Text Mining in Hebrew

Impact of Morphology Analysis on Topic Analysis and on Search Quality

Michael Elhadad, Meni Adler, Yoav Goldberg, Dudi Gabay and Yael Netzer

Text in Hebrew

- Extract information from text in Hebrew
- Major immediate obstacle
  - Rich morphology
  - Very high number of distinct word forms
  - Very high ambiguity

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)

How Critical is Morphological Analysis to Text Mining?

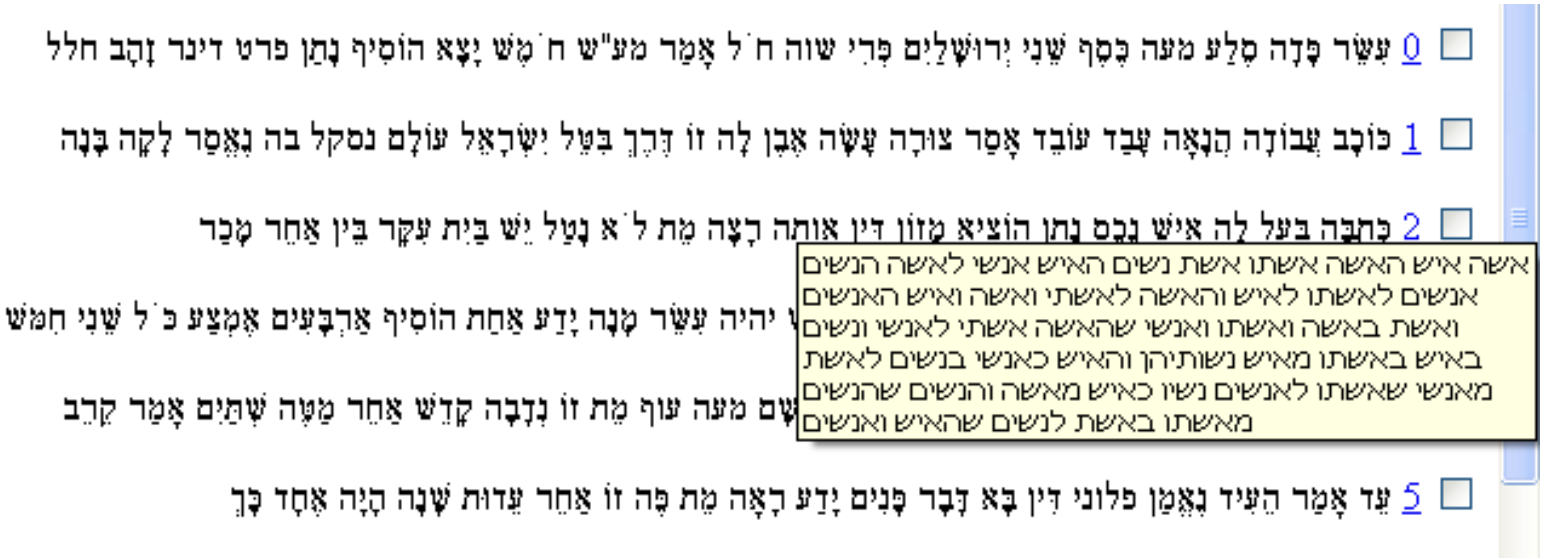
- How much does Hebrew morphology affect high-level text analysis tasks?
  - Named Entity Recognition
  - Information Extraction
  - Topic Analysis
  - Information Retrieval

Topic Analysis and Search

- Topic Analysis
  - Unsupervised discovery of topics in text collection
  - Useful to browse a large corpus by theme
  - Difficult to evaluate

- Faceted Search
  - Useful combination of search and browsing
  - Exploratory search (as opposed to fact finding)
  - Enabled by topic analysis

The Basic Idea



One word איש – about 50 distinct forms in the corpus

Outline

- Objectives
  - Topic Analysis in Hebrew
  - Improved Search
- Topic Analysis with LDA
- Obtaining Precise Morphology in Hebrew
- Combining LDA and Morphological Analysis
- Using Topic Models for Search
- Evaluating Topic Models
- Next Steps

Objectives

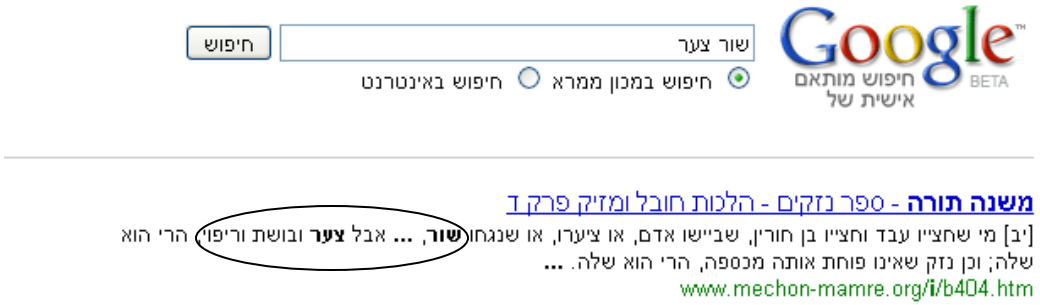
- Input:
  - Domain specific text corpus in Hebrew
- Output:
  - Topic model:
    - Discover “topics” discussed in the corpus
    - Recognize topics in unseen text
  - Index text collection by topic
- Task:
  - Search and browse text collection using topics

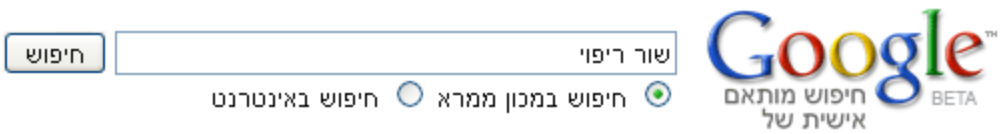
Example: Rambam's Mishne Torah

- Corpus of Mishne Torah
  - Exhaustive code of Halakha
  - Written by Maimonides 1170-1180
  - 14 books, 85 sections, 1000 chapters, 15K articles, 350K words.
- Creative compilation of laws from multiple sources:
  - Torah, Talmud (Bavli and Yerushalmi), Tosefta, halakhic midrashim (sifra and sifre), Geonim.
  - Synthetic hierarchical organization

Problems with Existing Search

- Morphology





A single “ו“ and the word is not found…

החיפוש שלך - **שור ריפוי** – לא תאם אף מסמך

Problems: Explore complex topics

- "שור" refers to many complex halakhic topics:
  - Damages (שור נוגח)
  - Kosher meat (שחיטה)
  - Sacrifices (קרבנות)
  - Shabat (שבת)
  - Calendar (מזל שור)

- Queries must be disambiguated
  - שור+שבת?

Exploratory/Faceted Search

- How to deal with ambiguous query terms?
  - Propose refinements according to contexts
  - "*Do you mean: damages, meat, shabat…*"
  - → Propose facets for query refinement
- Where do the topics (facets) come from?
- How do we disambiguate the query terms?
- Given a disambiguated topic, how do we refine the query?

Outline

- Objectives
- Topic Analysis with LDA
- Obtaining Precise Morphology in Hebrew
- Combining LDA and Morphological Analysis
- Using Topic Models for Search
- Evaluating Topic Models
- Next Steps

Discovering Topic Models: LDA

- Latent Dirichlet Allocation
  - Blei and Jordan 1993
- Discover (unsupervised) topic structures in a document collection
- Topics are modeled as distributions of words
- Probabilistic generative model of text

Topics for שור

- [ ] 65 שְׁלֵם הַזִּיק בַּעַל נזק שׁוֹר בְּהֵמָה רְשׁוּת חַיָּב נִזּוֹק חֵץ פָּטַר בּוֹר מוֹעֵד נָפַל הָיָה אדם את לוּ חבירו דִּין 0.0217290799815
- [ ] 83 הֵבִיא מִנְחָה מִן כְּלִי אָמַר עָלָה נָתַן שְׁמֶן לוג עשרון שַׂר לְבוֹנָה מִן ק׳ מִן יָצָא שָׁם חֵץ נָדַר מִזְבֵּחַ גַּב 0.00679501698754
- [ ] 72 בְּהֵמָה חָמוּר בה עָלָה פָּרָה צָרִיךְ אִנָּה אדם מַחְשָׁבָה עגלה כָּשֵׁר חֹשֵׁב חַי מְשַׁךְ הִכְנִיס שׁוֹר מְלָאכָה אֵין כְּלֵב בִּטֵּל 0.00636205899364
- [ ] 73 רָאָה מזל יָרֵחַ עֶשֶׂר יהיה רְאִיָּה מַעֲלָה שֵׁנִי מִן ר׳אשׁ פָּסוֹן יָדַע רִאשׁוֹן קֶשֶׁה גָּרַע דָּרוֹם לַיִל עוֹלָם חֵלֶק ממנו 0.00604686318972
- [ ] 54 בָּשָׂר מִן חַי חֵלֶב בְּהֵמָה עוף דם אבר זַיִת דָּג אָכַל טמא טהור בֵּיצָה בה אָסַר מין עֵצֶם תּוֹרָה שָׁחַט 0.002
- [ ] 35 מום בְּכוֹר בְּהֵמָה בַּעַל תְּמוּרָה זוֹ אָמַר כ׳הֵן קָדַשׁ בה בו אִנָּה תָּמִים עָשָׂה נוֹלַד קָבוּעַ בַּיִת הַקְדִּישׁ מִזְבֵּחַ ח׳ל 0.00176574455562
- [ ] 92 נֵר אֵשׁ הַדְלִיק אוֹר גַּב תַּנּוּר עֵץ כְּלִי קדרה בִּשֵּׁל חֲנֻכָּה- גַּחֶלֶת תַּבְשִׁיל הניח הוֹסִיף תּוֹךְ צָלָה 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 שַׂר אַחַר חֲבוּרָה עֲזָרָה דָּחָה דם ישחוט עליו ח׳ל אַרְבָּעָה 0.00125588697017
- [ ] 82 נָפַל כ׳ל מְאָה ח׳ל סאה עִסָּם אָסַר תרומה תּוֹךְ הִתְעָרֵב יַיִן מין אִפּוּר כְּרֵם אי נָתַן כָּלָא עָלָה דָּבָר שֵׁשׁ 0.00098167539267
- [ ] 81 קָדַשׁ מָעַל הִקְדִּישׁ נֶהֱנָה דם הקדש בְּדֶק בַּיִת פָּדָה פרט נָתַן מִזְבֵּחַ יָצָא בְּהֵמָה ח׳מֵשׁ יוֹבֵל דָּבָר אותה נָפַל הוֹסִיף 0.000669792364367
- [ ] 96 שְׁלֵם גָּנַב תַּשְׁלוּם קֶרֶן אַרְבָּעָה חֲמֵשׁ דם דִּין גַּנָּב כְּפֶל נָתַן בַּעַל חַיָּב את ח׳מֵשׁ פָּטַר מן מָכַר קָנָס ממון 0.000576701268743
- [ ] 49 מַיִם יָד רֶגֶל חָצַץ גּוּף ר׳אשׁ עין אדם נָעַל 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 קִדֵּשׁ אִישׁ אָמַר לי לָה זוֹ קִדֵּשׁ קִדּוּשִׁין את מִקְדָּשׁ נָתַן לוּ דָּבָר אִנָּה פרט סָפַק צָרִיךְ בַּת נִמְצָא מָנָה 0.000322684737012

LDA

Topics for a Document

:topics

- ☐ 65 שְׁלֵם הֲזִיק בַּעַל נזק שׁוֹר בְּהֵמָה רְשׁוּת חַיָּב נִזּוֹק חֵץ פָּטַר בּוֹר מוֹעֵד נָפַל הָיָה אדם את לוֹ חבירו דִּין 0.0822931114193
- ☐ 56 חַיָּב פָּטַר רְשׁוּת הוֹצִיא יָד הניח מַיִם רַב עָמַד תּוֹךְ פָּנִים שֵׁנִי חָפַץ זָרַק דֶּרֶךְ חוץ הַר עָבַר חבירו עָקַר 0.0153750259713
- ☐ 40 אַמָּה הָיָה עָפָר אַרְבָּעָה עשׂר אַרְבָּעָה הוא יֵשׁ כֹּ֫תֶל יָתֵר תּוֹךְ בֵּין עָשָׂה פָּחוֹת מָקוֹם מִן זוֹ שָׁלֹשׁ מְחָצָה בו 0.00681000097286
- ☐ 50 + ש נ' " prefix=כ עָשָׂה עָבַר שָׂתָה לָה שם '+ דִּבֵּר מָצָה דָּבָר ה' ' רֵעַ אָשָׁר נָתַן פרט 0.00546919976972

Submit

---

content

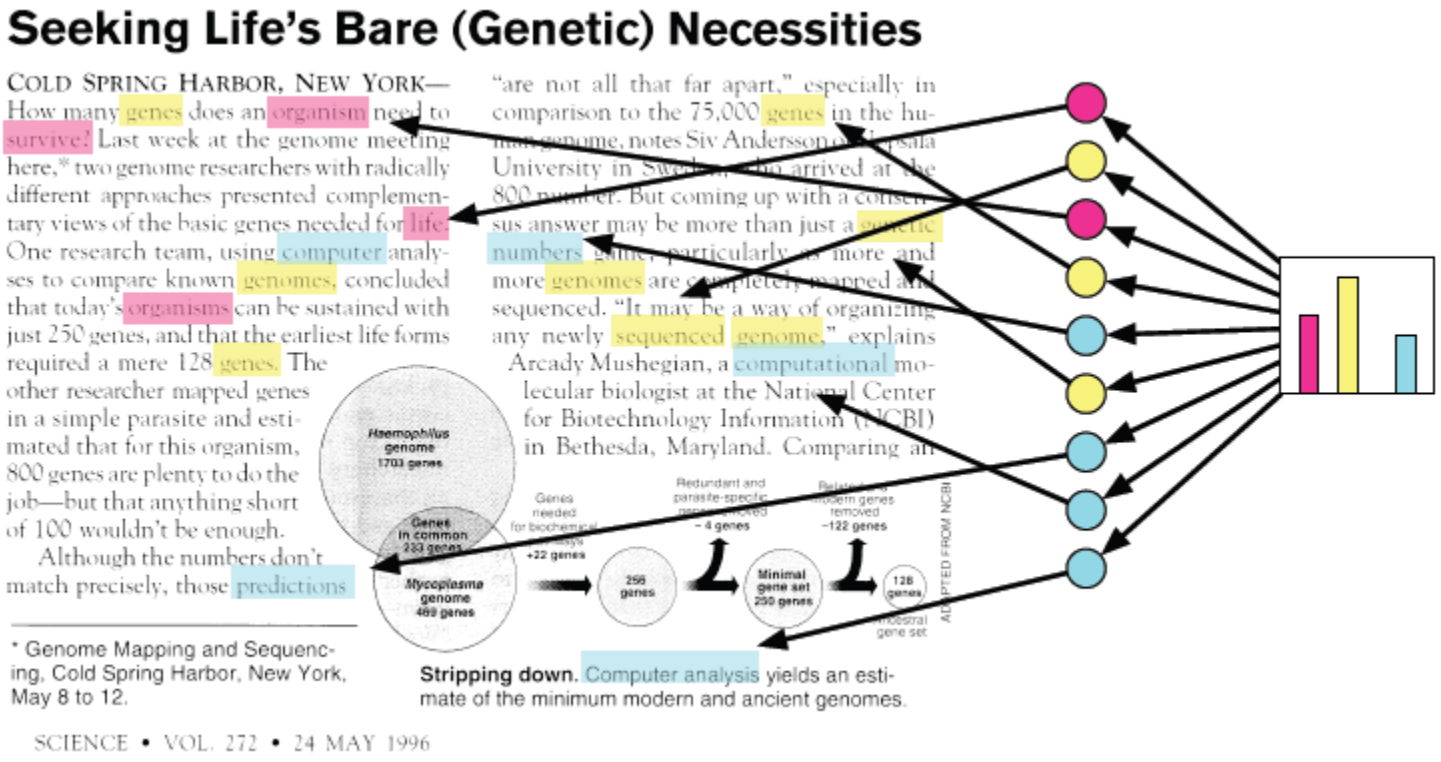
החופר בור ברשות הרבים ונפל לתוכו שור חמור ומת אפילו היה הבור מלא צמר וכיוצא בהן בעל הבור חייב לשלם נזק שלם שנאמר בעל הבור ישלם ואחד שור וחמור שאר מיני בהמה וחיה ועוף נאמר שור חמור בהווה אחד החופר בור ברשות הרבים החופר ברשותו ופתח לרשות הרבים פתח לרשות חבירו שחפר ופתח ברשותו רשותו הפקיר חייב בנזקיו הפקיר 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 כראוי אע"פ מתוכו ונפל לתוכו שור ומת 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 וזה האחרון פטור סתמו 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 הבעלים 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 נזק 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 בהמה הם בלבד חייב לשלם נזק שלם ואפילו גבוה שהוא בכל שהוא דבר מצוי וידוע ואין המיתה בכל שהוא מצויה והרי הוא כמו אונס אינו חייב מיתת בהמה

LDA

The LDA Model

- Observation: documents are composed of words.
- Intuition: documents exhibit multiple topics
- Generative probabilistic model:
  - Each document is a mixture of topics
  - Each word is drawn from the topics active in the document

Structure of the 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.

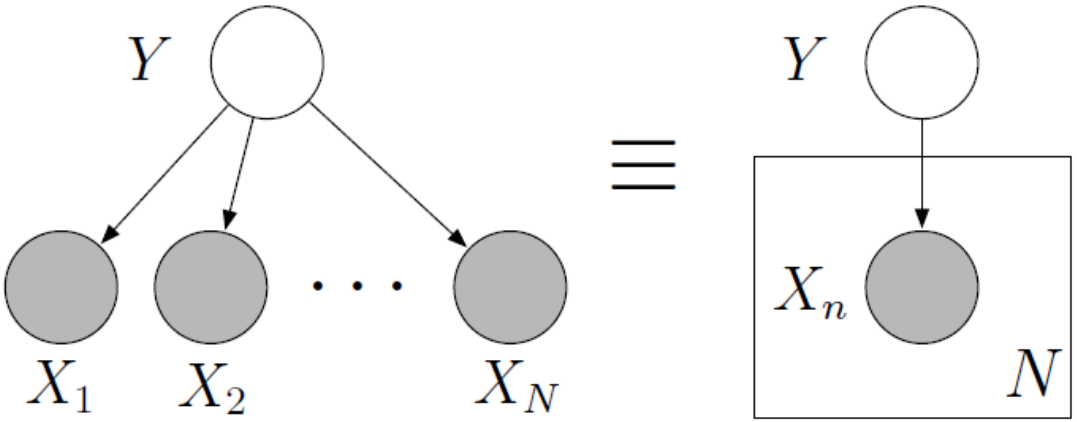
SCIENCE • VOL. 272 • 24 MAY 1996

From (Blei 2008)

Learning an LDA Model from Observations

- Observation: documents and words
- Objective: infer an underlying topic structure
  - What are the topics?
  - How are the documents divided according to those topics?

Graphical Models

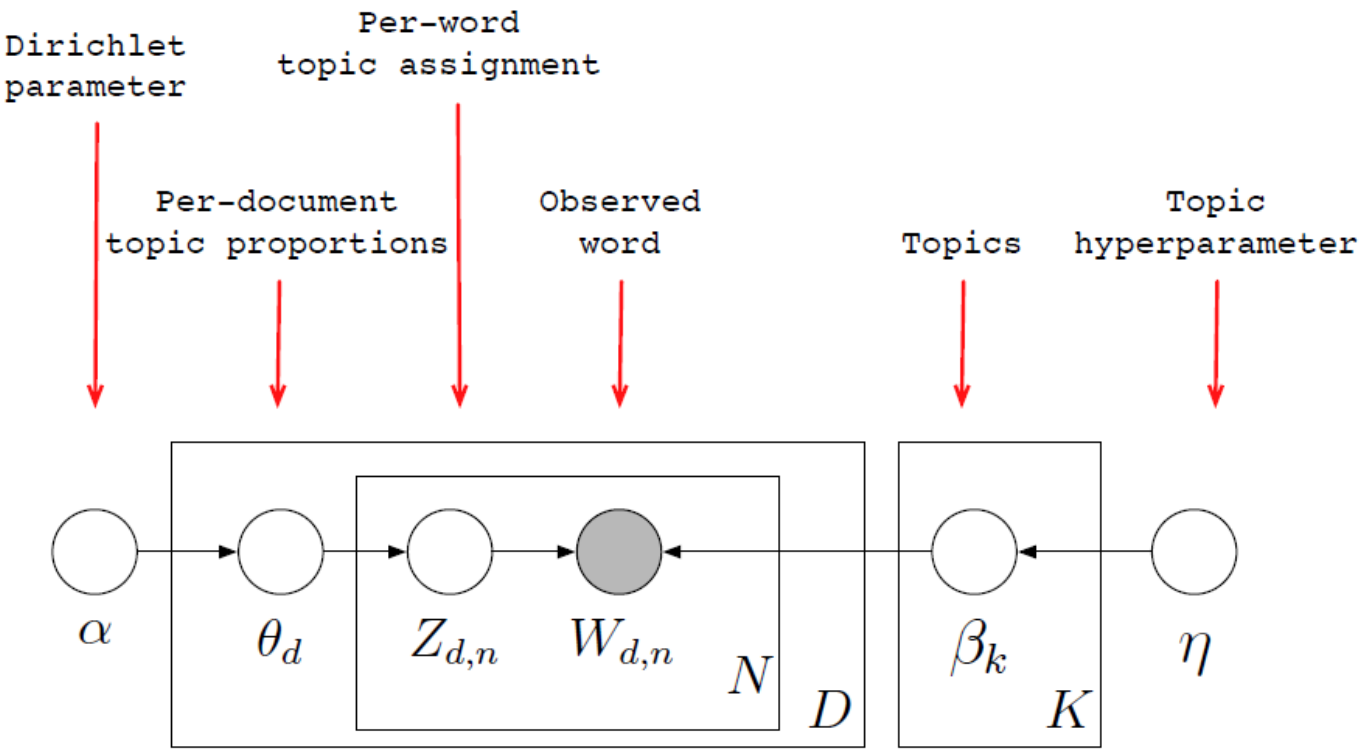


- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

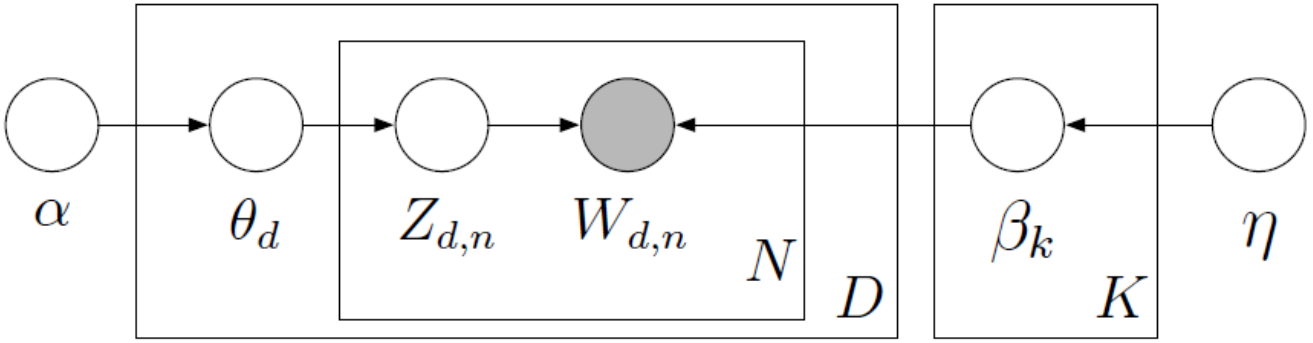
(Blei 2008)

LDA Graphical Model



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

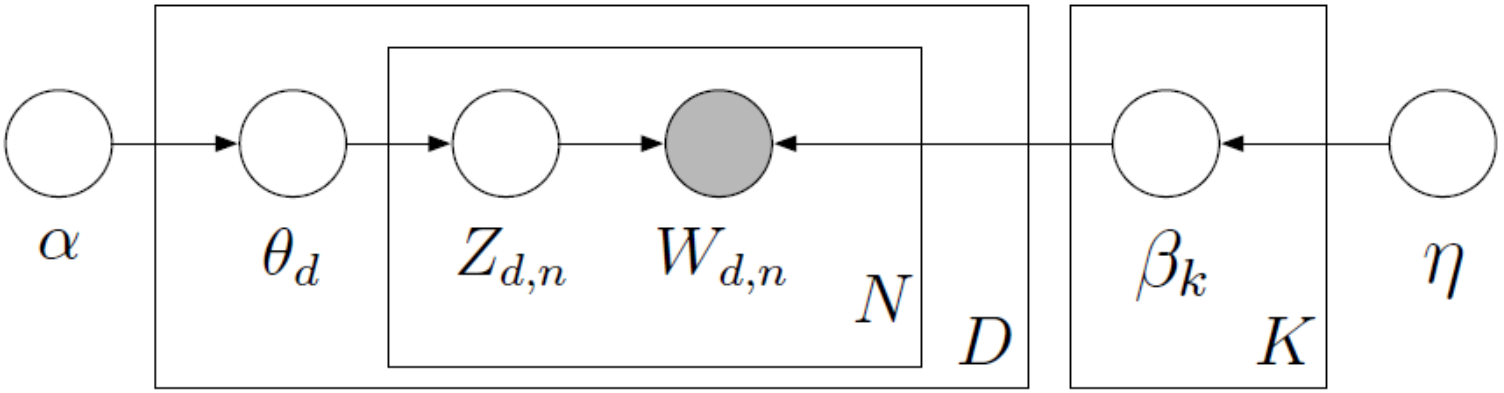
LDA Generative Process



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)

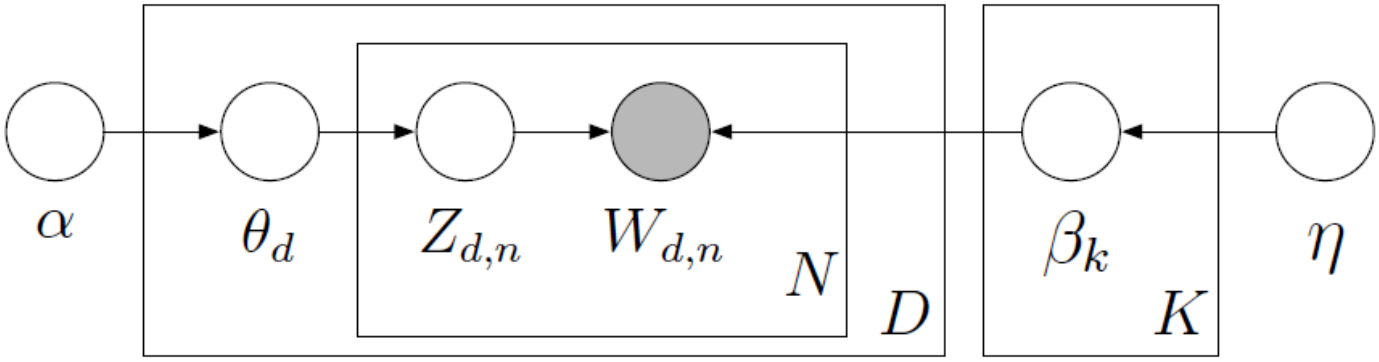
LDA Estimation



- 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 Approximation



- Computing the posterior is intractable:

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

- Several approximation techniques have been developed.

(Blei 2008)

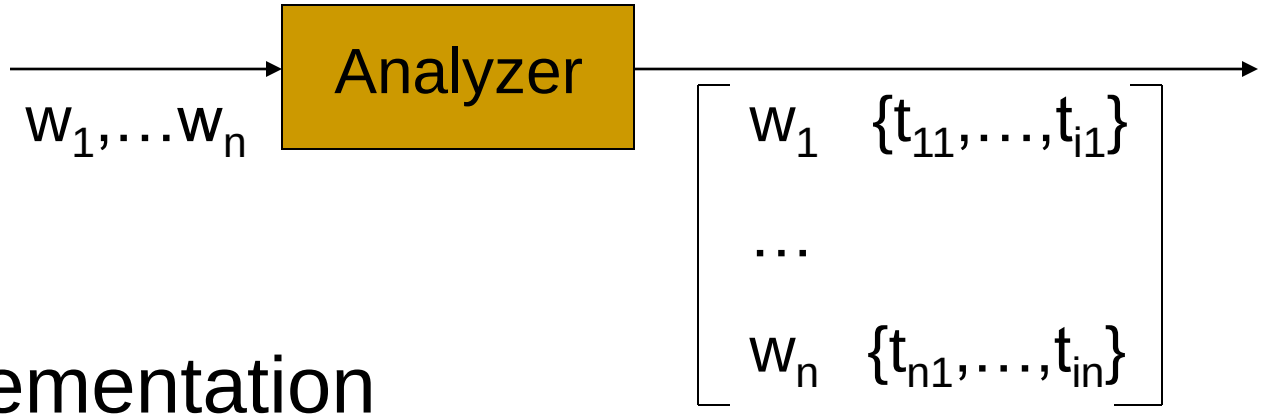
Outline

- Objectives
- Topic Analysis with LDA
- Obtaining Precise Morphology in Hebrew
- Combining LDA and Morphological Analysis
- Using Topic Models for Search
- Evaluating Topic Models
- Next Steps

Morphological Analysis

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

Morphological Analyzer

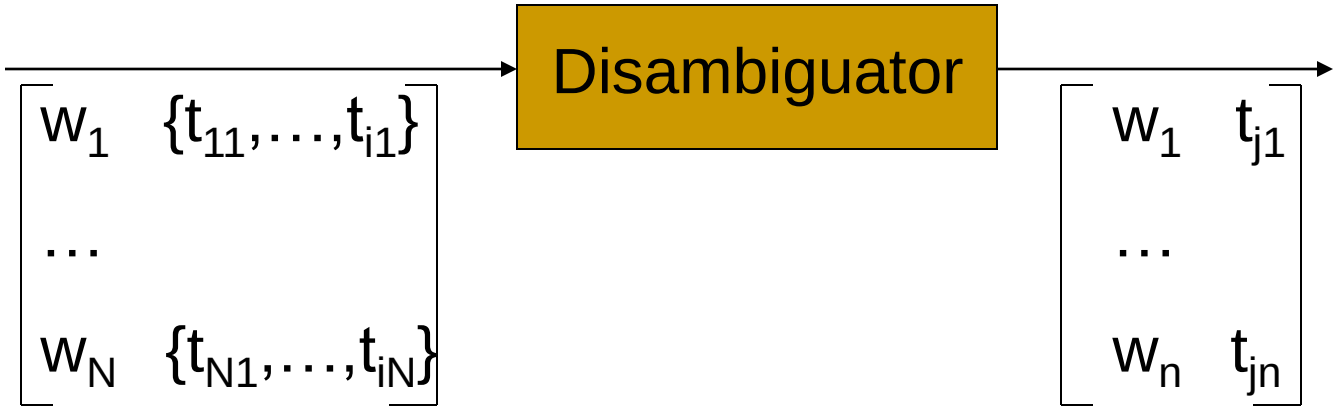


- Implementation
  - Corpus based
  - Lexicon based
    - Analytic
    - Synthetic

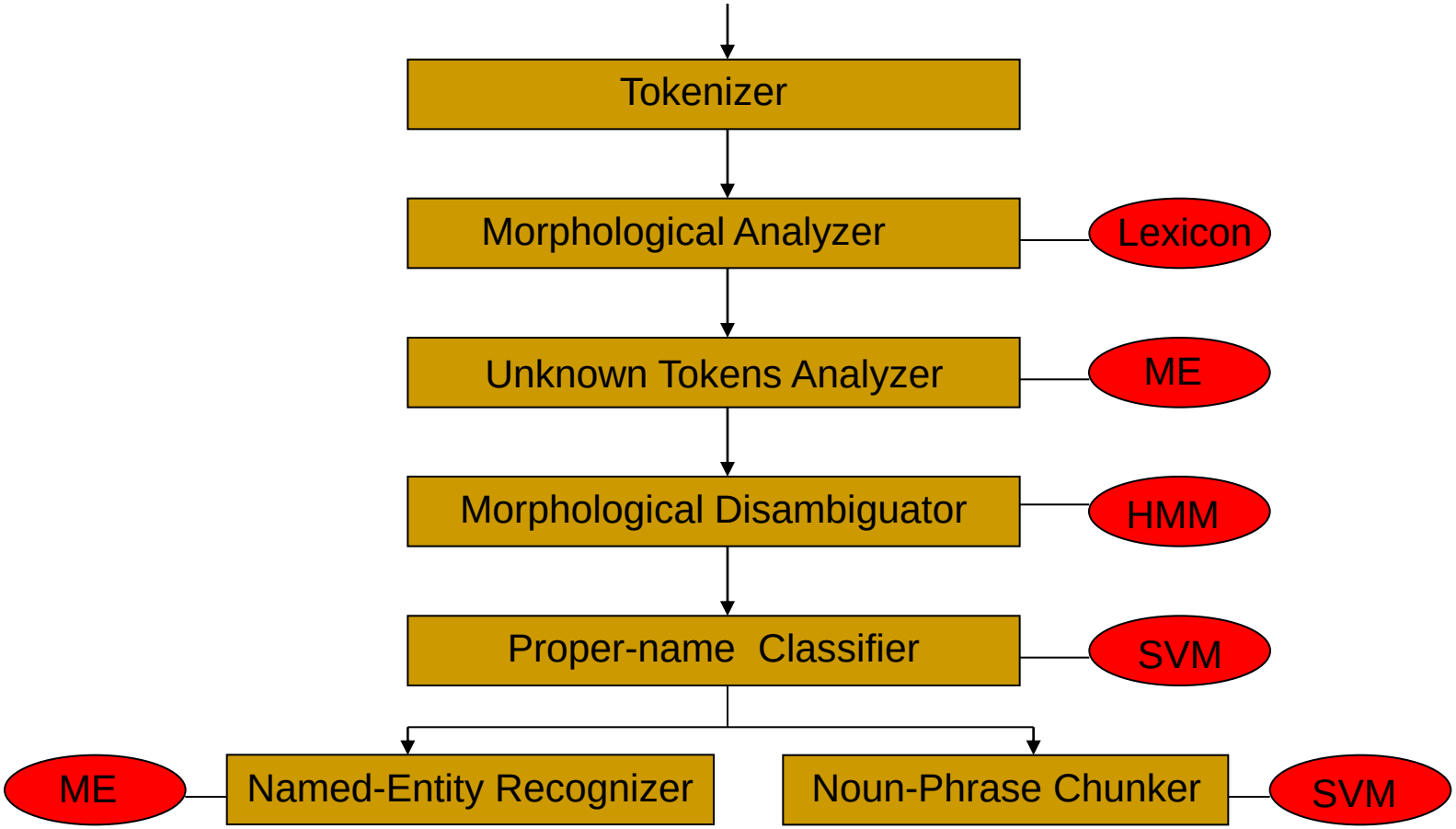
Morphological Disambiguation

- ארגון בצלם הודיע
- בצלם את המשחק הבחנתי בעוזר המאמן
- בצלם נחטף בשווקים
- בצלם הנעים חסינו
- נתקלתי בצלם מקצועי
- פגשתי בצלם חתונות
- נתקלתי בצלם המקצועי

Morphological Disambiguation



Hebrew Text Analysis System



http://www.cs.bgu.ac.il/~nlpproj/demo

Morphological Disambiguation - Methods

- Rule-based vs. Stochastic models
- Supervised vs. Unsupervised learning
- Exact vs. Approximate inference

Hidden Markov Model

- S – a set of states (= tags)
- O – a set of output symbols (= tokens)
- μ – a probabilistic model
  - State transition probabilities $A = \{a_{i,j}\}$
  - Symbol emission probabilities $B = \{B_{i,k}\}$

HMM- An Example

- S = {start, noun, verb}
- O = {ילד, ירח}
- μ = (A,B)

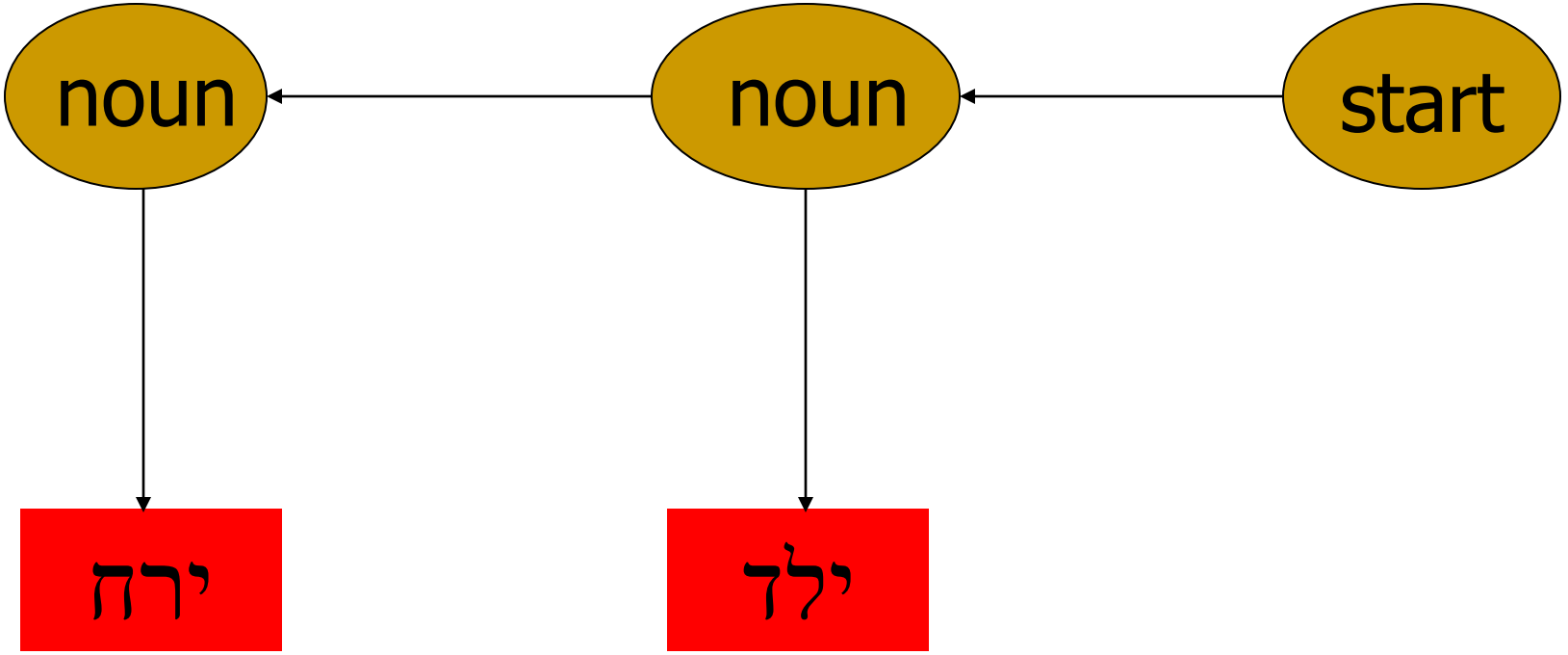
	noun	verb
start	0.8	0.2
noun	0.9	0.1
verb	0.9	0.1

A

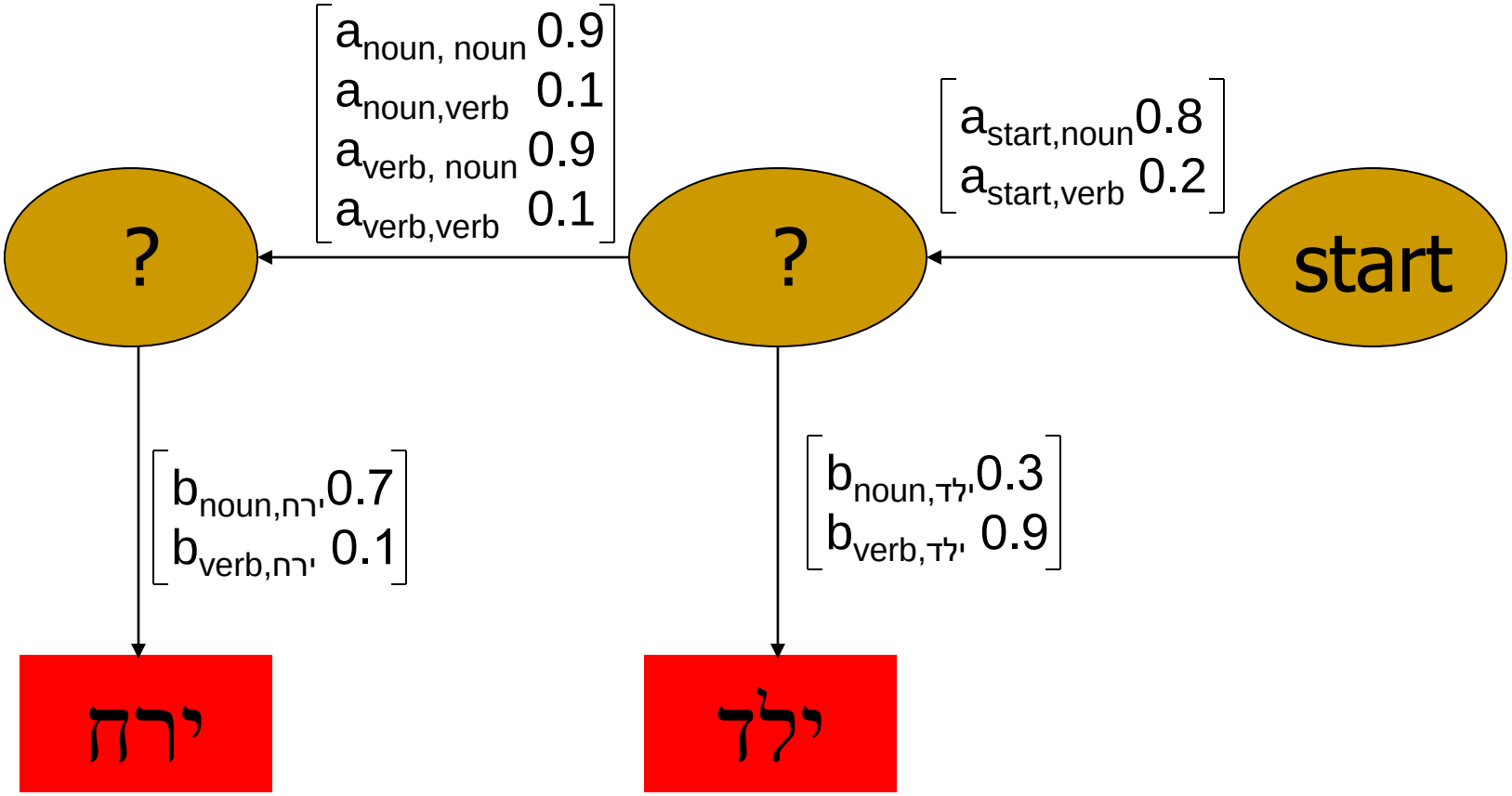
	ילד	ירח
noun	0.3	0.7
verb	0.9	0.1

B

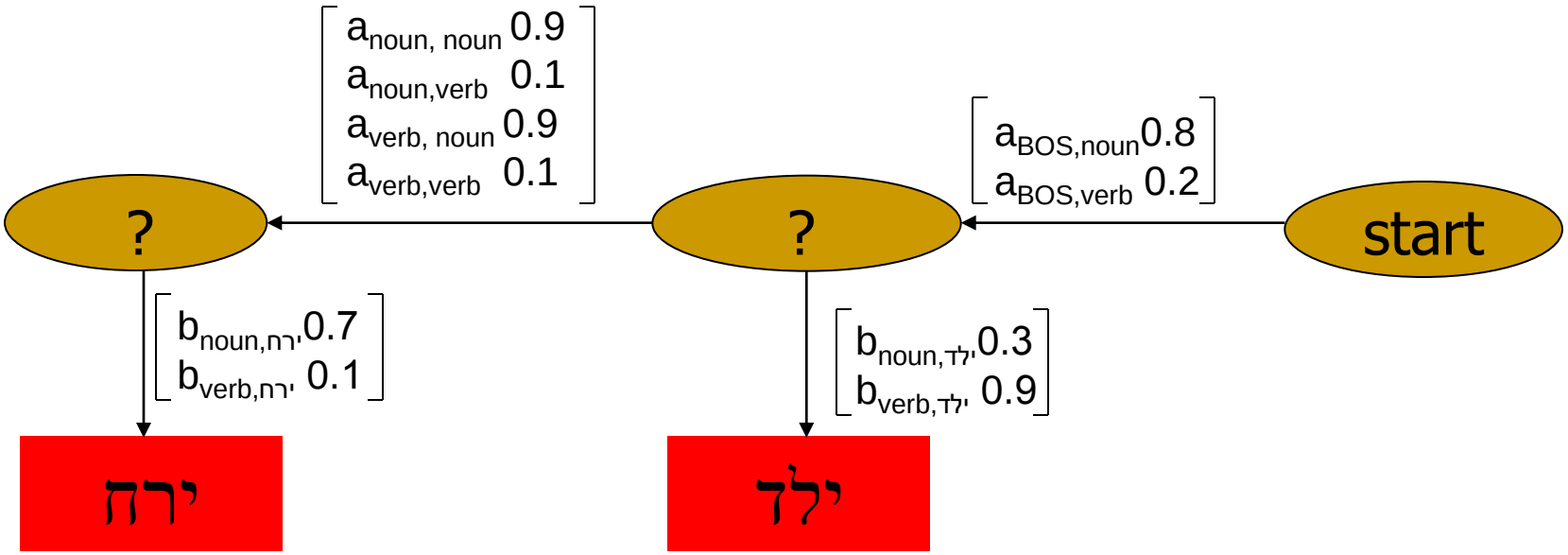
Markov Process



Decoding


$a_{noun,\ noun}$ 0.9
$a_{noun,verb}$ 0.1
$a_{verb,\ noun}$ 0.9
$a_{verb,verb}$ 0.1
$a_{start,noun}$ 0.8
$a_{start,verb}$ 0.2
?
?
start
$b_{noun,ירח}$ 0.7
$b_{verb,ירח}$ 0.1
$b_{noun,ילד}$ 0.3
$b_{verb,ילד}$ 0.9
ירח
ילך

Decoding



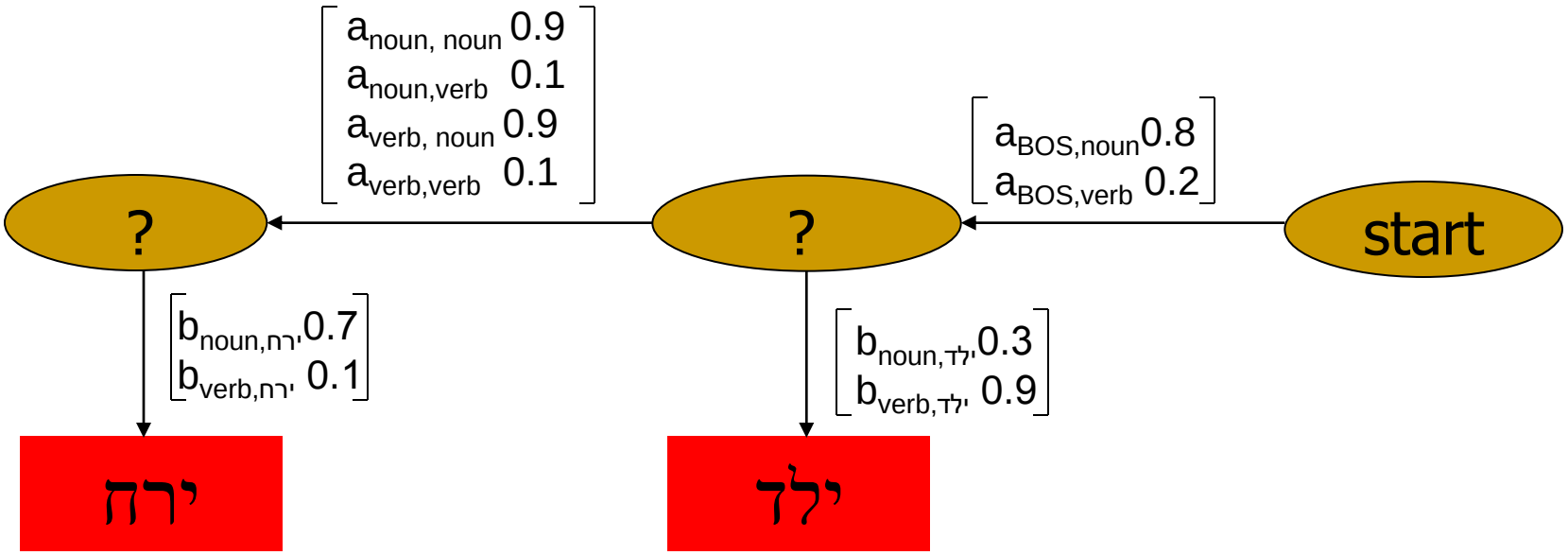
(noun, noun) = $a_{start,noun}b_{noun,ילד}a_{noun, noun}b_{noun,ירח}$= 0.8*0.3*0.9*0.7 = 0.1512

(noun, verb) = $a_{start,noun}b_{noun,ילד}a_{noun, verb}b_{verb,ירח}$ = 0.8*0.3*0.1*0.1 = 0.0024

(verb, noun) = $a_{start,verb}b_{verb,ילד}a_{verb, noun}b_{noun,ירח}$ = 0.2*0.9*0.9*0.7 = 0.1134

(verb, verb) = $a_{start,verb}b_{verb,ילד}a_{verb, verb}b_{verb,ירח}$ = 0.2*0.9*0.1*0.1 = 0.0018

Decoding



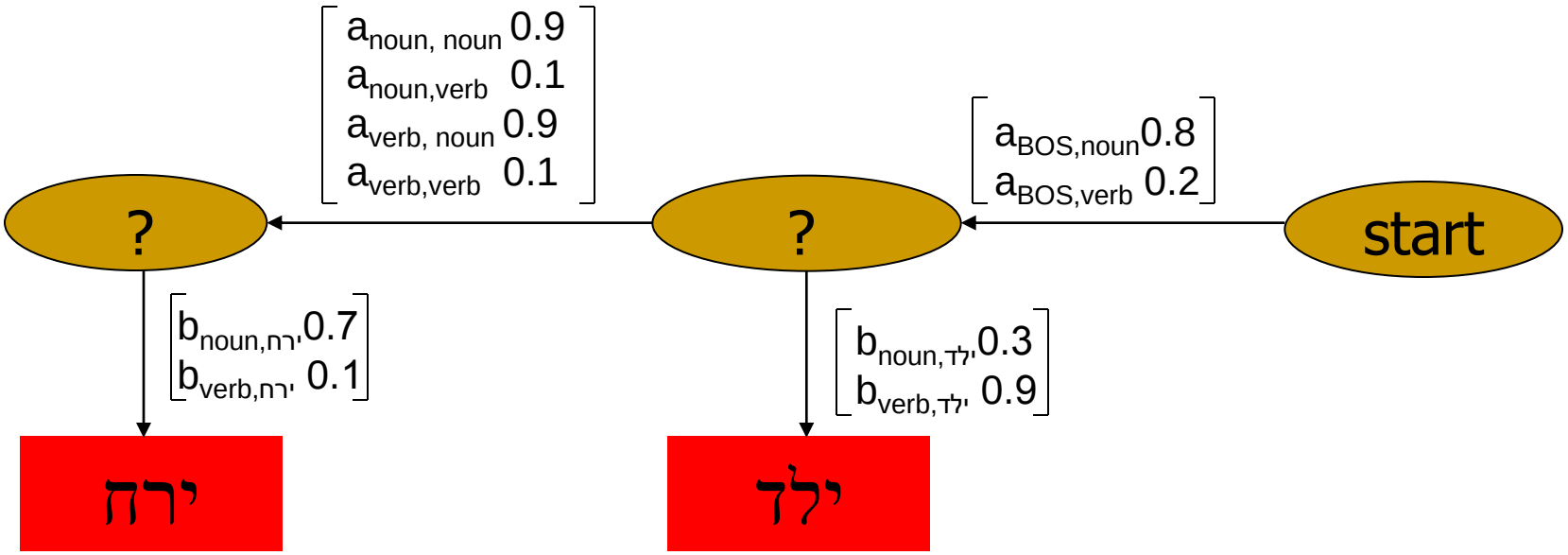
(noun, noun) = $a_{start,noun}b_{noun,ילד}a_{noun,\,noun}b_{noun,ירח}$ = 0.8*0.3*0.9*0.7 = 0.1512

(noun, verb) = $a_{start,noun}b_{noun,ילד}a_{noun,\,verb}b_{verb,ירח}$ = 0.8*0.3*0.1*0.1 = 0.0024

(verb, noun) = $a_{start,verb}b_{verb,ילד}a_{verb,\,noun}b_{noun,ירח}$ = 0.2*0.9*0.9*0.7 = 0.1134

(verb, verb) = $a_{start,verb}b_{verb,ילד}a_{verb,\,verb}b_{verb,ירח}$ = 0.2*0.9*0.1*0.1 = 0.0018

Decoding



(noun, noun) = $a_{start,noun}b_{noun,ילד}a_{noun,\ noun}b_{noun,ירח}$ = 0.8*0.3*0.9*0.7 = 0.1512

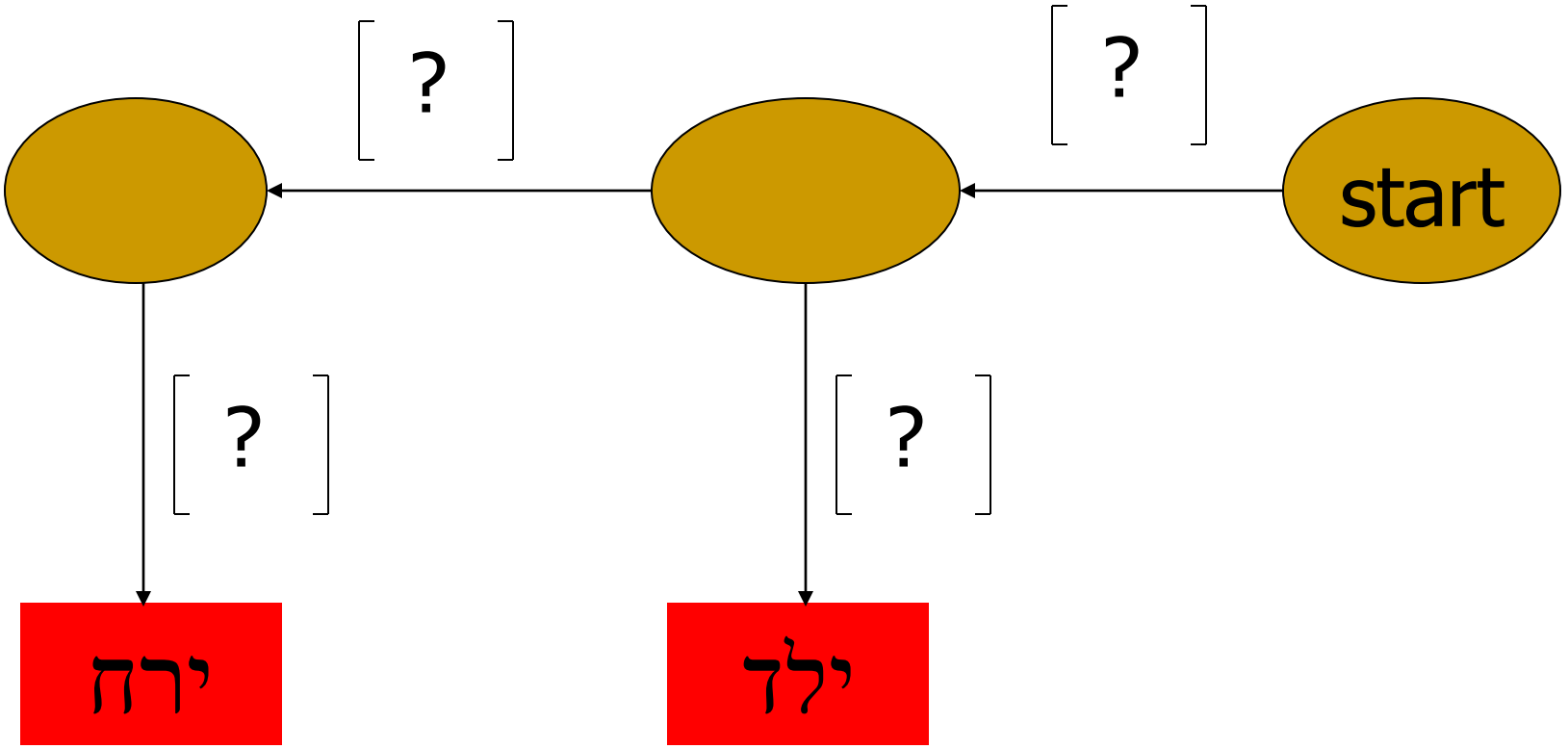
(noun, verb) = $a_{start,noun}b_{noun,ילד}a_{noun,\ verb}b_{verb,ירח}$ = 0.8*0.3*0.1*0.1 = 0.0024

(verb, noun) = $a_{start,verb}b_{verb,ילד}a_{verb,\ noun}b_{noun,ירח}$ = 0.2*0.9*0.9*0.7 = 0.1134

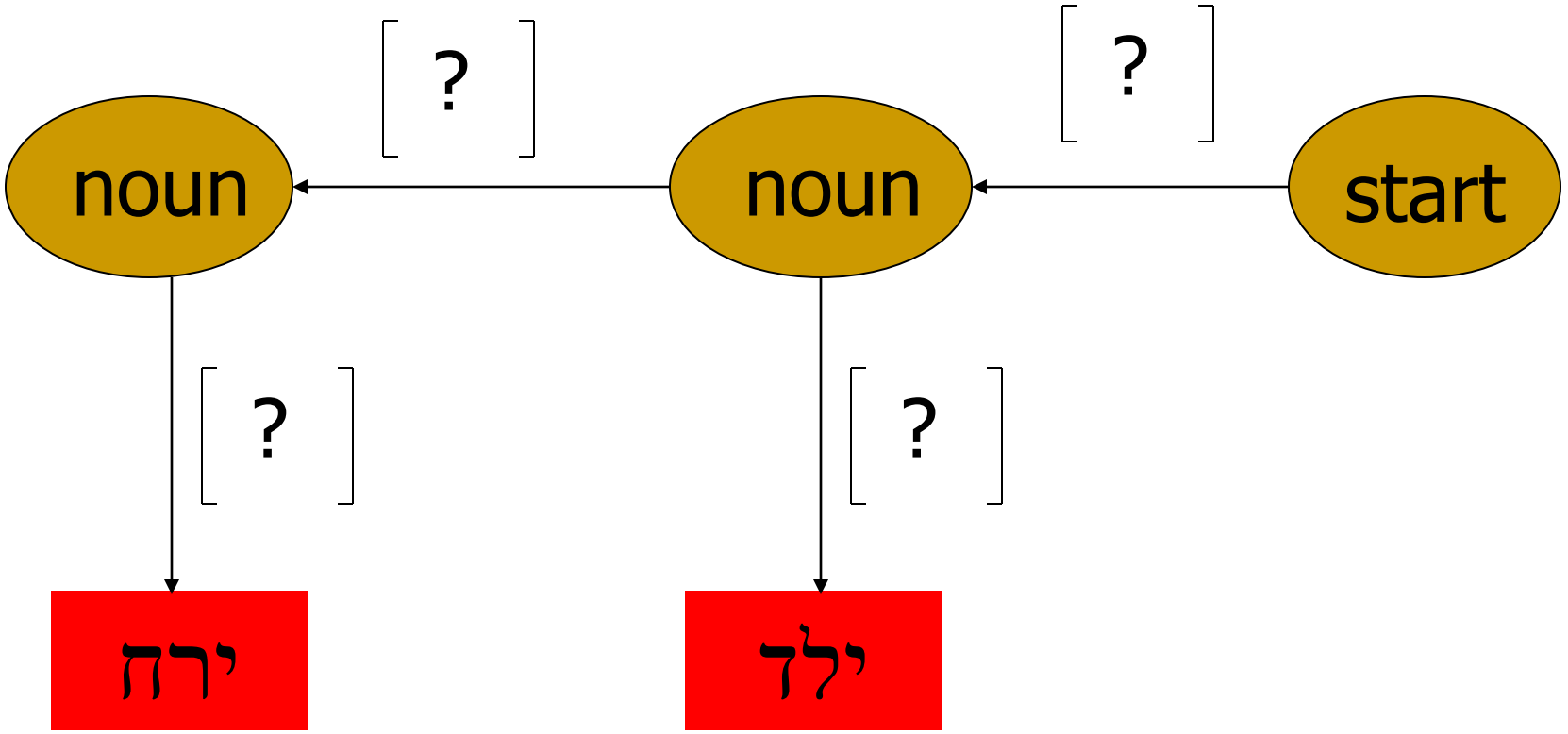
(verb, verb) = $a_{start,verb}b_{verb,ילד}a_{verb,\ verb}b_{verb,ירח}$ = 0.2*0.9*0.1*0.1 = 0.0018

Viterbi Algorithm (dynamic programming)

Parameter Estimation



Supervised Parameter Estimation

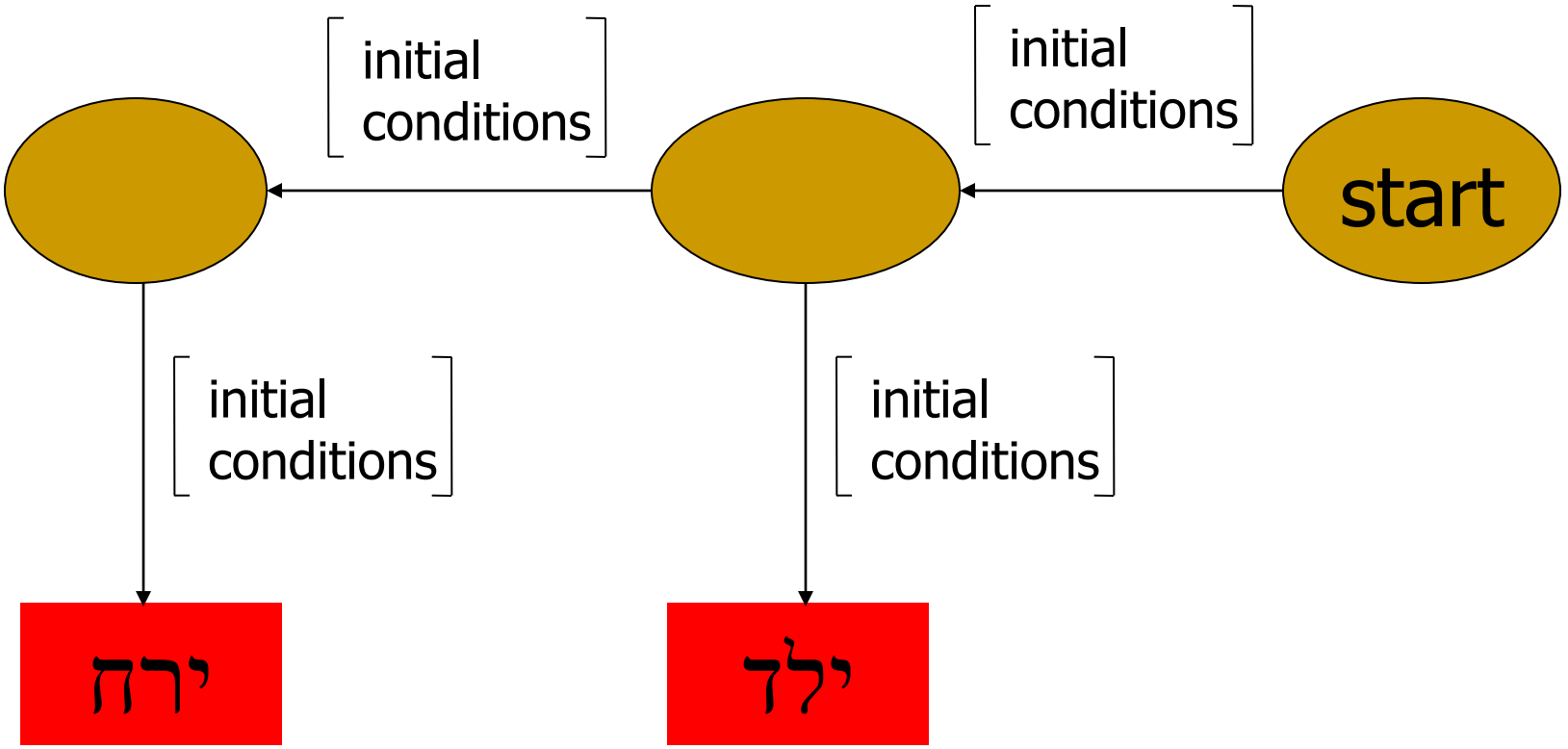


Supervised Parameter Estimation

$$a_{i,j} = \frac{\text{number of transitions from state i to state j}}{\text{number of transitions from state i}}$$

$$b_{i,k} = \frac{\text{number of lexical transitions from state i to symbol k}}{\text{number of transitions from state i}}$$

Unsupervised Parameter Estimation



Unsupervised Parameter Estimation

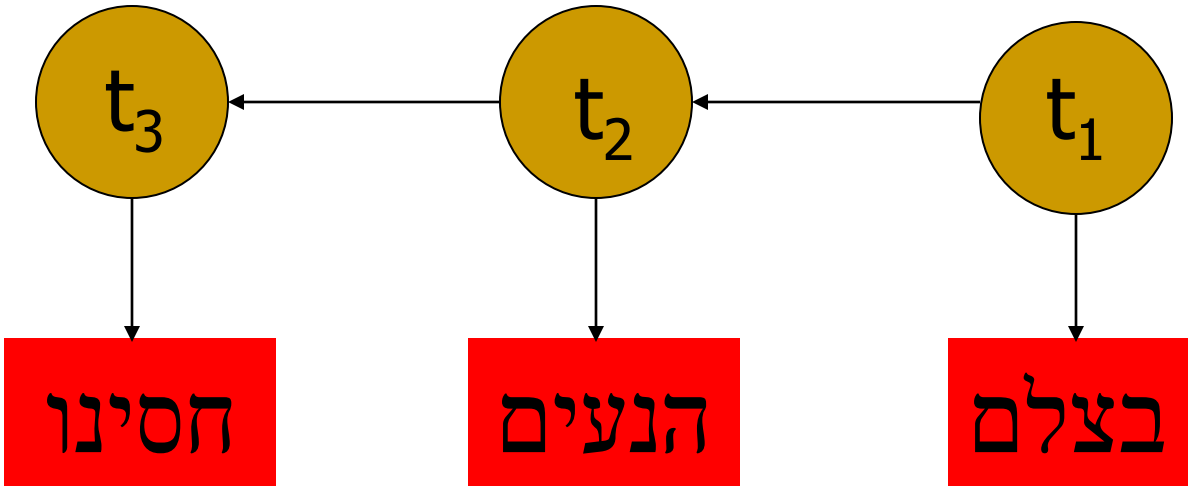
$$a_{i,j} = \frac{\textbf{expected} \text{ number of transitions from state i to state j}}{\textbf{expected} \text{ number of transitions from state i}}$$

$$b_{i,k} = \frac{\textbf{expected} \text{ number of lexical transitions from state i to symbol k}}{\textbf{expected} \text{ number of transitions from state i}}$$

Parameter Estimation

- Baum-Welch algorithm
  - Start with a model with **initial conditions**
  - do
    - Expectation: Calculate the expected number of transitions according to the corpus and the current model.
    - Maximization: Maximize the parameters of the model according to the expected transition number.

Token-based first order HMM

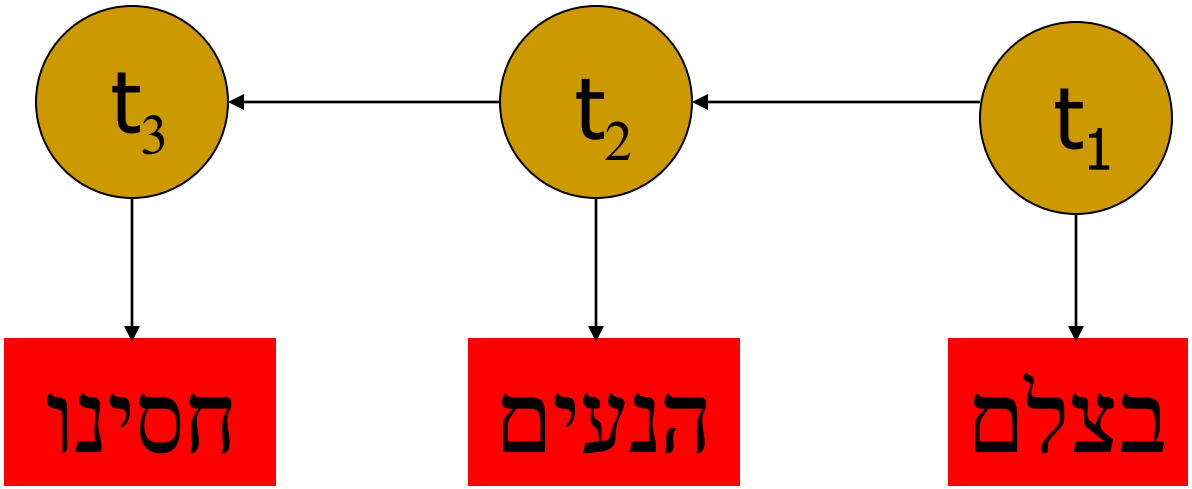


t_1 prep + noun.sing.masc + possessive

t_2 def + adj.sing.masc

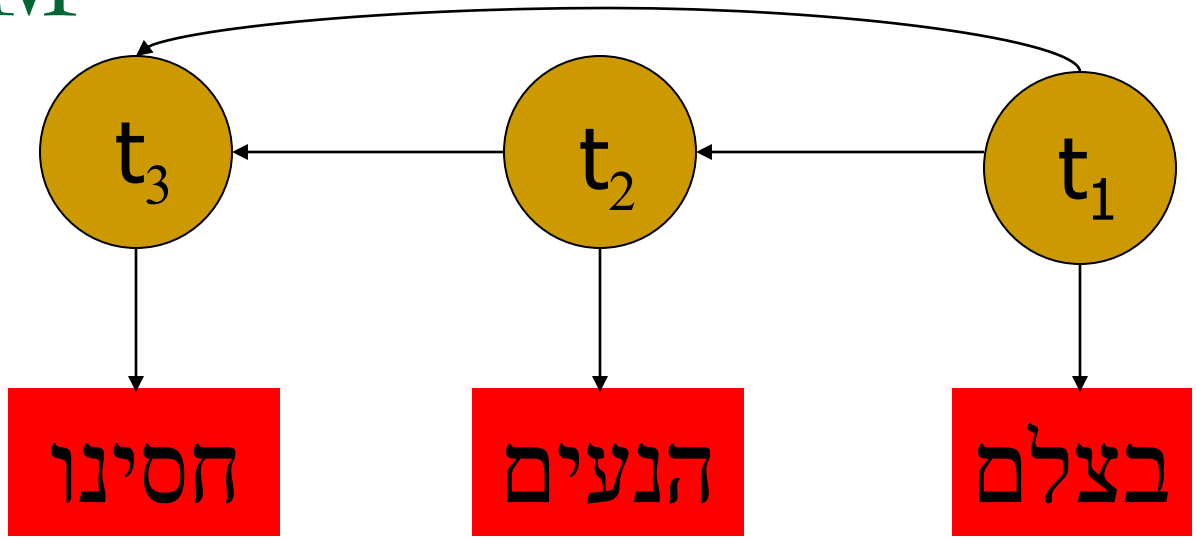
t_3 verb.plural.1.past

Token-based first order HMM



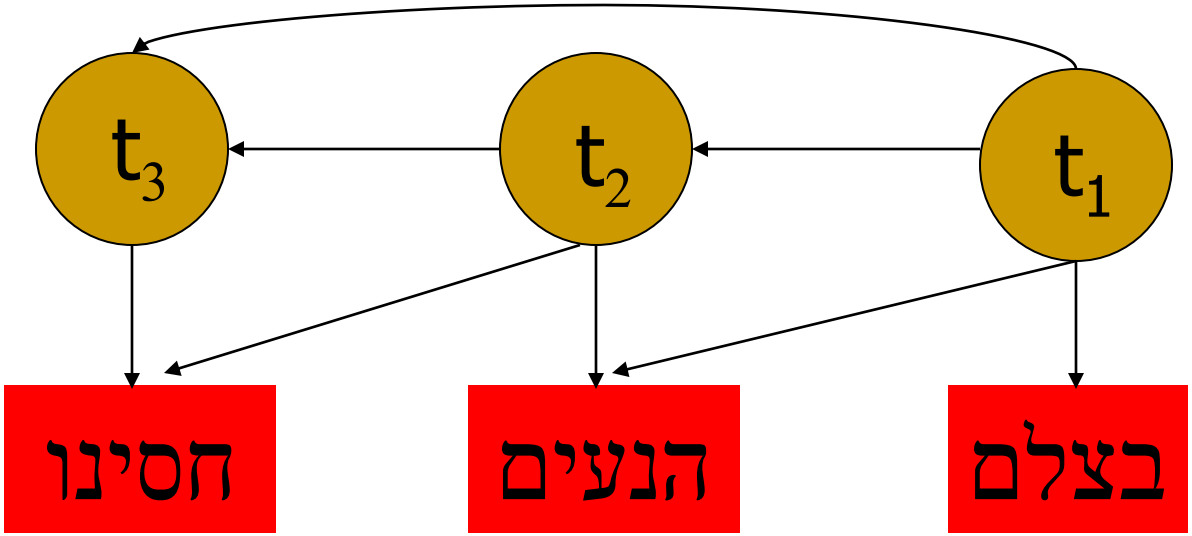
- Tags: English 48, Hebrew 3561
- State Transitions: English 1.8K, Hebrew 855K
- Lexical Transitions: English 57K, Hebrew 3.2M

Token-based partial second order HMM



- Tags: English 48, Hebrew 3561
- State Transitions: English 38K, Hebrew 41M
- Lexical Transitions: English 57K, Hebrew 3.2M

Token-based second order HMM



- Tags: English 48, Hebrew 3561
- State Transitions: English 38K, Hebrew 41M
- Lexical Transitions: English 300K, Hebrew 40M

Word-based Model

- Computational Considerations
  - Sparse data
  - Complexity (Number of parameters)
- Linguistic Motivation
  - Adequate representation
  - Dynamic nature of the language

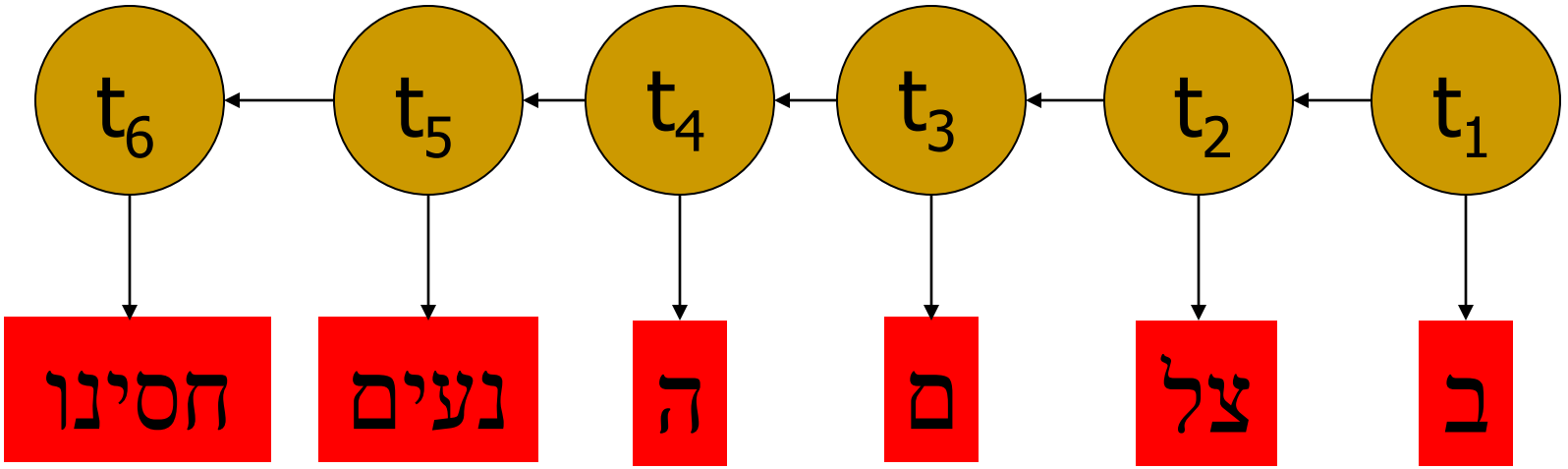
Hebrew Word Definition

- Preposition prefix ב כ ל מ
  - בית, בבית
- Conjunctions ו ש כש לכש בכש
  - בית, ובית שבית
- Definite article ה
  - בית, הבית
  - הלא כל כך נחמדים
- Pronoun suffix
  - בית, ביתו

Hebrew Word Definition

- Prepositions
  - לפני, על ידי
  - בעקבות, מטעם, בגדר, במסגרת
- Adverbs
  - במהירות, בחזרה, במשותף
- Inter token words
  - פה אחד, עורך דין
  - אף על פי, על ידי, לבד מ, פרט ל, בנוסף ל

Word-based first order HMM



t_1 prep

t_2 noun.sing.masc

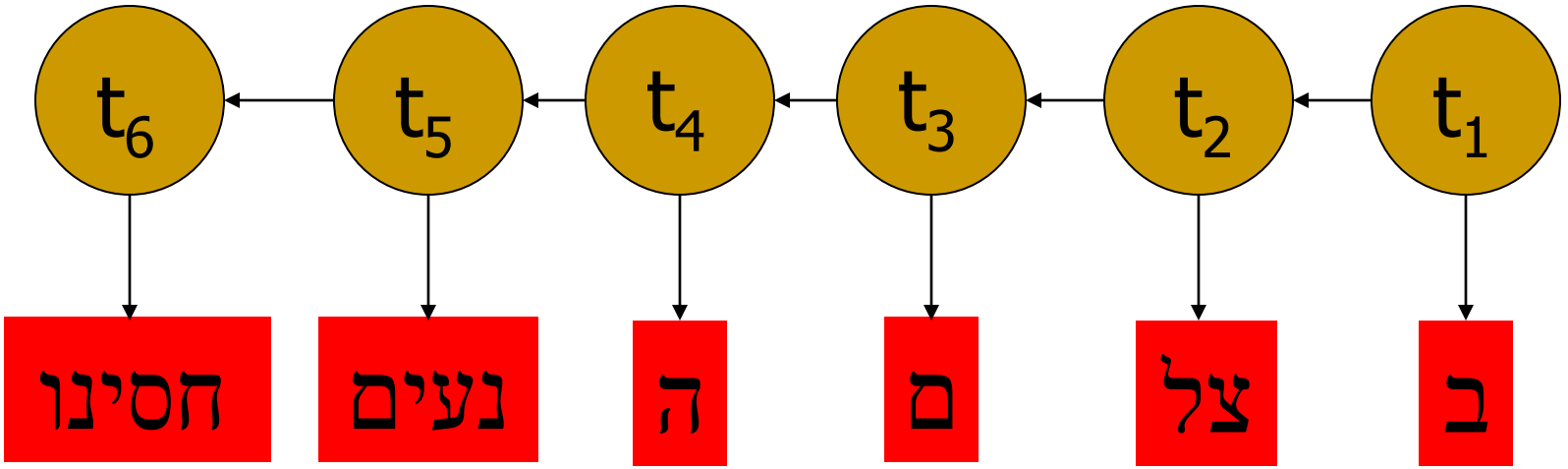
t_3 possessive

t_4 def

t_5 adj.sing.masc

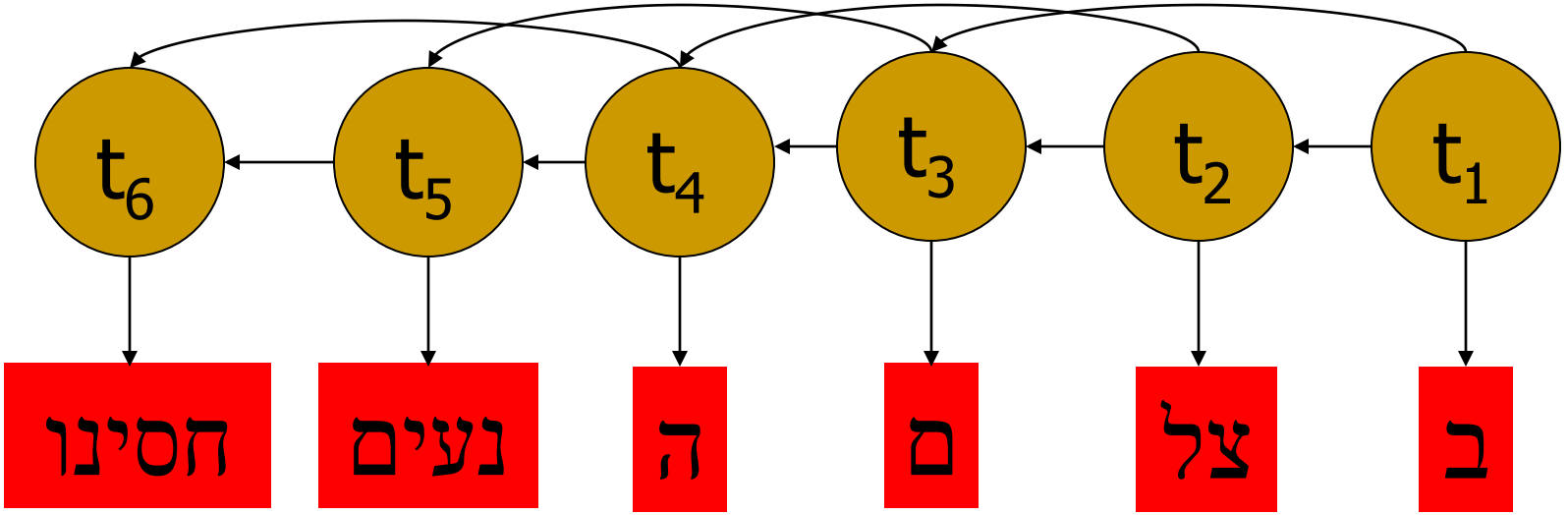
t_6 verb.plural.1.past

Word-based first order HMM



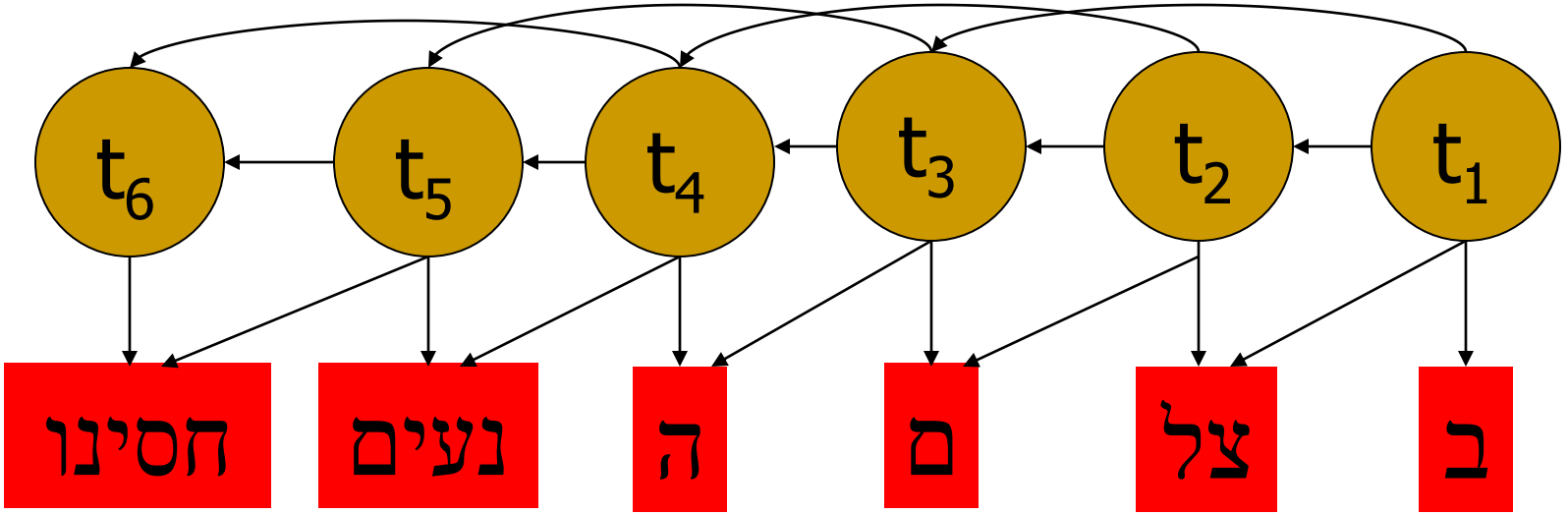
- Tags: English 48, Hebrew 362 (3561)
- State Transitions: English 1.8K, Hebrew 54K (855K)
- Lexical Transitions: English 57K, Hebrew 2.3M (3.2M)

Word-based partial second order HMM



- Tags: English 48, Hebrew 362 (3561)
- State Transitions: English 38K, Hebrew 2.5M (41M)
- Lexical Transitions: English 57K, Hebrew 2.3M (3.2M)

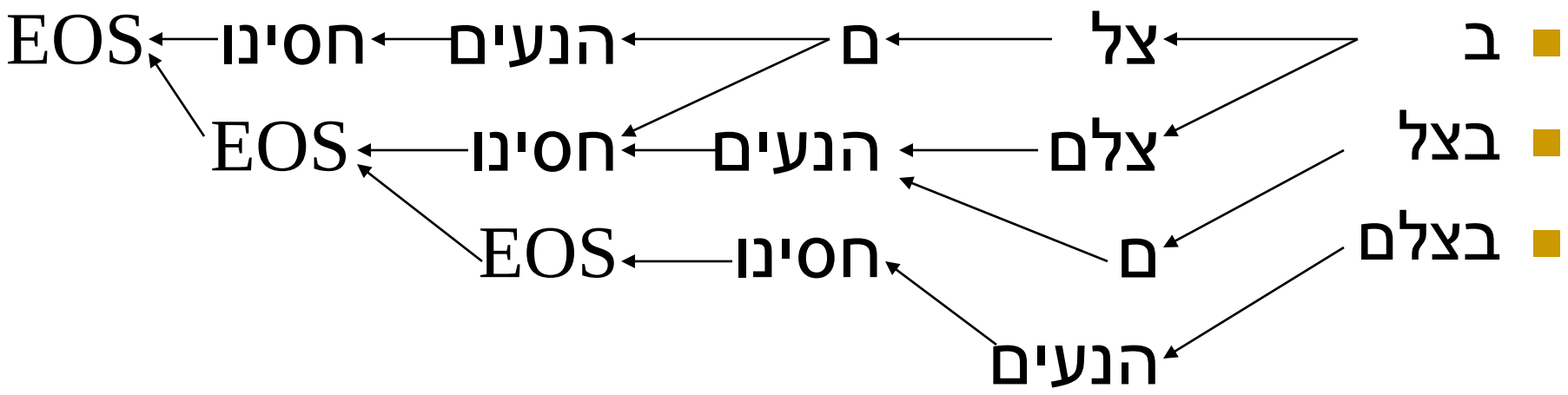
Word-based second order HMM



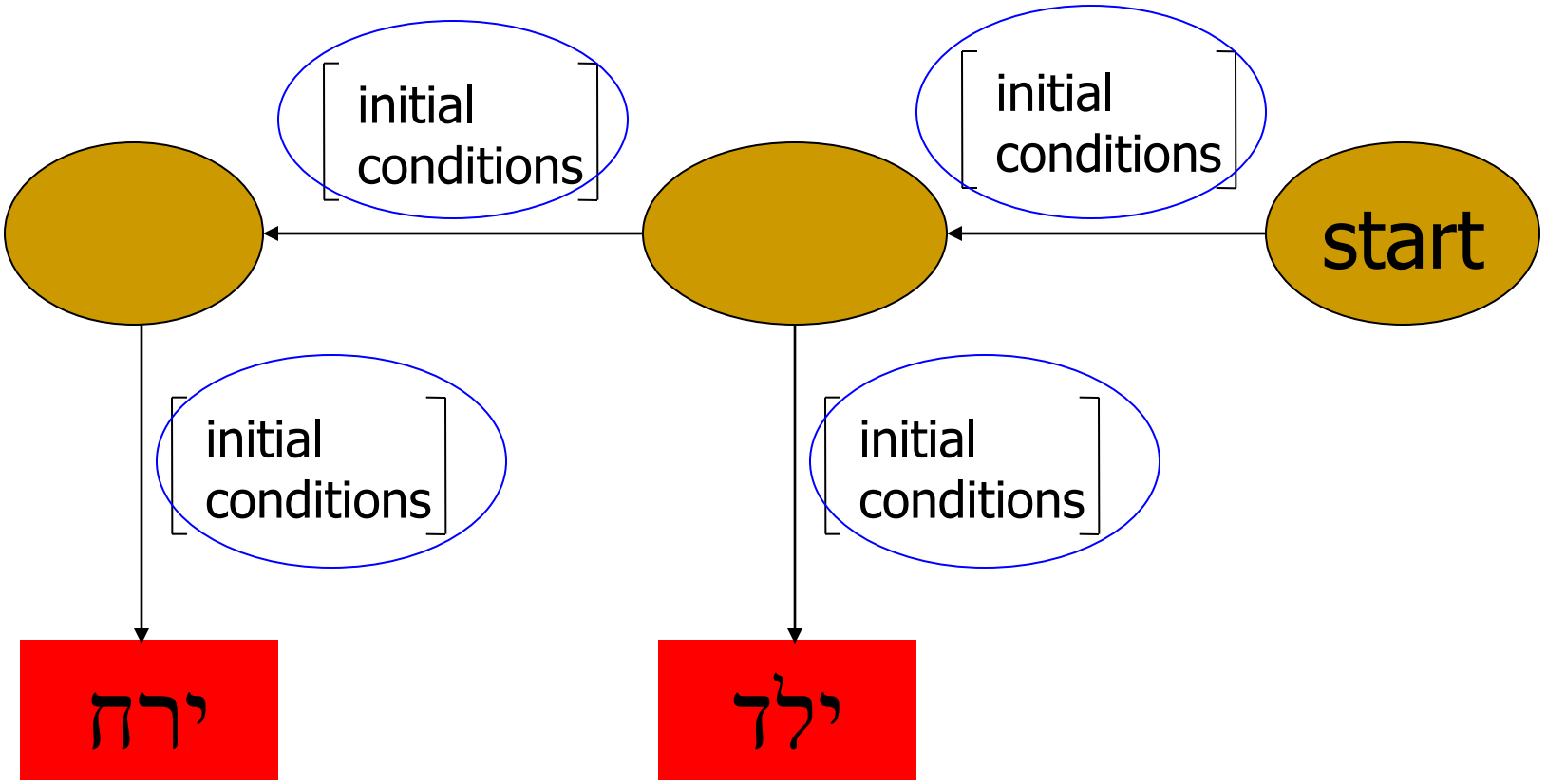
- Tags: English 48, Hebrew 362 (3561)
- State Transitions: English 38K, Hebrew 2.5M (41M)
- Lexical Transitions: English 300K, Hebrew 16M (40M)

Text Representation

- ב צל ם הנעים חסינו
- בצלם הנעים חסינו



Initial Conditions



Initial Condition Types

- Morpho-lexical
  - p(t|w)
  - את: noun, preposition, pronoun
- Syntagmatic
  - $p(t_i|t_{i-2}t_{i-1})$
  - בצלם הנעים חסינו: probability of three consecutive verbs

Morpho-lexical approximations

- Morphology-based [Levinger et al. 95]

את יכולה להעביר לי את את החפירה?

  - pronoun
  - preposition
  - noun

Morphology-based Approximations

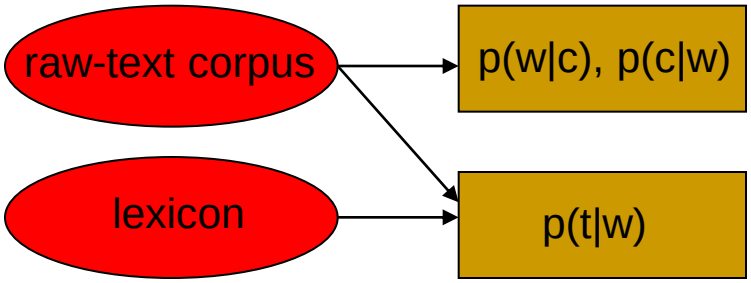
- Similar word sets
  - noun: האת, אתי, אתים
  - pronoun: אתה, אתם, אתן
  - Preposition: /
- The approximation of p(t|w) is based on the frequencies of the similar words of t in the corpus

Linear context-based approximations

- Motivation
  - שלך
    - preposition: שלך שלכם שלכן
    - noun: שלך שלכם שלכן
- Method
  - לכובע שלך שלוש פינות
    - p(preposition | שלוש ,לכובע)
    - p(noun | שלוש ,לכובע)
  - Observed Data
    - p(w|c), P(c|w) לכובע שלך שלוש
    - p(t|w) - similar words algorithm
  - Expectation/Maximization over p(t|w) and p(t|c)

Linear-context Model

- Notation: w –word, c – context of a word, t - tag
- Initial Conditions



- Expectation

$$p(t|c) = \sum_w p(t|w)p(w|c)$$

- Maximization

$$p(t|w) = \sum_c p(t|c)p(c|w)$$

Syntagmatic Approximations

- Syntagmatic Constraints
  - a construct state form cannot be followed by verb, preposition, punctuation, existential, modal, or copula

    שרות התקדמה, פרס את נעמי בע"מ
  - a verb cannot be followed by prep של

    ילד של חלום
  - copula and existential cannot be followed by a verb

    יש ילד בגן, דני הוא ילד חמוד
  - a verb cannot be followed by another verb (with some exceptions)

    ילד חלם חלום

Syntagmatic Approximations

- Initial Transitions
  - A small seed of randomly selected sentences (10K annotated tokens)
  - Tag trigram and bigram counts (ignoring the tag-word annotations) are used for initialization of $p(t|t_{-2},t_{-1})$ distribution

Syntagmatic	Morpho-lexical	Full	POS
Unif	Unif	87.1	91.9
	Morph	88.0	92.1
	Linear	87.5	92.4
	Morph+Linear	88.0	92.8
Pair Const	Unif	87.8	92.0
	Morph	88.1	91.8
	Linear	87.5	92.4
	Morph+Linear	88.0	92.8
Init Trans	Unif	89.4	92.6
	Morph	90.0	93.0
	Linear	89.7	93.0
	Morph+Linear	**90.0**	**93.1**

Baseline: Token-based, EM learning

Initial Condition

Hebrew Tagging - Analysis

- EM learning
  - Unsupervised HMM learning on word model
  - EM is very effective for Hebrew: error reduction of 65% over uniform initial conditions
- Morphology-based Initial Conditions:
  - Error reduction of 7.7% upon uniform distribution
- Syntagmatic Initial Conditions:
  - Pair constraints only have minor impact.
  - Initial transition frequencies: error reduction of 16.5% for full analysis; 12.5% for POS tagging

Outline

- Objectives
- Topic Analysis with LDA
- Obtaining Precise Morphology in Hebrew
- Combining LDA and Morphological Analysis
- Using Topic Models for Search
- Evaluating Topic Models
- Next Steps

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.

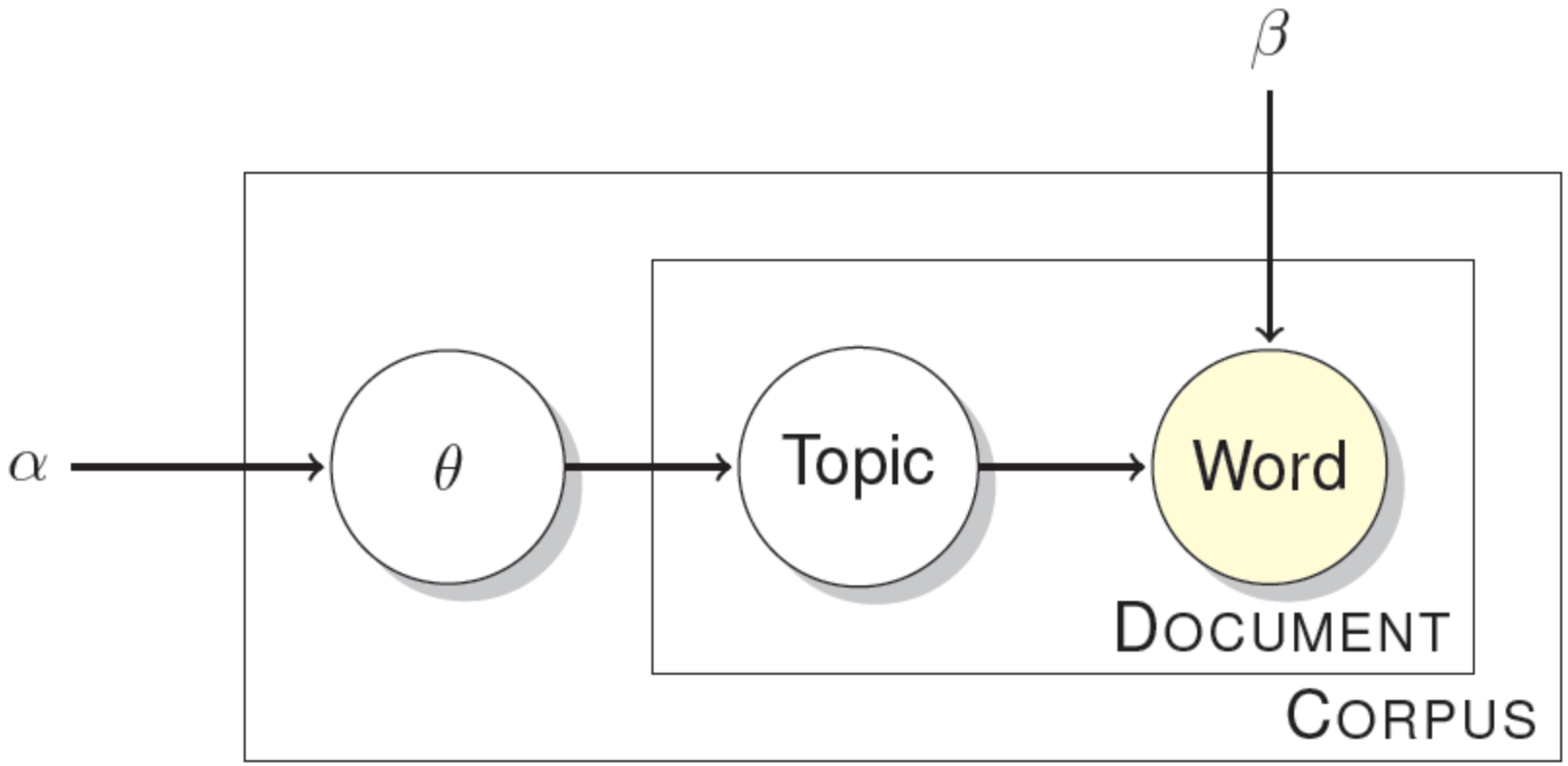
→What is the best method to combine LDA and Morphological analysis?

Combining LDA and Morphology

3 options:

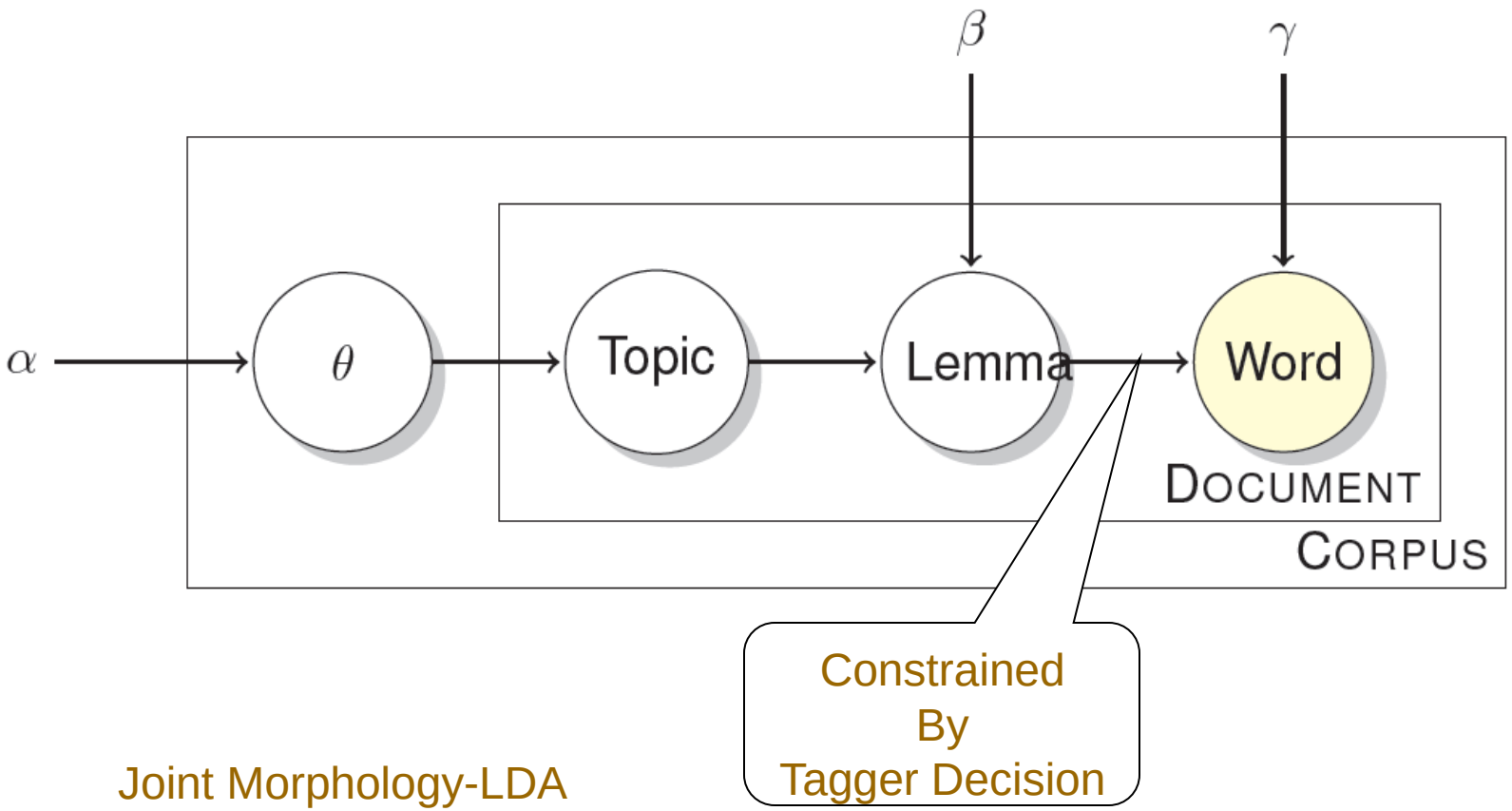
- **Ignore morphology** – token-based LDA
- **Pipeline –** resolve morphological ambiguities, then learn LDA.
- **Joint** – learn LDA on distributions of possible morphological analyses

Joint LDA-Morphology Learning



Standard token-based LDA

Joint LDA-Morphology Learning



Joint Morphology-LDA

Outline

- Objectives
- Topic Analysis with LDA
- Obtaining Precise Morphology in Hebrew
- Combining LDA and Morphological Analysis
- Using Topic Models for Search
- Evaluating Topic Models
- Next Steps

Searching with Topics

Combine search and browse:

- Word →Topics **Search**
- Topic →Documents **Browse**

Word → Topics **Disambiguate**

Cluster unseen documents based on topics

Mishne Torah Topics

- 100 topics (K parameter)
- Each word covers many variations

0 עֲשָׂר דָּה סַלַּע מעה כֶּסֶף שֵׁנִי יְרוּשָׁלַיִם פְּרִי שוה ח׳ ל אָמַר מע"ש ח׳ מֹשֶׁ יָצָא הוֹסִיף נָתַן פרט דינר זָהָב חלל

1 כּוֹתֵב עֲבוֹדָה הַנָּאָה עָבַד עוֹבֵד אָסַר צוּרָה עָשָׂה אֶבֶן לָהּ זוֹ דֶּרֶךְ בְּעַל יִשְׂרָאֵל עוֹלָם נסקל בה נֶאֱסַר לָקָה בָּנָה

2 כְּתֻבָּה בַּעַל לָהּ אִישׁ נָכְסֵי נָתַן הוֹצִיא מָזוֹן דִּין אותה רָצָה מֵת ל׳ א נָטַל יֵשׁ בַּיִת עִקָּר בֵּין אַחַר מָכַר

3 מַעֲלָה מֵאָה הָיָה חֵלֶק שֵׁשׁ הוא עֶשְׂרִים שְׁלֹשׁ לָשׁ יהיה עֲשָׂר מָנָה יָדַע אַחַת הוֹסִיף אַרְבָּעִים אֶמְצַע כ׳ ל שֵׁנִי חֲמֵשׁ

4 עֲלָה הֵבִיא קָרְבָּן חַטָּאָה שָׁלֵם דם הִפְרִישׁ נָפַל אָשָׁם מעה עוף מֵת זוֹ נִדְבָּה קָדַשׁ אַחַר מֵאָה שְׁתַּיִם אָמַר קָרֵב

5 עַד אָמַר הֵעִיד נֶאֱמָן פלוני דִּין בָּא דָּבָר פָּנִים יָדַע רָאָה מֵת פֶּה זוֹ אַחַר עֵדוּת שָׁנָה הָיָה אֶחָד כָּךְ

6 בְּעַל הֶחְזִיר גָּזַל לוֹ נָטַל שבח גָּזְלָן מָצָא יָד לָקַח חַיָּב סִימָן כְּלִי גְּזֵלָה נָתַן מָלֵךְ ממנו חזקה מִן רְשׁוּת

7 חָצֵר בֵּן מבוי עֶרֶב בַּיִת רְשׁוּת אֶחָד פתח חָלוֹן אִבָּה רָצָה חבירו יָכ׳ ל כ׳ חֵל עליו עֵרוּב עִם צָרִיךְ אִישׁ בְּעַל

8 טמא מֵת עֲמָאָה טהור נָגַע הוא אדם כְּלִי טִמֵּא מַגָּע א׳ הֶל אִבָּה זָב מַשָּׂא בֶּגֶד עֶצֶם בֵּין משכב כ׳ ל בו

9 כַּד הוֹצִיא שיעור חַיָּב תּוֹלָדָה שום פָּה כגרוגרת חַי נָתַן קָטָן גַּב חֻבָּל בְּהֵמָה עֵץ בְּשַׂל עוֹר חֵץ בֵּיצָה שְׁתַּיִם

10 תְּשׁוּבָה עֲשָׂה עֲבֵרָה רִשָּׁע חבירו חיטא יָד חָזַר קִבֵּל יִשְׂרָאֵל מָמוֹן בַּעַל בֵּין סָמַךְ לוֹ עון חוֹטֵא חֲשׁוּד מִיתָה גָּמוּר

11 לחם הִקְרִיב קָרְבָּן כִּבֵּשׂ עֶלָה הוא בָּא כ׳ ל מִן תּוֹדָה עִשָּׂר שָׁלֵם מִנְחָה קודם עִם נֶאֱכַל חַטָּאָה עָבַר מוֹסַף הִסְכִּים

12 שְׂעָר חוֹב לוֹ כָּתַב גָּבָה אָמַר הִלְוָה פָּרַע נָתַן המלוה יָד מִן בַּעַל לָוָה עֵרֶב דִּין יֵשׁ נָכְסֵי מָנָה עֵרָךְ

13 מַיִם נָפַל רְצוֹן תּוֹךְ סאה את אֵינוֹ הֶכְשֵׁר יָרַד אָבָה בו המקוה מקוה משק תְּלוּשׁ הוֹגָה הָיָה עליהן מהן גֶּשֶׁם

14 יוֹם דם רָאָה נָד טמא זָב טהור אִישׁ זוֹ אַחַר עֵת בָּדַק לָהּ תְּלָה כְּתֻם זיבה מָצָא שִׁמֵּשׁ נִמְצָא וָסֶת

15 עֻמְאָה הָיָה טמא עִפָּה בַּיִת טהור הוא תּוֹךְ גַּב בֵּין כְּלִי פתח א׳ הֶל כ׳ ל תַּחַת את נָגַד הַר פָּנִים אֶרֶץ

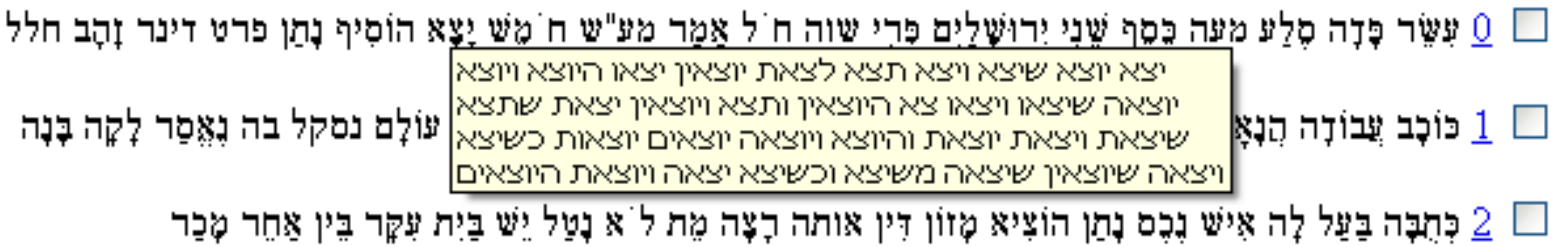
16 עֲשַׂר פארק הֶלְכָה יָצָא אַרְבָּעָה קְרָא חֲמֵשׁ מָנָה נִשְׁאַר עֲשִׂירִי תְּשָׁעָה דִּיר שָׁשׁ סִפֵּר פתח שם מְגִלָּה סְלִיק עֶשְׂרִים תּוֹךְ

17 שְׁבִיעִי פְּרִי אֶרֶץ שָׁכַח אילן ע׳ מֹר נָטַע שְׁדֵה עָרֵל שָׁנָה אילן עָשָׂה מוֹצָא ריבע הַתַּר עֲבוֹדָה זַיִת קָצַר ר"ה עִקָּר

Searching with Topics

Mishne Torah Topics

- Each word covers many variations



Query Disambiguation

- Various topics for 1 word

שור

Outline

- Objectives
- Topic Analysis with LDA
- Obtaining Precise Morphology in Hebrew
- Combining LDA and Morphological Analysis
- Using Topic Models for Search
- Evaluating Topic Models
- Next Steps

How Good are Discovered Topics?

- Difficult to evaluate LDA topics
  - Many parameters (many words, many topics)
  - Each run gives slightly different results
  - How to compare topic models?
- Task-based evaluation
  - Use topics for Summarization
- Ontology alignment evaluation
  - Compare topics with existing ontology
- Data-oriented evaluation

Ontology Alignment

- Mishne Torah has existing structure:
  - Hierarchy of Book/Section/Chapter
  - Excellent indexes exist
  - Compare indexes with existing ontology.

- We find excellent alignment topic/Book-Section
  - Some topics are “cross-concerns” (witnesses – topic 5)

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

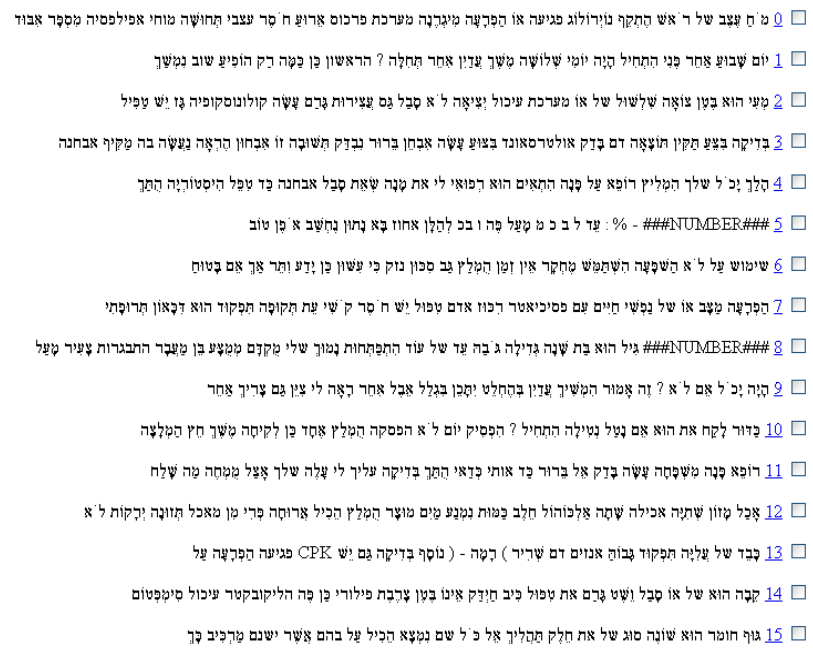
Submit

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

Other Domain: More Noise

- Applied LDA to InfoMed corpus (www.infomed.co.il) corpus of “popular medicine” articles.
- Need a different evaluation method



Data-oriented Evaluation

- Method derived from our previous work on ontology evaluation
- How well does an automatically constructed partition of entities into classes represent reality?
- How can a classification be improved via merge or split operations?
- **Example:** movies genres, taken from IMDB: *drama, comedy, war, sport, action…*

Text Classifier to Evaluate Classes

- **Idea**: use a set of texts from a given domain as a proxy of the “reality” represented by the ontology.
- ***Hypothesis:***

  If the ontology indicates that some movies are "clustered" according to one of the dimensions

  Then documents associated to these movies should also be found to be associated by a text-classification engine that has been trained on the classification induced by the ontology
- ***Procedure:***

  Train a classifier for each class we want to evaluate

  If a classifier can decide with good accuracy whether a given text belongs to a class – the class is well-defined.

  In the movies domain, reviews are used as representative

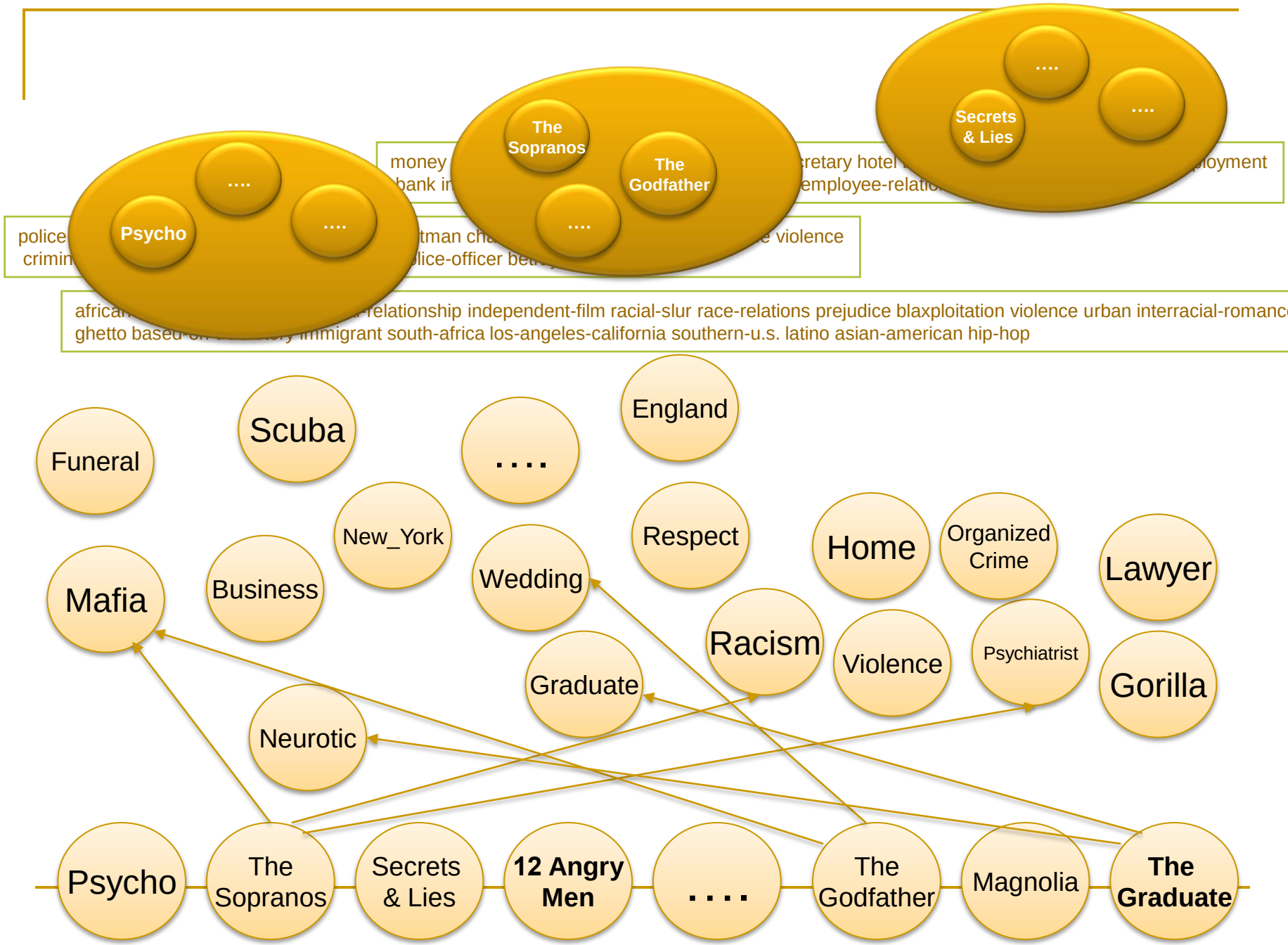
Family
Action
Romance
Comedy
M O V I E
Bah.
BOOM!
R E V I E W S
Action
Comedy
Crime
Drama
Movie Genres
War

Noisy Classes

- What can be done if the classes are very noisy?
- Example: Keywords in IMDB
  - Keyword lists are open to additions/deletions by users (yet moderated).
  - Examples of Keywords: *Mafia Business New_York Wedding Respect Home Organized_Crime Lawyer Violence …* *(and the film?)*
  - Too few movies are associated with any single keyword
  - Movies associated with a given keyword are not necessarily related
- ***Could be the same for LDA topics in noisy domains***

Cluster Keywords using LDA

- Apply LDA on the keywords
- Divide movies into classes according to LDA distributions of the movies reviews
- Construct classifiers to evaluate the quality of the LDA classes
- Example of the top keywords of a “good” class: *england inheritance london-england based-on-novel mansion london maid class-differences servant 19th-century period-piece butler orphan estate aunt uncle heir victorian-era love marriage*


....
Psycho
....
The Sopranos
The Godfather
....
....
Secrets & Lies
....
money
bank
police
violence
relationship independent-film racial-slur race-relations prejudice blaxploitation violence urban interracial-romance
immigrant south-africa los-angeles-california southern-u.s. latino asian-american hip-hop
Funeral
Scuba
....
England
New_York
Respect
Home
Organized Crime
Lawyer
Mafia
Business
Wedding
Racism
Violence
Psychiatrist
Gorilla
Graduate
Neurotic
Psycho
The Sopranos
Secrets & Lies
12 Angry Men
....
The Godfather
Magnolia
The Graduate

Conclusions

- Morphological analysis is a critical pre-processing step for most text mining applications in Hebrew
- We obtain accuracy in the 90-95% range on full segmentation/POS/morphological analysis – robust on unknown words.
- Enables effective LDA Topic Analysis in Hebrew
- More work needed on Topic Analysis Evaluation