Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling

Daichi Mochihashi

NTT Communication Science Laboratories, Japan

daichi@cslab.kecl.ntt.co.jp

ACL-IJCNLP 2009

Aug 3, 2009

Word segmentation: string→words

山花 貞夫・新 民連 会長 は 十六 日 の 記者 会見 で 、村山 富市 首相 ら 社会党 執行 部 と さきがけ が 連携 強化 を めざ した 問題 に ついて「 私 たち の 行動 が 新しい 政界 の 動き を 作った と いえる 。統一 会派 を 超え て 将来 の 日本 の …

今后 一段 时期 ， 不但 居民 会 更 多 地 选择 国债 ， 而且 一些 金融 机构 在 准备金 利率 调 低 后 ， 出于 安全性 方面 的 考虑 ， 也 会 将 部分 资金 用来 购买 国债 。 …

- Crucial for languages like Japanese, Chinese, Arabic, …
  - Useful for complex words in German, Finnish, …
- Many research→Mostly supervised

What’s wrong?

“Ungrammatical”

香港の現地のみんなが翔子翔子って大歓迎してくれとう!!!!アワ
わわわわ((°°дд°°
みんなのおかげでライブもギガントだったお(´;ω;`)まりがとう

Interjection

Word not in a dictionary

Face mark

Extraordinary writing for “thank you”

- Colloquial texts, blogs, classics, *unknown language*,…
  - There are no “correct” supervised segmentations
- New words are constantly introduced into language

This research..

"*The Tale of Genji*", written 1000 years ago,
Very difficult even for native Japanese!

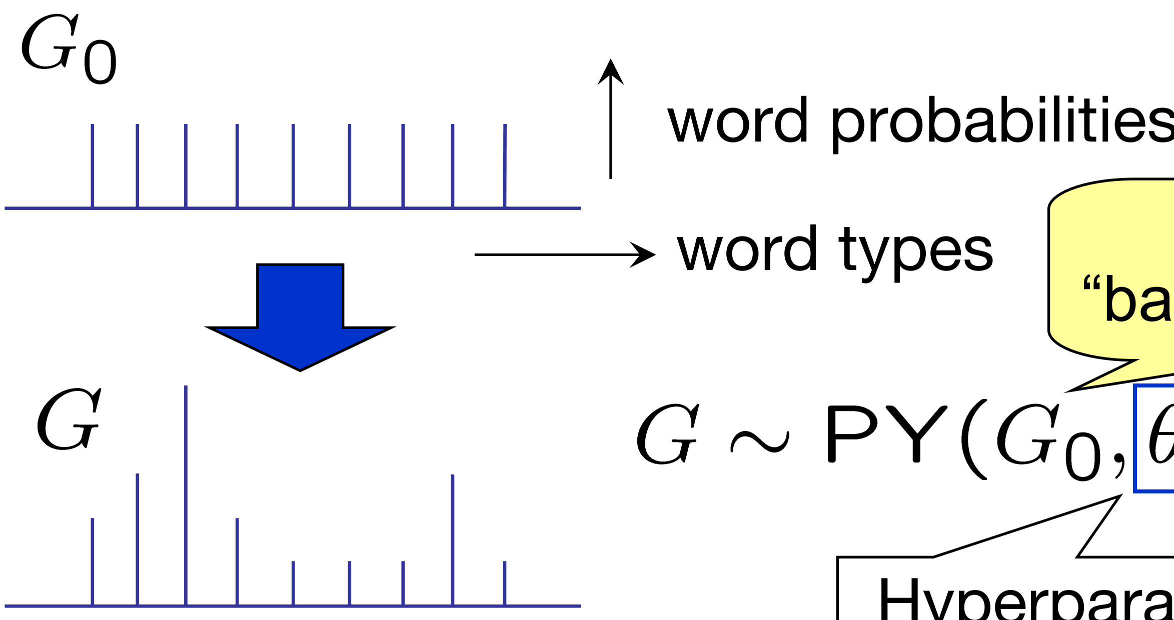
花の蔭にはなほ休らはまほしきにや、この御光を見たてまつるあたりは、ほどほどにつけて、わがかなしと思ふむすめを仕うまつらせばやと願ひ、もしは口惜しからずと思ふ妹など持たる人は、いやしきにても、なほこの御あたりにさぶらはせんと思ひよらぬ…

花 の 蔭 に は なほ 休らは まほし き に や 、こ の 御 光 を 見 たてまつる あたり は 、ほどほどにつけて 、わが かなし と 思ふ むすめ を 仕うまつ ら せ ば や と 願ひ 、もし は 口惜し から…

- Completely unsupervised word induction from a Bayesian perspective
  - Directly optimizes the performance of Kneser-Ney LM
- Extends: Goldwater+(2006), Xu+(2008), …
  - Efficient forward-backward+MCMC & word model

Pitman-Yor n-gram model

- The Pitman-Yor (=Poisson-Dirichlet) process:
  - Draw distribution from distribution
  - Extension of Dirichlet process (w/ frequency discount)

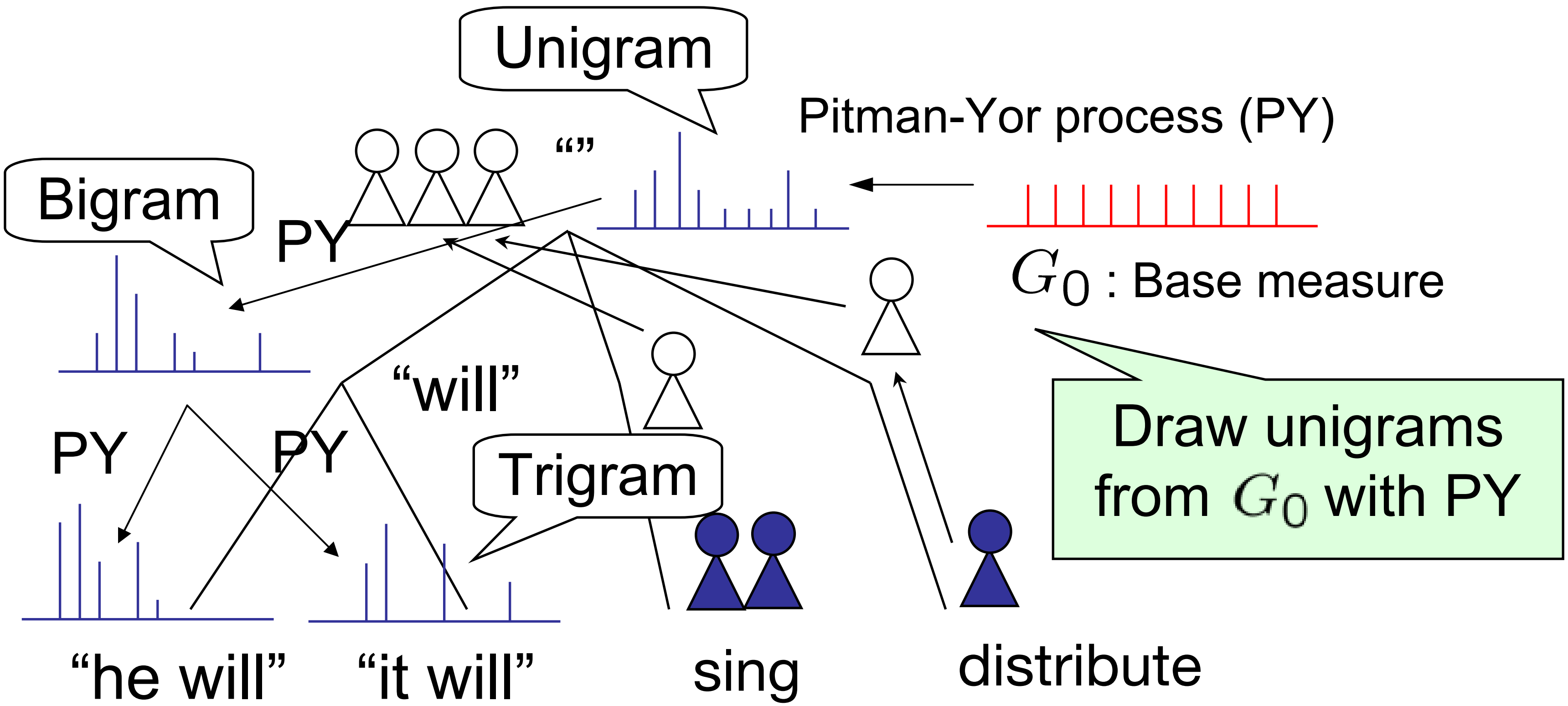


$$G \sim \mathrm{PY}(G_0, \boxed{\theta, d})$$

Called "base measure"

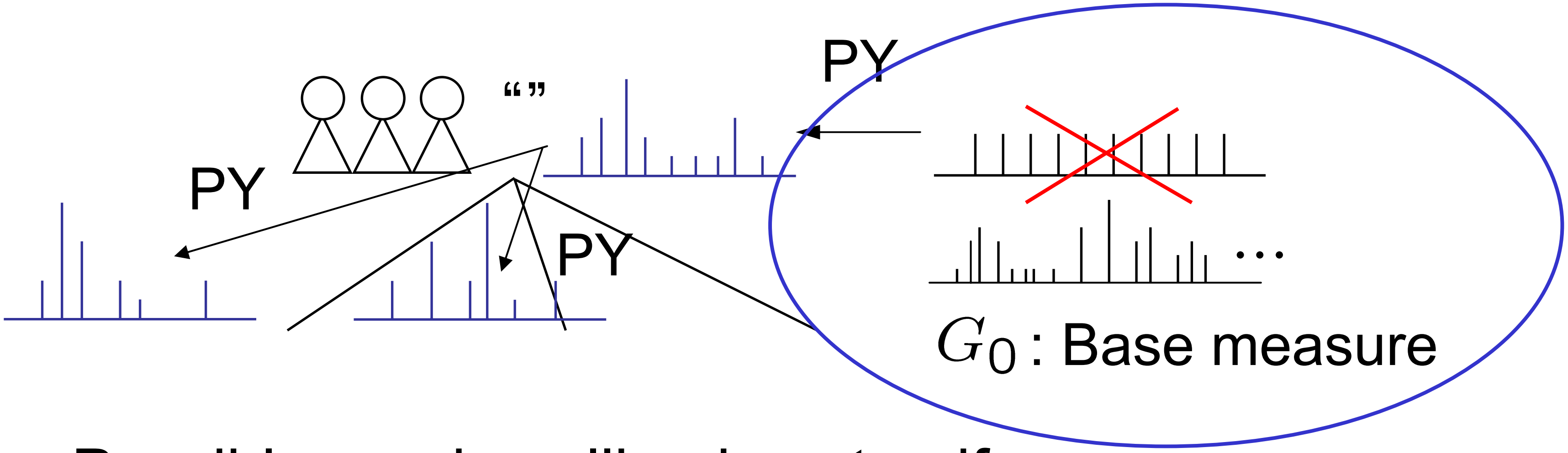
Hyperparameters (can be learned)

Hierarchical Pitman-Yor n-gram



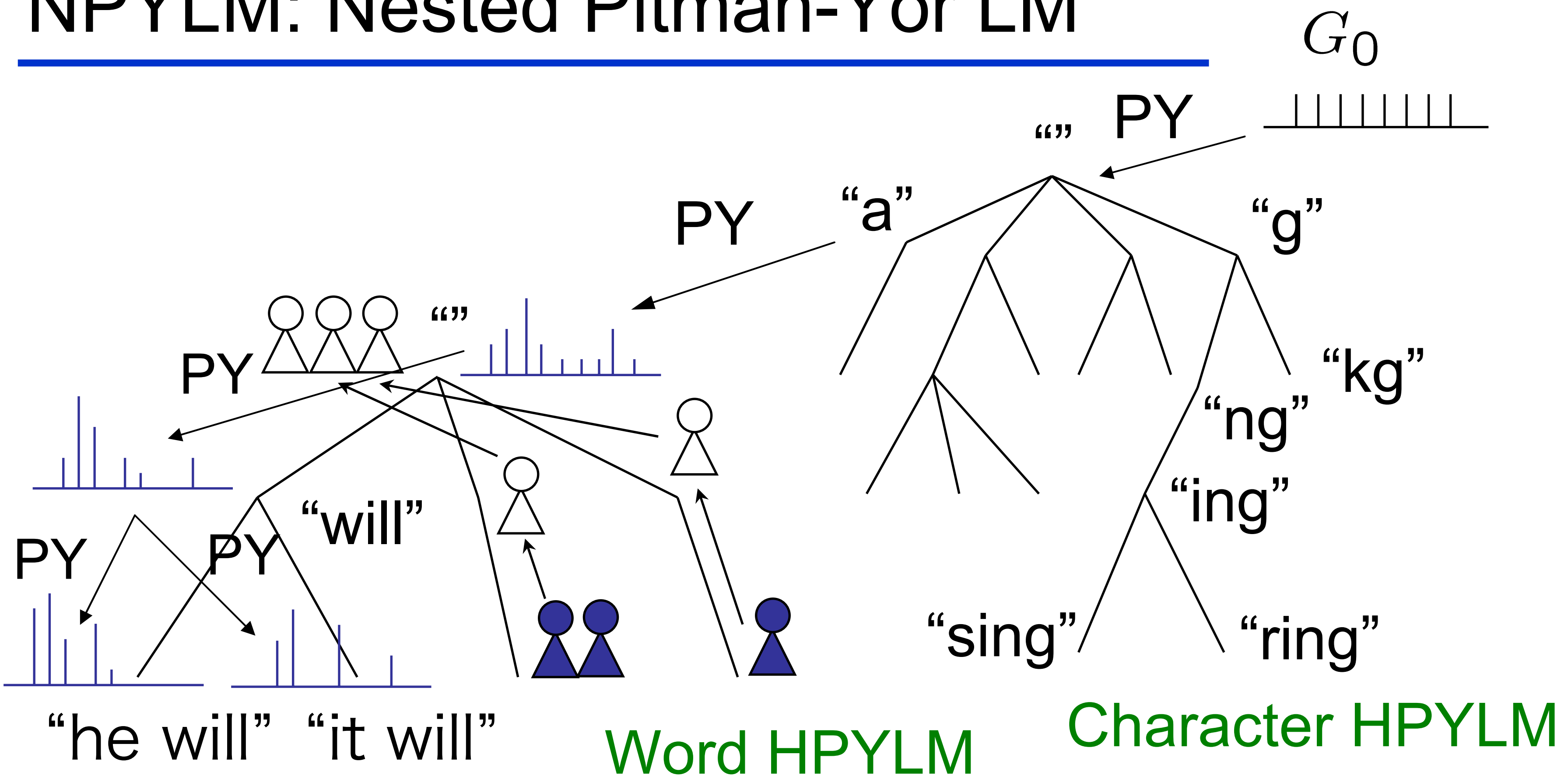
- Kneser-Ney smoothing is an approximation of hierarchical Pitman-Yor process (Teh, ACL 2006)
  - HPYLM = “Bayesian Kneser-Ney n-gram”

Problem: Word spelling



- Possible word spelling is not uniform
  - Likely: “will”, “language”, “hierarchically”, …
  - Unlikely: “illbe”, “nguag”, “ierarchi”, …
- Replace the base measure using character information
  →Character HPYLM!

NPYLM: Nested Pitman-Yor LM



- Character n-gram embedded in the base measure of Word n-gram
  - i.e. hierarchical Markov model
  - Poisson word length correction (see the paper)

Inference and Learning

- Simply maximize the probability of strings
  - i.e. minimize the perplexity per character of LM
- X : Set of strings $s_1, s_2, \cdots, s_N$
  Z : Set of hidden word segmentation indicators
  $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N$

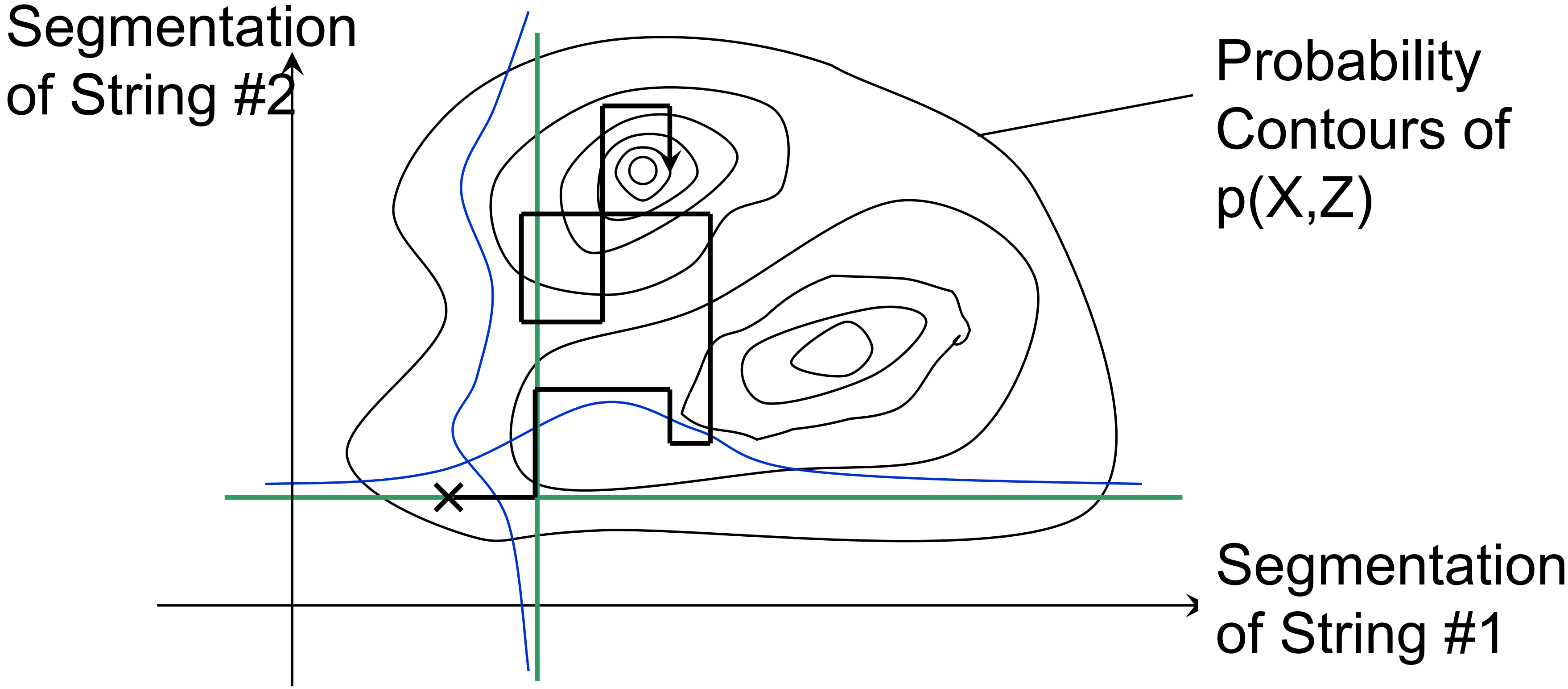
$$p(X) = \prod_n p(s_n)$$

$$p(s_n) = \sum_{\mathbf{z}_n} p(s_n, \mathbf{z}_n)$$

Hidden word segmentation of string s_n

  - Notice: *Exponential possibilities* of segmentations!

Blocked Gibbs Sampling



- Sample word segmentation block-wise for each sentence (string)
  - High correlations within a sentence

Blocked Gibbs Sampling (2)

- Iteratively improve word segmentations: words(s) of s

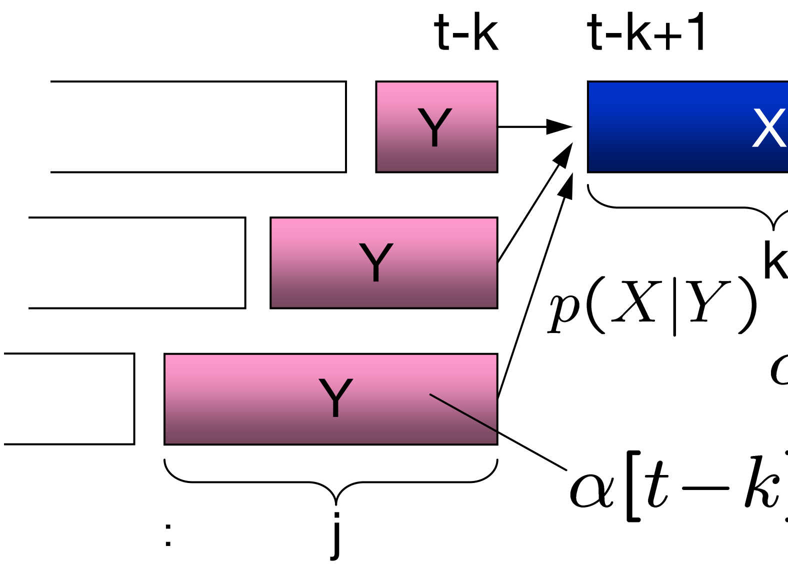
  0. For $s = s_1 \cdots s_N$ do
     parse_trivial(s,Θ).

  Whole string is a single "word"

  1. For j = 1..M do
     For s = *randperm*($s_1 \cdots s_N$) do
       Remove words(s) from NPYLM Θ
       Sample words(s) $\sim p(w|s,\Theta)$
       Add words(s) to NPYLM Θ
     done
     Sample all hyperparameters of Θ
   done

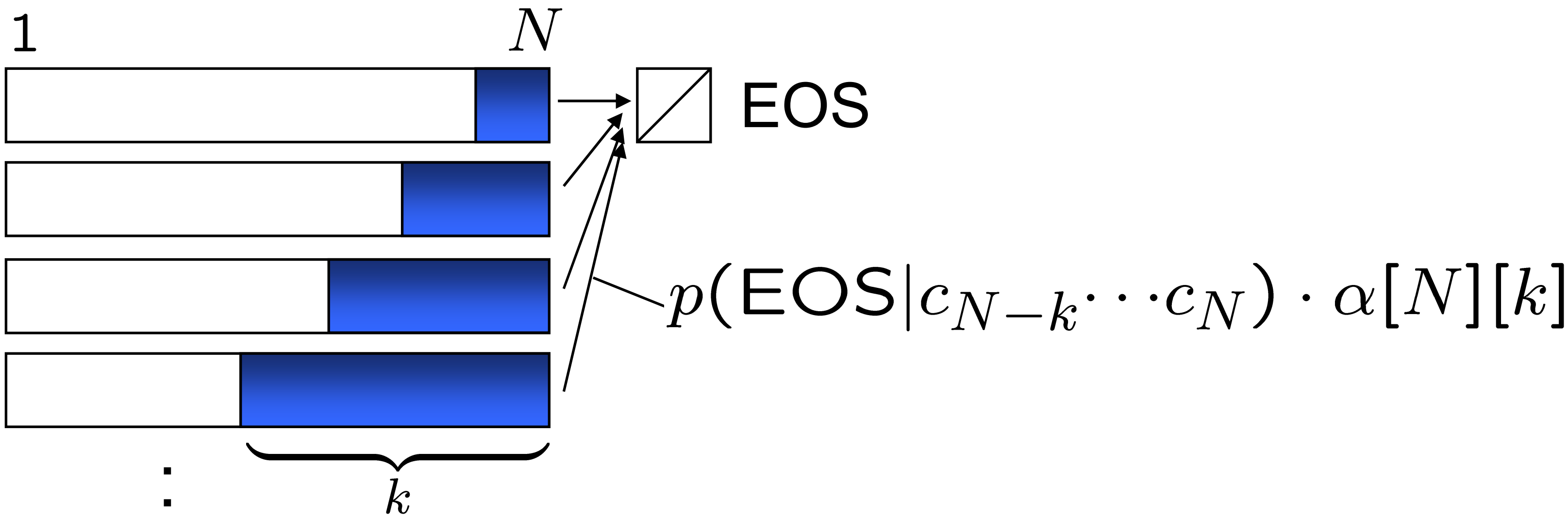
Sampling through Dynamic Programming

- Forward filtering, Backward sampling (Scott 2002)
- $\alpha[t][k]$: inside probability of substring $c_1 c_2 \cdots c_t$ with the last k characters constituting a word
  – Recursively marginalize segments before the last k



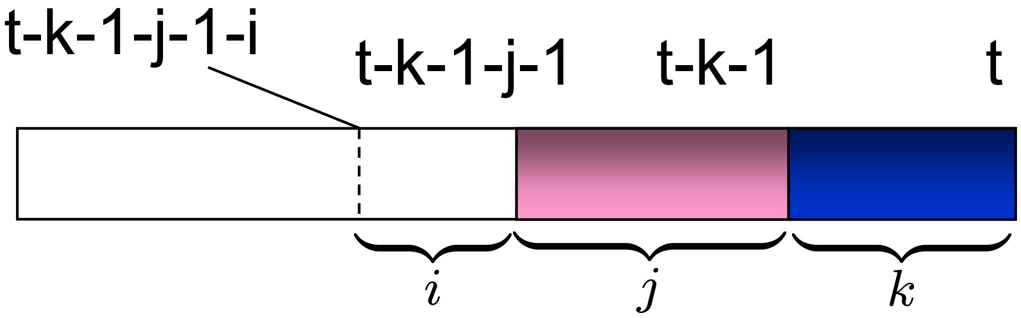
$$\alpha[t][k] = \sum_j p(X|Y)\,\alpha[t-k][j]$$

Sampling through Dynamic Programming (2)



- $\alpha[N][k]$ = probability of the entire string $c_1\cdots c_N$ with the last k characters constituting a word
  - Sample k with probability to end with EOS
- Now the final word is $c_{N-k}\cdots c_N$: use $\alpha[N-k-1][k']$ to determine the previous word, and repeat

The Case of Trigrams



- In case of trigrams: use $\alpha[t][k][j]$ as an inside probability
  - $\alpha[t][k][j]$ = probability of substring with the final k chars and the further j chars before it being words
  - Recurse using $\alpha[t-k-1][j][i]$ $(i = 0 \cdots L)$
- >Trigrams? Practically not so necessary, but use Particle MCMC (Doucet+ 2009 to appear) if you wish

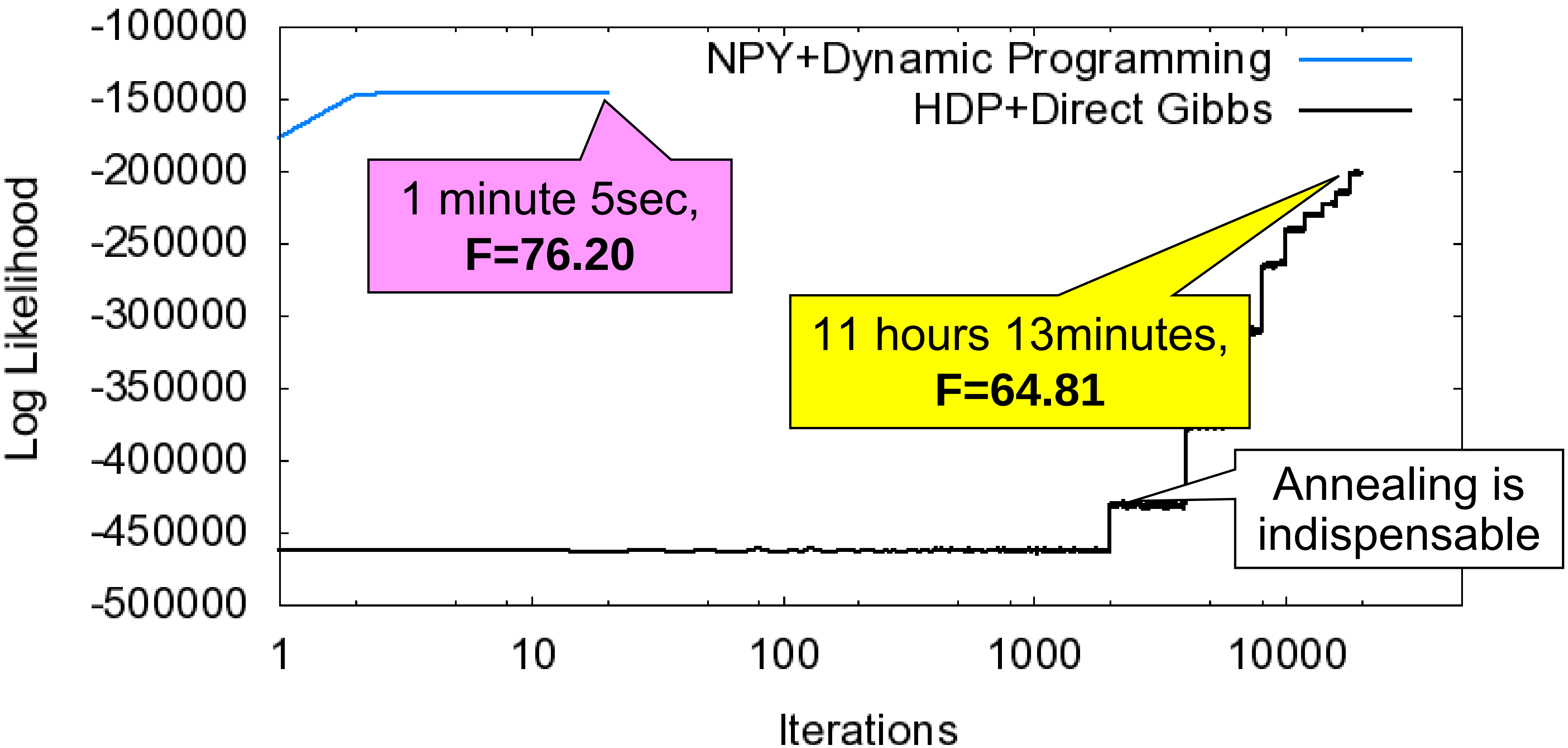
English Phonetic Transcripts

- Comparison with HDP bigram (w/o character model) in Goldwater+ (ACL 2006)
- CHILDES English phonetic transcripts
  - Recover “WAtsDIs”→”WAts DIs” (What’s this)
  - Johnson+(2009), Liang(2009) use the same data

Model	P	R	F	LP	LR	LF
NPY(3)	74.8	75.2	75.0	47.8	**59.7**	53.1
NPY(2)	74.8	**76.7**	**75.7**	57.3	56.6	57.0
HDP(2)	**75.2**	69.6	72.3	**63.5**	55.2	**59.1**

  - Very small data: 9,790 sentences, 9.8 chars/sentence

Convergence & Computational time



- NPYLM is very efficient & accurate! (600x faster here)

Chinese and Japanese

Perplexity per character

Model	MSR	CITYU	Kyoto
NPY(2)	0.802 (51.9)	**0.824** (**126.5**)	0.621 (23.1)
NPY(3)	**0.807** (**48.8**)	0.817 (128.3)	**0.666** (**20.6**)
NPY(+)	0.804 (**38.8**)	0.823 (**126.0**)	**0.682** (**19.1**)
ZK08	0.667 (—)	0.692 (—)	—

- MSR&CITYU: SIGHAN Bakeoff 2005, Chinese
- Kyoto: Kyoto Corpus, Japanese
- ZK08: Best result in Zhao&Kit (IJCNLP 2008)

Note: Japanese subjective quality is much higher
(proper nouns combined, suffixes segmented, etc..)

Arabic

- Arabic Gigawords 40,000 sentences (AFP news)

الفلسطينيبسببتظاهرةلانصارحركةالمقاومةالاسلاميةحماس.

و اذاتحققذلكفانكيسلوفسكيبكونقدحازثلاث ... فيابرزثلاثة

... صحية

... +قائد

... الايقل

وقالتدانييلتومسونالتيكتبتالسيناريو.وقداستغرقاعدادهخمسةاعوام."تاريخي

Google translate:
"Filstinebsbptazahrplansarhrkpalmquaompalaslamiphamas."

NPYLM

الفلسطيني بسبب تظاهرة ل انصار حركة المقاومة الاسلامية حماس .

و اذا تحقق ذلك ف ان كيسلوفسكي يكون قد حاز ثلاث ... فيابرز ثلاثة

... الصحية

... سطينية ... + " .

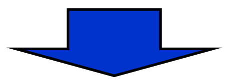
... ماالايقل

وقد استغرق اعداد ه خمسةاعوام . و قال ت دان ييل تومسون التي " تاريخي

Google translate:
"Palestinian supporters of the event because of the Islamic Resistance Movement, Hamas."

English (“Alice in Wonderland”)

first,shedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspedupo
nherknee,andthebrighteagereyeswerelookingupintohersshecouldheartheveryto
nesofhervoice,andseethatqueerlittletossofherheadtokeepbackthewanderinghai
rthatwouldalwaysgetintohereyesandstillasshelistened,orseemedtolisten,thewho
leplacearoundherbecamealivethestrangecreaturesofherlittlesister'sdream.thelo
Nggrassrustledatherfeetasthewhiterabbithurriedbythefrightenedmousesplashed
Hiswaythroughtheneighbouringpoolshecouldheartherattleoftheteacupsasthemar
chhareandhisfriendssharedtheirneverendingmeal,andtheshrillvoiceofthequeen…



first, she dream ed of little alice herself ,and once again the tiny hand s were clasped upon her knee ,and the bright eager eyes were looking up into hers -- shecould hearthe very tone s of her voice , and see that queer little toss of herhead to keep back the wandering hair that would always get into hereyes -- and still as she listened , or seemed to listen , thewhole place a round her became alive the strange creatures of her little sister 'sdream. thelong grass rustled ather feet as thewhitera bbit hurried by -- the frightened mouse splashed his way through the neighbour ing pool -- shecould hearthe rattle ofthe tea cups as the marchhare and his friends shared their never -endingme a l ,and the …

Conclusion

- Completely unsupervised word segmentation of arbitrary language strings
  - Combining word and character information via hierarchical Bayes
  - Very efficient using forward-backward+MCMC
- Directly optimizes Kneser-Ney language model
  - N-gram construction without any “word” information
  - Sentence probability calculation with all possible word segmentations marginalized out
    - Easily obtained from dynamic programming

Future Work

- Semi-supervised word segmentation with CRF
  - Generative model needed in semi-sup learning
  - Ongoing with Suzuki & Fujino (NTT)
- Bilingual word segmentation that optimizes SMT
  - Xu+ (COLING 2008) in semi-supervised, HDP & direct Gibbs
- *Now there are no need for Viterbi segmentation: let’s sample it or implicitly marginalize it!*