Power Law Discounting for N-Gram Language Models

Songfang Huang and **Steve Renals**

CSTR, University of Edinburgh

March 17th, 2010

Language Model Smoothing

- n-gram language model aims to estimate distribution over next word given a context

$$P(w_n|\mathbf{u}) \approx P(w_n|w_{n-2}, w_{n-1})$$

- Maximum likelihood – zero probability problem

$$P(w_n|\mathbf{u}) \approx P_{\mathbf{u}}^{\mathsf{ML}}(w) = \frac{c_{\mathbf{u}w}}{c_{\mathbf{u}\bullet}}$$

- Smoothed estimates
  - redistribute probability to unseen events
  - interpolate with lower order n-grams

Kneser-Ney Smoothing

- Interpolated Kneser-Ney (IKN) smoothing

$$P(w|\mathbf{u}) \approx P^{\text{IKN}}_{\mathbf{u}}(w) = \frac{c_{\mathbf{u}w} - d_{|\mathbf{u}|}}{c_{\mathbf{u}\bullet}} + \frac{d_{|\mathbf{u}|} t_{\mathbf{u}\bullet}}{c_{\mathbf{u}\bullet}} P^{\text{IKN}}_{\pi(\mathbf{u})}(w)$$

$c_{\mathbf{u}w} \sim$ count of $\mathbf{u}w$

$d_{|\mathbf{u}|} \sim$ discount

$t_{\mathbf{u}\bullet} \sim$ num word types observed after $\mathbf{u}$

$\pi(\mathbf{u}) \sim$ one word shorter context than $\mathbf{u}$

- Bayesian interpretation to IKN smoothing (*Teh, 2006*)
- Will more discount parameters help?
  - modified Kneser-Ney (MKN) smoothing improves over IKN by using 3 discounts for one, two, and three or more counts.

Power Laws

- Word frequencies tend to follow a power law distribution (Zipf's Law)

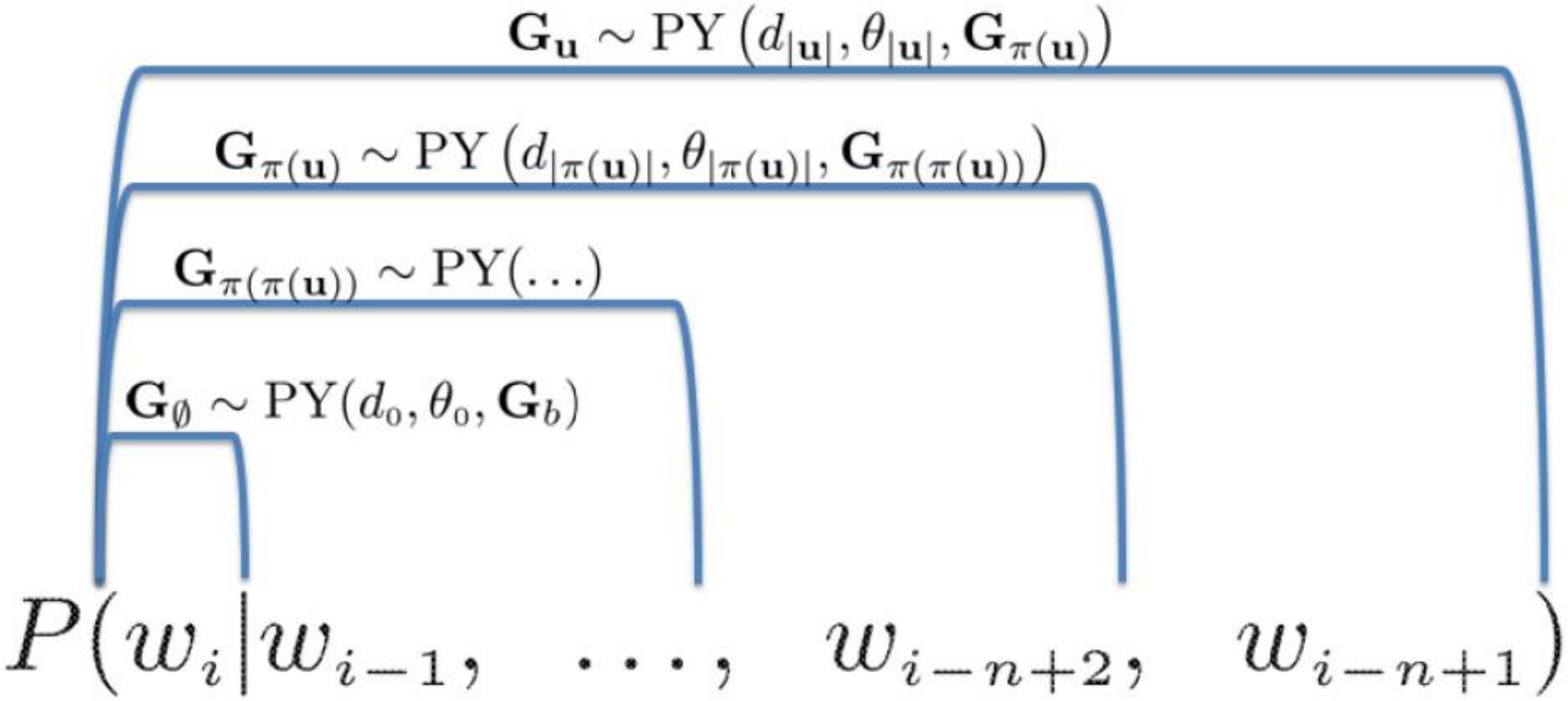
$$P(c_w = x) \propto x^{-d}$$

- Using Pitman-Yor process prior for LM to obtain power law behaviour (*Teh 2006*; *Goldwater et al 2006*)
- Hierarchical Pitman-Yor process LM (HPYLM) (*Teh 2006*)

Hierarchical Pitman-Yor Process LM

- Bayesian LM
  - place a prior over the predictive distribution of the LM
  - infer a posterior distribution from the training data
  - final predictive probability obtained by marginalizing out the latent variables and hyperparameters
- Pitman-Yor Process PY(d, θ, H)
  - a prior for non-parametric models
  - d is discount, θ is strength, and H is mean of PY
- Applying Pitman-Yor process prior recursively

Hierarchical Pitman-Yor Process LM



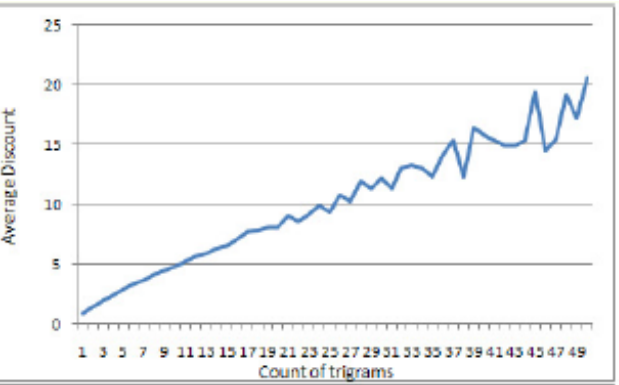
Hierarchical Pitman-Yor Process LM

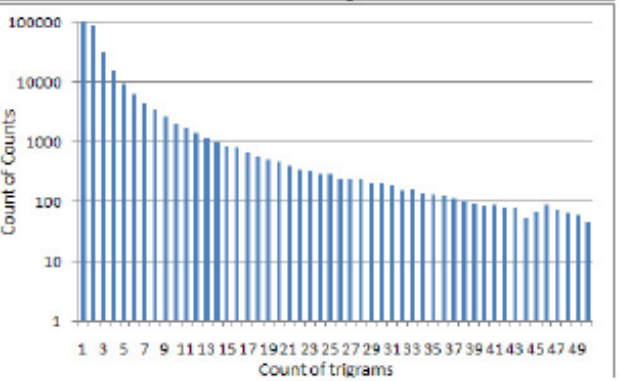
- Predictive probability:

$$P(w|\mathbf{u},\mathcal{S},\Theta) = \frac{c_{\mathbf{u}w\bullet} - d_{|\mathbf{u}|}t_{\mathbf{u}w}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\bullet\bullet}} + \frac{\theta_{|\mathbf{u}|} + d_{|\mathbf{u}|}t_{\mathbf{u}\bullet}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\bullet\bullet}} P(w|\pi(\mathbf{u}),\mathcal{S},\Theta)$$

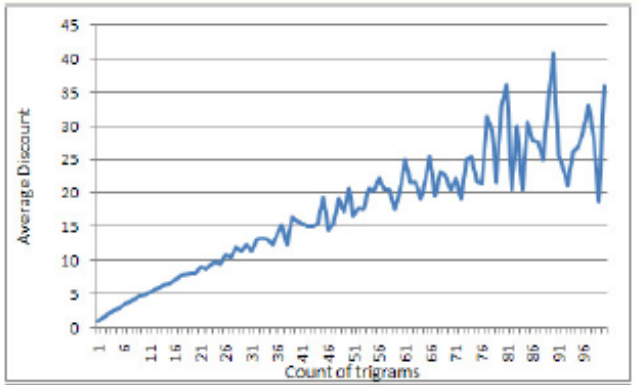
- HPYLM gives us …
  - a generalized form for IKN smoothing
  - power law distribution over t_{uw}
  - better perplexity results than IKN smoothing (*Teh 2006*)
- We have …
  - verified the effectiveness of HPYLM on ASR (*ASRU 2007*)
  - and proposed a parallel algorithm (*Interspeech 2009*)

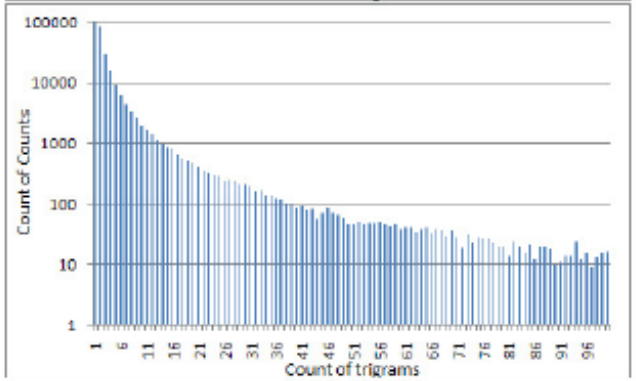
Discounting in the HPYLM



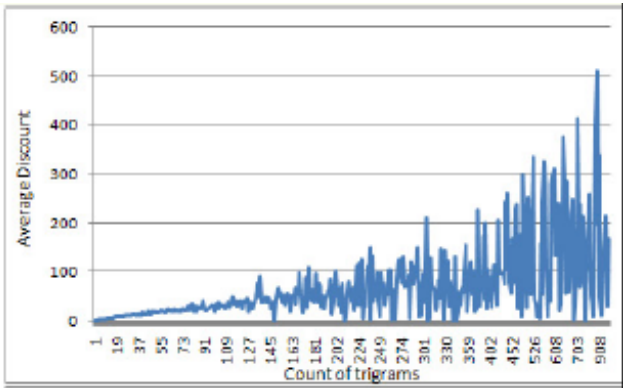


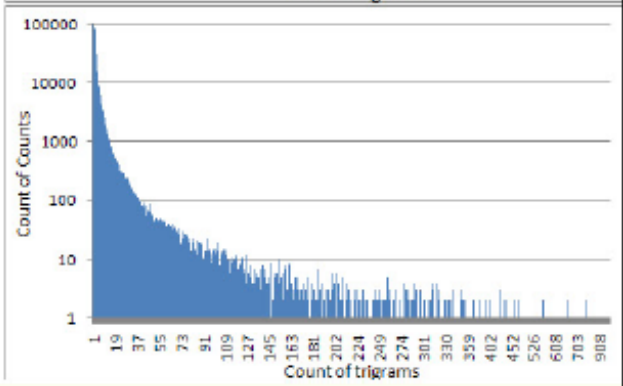
Counts 1-50





Counts 1-100





Counts 1-1000

- Each count has a different discount value
- Large discounts for trigrams with large counts

From HPYLM to PLDLM

- HPYLM is computationally heavy to estimate owing to sampling
- **Power Law Discounting** (PLD) LM:
  - an approximation to HPYLM
  - directly estimate t_{uw} rather than sampling

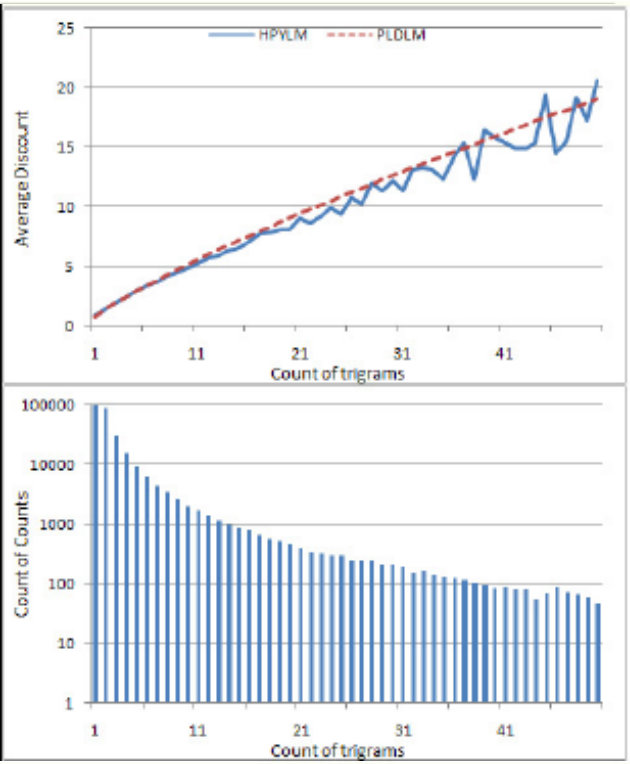
$$P^{\text{PLD}}(w|\mathbf{u}) = \frac{c_{\mathbf{u}w} - dt_{\mathbf{u}w}}{\theta + c_{\mathbf{u}\bullet}} + \frac{\theta + dt_{\mathbf{u}\bullet}}{\theta + c_{\mathbf{u}\bullet}} P^{\text{PLD}}(w|\pi(\mathbf{u}))$$

$$\begin{aligned} d &= \frac{n_1}{n_1 + 2n_2} \\ t_{\mathbf{u}w} &= f(c_{\mathbf{u}w}) = c_{\mathbf{u}w}^d \\ t_{\mathbf{u}\bullet} &= \sum_w t_{\mathbf{u}w} = \sum_w c_{\mathbf{u}w}^d \end{aligned}$$

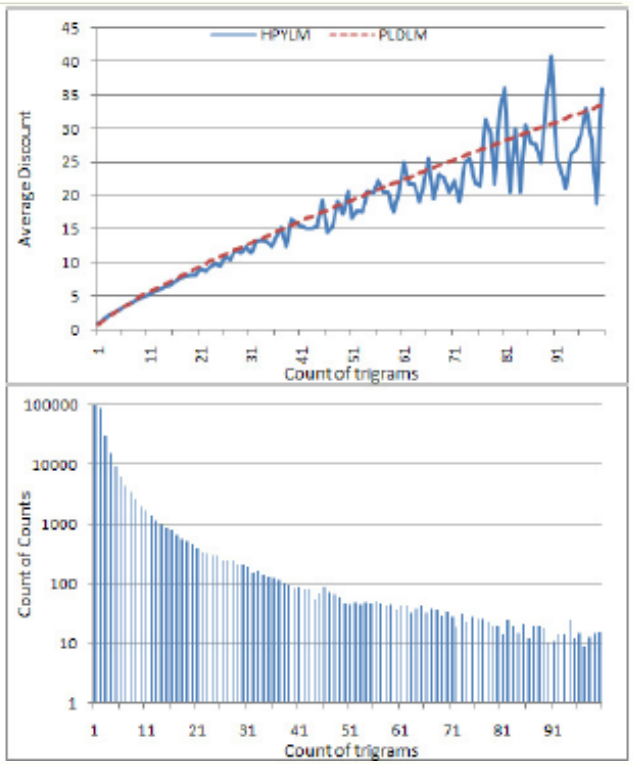
Power Law Discounting

- Hyperparameter estimation in PLDLM
  - set d similar to IKNLM: $d=n_1/(n_1+2n_2)$
  - we find smoothing is insensitive to value of θ
- PLDLM preserves the marginal constraints
- If cutoff=1, PLD reverts to IKN
- if cutoff=3, PLD is comparable to MKN

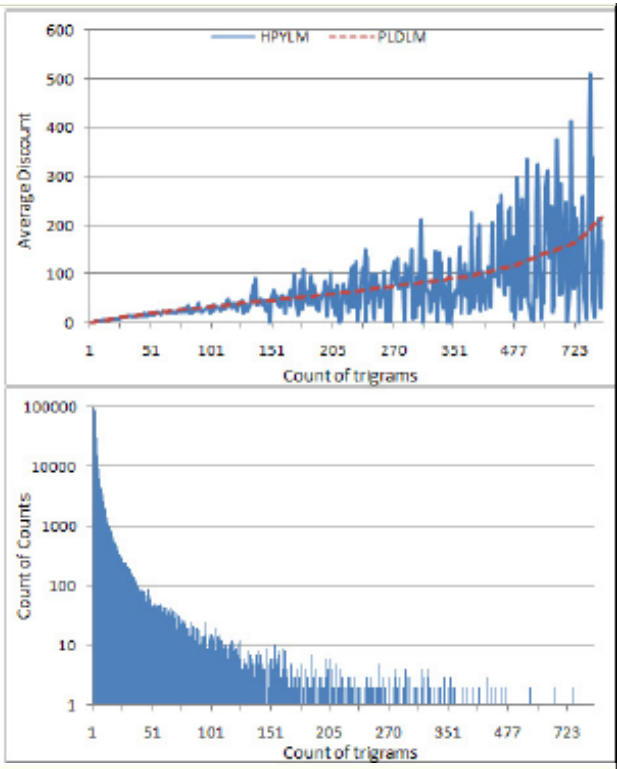
Discounting in the PLDLM



Counts 1-50



Counts 1-100



Counts 1-1000

- PLDLM well approximates the HPYLM for discounts
- but much computationally efficient to estimate

Perplexity on NIST RT06s

- train on several meeting transcripts, and test on *rt06seval* (31,810 words), 50K vocabulary

DATA	SIZE	IKN	MKN	HPY	PLD	PLD3
meeting-s1	1.8M	110.1	106.5	101.2	104.3	105.7
fisher-03-p1	10.6M	134.0	128.5	121.4	122.6	128.1
webmeet	36.1M	176.8	170.6	159.3	159.7	169.6
webconv	162.9M	135.4	131.8	120.2	120.8	130.5
ALL-1	**211.4M**	**107.0**	**105.2**	**98.9**	**100.7**	**104.6**

Word Error Rate on NIST RT06s

- AMI-ASR system (*Hain et al 2007*)
- LMs trained on ALL-1, used in first pass decoding

LMS	WER / %
IKNLM	27.0
MKNLM	26.8
HPYLM	26.5
PLDLM	**26.6**

- PLDLM is significant better than IKNLM ($p<0.01$) and MKNLM ($p<0.05$), but no significant different to HPYLM

Perplexity on AMI

- train on several meeting transcripts, and test on *amiseval* (175,302 words), 50K vocabulary

DATA	SIZE	IKN	MKN	HPY	PLD	PLD3
meeting-s2	1.7M	114.7	112.0	106.9	110.9	110.9
h5etrain03v1	3.5M	234.6	223.3	210.6	210.8	220.5
fisher-03-p1p2	21.2M	221.2	210.9	200.7	198.2	209.7
hub4-lm96	130.9M	321.1	301.3	289.1	282.5	282.5
ALL-2	**157.3M**	**168.6**	**163.9**	**158.8**	**157.9**	**163.7**

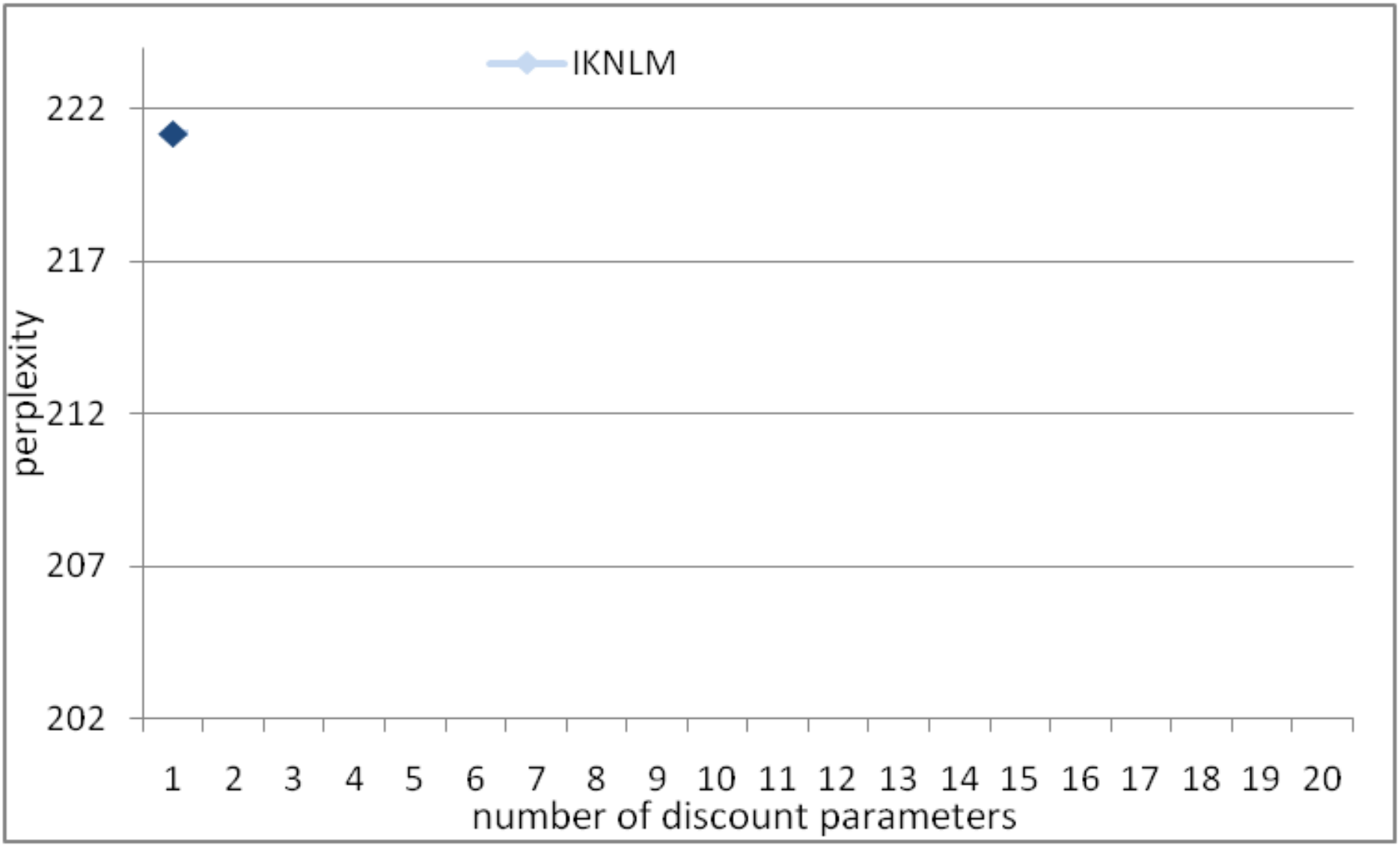
Word Error Rate on AMI

- AMI-ASR system (*Hain et al 2007*)
- LMs trained on ALL-2, used in first pass decoding

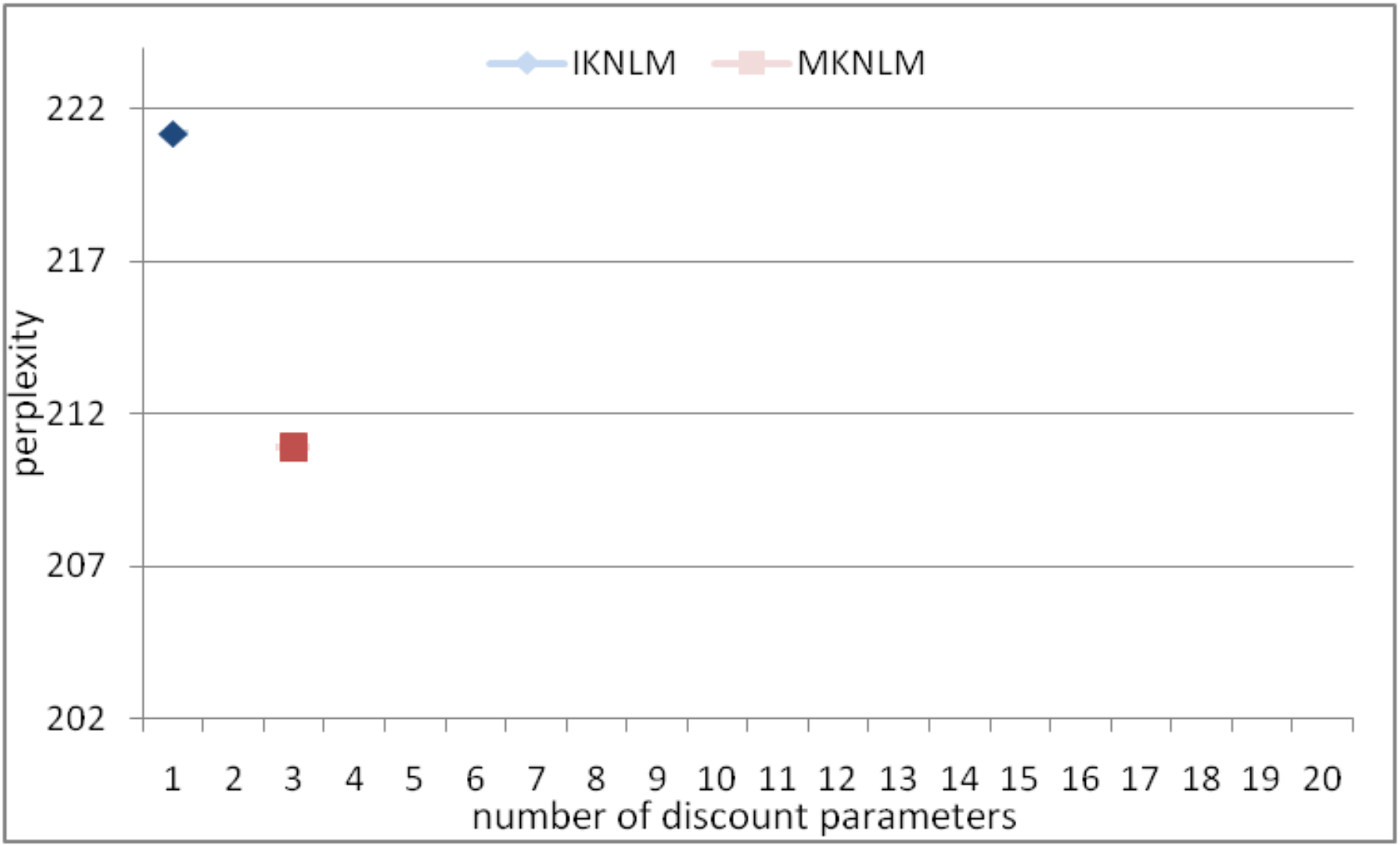
LMS	WER / %
IKNLM	38.6
MKNLM	38.5
HPYLM	38.2
PLDLM	**38.3**

- PLDLM is significant better than IKNLM and MKNLM ($p<0.001$), but again no significant different to HPYLM

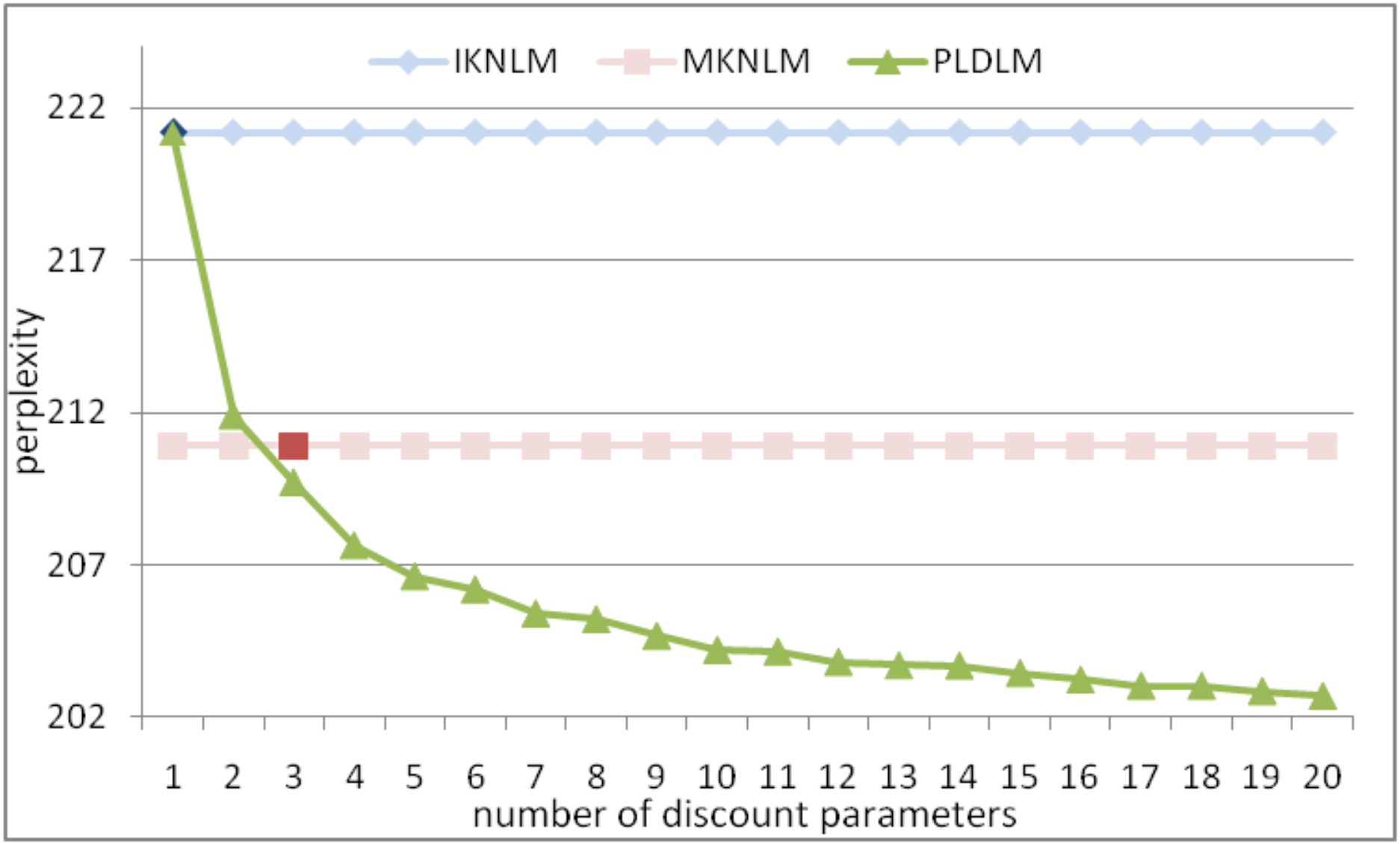
Effect of Discount Parameters



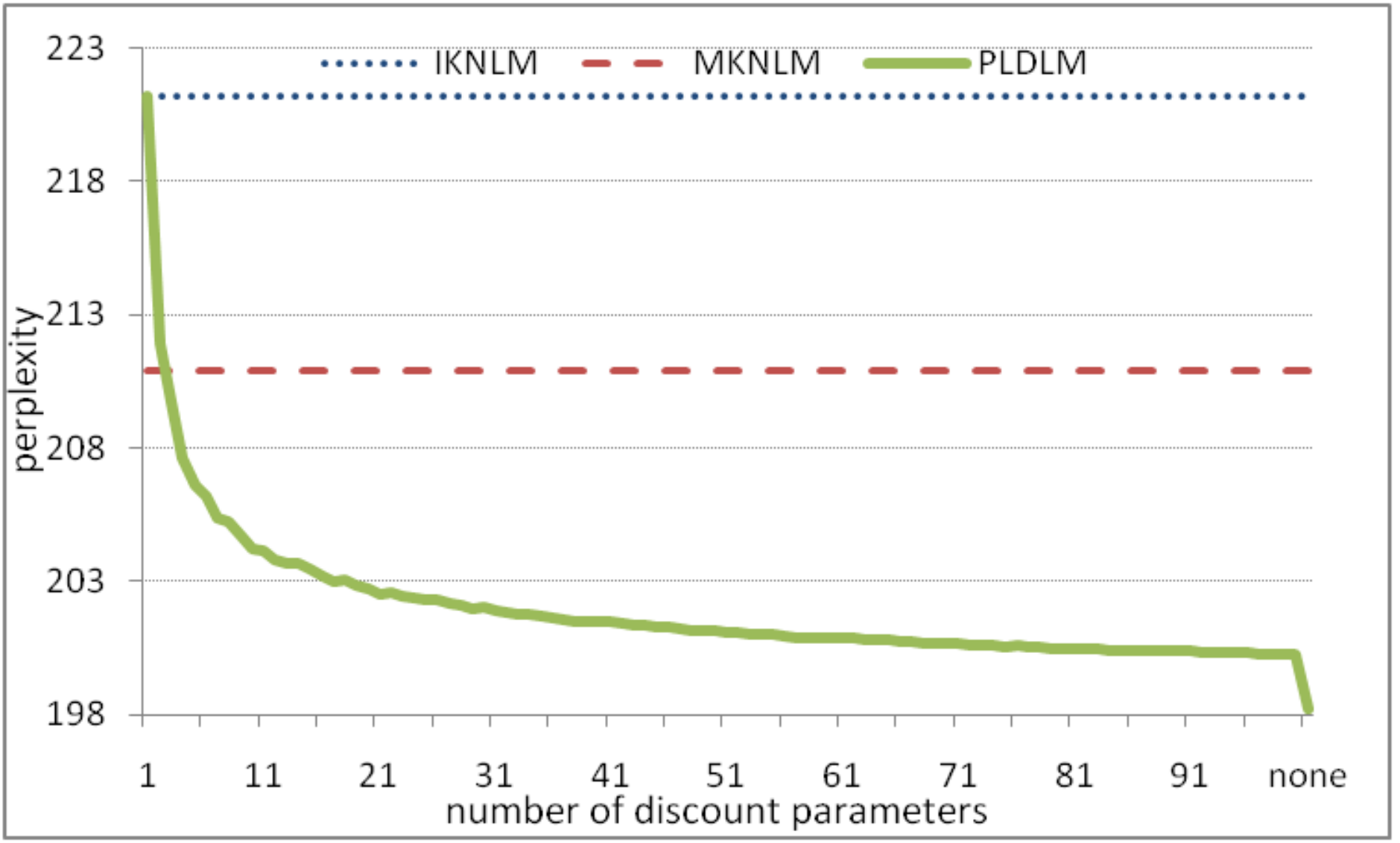
Effect of Discount Parameters



Effect of Discount Parameters



Effect of Discount Parameters



Conclusions

- We propose a simple but efficient smoothing technique for LM that generalizes IKN and MKN smoothing
- Power law behaviour of HPYLM directly approximated by power law discounting approach
- Small but significant decreases in word error rate on meeting corpora
- Program available from http://homepages.inf.ed.ac.uk/s0562315

Thank you for your attentions!