Topic Models for Morphologically Rich Languages

Michael Elhadad, Meni Adler, Yoav Goldberg, Rafi Cohen

23 Jan 2011, Haifa

Machine Translation and Morphologically-rich Languages

Topic Models

- Unsupervised discovery of topics in text collection
- Useful to browse/explore large corpora by theme
  - Topic evolution over time
  - Author-topic models
- Difficult to evaluate / Task-based evaluations help
  - WSD
  - Summarization
  - IR
  - Sentiment analysis
- Multilingual LDA could help as feature for MT

Topic Models and Rich Morphology

- Topic Models from text in Hebrew
  - Rich morphology
  - High number of distinct word forms
  - High ambiguity
- Halakhic Domain (Jewish Religious Law)
  - Mixture of languages (Hebrew / Aramaic)
  - Various Historical / Geographical / Subdomains
  - Existing metadata / Can we exploit it?
- Medical Domain
  - Patient letters / eHealth QA site
  - High level of mixture English/Hebrew (transliterations)
  - Existing metadata (UMLS) / Can we exploit it?
- Work in progress

Outline

- Topic Analysis with LDA
- Domain: Halakhic Sources / Medical dataset
- Combining LDA and Morphological Analysis
- Combining Semantic Priors and LDA
- Multilingual Topic Models
- Evaluating Topic Models

Objectives

- Input:
  - Domain specific text corpus in Hebrew
  - Metadata on documents (tags, alignment to English tags)
- Output:
  - Topic model:
    - Discover “topics” discussed in the corpus
    - Recognize topics in unseen text
  - Index text collection by topic
- Task:
  - Something where topics help:
    - WSD, IR, Text categorization, clustering
    - Some part of MT?

Term Ambiguity and What is a Topic?

- "שור" (ox/bull) refers to many complex halakhic topics:
  - Damages (שור נוגח – goring ox)
  - Kosher meat (שחיטה – slaughter)
  - Sacrifices (קרבנות)
  - Shabbat (שבת – domestic animals must rest)
  - Calendar (מזל שור – Zodiac sign Taurus)

  → What are these "topics"?

- Terms are disambiguated in context
  - שור+שבת (Ox + Shabbat)

  → Associate a word to a topic
  → Associate a document to topics

Discovering Topic Models: LDA

- Latent Dirichlet Allocation
  - Blei and Jordan 2003
- Discover (unsupervised) topic structures in a document collection
- Topics are modeled as distributions of words
- Probabilistic generative model of text

What can be done with an LDA Topic Model?

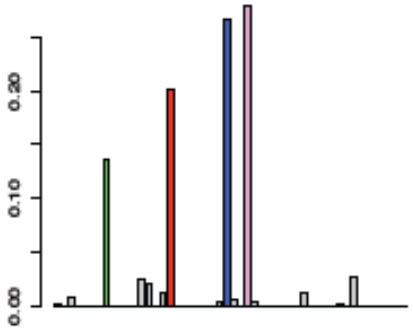
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions

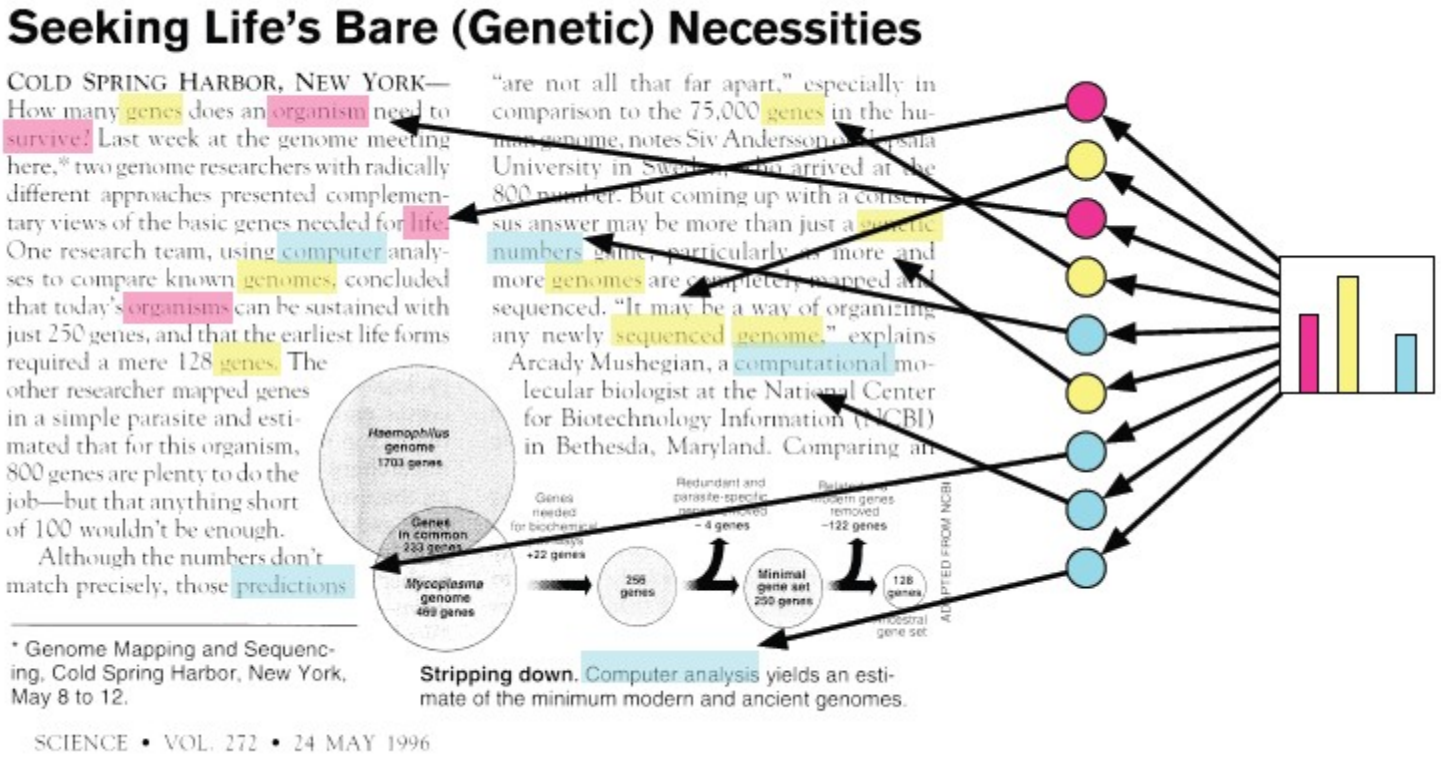


Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

D. Blei and J. Lafferty. Topic Models.
In A. Srivastava and M. Sahami, editors,
Text Mining: Theory and Applications. 2009

Structure of an LDA Model



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

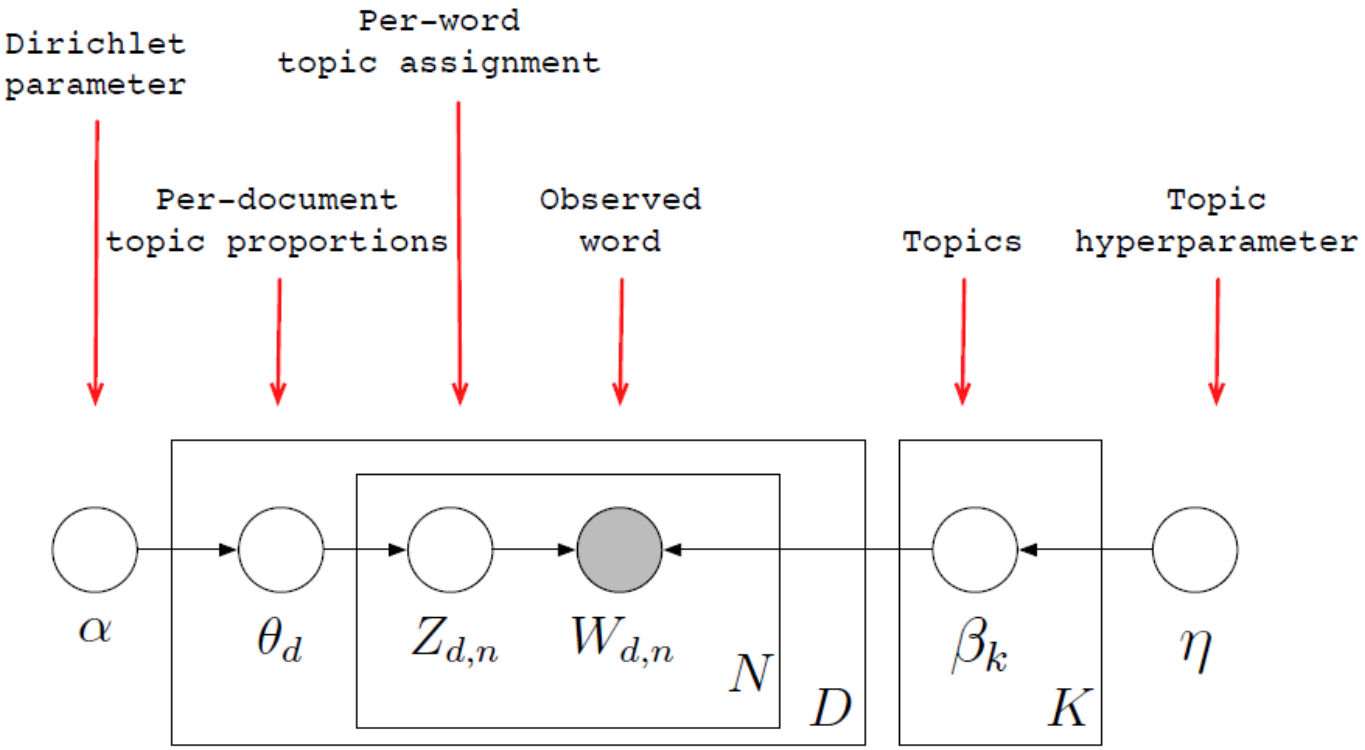
SCIENCE • VOL. 272 • 24 MAY 1996

)From (Blei 2008

The LDA Model

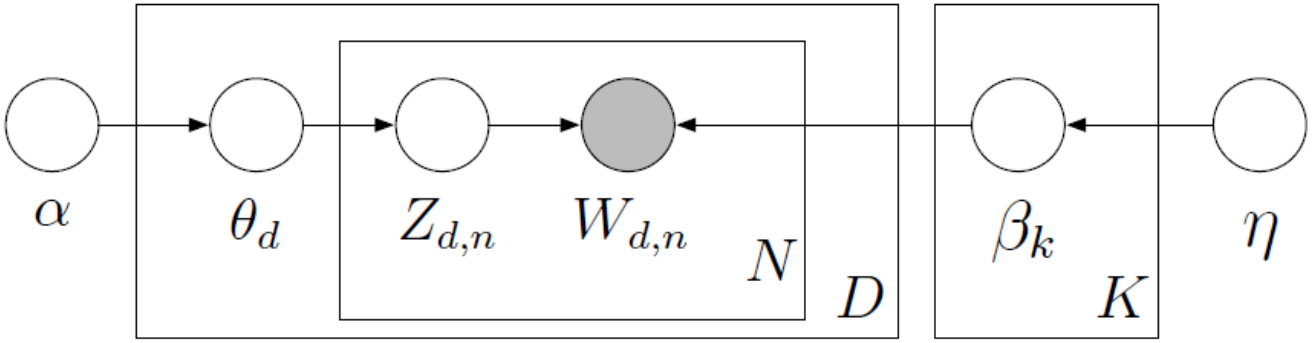
- Observations: documents are composed of words.
- Latent variable: each document expresses a few topics
- Generative probabilistic model:
  - Each document is a mixture of topics
  - Each word is drawn from the topics active in the document

LDA Graphical Model



)Blei 2008(Each piece of the structure is a random variable.

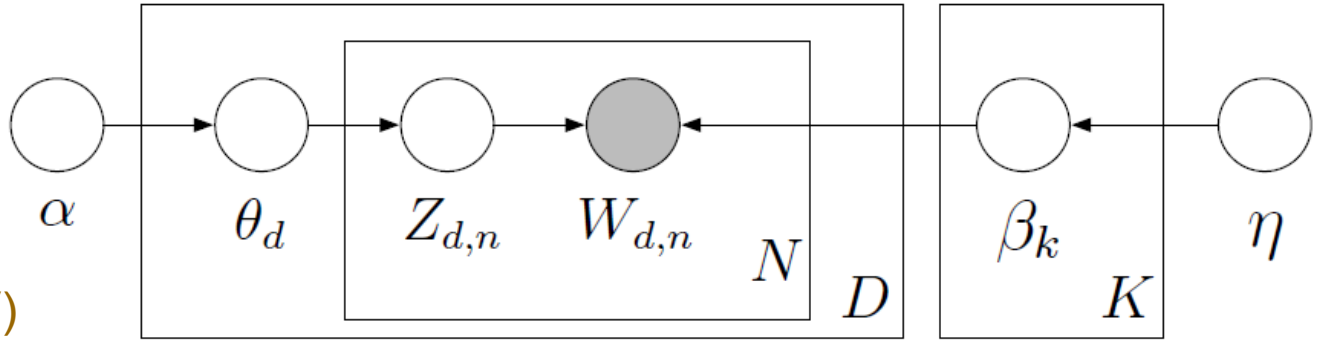
LDA Generative Process



1. Draw each topic $\beta_i \sim \text{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.
2. For each document:
   1. Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   2. For each word:
      1. Draw $Z_{d,n} \sim \text{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}})$.

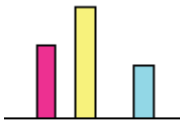
)Blei 2008(

LDA Generative Process



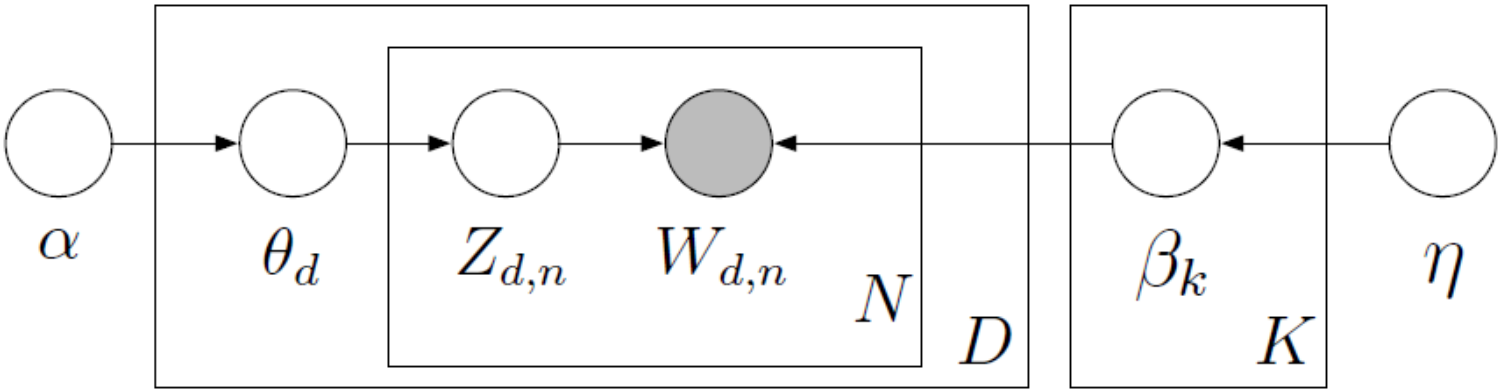
(w1, w2,…. wV)

1. Draw each topic $\beta_i \sim \text{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.
2. For each document:
   1. Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$. (t1,…,tK)
   2. For each word:
      1. Draw $Z_{d,n} \sim \text{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}})$.



)Blei 2008(

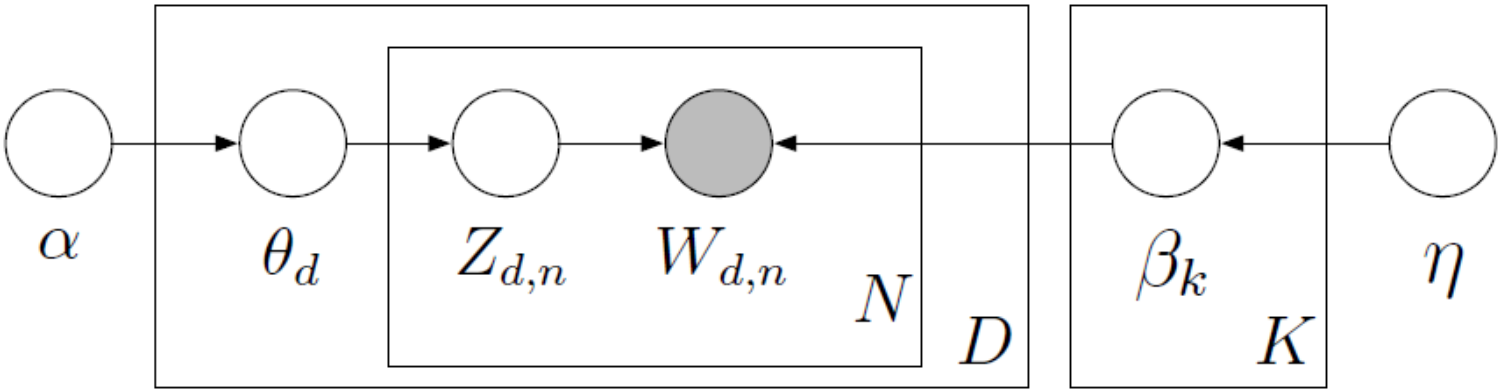
LDA Estimation



- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k

)Blei 2008(

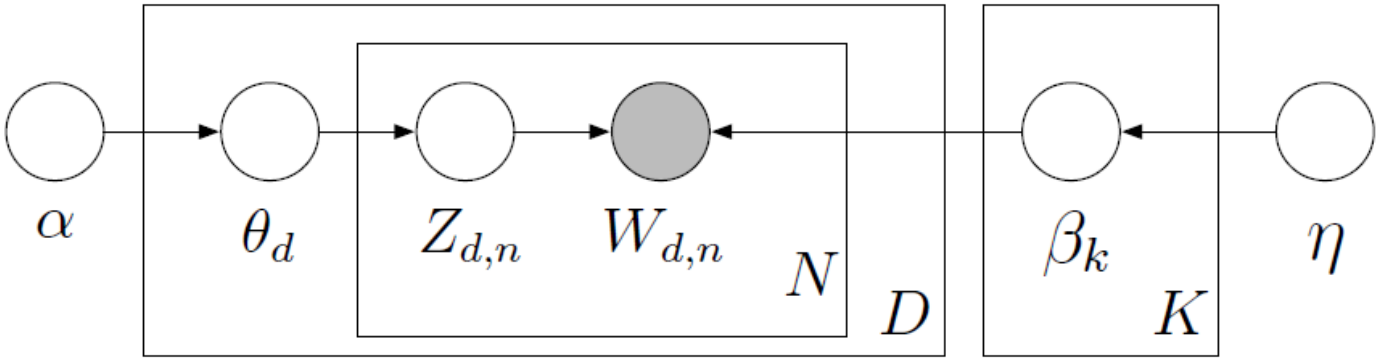
LDA Estimation



- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k ← Matrix KxV

)Blei 2008(

LDA Approximation



- Computing the posterior is intractable:

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

Generally use Gibbs Sampling to estimate

)Blei 2008(

Gibbs Sampling

- Represent corpus as:
  - Array of words w[i] ← fixed
  - Document indices d[i] ← fixed
  - Topics z[i] ← change
- Markov chain where states = topic assignments to words
- Macro-steps: assign a new topic to all the words
- Micro-steps: assign a new topic to each word w[i]

Outline

- Topic Analysis with LDA
- Domain: Halakhic Sources / Medical dataset
- Combining LDA and Morphological Analysis
- Combining Semantic Priors and LDA
- Multilingual Topic Models
- Evaluating Topic Models

LDA in Hebrew

- Explore various datasets in Hebrew
- How well does LDA work on Hebrew?

Domain: Halakhic Sources

Various Historical / Geographical background

	Period	**Work**	**Region**	**Language**
Tanaim	-200-200	Midrash, **Mishna**	Israel	Hebrew
Amoraim	200-500	Talmud	Babylonia, Israel	Aramaic
Geonim	500-1000	Responsa	Babylonia	Aramaic, Hebrew
Rishonim	1000-1500	Responsa **Codes**	Europe, North-Africa	Hebrew, Arabic
Aharonim	1500-now	**Responsa**	All	Hebrew

The Mishna

- Mishna (Tanaim)
  - Exhaustive code of Jewish Law
  - Written by R. Yehuda Hanasi (220 CE)
  - 6 orders, 63 tractates, 524 chapters, 6K paragraphs, 350K words.
  - Hierarchical thematic organization by topics

Rambam's Mishne Torah

- Corpus of Mishne Torah (Rishonim)
  - Exhaustive code of Halakha
  - Written by Maimonides 1170-1180
  - 14 books, 85 sections, 1,000 chapters, 15K articles, 600K words.

Responsa Corpus

We manually constructed a reference corpus for testing purposes.

Team of 5 Jewish Law experts with metadata associated to each QA document.

- **Documents**
  - 8,000 responsa from 35 distinct books of various origins (geographical, historical)
  - 3.6M words (avg 450 tokens per document)
  - On average 4.5 tags per document (from the ontology)
- **Ontology of Halakha**
  - ~2,000 concepts
  - ~5,000 relations among concepts of 14 distinct types
- **Metadata**
  - Per book: Author, Location, Publication Date
  - Per document:
    - Topics from index
    - References to "sources" (Bavli, Yerushalmi, Mishna, Tanakh, Shulhan 'arukh) (In progress)
    - References to other responsa (In progress)

Halakhic Corpus Specificity

- Language
  - Mixture (Hebrew + Aramaic)
  - Semitic languages: rich morphology
  - Many acronyms / abbreviations
- Wide variety of domains / historical background
- Various Genres
  - Codes (hierarchical, synthetic)
  - Commentaries (segmented, linear)
  - Responsa (implicitly hypertextual – complex citations)
- Layers of corpus (derivation, authority)
  - Mishna → Gmara → Mishne Tora → Responsa

Medical Corpus

- Infomed.co.il
  - Popular QA Health site
  - 2M words / 4K documents
  - Annotated by site categories
    - 6,000 concepts / 3,000 mapped to UMLS
- Hospital Patient release letters
  - Neurology department
  - 150K words / 1K documents
  - Manual UMLS concept annotation (in progress)

Medical Corpus Specificity

- Many unknown words (~20% token types)
- Many transliterations (Rafi's talk)
- Many named entities

Outline

- Topic Analysis with LDA
- Domain: Halakhic Sources / Medical dataset
- Combining LDA and Morphological Analysis
- Combining Semantic Priors and LDA
- Multilingual Topic Models
- Evaluating Topic Models

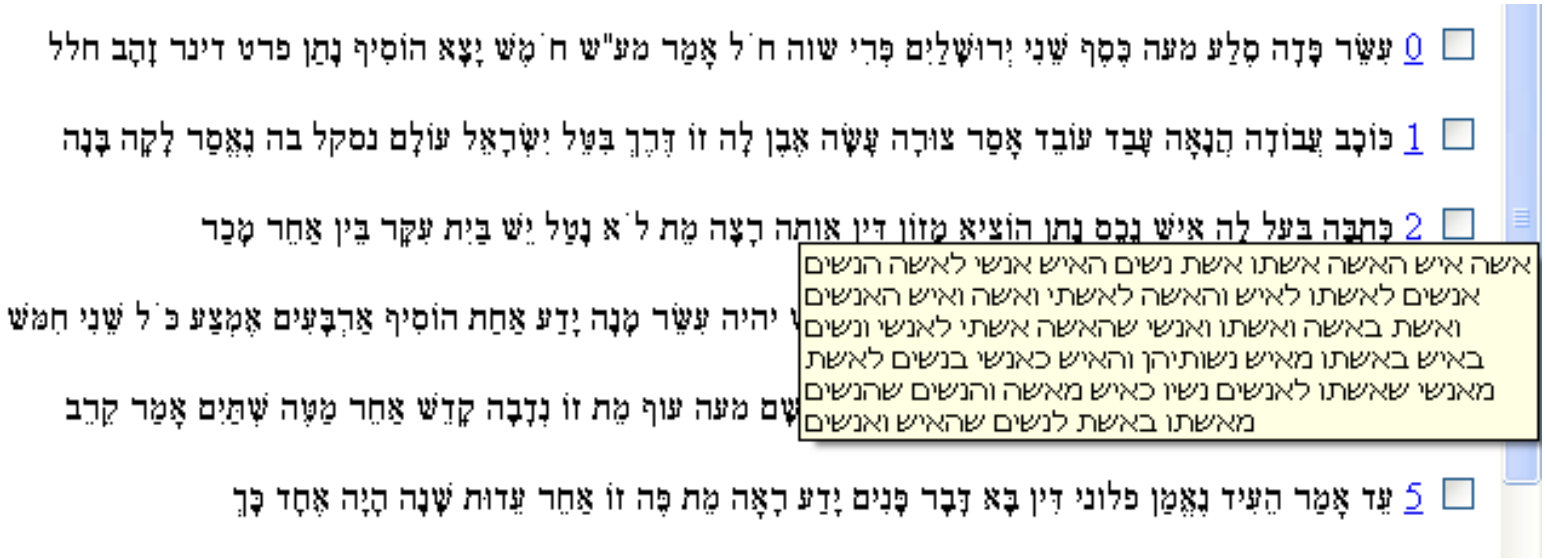
Hebrew Morphological Analysis

- בצלם
  - בְּצֶלֶם (name of an association)
  - בְּצַלֵּם (while taking a picture)
  - בְּצָלָם (their onion)
  - בְּצִלָּם (under their shades)
  - בְּצַלָּם (in a photographer)
  - בַּצַּלָּם (in the photographer(
  - בְּצֶלֶם (in an idol(
  - בַּצֶּלֶם (in the idol(

Morphological Analysis

- בְּצֶלֶם
  - בצלם proper-noun
- בְּצַלֵּם
  - בצלם verb, infinitive
- בְּצָלָם
  - ם-בצל noun, singular, masculine
- בְּצְלָם
  - ם-צל-ב noun, singular, masculine
- בְּצַלָּם בְּצֶלֶם
  - ב-צלם noun, singular, masculine, absolute
  - ב-צלם noun, singular, masculine, construct
- בַּצַּלָּם בַּצֶּלֶם
  - ב-צלם noun, definitive singular, masculine

…Many morphological variants



One word איש – about 50 distinct forms in the corpus
(12 forms average)

Combining LDA and Morphology

- LDA picks up patterns of word co-occurrence in documents.
- Heavy variations in Hebrew could mean we "miss" co-occurrence if we do not first analyze morphology.

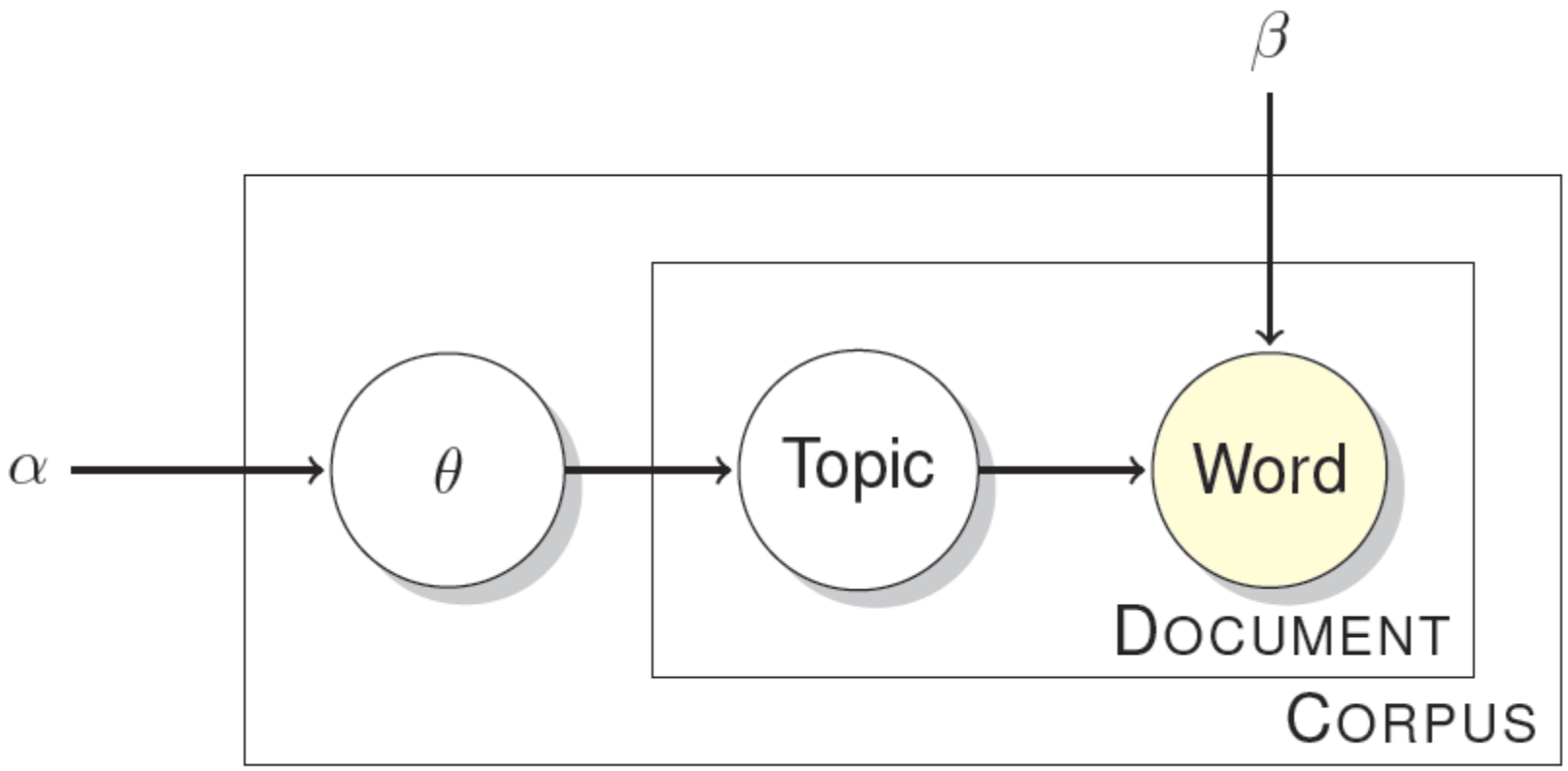
→What is the best method to combine LDA and Morphological analysis?

Combining LDA and Morphology

3 options:

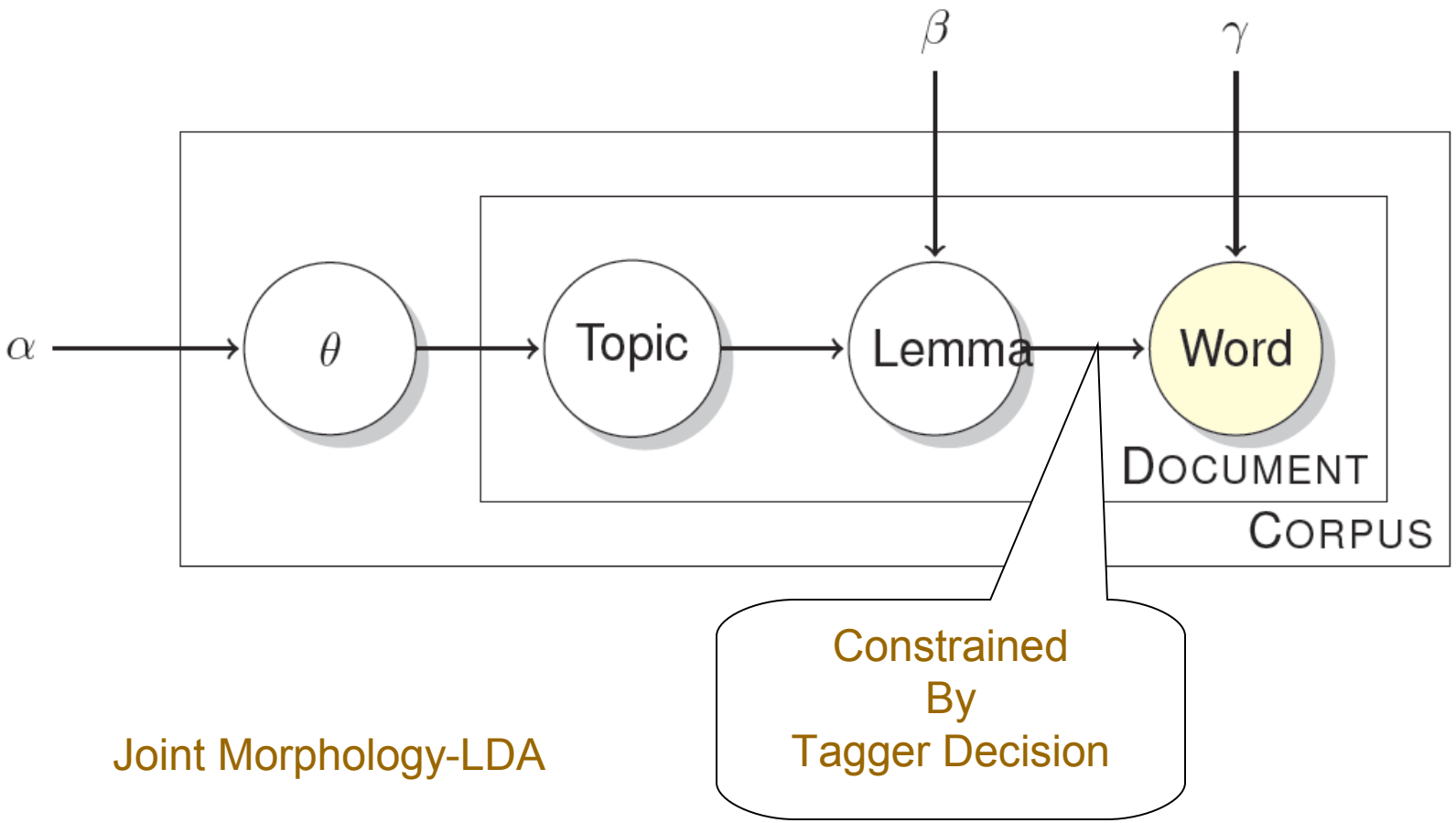
- **Ignore morphology** – token-based LDA
  - English LDA: stemming, filter POS (nouns)
- **Pipeline –** resolve morphological ambiguities, then learn LDA.
  - Morphology→Lemma is ambiguous
- **Joint** – learn LDA on distributions of lemma conditioned by morphological analysis

Joint LDA-Morphology Learning



Standard token-based LDA

Joint LDA-Morphology Learning



Joint Morphology-LDA

Joint LDA-Morphology works

- Token-based LDA in Hebrew gives no useful topics:
  - No semantic coherence (less than 1/3 topics)
  - No alignment with semantic annotations
- LDA-Morphology “works”
  - Semantic coherence
  - More on evaluation…

Morphology Variants

- Semantic Coherence Evaluation
  - Ask experts if they recognize a topic as coherent and to label it.
  - Test on Rambam 128 topics
    - 108 coherent topics with short label
    - 20 unrecognized [2 taggers / high agreement]
  - Test on Medical Data 128 topics
    - 115 coherent topics
  - Test on Mishna 128 topics
    - 60 coherent topics

Morphology Variants

- Variant models on Mishna Dataset
  - LDA on Nouns only
  - LDA on Nouns and Compound nouns (smixut)
- Semantic coherence only for Compound model
  - 80 coherent topics / 128 topics
  - Unstable: 75 coherent / 150 topics
- Marked Compounds
  - 45 compounds appear as top terms in topics (out of 6,500 distinct compounds)
  - All recognized as key concepts by domain experts
  - More evaluation needed on term extraction
  - Why such a difference with Rambam?

Outline

- Topic Analysis with LDA
- Domain: Halakhic Sources / Medical dataset
- Combining LDA and Morphological Analysis
- Evaluating Topic Models
- Combining Semantic Priors and LDA
- Multilingual Topic Models

?How Good are Discovered Topics

- Difficult to evaluate LDA topics
  - Many parameters
  - Each run gives slightly different results
  - How to compare topic models?
- Methods
  - Data-oriented evaluation
  - Semantic Coherence
  - Ontology alignment evaluation
  - Task-based evaluation

Topic Evaluation Methods 1

- Data-oriented:
  - Measure fit between dataset and generative model seen as language model (perplexity)
  - Seems to “miss” what is “good” about topics
- Semantic coherence
  - Subjective judgment
    - Individual topics meaningful? Can be labeled?
    - Assignments topic/docs meaningful?
  - Find the intruder tests
    - Rank best word / worst word – find the intruder word

Evaluating Topic Model

- "שור" (ox/bull) refers to many complex halakhic topics:
  - Damages (שור נוגח – goring ox)
  - Kosher meat (שחיטה – slaughter)
  - Sacrifices (קרבנות)
  - Shabbat (שבת – domestic animals must rest)
  - Calendar (מזל שור – Zodiac sign Taurus)

Topics for שור (Ox) on Rambam Corpus

Damages

Sacrifices

Calendar

Meat

65 שֹׁלֵם הַזִּיק בַּעַל נזק שׁוֹר בְּהֵמָה רִשׁוּת חַיָּב נִזּוֹק חֵץ פָּטַר בּוּר מוֹעֵד נָפַל הָיָה אדם את לוֹ חבירו דִּין 0.0217290799815

83 הַבִּיא מִנְחָה מִן כְּלִי אָמַר עֶלָה נָתַן שְׁמֶן לוֹג עשרון שֹׂר לְבוֹנָה מִן ק׳מֶץ יָצָא שָׁם חֵץ נָדַר מִזְבֵּחַ גַּב 0.00679501698754

72 בְּהֵמָה חָמוּר בה עֶלָה פָּרָה צָרִיךְ אִנָּה אדם מַחֲשָׁבָה עגלה כָּשֵׁר חֹשֶׁב חַי מְשַׁךְ הִכְנִיס שׁוֹר מְלָאכָה אַיִן כְּלֵב בְּעֵל 0.00636205899364

73 רָאָה מזל יָרֵחַ עִשָּׂר יהיה רְאִיָּה מַעֲלָה שֵׁנִי מִן ר׳אשׁ עָפוֹן יָדַע רִאשׁוֹן קֶשֶׁה גָּרַע דָּרוֹם לַיִל עוֹלָם חֵלֶק ממנו 0.00604686318972

54 בָּשָׂר מִן חֵי חֵלֶב בְּהֵמָה עוף דם אבר זַיִת דָּג אָכַל טמא טהור בֵּיצָה בה אָסַר מִין עִמָּם תּוֹרָה שָׁחַט 0.002

35 מוּם בְּכוֹר בְּהֵמָה בַּעַל תְּמוּרָה זוֹ אָמַר כ׳הֵן קָדַשׁ בה בו אִנָּה תָּמִים עָשָׂה נוֹלַד קָבוּעַ בַּיִת הִקְדִּישׁ מִזְבֵּחַ ח׳ל 0.00176574455562

92 גֵּר אֵשׁ הִדְלִיק אוֹר גַּב תַּנּוּר עֵץ כְּלִי קדרה בִּשֵּׁל חֲנֻכָּה- גַּחֶלֶת תַּבְשִׁיל הניח הוֹסִיף תּוֹךְ עֶלָה צָרִיךְ חִשְׁכָּה חַם 0.00129449838188

66 פָּתוּחַ סתום כִּי אֶל דִּבֶּר משה אָמַר ' שְׁתַּיִם בֶּן יי אִישׁ את שְׁלֹשׁ עָשָׂה ויהי כולן יוֹם הוּא שֵׁשׁ 0.00125680770842

53 שָׁחַט פֶּסַח שְׁחִיטָה אָכַל שֵׁנִי את סָכִין שם עָשָׂה רִאשׁוֹן שֹׂר אַחַר חֲבוּרָה עֲזָרָה דָּחָה דם ישחוט עליו ח׳ל אַרְבָּעָה 0.00125588697017

82 נָפַל כ׳ל מֵאָה ח׳ל סאה עִמָּם אָסַר תרומה תּוֹךְ הִתְעָרֵב יַיִן מין אִפּוּר כְּרֵם אי נָתַן כָּלָא עֶלָה דָּבָר שֵׁשׁ 0.00098167539267

81 קָדַשׁ מָעַל הִקְדִּישׁ נֶהֱנָה דם הקדש בְּדֶק בַּיִת פָּדָה פרט נָתַן מִזְבֵּחַ יָצָא בְּהֵמָה ח׳מֵשׁ יוּבַל דָּבָר אותה נָפַל הוֹסִיף 0.000669792364367

96 שֹׁלֵם גָּנַב תַּשְׁלוּם קֶרֶן אַרְבָּעָה חֲמֵשׁ דם דִּין גַּנַּב כְּפֶל נָתַן בַּעַל חַיָּב את ח׳מֵשׁ פָּטַר מִן מָכַר קָנָם ממון 0.000576701268743

49 מַיִם יָד רֶגֶל חָצַץ גּוּף ר׳אשׁ עין אדם נָטַל ידים עָלָה : נָתַן עַל בָּשָׂר עָבַל צָרִיךְ טְבִילָה גַּב שער 0.000497636227917

84 ( ) א׳כָל חוּץ אָכַל : נִכְנַס ד ו ב ג א אֶל יָד י נָתַן ה ז ח עֲנִי 0.000467508181393

64 קָדַשׁ אִישׁ אָמַר לי לָה זוֹ קִדֵּשׁ קִדּוּשִׁין את מִקְדָּשׁ נָתַן לוּ דָּבָר אִנָּה פרט סָפַק צָרִיךְ בַּת נִמְצָא מָנָה 0.000322684737012

Evaluation

Topics for שור (Ox) on Rambam Corpus

Evaluation

Topics for שור (Ox) on Rambam Corpus

Evaluation

Topics for שור (Ox) on Rambam Corpus

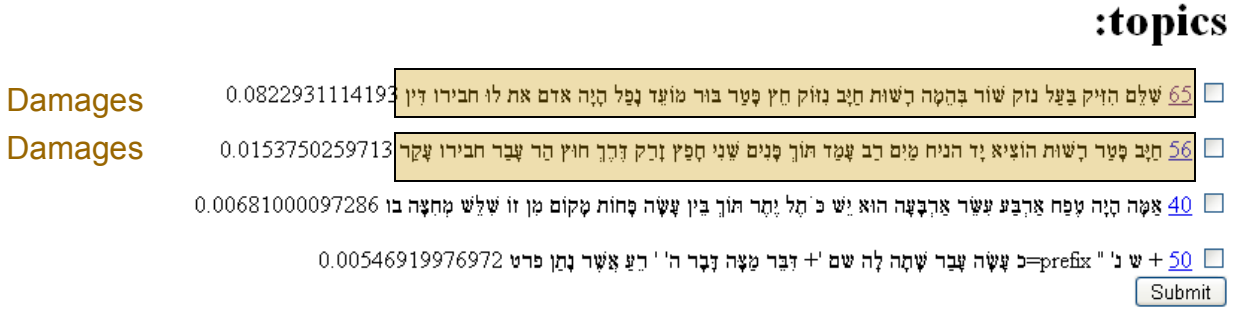
Evaluation

Topics for a Document

Damages

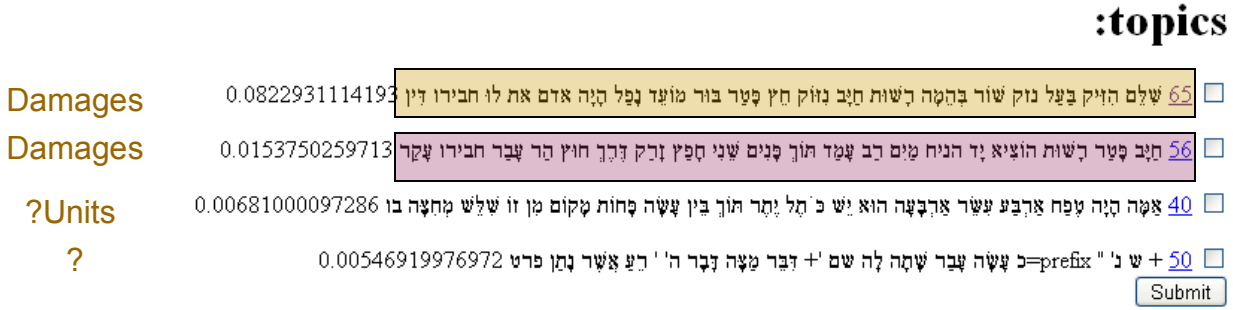
Damages

Rambam
Book of Damages
Damages by Property
Chapter 12



Evaluation

Topics for a Document



Rambam
Book of Damages
Damages by Property
Chapter 12

content משנה תורה - ספר נזקים - הלכות נזקי ממון פרק יב

החופר בור ברשות הרבים ונפל לתוכו שור חמור ומת אפילו היה הבור מלא צמר וכיוצא בהן בעל הבור חייב לשלם נזק שלם שנאמר בעל הבור
ישלם ואחד שור וחמור שאר מיני בהמה וחיה ועוף נאמר שור חמור בהווה אחד החופר בור ברשות הרבים והחופר ברשותו ופתח לרשות הרבים
פתח לרשות חבירו שחפר ופתח ברשותו רשותו הפקיר חייב בנזקיו הפקיר רשותו שהפקיר שברשותו הקדישו פטור שנאמר בעל הבור ישלם מי
שיש לו בעלים וזה הפקר ברשות חפר מפני שחפר ברשותו אחד החופר הבור מאליו בהמה חיה הואיל והוא חייב עשה חייב בנזקיו ואחד החופר
הלוקח שנתן לו במתנה שנאמר בעל הבור ישלם מי שיש לו בעלים מכל מקום אחד החופר מקום שהיה מכוסה שנאמר + שמות כ"א prefix=ל " ג
+ וכי יפתח איש בור כי איש בור כראוי אע"פ מתוכו ונפל לתוכן שור מת פטור שנאמר הא פטור בדבר שיכול לעמוד בפני שוורים ואינו יכול
לעמוד בפני גמלים והלכו עליו גמלים והלכו עליו שוורים ונפלו בו הגמלים מצויין באותו מקום פטור מפני שזה אונס יבואו שם גמלים אפילו חייב
מתוכו ונפלו בו שוורים אע"פ מצויין שם תמיד והרי הוא פושע הואיל ומחמת נפלו בו פטור כיוצא בזה המוצא בור וחזר בעל הבור חייב זה האחרון
פטור מתמו בעפר וחזר והוציא את העפר האחרון חייב שכיון בעפר מעשה הראשון בור שני שותפין ועבר עליו הראשון והשני הראשון חייב לשני
לשני ממנו נפטר הראשון ונתחייב השני הראשון ובא השני ומצאו מגולה השני חייב ועד אימתי יהיה השני לבדו חייב שידע הראשון מגולה וכדי
פועלים וכל שימות בו בתוך זמן השני לבדו חייב בו וכל שימות בו אחר זמן כזה שניהן חייבין לשלם שהרי שניהן בו המוסר לשומר השומר חייב
בנזקיו לחרש שוטה וקטן אע"פ שהיה מכוסה בבעלים חייבין עשוי ואלו בהן דעת חבירו ובא בעל הדלי ונטל בעל הבור חייב אחד החופר בור
מערה חריץ ולמה נאמר בור שיהיה בו כדי להמית וכמה כדי להמית עומק עשרה טפחים היה פחות מעשרה ונפל לתוכו שור שאר בהמה חיה ועוף
ומת פטור חייב בעל בור שלם היה עומק הבור תשעה ומהן טפח אחד מים חייב מים חשוב שני טפחים ביבשה היה שמונה ומהן שני טפחים מים
שהיה שבעה ומהן שלשה טפחים מים ונפל לתוכו שור וכיוצא בו ומת מחייבין אותו לשלם תפש הניזק מוציאין מידו שהדברים האלו ספק יש בהן
החופר בור עמוק עשרה טפחים ובא אחר לעשרים ובא שלישי לשלשים כולן חייבין חפר הראשון פחות מעשרה אפילו טפח ובא האחרון לעשרה
בין שחפר בו טפח שהגביה בנין שפתו טפח האחרון חייב סתם טפח שהוסיף טפח שבנה ספק כבר מעשה הראשון עדיין חפר הראשון בור עמוק
ובא האחרון ונפל לתוכו שור ומת מחמת מת האחרון פטור שהרי מיעט מחמת מת האחרון חייב שהרי הוא הקריב בור נפל השור מאותו הצד
האחרון האחרון חייב שהרי הקריב בור אע"פ שמת מן ההבל מן הצד שחפר הראשון נפל הראשון חייב שזה האחרון מיעט בור עליו התורה אפילו
מתה בהמה ואין צריך לומר מתה לפיכך היה עומק הבור לו הבל בו בהמה ומתה פטור היה יתר רחבו יש לו הבל מתה בו בהמה חייב אע"פ
שלא עשה תל גבוה ברשות הרבים בו בהמה ומתה היה גבוה עשרה טפחים חייב לשלם היה פחות מעשרה פטור מיתת הבהמה הם בלבד חייב
לשלם נזק שלם ואפילו גבוה שהוא שהוא בכל שהוא דבר מצוי וידוע ואין המיתה בכל שהוא מצויה והרי הוא כמו אונס אינו חייב מיתת הבהמה

Evaluation

Topic Evaluation Methods 2

- Alignment Topic Model / Ontology
  - Does the topic model reproduce existing metadata classification
- Task-based Evaluation
  - Do topics facilitate search or navigation?
  - For IR, relevance models with semantic smoothing
  - Do multilingual topics capture word alignments?

Semantic Coherence

- Subjective evaluation
  - Topic is meaningful / can be labeled?
- Highly positive on Rambam and Medical
- Low on Mishna until restricted to Compound+N / Marked morphologically

- Can topic semantic coherence be predicted?
  - (Newman et al 2010) using PMI measure

Ontology Alignment

- Rambam Mishne Torah has existing structure
  - Hierarchy of Book/Section/Chapter
- We find good alignment Topic/Book
  - Some topics are “cross-concern” (*witnesses*)

Topic → Documents

- Fits the Rambam's classification

65 שְׁלֵם הַזִּיק בַּעַל נזק שׁוֹר בְּהֵמָה רָשׁוּת חַיָּב נִזּוֹק חֵץ פָּטַר בּוֹר מוֹעֵד נָפַל הָיָה אדם את לו חבירו דִּין

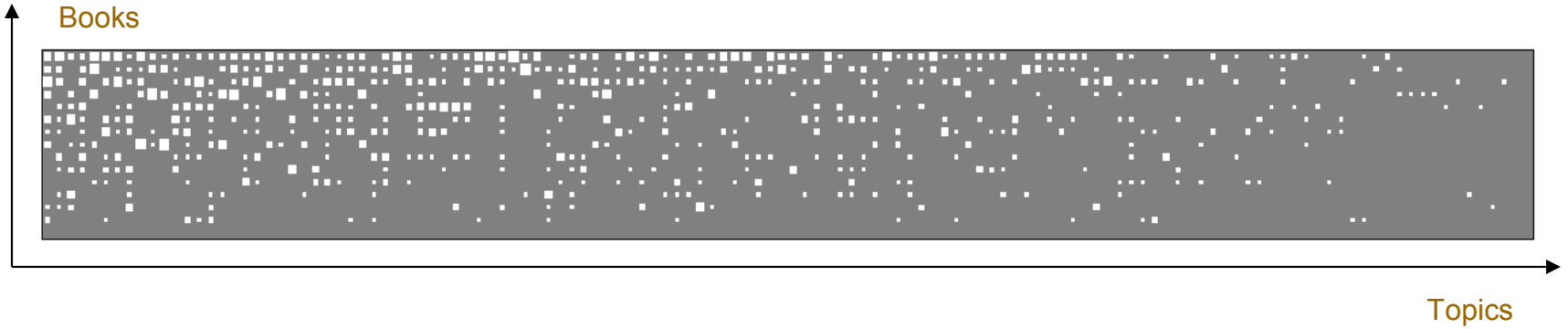
Submit

p(doc|topic):

('all', ('NZIKIN', 'nzki-mmown', '000012td00017000793000000prk_ib.txt')),0.0822931114193
('all', ('NZIKIN', 'nzki-mmown', '000007td00017000788000000prk_z.txt')),0.0721220527046
('all', ('NZIKIN', 'nzki-mmown', '000009td00017000790000000prk_t.txt')),0.0686546463245
('all', ('NZIKIN', 'nzki-mmown', '000003td00017000784000000prk_g.txt')),0.0674988441979
('all', ('NZIKIN', 'nzki-mmown', '000002td00017000783000000prk_b.txt')),0.0665742024965
('all', ('NZIKIN', 'nzki-mmown', '000001td00017000782000000prk_a.txt')),0.0605640314378
('all', ('NZIKIN', 'chowbl-owmzik', '000007td00017000832000000prk_z.txt')),0.0485436893204
('all', ('NZIKIN', 'chowbl-owmzik', '000006td00017000831000000prk_ow.txt')),0.0483125288951
('all', ('NZIKIN', 'nzki-mmown', '000010td00017000791000000prk_i.txt')),0.043458159963
('all', ('NZIKIN', 'nzki-mmown', '000013td00017000794000000prk_ig.txt')),0.0429958391123
('all', ('NZIKIN', 'nzki-mmown', '000008td00017000789000000prk_ch.txt')),0.042071197411
('all', ('NZIKIN', 'nzki-mmown', '000004td00017000785000000prk_d.txt')),0.0390661118816
('all', ('NZIKIN', 'nzki-mmown', '000014td00017000795000000prk_id.txt')),0.0314378178456
('all', ('NZIKIN', 'nzki-mmown', '000006td00017000787000000prk_ow.txt')),0.0298196948682
('all', ('NZIKIN', 'nzki-mmown', '000011td00017000792000000prk_ia.txt')),0.0235783633842
('all', ('SHOFTIM', 'snhdrin', '000005td00017001013000000prk_h.txt')),0.0228848821082

Alignment Topic / Books

Document – Book on Rambam's topic model
Document = (book[1-14] / section[1-85] / document)



5 general topics / 20 focus on 2 books / 30 skinny / 65 focus on 1 book
1 book covers many topics / 2 books very few
ZRAIM MADA ZMANIM NZIKIN AVODA KINYAN TAHARA KORBANOT AHAVA MISHPATIM SHOFTIM NASHIM HAFLAA KDUSHA

Outline

- Topic Analysis with LDA
- Domain: Halakhic Sources / Medical dataset
- Combining LDA and Morphological Analysis
- Evaluating Topic Models
- Combining Semantic Priors and LDA
- Multilingual Topic Models

Semantics and LDA

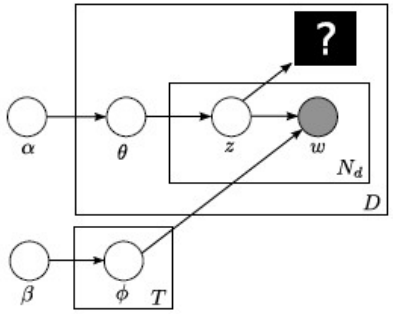
- LDA is fully unsupervised
- Learn better models with underlying semantic knowledge?
- Active field of research
  - Excellent survey: *Incorporating domain knowledge in latent topic models* (Andrzejewski 2010)

Semantics and LDA: 3 Types of Approaches

- LDA+X:
  - Model additional observed data (Document+Tag)
  - SupervisedLDA, Author-Topic, Topic-Link LDA
- Word-Topic Constraints
  - Prior constraints on word-topic association
  - Syntax: Syntactic Topic Model, HMM-LDA
  - Concept-Topic Model (semantic fields), LDAWN, Dirichlet Forest, Topic-in-Set
- Document-Topic Constraints
  - Prior constraints on document-topic association and among topics
  - Topic relations: hLDA, Correlated Topic Models, PAM
  - Document-Topic: Dirichlet Multinomial Regression, labeled LDA, Logic LDA
  - Topics over time: DTM, TOT

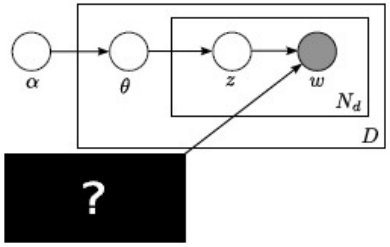
Semantics and LDA: 3 Types of Approaches

Document-Tag Observed



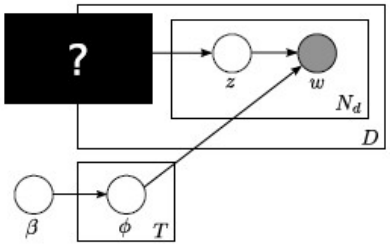
Images, labels

Word-Topic Conditions



Concepts, WordNet

Topic-Topic Conditions



Topic correlations

Which Method for our domain

- Document-Tags are available
  - Labeled LDA and DMR
  - Hierarchical topic models (PAM)
- Hyperlinks exist but are difficult to extract
  - LinkLDA
- Currently experimenting with Labeled-LDA on our datasets.

Outline

- Topic Analysis with LDA
- Domain: Halakhic Sources / Medical dataset
- Combining LDA and Morphological Analysis
- Evaluating Topic Models
- Combining Semantic Priors and LDA
- Multilingual Topic Models

Multilingual Topic Models

- Assume bilingual document set (di, li)
- Can we catch patterns of word co-occurrence across languages?

MUTO (Boyd-Graber & Blei 2009)

- Combine 2 aspects in one generative model:
  - Align words across languages
  - Group words into topics

MUTO Generative Process

- Choose matching m (m_{st} weight of (w_s, w_t))
- Choose multinomial term distributions:
  - Choose background distributions for words not in m for (S,T) ρl
  - Choose topic Ti ~ Dir(λ) – i in (1..K) over the pairs in m
  - For each document d (1..D) with language ld
    - Choose topics θd ~ Dir(α)
    - For each n in (1..Md)
      - Choose topic assignment zd ~ Mult(θd)
      - Choose cn from (matched, unmatched) uniformly
      - If cn = matched: choose a pair ~ Mult(βzn(m)) / project on ld
      - If cn = unmatched: choose wn ~ Mult(ρl)

Learned bi-lingual topic (En/Ge)

- time:schatten
- world:kontakt
- history:roemisch
- **number:nummer**
- math:with
- term:zero
- **axiom:axiom**
- **system:system**
- **theory:theorie**

Learned bi-lingual topic (En/Ge)

- time:schatten
- world:kontakt
- history:roemisch
- **number:nummer**
- math:with
- term:zero
- **axiom:axiom**
- **system:system**
- **theory:theorie**

Edit distance prior
A bilingual dictionary helps
Does much better on aligned corpora

- Could topic models over documents help MT with document level features?

Conclusions

- Morphological analysis is critical to start exploring topic models in MRLs
- Topic models are hard to evaluate
- Semi-supervised topic models improve quality of topics
- Multi-lingual topics can be learned

- Could help provide “document level” direction in MT