ISciDE 2012

Nanjing

Autoencoder for Polysemous Word

Wei-Chen Cheng
Jiun-Wei Liou
Daw-Ran Liou
Cheng-Yuan Liou *

Dept. of Computer Sci and Information Eng
National Taiwan University

Introduction & review

Generating a code for each word

Modeling word perception using the Elman network. Cheng-Yuan Liou, Jau-Chi Huang, Wen-Chie Yang: Neurocomputing 71 (2008) 3150– 3157

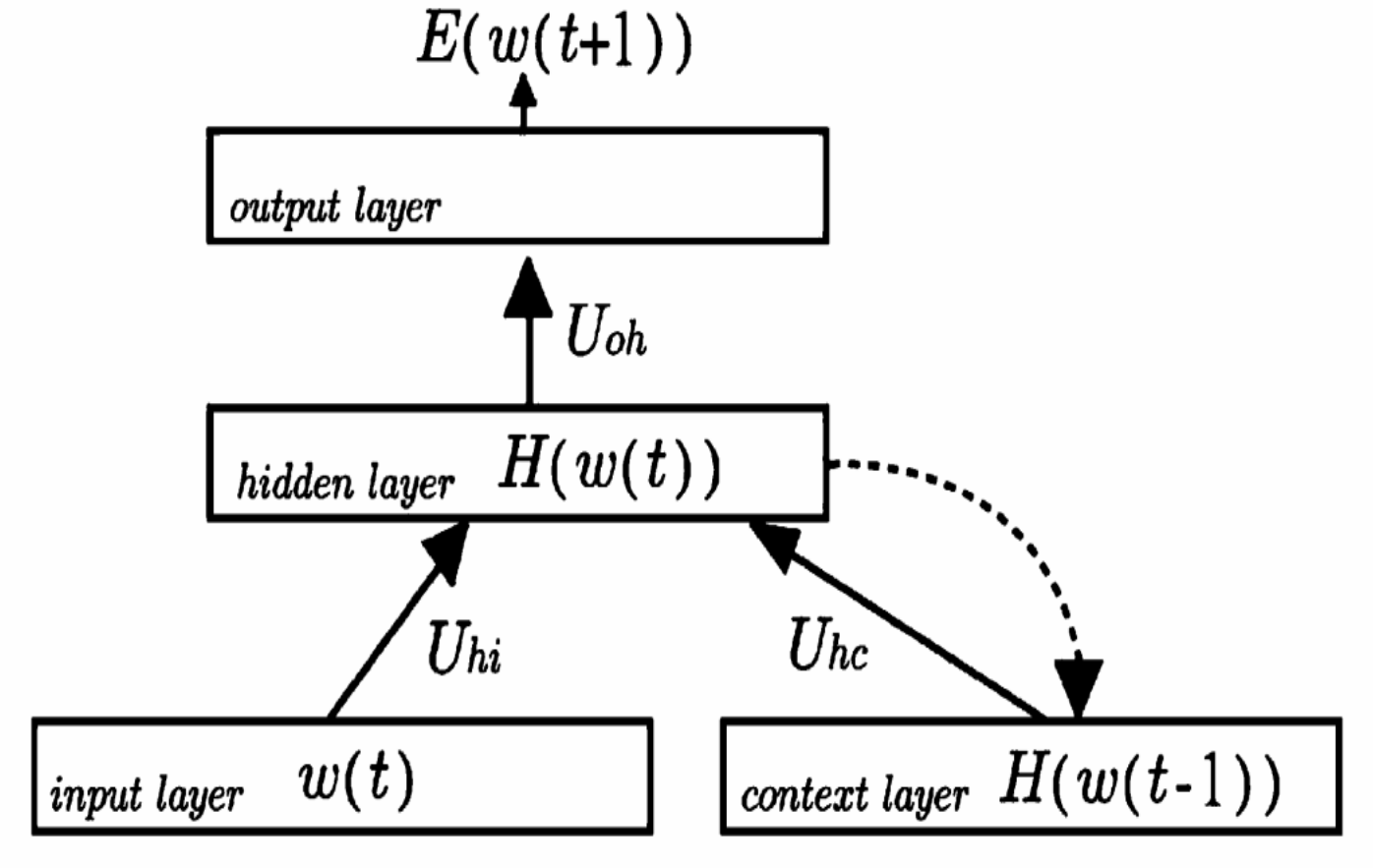
Generating attributes for each word by linguistics experts

Bank = attribute 1 = water,
attribute 2 = earth,
…,
attribute R = … ,

Cost and controversy for these manually assigned attributes.

Generating them Automatically !

Predicting next word's attributes



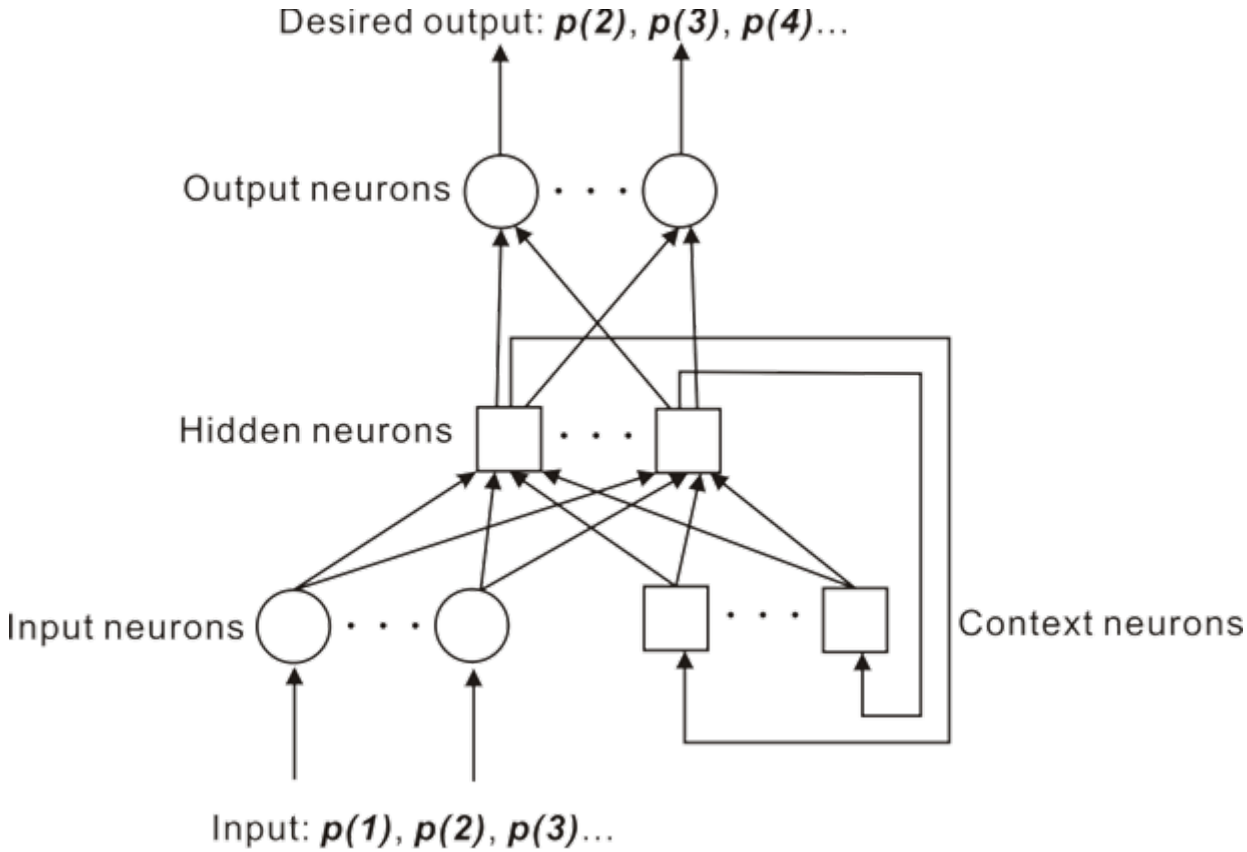


Figure: Illustration of Elman network.

Renewed attributes

- Updating network's weights after presentation of each word to reduce the prediction error.

- Averaged prediction for each word is used as the renewed attributes after each training pass.

Generated attributes

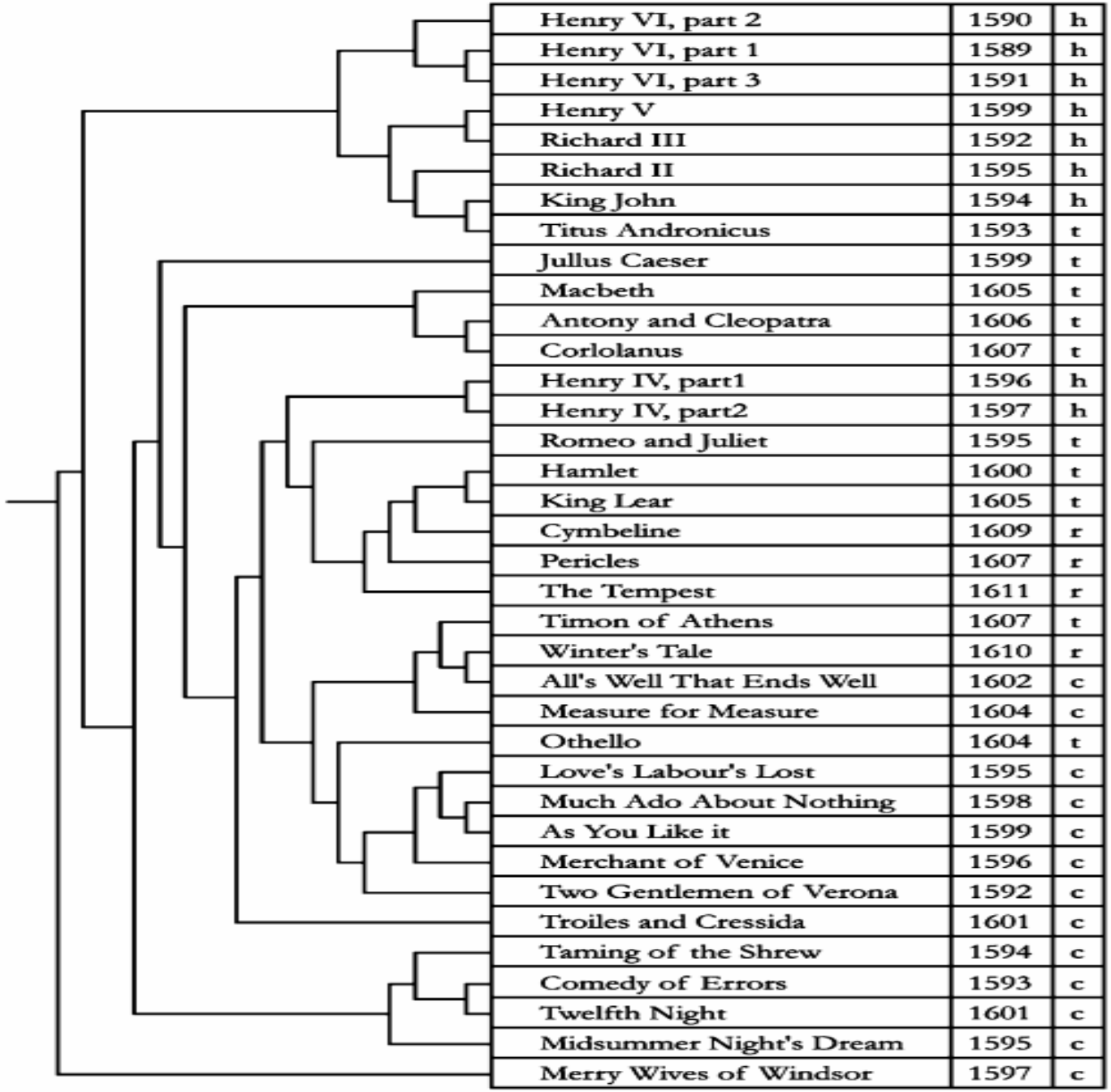
Semantic categorization

Indexing

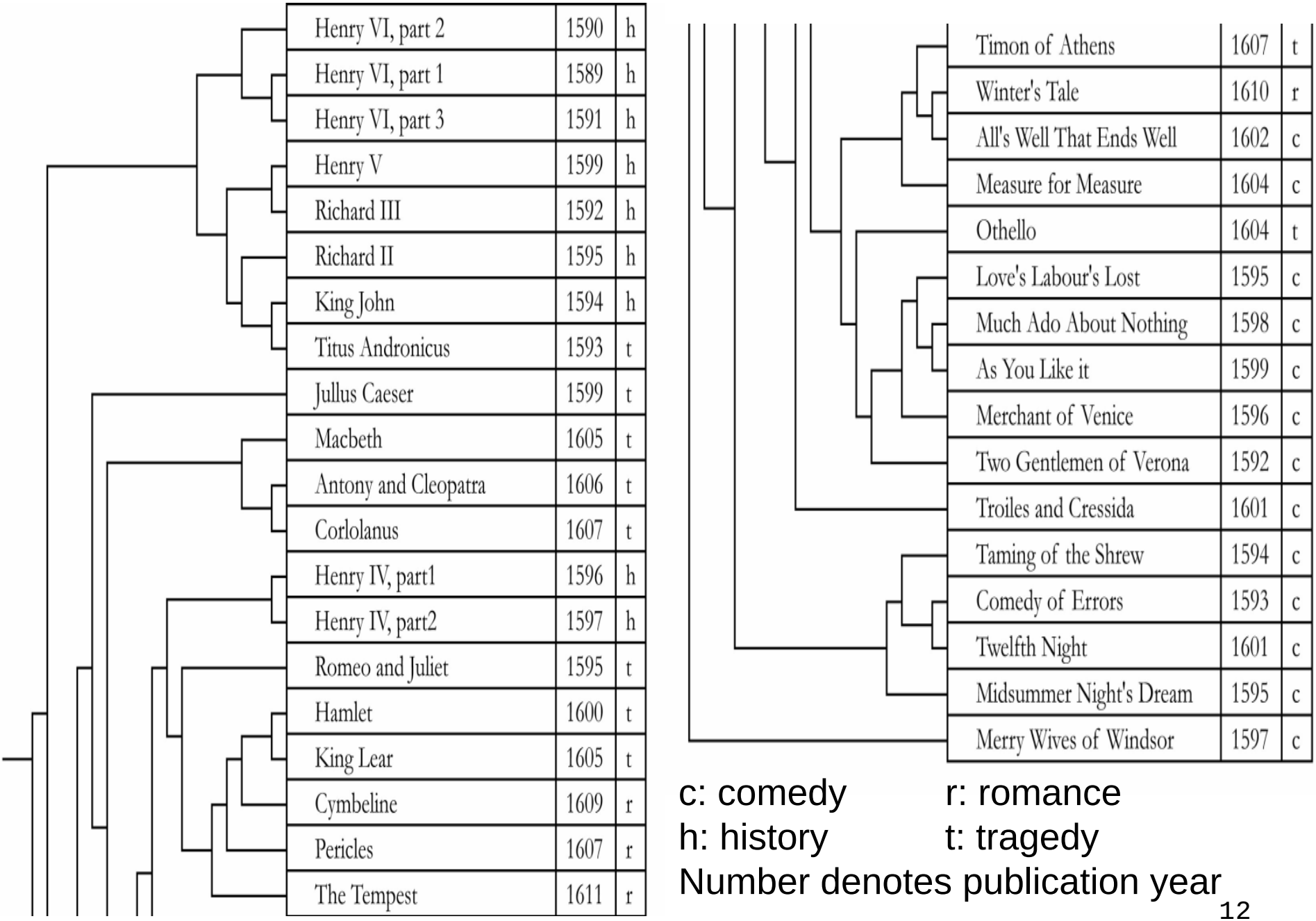
Ranking

Stylish analysis

Categorization of Shakespeare's 36 plays



Henry VI, part 2	1590	h
Henry VI, part 1	1589	h
Henry VI, part 3	1591	h
Henry V	1599	h
Richard III	1592	h
Richard II	1595	h
King John	1594	h
Titus Andronicus	1593	t
Jullus Caeser	1599	t
Macbeth	1605	t
Antony and Cleopatra	1606	t
Corlolanus	1607	t
Henry IV, part1	1596	h
Henry IV, part2	1597	h
Romeo and Juliet	1595	t
Hamlet	1600	t
King Lear	1605	t
Cymbeline	1609	r
Pericles	1607	r
The Tempest	1611	r
Timon of Athens	1607	t
Winter's Tale	1610	r
All's Well That Ends Well	1602	c
Measure for Measure	1604	c
Othello	1604	t
Love's Labour's Lost	1595	c
Much Ado About Nothing	1598	c
As You Like it	1599	c
Merchant of Venice	1596	c
Two Gentlemen of Verona	1592	c
Troiles and Cressida	1601	c
Taming of the Shrew	1594	c
Comedy of Errors	1593	c
Twelfth Night	1601	c
Midsummer Night's Dream	1595	c
Merry Wives of Windsor	1597	c



c: comedy r: romance
h: history t: tragedy
Number denotes publication year

Indexing result without keywords

http://red.csie.ntu.edu.tw/demo/literal/SAS.htm

Query	Search result, Shakespeare plays
she loves kiss	**BENVOLIO**: Tut, you saw her fair, none else being by herself poised with herself in either eye; but in that crystal scales let there be weigh.d. Your lady.s love against some other maid that I will show you shining at this feast, and she shall scant show well that now shows best. -Romeo and Juliet
armies die in blood	**MARCUS AND RONICUS**: Which of your hands hath not defended Rome, and rear.d aloft the bloody battle-axe, writing destruction on the enemy.s castle? O, none of both but are of high desert my hand hath been but idle; let it serve. To ransom my two nephews from their death; then have I kept it to a worthy end. -Titus Andronicus

Ranking Shakespeare's 36 plays

Authorship

Title	Year	Title	Year
Henry VI, part 1	1589	Henry V	1599
Henry VI, part 2	1590	Julius Caeser	1599
Henry VI, part 3	1591	As You Like It	1599
Richard III	1592	Hamlet	1600
Two Gentlemen of Verona	1592	Twelfth Night	1601
Comedy of Errors	1593	Troiles and Cressida	1601
Titus Andronicus	1593	All's Well That Ends Well	1602
Taming of the Shrew	1594	Measure for Measure	1604
King John	1594	Othello	1604
Love's Labour's Lost	1595	King Lear	1605
Richard II	1595	Macbeth	1605
Romeo and Juliet	1595	Antony and Cleopatra	1606
Midsummer Night's Dream	1595	Coriolanus	1607
Merchant of Venice	1596	Timon of Athens	1607
Henry IV, part 1	1596	Pericles	1607
Merry Wives of Windsor	1597	Cymbeline	1609
Henry IV, part 2	1597	Winter's Tale	1610
Much Ado About Nothing	1598	The Tempest	1611

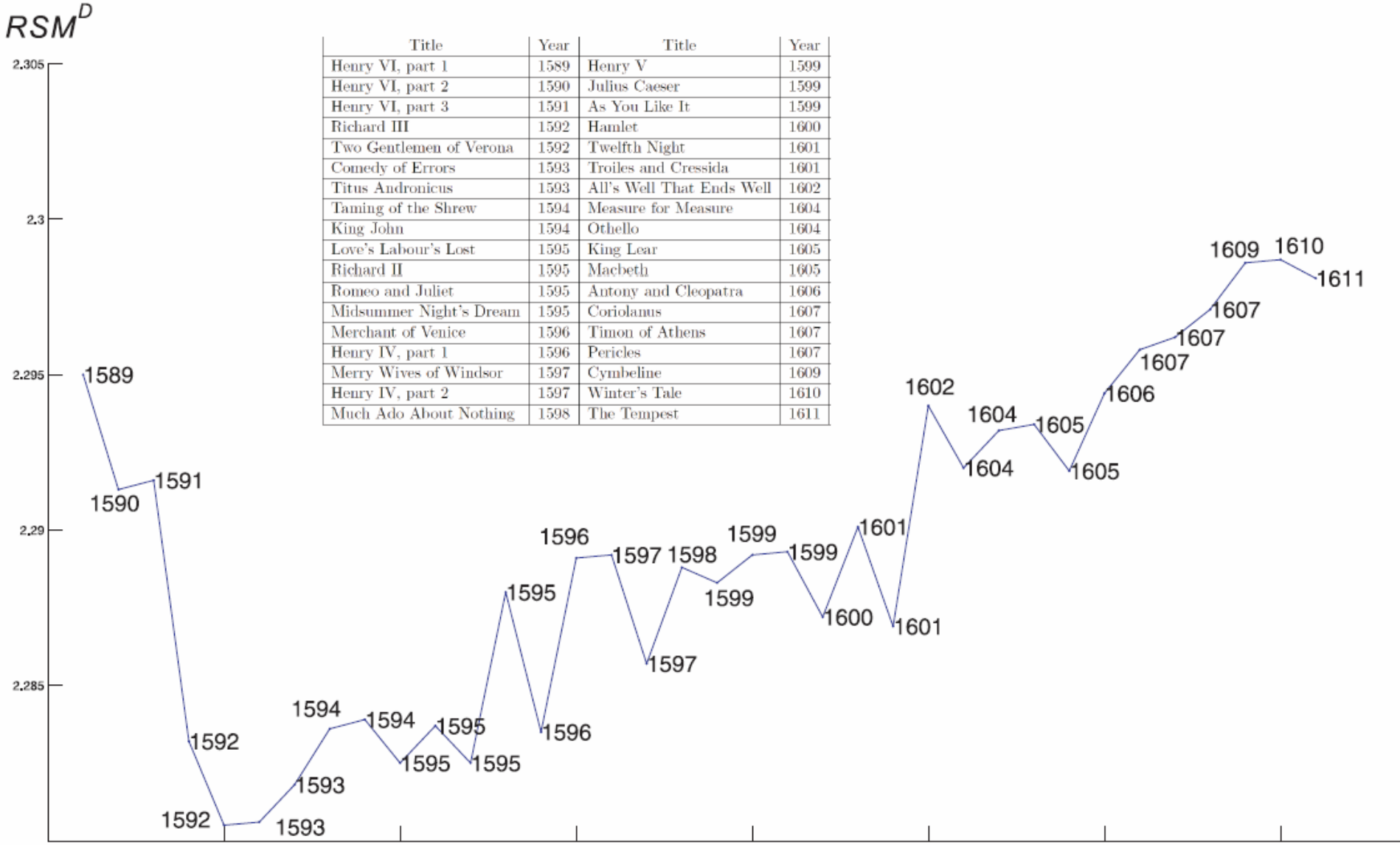


Figure 5-4: The RSM^D of each Shakespeare's work over time.

Stylish analysis

Table: 'RSM^D' values of William Shakespeare's plays

Title	Year	Code WS	Code JF	Code MT	Code SL	Grade
Henry VI, part 1	1589	2.2950	1.3379	1.1235	1.1424	74.24
Henry VI, part 2	1590	2.2913	1.4015	1.2489	1.2625	78.44
Henry VI, part 3	1591	2.2916	1.4393	1.3238	1.2750	80.17
Richard III	1592	2.2832	1.4639	1.3378	1.3084	81.06
Two Gentlemen of Verona	1592	2.2805	1.5195	1.0887	1.3018	78.21
Comedy of Errors	1593	2.2806	1.6138	0.8666	1.2728	75.98
Titus Andronicus	1593	2.2818	1.4603	1.0045	1.1484	74.16
Taming of the Shrew	1594	2.2836	1.4448	0.7780	1.2535	72.23
King John	1594	2.2839	1.4601	1.3679	1.3551	82.07
Love's Labour's Lost	1595	2.2825	1.4962	1.1593	1.2910	78.76
Richard II	1595	2.2837	1.4791	1.3059	1.3137	80.90
Romeo and Juliet	1595	2.2825	1.5085	1.3037	1.2458	80.33
Midsummer Night's Dream	1595	2.2880	1.3977	0.9564	1.1792	73.13
Merchant of Venice	1596	2.2835	1.4813	1.2138	1.3739	80.45
Henry IV, part 1	1596	2.2891	1.4723	1.2357	1.2910	79.57
Merry Wives of Windsor	1597	2.2892	1.4904	1.2515	1.2705	79.76
Henry IV, part 2	1597	2.2857	1.4954	1.2951	1.3690	81.74
Much Ado About Nothing	1598	2.2888	1.5120	1.2034	1.3851	80.93
Average *RSM^D*		2.2858	1.4708	1.1702	1.2799	

Table: 'RSM^{D}' values of William Shakespeare's works

Title	Year	Code WS	Code JF	Code MT	Code SL	Grade
Henry V	1599	2.2883	1.4409	1.1987	1.3040	78.80
Julius Caeser	1599	2.2892	1.4500	0.9923	1.2715	75.61
As You Like It	1599	2.2893	1.5115	1.2497	1.4021	81.81
Hamlet	1600	2.2872	1.4710	1.3317	1.3843	82.14
Twelfth Night	1601	2.2901	1.4949	1.2007	1.3469	80.16
Troiles and Cressida	1601	2.2869	1.4068	0.7687	1.1462	70.17
All's Well That Ends Well	1602	2.2940	1.5647	1.4159	1.4103	85.01
Measure for Measure	1604	2.2920	1.5114	1.2466	1.3190	80.66
Othello	1604	2.2932	1.4826	1.1933	1.2732	78.92
King Lear	1605	2.2934	1.5260	1.3456	1.3121	82.17
Macbeth	1605	2.2919	1.4771	1.1931	1.2977	79.16
Antony and Cleopatra	1606	2.2944	1.4473	1.1919	1.3102	78.94
Coriolanus	1607	2.2958	1.5119	1.2147	1.3306	80.43
Timon of Athens	1607	2.2962	1.5284	1.0982	1.3295	79.01
Pericles	1607	2.2971	1.5141	1.3184	1.4117	83.02
Cymbeline	1609	2.2986	1.5178	1.2779	1.3663	81.90
Winter's Tale	1610	2.2987	1.5310	1.2886	1.3973	82.65
The Tempest	1611	2.2981	1.4847	1.2412	1.2822	79.80
Average *RSM*^{D}		2.2930	1.4929	1.2093	1.3275	

Polysemous word

- Difficulty of concept

- Many-to one is a function,
  one to many isn’t a function.

Polysemous word

Building a meaning pool matrix, *M*, for each word.

M contains *B* meanings (*B* candidates) in its column vectors.

B=2 for Polysemous word ‘bank’

	money	river
Bank =	attribute 1, attribute 2, …, attribute R,	attribute 1, attribute 2, …., attribute R

Predicting next word's meaning

The code of the best predicated meaning in '*M*' is used for the next input word.

Renewed attributes

- Updating network's weights after presentation of each word to reduce the prediction error.

- Averaged prediction for a specific meaning of the next word is used as the renewed attributes of that meaning after each training pass.

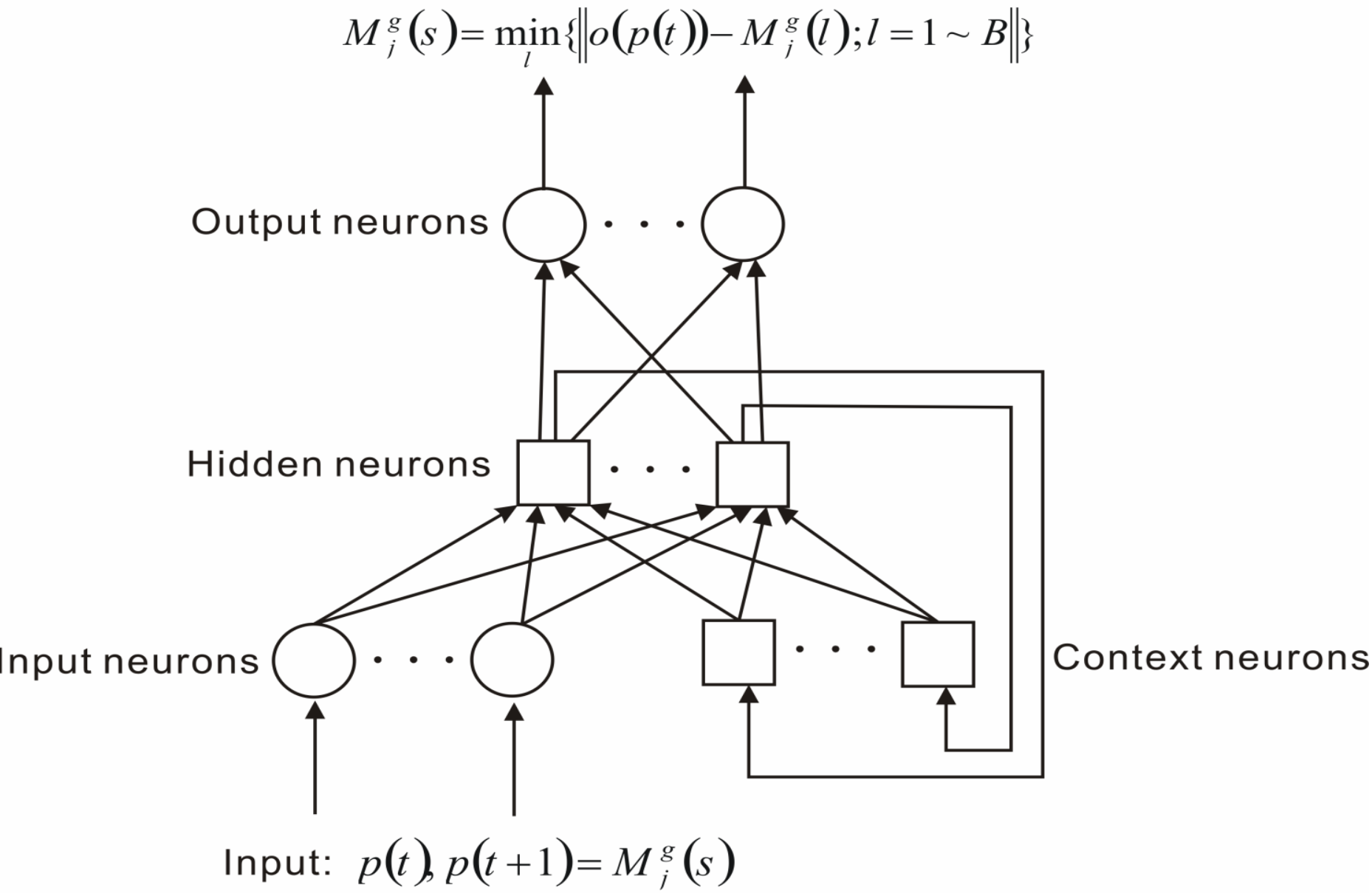


Figure: Illustration of Elman network for multi-code.

Experiments

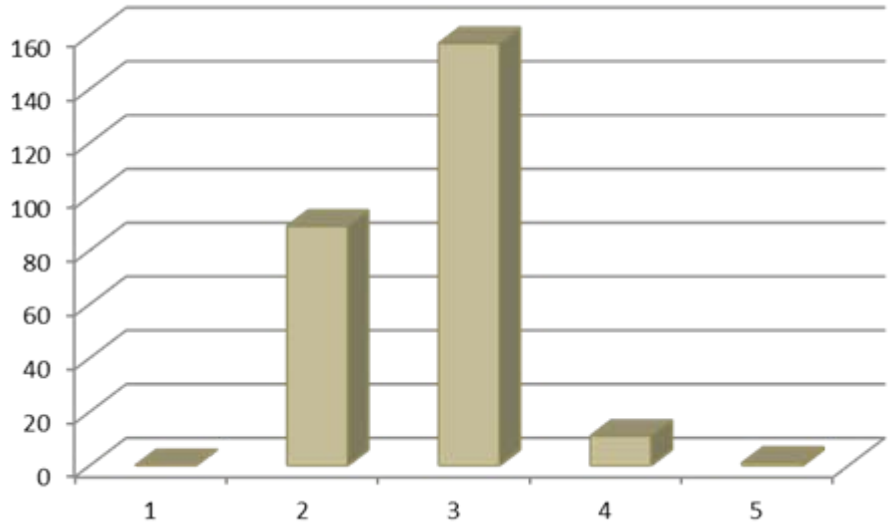
Dream of the Red Chamber
紅樓夢

Romance of the **Three Kingdoms**
三國演義

Red Chamber has more than 841 thousands of characters and uses 5069 different Chinese characters

Pick 246 words (<5%) with fq in {fq $\geq$ 300 and $\leq$ 1200}

Dream of Red Chamber
246 words {fq $\geq$ 300 and $\leq$ 1200}



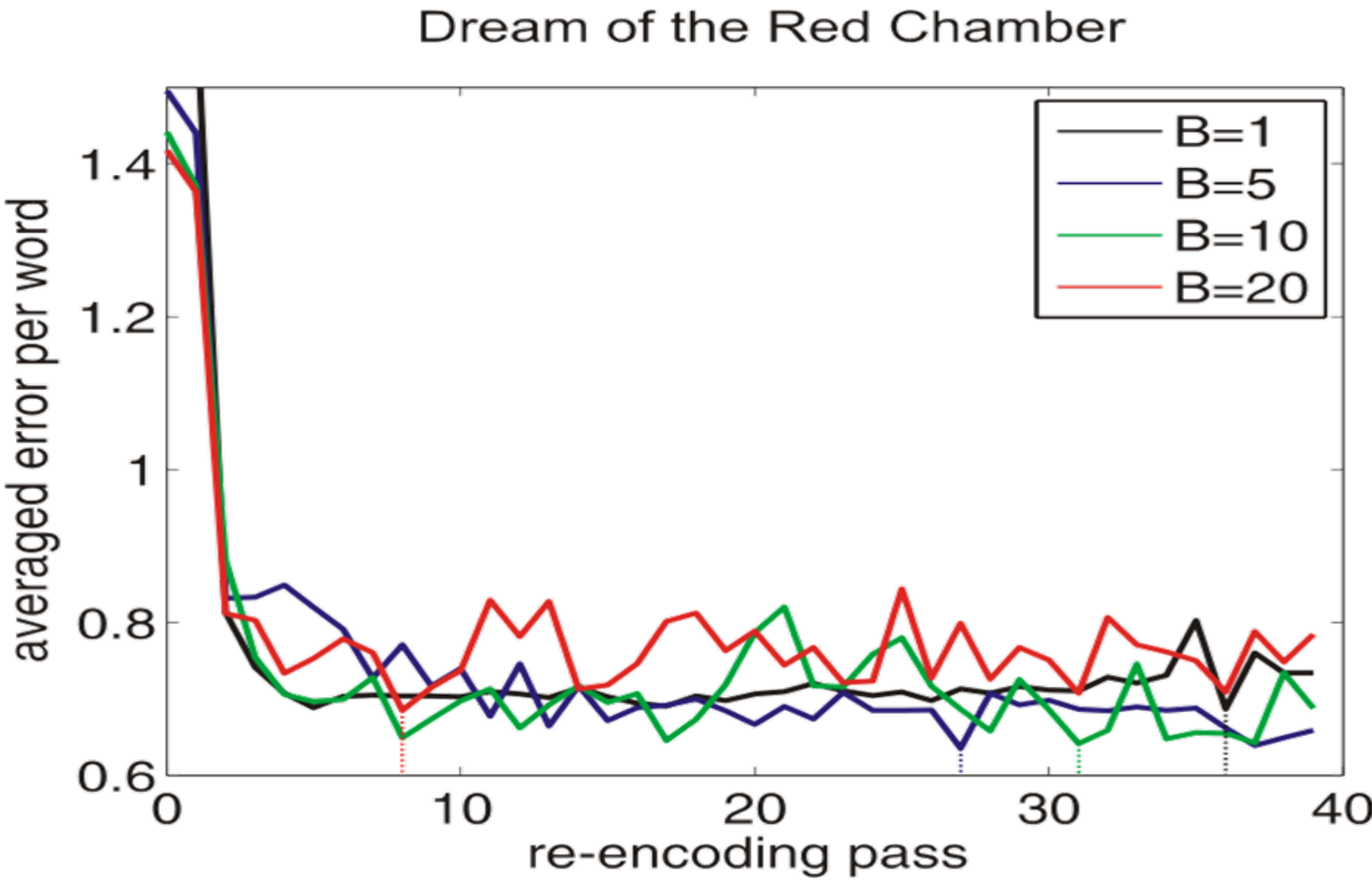
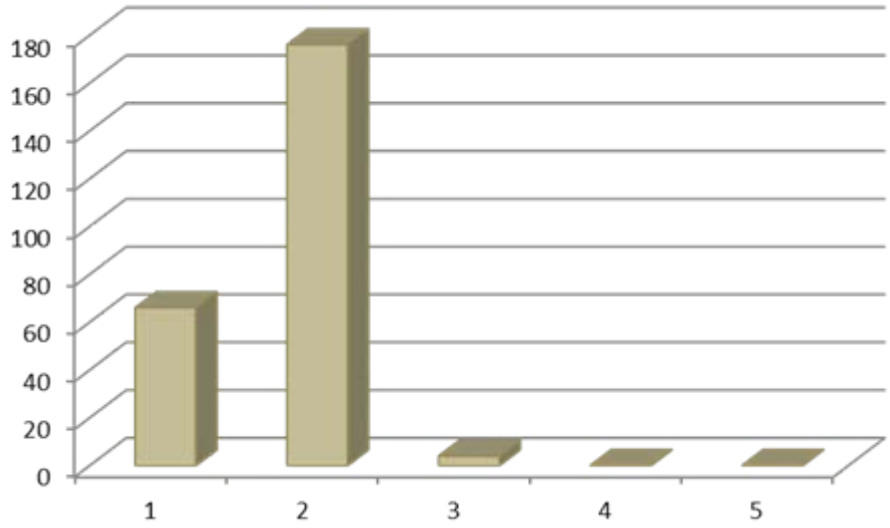


Figure: Training errors using different pool sizes. Color vertical lines mark the minimum pass.

Three Kingdoms has more than 570 thousands of characters and uses 5071 Chinese characters.

Pick 258 words in
{fq$\geq$ 225 and $\leq$ 525}

Romance of the Three Kingdoms 258 words {fq≥ 225 and ≤ 525}



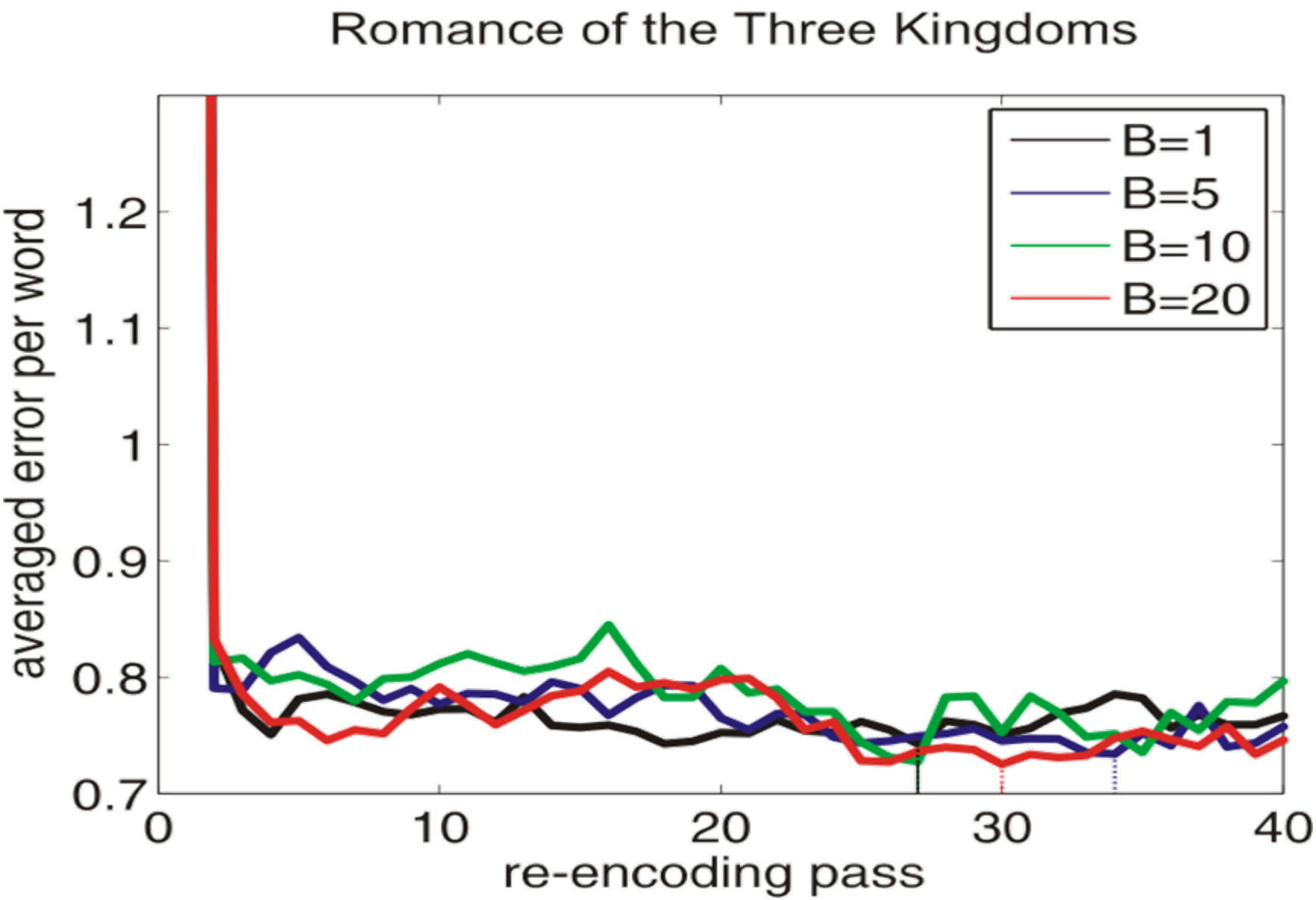


Figure: Training errors using different pool sizes. Color vertical lines mark the minimum pass.

Table: Characters have multiple codes. The total number of meanings of a character is labeled next to its character.

紅樓夢	三國演義
張(2)、真(4)、輕(2)、 花(4)、分(2)、長(3)、 把(3)、思(2)、答(2)、 紅(2)、經(3)、方(5)、	車(2)、發(2)、差(2)、 合(2)、陣(2)、投(2)、 成(3)、禮(3)、飛(2)、 老(2)、騎(2)、勢(2)、

Table: Sentences in Red Chamber which contain the same character with two different meanings, s=1 and s=4.

紅樓夢

Sentences	Meaning
士隱接了看時，原來是塊鮮明美玉，上面字跡分明，鐫著「通靈寶玉」四字， 後面還有幾行（4）小字。	排
後有幾行（4）字跡，寫的是：霽月難逢，彩雲易散。心比天高，身為下賤。 風流靈巧招人怨。壽夭多因毀謗生，多情公子空辜念。	排
說畢，在前導引，大家攀藤撫樹過去。只見水上落花愈多，其水愈清，溶溶蕩 蕩，曲折縈迂。池邊兩行（4）垂柳，雜著桃杏，遮天蔽日，真無一些塵土。	排
說著，大家出來。行（1）不多遠，則見崇閣巍峨，層樓高起，面面琳宮合抱， 迢迢複道縈紆；青松拂檐，玉欄繞砌，金輝獸面，彩煥螭頭。	走
回頭再走，又有窗紗明透，門徑可行（1）；及至門前，忽見迎面也進來了 一群人，都與自己形相一樣，卻是一架玻璃大鏡相照。	走
後面方是八個太監抬著一頂金頂金黃繡鳳版輿，緩緩行(1)來。賈母等連忙 路旁跪下。早飛跑過幾個太監來，扶起賈母、邢夫人、王夫人來。	走

Table: Sentences in Three Kingdoms which contain the same character with two different meanings, s=2 and s=4.

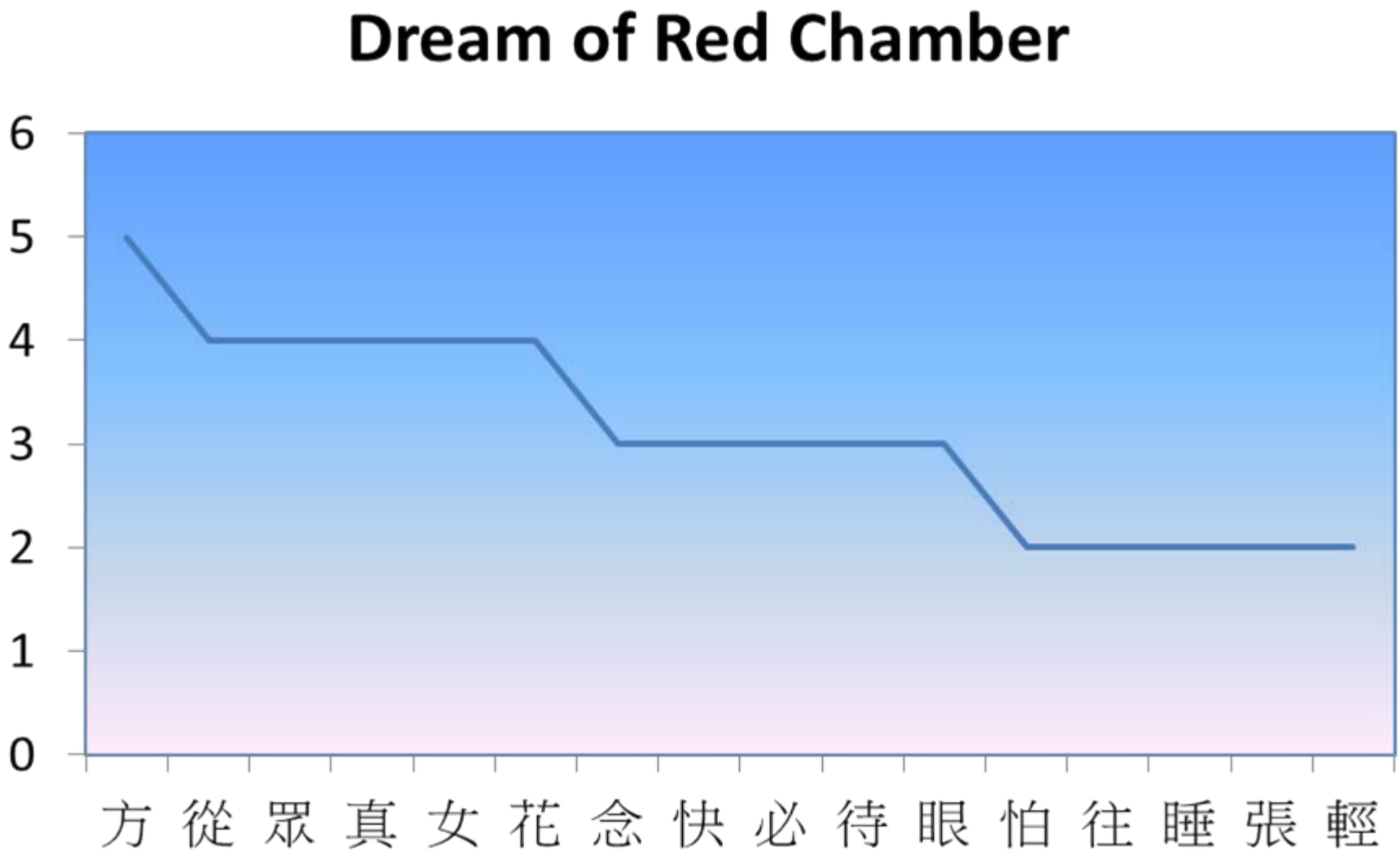
三國演義

Sentences	Meaning
張寶遣副將高昇出馬搦戰。玄德使張飛擊之。飛縱馬挺矛，與昇交戰，不數合（4），刺昇落馬。玄德麾軍直衝過去。	一場
堅曰：「汝言正合（2）吾意。明日便當託疾辭歸。」商議已定，密諭軍士勿得洩漏。	相配
操見其人威風凜凜，心中暗喜，分付典韋，今日且詐敗。韋領命出戰；戰到三十合（4），敗走回陣。壯士趕到陣門中，弓弩射回。	一場
漢以火德王，而明公乃土命也。許都屬土，到彼必興。火能生土，土能旺木：正合（2）董昭、王立之言。他日必有興者。	相配
操大喜曰：「君言正合（2）吾心。」次日，即表薦劉備領豫州牧。程昱諫曰：「劉備終不為人之下，不如早圖之。」	相配
操怒，使張遼出戰。張郃躍馬來迎。二將鬥了四五十合（4），不分勝負。曹操見了，暗暗稱奇。許褚揮刀縱馬，直出助戰。	一場

Table: Samples of two names having multiple codes.

紅樓夢	三國演義
周瑞家的不敢驚動，遂進裏間來，只見薛寶釵穿著家常衣服，頭上只散挽著簪兒，坐在炕裏邊，伏在小炕兒上同丫鬟鶯兒正描花樣子呢。見她進來，寶釵（２）便放下筆，轉過身來，滿面堆笑讓：「周姐姐坐。」....寶釵（４）笑道：「那裏的話！只因我那種病又發了兩天，所以靜養兩日。」....周瑞家的聽了點頭兒，因又說：「這病發了時到底覺怎樣？」寶釵（４）道：「也不覺甚什麼，只不過喘嗽些，吃一丸也就罷了。」....寶玉亦湊了上去，從項上摘了下來，遞在寶釵手內。寶釵（２）托於掌上，只見大如雀卵，燦若明霞，瑩潤如酥，五色花紋纏護。....王夫人忙叫了兩個人跟寶姑娘去。一時寶釵（２）取了衣服回來，只見寶玉在王夫人旁邊坐著垂淚。賈母因看見有個赤金點翠的麒麟，便伸手翻弄拿了起來，笑道：「這件東西，好像我看見誰家的孩子也戴著這麼一個。」寶釵（４）笑道：「史大妹妹有一個，比這個小些。」...	城中忽有一將，引數百人逕上城樓，大喝：「蔡瑁、張允，賣國之賊！劉使君乃仁德之人，今為救民而來投，何得相拒！」眾觀其人，身長八尺，面如重棗；乃義陽人也，姓魏，名延，字文長。當下魏延（２）輪刀砍死守門將士，開了城門，放下吊橋，大叫：「劉皇叔快領兵入城，共殺賣國之賊！」張飛便躍馬欲入。玄德急止之曰：「休驚百姓！」魏延（４）只管招呼玄德軍馬入城。只見城內一將飛馬引軍而出，大喝：「魏延（４）無名小卒，安敢造亂！認得我大將文聘麼！」魏延（４）大怒，挺槍躍馬，便來交戰。兩下軍兵在城邊混殺，喊聲大震。玄德曰：「本欲保民，反害民也！吾不願入襄陽！」孔明曰：「江陵乃荊州要地，不如先取江陵為家。」玄德曰：「正合吾心。」於是引著百姓，盡離襄陽大路，望江陵而走。襄陽城中百姓，多有乘亂逃出城來，跟玄德而去。魏延（２）與文聘交戰，從巳至未，手下兵卒，皆已折盡。延乃撥馬而逃，卻尋不見玄德，自投長沙太守韓玄去了。

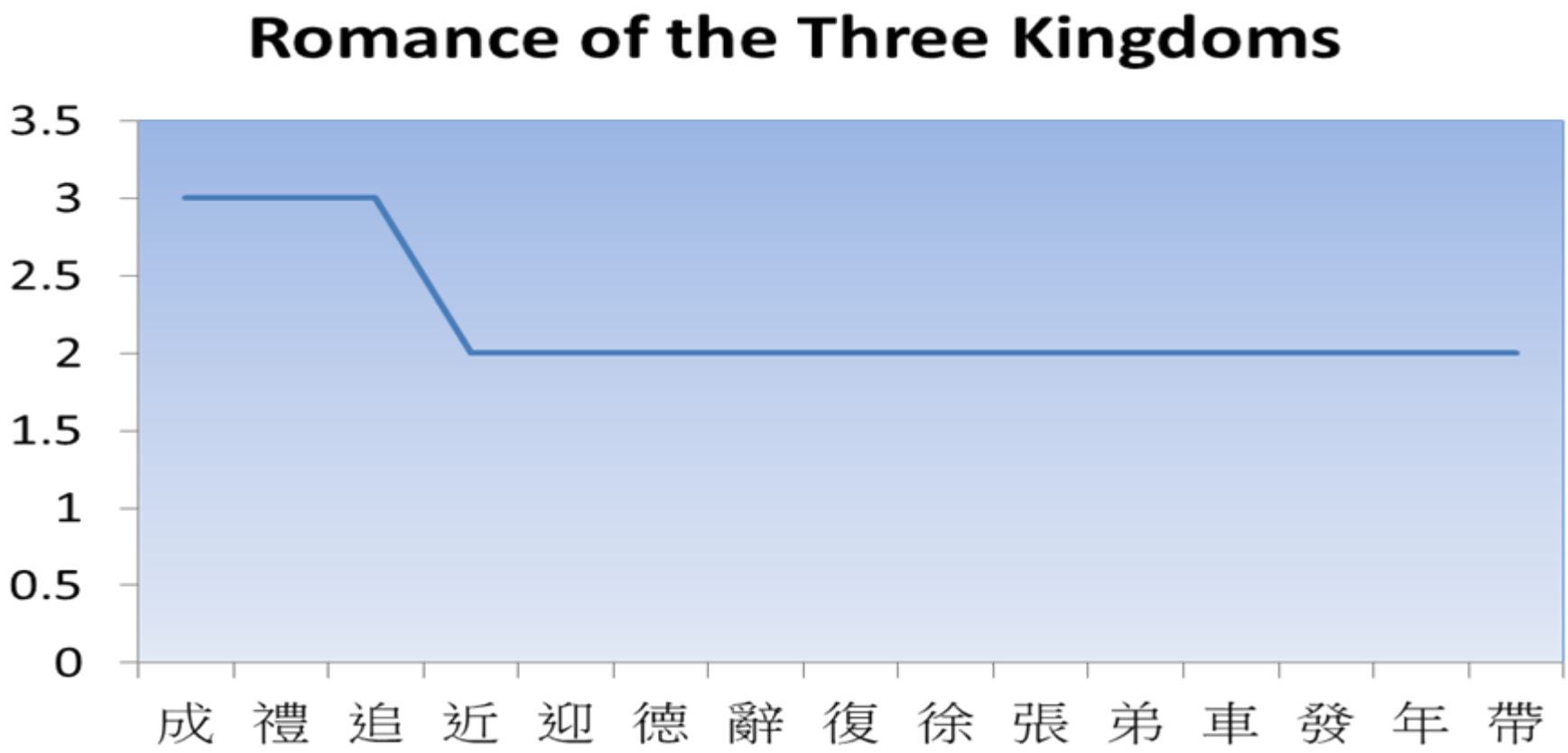
The number of meanings



Examples

- 話說王夫人見中秋已過，鳳姐病已比先減了，雖未大愈，然亦可出入行走得了，仍命大夫每日診脈服藥，又開了丸藥方(1)子來，配調經養榮丸。 (意：單子)
- 說著，便袖了這石，同那道人飄然而去，竟不知投奔何方(2)何捨。 (意：方向)
- 自取了筆硯紙墨出來，將方(3)才的詩，命她二人念著，遂從頭寫出來。(意：剛剛)
- 妙玉送至門外，看她們去遠，方(4)掩門進來。(意：才)
- 賈珍等拿了藥方(5)來，回明賈母原故，將藥方放在桌上出去，不在話下。(意：帖)

The number of meanings



Examples

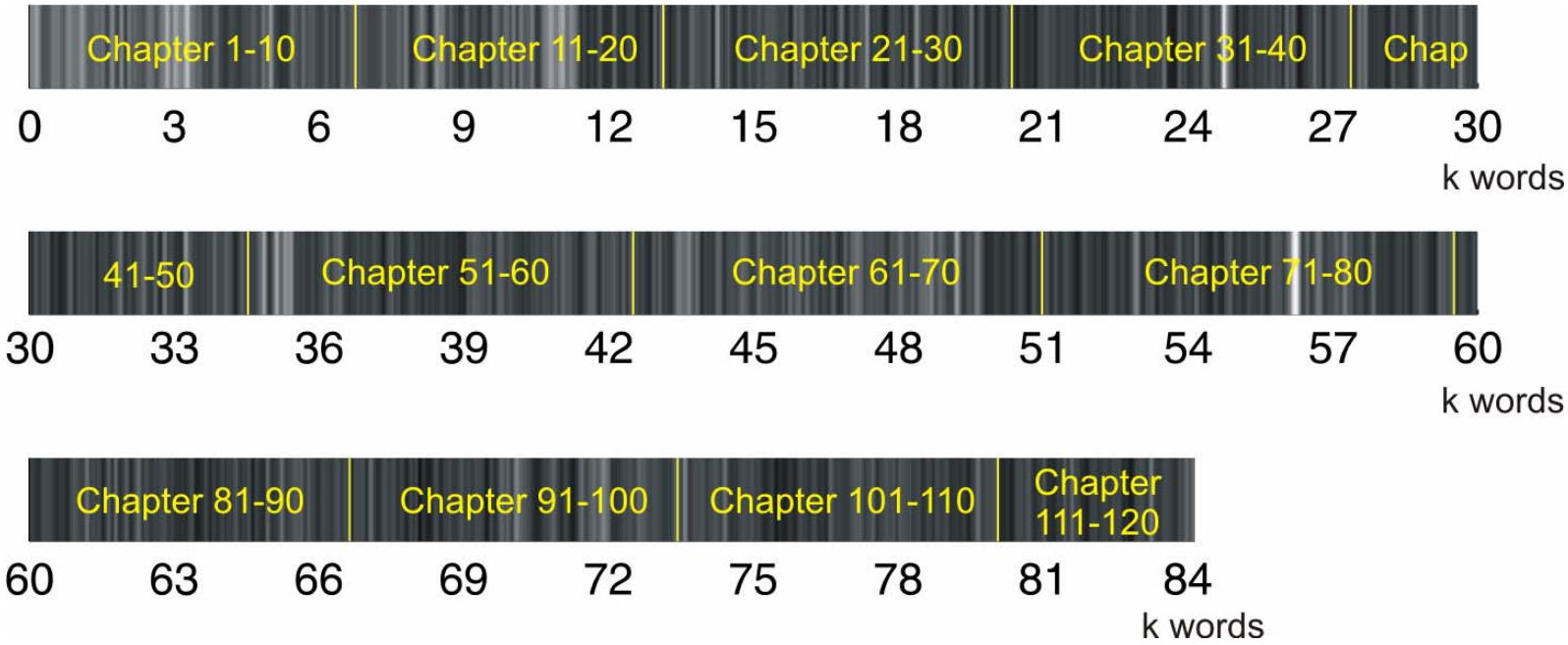
- 是非成(1)敗轉頭空：青山依舊在，幾度夕陽紅。 (意：成功)
- 孔明曰：「曹操幼子曹植，字子建，下筆成(4)文。操嘗命作一賦，名曰銅雀臺賦。賦中之意，單道他家合爲天子，誓取二喬。」 (意：形成)
- 成(5)功不必添蛇足，討賊猶思奮虎威。 (意：勝利)

Stylish analysis

Authorship

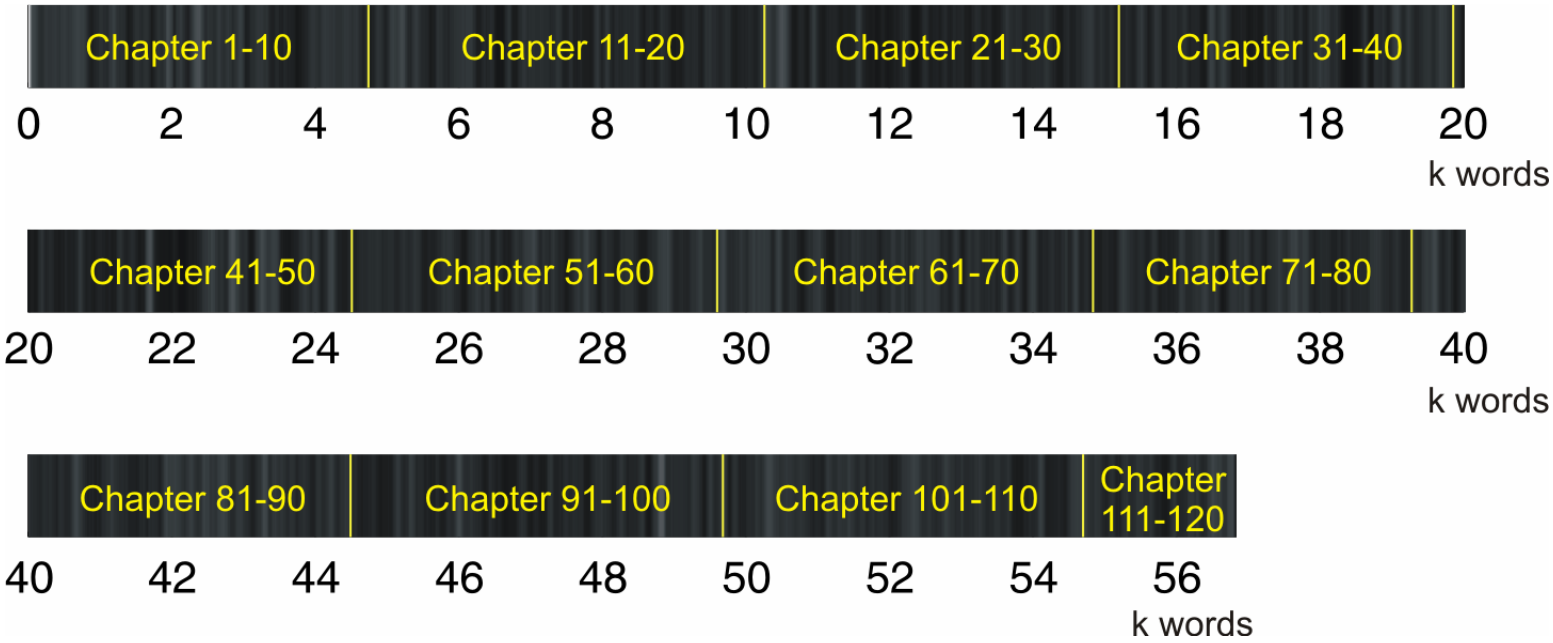
Dream of Red Chamber

- Prediction error along each word:



Romance of the Three Kingdoms

- Prediction error along each word:



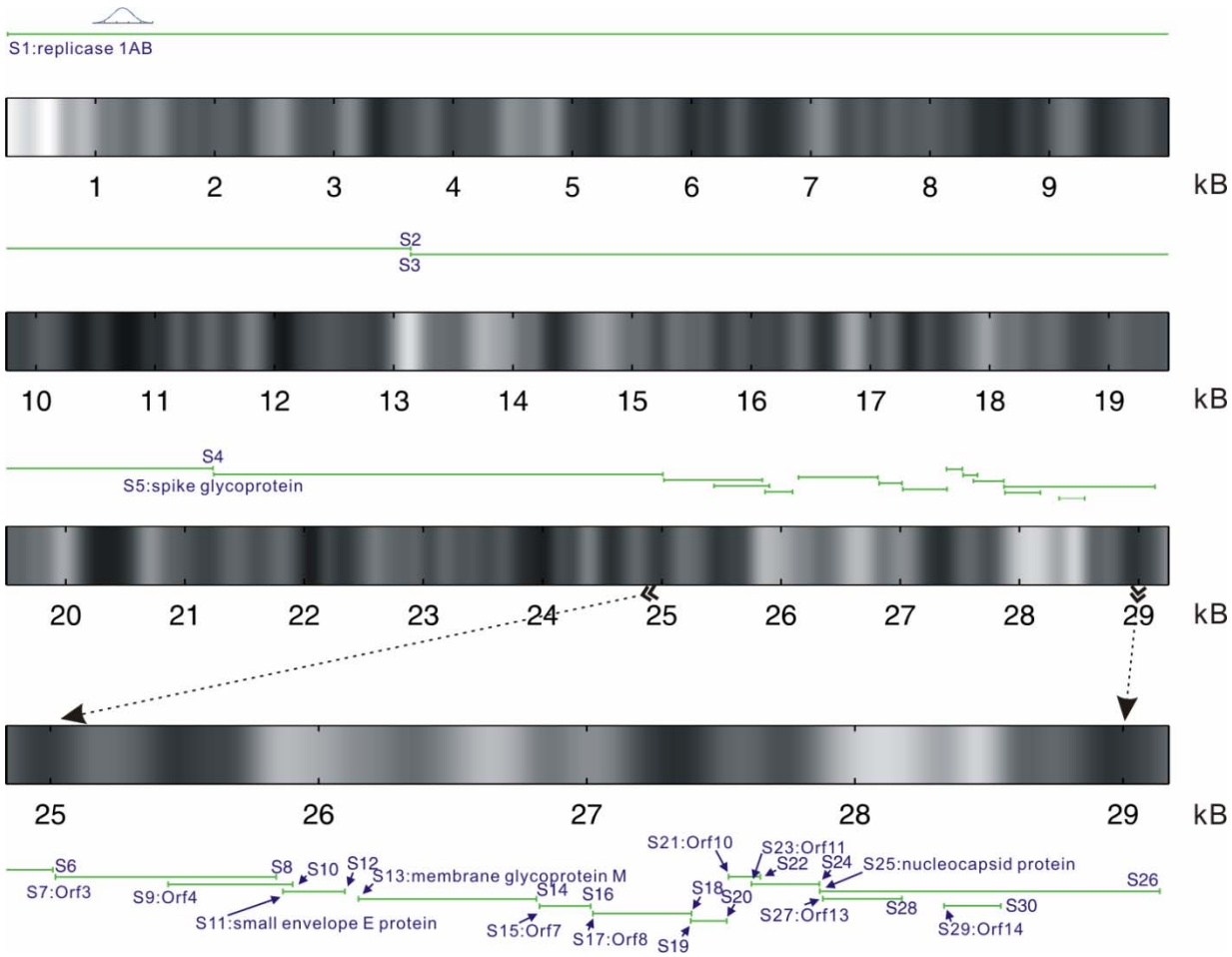
Summary

- Context-based method (Changing scenario)
- Symbol–free sequence
- Meaning of a learned attribute can be calibrated by its similar words.
- Predicating the next word (symbol) of a given word sequence.

Applications

- Stylish analysis
- Authorship
- Semantic indexing,
  ranking, and categorization
- Internet
- DNA, gene, or protein
- Cryptography
- Ancient language, machine translation

SARS ‘‘AY274119.3’’ genome. ‘white represents the largest error’



Influenza (1918)

- ACCESSION: AF116575.1

- (100)...GACACAGTACTCGAAAAGAATGTGACCGTGACACACTCTGTTAACCTGCTC...(150)
- (100)...1121211253531151555155555325555353535552512355222...(150)

- (500)...GGCTGACAAAGAAGGGAAGCTCATACCCAAAGCTTAGCAAGTCCTATGTGA...(550)
- (500)...15221121511155155113535253235155225113511223515555 5...(550)

- (1000)...GGACTAAGAAACATTCCATCTATTCAATCCAGGGGTCTATTTGGAGCCATT...(1050)
- (1000)...1553515551535253215351522152235515122525215552352 5...(1050)

A: 1, 5
T: 2, 5
C: 2, 3
G: 1, 5

Influenza (2009)

- ACCESSION: FJ966082.1

- (100)...GACACAGTACTAGAAAAGAATGTAACAGTAACACACTCTGTTAACCTTCTA...(150)
- (100)...13434315345313333133515334315334343454155334455453...(150)

- (500)...GGCTAGTTAAAAAAGGAAATTCATACCCAAAGCTCAGCAAATCCTACATTA...(550)
- (500)...11453155333333113335543534443331454314333544534355...(550)

- (1000)...GGATTGAGGAATATCCCGTCTATTCAATCTAGAGGCCTATTTGGGGCCATT...(1050)
- (1000)...11355131133535444154535543354531311445355511114435...(1050)

A: 1, 3
T: 3, 5
C: 4
G: 1

Detailed techniques and settings in ISclDE 2012 paper.

Museum of Cao Xueqin

- Born and grown in Nanjing

  1715 or 1724 — 1763 or 1764

- 曹雪芹故居 江宁织造府 (大行宫)

Thanks

http://www.csie.ntu.edu.tw/~cyliou/