

High-Recall Gene Mention Recognition by Unification of Multiple Backward Parsing Models

Han-Shen Huang¹

hanshen@iis.sinica.edu.tw

Yu-Shi Lin¹

bathroom@iis.sinica.edu.tw

Kuan-Ting Lin¹

woody@iis.sinica.edu.tw

Cheng-Ju Kuo²

cju.kuo@gmail.com

Yu-Ming Chang¹

porter@iis.sinica.edu.tw

Bo-Hou Yang^{1,3}

ericyang@iis.sinica.edu.tw

I-Fang Chung²

ifchung@ym.edu.tw

Chun-Nan Hsu¹

chunnan@iis.sinica.edu.tw

¹ Institute of Information Science, Academia Sinica, Taipei, Taiwan

² Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan

³ Department of Electrical Engineering, Chang-Gung University, TaoYuan, Taiwan

1 Introduction

We considered the gene mention tagging task as a classification problem and applied support vector machines (SVM) to solve it. We selected a large set of features as the input and trained two SVM models with different multiclass extension methods. We found that backward parsing constantly outperformed forward parsing regardless of the multiclass extension methods and obtained high precision rates, but recall rates were not as satisfactory. To enhance recall rates, our approach is to construct divergent but high performance models to cover different aspects of the feature space, and then combine them into an ensemble. We applied union and intersection to combine the outputs of SVM models with that of a CRF model, which was trained with the same feature set, and successfully enhanced recall rates without degrading too much precision.

2 Method and Results

SVM has been shown to perform well for name entity chunking in the literature (see, e.g., [2, 5]). Name entity chunking is a problem of supervised sequential learning. To apply SVM to this problem, We used a sliding window to convert the problem into a supervised classifier learning problem [1]. We chose five as the width of the window. During the parsing, the information from the two preceding tokens and the two following tokens are used to construct a feature vector for the classifier to assign a class label to the current token. We chose Yet Another Multipurpose Chunk Annotator (YamCha) [2] to build our SVM models because it is tuned for name entity chunking tasks.

We designed our features based on our experience and previous work on named entity recognition [4, 5, 6]. Table 1 shows the set of features. There are 10 feature types with 617,515 feature values in our feature set. As a preprocessing step, we used the GENIA tagger [7] to tokenize sentences and tag part-of-speech (POS) for training and test data. Then we can extract features from the data.

We used an Inside/Outside representation for name entity chunking with B , I , and O class labels. Since SVM is an intrinsic binary classifier, we must extend SVM to handle multiclass problems. We used two popular methods to extend a binary classifier to multiclass:

- one vs. all: Train a binary classifier for each class against all other classes.

Table 1: Types of features and their possible values

Feature	Value
word	all words in the training data
POS	part-of-speech tagging by GENIA tagger
orthographic	see Table 2 for details
vowel	it is a list of the vowel(s)(a,e,i,o,u) in a word
length	1,2,3~5,≥6
morphological I	replacing digits with a "*" (e.g., Abc123→Abc*)
morphological II	replacing each letter and digit with a morphological symbol (e.g., AbcD123→AaaA111)
prefix	1~6 gram of the starting letters of the token
suffix	1~6 gram of the ending letters of the token
preceding class	class labels(B,I,O) of the two preceding tokens

Table 2: Types of orthographic features and their examples

Feature	Ex.	Feature	Ex.	Feature	Ex.	Feature	Ex.
InitCaps	Abc	SingleDigit	1	BackSlash	/	Apostrophe	'
EndCaps	abC	TwoDigits	12	OpenSquare	[QuotationMark	"
AllCaps	ABC	ThreeDigits	123	CloseSquare]	Greek	α
LowerCase	abc	FourDigits	1234	Colon	:	AminoAcidLong	lysine
WordAndDigits	A1	MoreDigits	12345	SemiColon	;	AminoAcidShort	Lys
InitCapsEndCaps	AbC	Floatpoints	1.2	Percent	%	Nucleoside	Uracil
SingleCap	A	Star	*	OpenParen	(Nucleotide	ATP
TwoCaps	AB	Equal	=	CloseParen)	Roman	V
ThreeCaps	ABC	Plus	+	Comma	,		
MoreCaps	ABCD	Hyphen	-	FullStop	.		

- one vs. one: Train a binary classifier for each pair of classes and select the class appearing in the most outputs.

We also trained a conditional random field (CRF) model to increase the divergence of our ensemble. The CRF model was trained using MALLET [3] with the same features as those used by our SVM models.

We compared two parsing directions: forward and backward. The direction which parses from left to right is forward parsing, while the direction which parses from right to left is backward parsing. Table 3 shows the results of our comparison, which show that backward parsing performed better than forward parsing for both SVM and CRF models, but there is no evidential difference between the SVM models with different multiclass extensions. For all models, precision is substantially better than recall. These models were trained by 10,000 examples selected at random from the data set provided by the organizer and tested by the remaining 5,000 examples.

Our final step is to determine how to integrate results of the three models mentioned above to enhance recall. Weighted majority vote may be a good idea, but regulating the weights for different models is difficult. Poorly assigned weights may degrade the performance. Instead, we simply applied union and intersection to combine these models. Usually, union can enhance recall because it includes more tagging results from different models, but it also degrades precision. In contrast, intersection can filter out false positives and therefore increase precision, but at the expense of recall. To take advantage of both operations but avoid their pitfalls, we applied intersection to the tagging results

Table 3: Performance comparison for different models and parsing directions

Model	Forward			Backward		
SVM+One vs.All	P:82.81%	R:78.27%	F:80.48%	P:86.99%	R:75.79%	F:81.01%
SVM+One vs.One	P:82.41%	R:78.11%	F:80.20%	P:85.49%	R:79.25%	F:82.25%
CRF	P:86.52%	R:79.44%	F:82.83%	P:86.77%	R:80.39%	F:83.46%

P,R and F denote precision, recall, and f-score, respectively.

of the two SVM models and then union with the tagging results of the CRF model as our ensemble model. Table 4 shows the final test results of this model, as well as the final results of the unions of CRF with the two SVM models, as reported by the organizer. The results show that our simple ensemble model remarkably enhanced recall, with all recall results ranked in the top quartile, while precision results dropped slightly, compared with the results in Table 3. All f-score results were ranked in the top quartile, too.

Table 4: Experimental results

Run	Ensemble	Performance		
1	M1UM3	P:83.27%(3)	R:89.34%(1)	F:86.20%(1)
2	M2UM3	P:82.98%(3)	R:89.58%(1)	F:86.15%(1)
3	(M1∩M2)UM3	P:84.93%(3)	R:88.28%(1)	F:86.57%(1)

The number in the parentheses is the quartile among 21 participants

References

- [1] Dietterich, T. G. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, 2002.
- [2] Kudo, T. and Matsumoto, Y. Chunking with support vector machines. In *Proceeding of Second Meeting of North American Chapter of the Association for Computational Linguistics(NAAACL)*, pages 192–199, 2001.
- [3] McCallum, A. K. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>, 2002.
- [4] McDonald, R. and Pereira, F. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6(S6), 2005.
- [5] Mitsumori, T., Fation, S., Murata, M., Doi, K., and Doi, H. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6(S8), 2005.
- [6] Takeuchi, K. and Collier, N. Bio-medical entity extraction using support vector machines. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, pages 57–64, 2003.
- [7] Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pages 382–392, 2005.