SVM を用いたインドネシア語連体従属接続詞の判定システム

Wahyu Purnomo（東京農工大学　工学部　情報工学科）†
古宮　嘉那子（東京農工大学　工学研究院　先端情報科学部門）
小谷　善行（東京農工大学　工学研究院　先端情報科学部門）

SVM-based System for the Determination of Adnominal Subordinating Conjunction in the Indonesian Language

Wahyu Purnomo (Department of Computer and Information Sciences, Faculty of Engineering, Tokyo University of Agriculture and Technology)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

１．はじめに

インドネシア語において，いろいろな接続詞が存在する．その中で，連体従属接続詞「Yang」が最もよく使用されるものである．インドネシア語の「A Yang B」は，B が A を修飾しており，日本語の「B の A」,「BA」又は「B のほうの A」や英語の「BA」又は「A which (to be/ verb) B」に意味が似ている．A には，名詞や名詞フレーズが来る場合が多く，B には，名詞だけでなく動詞や形容詞や長いフレーズなどが来る可能性がある．

インドネシア語には「A Yang B」の他にも「AB」という書き方もある．しかし，B に来るものが長ければ「Yang」があった方が自然である．人間でも「Yang」の有無を判定するのは難しい．そのため，本稿では，文書から「Yang」の使用方法をコンピュータに学習させ，自動的に「Yang」の有無を判定させるシステムを提案する．コンピュータに手掛かりの素性を変えたりすることで，正解率が変わると見られた．

２．関連研究

（宋，浅原，古宮，小谷（2013））がインドネシア語の「Yang」に近い意味を持っている中国語の「的」の研究を報告している．（宋，浅原，古宮，小谷（2013））では，SVM を用いて中国語助詞の用法を解析した．その結果，コンピュータが判定した場合は正解率が 97.4%で，人間が判定した場合は正解率が 96.2%だった．

また，（竇，古宮，小谷（2012））では，中国語においてインターネット上でよく使用される表現を判定するシステムを提案した．（竇，古宮，小谷（2012））の結果では，100 文の語を対象にアンケートを用いて人間で判定する際に 42%の正解率が得られた．一方，同じ 100 文の語を SVM で判定する際に 92%の正解率が得られた．

３．連体従属接続詞の有無を判定するシステム

本システムは，大きく分けると文書をコンピュータに学習させる側の学習部分と学習させた上で未知の文書を判定する側の実行部分から成る．さらに，学習部分と実行部分のそれぞれは，入力文書を扱うコーパス部分，文書から手掛かりの素性を抽出する素性抽出部分，機械学習を行う SVM の部分から成る．

本システムのコーパス部分では品詞タグつきコーパスを扱う．素性抽出部分では，Python 2.7.4 で書いた素性ベクトルリストを作成し，SVM の部分では LIBSVM 3-17(Chang and Lin (2001))を利用した．

学習部分の構成は下記の図 1 のとおりである．学習部分では，コンピュータに学習させようとする品詞タグつきのインドネシア語コーパスを，素性ベクトルのリストを作成するプログラムに送り，素性の抽出を行う．このプログラムにより作成された，素性ベクトル

†50011268509@st.tuat.ac.jp

のファイルから SVM によって，モデルファイルが生成される．モデルファイルは次に述べる実行部分で利用される．

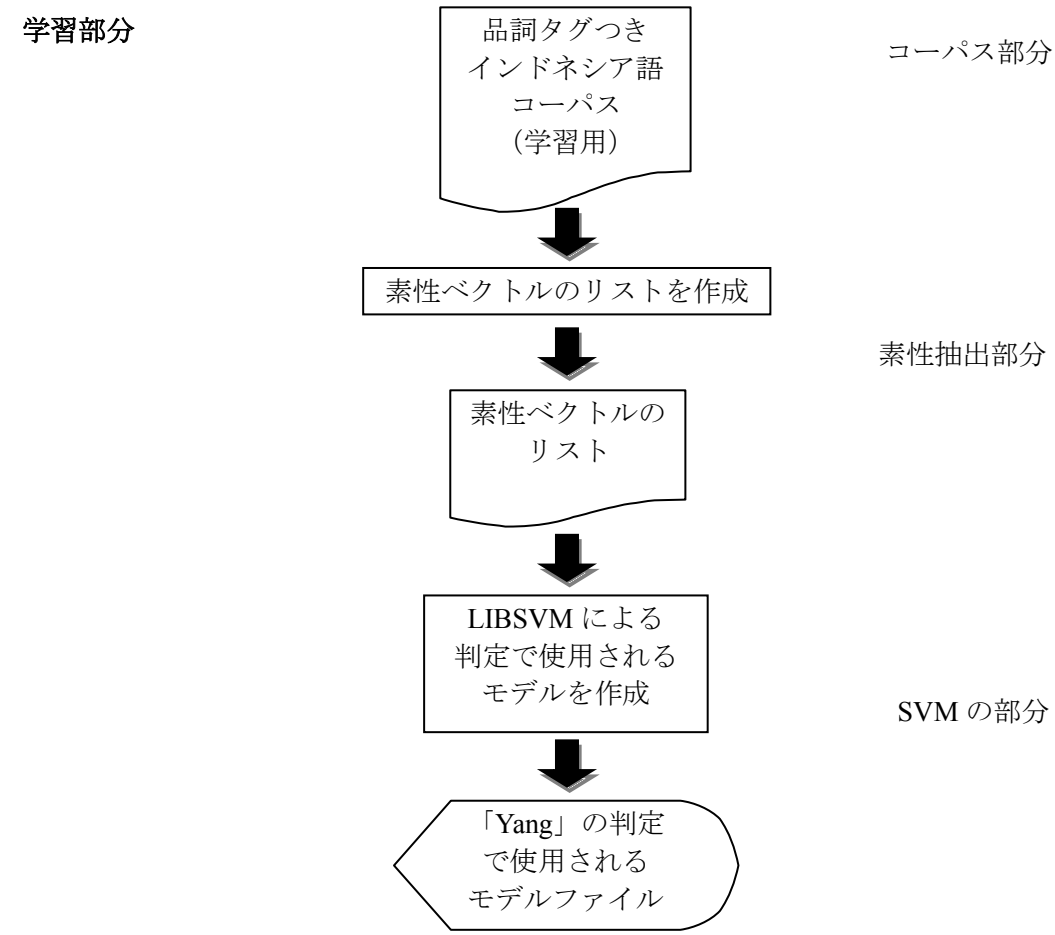

図 1　本システムの学習部分の構成

次に，実行部分では，判定のために，学習部分で使用されていない品詞タグつきコーパスを入力としている．学習の際と同様に，対象となるコーパスの素性抽出を行う．学習部分で生成されたモデルファイルと判定用の素性ベクトルリストのファイルに基づいて SVM による判定を行い，「Yang」の判定結果が得られる．本システムの実行部分は下記の図 2 のとおりである．

なお，本システムは「Yang」が名詞の直後に来ると想定しているため，コーパス中の名詞ごとに「Yang」の有無を判定した．

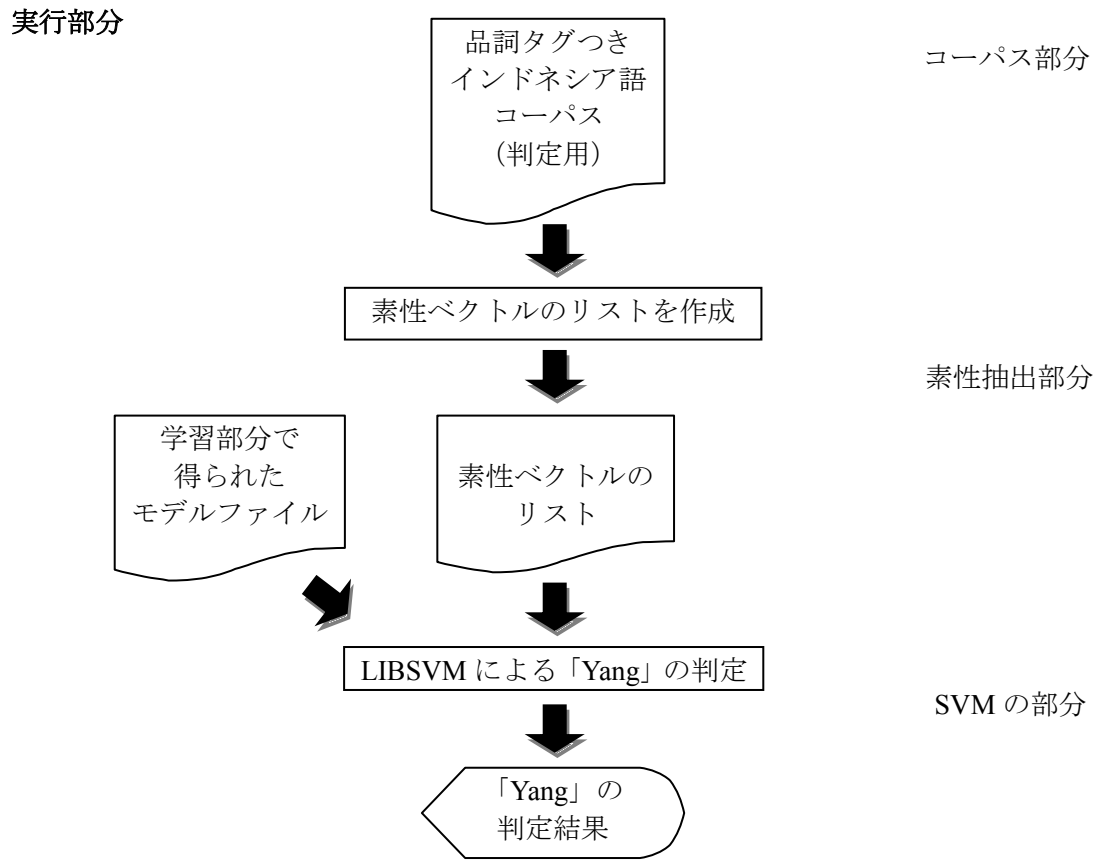

図 2　本システムの実行部分の構成

４．データ

本システムでは，One Million POS Tagged Corpus of Bahasa Indonesia[1]を利用した．このコーパスは, インドネシアの新聞から収集された 100 万単語のインドネシア語から成り，単語だけでなくその単語の品詞情報も付与されている．コーパスの例文を図 3 に示す．また，コーパス中の品詞の種類を表 1 に示す.

... pemegang/nnc saham/nnc publik/nnc ./. **Dan/nn** riil/nn **estat/nn** biasanya/jj tidak/neg diperdagangkan/nn dengan/in baik/jj dibawah/nn **kepemilikan/nn** publik/nnc ./. Salomon/nn **Brothers**/nn mengatakan/vbi ,/, Kami/prp yakin/nn **properti-properti/nn** riil/nn estat/nn akan/md ...

図 3　コーパスから取ってきた例文

[1] http://www.panl10n.net/english/OutputsIndonesia2.htm

表 1　コーパスで使用されている品詞の種類

タグ	説明	例
,	コンマ	, ;
.	文章の区切り	.
-	ダッシュ	-
SYM	記号	%
NN	普通名詞	Indeks, biaya, tenaga
PRP	人称代名詞	Kita, mereka, ia
PRN	数字代名詞	Satunya, keduanya, ketiganya
PRL	位置格代名詞	Sana, situ, sini
WRB	Wh 副詞	Apa, bagaimana, mengapa
WP	Wh 代名詞	Apa, apakah, apa-apa
VBI	自動詞	Ada, berakhir, berkata
VBT	他動詞	Membantu, menolak, menjadi
MD	助動詞	Akan, bias, telah
JJ	形容詞	Swasta, jauh, baik
CDP	基数	Satu, juta, milyar
CDO	順序	Pertama, kedua, ketiga
NEG	否定詞	Belum, bukan, tidak
IN	前置詞	Dengan, kepada, untuk
CC	等位接続詞	Atau, dan, karena
SC	従属接続詞	Bahwa, sekaligus, yang
RB	副詞	Hanya, mungkin, sebagaimana
DT	限定詞	Ini, para, tersebut
FW	洋語	Few, fiscal, for

また, 名詞には細分類がある．表 2 に細分類を示す.

表 2　名詞の細分類

タグ	説明	例
NNC	加算普通名詞	Cara, laut, tahap
NNU	不加算普通名詞	Peringatan, pikiran, system
NNG	属格普通名詞	Adanya, lainnya, misalnya
NNP	固有普通名詞	Desa, dunia, lembaga

５．　実験

本システムでは素性に単語の位置を考慮した．単語列 W：{A, B, C, D, E, F, G} とこれに対応する品詞 P：{p_A, p_B, p_C, p_D, p_E, p_F, p_G} があるとする．ここで，p_D を名詞とすると，「A B C D **Yang** E F G」という文章において，名詞の D の位置を 0，A の位置を-3，B の位置を-2，C の位置を-1，E の位置を+1，F の位置を+2，G の位置を+3 と置き, A, B, C, E, F, G の形態素と品詞を素性とした．なお,「Yang」の有無を判定するシステムであるため, 実際に E の位置に「Yang」があった場合でも素性には含まなかった．また,「Yang」が挿入される直前の名詞の形態素は素性に含めたが, 品詞は常に名詞であるため, 素性に含めなかった.

また，本システムでは関連性の低い単語を素性ベクトルに入れないために，文節（文の始まりまたは終わりやコンマの前または後）を超えた単語およびその品詞は素性に含めず,

それを超えた形態素は「NONE/none」として扱った．また，文頭と文尾は特殊な形態素として扱った．

本稿では，ふたつの実験を行った．一つ目は素性の種類を変えた実験である．具体的には，形態素だけを素性にした場合と，形態素に加えて品詞を素性にした場合の実験を行い比較した．

図 3 の例文を例に，形態素だけを利用した場合の「estat/nn」の素性ベクトルのパターンを表 3 に，形態素と品詞を利用した場合の「estat/nn」の素性ベクトルを表 4 に示す．

表 3　「estat/nn」の単語だけの素性ベクトル

「Yang」の有無	単語の位置						
	-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
ない	.	Dan	riil	estat	biasanya	tidak	diperdagangkan

表 4　「estat/nn」の単語と品詞が入った素性ベクトル

「Yang」の有無	単語の位置							品詞の位置					
	-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**
ない	.	Dan	riil	es tat	bia sa nya	ti dak	Diperda gangkan	bos	nn	nn	jj	neg	nn

二つ目の実験として，ウィンドウサイズを変えた実験を行った．連体従属接続詞「Yang」は，直後に長いフレーズが，また直前にはより短いフレーズが来る可能性が高い．そこで，ウィンドウサイズを変えて，本システムの正解率がどれぐらい変わるのかを確認した．この際，問題の名詞の位置より前の単語は 3 個に固定し，後ろの単語として 3 個と 5 個単語の両方を試した．

６．実験の結果

実験で使用したコーパスには，42,451 種類，計 451,339 個の名詞を含んでいた．そして，その中から「Yang」がつくものは 17,588 個であった．これは全体の 3.90%を占めているため，69.10%が最頻出ベースラインである．実験の結果は表 5 のとおりである．

表 5　実験の結果

実験		正解率 [%]
素性の種類	単語のみ	96.90
	単語と品詞	97.07
ウィンドウサイズ	前 3 後 3	97.07
	前 5 後 5	97.37

７．考察

単語と品詞の情報を素性ベクトルに入れた場合は単語のみを入れた場合より正解率が高いことが表 5 から分かる．すなわち，コンピュータに単語だけ覚えさせるのでなく，その単語の品詞も覚えさせたほうがより効果的であることが分かった．

表 5 より，ウィンドウサイズは 3 よりも 5 のときの方が正解率が高いことが見て取れる．この差は 0.30%であった．

次に，「Yang」がある際にシステムが間違って判定し 0 と出力した例を(1)と(2)に示す．

(1)「…naik pada *tingkat* **yang** jauh lebih cepat…」
「…より速いペースの段階に上がった…」
(2)「…dan *12 bulan* **yang** berakhir pada September 1988…」
「…1988 年 9 月に終わる 12 カ月や…」

上記の(1)と(2)では，「Yang」がなければ文章の構成が変わる他に，文章の意味もおかしくなるため，システムが 0 と出力する場合は，判定が間違っている．(1)における「tingkat（段階）」と(2)における「12 bulan（12 カ月）」の使用例が学習データになかったため，システムが「Yang」がないと判定したと考えられる．

コーパスを見ると，「Yang」が実際にはあるのにもかかわらず，システムはないと判定している．これは，学習データのうち，「Yang」があるデータが非常に少ないためであると考えられる．これは「oversampling」や「undersampling」によってデータの割合を変えることで改良の可能性があるが，（宋，浅原，古宮，小谷（2013））によれば，データの割合を変えない方が良い結果が出ている．

８．まとめ

本稿では，インドネシア語連体従属接続詞「Yang」の有無を判定するために，名詞の直前にある 3 個単語と名詞の直後にある「Yang」を含めず 5 個単語を素性ベクトルに入れる場合は正解率が最も高いと見られた．そして，ピリオードやコンマにより関連性の低い単語を素性ベクトルに入れないことの重要さが見られた．

文献

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

宋 東旭，浅原 正幸, 古宮 嘉那子，小谷 善行（2013），機械学習による中国語助詞の用法解析，第三回コーパス日本語学ワークショップ予稿集，pp. 111-116.

寶 梓瑜，古宮 嘉那子，小谷 善行（2012），コーパスを用いた中国語ネット語の判定システム，　第一回コーパス日本語学ワークショップ予稿集, pp.161-166.