Semi-Markov Model for Sequence Segmentation

Automatic Paragraph Segmentation via Max-Margin Semi Markov Models

Qinfeng (Javen) Shi,
Y. Altun, A. Smola and S. V. N. Vishwanathan

Statistical Machine Learning Program, NICTA, Canberra, Australia
Australian National University, Canberra, Australia
Toyota Technological Institute, Chicago,US

29 June 2007

Automatic Paragraph Segmentation (APS)

- Related to text segmentation, discourse parsing, topic shift detection
- Useful for speech recognition, summarization,OCR, machine translation(Zhu 1999)
- Previous approaches formulate it as a labeling problem
- We use Semi-Markov models that enable to encode properties of whole paragraph
- We apply a Maximum Margin learning algorithm and improve over previous results

Automatic Paragraph Segmentation

"I'm afraid," said Holmes, smiling, "that all the queen's horses and all the queen's men cannot avail in this matter."
He had spread out his big map of London and leaned eagerly over it.
"Well, well," said he presently with an exclamation of satisfaction, "things are turning a little in our direction at last.
Why, Watson, I do honestly believe that we are going to pull it off, after all."
He slapped me on the shoulder with a sudden burst of hilarity.
"I am going out now.
It is only a reconnaissance.
I will do nothing serious without my trusted comrade and biographer at my elbow.
Do you stay here, and the odds are that you will see me again in an hour or two.
If time hangs heavy get foolscap and a pen, and begin your narrative of how we saved the State."
I felt some reflection of his elation in my own mind, for I knew well that he would not depart so far from his usual austerity of demeanour unless there was good cause for exultation.
All the long November evening I waited, filled with impatience for his return.
At last, shortly after nine o'clock, there arrived a messenger with a note:
Am dining at Goldini's Restaurant, Gloucester Road, Kensington.
Please come at once and join me there.
Bring with you a jemmy, a dark lantern, a chisel, and a revolver.
It was a nice equipment for a respectable citizen to carry through the dim, fog-draped streets.
I stowed them all discreetly away in my overcoat and drove straight to the address given.
There sat my friend at a little round table near the door of the garish Italian restaurant.
"Have you had something to eat?
Then join me in a coffee and curacao.
Try one of the proprietor's cigars.
They are less poisonous than one would expect.
Have you the tools?"

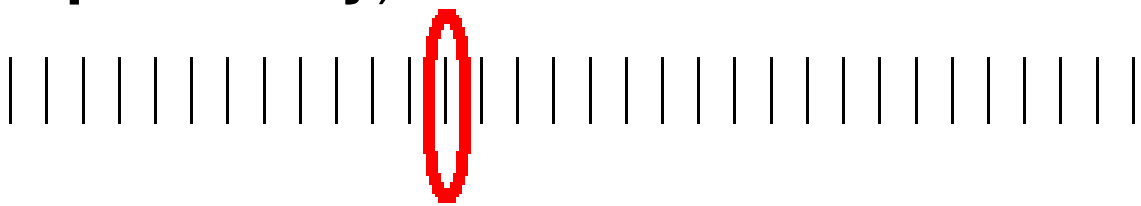
"I'm afraid," said Holmes, smiling, "that all the queen's horses and all the queen's men cannot avail in this matter." 0
He had spread out his big map of London and leaned eagerly over it.
"Well, well," said he presently with an exclamation of satisfaction, "things are turning a little in our direction at last.
Why, Watson, I do honestly believe that we are going to pull it off, after all."
He slapped me on the shoulder with a sudden burst of hilarity.
"I am going out now.
It is only a reconnaissance.
I will do nothing serious without my trusted comrade and biographer at my elbow.
Do you stay here, and the odds are that you will see me again in an hour or two.
If time hangs heavy get foolscap and a pen, and begin your narrative of how we saved the State."
I felt some reflection of his elation in my own mind, for I knew well that he would not depart so far from his usual austerity of demeanour unless there was good cause for exultation. 10
All the long November evening I waited, filled with impatience for his return.
At last, shortly after nine o'clock, there arrived a messenger with a note:
Am dining at Goldini's Restaurant, Gloucester Road, Kensington. 13
Please come at once and join me there.
Bring with you a jemmy, a dark lantern, a chisel, and a revolver.
It was a nice equipment for a respectable citizen to carry through the dim, fog-draped streets. 16
I stowed them all discreetly away in my overcoat and drove straight to the address given.
There sat my friend at a little round table near the door of the garish Italian restaurant.
"Have you had something to eat? 19
Then join me in a coffee and curacao.
Try one of the proprietor's cigars. 130
They are less poisonous than one would expect.
Have you the tools?"

Labeling approach:
Assign boundary/non-boundary labels to every sentence (**independently** or **jointly**)

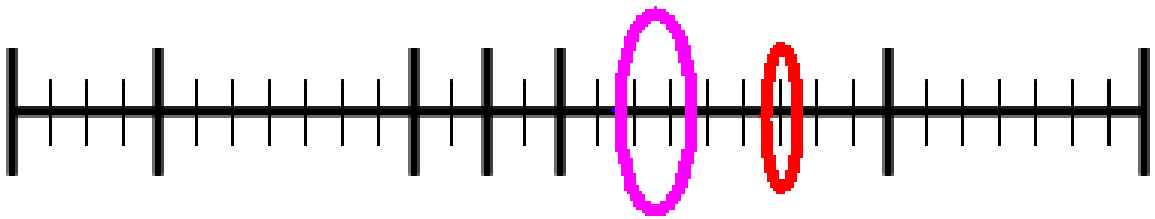
Segmentation approach:
Find the consecutive sentences forming a segment.

Two approaches

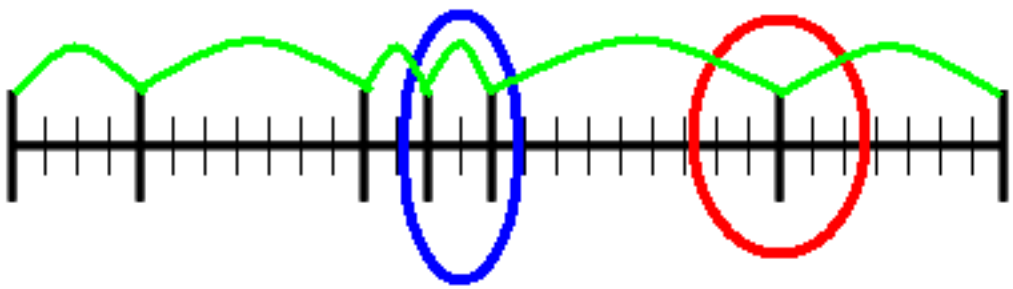
SVM,BoosTexter(BT),Perceptron(Labeling approach, **independently**)



HMM(Labeling approach, **jointly**)



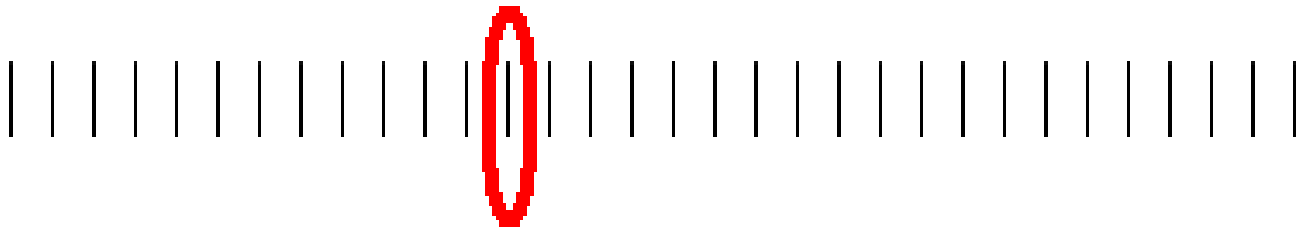
SMMs(Segmentation approach)



Previous Work

Labeling approach (Sporleder & Lapata 2004,2006, Genzel 2005, Filippova & Strube 2006)

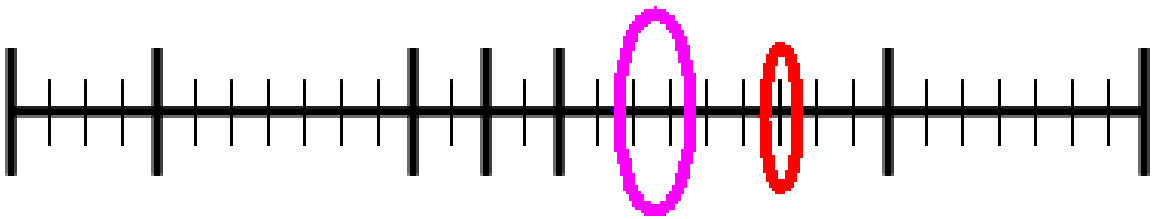
- Sentences are labeled independently via binary classifiers.
- Features capture properties of the boundaries/sentences.
  - Length of current sentence, punctuation, word overlap with previous/next sentences, similarity with previous/next sentences, etc.



- Disadvantages:
  - It can lead to very short or long paragraphs.
  - It cannot capture properties of a paragraph.

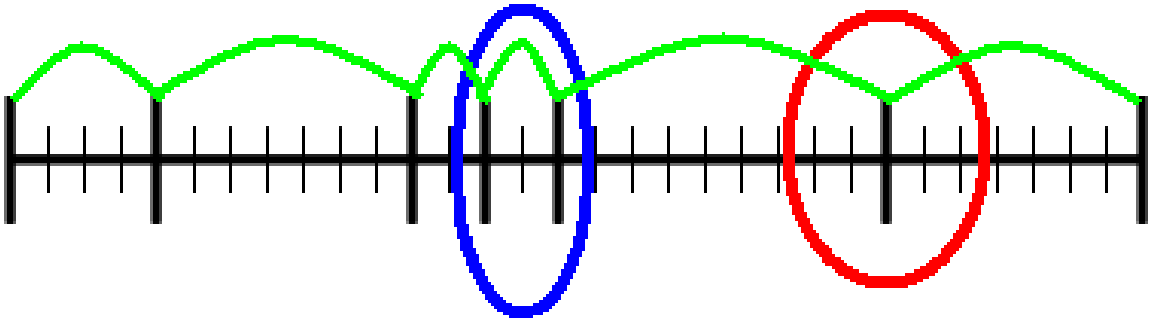
Why not a Markov Chain?

- Sentences have a sequence dependency structure.
- Why not use a Markov chain to overcome disadvantages of independent classification!
- We can apply generalizations of SVMs or Boosting on this structure for discriminative learning.



- Disadvantages:
  - It still cannot use features for properties of paragraphs even for higher order.

Use Semi-Markov Models (SMMs)



$$y = (0, 4, 11, 13, 15, 24, ...)$$

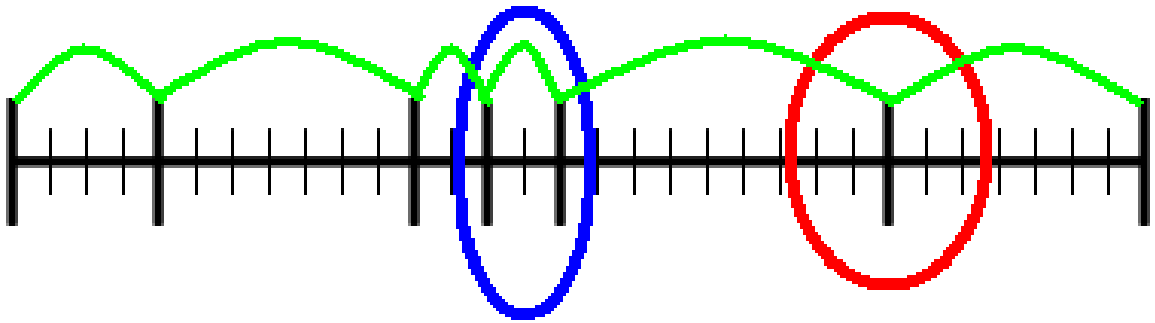
- Model the text as sequence of **paragraphs**, not a sequence of sentences(Sarawagi&Cohen 2004)

$$y = \{n_i\}_{i=0}^{L-1},$$

n_i = beginning of paragraph i
L = number of boundaries

- Can encode features of boundaries as well as features over paragraphs, e. g.
  - Paragraph coherence (via within paragraph similarity)
  - Content change across paragraphs (via dissimilarity across paragraphs)

SMMs (cont.)



- Find boundaries with best score,
  $y^*(x) := \text{argmax}_y \langle w, \phi(x, y) \rangle$
- Use stationarity $\phi(x, y) = \sum_i \phi(x, n_i, n_{i+1})$
- Three kinds of features:
  - Constant feature: $\phi_0 = L \quad L =$ number of boundaries
  - Boundary features: $\phi_1(x, n_i)$
  - Paragraph features: $\phi_2(x, n_i, n_{i+1})$

Boundary Features

Features from previous work (Sporleder & Lapata 2004,2006, Genzel 2005)

- Length of current sentence $x(n_i)$.
- Relative Position of $x(n_i)$.
- Final punctuation of $x(n_i)$.
- Number of capitalized words in $x(n_i)$.
- Bag of Words of $x(n_i)$.
- Word Overlap of $x(n_i)$ with the next (previous) one
- Cosine similarity of $x(n_i)$ with the next (previous) one
- Entropy of $x(n_i)$

Paragraph Features

Specific for SMM models

- Length of the paragraph
- Entropy of the paragraph
  (for **within paragraph similarity**)
- Cosine similarity of current and next paragraph
  (for **dissimilarity across paragraphs**)
  $P_i = (\underline{n_i}, ..., \underline{n_{i+1}} - 1) \quad P_{i+1} = (\underline{n_{i+1}}, ..., \underline{n_{i+2}} - 1),$

$$CS(P_i, P_{i+1}) = \frac{\langle BOW(P_i), BOW(P_{i+1}) \rangle}{\mid BOW(P_i) \mid \times \mid BOW(P_{i+1})) \mid}$$

approximated by similarity of two boundaries

Max Margin Semi-Markov Models

- Primary(Tsochantaridis et al 2004)

$$\min_{w,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i$$

$$\forall i \max_{y} \langle w, \Phi(x_i, y_i) - \Phi(x_i, y) \rangle \geq \Delta(y_i, y) - \xi_i$$

- Cost: Count number of missed and added boundaries

$$\Delta(y, y') = |y| + |y'| - 2|y \cap y'|$$

- Dual formulation to use kernels
- Polynomial time training algorithm which requires $\text{argmax}_y \Delta(y_i, y) + \langle w, \Phi(x_i, y) \rangle$.

Dynamic Programming

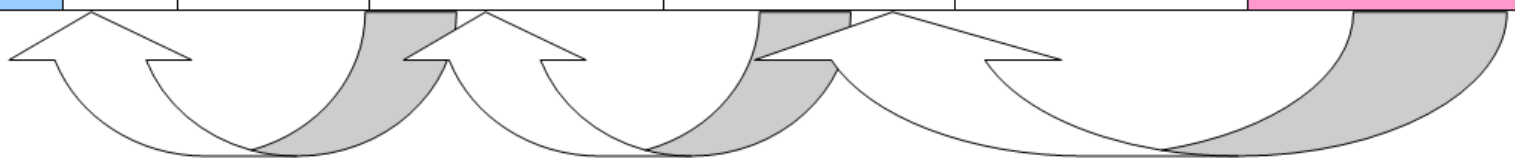
R_i :the best boundary if current boundary is at position i
T_i :score of best segmentation at position i

$$R_i = \underset{\max(0,i-M)\le j<i}{\operatorname{argmax}} \{T_j + g(j,i)\}$$

$$T_i = T_{R_i} + g(R_i, i), \quad M = \text{max length of paragraphes}$$

where $g(j,i) = \langle w, \phi(x,j,i)\rangle + 1 - 2\{(j,i) \in y\}$

sentences	0	1	2	3	4(Last sentence)	5(Dummy)
T (score)	$T(0)$	$T(0)+g(0,2)$	$\max_{j=0}^{1}\{T(j) + g(j,2)\}$	$\max_{j=0}^{2}\{T(j) + g(j,3)\}$	$\max_{j=0}^{3}\{T(j) + g(j,4)\}$	$\max_{j=0}^{4}\{T(j)\}$
R (indicator)	-1	0	$\arg\max_{j=0}^{1}\{T(j) + g(j,2)\}$ 0	$\arg\max_{j=0}^{2}\{T(j) + g(j,3)\}$ 2	$\arg\max_{j=0}^{3}\{T(j) + g(j,4)\}$ 2	$\arg\max_{j=0}^{4}\{T(j)\}$ 3



$y = \{0,2,3\}$ $O(MN)$

Experiment and Analysis

- Training (72%), development (21%), and test set (7%)
- Two languages (English, German)
- SMM, SVM,BoosTexter(BT)

Table: Test results on ENG and GER data after model selection.

DATA	ALGO.	ACC.	REC.	PREC.	F_1
ENG	SMM	**75.61**	**46.67**	**77.78**	**58.33**
	SVM	58.54	26.67	40.00	32.00
	BT	65.85	33.33	55.56	41.67
GER	SMM	70.56	46.81	65.67	54.66
	SVM	39.92	**100.00**	38.68	55.79
	BT	**72.58**	54.26	**67.11**	**60.00**

Experiment and Analysis

Conjecture

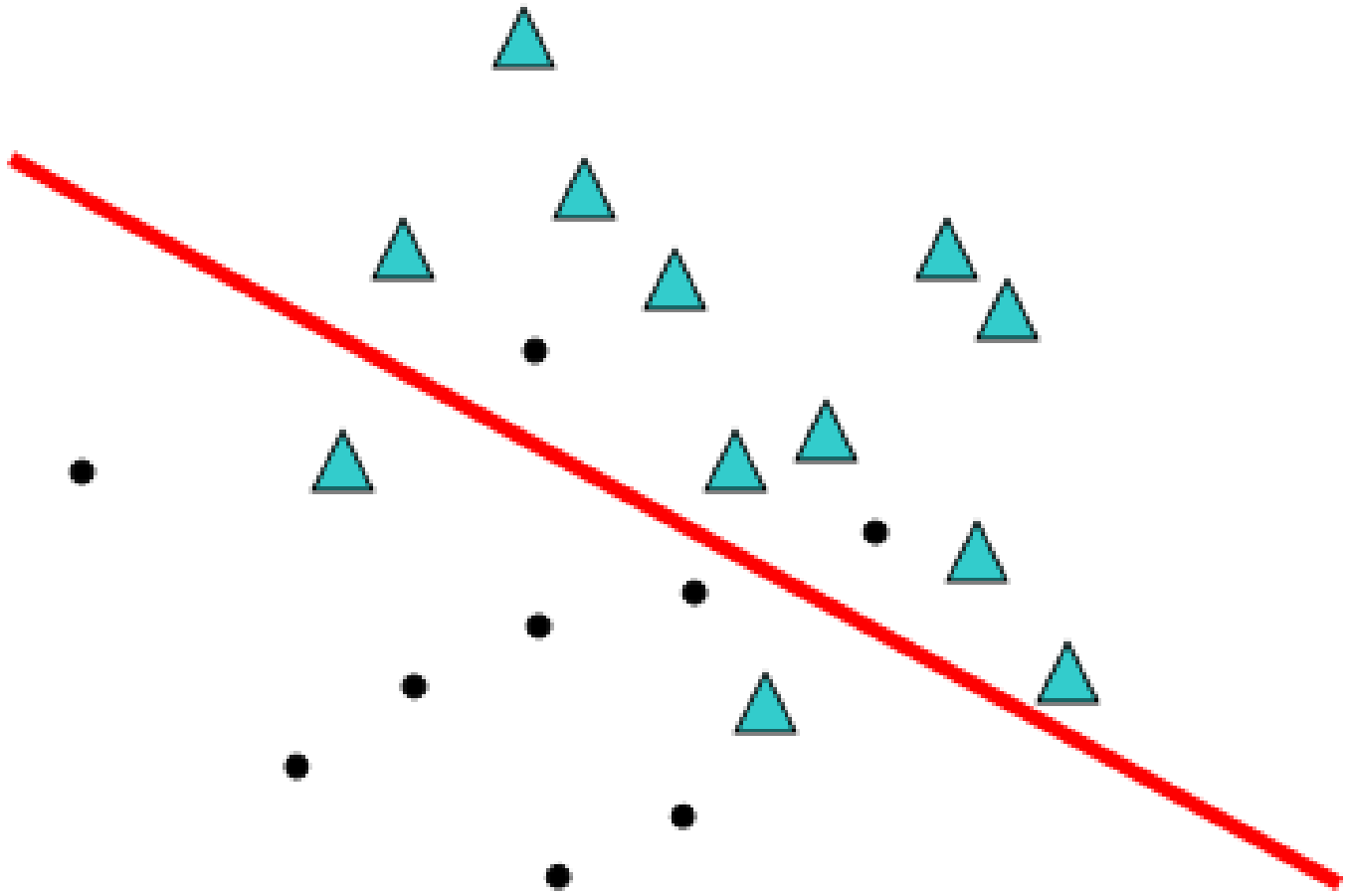
- The styles from dev and test may be different
- Test on texts from the same book and different books from the same author

Table: Comparison on different datasets.

DATASET	ACC.	REC.	PREC.	F_1-SCORE
SB (DEV)	92.81	86.44	92.73	89.47
SB (TEST)	96.30	96.00	96.00	96.00
SA (DEV)	82.24	61.11	82.38	70.17
SA (TEST)	81.03	79.17	76.00	77.55

Experiment and Analysis

Play with Offset

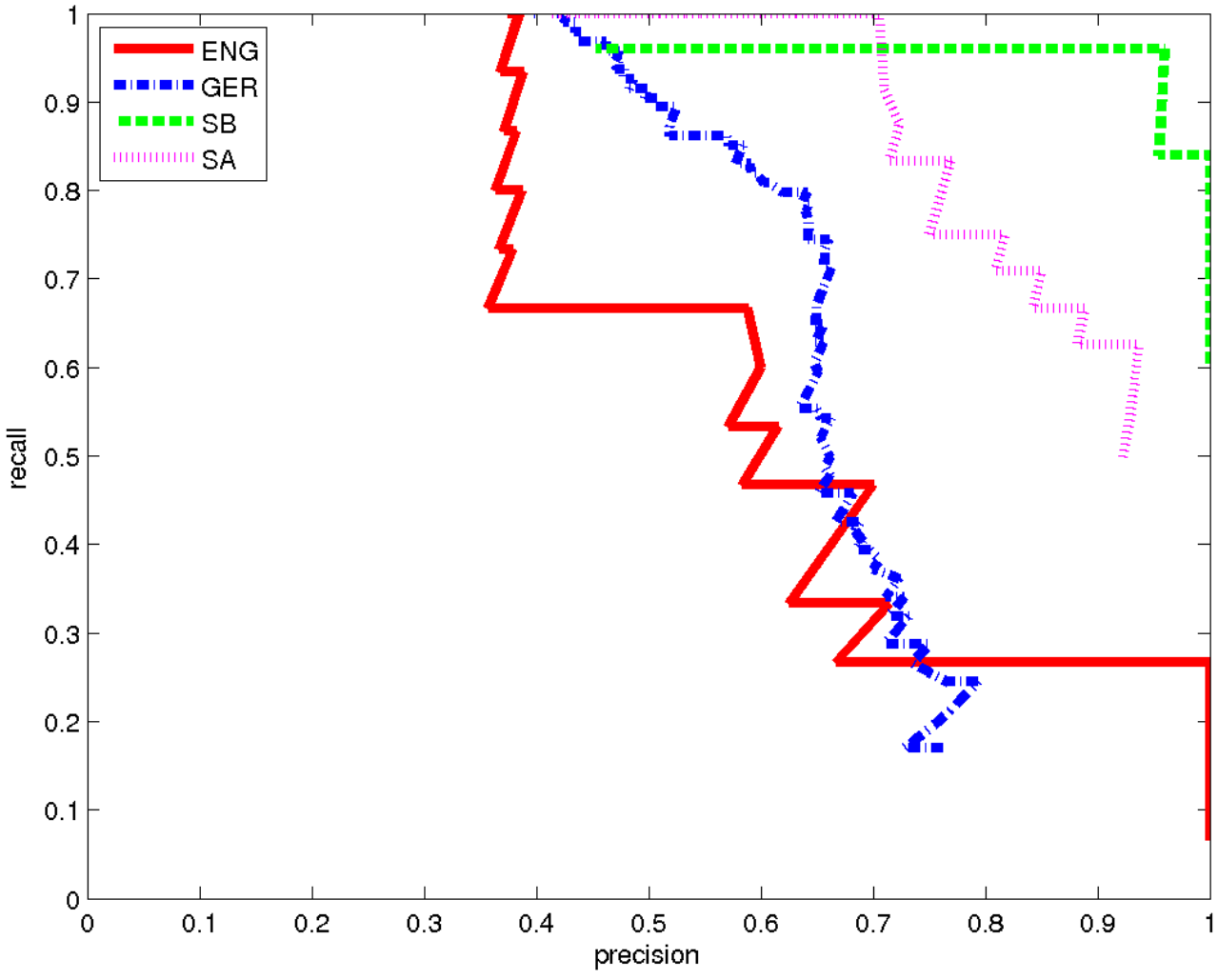


Experiment and Analysis

Table: Tuning the offset for best F_1-score.

DATASET	ACC.OLD	ACC.	F_1 OLD	F_1
ENG (DEV)	69.84	55.02	29.90	60.16↑
ENG (TEST)	75.61	39.02	58.33	52.28↓
GER (DEV)	73.41	64.29	39.98	43.14↑
GER (TEST)	70.56	75.40	54.66	69.35↑

Recall VS Precision



Conclusion

- We presented Semi-Markov models for paragraph segmentation.
  - We can use features for paragraphs.
  - We applied maximum margin structured learning framework to Semi-Markov models.
  - We derived the linear time dynamic program for segmentation.
- Can be applied to video, other sequence simply by providing sufficient features

The End

Acknowledgement

- NICTA
- PASCAL Research Networks

Appendix

Table: Performance of various algorithms on our test corpus.

dataset	Algo.	Acc.	Rec.	F_1-score
Eng	SMM	**77.71±1.95**	33.44±4.42	**40.12±3.55**
	SVM	66.95±1.67	**37.06±3.14**	34.72±1.53
	BT	75.44±2.30	23.43±4.03	29.47±3.98
Ger	SMM	76.68±1.17	**50.87±3.41**	**55.15±3.42**
	SVM	67.22±2.37	19.88±2.72	24.70±2.67
	BT	**77.29±0.76**	47.06±5.28	51.85±4.52
SB	SMM	86.46±2.66	73.62±6.14	78.46±4.78
	SVM	63.73±3.18	41.47±4.17	43.48±3.35
	BT	**87.99±1.97**	**77.51±4.44**	**81.23±3.02**
SA	SMM	**82.96±1.96**	**65.60±4.44**	**71.13±3.97**
	SVM	58.26±2.81	49.92±4.16	41.05±3.29
	BT	78.41±2.32	57.75±4.87	62.46±4.81