Of Space Piracy and Secret Baby Romances: Deep Multi-View Book Representations and a Scalable Evaluation Framework

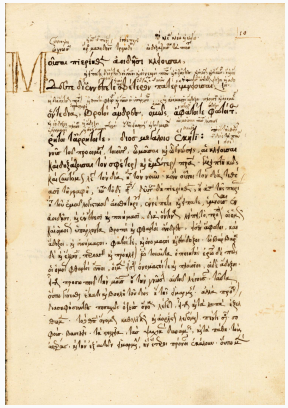
Lea Frermann

August 24, 2017

joint work with Gyuri Szarvas

amazon

A Short History of Books



Library of Alexandria

'concurrent copying'

Ancient Egypt

Ancient Greece

Roman Empire

Middle Ages

3000 BC

2000 BC

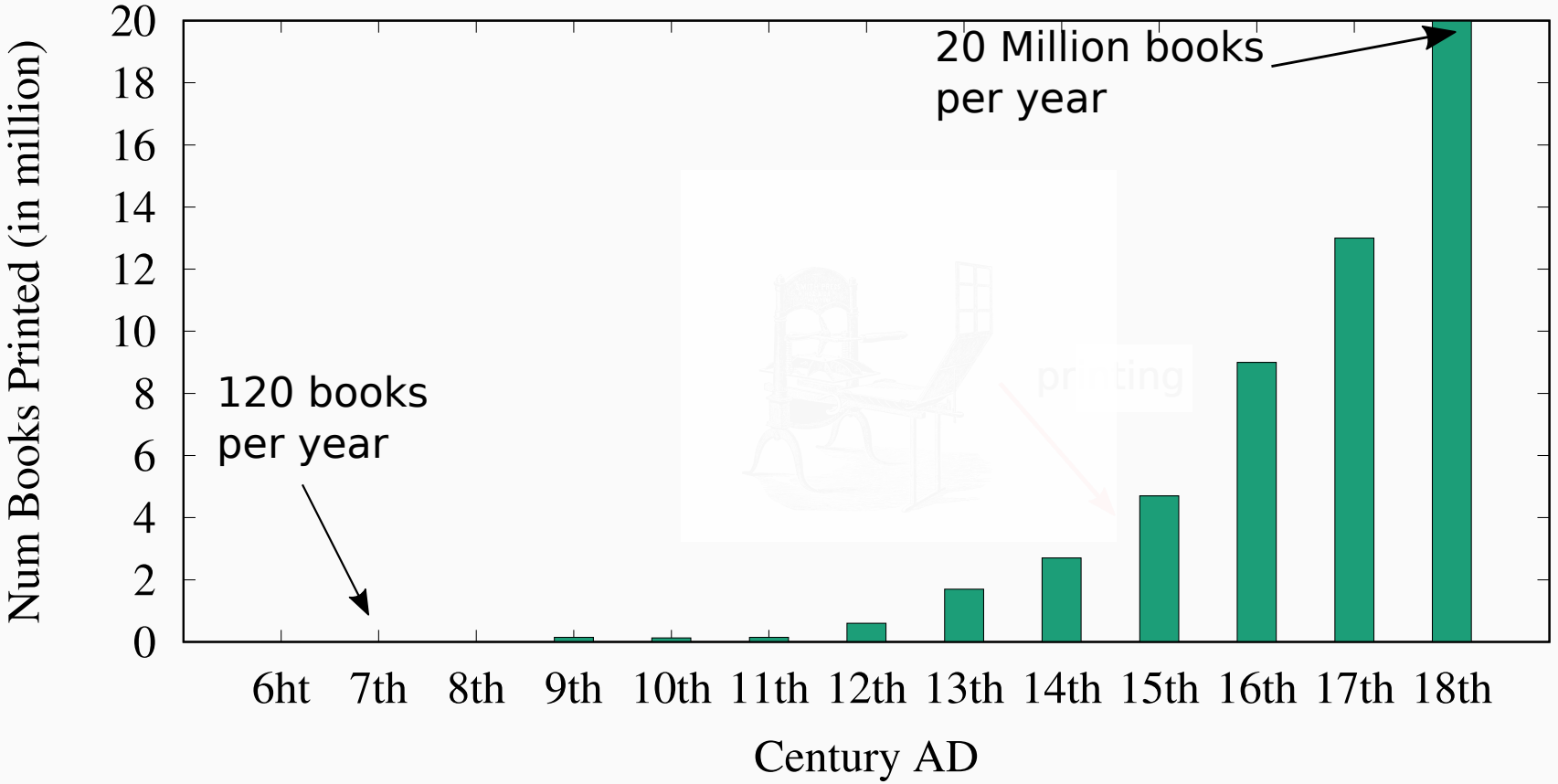
1000 BC

0 AD

500 AD

A Short History of Books

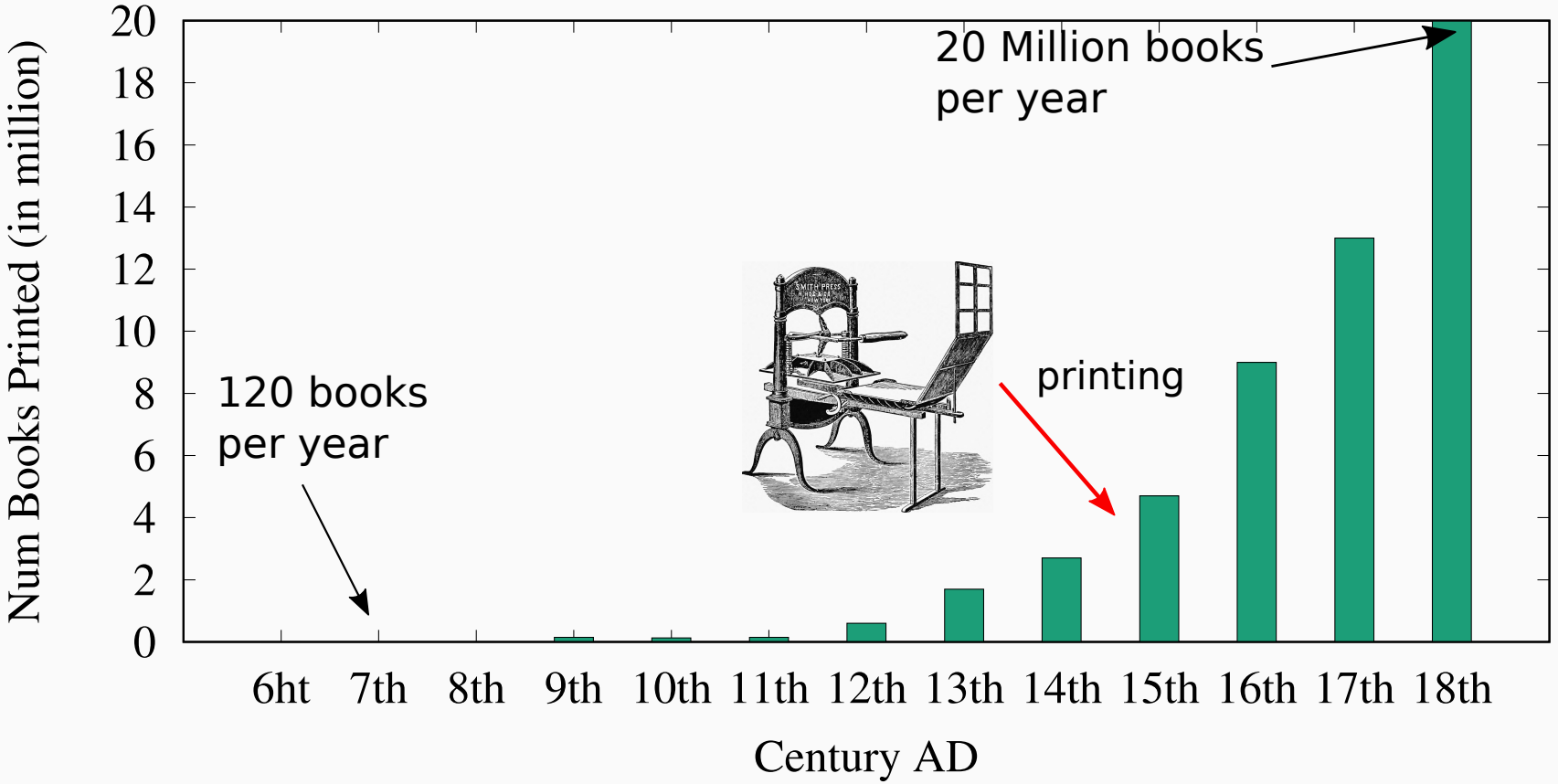
Book Production in Europe



Data from https://ourworldindata.org/books/

A Short History of Books

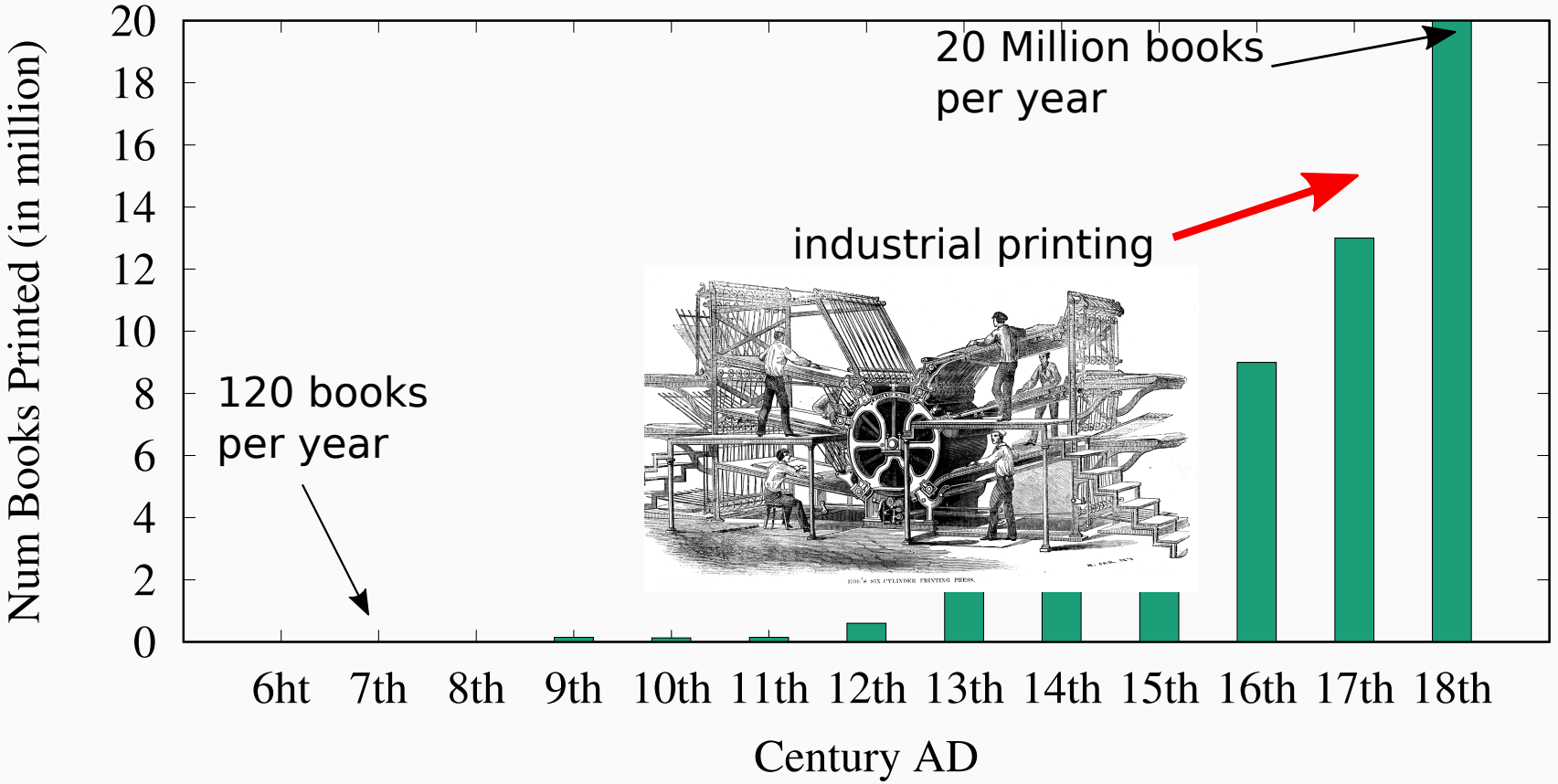
Book Production in Europe



Data from https://ourworldindata.org/books/

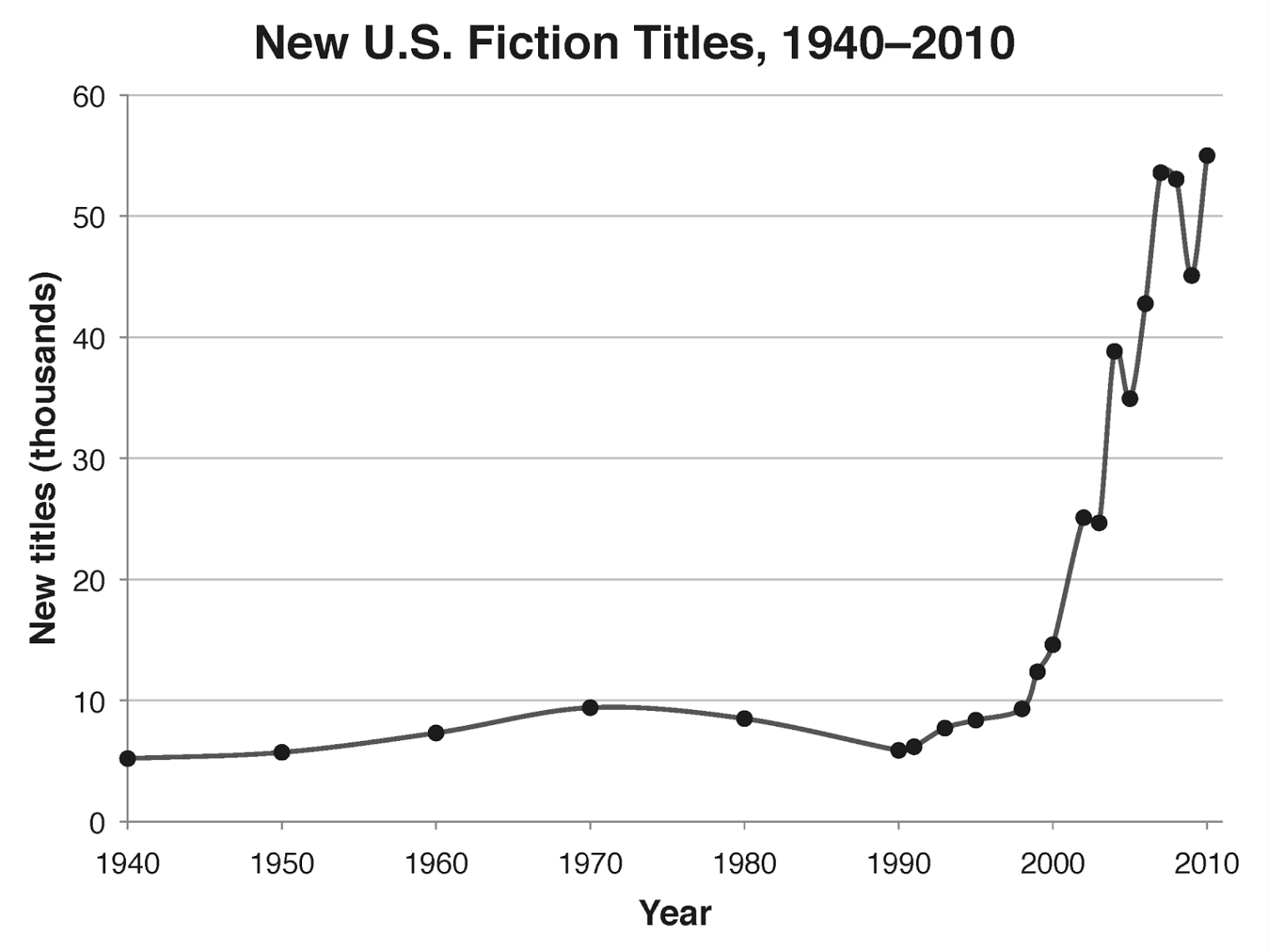
A Short History of Books

Book Production in Europe



Data from https://ourworldindata.org/books/

What should I read next?

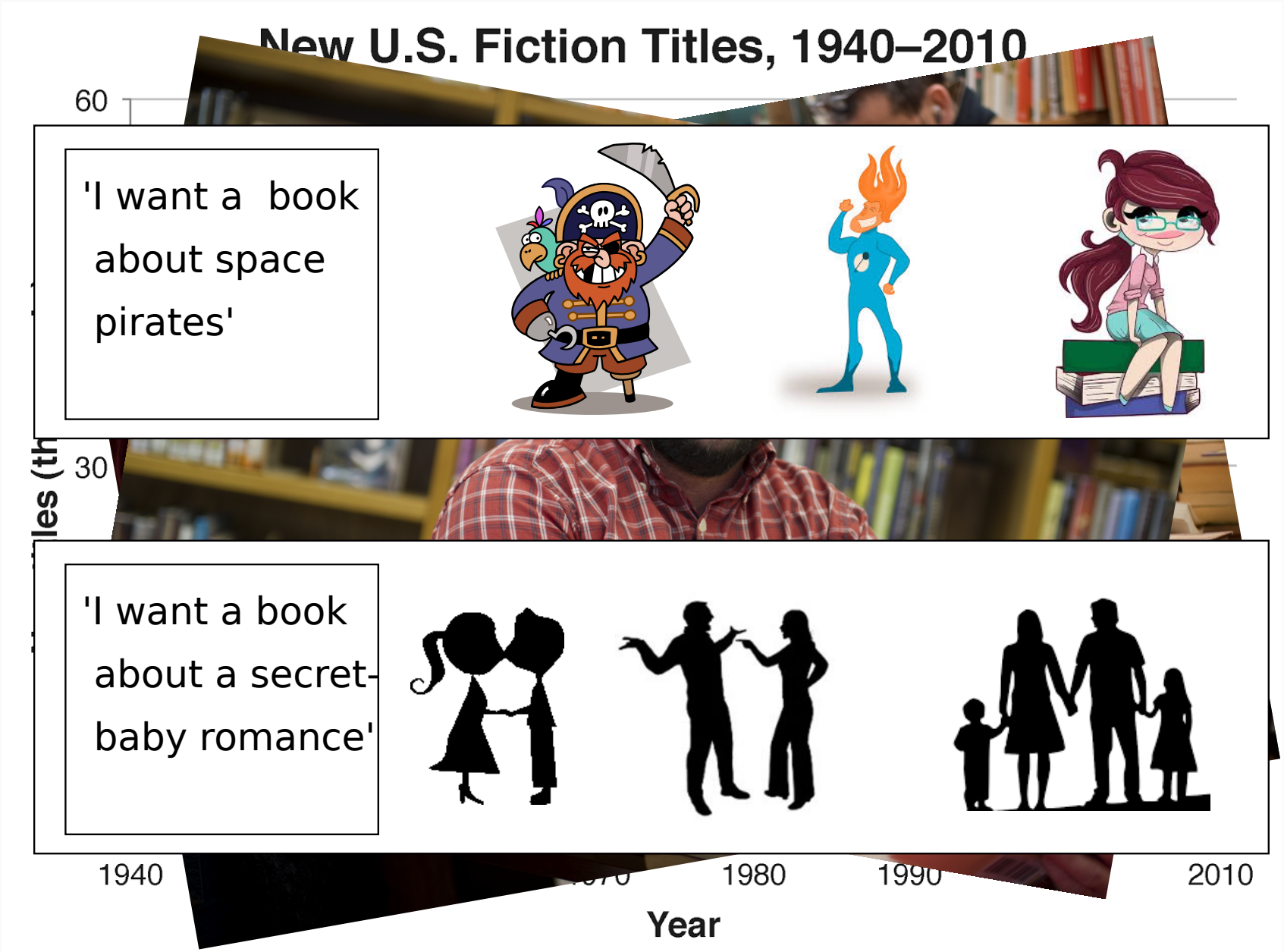


http://www.thewritermoibrahim.com/2016/09/over-50000-fiction-books-published-per.html

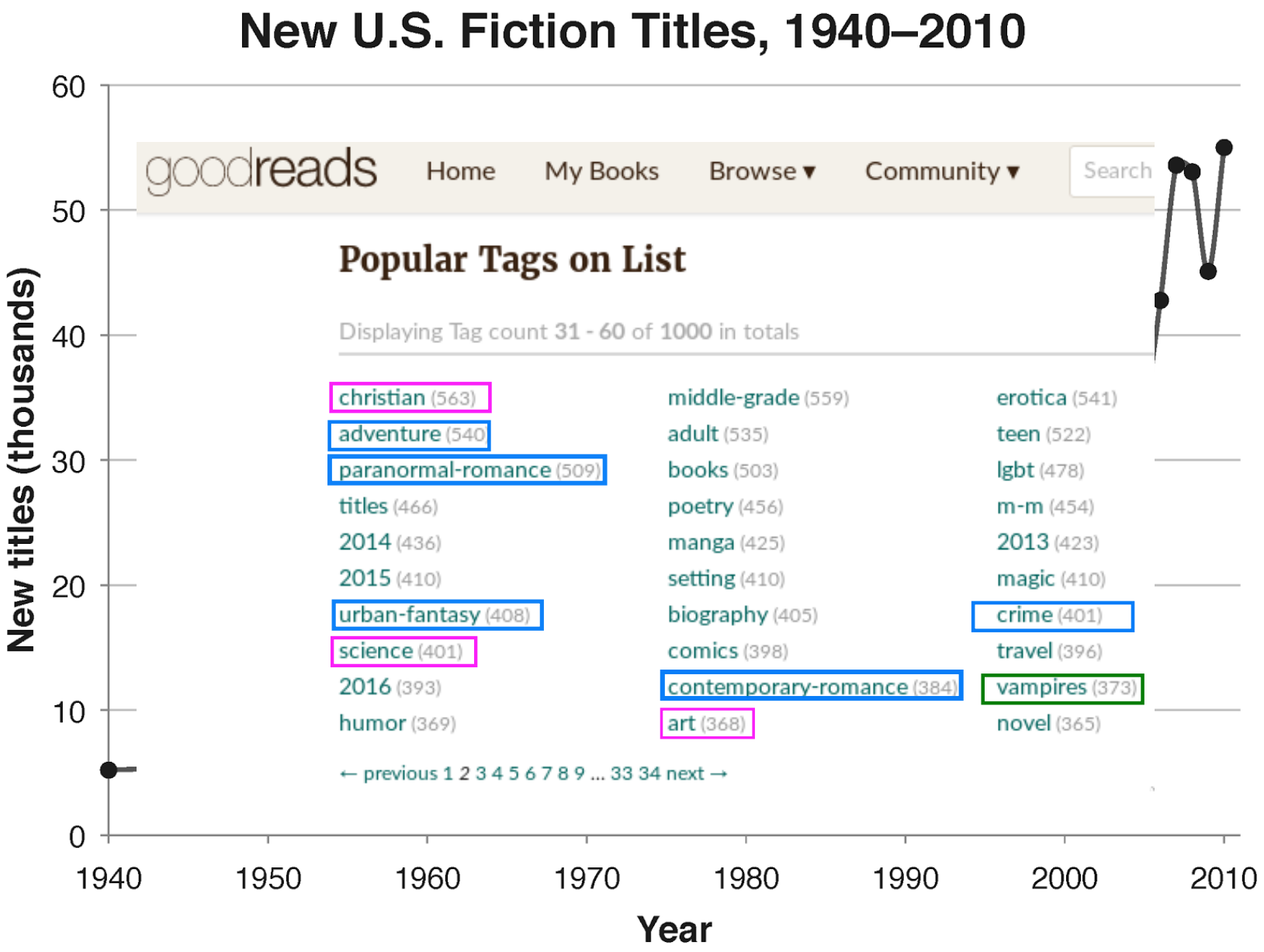
What should I read next?

What should I read next?

What should I read next?



What should I read next?



What should I read next?

Book tags the Amazon Catalog

Romantic Themes
- Secret Baby (691)
- Second Chances (172)
- Wedding (207)
- Beaches (34)
- International (30)
- Medical (16)
- Vacation (42)
- Workplace (52)

Book Format

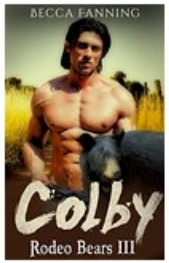
Kindle Edition (702)
Audible Audio Edition
Paperback (22)
Large Print

Romantic Heroes
- Wealthy (463)
- Cowboys (78)
- Criminals & Outlaws (156)
- Doctors (13)
- Royalty & Aristocrats (65)
- Spies (8)

Book Language
- English (724)

Kindle Unlimited
- Kindle Unlimited Eligible

$2.99 to buy
Auto-delivered wirelessly



Colby (BBW Western Bear Shifter Romance) (Rodeo Bears Book 3) Sep 8, 2016

by Becca Fanning

Kindle Edition
$0.00 kindleunlimited
Subscribers read for free. Learn more.

$0.99 to buy
Auto-delivered wirelessly

4.5 ▾ 145
Borrow for free from your Kindle dev

Fury: A Secret Baby Romance Jun 17, 2016

by Kira Ward and Aubrey Sage

Kindle Edition
$0.00 kindleunlimited
Subscribers read for free. Learn more.

$0.99 to buy
Auto-delivered wirelessly

Audible Audio Edition
$0.00
Free with Audible trial

4.5 ▾ 43
Borrow for free from your Kindle dev

Football Baby: A Secret Baby Romance Sep 12, 2016

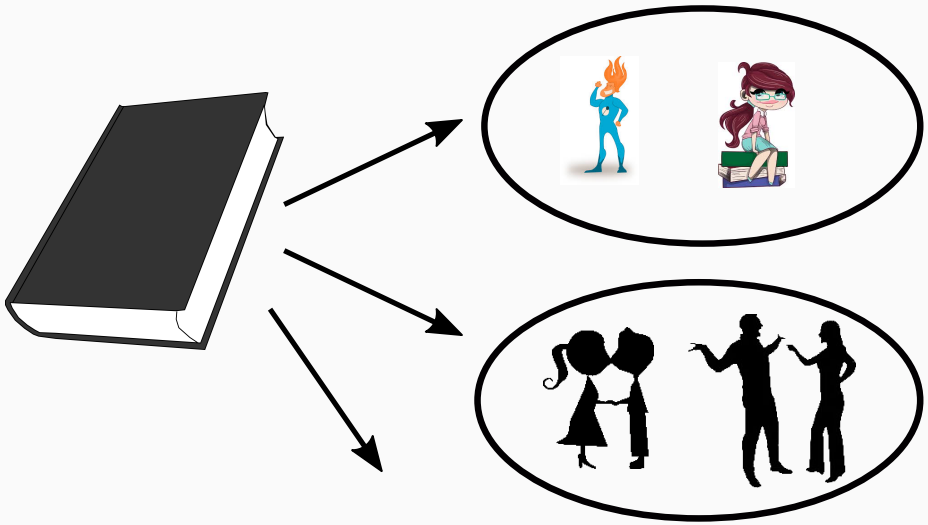
by Roxeanne Rolling

Kindle Edition
$0.00 kindleunlimited
Subscribers read for free. Learn more.

5 ▾ 29
Borrow for free from your Kindle dev

Motivation and Goals

Goal: Flexible, fine-grained book recommendation



Properties of characters

+ development over the plot

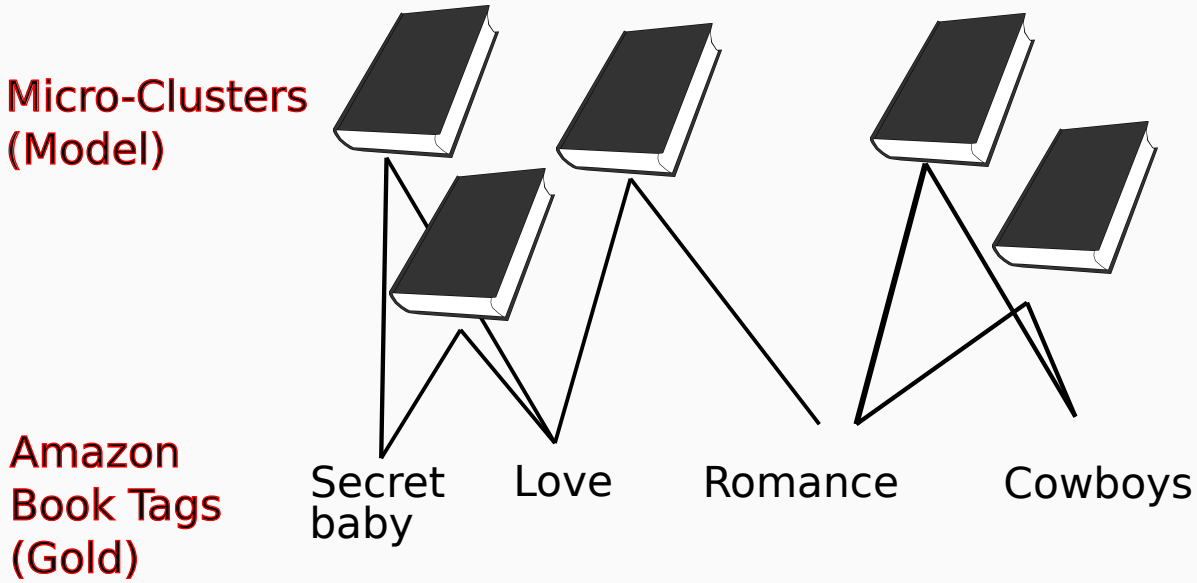
Relations between pairs of characters

+ development over the plot

- induce **multi-view** representations of books
- views encode **relevant** and **distinct** information
- use representations to group books into locally coherent groups (**micro-genres**)

Motivation and Goals

But, how do we know our clusters are meaningful?



- books that share gold tags are similar
- a scalable and principled evaluation framework

Prior Work

Prior work – Unsupervised learning of plot representations

- Typical event chains (McIntyre and Lapata, 2010)
- Typical characters (Bamman et al., 2014; Elsner, 2012)
- Social network analysis (Elson et al., 2010)
- Character relations and their development (Iyyer et al., 2016)

Prior work – Unsupervised learning of plot representations

- Typical event chains (McIntyre and Lapata, 2010)
- Typical characters (Bamman et al., 2014; Elsner, 2012)
- Social network analysis (Elson et al., 2010)
- **Character relations and their development** (Iyyer et al., 2016)
  Relationship Modeling Networks (RMN)

We extend RMNs to induce multi-view representations and introduce a new loss function

Prior work – Evaluation

- manual analysis
- crowd-sourced interpretability judgments
- small manual gold-standards of character types
- distinguish natural vs. scrambled stories

Elsner (2012); Bamman et al. (2013, 2014); Iyyer et al. (2016),...

→ *intrinsic* measures

→ utility for downstream tasks?

→ artificial data

Prior work – Evaluation

- manual analysis
- crowd-sourced interpretability judgments
- small manual gold-standards of character types
- distinguish natural vs. scrambled stories

Elsner (2012); Bamman et al. (2013, 2014); Iyyer et al. (2016),...

→ *intrinsic* measures

→ utility for downstream tasks?

→ artificial data

We propose a general, scalable and principled framework utilizing Amazon catalog book tags.

Prior work – Language-based recommendation

Prevalent recommendation paradigm: social filtering

- bags-of-words from book information data bases
- semantic frames from book plot summaries
- topic modelling for scientific article recommendation

Mooney and Roy (1999); Clercq et al. (2014); Wang and Blei (2011),...

Prior work – Language-based recommendation

Prevalent recommendation paradigm: social filtering

- bags-of-words from book information data bases
- semantic frames from book plot summaries
- topic modelling for scientific article recommendation

Mooney and Roy (1999); Clercq et al. (2014); Wang and Blei (2011),...

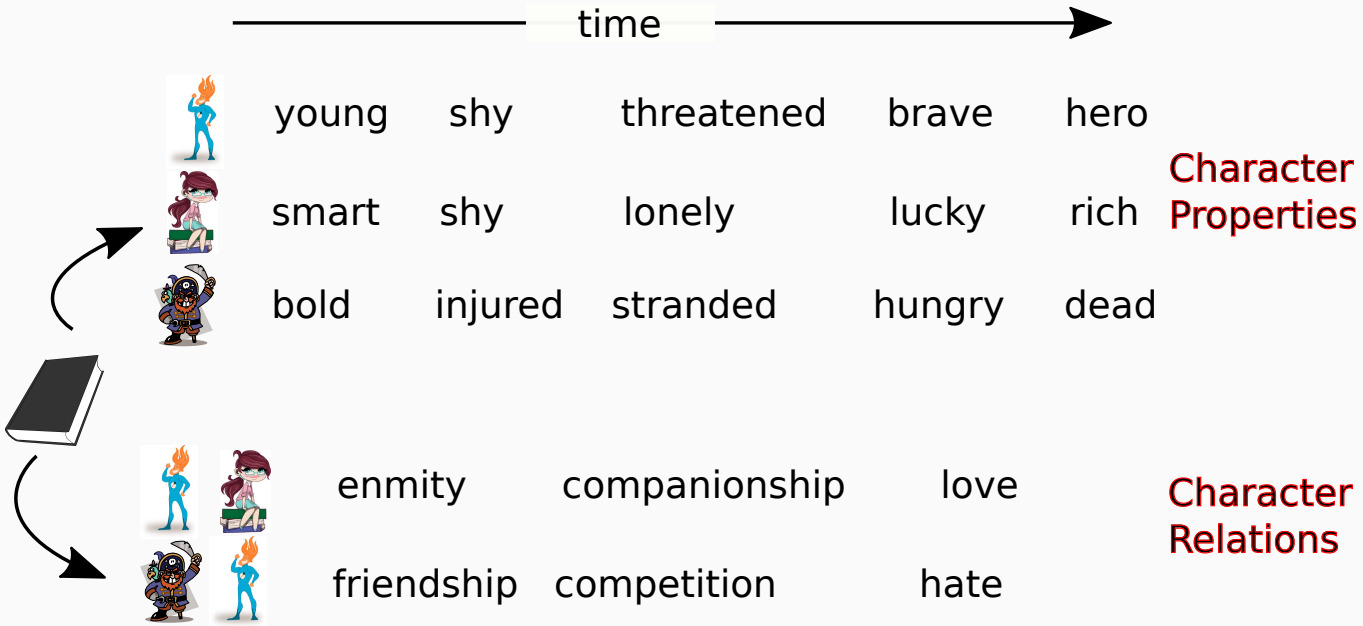
We use richer information from the full book text, and learn structured multi-view representations

MVPlot – A Model of Multi-View Plot Representations

The Idea

view **v1** characteristics of individual characters

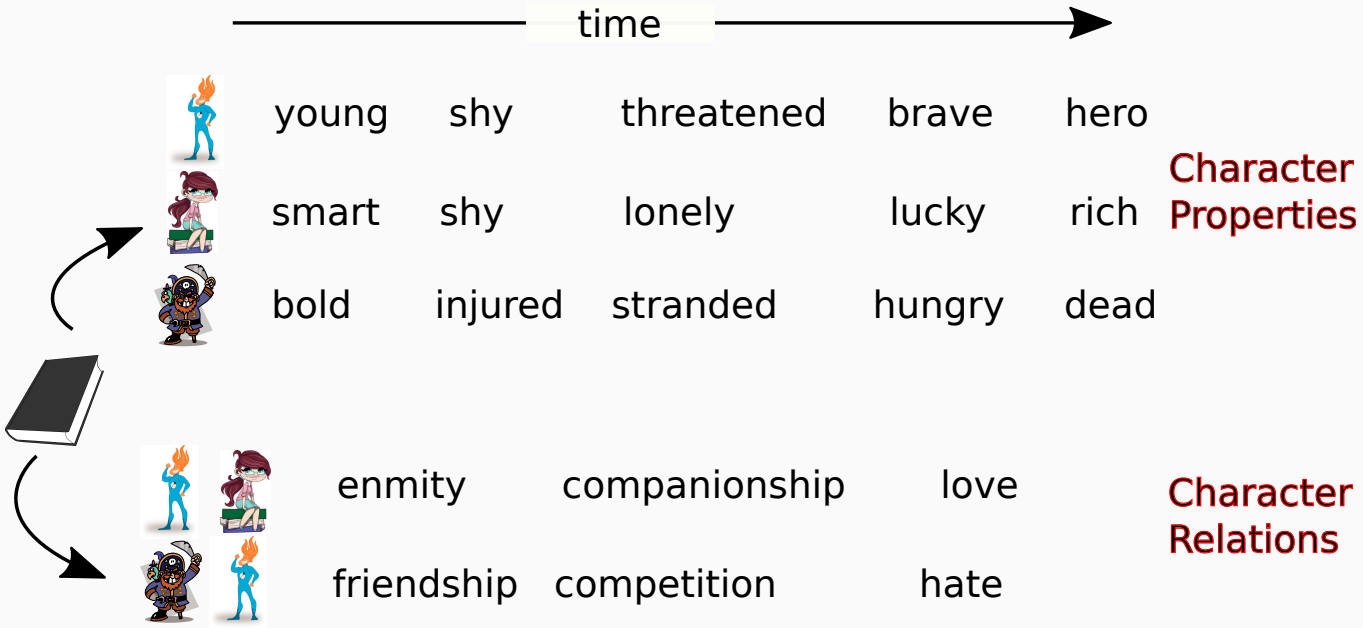
The Idea



view **v1** characteristics of individual characters

view **v2** relations between character pairs

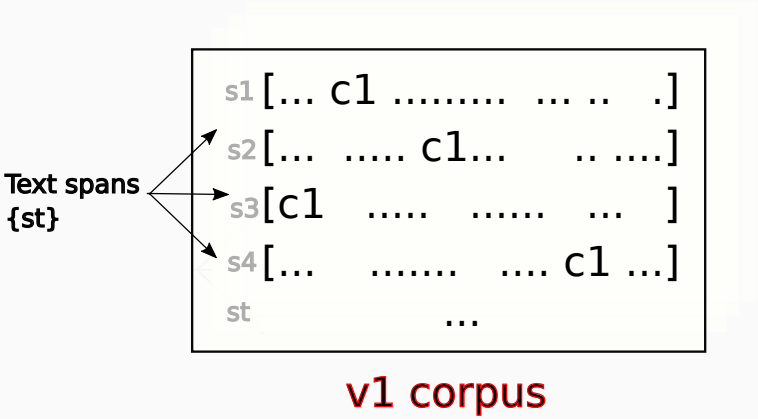
The Idea



Properties and Relations: word clusters ($\sim$ topics)

Objective: triggers views to capture distinct information

MVPlot: Input

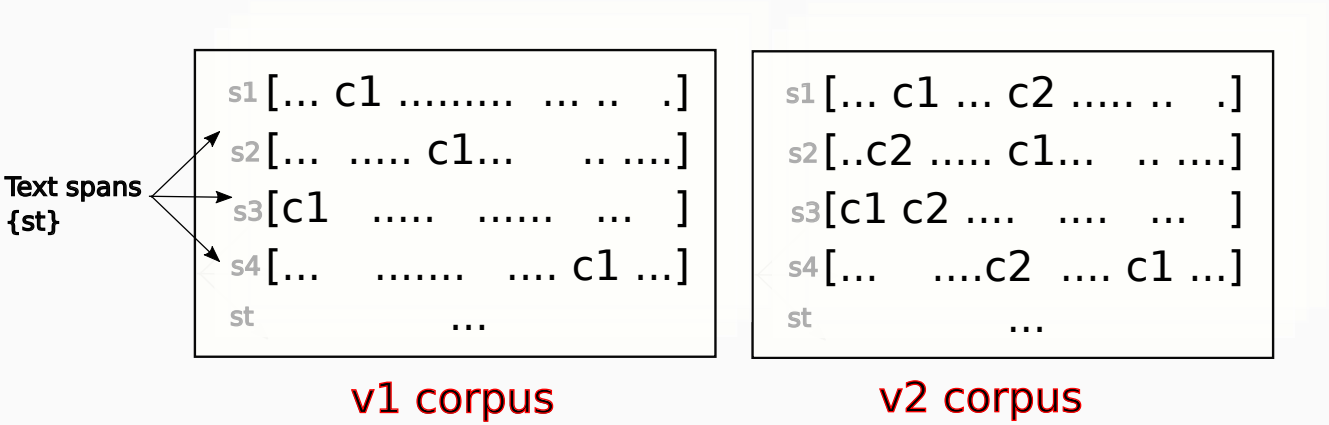


v1 corpus

v1 s_t 20-word text spans mentioning **only** character c1

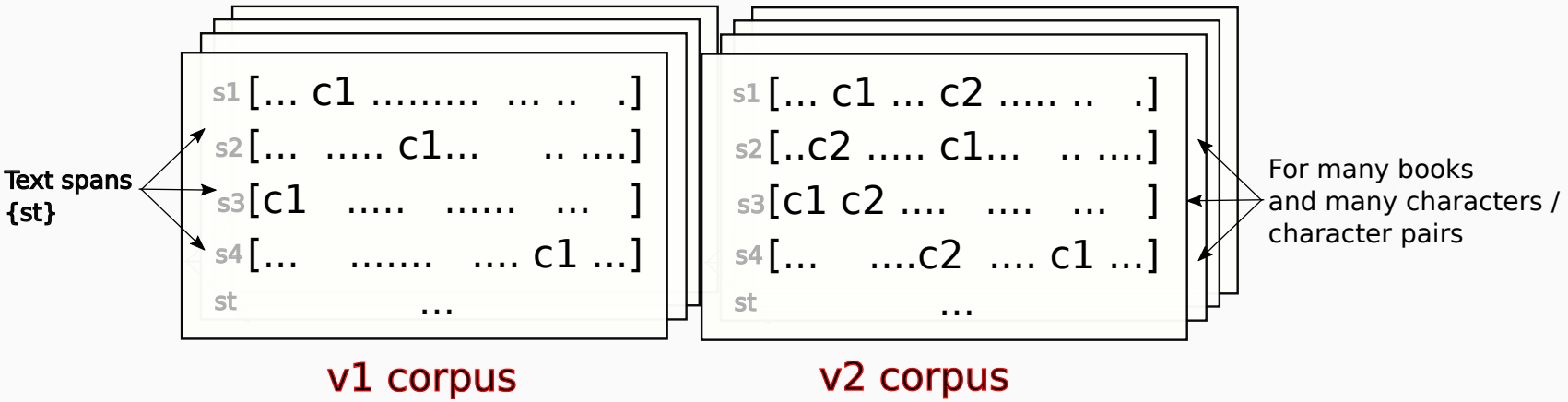
$\{s_t\}$ chronologically ordered sequence of spans $\{s_1, s_2, \ldots, s_T\}$

MVPlot: Input



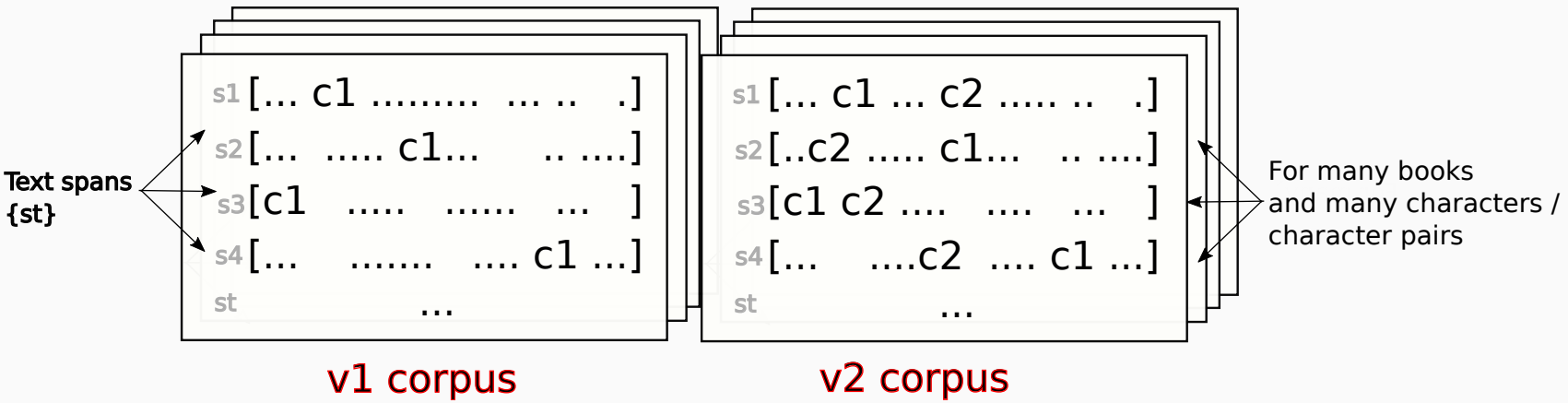
v1	s_t	20-word text spans mentioning **only** character c1
v2	s_t	20-word text spans mentioning **only** characters c1 and c2
	$\{s_t\}$	chronologically ordered sequence of spans $\{s_1, s_2, \ldots, s_T\}$

MVPlot: Input



v1	s_t	20-word text spans mentioning **only** character c1
v2	s_t	20-word text spans mentioning **only** characters c1 and c2
	$\{s_t\}$	chronologically ordered sequence of spans $\{s_1, s_2, \ldots, s_T\}$

MVPlot: Input



- map words in spans to pre-trained GloVe embedding
- average word embeddings to obtain span embedding

MVPlot: Output

1. A set of global **property descriptors v1**, e.g.,

despair	alone desperate sad
intellect	smart clever knowledge
look	pretty small slim
	...

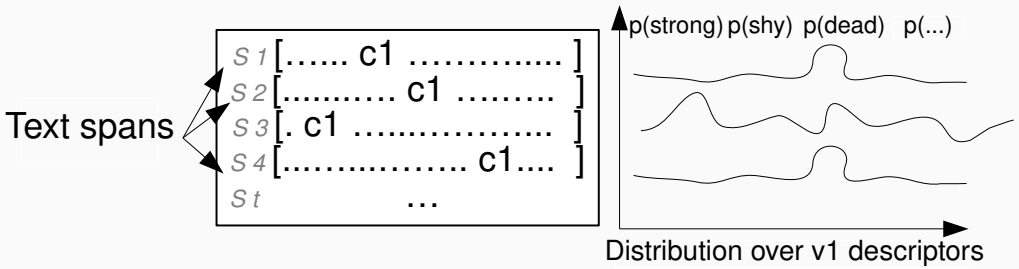
2. A set of global **relation descriptors v2**, e.g.,

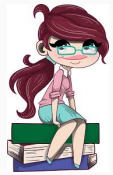
love	dearest loving attracted like
crime	murder chase escape prosecute
hate	hatred dislike fear anger rage
	...

MVPlot: Output

3. For each **character**: a **property trajectory**

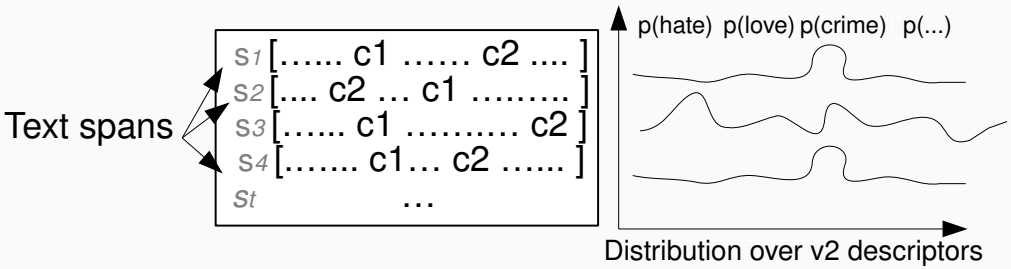
   → a sequence of probability distributions over **v1** descriptors

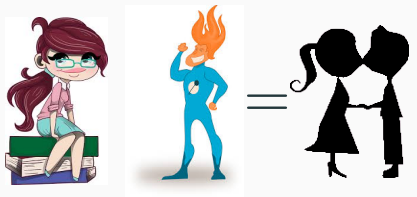




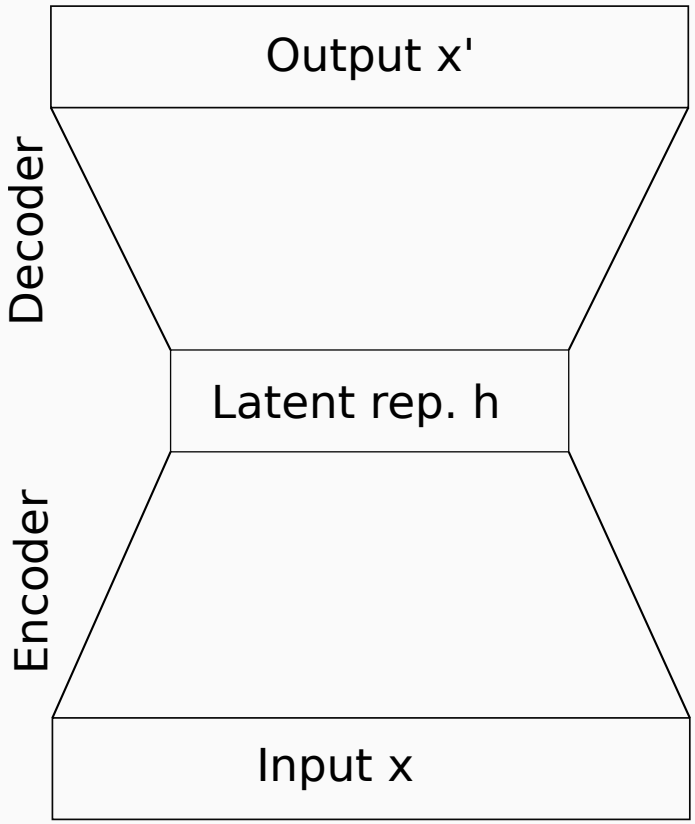
4. For each **character pair**: a **relationship trajectory**

   → a sequence of probability distributions over **v2** descriptors



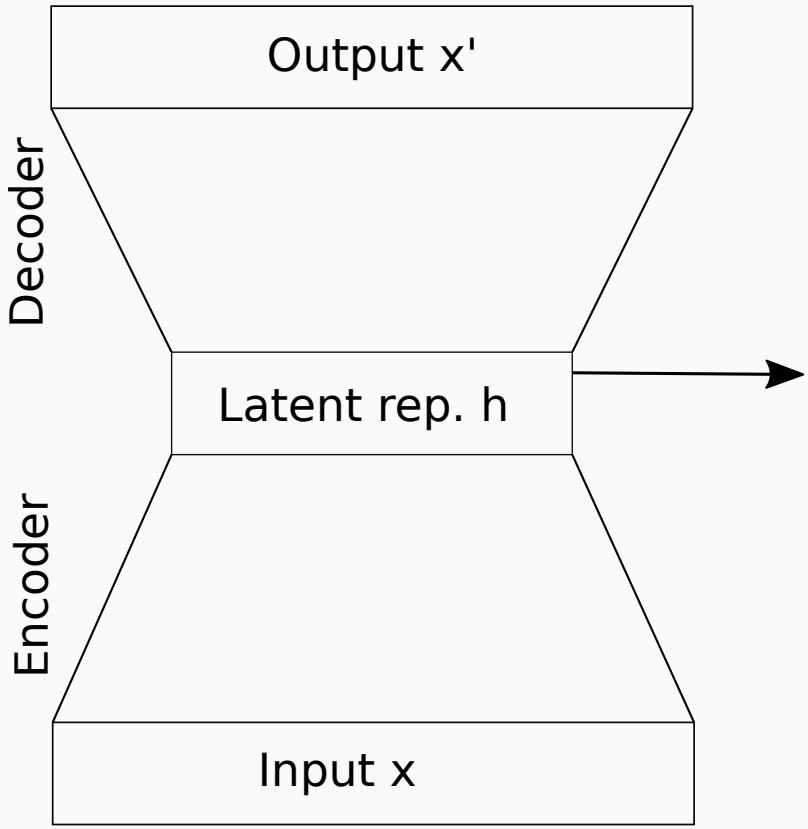


MVPlot: Intuition



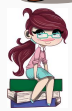
Model architecture based on Iyyer et al (2016)

MVPlot: Intuition



v1 dictionary


[0.7 8 0.002 4]


[0.1 0.008 2 9]


[74 38 6 0.004]

v2 dictionary

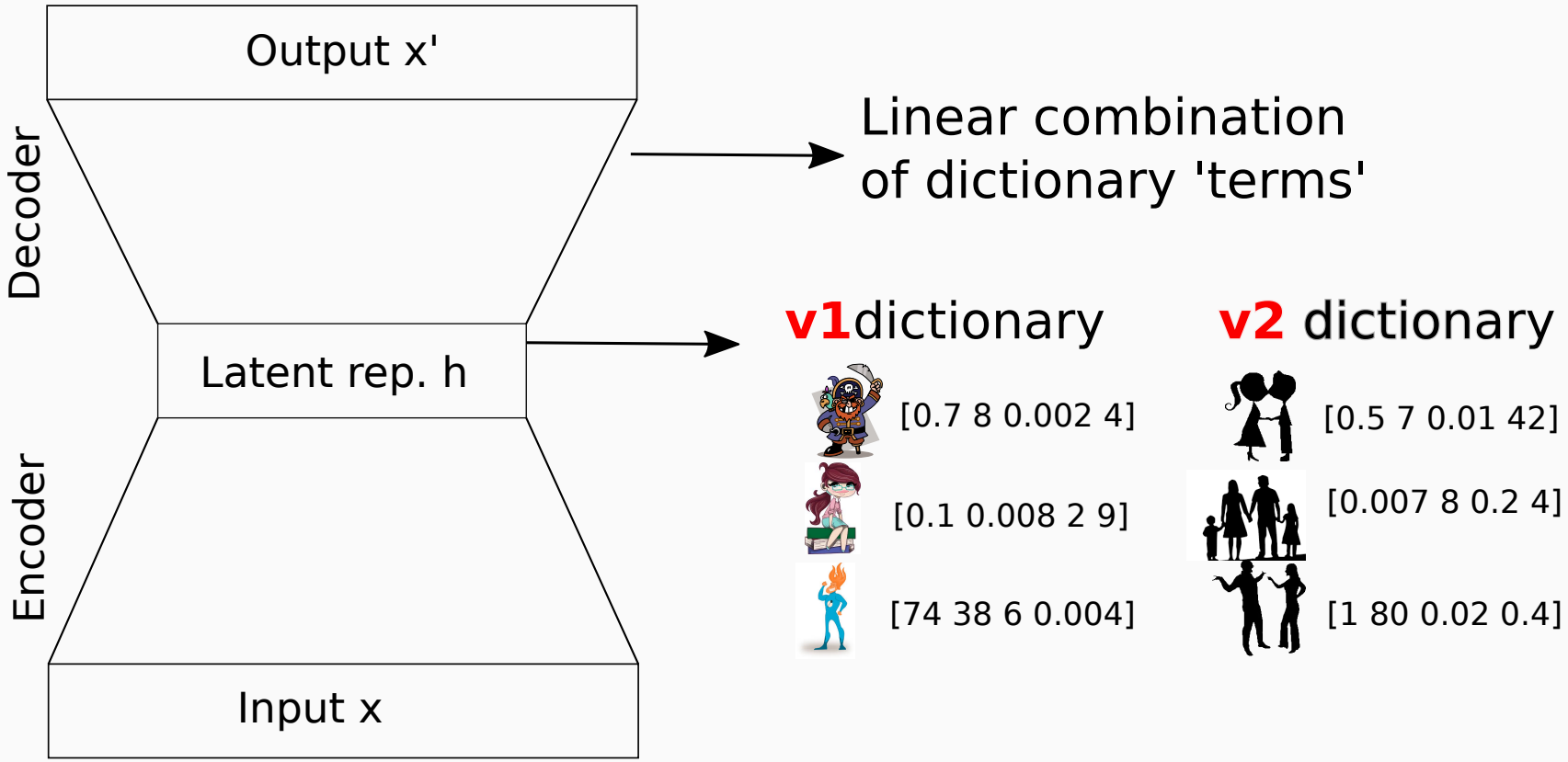

[0.5 7 0.01 42]


[0.007 8 0.2 4]


[1 80 0.02 0.4]

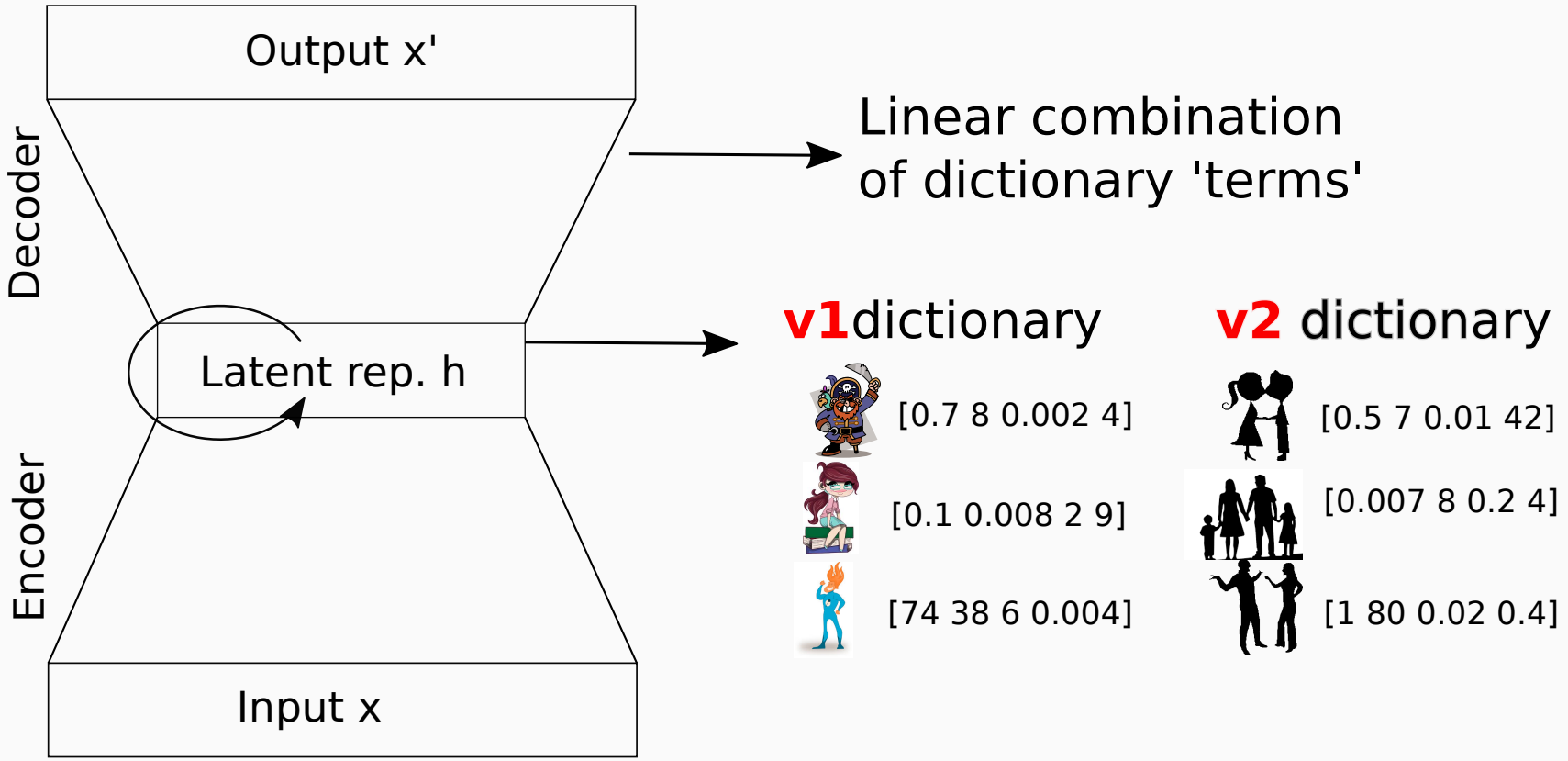
Model architecture based on Iyyer et al (2016)

MVPlot: Intuition



Model architecture based on Iyyer et al (2016)

MVPlot: Intuition

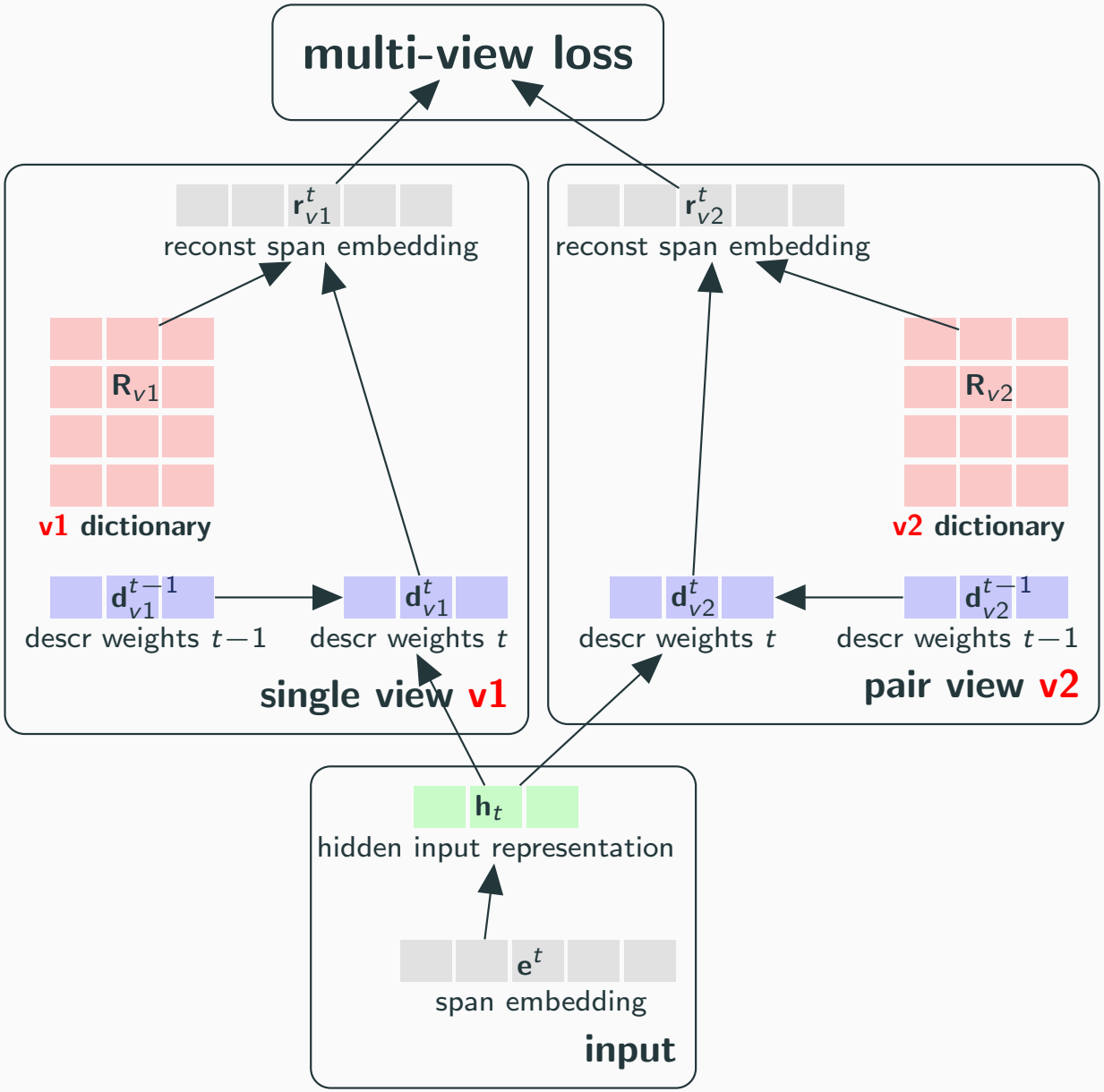


Properties, Relations change smoothly over time

hidden rep. at time t depends on hidden rep. at time $t-1$

Model architecture based on Iyyer et al (2016)

MVPlot: Architecture



The Multi-View Loss

Intuition

1) Within each view,

$\rightarrow$ parameters that efficiently represent and **accurately reconstruct** the input

2) Across views,

$\rightarrow$ parameters that encode relevant and **distinct aspects**

Term 1: View-specific Hinge Loss with Negative Sampling

1) Within each view,

$\rightarrow$ parameters that **accurately reconstruct** the input

Loss Term 1 based on Iyyer et al (2016)

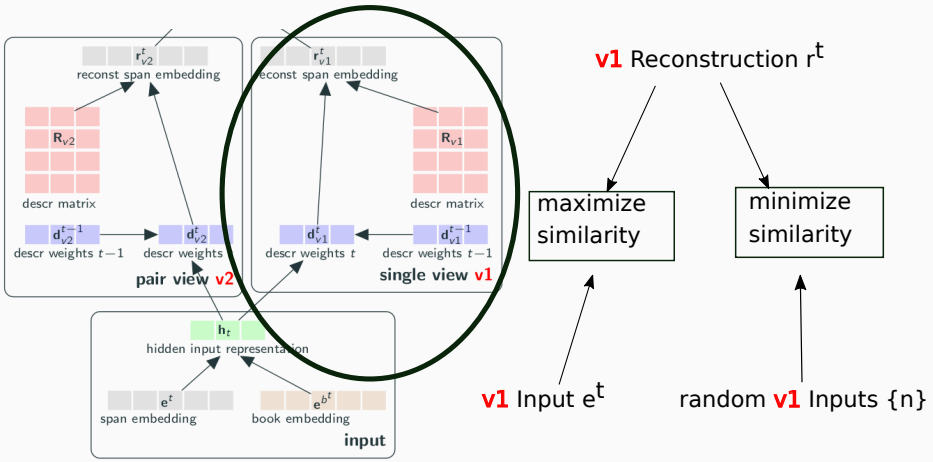
Term 1: View-specific Hinge Loss with Negative Sampling

1) Within each view,

→ parameters that **accurately reconstruct** the input

v1 parameters reconstruct **true v1** input **well**

v1 parameters reconstruct **random v1** inputs **poorly**



$$J(\theta) = \sum_t \beta \sum_{n \in N} \underbrace{max(0, 1 - \mathbf{r}_{v1}^t \mathbf{e}_{v1}^t + \mathbf{r}_{v1}^t \mathbf{e}_{v1}^{n_i})}_{\text{maximize margin between true input and negative samples}} + (1-\beta) \underbrace{max(0, 1 - \mathbf{e}_{v_1}^t \mathbf{r}_{v1}^t + \mathbf{e}_{v_1}^t \mathbf{r}_{v2}^t)}_{\text{maximize margin between true view's and other view's reconstruction}} + \lambda \underbrace{||\mathbf{R}_{v1} \mathbf{R}_{v1}^T - \mathbf{I}||}_{\text{Regularizer}}$$

Loss Term 1 based on Iyyer et al (2016)

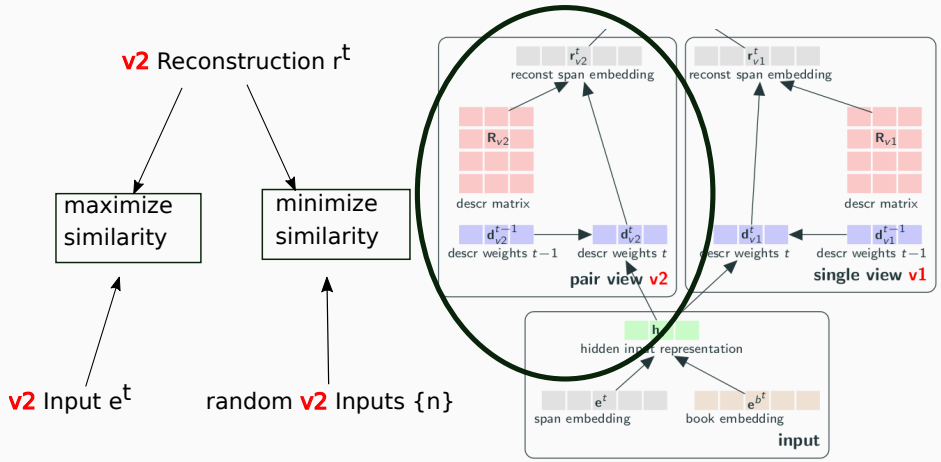
Term 1: View-specific Hinge Loss with Negative Sampling

1) Within each view,

→ parameters that **accurately reconstruct** the input

v2 parameters reconstruct **true v2** input **well**

v2 parameters reconstruct **random v2** inputs **poorly**



$$J(\theta) = \sum_t \beta \sum_{n \in N} \underbrace{max(0, 1 - \mathbf{r}^t_{v1}\mathbf{e}^t_{v1} + \mathbf{r}^t_{v1}\mathbf{e}^{n_i}_{v1})}_{\text{maximize margin between true input and negative samples}} + (1-\beta) \underbrace{max(0, 1 - \mathbf{e}^t_{v_1}\mathbf{r}^t_{v1} + \mathbf{e}^t_{v_1}\mathbf{r}^t_{v2})}_{\text{maximize margin between true view's and other view's reconstruction}} + \lambda \underbrace{||\mathbf{R}_{v1}\mathbf{R}^T_{v1} - \mathbf{I}||}_{\text{Regularizer}}$$

Loss Term 1 based on Iyyer et al (2016)

Term 2: Cross-View Max-Margin Loss

2) Across views,

$\rightarrow$ parameters that encode relevant and **distinct aspects**

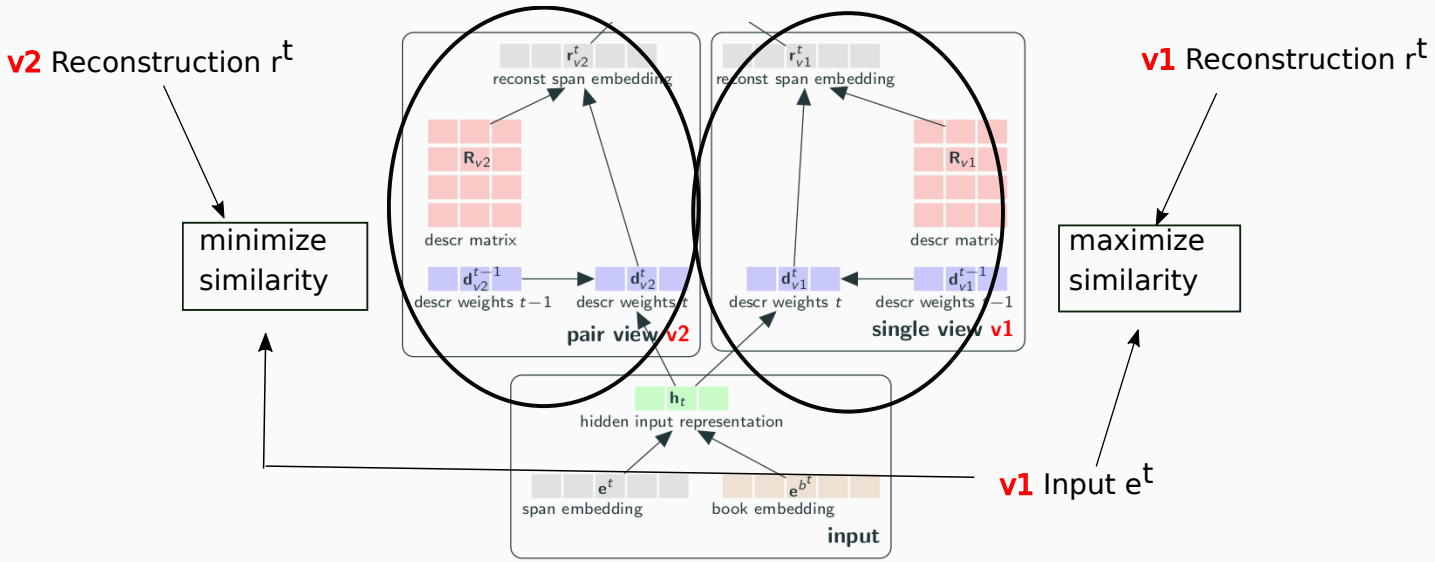
Term 2: Cross-View Max-Margin Loss

2) Across views,

→ parameters that encode relevant and **distinct aspects**

v1-parameters reconstruct **true v1** input **well**

v2-parameters reconstruct **true v1** input **poorly**



$$J(\theta) = \sum_t \beta \sum_{n \in N} \underbrace{max(0, 1 - \mathbf{r}_{v1}^t \mathbf{e}_{v1}^t + \mathbf{r}_{v1}^t \mathbf{e}_{v1}^{n_i})}_{\text{maximize margin between true input and negative samples}} + (1-\beta) \underbrace{max(0, 1 - \mathbf{e}_{v_1}^t \mathbf{r}_{v1}^t + \mathbf{e}_{v_1}^t \mathbf{r}_{v2}^t)}_{\text{maximize margin between true view's and other view's reconstruction}} + \lambda \underbrace{||\mathbf{R}_{v1} \mathbf{R}_{v1}^T - \mathbf{I}||}_{\text{Regularizer}}$$

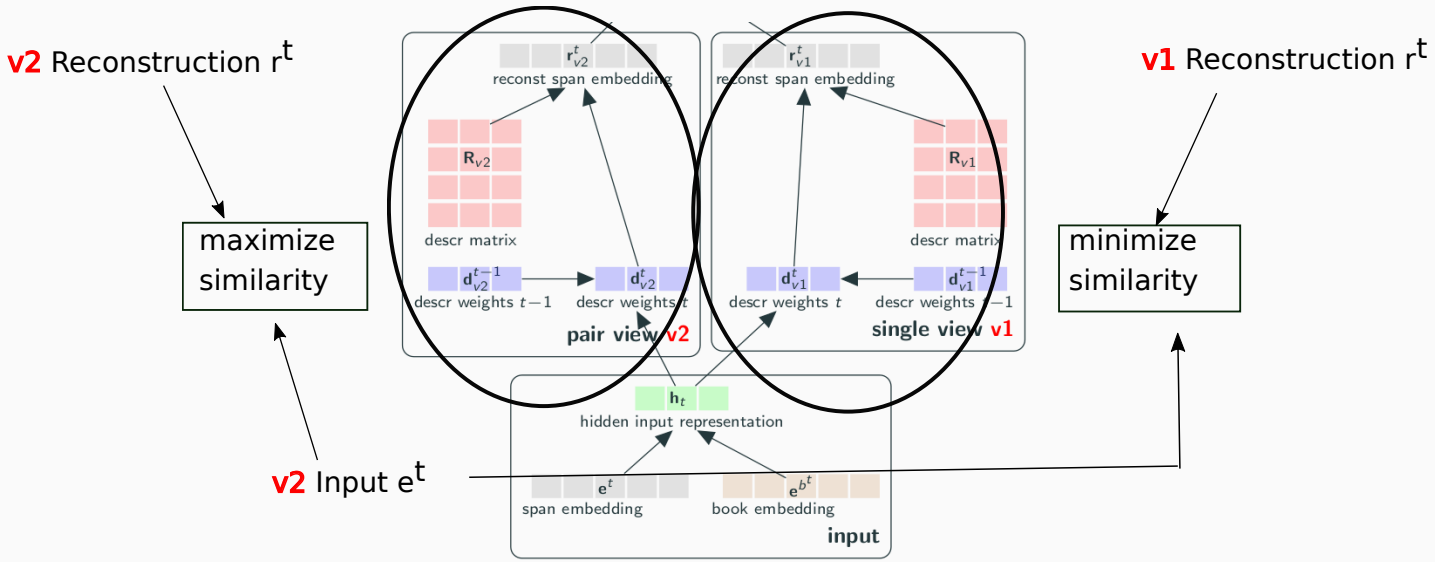
Term 2: Cross-View Max-Margin Loss

2) Across views,

→ parameters that encode relevant and **distinct aspects**

v2-parameters reconstruct **true v2** input **well**

v1-parameters reconstruct **true v2** input **poorly**



$$J(\theta) = \sum_t \beta \sum_{n \in N} \underbrace{max(0, 1 - \mathbf{r}^t_{v1}\mathbf{e}^t_{v1} + \mathbf{r}^t_{v1}\mathbf{e}^{n_i}_{v1})}_{\text{maximize margin between true input and negative samples}} + (1-\beta) \underbrace{max(0, 1 - \mathbf{e}^t_{v_1}\mathbf{r}^t_{v1} + \mathbf{e}^t_{v_1}\mathbf{r}^t_{v2})}_{\text{maximize margin between true view's and other view's reconstruction}} + \lambda \underbrace{||\mathbf{R}_{v1}\mathbf{R}^T_{v1} - \mathbf{I}||}_{\text{Regularizer}}$$

Evaluation

Evaluation Overview

Data

	Amazon corpus
# books	10,000
# *v*1 sequences	91,511
# *v*2 sequences	70,156

Evaluation Overview

Data

	Amazon corpus	Gutenberg corpus
# books	10,000	3,500
# $v1$ sequences	91,511	45,182
# $v2$ sequences	70,156	60,493

Evaluation Overview

Data

	Amazon corpus	Gutenberg corpus
# books	10,000	3,500
# $v1$ sequences	91,511	45,182
# $v2$ sequences	70,156	60,493

- Preprocessing with BooksNLP pipeline (Bamman et al., 2014)
- keep sequences s iff $5 \leq |s| \leq 250$

Evaluation of quality of induced descriptors

Experiment 1 – Evaluation of induced Descriptors

Data: Gutenberg corpus

Descriptors represented as 10 closest words in GloVe space

Property Descriptors (v1)
laugh scream laughing yell joke cringe disgrace embarrassment hate cursing
snug fleece warm comfortable wet blanket flannel cozy comfort roomy
excellency mademoiselle monsieur majesty duchess empress madame countess madam dowager
love loving lovely dear sweetest dearest thank darling congratulation hello
associate assistant senior chairman executive leadership vice director liaison vice-president

Experiment 1 – Evaluation of induced Descriptors

Data: Gutenberg corpus

Descriptors represented as 10 closest words in GloVe space

Property Descriptors (v1)
laugh scream laughing yell joke cringe disgrace embarrassment hate cursing
snug fleece warm comfortable wet blanket flannel cozy comfort roomy
excellency mademoiselle monsieur majesty duchess empress madame countess madam dowager
Relation Descriptors (v2)
love loving lovely dear sweetest dearest thank darling congratulation hello
associate assistant senior chairman executive leadership vice director liaison vice-president

Experiment 1 – Evaluation of induced Descriptors

Do *v2* relation descriptors indeed capture character relations?

- MVPlot *v2* descriptors vs RMN descriptors (Iyyer et al., 2016)
- crowd-sourced plausibility judgments (Chaturvedi et al., 2017)

Experiment 1 – Evaluation of induced Descriptors

Do *v2* relation descriptors indeed capture character relations?

- MVPlot *v2* descriptors vs RMN descriptors (Iyyer et al., 2016)
- crowd-sourced plausibility judgments (Chaturvedi et al., 2017)

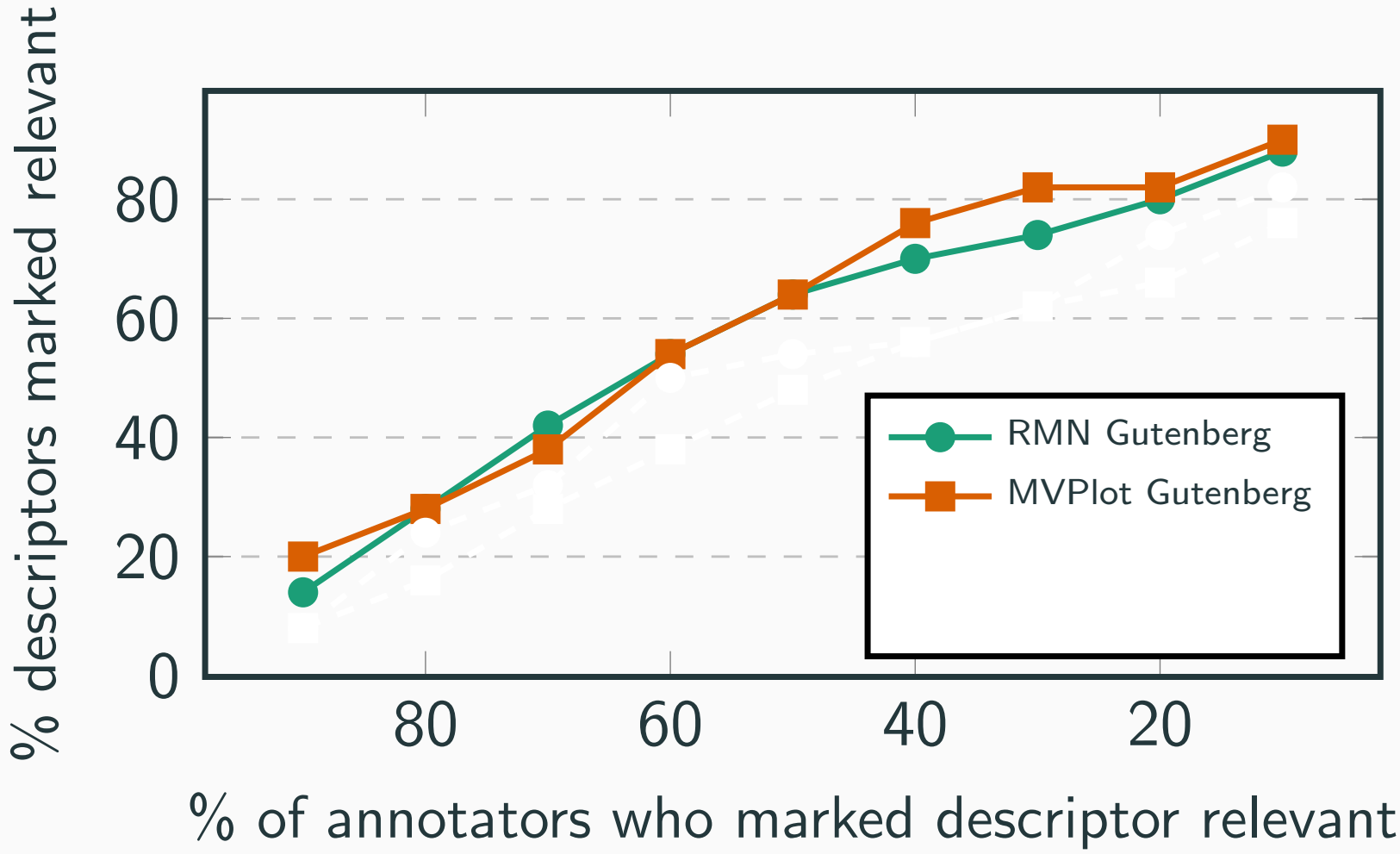
love loving lovely dear sweetest dearest thank darling congratulation hello

Do the words describe a relation, event or interaction between people?

○ yes ○ no

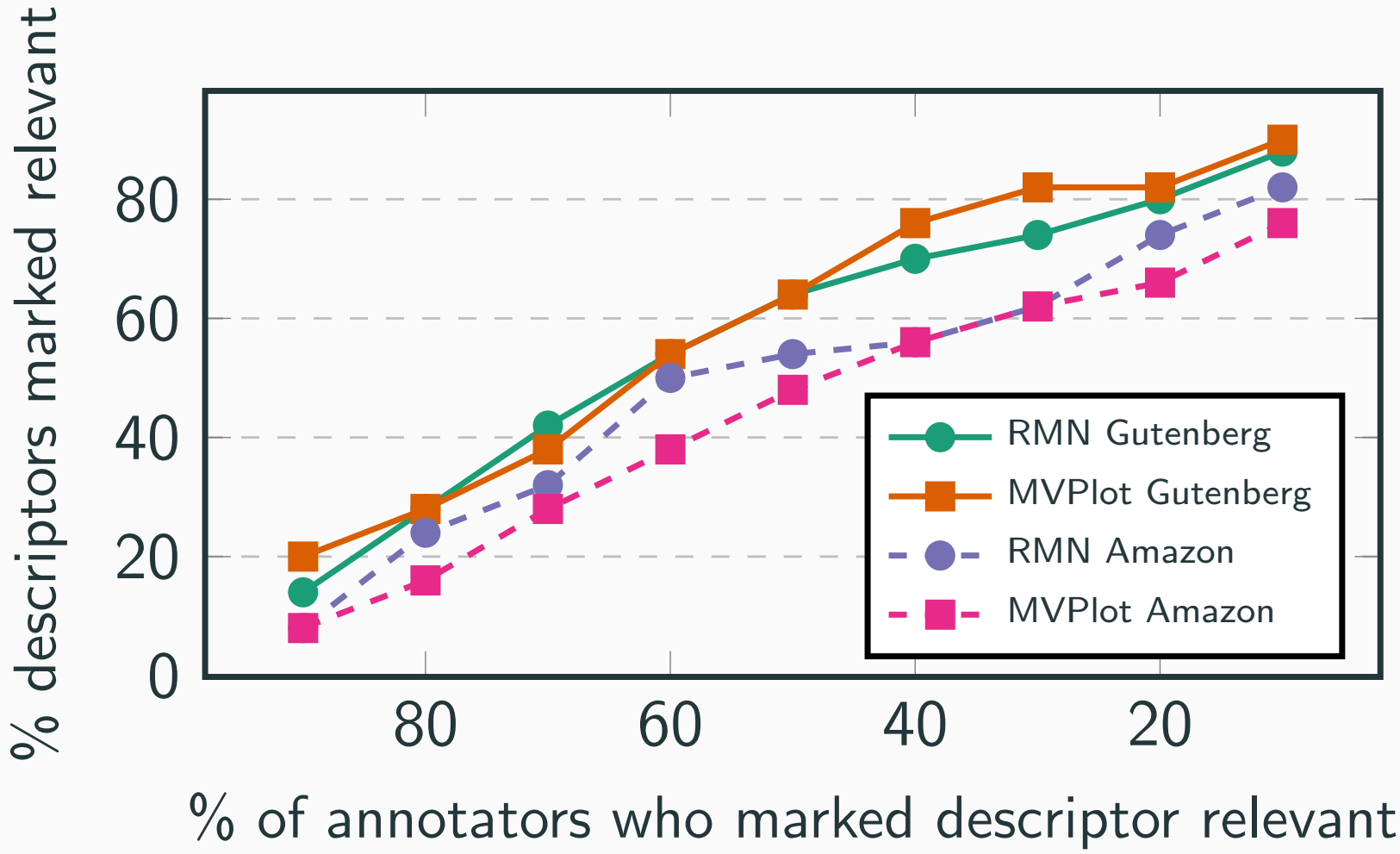
Experiment 1 – Evaluation of induced Descriptors

Gutenberg corpus



Experiment 1 – Evaluation of induced Descriptors

Gutenberg corpus and Amazon corpus



Large-scale evaluation of local neighborhoods in model space

Experiment 2: Large-scale evaluation of book neighborhoods

Are novels with similar induced representations indeed similar?

- Quantify the quality of induced micro-clusters
- Utilize expert-tags from the Amazon catalog: do similar books share tags?

Experiment 2: Large-scale evaluation of book neighborhoods

Are novels with similar induced representations indeed similar?

- Quantify the quality of induced micro-clusters
- Utilize expert-tags from the Amazon catalog: do similar books share tags?

Genre	Refinement: Character Type
Thriller	British Detectives; FBI Agents; Female Protagonists
Romance	Cowboys ; Criminals & Outlaws ; Doctors ; Spies ; Wealthy
SciFi	AIs ; Aliens ; Clones ; Mutants ; Psychics ; Robots & Androids

Experiment 2: Large-scale evaluation of book neighborhoods

Are novels with similar induced representations indeed similar?

- Quantify the quality of induced micro-clusters
- Utilize expert-tags from the Amazon catalog: do similar books share tags?

Genre	**Refinement: Mood and Theme**
Thriller	Action-packed ; dark ; fun ; racy & risque
Romance	Second-chances ; secret-baby ; international ; workplace
SciFi	horror ; humor ; mystery ; non-romantic

Experiment 2: Large-scale evaluation of book neighborhoods

We select 50 tags

- from popular genres (thriller, romance, scifi)
- frequent refinements (character, category, mood/theme)
- Not tuned at all.

Experiment 2: Large-scale evaluation of book neighborhoods

Each book b is represented through two views

v1 For each character
property representation c (averaged trajectory)

v2 For each character pair
relation representation p (averaged trajectory)

Experiment 2: Large-scale evaluation of book neighborhoods

Each book b is represented through two views

v1 For each character
property representation c (averaged trajectory)

v2 For each character pair
relation representation p (averaged trajectory)

We can cluster books based on different views and ask

- Which books have similar character types (v1)?
- Which books have similar character pair relations(v2)?
- Which books are similar overall (v1,v2)?

Experiment 2: Large-scale evaluation of book neighborhoods

Each book b is represented through two views

v1 For each character
property representation c (averaged trajectory)

v2 For each character pair
relation representation p (averaged trajectory)

View-specific book similarities, given books $b1, b2$

$$sim(\text{v1}) = \sum_{c \in b1} \sum_{c' \in b2} min(dist\_euclid(c, c'))$$

$$sim(\text{v2}) = \sum_{p \in b1} \sum_{p' \in b2} min(dist\_euclid(p, p'))$$

$$sim(\text{v1,v2}) = \frac{1}{2}(sim(v1) + sim(v2))$$

Experiment 2: Large-scale evaluation of book neighborhoods

Setup

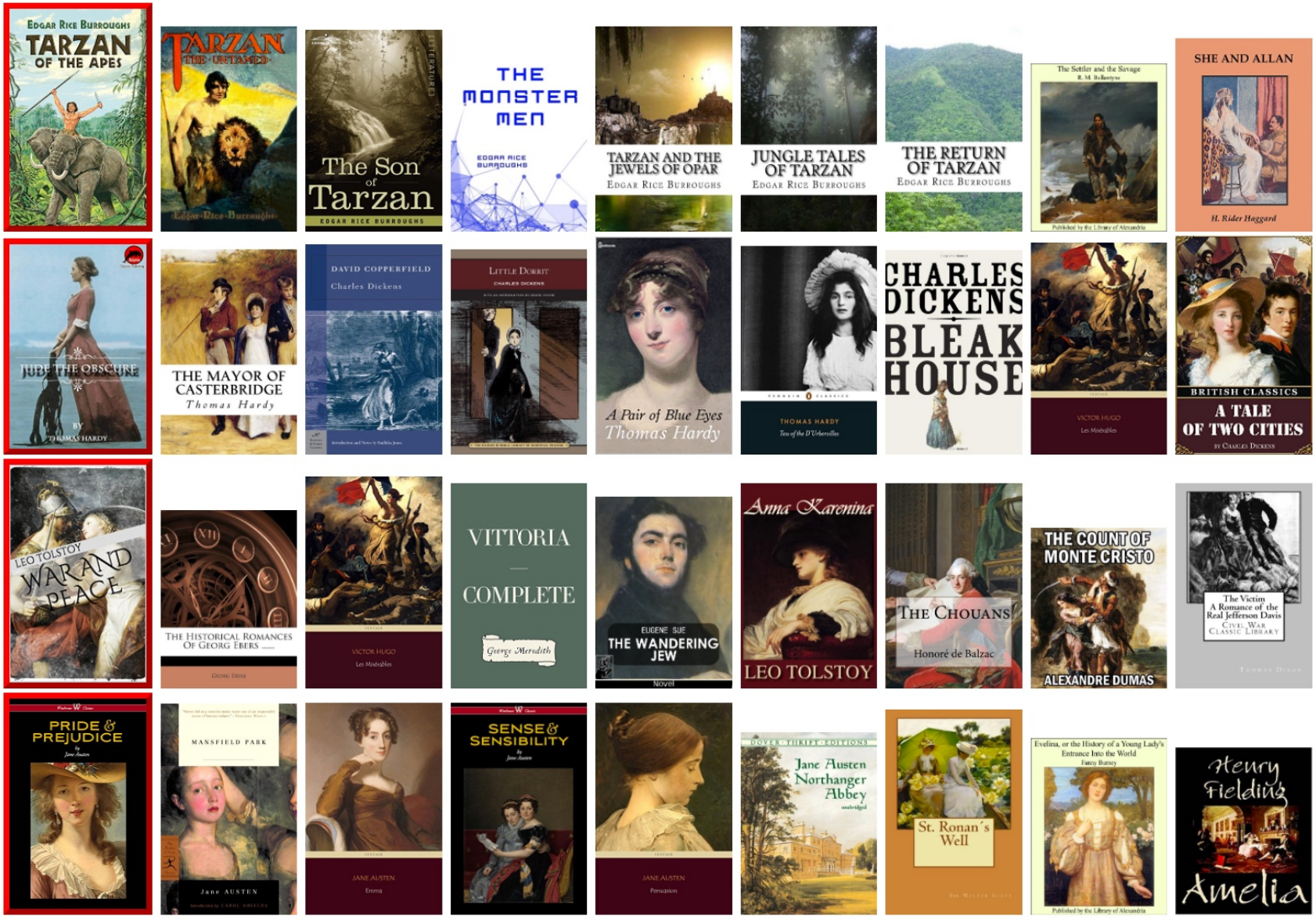
- Reference set: 500 most popular books in Amazon dataset (by # reviews)
- For each, compute the 10 nearest neighbors
- Do nearest neighbors share a gold tag with the reference? (precision@10)

Results

Model	View	P@10	MAP
MVPlot	v_1	0.529 †	0.401 †
	v_2	0.496 ‡	0.367 ‡
	v_1, v_2	**0.546**	**0.421**
RMN	v_2	0.479 ‡	0.347 ‡
cosine	v_1	0.516 ‡	0.392 †
	v_2	0.468 ‡	0.339 ‡
	v_1, v_2	0.512 ‡	0.390 ‡

significance with $p < 0.05$ (†) or $p < 0.01$ (‡) (paired t-test).

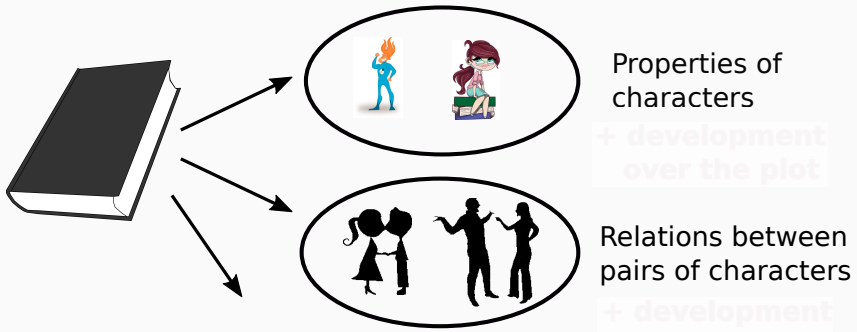
K-Nearest Neighbors Visualization



Conclusions

Conclusions

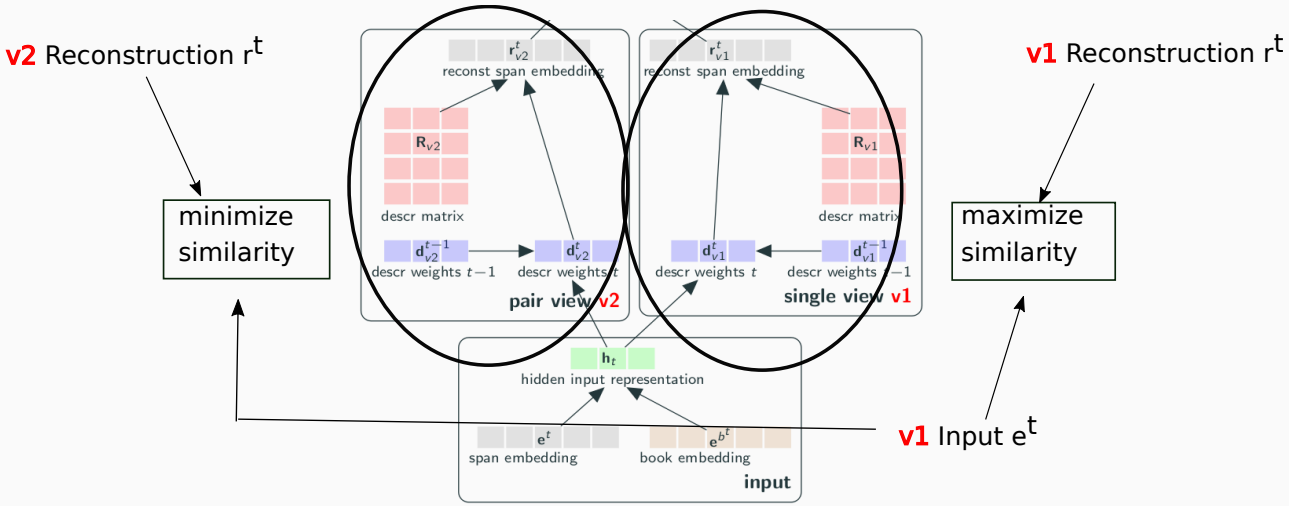
MVPlot – Deep multi-view book representations



- learns sets of **property** and **relation** descriptors
- efficient on both large-scale (Amazon) and smaller (Gutenberg) data
- multi-view model performs best
  $\rightarrow$ views capture *distinct* information

Conclusions

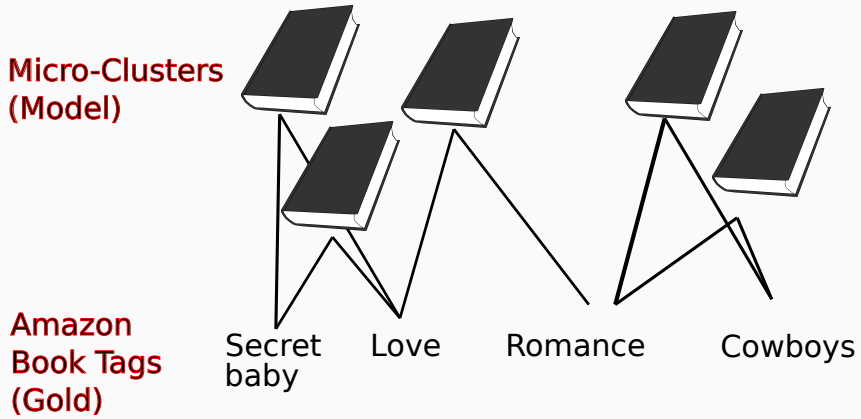
Multi-view Loss function



- novel multi-view loss function
- triggers views to capture **relevant** and **distinct** information
- generalizes to arbitrary # views and different input data

Conclusions

Evaluation of plot representations



- task-based extrinsic evaluation
- scalable and principled evaluation framework

Thank you!

www.frermann.de

References

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361. Association for Computational Linguistics, Sofia, Bulgaria. URL `http://www.aclweb.org/anthology/P13-1035`.

References ii

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379. Association for Computational Linguistics, Baltimore, Maryland. URL `http://www.aclweb.org/anthology/P/P14/P14-1035`.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

References iii

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610. Association for Computational Linguistics, Suntec, Singapore. URL `http://www.aclweb.org/anthology/P/P09/P09-1068`.

Jonathan Chang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Connections between the lines: augmenting social networks with text. In John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Zaki, editors, *KDD*, pages 169–178. ACM.

References iv

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daumé III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Association for the Advancement of Artificial Intelligence*.

Orphée De Clercq, Michael Schuhmacher, Simone Paolo Ponzetto, and Véronique Hoste. 2014. Exploiting framenet for content-based book recommendation. In *Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems co-located with the 8th ACM Conference on Recommender Systems, CBRecSys@RecSys 2014, Foster City, Silicon Valley, California, USA, October 6, 2014.*, pages 14–21.

M. Elad and M. Aharon. 2006. Image denoising via sparse and redundant representations over learned dictionaries. *Trans. Img. Proc.*, 15(12):3736–3745.

References v

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644. Association for Computational Linguistics, Avignon, France. URL `http://www.aclweb.org/anthology/E12-1065`.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147. Association for Computational Linguistics, Uppsala, Sweden. URL `http://www.aclweb.org/anthology/P10-1015`.

References vi

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1180–1189. JMLR Workshop and Conference Proceedings. URL `http://jmlr.org/proceedings/papers/v37/ganin15.pdf`.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In Marina Meila and Xiaotong Shen, editors, *AISTATS*, volume 2 of *JMLR Proceedings*, pages 163–170. JMLR.org.

References vii

A. Hyvärinen and E. Oja. 2000. Independent component analysis: Algorithms and applications. *Neural Netw.*, (4-5):411–430.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *North American Association for Computational Linguistics*.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

References viii

Philip Massey, Patrick Xia, David Bamman, and Noah A. Smith. 2015. Annotating character relationships in literary texts. *CoRR*, abs/1512.00728. URL `http://arxiv.org/abs/1512.00728`.

Neil McIntyre and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572. Association for Computational Linguistics, Uppsala, Sweden. URL `http://www.aclweb.org/anthology/P10-1158`.

Raymond J. Mooney and Loriene Roy. 1999. Content-based book recommending using learning for text categorization. In *Proceedings of the SIGIR-99 Workshop on Recommender Systems: Algorithms and Evaluation*. Berkeley, CA.

Bruno A. Olshausen and David J. Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. URL `http://www.aclweb.org/anthology/D14-1162`.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL)*, pages 410–420.

References x

Hardik Vala, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths. 2016. Annotating Characters in Literary Corpora: A Scheme, the CHARLES Tool, and an Annotated Novel. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456. ACM, New York, NY, USA.