E-018

漸進的な言語処理のための残存文長の推定
Estimation of Remaining Sentence Length for Incremental Language Processing

河村 天暉[†,a)]　大野 誠寛[‡,b)]　松原 茂樹[‡]
Takaki Kawamura　Tomohiro Ohno　Shigeki Matsubara

1　はじめに

同時通訳や字幕生成，入力予測などのリアルタイム音声言語システムでは，入力に対して漸進的に処理を行う必要があり，処理の正確さを保ちつつ，遅延時間を抑えることが求められる．このような処理を実現するにあたり，意味的なまとまりをもつ文が今後どれだけ続くかという情報は，重要な手がかりとなりうる．例えば，文がもう少しで終わることが分かれば，処理の正確さを保つため，同時通訳における訳出タイミング [1] や，読みやすい字幕とするための改行挿入タイミング [2] を遅らせるという判断が可能となる．しかし，管見の限り，文の残りの長さ（以下，残存文長）を推定する研究はない．

そこで本稿では，文節が入力されるごとに残存文長を推定する手法を提案する．提案手法は，入力済みの文字列から得られる情報をもとに，残存文長が，1 文節，2〜3 文節，4 文節以上，のいずれであるかを判定する．

2　残存文長

本研究では，文全体が一度に入力されるのではなく，文頭から徐々に 1 文の要素が入力される状況を前提とし，文節が入力されるごとに残存文長を推定することを試みる．本節では，残存文長，及び，平均残存文長を定義する．

2.1　残存文長の定義

本研究では，文 s が n_s 個の文節から成り，文頭から x 番目の文節まで既に入力されているとき（すなわち，既入力文節数が x であるとき）の残存文長 $RL(s, x)$ を，式 (1) により定義する．なお，文長は文節単位で計測するものとする．

$$RL(s, x) = n_s - x \tag{1}$$

2.2　平均残存文長

残存文長の推定において手がかりとなりうる統計量として，残存文長の平均（以下，平均残存文長）を使用する．平均残存文長 $ARL(\mathbf{S}, x)$ は，文の集合 $\mathbf{S}$ に含まれる各文が文頭から x 番目の文節まで入力されているときの残存文長の平均であり，式 (2) より算出されるものとする．なお，この値は残存文長の期待値ともみなせる．

$$ARL(\mathbf{S}, x) = \frac{\sum_{s' \in \{s | n_s > x, s \in \mathbf{S}\}} RL(s', x)}{|\{s | n_s > x, s \in \mathbf{S}\}|} \tag{2}$$

3　残存文長の推定

提案手法では，SVM（Support Vector Machine）を用いて，1 文を構成する文節が文頭から 1 つ入力されるごとに残存文長が 1 文節，2〜3 文節，4 文節以上の 3 クラスのいずれであるかを分類する．2〜3 文節のクラスは，節の平均長（2.60 文節 [3]）を考慮して設けたものであり，文の残りが最終節だけであることの判別を意図したものである．なお，入力は形態素解析，文節まとめあげ，漸進的係り受け解析 [3] が施されているものとする．以下では，SVM で用いる素性について，図 1 に例を示しつつ説明する．

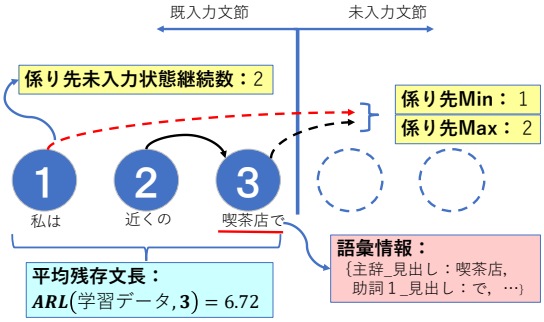

図 1　SVM で用いる各素性の例

平均残存文長（以下，ARL）

残存文長の期待値として，平均残存文長を素性として用いる．図 1 の場合，既入力文節数 $x = 3$ であるときの平均残存文長（6.72）を素性として用いる．ただし，実際には 0 から 1 の間でスケーリングした値を使用する．

語彙情報（以下，LEX）

言語情報のうち語彙レベルの情報として，既入力文節列の最終文節（図 1 では文節「喫茶店で」）がどのような文節であるかの情報を素性として用いる．具体的には，当該最終文節から得られる，内元らの論文 [4] における素性番号 1〜15 の情報[*1]（例：主辞形態素などいくつかの形態素の見出し・品詞・活用など）を用いる．ただし，これらの素性値には出現頻度が 3 以上のものを採用し，また主辞品詞の数詞，固有名詞については抽象化したものを使用する．

構文情報（以下，SYN）

言語情報のうちの構文レベルの情報として，既入力文節列に対する漸進的係り受け解析 [5] の出力結果（図 1 の矢印で表された係り受け構造）から得られる 3 つの情報を素性として用いる．1 つ目は係り先が未だ入力されていない文節（図 1 では「私は」と「喫茶店で」）の係り先となりうる文節の数の最大値であり，これらの文節が互いに異なる文節に係るとした場合の数（図 1 では 2）である．2 つ目は，その最小値であり，係り先が未入力の文節が互いに同一の格助詞を持っている場合のみ，それらは異なる文節に係るとした場合の数（図 1 では 1）である．これら 2 つの素性は，今後出現する必要のある文節数を意味するものである．3 つ目は，係り先が未入力である文節のうち，最も文頭に近い文節（図 1 では「私は」）と現在の入力文節（図 1 では「喫茶店で」）との距離（図 1 では 2）を用いる．これは，係り先が未入力の状態が長く継

*1　括弧については括弧の有無のみを素性として使用した

† 東京電機大学大学院未来科学研究科, Graduate School of Science and Technology for Future Life, Tokyo Denki University.
‡ 名古屋大学情報連携総括本部, Information and Communications, Nagoya University.
a) 18fmi07@ms.dendai.ac.jp
b) ohno@mail.dendai.ac.jp

表 2　実験結果

Class	1 文節			2〜3 文節			4 文節以上			ALL
Metrics	P(%)	R(%)	F	P(%)	R(%)	F	P(%)	R(%)	F	A(%)
Chance Rate	11.42 (269/2,420)	11.74 (269/2,291)	11.42	21.19 (978/4,615)	22.39 (978/4,368)	21.77	67.67 (9,173/13,556)	65.83 (9,173/13,934)	66.74	50.60
ARL	15.92 (975/6,124)	42.56 (975/2,291)	23.17	28.33 (393/1,387)	9.00 (393/4,368)	13.66	73.61 (9,630/13,082)	69.11 (9,630/13,934)	71.29	53.41
ARL +LEX	26.12 (1,403/5,318)	61.24 (1,403/2,291)	36.62	28.80 (1,365/4,793)	31.25 (1,365/4,368)	29.98	78.91 (8,271/10,482)	59.36 (8,271/13,934)	67.75	53.61
ARL +LEX +SYN	26.66 (1,511/5,667)	65.95 (1,511/2,291)	37.97	30.36 (1,321/4,394)	30.24 (1,321/4,368)	30.15	79.17 (8,338/10,532)	59.84 (8,338/13,934)	68.16	54.25

表 1　学習データにおける残存文長の割合

1 文節	2〜3 文節	4 文節以上
11.75% (34,612/294,285)	22.40% (65,971/294,285)	65.82% (193,702/294,285)

続するにつれて，ワーキングメモリなどの影響から，係り先が入力される可能性が高まると考えられ，文末が近づいていることを示唆すると考えたためである．なお，いずれの値も実際には，0 から 1 の間でスケーリングした値を使用する．

4　評価実験

提案手法の有効性を確認するために，新聞記事文を用いて残存文長の推定実験を実施した．なお，同時通訳などの音声言語システムへの応用を考える場合は，話し言葉特有の非文法的な現象等にも対処する必要があるが，本稿では，まずは整った文を対象として検証する．

4.1　実験概要

実験には，京都大学テキストコーパス Ver.4.0[*2]（毎日新聞 95 年 1 月 1 日から 17 日までの全記事と，1 月から 12 月までの社説記事）のうち，文長が 1 であるものを除いた 38,115 文を使用する．それらのデータのうち，1 月 4 日，1 月 5 日の全記事 2,291 文をテストデータ，1 月 9 日の全記事 1,212 文を開発データ，残りの 34,612 文を学習データとして用いた．また，ARL の算出には本実験における学習データを用いた．

提案手法 [ARL+LEX+SYN] との性能比較のために，以下の 3 つの比較手法を用意した．

比較手法 [ChanceRate]：学習データより，各入力文節における 1 文の残存文長を計測し（表 1 に示す），その割合に従ってランダムに 3 クラスを出力する．

比較手法 [ARL]：SVM の素性として言語情報を用いず，ARL のみを用いる．その他は提案手法と同一である．

比較手法 [ARL+LEX]：SVM の素性として SYN を用いず，ARL と LEX のみを用いる．その他は提案手法と同一である．

SVM のツールには scikit-learn[*3]を用い，カーネル関数は RBF を使用した．また本研究の問題設定では，残存文長の割合の偏りにより学習データが不均衡データとなるため，学習データに各クラスの頻度割合に基づいて重み付けを行い学習した．なお，クラス分類方式については oneVSrest 方式を採用した．さらに，SVM におけるコストパラメータ C については，手法ごとに開発データを用いてグリッドサーチ（$2^{-4} \leq x \leq 2^{15}$）を行い [6]，比較手法 [ARL] は 2^{14}，比較手法 [ARL+LEX] は 2^{12}，提案手法は 2^{14} とした．

評価では，各クラスの判別における適合率（P），再現率（R），F 値（F），及び，全判定に対する正解率（A）を測定する．

*2　http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス
*3　http://scikit-learn.org/stable

4.2　実験結果

実験結果を表 2 に示す．正解率に着目すると，提案手法が最も高い値を達成しており，提案手法の有効性が確認できる．一方，F 値に着目すると，1 文節や 2〜3 文節のクラスでは提案手法が最も高いが，4 文節以上のクラスでは比較手法 [ARL] の F 値が最も高く，提案手法は次点となった．適合率の分母をみると，比較手法 [ARL] は残存文長を 4 文節以上と推定する傾向が強いことがわかる．正解データにおいて，残存文長が 4 文節以上となる割合は全体の 7 割近くを占めており，4 文節以上と推定する割合が高いほど，このクラスでの F 値が高まる傾向があることが影響したものと考えられる．

5　おわりに

本論文では，文節が入力されるごとに残存文長が 1 文節，2〜3 文節，4 文節以上のいずれであるかを，SVM を用いて判定する手法を提案した．提案手法は，平均残存文長と語彙情報，構文情報を組み合わせた素性を用いており，評価実験の結果，残存文長の予測における各素性の有効性を確認した．今後は，節情報の使用などによる精度向上，また話し言葉への対応に取り組みたい．

謝辞　本研究は一部，科研費基盤研究 (C)（No. 16K00300）により実施した．

参考文献

[1] 笠ら，“英日同時翻訳のための依存構造に基づく訳文生成手法，”信学論, Vol. J92-D, No. 6, pp. 921–933, 2009.
[2] 村田ら，“読みやすい字幕生成のための講演テキストへの改行挿入，”信学論, Vol. J92-D, No.9, pp. 1621–1631, 2009.
[3] 丸山ら，“日本語節境界検出プログラム CBAP の開発と評価，”自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.
[4] 内元ら，“最大エントロピー法に基づくモデルを用いた日本語係り受け解析，”情処学論, Vol. 40, No. 9, pp. 3397–3407, 1999.
[5] 大野ら，“文節間の依存・非依存を同定する漸進的係り受け解析，”信学論, Vol. J98-D, No. 4, pp. 709–718, 2015.
[6] 廣川，“文単位の有価証券報告書分析による利益伸び率の予測，”信学技報, Vol. 113, No. 213, pp. 77–82, 2013.