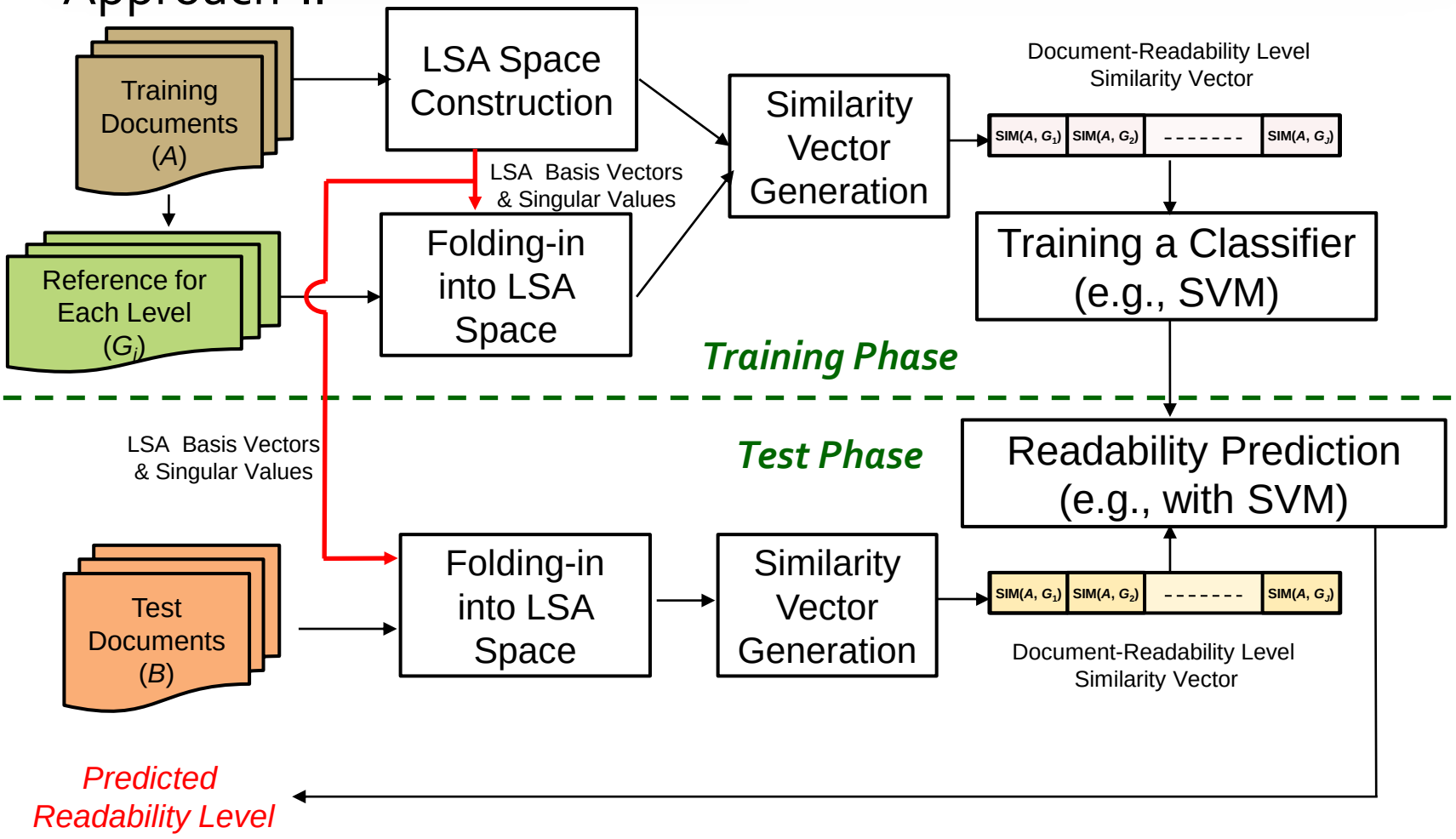
Latent Semantic Analysis (3/5)

- Approach-I



Latent Semantic Analysis (4/5)

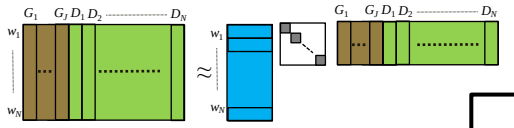
- Approach-II



A. R. Razon & J. A. Barnden, "A new approach to automated text readability classification based on concept indexing with integrated part-of-speech n-gram features," *RANLP*, 2015

Latent Semantic Analysis (5/5)

- Approach-III



Documents for Each Level (G_i)

Training Documents (A)

LSA Space Construction

Selection of Anchor Word Types for Each Level G_i

Anchor-Word Type List for Level G_1

Anchor-Word Type List for Level G_J

Matched Anchor-Word Type Count Accumulation for Each Level

Document Vector

Anchor-Word Lists for All Levels

Matched Counts for Level G_J

Training a Classifier (e.g., SVM)

Training Phase

Test Phase

Test Documents (B)

Matched Anchor-Word Type Count Accumulation for Each Level

Document Vector

Readability Prediction (e.g., with SVM)

Anchor-Word Lists for All Levels

Predicted Readability Level

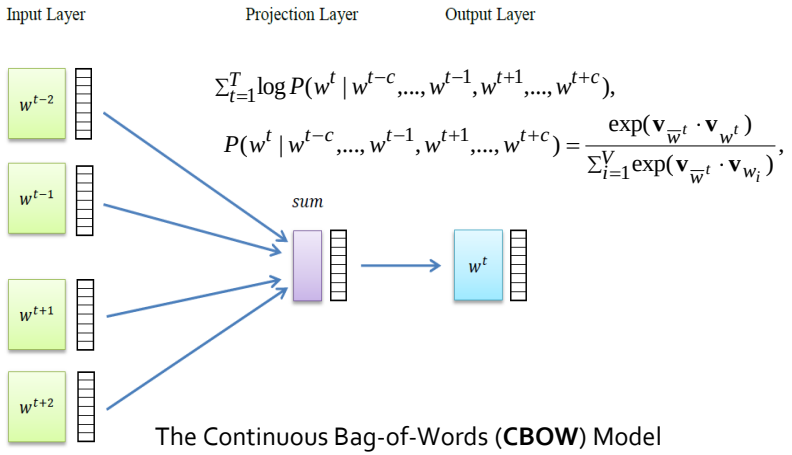
Representation Learning (1/4)

- Word/Paragraph Embedding (WE/PE)
  - Instead of a one-hotspot vector, a word is represented by a real-valued vector with a much smaller size (normally by several hundreds)
  - The syntactic and semantic regularities of words can be encoded in the distributed vector space: the Euclidean distance between two words in the lower-dimensional vector space represents the syntactic or semantic similarity between them
    - E.g., vector("***king***")-vector("***man***")+vector("***woman***") results in a vector that is closest to vector("***queen***")
  - A common thread of leveraging word embeddings to NLP-related tasks is to represent the document (or query and sentence) by averaging the word embeddings corresponding to the words occurring in the document (or query and sentence)

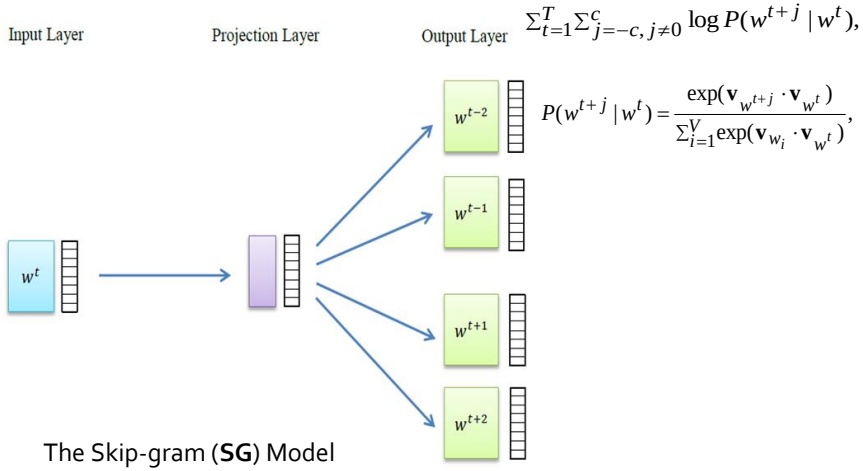
1. Bengio et al., "Representation Learning-A Review and New Perspectives ," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013
2. Mikolov et al., "Distributed representations of words and phrases and their compositionality," NIPS 2013

Representation Learning (2/4)

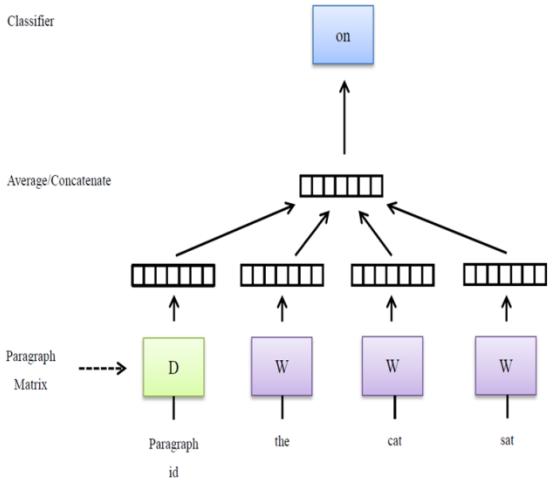
- Some Typical Learning Architectures



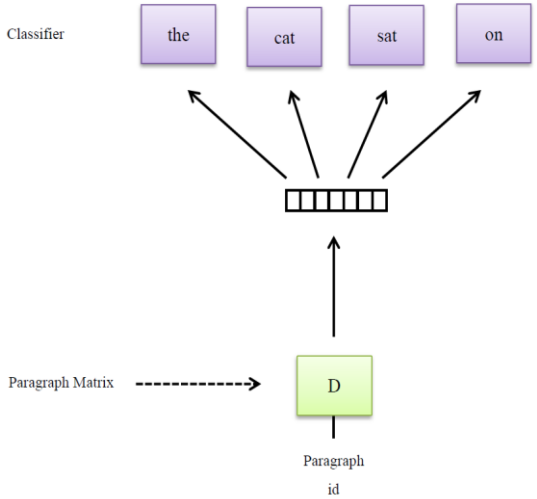
The Continuous Bag-of-Words (**CBOW**) Model



The Skip-gram (**SG**) Model

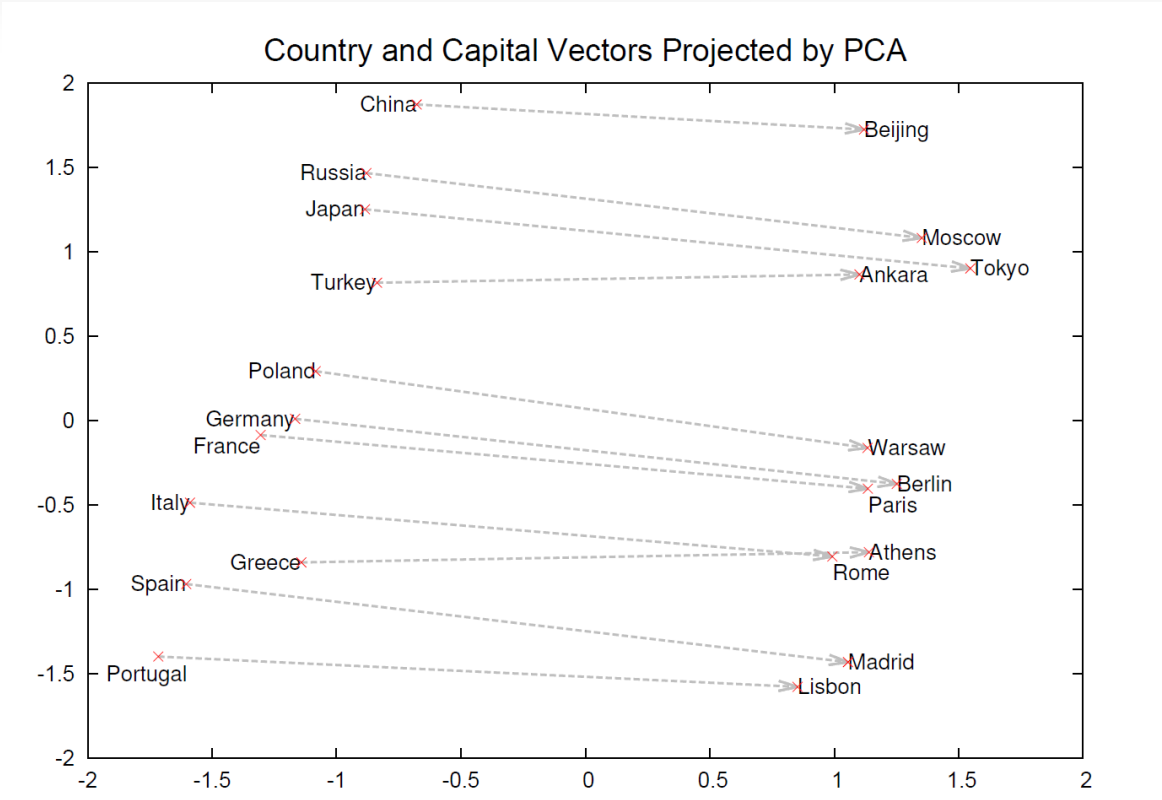


The Distributed Memory of Paragraph Vector(**PV-DM**) Model



The Distributed Bag-of-Words of Paragraph Vector (**PV-DBOW**) Model

Representation Learning (3/4)



Two-dimensional PCA projection of the 1,000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as no any supervised information about what a capital city means was provided during the training .

This figure is adopted from Mikolov et al., "Distributed representations of words and phrases and their compositionality," NIPS 2013

Representation Learning (4/4)

- The derived word/paragraph embeddings (vector representation) for documents (as well as the reference of each grade level) can serve to perform readability prediction in a similar way as LSA-based Approaches (I, II & III)

Comparison between LSA and WE (1/2)

- Corpus
  - A total of 4,648 texts as the experimental corpus
    - Selected from the Chinese, social studies and natural science textbooks for grades 1-12 published in 2009 by three major publishers in Taiwan
- Readability Prediction Accuracy (%)
  - Adopt a similar modeling framework as LSA-I

Dimensionality	LSA	CBOW	Skip-gram	GloVe
100	65.02	70.83	70.59	68.33
200	65.94	71.60	71.79	70.42
300	66.37	72.53	72.35	70.89
400	66.44	72.89	73.15	71.92
500	66.80	73.56	73.71	73.56
600	66.05	73.86	74.05	73.26

with 5-fold cross-validation

H.-C. Tseng et al., "Classification of text readability based on representation learning techniques," the 26th Annual Meeting of the Society for Text and Discourse (ST&D 2016), 2016

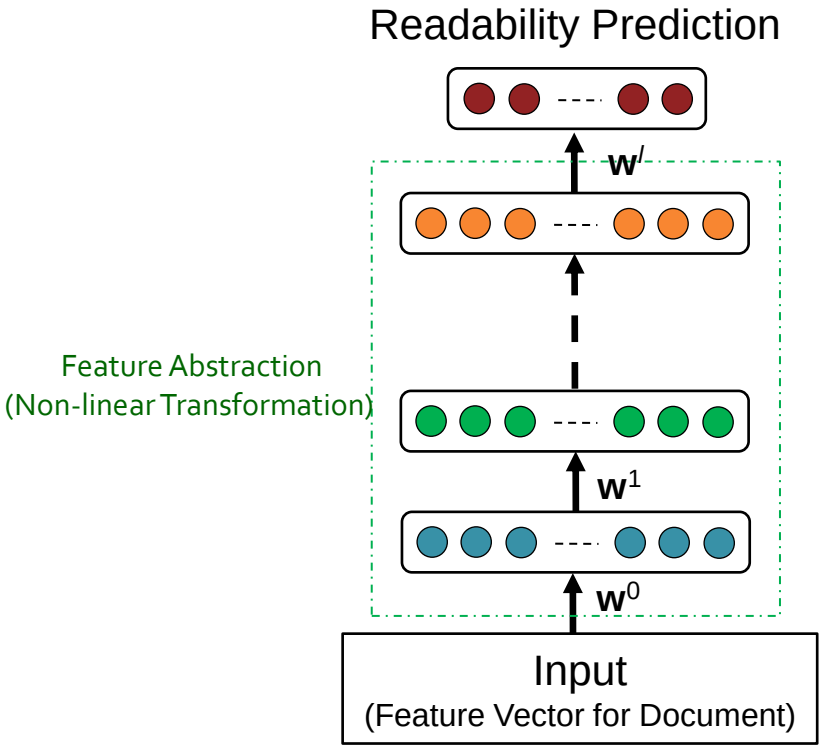
Comparison between LSA and WE (2/2)

- Synergies of Various Types of Embeddings

Dimensionality	CBOW +Skip-gram	GloVe +Skip-gram	GloVe +CBOW	GloVe+CBOW +Skip-gram
100	71.34	71.60	70.74	72.65
200	73.69	73.17	72.96	74.27
300	73.67	73.71	73.52	74.33
400	73.11	74.44	73.49	75.13
500	74.29	75.24	74.78	75.71
600	74.87	75.24	75.15	75.86
700D	75.00	75.15	74.83	75.99
800D	75.52	75.41	75.58	75.84

Deep Neural Networks (DNN)

- Leveraging DNN as the predictor and word/paragraph embeddings as document representations shows promising results



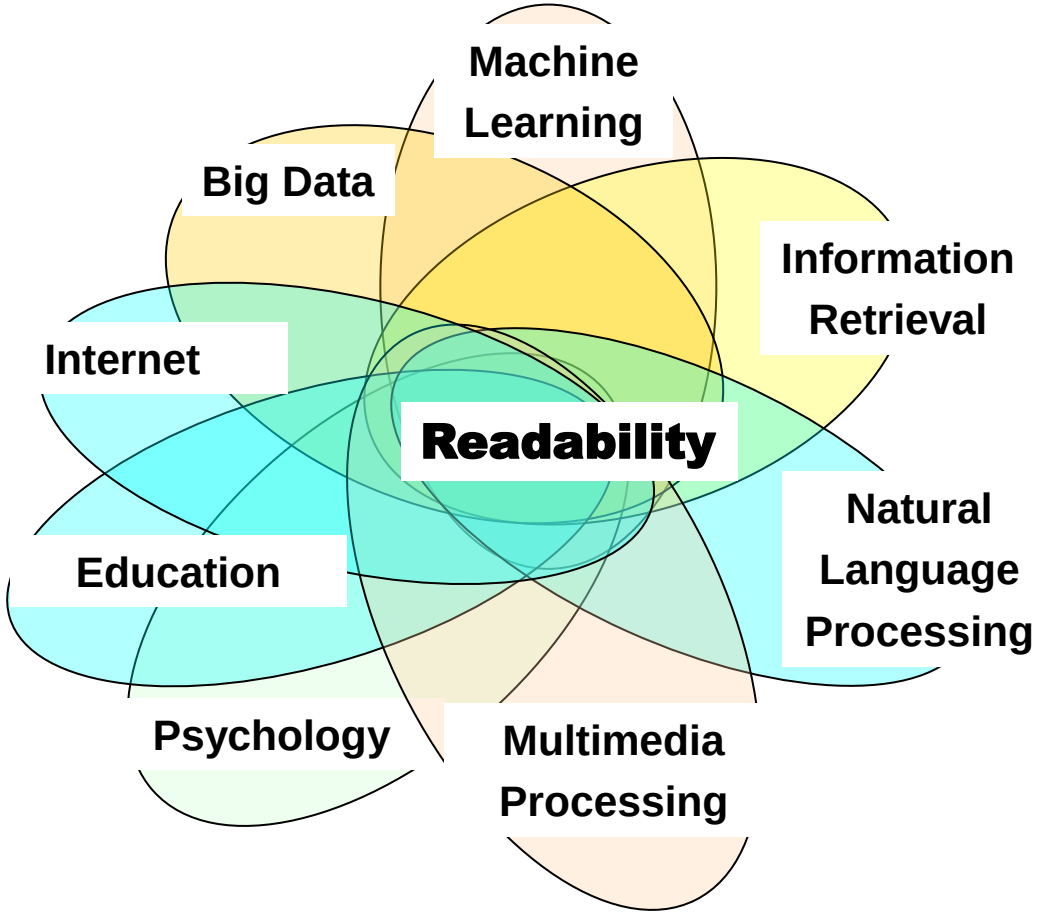
  - We are currently endeavoring to exploit more sophisticated DNN or deep learning techniques for use in readability assessment

Conclusion and Outlook (1/2)

- Readability assessment emerges to be an attractive realm of research that is confluence of multiple disciplines
- Many low-hanging "fruits" (e.g., features/models/tasks) have been taken, while high-hanging "fruits" are still difficult to achieve and demand extensive research and experimentation
- We should make as few assumptions as possible
  - Bag-of-words & bag-of-sentence assumptions
  - Feature independence assumption
  - …
- We may seek the possibility to extend readability assessment research from text to speech

Conclusion and Outlook (2/2)

- *Exploring Known Unknowns vs. Exploring Unknown Unknowns*



Thank You!