NAACL2001

Chunking with Support Vector Machines

Graduate School of Information Science,

Nara Institute of Science and Technology, JAPAN

Taku Kudo, Yuji Matsumoto

{*taku-ku,matsu*}*@is.aist-nara.ac.jp*

Chunking (1/2)

Dividing sentences into syntactically related non-overlapping groups

```
- Example of BaseNP chunking:

 In [ early trading ] in
 [ Hong Kong ] [ Monday ] , [ gold ] was
 ...


- Example of Base Phrase chunking:

 [ In ]/PP  [ early trading ]/NP [ in ]/PP
 [ Hong Kong ]/NP [ Monday ]/NP , [ gold ]/NP [ was ]/VP
 ...
```

Chunking (2/2)

Other Chunking Tasks:

- Named Entity extraction
- Japanese *bunsetsu* identification
- Tokenization
- Part-of-speech tagging

Our approaches

- Propose a general framework for chunking based on SVMs
- Apply the weighted voting from 8 SVMs-based systems trained with distinct chunk representations

Outline

- Brief introduction to Support Vector Machines
- How do we apply SVMs to Chunking?
- Weighted Voting from 8 SVM-based systems
- Experiments and Evaluation
- Summary and future work

Support Vector Machines (1/3)

- V.Vapnik 1995
- Two strong properties
  - High generalization performance independent of feature dimension
  - Training with combinations of multiple features by using a Kernel Function.

Support Vector Machines (2/3)

- Separate positive and negative (binary) examples with a **Linear Hyperplane**: $(\mathbf{w} \cdot \mathbf{x} + b, \quad \mathbf{w}, \mathbf{x} \in \mathbf{R^n}, b \in \mathbf{R})$
- Find an optimal hyperplane (parameter $\mathbf{w}, b$) with the **Maximal Margin Strategy**

Support Vector Machines (3/3)

- Potential to carry out non-linear classification.
- Replace every dot product in optimization formula with some **Kernel Function**
- Build a linear classifier in a higher-dimensional feature space

d-th polynomial kernel

$$K(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + \mathbf{1})^{\mathbf{d}}$$

considering combinations of up to d features

Chunk representation (1/2)

- Regard Chunking as a Tagging task
- Inside/Outside (IOB1) representation

I	Current token is inside a chunk.
O	Current token is outside any given chunk.
B	Current token is the beginning of a chunk which immediately follows another chunk.

Tjong Kim Sang (1997) introduces three alternative versions — IOB2, IOE1 and IOE2

Chunk representation (2/2)

	Base NP Chunking				Base Phrase Chunking
	IOB1	IOB2	IOE1	IOE2	IOB2
In	O	O	O	O	B-PP
early	I	B	I	I	B-NP
trading	I	I	I	E	I-NP
in	O	O	O	O	B-PP
Hong	I	I	I	I	B-NP
Kong	I	I	E	E	I-NP
Monday	B	B	I	E	B-NP

Applying SVMs to Chunking

- Chunking as a classification task of the IOB tags
- We use the pair-wise method to extend a binary classifier (SVMs) to a multi-class classifier

Feature Sets for Learning

- Parsing from left to right,
  $i-1, i-2$ IOB tags are added dynamically(**Forward Parsing**)
- Parsing from right to left,
  $i+1, i+2$ IOB tags are added dynamically(**Backward Parsing**)

Chunking with Weighted Voting (1/3)

- 8 SVM-based classifiers can be built:
  $\{$IOB1/IOB2/IOE1/IOE2$\} \times \{$Forward , Backward$\}$
- Final IOB tag is obtained from the weighted voting
- How can we assign voting weights to individual classifiers?
  - Uniform weights (baseline)
  - 5-fold Cross Validation
  - **VC bound**
  - **Leave-One-Out bound**

Chunking with Weighted Voting (2/3)

Estimate the accuracy of test data (not training data)

- From the theoretical background of SVMs.
- Only using the training data
- Without re-sampling: training and estimation simultaneously

- **VC bound**
  Estimate the accuracy from the size of the **margin**
- **Leave-One-Out bound**
  Estimate the accuracy from the number of **support vectors**

Experiments

- **baseNP-S**: Penn Tree Bank/WSJ
  A standard data set for baseNP chunking
- **baseNP-L**: Penn Tree Bank/WSJ
- **base Phrase chunking**: Penn Tree Bank/WSJ
  Total 10 types of base phrase classes VP, PP, ADJP..
  Data set for CoNLL-2000 Shared Task

- Evaluation measure: *F-measure*
- Kernel Function: 2nd-polynomial kernel

Results of Weighted Voting

	A	B	C	D
baseNP-S	94.16	**94.22**	**94.22**	94.18
baseNP-L	**95.77**	-	95.66	95.66
base Phrase chunking	93.77	93.89	**93.91**	93.85

A:Uniform B:Cross Validation C:VC bound D:L-O-O bound

Results of individual representations

baseNP-S:

	$F_{\beta=1}$	Cross Validation	VC bound	L-O-O bound
IOB1-F	93.76	.9394	.4310	**.9193**
IOB1-B	93.93	.9422	.4351	.9184
IOB2-F	93.84	.9410	.4415	.9172
IOB2-B	93.70	.9407	.4300	.9166
IOE1-F	93.73	.9386	.4274	.9183
IOE1-B	93.98	.9425	.4400	.9217
IOE2-F	93.98	.9409	.4350	.9180
IOE2-B	**94.11**	**.9426**	**.4510**	**.9193**

Results of individual representations

baseNP-L:

	$F_{\beta=1}$	VC bound	L-O-O bound
IOB2-F	**95.34**	.4500	.9497
IOB2-B	95.28	.4362	.9487
IOE2-F	95.32	.4467	.9496
IOE2-B	95.29	**.4556**	**.9503**

Results of individual representations

base Phrase Chunking:

	$F_{\beta=1}$	Cross Validation	VC bound	L-O-O bound
IOB1-F	93.48	.9342	.6585	.9605
IOB1-B	93.74	.9346	.6614	.9596
IOB2-F	93.46	.9341	.6809	.9586
IOB2-B	93.47	.9355	.6722	.9594
IOE1-F	93.45	.9335	.6533	.9589
IOE1-B	93.72	.9358	.6669	**.9611**
IOE2-F	93.45	.9341	.6740	.9606
IOE2-B	**93.85**	**.9361**	**.6913**	.9597

Discussion

- Accuracy improved regardless of the voting scheme used
- Cross-Validation and VC bound outperform Leave-One-Out bound and Uniform in almost all cases
- Comparing VC bound to Cross Validation
  - comparable accuracy
  - both provide good criteria for classifier selection
  - but, Cross Validation requires a larger amount of computational resources

Comparison with related work

	Outline of System	F-measure
Tjong Kim Sang 2000	Weighted voting of different Machine Learning algorithms (MBL, ME, IGTree) and distinct chunk representations (IOB1/IOB2/IOE1/IOE2)	baseNP-S 93.86 baseNP-L 94.22 base Phrase 92.50
Proposed method	Weighted voting of 8-SVMs based systems trained with distinct chunk representations (IOB1/IOB2/IOE1/IOE2)	baseNP-S 94.22 baseNP-L 95.77 base Phrase 93.91

Summary

- We proposed a general framework for chunking based on SVMs.
- We can achieve higher accuracy compared to previous methods
- We can also improve the accuracy by applying weighted voting from 8 SVMs-based classifiers trained with distinct chunk representations
- For the weights assigned to the individual classifiers, we applied methods stemming from the theoretical background of SVMs (VC bound and Leave-One-Out bound)

Future Work

- Application to other chunking tasks
  (NE, POS tagging, *bunsetsu* identification)
- Consider more predictable bounds such as Span SVM
  [Chapelle,Vapnik 2000]
- Incorporate variable length models
  The context length features were selected ad-hoc
  But, the optimal context length depends on the task