自由対話からの話題遷移検出のためのタグ付け調査

谷津 元樹　ジェプカ ラファウ　荒木 健治
北海道大学 大学院情報科学研究科
{my, kabura, araki}@media.eng.hokudai.ac.jp

1　はじめに

近年，音声対話インターフェースにおいて同時に処理可能な情報源の容量が飛躍的に増加した．機械対話の分野においても，処理できる情報量の増加により，対象とすることのできるドメインが広範囲化した．しかし，言語資源の規模によらず，異なるドメインについて選択的に言語資源を活用するためには，情報源の選択が必要となる．対話の場合，話者が話題の変遷を認識するとともに，用いる応答出力手法および情報源が変更される [1].

本研究では，2 者対話において話題が遷移した時点を話題遷移境界と定義する．話題遷移の発生した発話にラベルを付与する．ラベルを付与した発話の素性情報および前後の文脈情報を素性として，教師有り機械学習を行う．本稿では，学習データを準備するための自由対話への話題遷移タグの付与について述べる．

はじめに，自由対話データに付与した話題遷移タグの位置の分布を考察する．話題遷移タグの付与実験では，複数人において『日本語話し言葉コーパス』[2] 第 3 版に収載された 16 の自由対話データに複数人数により話題遷移タグを付与した．一つのデータの分布の偏りを表す指標である，分布の尖度に着目し，発話の表記単位に基づく分布の性質の異なりについて考察する．

2　関連研究

本研究に関連する分野は，機械学習あるいはルールベースを用いたテキスト情報に対する談話構造の再現および話題遷移のモデル構築である．

談話構造の付与は一連のテキストにおける話題の変化を区切りとした話題上のセグメントを形成する．対話行為タグの心的状態を表す対話行為タグなどの試み [4] もあるが，ここでは話題の変遷に着目する必要がある．

非自由対話の談話構造の付与 『日本語話し言葉コーパス』には，収載された 40 のモノローグ講演に対し，談話境界情報が付加されている [5]．談話境界は，講演音声を階層性を意識せずに 5〜15 程度の話題のまとまりに分割することで付与されている．またタスク指向型の音声対話では，観光案内タスクのおける対話行為タグの付与が行われている [6]．英語の談話における談話構造に対しては，会議録に対して対話行為タグの付与および話題セグメント情報を付与する研究 [3] がある．

非自由対話の話題遷移 菊池らは，EPG やニュース記事を例とする日本語の時系列テキストを対象にした話題遷移抽出 [7] において，95.8%の精度で話題遷移を抽出している．

自由対話への談話構造の付与，話題遷移 日本語対話に対する詳細な対話行為タグの付与のため，ISO DIS 24617-2 対話行為タグ [8] の日本語自由対話への付与 [9] を試みた例がある．また，非タスク指向型対話を対象とした自動的タグ付与の研究 [10] では，SWBD-DAMSL タグを基にしたタグの自動的付与に成功している．

自由対話のデータを元として機械学習の分類結果を用いた動的な話題変更の検出のために，実対話データから学習データを構築する試みは，未だ多く行われてはいない．本研究は，話題遷移の発生間隔の分布から話題遷移の発生条件の統計的な推定を目標とする．

3　話題遷移タグと話題の付与

3.1　実験の方法

公募により表 2 の通り構成された作業参加者に，次のような作業を依頼した．すなわち，表示された対話のログで「話題が変わった」と感じた箇所に話題遷移タグを付与し，話題遷移タグ間の区間（以下，セグメ

表 1: 作業参加者の分布．(計 17 名)

	女	男
10 代	1	0
20 代	4	10
50 代	0	1
60 代	0	1

職業	
社会人	3
学生 (自然言語処理関連分野)	5
学生 (上記以外文系)	3
学生 (上記以外理系)	5
その他・回答無し	1

ントと呼称）から読み取れる話題を記入していただいた．

3.2 自由対話データの準備

作業参加者に呈示された対話データについて述べる．元となった対話は，『日本語話し言葉コーパス Version 1.2』[2](以降，CSJ と表記) に収録された，話者 16 名による自由対話 16 対話の天気情報である．実験画面に表示する際，元データではタグ付けの形で表記されている，感情の表出，言い直し， 聞き取りにくい表現，メタ的表現， 他の特殊な表現を色分けして表示した．以降，自由対話データを**対話**と表記する．なお，作業参加者に呈示される対話は最も回答数の少ないものからランダムに選択した．

3.2.1 発話の表示

CSJ における対話の転記は，その転記基本単位に，対話の録音により取得された物理的指標 (無音区間等) が用いられている [11]．このため，表 3.3 の対話のように発話（書き言葉における文に相当する単位として扱う）は分割して転記される．以降，話者の発声した句の個々の転記基本単位に基づく転記 (転記テキストにおける 1 行の記述) を**発話断片**と呼ぶ．

話題遷移タグの付与において，セグメントの数，タグ付与の基準となる話題のとらえ方には特に制約を設けなかった．これは，作業参加者が記述する話題が様々な粒度をとることを期待したものである．

3.3 調査の結果

本実験により得られたデータについて述べる．データ数の分布を表 2 に示す．

調査の結果にみられるデータ数と対話数の関係から，取得したデータを 3 つのグループへの分類ができることがわかった．データ数が多く，対話の数が少ない，データ数が 12 または 11 の対話 (グループ A) および，

表 2: 対話別のデータ数の分布．

グループ	データ数	対話数	平均対話長
A	12	1	539.3
	11	2	
B	10	6	568.3
	9	6	
C	8	1	769.0

(計 16 対話，データ数 156)

表 3: 話題遷移タグの付与された自由対話の例．

話者	発話内容	付与数
R	ごく普通に使う	0
R	状況でしょ	0
L	(F はい)	0
L	(F え) じゃ(F あの) もしも さっきの話の続きですけど	1
R	(F うん)	7
R	(F うん)(F うん)	0
L	(F あのー) 機械が喋るように なった時に (D い)	0
R	(F うん)	0
L	対話もできるように なったりとかしますか	0

計 12 対話がデータ数が 10 または 9 の対話 (グループ B) であり，残りの 1 対話は，データ数が必ずしも少ないとはいえないが，対話長が他グループと比較して外れ値ともとれる長さを持つため，もう一つのグループに分類される．

なお，本稿では，自由対話において発話を行う時間的頻度を一定と仮定する．話者による個人差，環境や対話時の状況による発声時間への影響の考慮は今後の課題とする．

4 考察

4.1 データの性質

本節での考察の対象は，自由対話の種類ごとに，付与の行われた全データを合わせたものである．話題遷移タグ数およびセグメント長を指定していないので，

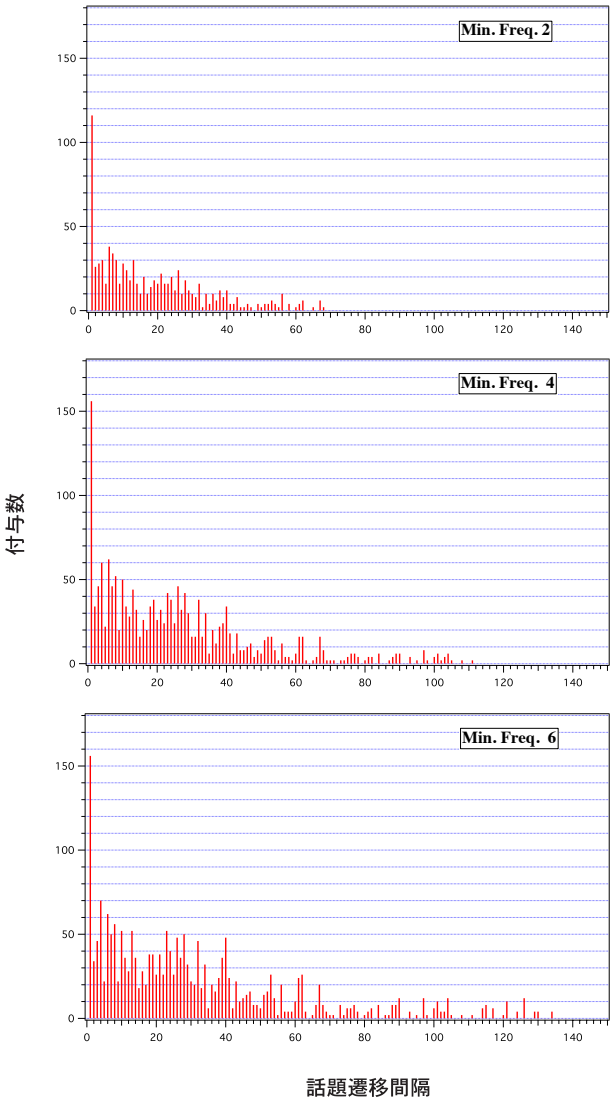

図 1: タグ付与数の下限 $Y_{min} = \{2, 4, 6\}$ の話題遷移間隔の分布．横軸は話題遷移間隔，縦軸は度数を示す．

話題遷移タグの付与位置に法則性が認められるとしても，各位置で付与数に差異が生じることが考えられる．会議録データへの話題境界タグ付与の際，対話データを通しての付与数の少ない作業参加者は，セグメント毎の話題の粒度[1]が大きくなることが報告されている [3]．本研究では，この性質を 2 者間の自由対話に適用して粒度の異なる話題遷移の検出を試みる．

4.2 発話の話題遷移間隔の度数分布

話題遷移タグ付与数の下限 Y_{min} の値を変え，話題遷移間の間隔と作業参加者によるタグ付与の頻度の分布がどのように変化するかを調査する．今回の調査で得られた度数分布は図 1 である．ここで，分布の左端に度数の偏りが見られる．

[1]話題や対話セグメントの粒度に関して共通の指標は確立されていない．粒度を変数とした話題遷移の発生確率が連続的な分布関数をとるものと仮定する．

表 4: 発話の統合操作前後の最小付与数 $Y_{min} = \{2, 4, 6\}$ の分布の統計量．β_{2std} は正規分布における尖度を，β_{2pois} はポアソン分布における尖度を示す．

統合	Y_{min}	データ長	平均付与数	β_{2std}	β_{2pois}
統合前	2	73	11.0	28.4	0.091
	4	111	15.0	21.8	0.067
	6	138	15.0	18.7	0.067
統合後	2	57	9.0	1.87	0.111
	4	90	12.0	3.63	0.083
	6	108	12.0	4.10	0.083

4.3 分布の仮定

観測されたデータは，データ量が十分に大きい場合，正規分布またはポアソン分布のいずれかの分布に従う可能性があると考え，両者の標準的な分布との差異を測定する．正規分布は，確率的に独立な事象がランダムな確率変数の値を持つ際に仮定される分布 (確率密度関数は $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$) である．ポアソン分布 (確率質量関数は $f(x) = \frac{\lambda^k e^{-\lambda}}{k!}$) は，事象の発生間隔がランダムである場合にその間隔の度数がとりうる分布である．理想的なポアソン分布に従う場合，話題遷移間隔はランダムな値をとることになる．

4.4 発話の統合操作

3.2 で述べたように，実験で用いた CSJ の自由対話データは，1 つの発話が複数位置にわたって標記されている．図 1 に，最小付与数 Y_{min} を 3 から 6 にとった場合の話題遷移タグ間の間隔の度数分布を示す．間隔 1 が大きな値となっており，発話の分割表記によるゆれと考えられる．

結果データにおいて分断された発話を統合し 1 つの単位とすることにより，この影響の低減を試みた．具体的には，話題遷移タグの付与データにおいて，同じ話者による連続した発話断片を一つの発話と見なし，その内のひとつの発話断片へのタグ付与を発話へのタグ付与とした．

尖度は，特定の確率分布の上の異常な値の多さを示す値である．正規分布およびポアソン分布を仮定した

尖度はそれぞれ，

$$\beta_{2std} = \frac{\sum_x \left(\frac{(x-\mu)^4}{\sqrt{\sigma^2}^3} \right)}{||X||} - 3, \quad (1)$$

$$\beta_{2pois} = \frac{1}{\mu} \quad (2)$$

で表される [12](ただしデータ $x \in X$ について，μ は平均，σ^2 は分散)．これらの指標を用いて，操作前と後の話題遷移タグ間隔の分布の形状を比較した．表 4 の通り，処理の前後の分布の尖度に差異が認められる．このことから，発話の分割表記による話題遷移タグの付与位置のゆれが示唆される．また，より尖度の値が小さいということは，観測値の分布が仮定した標準的な分布により近いことを意味する．

5 おわりに

本稿では，発話を統合することにより分布の観測にどのような影響が及ぼされるかを考察した．結果，同一の発話者ごとに発話を統合し，話題遷移タグの付与を発話の単位で扱えば，度数分布がより高い正規性を持つことが示唆される．度数分布がどのような確率密度関数に従うかの検定，A・B・C グループ間での話題遷移間隔の分布の異同の発見，そして機械学習のデータとしての適性の検討については，今後の課題である．

参考文献

[1] 谷津元樹, ジェプカラファウ, 荒木健治. トピック推定を用いたタスクドメインを選択するための発話生成. 第 19 回言語処理学会年次大会, 2012.

[2] 前川喜久雄, 籠宮隆之, 小磯花絵, 小椋秀樹, 菊池英明:. 日本語話し言葉コーパスの設計. 音声研究, Vol. 4, No. 2, pp. 51–61, 2000.

[3] Alexander Gruenstein, Jojn Niekrasz, and Matthew Purver. Meeting structure annotation: Data and tools. In *In Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pp. 117–127, 2005.

[4] 徳久雅人, 前田浩佑, 村上仁一, 池原悟. 心的状態を表す対話行為タグ付きテキスト対話コーパスの構築. 電子情報通信学会技術研究報告, 思考と言語, TL2007-45, pp. 25–30, 2007.

[5] 竹内和広, 森本郁代, 高梨克也, 井佐原均. 『日本語話し言葉コーパス』の談話境界情報について version 1.0. Technical report, 独立行政法人 情報通信研究機構.

[6] 翠輝久, 大竹清敬, 堀智織, 柏岡秀紀, 中村哲. 京都観光案内対話コーパスにおける対話行為タグの設計と分析 (理解). 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2009, No. 10, pp. 39–44, 2009.

[7] 菊池匡晃, 岡本昌之, 山崎智弘. 階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出. データ工学ワークショップ DEWS, 2008.

[8] Harry Bunt and et al. Iso 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of LREC*, pp. 430–437, 2012.

[9] 井岡孝徳. 日本語コーパスに対する iso : Dis 24617-2 に基づく対話行為情報を用いたアノテーションとその分析. 奈良先端科学技術大学院大学課題研究, 2012.

[10] 磯村直樹, 鳥海不二夫, 石井健一郎. 対話エージェント評価におけるタグ付与の自動化. 電子情報通信学会論文誌. A, 基礎・境界, Vol. 92, No. 11, pp. 795–805, 2009.

[11] 国立国語研究所. 日本語話し言葉コーパスの構築法. Technical Report 124, 国立国語研究所報告, 3 2006.

[12] Maurice G. Kendall and William R. Buckland. *A Dictionary of Statistical Terms*. No. 4th ed., rev. and enl. 1982.