社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (30)
2008－SLP－70 (30)
2008／2／9

統計学習を用いた対話からの非語彙的表現の抽出

牧本慎平† 吉川哲史† 柏岡秀紀†‡§ ニック キャンベル†‡§
†奈良先端科学技術大学院大学 情報科学研究科
‡情報通信研究機構 知識創成コミュニケーション研究センター
§国際電気通信基礎技術研究所 音声言語コミュニケーション研究所
{shimpei-m, satoshi-yo, kashioka, nick}@is.naist.jp

あらまし 音声対話を処理する上で，フィラーや言い淀みなどの非語彙的な表現の出現箇所を同定・抽出するというタスクは重要な課題であるが，それらは非常に多くのバリエーションを持つため，困難な問題とされてきた．そこで，対話書き起こしを一連の系列とした系列ラベリング問題として定式化し，サポートベクターマシンと条件付確率場の２つの学習器を適用した．文字情報や形態素解析による品詞の n-best 解，相手話者の状態などを素性として用いて評価実験を行ない，学習アルゴリズム毎や素性毎の性能の比較を行なった．

Extraction of Non-lexical Expressions in Dialogues using Statistical Learning

Shimpei Makimoto† Satoshi Yoshikawa† Hideki Kashioka†‡§ Nick Campbell†‡§
†Nara Institute of Science and Technology
‡National Institute of Information and Communications Technology
§Advanced Telecommunications Research Institute International
{shimpei-m, satoshi-yo, kashioka, nick}@is.naist.jp

Abstract In processing spoken dialogues, extraction of non-lexical expressions (eg. fillers, disfluencies and so on) is an important challenge. However, the large number of non-lexical expressions make the challenges difficult. We formalize the problem as a sequential labeling problem, and adapt two learning algorithms; Support Vector Machines and Conditional Random Fields. We use character information, n-best results of a morphological analyzer, and the partner speaker's speech activities as the features for machine learning. We draw comparisons among learning algorithms and adapted features in extraction of non-lexical expressions.

1 はじめに

1.1 背景と目的

話し言葉にはフィラーや言い淀み，笑いなどの書き言葉にはない非語彙的な表現が現れる．これらは一般的な音声言語処理のアプリケーションにおいてはノイズとして処理されているが，実際の人間の言語活動においては，対話を円滑にするための重要な要素となりうる．

そのため，そのような非語彙的表現を同定・抽出し，その情報を用いることで，音声認識の精度向上や音声自動翻訳などのアプリケーションの機能向上を図ることができると考えられる．また，未知の音声リソースからそれらの表現を獲得することによっ

て，より人間らしい表現豊かな音声合成や自然な対話の中での発話意図認識 [13] などの実現に寄与できるものと考えられる．

そこで本稿では，特に音声対話の書き起こしテキストに対する，統計的機械学習に基く非語彙的な表現の抽出手法を提案する．

提案手法は，非語彙的表現抽出を系列ラベリング問題として定式化し，教師あり統計的機械学習の枠組みによって学習を行なった．学習器としては，近年様々な分野で応用されている，サポートベクターマシンと条件付確率場の 2 つを適用し，性能を比較した．対象データは，音声対話コーパス ESP_C を用いている．

1.2 関連研究

従来，フィラーや言い淀みなどの表現に対する処理として，予め想定した表現を辞書に登録するなどと言った方法が取られてきた．しかしながら，この方法では，未知のデータに出現する新しい表現に対応することができない．そこで，近年では統計的機械学習による，それら表現の抽出が提案されてきている．

Liu ら [8] は英語による対話および独話の文境界と言い淀み箇所の機械学習ベースの同定による音声認識精度の向上を試みた．この研究では，音響的特徴なども素性として使用しており，書き起こしテキストから抽出を行なうとしている我々の研究とは異なっている．

Asahara and Matsumoto[4] は日本語話し言葉コーパスに付せられたフィラータグ，言い淀みタグをサポートベクターマシンを用いて同定する手法を提案した．これは，形態素解析器を冗長的に使用して得た n-best 解を素性として組み込むことによって性能を向上させている．

2 音声対話コーパス

2.1 ESP_C コーパス

我々は **ESP_C** コーパス[5] に対し，人手で非語彙的表現情報を付与した．ESP_C コーパスは JST/CREST の「表現豊かな発話音声のコンピュータ処理システム」プロジェクト [1] によって作成された JST/ATR Expressive Speech Processing Corpora のサブセットであり，2 人の話者の電話による音声対話が収録されている．対話は日本語により行なわれており，内容に制限はない．1 セッション 30 分の単位で，それぞれのペアで約 10 セッションの収録が行なわれた．話者は全 10 名で，母語が日本語の者が 6 名，英語の者が 2 名，中国語の者が 2 名である．本稿では日本語が母語の話者で収録されたデータを対象とした．

```
JFA-JFB-322.625-1.754 うーんだからすごく若い
JFB-JFA-324.679-1.104 あそうハハハ
JFA-JFB-324.735-3.498 エハハハハハ
JFB-JFA-326.240-2.547 じゃ話の内容も若返るんじゃないですか
JFA-JFB-328.298-0.548 ヒーハ
JFA-JFB-328.861-1.337 ハーちょっとエネルギーを
JFA-JFB-330.210-0.478 アハア
JFB-JFA-330.423-0.933 ハハハ
JFA-JFB-330.708-0.206 ハッ
JFA-JFB-330.925-1.074 いただこうかと
```

図 1 An example for ESP_C transcriptions

図 1 は ESP_C コーパスの一部である．コーパスには発話話者，発話開始時間，発話時間が含まれる．

2.2 アノテーション

本稿で用いたアノテーションのスキームは NIST Rich Text Evaluation Project[3] の Metadata Extraction (MDE) のタスクにて使用されている Simple Metadata Annotation Specification[9] を元に作成した．

MDE では，フィラーや言い淀みの抽出・訂正，節境界の抽出を行なうタスクであり，我々のコーパスにも同等の情報を付した．

我々は 5 話者 4 セッションの対話書き起こしに対し人手でラベル付けを行なった．結果，4,107 トークン，1,111 タイプの非語彙的表現に対しラベル付けした．ラベル付けしたコーパス内で，出現頻度 1 だったトークンの数は 843，出現頻度 2 だったもの

は 93 となっており，低頻度な表現のバリエーションが多くなっている．

3 系列ラベリングと統計的機械学習

系列ラベリング (sequential labeling) とは，観測系列 (observed sequence) $\boldsymbol{X} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_l\}$ とそれに対応した隠れ変数であるラベル系列 (label sequence) $\boldsymbol{Y} = \{y_0, y_1, ..., y_l\}$ の組み合わせ $(\boldsymbol{X}, \boldsymbol{Y})$ から学習を行ない，未知の観測系列 $\boldsymbol{X}^+$ が入力された時，最適なラベル系列 $\boldsymbol{Y}^*$ を出力する手法である．主に自然言語処理やバイオインフォマティクスなどの分野で広く用いられている．

本稿では，対話の書き起こしを観測系列とした系列ラベリング問題として，非語彙的表現抽出を行う．学習器としては，品詞タグ付けや固有表現抽出など様々なアプリケーションで高い性能を示しているサポートベクターマシンと条件付確率場を適用した．以下に，各学習器について概説する．

3.1 サポートベクターマシン

サポートベクターマシン (support vector machines, 以下 SVMs)[11] はマージン最大化による決定的な 2 値分類器である．SVMs は n 次元の素性ベクトル $\boldsymbol{x}_t$ と 2 値のラベル $y_t = \{+1, -1\}$ のペア $(\boldsymbol{x}_t, y_t)$ で表現される訓練事例について，正負のラベルを分離する超平面 $\boldsymbol{w} \cdot \boldsymbol{x} + b$ (ここで $\boldsymbol{w}, \boldsymbol{x} \in R^n$, $b \in R$) を求める分類器である．図 2 のように，訓練事例の正例・負例を正しく分類する超平面のなかから，分離超平面とその分離超平面に近い事例のマージンが最大化されるものを求めることによってモデル化を行なう．未知の事例 $\boldsymbol{x}^+$ について識別を行なう場合は，事例とモデル化された分離超平面との位置から求められる．

$$f(\boldsymbol{x}^+) = \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x}^+ + b)$$

SVMs 自体は汎用な分類器であるが，推定すべきラベルの前後のトークンの観測変数やラベル情報などを素性として用いる memory-based parsing[6] のアプローチを適用することによって，系列ラベリング問題に応用することが可能となる．

本稿では，2 次の多項式カーネルを用い，Pairwise 法によってマルチクラス問題に対応させた．

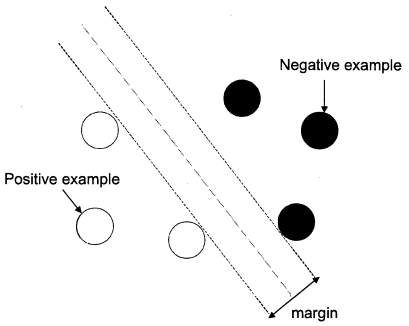

図 2 Margin Maximization

3.2 条件付確率場

条件付確率場 (conditional random fields, 以下 CRFs)[7] は，系列データのセグメンテーションとラベリングのために提案された確率モデルの枠組みであり，テキストチャンキングや日本語形態素解析など広く自然言語処理のタクスで適応され，いずれも高い精度を示している．条件付分布 $p(\boldsymbol{Y}|\boldsymbol{X})$ を得るために，$p(\boldsymbol{X}, \boldsymbol{Y})$ を最初に求める必要がある Hidden Markov Models (HMMs) に代表される生成モデル (generative models) とは異なり，CRFs は条件付分布 $p(\boldsymbol{Y}|\boldsymbol{X})$ の直接的推定を行なう識別モデル (discriminative models) であり，柔軟な素性設計ができるなどといった利点がある．

前節で述べた SVMs がウィンドウ幅内の局地的な情報を用いて各トークンで個別に分類を行なうのに対し，CRFs は入力された観測系列 $\boldsymbol{X}$ について，最尤なラベル系列 $\boldsymbol{Y}$ を出力する．

CRFs は以下のような指数モデルから条件付確率分布 $p(\boldsymbol{Y}|\boldsymbol{X})$ を表現する．

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp(W \cdot \boldsymbol{F}(\boldsymbol{X}, \boldsymbol{Y}))}{\sum_{\tilde{Y}} \exp(W \cdot \boldsymbol{F}(\boldsymbol{X}, \tilde{\boldsymbol{Y}}))}$$

ここで，W はモデルパラメータ，$\tilde{\boldsymbol{Y}}$ は全ての可能なラベル系列である．$\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{Y})$ は素性ベクトルであり，l 個の素性が $\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{Y}) = \{f_i(\boldsymbol{X}, \boldsymbol{Y}), f_i(\boldsymbol{X}, \boldsymbol{Y}), ..., f_l(\boldsymbol{X}, \boldsymbol{Y})\}$ という形で含まれている．適切な W を学習することによって，

新たな入力に対し，$P(\boldsymbol{Y}|\boldsymbol{X}))$ を最大化するような $\boldsymbol{Y}$ を求めることにより，ラベルの推定が可能となる．

我々が今回採用した Linear-chain CRFs は，Lafferty らの原論文 [7] にある最も単純なモデルであり，無向グラフィカルモデルとなっている．素性は，対応する観測変数と，ラベル変数の組上に定義される観測素性と隣り合ったラベル変数の組上に定義される遷移素性からなっている．

4 提案手法

本稿で提案する手法は，統計的機械学習の枠組みを用いて対話内の非語彙的表現の抽出を行なうものである．

非語彙的表現の抽出は，固有表現抽出や品詞タグ付けなどと同様のチャンキング問題として捉えることができる．本稿では，チャンクが一意に定まるように **IOB2** チャンクラベルセット [10] を採用した．このチャンクラベルセットを，隠されたラベル系列とするラベリング問題として定式化している．IOB2 ラベルセットでは，チャンクの先頭では **B**(eginning)，チャンクの先頭以外の部分では **I**(nside)，チャンクの外では **O**(utside) のラベルが用いられ，一意のチャンクを定義することができる．

教師あり機械学習の枠組みでは，学習を行なう上で，観測情報 (ここでは対話の書き起こし) から得られる素性セットを選択する必要がある．我々は，文字情報，形態素情報，相手話者情報の 3 種類の情報を用いた．以下にそれぞれの詳細を述べる．

文字情報 (CHARA)　まず，一番基本的な情報として，文字そのものから得られる情報を用いた．用いた情報は文字そのもの，文字のクラス (ひらがな，カタカナ，その他)，文字の対応する母音 (a,i,u,e,o,nn)，子音 (a,k,s,...) である．窓幅を前後 4 文字と設定し，推定を行なう文字を含め全 9 文字の情報を素性として用いた．尚，ひらがなやカタカナで記述されていない文字の母音・子音については推定せずに特殊なクラスを用いた．

形態素情報 (MORPH)　非語彙的表現抽出を行なう上で，予め辞書に登録されている語彙や形態素の情報は有効であると考えられる．本研究では，Asahara and Matsumoto[4] の手法で用いられた，形態素解析による冗長な形跡による n-best 解を素性に組み込む手法を導入した．形態素解析器としては IPAdic version 2.7.0 が組み込まれた *MeCab version 0.9.6*[2] を用い，素性に用いる解の数を $n = 3$ とした．尚，MeCab による未知語の品詞推定の機能は用いなかった．

相手話者情報 (SPEAK)　音声対話においては発話に相手話者の状態が影響されると考えられる．そこで，本稿では相手話者の状態を素性として組み入れた．相手話者の状態を示す情報として，推定を行なう時点での相手話者が発話をしているかどうかを用いる．相手話者の状態を発話開始，発話中，発話終了，発話なし，発話引き継ぎ (発話開始時には相手の発話が終了しておらず，発話終了時には相手の新しい発話が開始している状態) の 5 つのクラスに分類し，その状態を素性として用いた．

系列ラベリングでは，観測系列の入力の単位を決める必要がある．一般に，音声発話には書き言葉にあるようなセンテンスの概念は存在しないが，本稿では，一定の発話のない時間 (ポーズ) で区切られた発話を単位して入力に用いた．本稿では，実験的に定めた 2,000ms. のポーズを発話を区切る指標とした．

5 評価実験

提案手法の有効性を確認するために評価実験を行なった．行なった実験は学習器ごとの性能の比較と各素性毎の性能の比較である．

5.1 ベースライン

性能を評価する上でのベースラインとして，形態素解析器による推定結果を用いた．本稿で抽出の対象とする非語彙的表現が，一般に辞書などに登録されていない未知語として扱われ得ることを考慮に入れ，MeCab によって未知語として推定された箇所

を非語彙的表現であると考えた．また，フィラーおよび感動詞として辞書登録されており，それらの品詞として推定されたものについても，同様に非語彙的表現と捉え，これらの結果を正解ラベルと比較して性能を評価した．

5.2 評価尺度

非語彙的表現抽出の性能評価の尺度として，情報検索などの分野で用いられる F 値 (F-measures) を用いる．

$$P(\text{recision}) = \frac{\text{推定が成功したラベルの数}}{\text{推定したラベルの数}}$$

$$R(\text{ecall}) = \frac{\text{推定が成功したラベルの数}}{\text{正解ラベルの数}}$$

$$F_{\beta=1} = \frac{2P \times R}{P + R}$$

5.3 実験設定

評価実験は，学習器の性能比較と使用素性セット毎の性能比較を行なった．

学習器毎の比較では，学習器として SVMs と CRFs をそれぞれ用いた時の性能を比較し，使用素性セット毎の比較では，文字情報，形態素情報，相手話者情報の順で素性を追加して性能を比較した．

対象となったコーパスは，人手でラベル付けされた ESP_C コーパス 4 セッション (話者数 5，合計約 120 分) である．

ラベル付けされたコーパスに対し 4 分割交差検定を行ない，各ラベルについての評価を求めた．

5.4 実験結果

以下に実験を行なった結果を示す．

5.4.1 学習器毎の比較

4 節で述べた全ての素性を使用した際の SVMs と CRFs での非語彙的表現抽出の性能を比較した．その結果を表 1 に示す．

この結果から，両学習器ともベースラインを越えた性能を示した．両者に共通する特徴は，どちらも高い精度を持ち，再現率は精度と比較して低くなっていることである．全般的に SVMs の方が CRFs よりも非語彙的表現抽出の性能が高いことが分かる．

表 1 Comparison among Learning Algorithms

	B labels			I labels		
	P	R	F	P	R	F
baseline	32.8	53.1	40.6	57.6	67.2	62.0
SVMs	83.1	69.9	**75.9**	85.7	80.4	**83.0**
CRFs	78.1	66.3	71.7	84.6	75.6	79.9

5.4.2 使用素性セット毎の比較

続いて，学習に用いた素性を変更することによる性能の変化を調べた．

表 2 Comparison among Adapted Features

	SVMs		CRFs	
	B	I	B	I
CHARA	73.7	81.4	63.9	75.8
+ MORPH	75.2	82.6	70.2	79.1
+ SPEAK	**75.9**	**83.0**	**71.7**	**79.9**

表 2 は，文字情報 (CHARA)，形態素情報 (MORPH)，相手話者情報 (SPEAK) を順に加えていった時の F 値の変化を示すものである．いずれの条件でも，表 1 で示したベースラインを越える性能を示した．また，表が示す通り，それぞれの情報が加わる毎に，いずれの学習器においても性能が向上していることが分かる．

5.5 考察

実験の結果，両学習器ともベースラインを遥かに越える性能を示した．

学習器毎の性能比較では SVMs が CRFs を越える性能を示した．本稿での実験設定では，SVMs と比較して CRFs が学習時に高速で動作するという利点がある．そのため，大量の学習データを扱う場合などは，CRFs の使用が有効であると考えられる．

また，素性を文字情報，形態素情報，相手話者情報と追加していくことによって，性能向上させることができる．Asahara and Matsumoto[4] で述べられているように，文字毎の素性と形態素毎の素性を両方とも利用することが，形態素情報による性能向上の原因であると考えられる．

また，対話では相手話者の働きかけにより発話内容が変化することが考えられる．例えば，相手話者の発話により自分の発話が中断してしまうなどというケースがある．非語彙的表現は話者同士のインタラクションにおいて出現することがあるので，このような相手話者の状態を素性として組み込むことによって，性能向上に貢献できたのだと考えられる．

CRFsでは，これらの素性を使用することで飛躍的に性能が向上している．これは，CRFsにおいては入力系列全体を最適化するラベルを出力するので，形態素単位や相手話者の状態など広い範囲の情報を有効に扱うことができるためであると考えられる．逆に，SVMsは局地的な情報のみで推定を行なうので，素性追加による性能向上はCRFsと比べて小さい．しかし，文字情報のみでも比較的高性能であるので，リソースが少ない場合は特に有効であると考えられる．

6 おわりに

本稿では，教師あり機械学習の枠組みを用いた，対話内の非語彙的表現を抽出手法を提案した．本手法も用いることによって，未知対話の中から非語彙的な表現を抽出することが可能となる．

非語彙的な表現を抽出することによって，人間の対話構造の理解に貢献できると考えられる．例えば，節境界同定のタスクにおいて，非語彙的表現の情報を用いることによる性能を向上できる [12]．今後の課題としては，本手法により抽出された情報を意図識別 [13] などのタスクへ応用していきたい．

参考文献

[1] Expressive Speech Processing project, `http://feast.atr.jp/esp/`.

[2] MeCab: Yet another part-of-speech and morphological analyzer, `http://mecab.sourceforge.net/`.

[3] RT evoluation project, `http://www.nist.gov/speech/tests/rt/`.

[4] M. Asahara and Y. Matsumoto. Filler and disfluency identification based on morphological analysis and chunking. In *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 163–166, Tokyo, Japan, 2003.

[5] N. Campblell. Selecting speech fragments for affect display in concatenative expressive speech synthesis. 日本音響学会 2007 年春季研究発表会, 2007.

[6] S. Kübler. *Memory-Based Learning*, Vol. 7 of *Natural Language Processing*. John Benjamins Publishing Company, 2004.

[7] J. Lafferty, A. McCallun, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289, San Francisco, CA, 2001.

[8] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1526–1540, 2006.

[9] NIST Speech Group. Simple metadata annotation specification version 6.2 – february 3, 2004. Technical report, Linguistic Data Consortium, 2004.

[10] E. Sang and J. Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pp. 173–179, Bergen, Norway, 1999.

[11] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience Publication, 1998.

[12] 牧本, 柏岡, キャンベル. 非語彙的な表現を利用した音声対話の節境界同定の検討. 言語処理学会第 14 回年次大会, 2008 (to apper).

[13] 吉川, 牧本, 柏岡, キャンベル. 音響的特徴に基づくノンバーバル発話の意図識別. 本誌, 2008.