サポートベクタマシンを用いたキーワードからの文生成

肥塚 真輔　岡本 紘幸　斎藤 博昭
慶應義塾大学 理工学部
Email: {shizuka,motch,hxs}@nak.ics.keio.ac.jp

1　はじめに

キーワードからの文生成とは，{ 国, 政策, 発足 } のようなキーワード集合を入力に受け，各キーワードに機能語付加などの整形を加えることで表層文を生成するタスクである．入力からは目的とする文の意味が取得不可能であることから，本タスクは文法的に正しく意味が理解可能な文の生成を目的とする．これは自然言語生成の基本タスク（text planning, sentence planning, realization）[1] のうち，主に realization に該当する．

自然言語生成は機械翻訳や対話システムなど，多くの自然言語処理アプリケーションの構成要素である．各々の自然言語処理アプリケーションによって入力形式など自然言語生成部への要求は異なるが，その中には語彙選択以降のタスクを自然言語生成部に要求するアプリケーションも考えられる上，発想支援などを目的とした情報の補完にも利用可能である．

2　先行研究

内元らはキーワードからの文生成を，順序付きキーワード集合 K，順序付き形態素集合 M，順序付き依存関係集合 D，テキスト T を用い，式 (1), (2) のように定式化した [2]．定式化に際し内元らは，テキストが与えられた時，キーワード集合を生成する品詞の列と係り受け関係は一意に決まると仮定している．

$$\begin{aligned} T_{best} &= \arg\max_{T} P(T|K) \\ &= \arg\max_{T}\{P(K|T)\times P(T)\} \qquad (1) \\ P(K|T) &\approx P(K,D,M|T) \\ &= P(K|D,M,T)\times P(D|M,T)\times P(M|T) \quad (2) \end{aligned}$$

式 (2) 右辺の確率モデルは左からそれぞれキーワード生成モデル，係り受けモデル，形態素モデルである．式 (2) により，キーワードからの文生成は，これらの確率モデルを用いて候補となる表層文の尤度を算出，順位付けるタスクに置き換えられる．候補文は各キーワードが含まれる文節をコーパス内から取り出して組み合わせることにより生成されるため，式 (1) の言語モデル $P(T)$ は無視される．

内元らは式 (2) 右辺の確率モデルを最大エントロピー法を用いて獲得した．これに対し本稿では，サポートベクタマシン（SVM）[3] を用いて確率モデルを獲得，利用するシステムを提案する．SVM はデータスパースネスに頑健な統計的機械学習法であり，質問応答の分野では最大エントロピー法より優れているという報告もなされている [4]．

3　システム概要

本システムは，大きく二つのモジュールにより構成されている．一つ目はキーワード集合を入力に受けて候補文の集合を出力する候補文生成部，二つ目は候補文の集合を受けてそれらを順位付けし，上位 λ 文を出力する評価部である．本システムの概要を図 1 に示す．

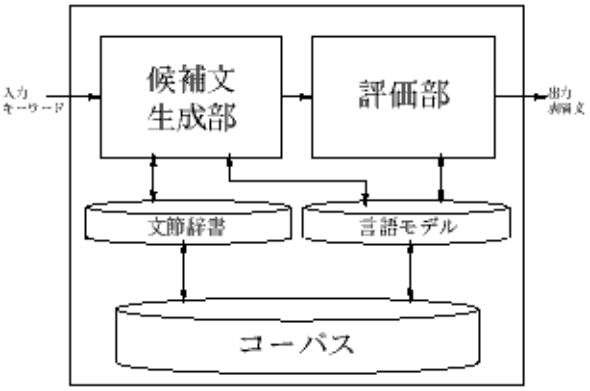

図 1: キーワードからの文生成システム概観

3.1　候補文生成部

本節では，本システム候補文生成部の処理を示す．

3.1.1　ステージ 1

図 2 は，本稿における日本語文法の定義を表現したものである．本稿ではキーワードを，各文節内の主辞単語と定義する．文節は，いくつかの内容語とそれに続くゼロ個以上の機能語により構成され，主辞単語は文節内で文末に一番近い内容語と定義する．内容語は，その品詞が動詞，形容詞，名詞，指示詞，副詞，接続詞，連体詞，感動詞，未定義語である形態素とし，それ以外の品詞である形態素を機能語とする．但し，サ変動詞，動詞「なる」，形式名詞「の」については，文節内に他の内容語が存在しない場合には内容語とみなす．品詞の体系は京都大学テキストコーパス version3.0[6] に従った．

キーワード ::= 主辞単語 1 | 主辞単語 2
主辞単語 1 ::= 内容語
主辞単語 2 ::= サ変動詞 | 動詞「なる」| 形式名詞「の」
内容語 ::= 動詞 | 形容詞 | 名詞 | 指示詞 | 副詞 | 接続詞 | 連体詞 | 感動詞 | 未定義語
機能語 ::= 判定詞 | 助動詞 | 助詞 | 接頭辞 | 接尾辞
文節 ::= 内容語* 主辞単語 1 機能語* | 主辞単語 2 機能語*
文 ::= 文節 文節 文節*

図 2: 本稿における日本語文法定義

この定義のもとに本システムは，入力として与えられたキーワード集合から候補文節を生成する．候補文節の生成は，コーパス中に現れた文節のうち，キーワードを主辞単語として含む全ての文節を抽出することで実現する．そのため，キーワードを見出しとして，それを主辞とする全ての候補文節を引くことが可能な文節辞書をあらかじめ用意する．

しかしながら，図 2 に示した文節の定義をそのまま適用すると，入力されたキーワード以外の内容語が含まれた候補文節も生成される．そこで本稿では，図 2 に示した文節を以下のように再定義した．

$$\text{文節} \quad ::= \quad \text{キーワード 機能語}^* \tag{3}$$

ステージ 1 において，システムは入力として与えられた各キーワードを文節辞書にて調べ，式 (3) を満たす全ての候補文節を抽出する．例えば，キーワード群として { 国, 政策, 発足 } が与えられた場合，システムは以下のような候補文節を獲得する．

- 国 → 国の，国側の，国からの，...
- 政策 → 政策に，政策などを，政策を，...
- 発足 → 発足した，発足させる，...

3.1.2 ステージ 2

システムは次に，ステージ 1 で獲得した候補文節をあらかじめ用意した確率モデルを用いて評価し，上位 ϵ 個を選択する．これは，ステージ 1 で獲得した全候補文節を用いた場合の組み合わせ爆発を回避するためである．確率モデルの詳細は 4 章に示す．

3.1.3 ステージ 3

続いてシステムは，とり得る全ての構文木を生成する．本稿では，各文節の間に依存関係があると仮定して日本語の係り受け関係を以下のように定義し，二分木で表現することとした．

- 係り受けは前方から後方に向いている（後方修飾）
- 係り受け関係は交差しない（非交差条件）
- 係り要素は受け要素を一つだけ持つ

生成される構文木の形態と数は，入力として与えられたキーワードの数により変化する．例えば入力されたキーワードが 3 個であった場合，生成される構文木は，キーワード k_j を含む文節 c_j を用い，$((c_1, c_2), c_3)$, $(c_1, (c_2, c_3))$ のように表される．

3.1.4 ステージ 4

システムは次に，ステージ 3 で生成した各構文木の葉をステージ 2 で選択した候補文節で置換することで，候補文を生成する．

3.2 評価部

最後にシステムは，あらかじめ用意された確率モデルを用い，候補文生成部が出力した候補文を評価する．確率モデルの詳細は，4 章に示す．

4 言語モデル

2 章で述べたように，内元らはキーワードからの文生成を，式 (1), (2) のように定式化している．しかし，係り受け関係が形態素の活用形にも依存することや，文の尤もらしさが文節区切りにも依存することが考えられることから，本稿では式 (2) に順序付き活用形集合 I，順序付き文節区切り集合 C を挿入，キーワードからの文生成を以下のように定式化した．

$$\begin{aligned} P(K|T) &\approx P(K, D, I, M, C|T) \\ &= P(K|D, I, M, C, T) \times P(D|I, M, C, T) \times \\ &\quad P(I|M, C, T) \times P(M|C, T) \times P(C|T) \end{aligned} \tag{4}$$

式 (4) 右辺に示された各確率モデルは左から，キーワード生成モデル，係り受けモデル，活用モデル，形態素モデル，文節生成モデルである．本システムはこれらの確率モデルのうち，文節生成モデルを候補文生成部の文節選択に利用，残り 4 つのモデルを評価部における表層文評価に利用した．

本稿では文節生成モデルの一部を除いたこれら全ての確率モデルを，TinySVM[5]（2 次多項式カーネル）を用いて学習した．本来 SVM は二値分類器であることから，分類の尤度としての確率は返さない．そこで本稿では，候補文を表す各事例ベクトルの超平面からの距離を最も超平面から遠い事例の距離で正規化し，それを分類の尤度確率とみなした．

4.1 キーワード生成モデル

キーワード生成モデルは，入力として与えられたキーワードが表層文中で「鍵」になる単語かどうかを，キーワードにかかる文節の格に注目して評価するモデルである．本稿では，T, C, M, I, D が与えられた時に K が与えられる確率を，各々のキーワード k_j $(1 \leq j \leq \kappa)$ がそれに係る文節のうち最も文末側の文節の末尾二単

語 w_l, w_{l-1} にのみ依存すると仮定し，以下のように定義した．

$$P(K|D,I,M,C,T) = \left(\prod_{j=1}^{\kappa} P(k_j|w_l, w_{l-1})\right)^{\frac{1}{\kappa}} \quad (5)$$

4.2 係り受けモデル

係り受けモデルは，システムが構文木として自動生成した文節間の係り受けが文として尤もらしいかを評価するモデルである．本稿では，T, C, M, I が与えられた時に D が与えられる確率を，各々の係り受け関係 d_j $(1 \leq j \leq \kappa - 1)$ が独立であると仮定し，以下のように定義した．

$$P(D|I,M,C,T) = \left(\prod_{j=1}^{\kappa-1} P(d_j|I,M,C,T)\right)^{\frac{1}{\kappa-1}} \quad (6)$$

4.3 活用モデル

活用モデルは，各形態素の活用形の並びが文として尤もらしいかを評価するモデルである．本稿では，T, C, M が与えられた時に I が与えられる確率を，各々の活用形 i_j $(1 \leq j \leq \iota)$ が後方に連接した形態素活用形にのみ依存すると仮定し，以下のように定義した．

$$P(I|M,C,T) = \left(\prod_{j=1}^{\iota} P(i_j|i_{j+1},M,C,T)\right)^{\frac{1}{\iota}} \quad (7)$$

4.4 形態素モデル

形態素モデルは，各形態素の品詞の並びが文として尤もらしいかを評価するモデルである．本稿では，T, C が与えられた時に M が得られる確率を，各形態素 m_j $(1 \leq j \leq \mu)$ が前方及び後方に連接した形態素にのみ依存すると仮定し，以下のように定義した．

$$P(M|C,T) = \left(\prod_{j=1}^{\mu} (P(m_j|m_{j-1},C,T) \times P(m_j|m_{j+1},C,T))\right)^{\frac{1}{\mu}} \quad (8)$$

4.5 文節生成モデル

文節生成モデルは，候補文生成部で文節辞書から獲得した文節の尤度を算出するモデルである．本稿では，j 番目の文節 c_j が尤もらしい文節となるには，c_j の末尾文字 t_j が後方に隣接した文節 c_{j+1} の末尾文字 t_{j+1} に対して尤もらしいことが必要であると仮定し，文節生成モデルを文節末尾文字 bigram として定義した．また，文末の文節 $(j = \kappa)$ の後方連接尤度は算出できないため，代わりに文末文節生成モデル $P_e(c_j|T)$ を用意した．これを以下に示す．

$$P(C|T) = \prod_{j=1}^{\kappa} P(c_j|T) \quad (9)$$

$$P(c_j|T) = \begin{cases} P_e(c_j|T) & if \quad j = \kappa \\ P(t_j|t_{j+1}) & otherwise \end{cases} \quad (10)$$

本モデルの主な目的は，以下のような非文の排除である．これらの文は連接文節の末尾文字の不自然な連続や不完全な文節，文中の文末表現の挿入，不完全な文末などの文法的非整合性を伴っている．

- 私は家は行く
- 私家に行って

5 実験・結果

システムの評価にあたり本稿では，独自に規定した以下の評価項目を全て満たす文を正解文と定義し，精度の算出および先行研究との比較を行った．

- 文末で文が完結していること（体言止め不可）
- 文の途中で完結していないこと
- 文節の意味役割に重複がないこと
- 存在すべき機能語が存在すること
- 2 文節の機能語の関係が自然であること
- 文意が理解可能であること

5.1 精度

本稿では，正解文がどの程度上位に集められているかを，生成された候補文集合における精度，R 精度，1 位の精度の比較により評価した．またベースラインとして，候補文節の全ての組み合わせにより生成された候補文集合の精度も算出した．情報検索タスクにおける R 精度の R は「全文書集合」中の適合文書数であるが，キーワードからの文生成タスクにおいては「全文集合」の再現が不可能なことから,「生成された候補文集合」をその代わりとみなした．

実験は，無作為に選択された文の 2 キーワード 60 通りの入力で，本稿で評価部に新たに提案した活用モデルの有無を区別して行った．使用したコーパスは京大コーパスの 1995 年 1 月 1 日および同月 3 日〜11 日分で，1 月 6 日分をテストデータとし，残りを学習データとした．

表 1 に，精度実験の結果を示す．結果より，本システムの候補文生成部が尤もらしい文を構成する文節を選択して文を生成していること，評価部が正解文をより上位に順位付けしていること，ならびに活用モデルの有効性が確認された．

5.2 先行研究との比較

本稿では精度に加え，先行研究である内元らの手法との比較を行った．内元らは出力文を意味的・文法的な観点で主観評価し，以下の 2 つの基準を用いてシステムを評価している．

a. 1 位の候補文が意味的・文法的に適切である
b. 上位 10 位に意味的・文法的に適切な候補文がある

実験は，京大コーパス 1 月 1 日分の記事に 10 回以上現れた主辞単語の集合から選択されたキーワードの 2 および 3 語の組み合わせ各 30 通りを，内元らの実験と同じ組み合わせで行った．確率モデル学習用のコーパスは，以下の表 2 の構成とした．

表 3 に，比較実験の結果を示す．結果より，本システムは内元らの手法と比較して基準 a（上位 1 位）の性能は劣るが，基準 b（上位 10 位）の性能は同等もしくは優れていることがわかった．

表 1: 精度実験の結果

項目	活用モデルなし	活用モデルあり
ベースライン 候補文集合の精度	0.11 0.28	
R 精度 1 位の精度	0.30 0.33	0.31 0.37

表 2: 比較実験の環境

項目	内元らのシステム	本システム セット 1	本システム セット 2
キーワード生成モデル	1 月 1 日		3〜11 日
係り受けモデル	1 月 1，3〜9 日	3〜11 日	
活用モデル	（モデルなし）	3〜11 日	
形態素モデル	1 月 1，3〜9 日	3〜11 日	
文節生成モデル	（モデルなし）	3〜11 日	
文節辞書	1 月 1，3〜16 日		

表 3: 比較実験の結果

キーワード数	評価基準	内元らのシステム	本システム セット 1	本システム セット 2
2	a	0.90	0.63	0.73
	b	0.96	0.96	0.96
3	a	0.63	0.50	0.53
	b	0.80	0.83	0.83

6 考察

本章では，実験結果について考察する．

6.1 精度

実験より，本システムは正解文を上位に出力していることが確認できたが，精度には改善の余地がある．精度低下の原因としてまず考えられるのは，文節辞書のデータ不足である．個別の結果を見ると，文節辞書から候補文節が得られなかった場合が 11 通り，文節生成モデルの閾値を越える評価の候補文節が得られなかった場合が 6 通り，候補文は生成されたが正解文は生成されなかった場合が 8 通りと，候補文生成で精度を落していることが分かる．このことから，本システムは文節辞書の改善により，精度向上が見込める．尚，正解文が生成された場合のみの精度は，候補文集合の精度 0.49，R 精度 0.54，1 位の精度 0.65 であった．

文節辞書生成の改善手法としては，本システムの単独利用の場合はキーワードを品詞などの構文情報でクラスタリングして不足した文節データを補う手法が，sentence planner と連結する場合はそちらでの意味情報を考慮した文節データの補完が考えられる．

精度低下の原因として次に考えられるのは，文節生成モデルである．出力文の上位には「疑いを逮捕した」のような，各文節の意味役割が誤った文が多く生成された．これは，文節生成モデルの文字 bigram が文節を表層的な尤度でしか評価せず，文節間の意味的関係を考慮しないことに原因があると考えられる．

この問題の解決には，キーワードの意味的関係をシソーラスなどの利用で獲得し，誤った意味役割の候補文節を排除する手法が考えられる．しかしながら，キーワード集合という限られた入力から語彙間関係を決定することは困難であることが予想される．sentence planner との接続を考えるならば，そちらでキーワードに意味役割を付与し，生成される文節に制限をかけることは可能である．

6.2 先行研究との比較

実験により，本システムは内元らの手法と比較して，基準 a の性能に劣り，基準 b の性能は同程度もしくはそれ以上であることがわかった．

基準 a が劣る原因としては，6.1 節で述べた文節の意味役割の誤りを評価部が補正できなかったことにあり，特にキーワードの格との結び付きを重視して学習しているキーワード生成モデルに改善の余地がある．

また，基準 b の性能が同程度となったのは，内元らのシステムに定義されていない文節生成モデルが表層的な特徴を重視して学習し，文として尤もらしい文字列を上位に順位付けたためと考えられる．

7 おわりに

本稿では，キーワードからの文生成，特に文法的に整合性のある文の生成について，コーパスに記載された文の表層的な情報を統計的に用いる手法の有効性を示した．今後取り組むべき課題として，入力キーワードの意味情報に基づいた候補文生成や順位付け，意味表現を入力に受ける自然言語生成システムへの拡張などが挙げられる．

参考文献

[1] Owen C. Rambow, and Tanya Korelsky: Applied text generation, in *Proceedings of Conference on Applied Natural Language Processing*, pp. 40–47, 1992.

[2] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara: Text generation from keywords. in *Proceedings of the Main Conference on COLING*, 2002.

[3] Vladimir N. Vapnik: *The Nature of Statistical Learning Theory*. Springer, 2nd edition, 1999.

[4] Jun Suzuki, Yutaka Sasaki and Eisaku Maeda: Svm Answer selection for open-domain question answering. in *Proceedings of the Main Conference on COLING*, 2002.

[5] 工藤 拓: *TinySVM: Support Vector Machines*, http://cl.aist-nara.ac.jp/˜taku-ku/software/TinySVM/, 2002.

[6] 黒橋 禎夫, 長尾 眞: 京都大学テキストコーパス・プロジェクト. 人工知能学会 第 11 回全国大会, pp. 58–61, 1997.