Vol.2010-SLP-83 No.4
2010/10/29

日本語自由発話電話音声からの固有表現抽出

伊東伸泰[†]　倉田岳人[†]　西村雅史[†]

固有表現抽出は自然言語処理を対象とした情報抽出において基本的な処理であり，多くの先行研究がある．しかしながら音声，なかでも自由発話を対象とする場合は話し言葉特有のフィラーや言い淀みに加え，音声認識の認識誤りがノイズとなるため非常に困難とされている．また特徴としてよく用いられる品詞を同定するためには形態素解析を行う必要があるが，認識誤りを伴った音声認識結果に対して頑健に動作させることは難しい．本研究ではコールセンターにおける実際の発話を対象として，音声認識を行い，固有表現抽出を試みた．モデリングには単語 3-gram を素性とした条件付確率場（CRF）を用い，加えて各固有表現との相互情報量に基づくクラス，および音声認識から得られる *N*-best を利用した追加素性について検討した．実験によれば音声認識結果に対し金額表現，組織（法人），商品について，それぞれ F 値で 0.89，0.56，0.74 という結果を得た．

Named Entity Recognition from Conversational Telephone Speech in Japanese

Nobuyasu Itoh[†]　Gakuto Kurata[†]
and Masafumi Nishimura[†]

Named entity recognition (NER) is an important task for natural language processing and many research works have been reported. It is still, however, challenging to extract them from an output of automatic speech recognizers (ASR) output due to recognition errors. Binary features obtained from look up dictionaries with parts-of-speech are often used for text based NER. But part-of-speech taggers for Japanese usually assume clean texts and may produce unexpected results for ASR output. In this paper, we report NER for real conversational telephone speech in Japanese. We created Conditional Random Field (CRF) models with the word 3-grams. Therefore we used classes based on mutual information criteria between the classes and the named entities are introduced instead of parts-of-speech. We also tested *N*-best results in a word confusion network. According to our test, F-measures for prices, company names, and product names were 0,89, 0.56, and 0.74 respectively.

1. はじめに

音声認識技術の発達に伴って，その制約がより小さくなったことから応用範囲が広がっている．従来からの口述筆記（ディクテーション）や特定分野の情報検索のみならず講演を書き起こし音声ドキュメントを検索可能な形でアーカイブする，分野を特定せず検索のインタフェースとする，といったことが実現され始めた[1,2]．このように「情報抽出・検索のためのメディア変換技術」として音声認識の役割は今後さらに大きくなると期待されるが，その重要な対象の１つとして電話音声がある．実際，多くの企業がコールセンターをはじめとする電話を重要な「顧客接点」の場と位置づけており[3]，その品質改善とともにそこから製品サービスに関するさまざまな有用な情報を抽出することが試みられている[4]．しかしながらコールセンターにおける電話音声はその音響環境上の制約と同時に人間同士が多くの場合何らのメモ・原稿なしに会話する自然発話でかつ大語彙であることから認識が難しく，日本語ではもちろん英語においても長い間ベースラインの認識精度を上げることが研究の中心であった[5][a]．しかしここ数年特に英語において電話音声，自由発話，かつ大語彙タスクの認識精度向上が見られることから[6]，本研究ではコールセンターでの自由発話を対象として日本語を音声認識し，そこから情報抽出の第一ステップとなる「固有表現」抽出を試みた．

2. 自由発話電話音声のコーパス

本節では本研究で用いたコーパスについて記述し，対象とするタスクの概略を示す．

2.1 書き起こし

日本語自由発話コーパスは「日本語話し言葉コーパス」（CSJ [7]）をはじめとしていくつかの報告があるが，電話音声の実発話は機密保持，さまざまな権利保護上取り扱いが難しいことから米国における Fisher[5]のようなコーパスは提供されていない．そこでまずあるコールセンターの対話を人手で書き起こし，ベースとなるデータを作成した．書き起こし規約は概ね[8]にしたがっているが，その概略と一部異なる箇所を以下に述べる．

A. 一定時間以上のポーズで分割し，それを１発話とする．
B. 不要語は単語断片，フィラーともに書き起こし（それを示す）タグを付ける．

[†] 日本アイ・ビー・エム株式会社，東京基礎研究所
IBM Research – Tokyo, IBM Japan, Ltd.

a 一方限定された分野で人がマシンと対話することにより所定のタスクを人手を介さずに達成することを目的とする音声自動応答は情報抽出（言語処理）部分を含め多くの報告がある．

Vol.2010-SLP-83 No.4
2010/10/29

C. 聴取不能部分は<?>で示す.
D. 送り仮名はいずれかに統一（例: 組み合わせ, 組合せ）, 擬音・擬態語の表記（例: 「バタバタする」「ばたばたする」）も統一する.
E. 法人名は日本語正式表記に準ずる.（例: 「キヤノン」「キャノン」, 前者に統一）
F. 数字表現は読み方（桁読み）を反映し，漢数字で表記する.（例: 「ニヒャクニジュウエン」→「二百二十円」,「ニーサンゴ」→「二三五」）

2.2 単語単位

前節のルールに従って書き起こした結果を文献[9]で用いた単語単位[b]により分割した. 具体的には[9]で用いたデータや講義コーパス[8]から学習した言語モデルを用いて尤度最大になるように一次分割を行い，明らかに誤っている箇所について修正を加えるという方法をとった．その際本研究の対象となる「固有表現」についてはどのような単位を用いるかが固有表現抽出にとっても重要であるが，ここでは会話に出現した単位をすべて採用するという方針を採用した．たとえば「ＩＢＭグローバルファンド」という金融商品があったとすると，それは前記の正式名称で呼ばれる場合もある一方,「ＩＢＭグローバル」と発話されたり，文脈によっては単に「ＩＢＭ」「グローバル」と呼ばれている場合もある．この場合のべ４個の単語を登録する．図にコーパスの例を示す．ただし機密保護のためコーパスの特徴を歪めない範囲で改変を行っている.

ニュージーランド の 通貨 が <エー> 上がって おりまして 、ええ 、
元本 が <アノ> *MONEY*[十 一万] ぐらい 上がって るん です けど 、
これ 、 つまり その <?> せい で *PRODUCT*[ＩＢＭ グローバル] 十
二月 で 償還 に なって しまう ん です よ 。
<途中省略>
:
そちら に お 振込み いただければ 大丈夫 なん です けれど 。
お 手数 かけます 。 ありがとう ございます 。

図 1 電話音声コーパスのサンプル

‘[ ]‘ （かぎ括弧）は固有表現の範囲を示し，斜体の英語（*MONEY, PRODUCT*）は固有表現の種類を示す.

b 活用語については活用語尾，音便，接続助詞などがしばしば結合された単位となっている一方，名詞は「短単位」が多くを占める.

3. 固有表現抽出

3.1 固有表現抽出手法

自然言語処理において固有表現，あるいは Named Entity 抽出は情報抽出の基本処理であるため数多くの報告があるが，いわゆるルールベースを除くと大きく生成モデルをベースとする手法と識別モデルに基づく手法に分けることができる．前者は入力単語列 W に対して，当該単語列とそれに対する固有表現を示すラベル(T)の同時生起確率 $P(W, T)$を最大とするラベル列 ($\overline{T}$) を求めようとする手法であり，その計算には HMM がよく用いられる[10]．一方識別モデルでは SVM [11], 最大エントロピー法[12], その拡張である条件付確率場（CRF）[13]による報告がある．自然言語処理のコンテストとして著名な CoNLL は 2003 年度に言語に依存しない手法による Named Entity Recognition タスクについて報告を行っており[14]，識別モデルが上位システムの多くを占めた．本研究ではその中でも系列ラベリング問題に適しているとされる条件付確率場を用いることにした.

3.2 条件付確率場 （CRF）

本節では条件付確率場について簡単に説明する[13]．入力単語列を $\boldsymbol{W}=(w_1 w_2,\ldots,w_N)$, 固有表現であるかないか，ある場合はその種類を示すラベル列（$\boldsymbol{T}=(t_1 t_2,\ldots,t_{N+1})$）[c]とすると，CRF では以下のように求められる条件付確率 $P(\boldsymbol{T} \mid \boldsymbol{W})$を最大にするラベル列（$\overline{T}$ ）を最適なラベル列とみなす.

$$\overline{T} = \underset{T}{argmax}\, P(\boldsymbol{T} \mid \boldsymbol{W}) = \underset{T}{argmax} \frac{\exp(\langle \Theta, \Phi(\boldsymbol{W},\boldsymbol{T}) \rangle)}{Z(\boldsymbol{W})}$$

$$= \underset{T}{argmax} \frac{\exp\left(\sum_{i=1}^{N} \lambda_i \cdot f_i(\boldsymbol{W}, T_i^{i+1}) \right)}{Z(\boldsymbol{W})}$$

$$Z(\boldsymbol{W}) = \sum_{\tilde{T}} \exp(\langle \Theta, \Phi(\boldsymbol{W}, \tilde{\boldsymbol{T}}) \rangle), \quad T_i^{i+1} = (t_i, t_{i+1})$$

ここで $\langle\ \rangle$ は内積，$\Theta = (\lambda_1 \lambda_2,\ldots,\lambda_N)$ は素性 $\Phi(\boldsymbol{W},\boldsymbol{T}) = (f_1, f_2,\ldots, f_N)$ に対する重みを表す. 予測時の計算では Z は T に依存しないので無視される. f は入力単語列（$\boldsymbol{W}$）から得られる素姓（入力特徴）と，デコード時に得られる素性（遷移特徴）に分けら

c t_{N+1} は文または発話終端を示すラベル.

Vol.2010-SLP-83 No.4
2010/10/29

れるが，ここでは遷移特徴としてラベル列（*T*）の 2-gram が用いられている．

3.3 音声認識結果に対する固有表現抽出

技術が進歩したとはいえ，音声認識にはある程度の認識誤りが含まれる．また固有表現抽出上有力な特徴となる句読点，（英語他における）大/小文字の別もそのままでは得られない．音声認識結果特有の手法として Palmer[10]は音声認識結果のエラーをモデル化することを提案している．Zhai[15]は音声認識の *N*-best 結果を利用する試みを報告，須藤[11]は認識結果の確信度を素性に導入している．本研究ではこれらを踏まえて，認識結果の *N*-best 結果をトレリスとして表現した単語コンフュージョンネットワーク（Word Confusion Network）を抽出器への入力とし，各単語に付随するスコア（事後確率）を連続値の入力特徴（素性）として表現することを試みる．

4. 単語コンフュージョンネットワークに基づく素性

4.1 単語コンフュージョンネットワーク

音声認識結果を処理するにあたって，精度向上のため *N*-best 結果を用いることはしばしば行われるが，音声認識デコーダーが出力する *N*-best 結果は探索パスを尤度順に並べたものであり，各結果の単語・境界の対応関係が明らかでないため，１つのまとまった系列として扱うことが非常に困難である．Zhai[15]は中国語において *N*-best 結果を用いるさい，すべての *N*-best パスを 1 つずつ固有表現抽出システムへ入力し，得られた結果から多数決により最終的なラベルを決定しているが，対応付けの詳細は記述されていない．

お待ち　いただいて　申し訳　ありません　</s>
お持ち　いただいて　申し訳　あり　ません　</s>
お　待ち　いただいた　申し訳　あり　ません　</s>
お　待ち　いただいた　申し訳　始まり　ません　</s>
：

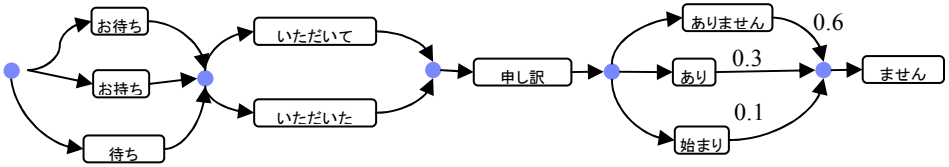

図 2　*N*-best 結果と単語コンフュージョンネットワーク

本研究では Mangu 他[16]によって提案された単語コンフュージョンネットワーク（以降 WCN）に基づいて入力となる系列を構成する．WCN とは *N*-best 結果の各単語列について Multiple Alignment の手法により単語同士の対応関係を求め，その後で時間情報を元にオーバーラップする単語群をまとめることにより，単語候補集合の系列として表現された構造である．図 2 に例を示す．図の上が *N*-best 候補，下がそれを WCN に変換した例である．*N*-best の各候補パス（単語列）の出力単語数は一般に一定ではないが，WCN では「いただいて / いただいた」や「ありません / あり / 始まり」のように対応付けられ，単語候補集合が時刻により順序付けられた形式となっている．各候補単語集合を 1 つの素性ベクトルと考えれば，テキストの場合同様，入力となる時系列を容易に構成することができる[d]．かつ各単語候補には元の *N*-best パスの尤度を当該単語に再配分し，正規化した確率値が付けられる．たとえば「ありません / あり / 始まり」に付随する 0.6，0.3，0.1 という数値は当該位置における各単語の事後確率を示す．したがって 1 位パスのみを用いて本ネットワークを構成する場合は，各候補集合に唯一の単語が含まれ，その事後確率を 1 とする素性ベクトルに相当し，*N*-best 結果が 1 位結果の自然な拡張となっている．

4.2 素性（入力特徴）

識別モデルを用いた固有表現抽出ではさまざまな素性が提案されているが，一般によく用いられるのは単語 n-gram やそこから抽出された品詞，さらに英語の場合は「大文字/小文字」の別，日本語では文字種といった表層特徴である[12]．しかしながら単語 n-gram を除くと，いずれも音声認識結果には適用しづらい．品詞を入力素性とすると後処理で形態素解析を行うことになるが，コールセンターにおける自由発話を精度よく解析することは，十分な認識精度が確保できたとしても難しいと考えられる．また文字種が効果的であるのは主として未知語における対応であるが，音声認識においては未知語が表記として正しい結果に変換されることはほとんど期待できない．そこで本研究では通常用いられる単語 3-gram に加えて，固有表現をあらかじめ分類したクラスと各単語との相互情報量（*I*）を用いることにした．単語 3-gram では現在位置(*i*)および前後 1 個ずつの単語（w_{i-1}, w_i, w_{i+1}）を用いる．相互情報量はクラスを C_k，現在位置 *i* の WCN を $\boldsymbol{W}_i = (w_{i1}, w_{i2}, \ldots, w_{in})$ として以下のように定式化される．

$$I(C_k;\boldsymbol{W}_i) = \log\frac{P(C_k,\boldsymbol{W}_i)}{P(C_k)P(\boldsymbol{W}_i)} = \log\frac{P(C_k\,|\,\boldsymbol{W}_i)P(\boldsymbol{W}_i)}{P(C_k)P(\boldsymbol{W}_i)}$$

$$= \log\frac{P(C_k\,|\,\boldsymbol{W}_i)}{P(C_k)} \cong \log\frac{\sum_j P(C_k\,|\,w_{ij})P(w_{ij}\,|\,\boldsymbol{W}_i)}{P(C_k)}$$

d あるパスで 1 単語であった部分が別のパスでは複数単語となることはしばしば起こるが，単語コンフュージョンネットワーク作成においてはクラスタリングされた結果を枝狩りし，残ったグループを時間情報で整列化する．したがって元の候補に存在した「お」(接辞)が削除されたり，ある単語系列を見るとオーバーラップすることがあり得る（「ありません」の後に「ません」が続くなど）．

$P(C_k|w_{ij})$は単語を条件とする固有表現の出現確率であるからコーパスから学習でき，$P(w_{ij}|\boldsymbol{W}_i)$は WCN における候補単語に付随する事後確率に相当する．

C_kは金額表現およびその前後，金額表現以外の固有表現（法人名，商品名など）とその前後，それ以外で，のべ 7 クラスを作成した．したがって語彙サイズを|V|として，各位置(i)から 3・|V|＋7 次元の素性ベクトルが得られることになる．各単語に相当する要素には WCN から得られる事後確率，相互情報量に基づいたクラスに対応する要素には当該相互情報量（I）が入る[e].

5. 実験

本節ではこれまで述べた手法により固有表現抽出を行った実験，およびその結果について述べる．

5.1 予備認識実験

第 2 節で作成した電話音声コーパスから語彙を作成した．コーパスおよび語彙の緒元を次に示す．

表 1　電話音声コーパスの緒元

文数（句点数）	431K
単語数（のべ数）	2,010K
フィラー	175K
語彙サイズ	20,650
聴取不能（箇所）	32K

このコーパスから文献[6]のシステムに基づく言語モデル（単語 3-gram）を作成し，上記データとは別の８コール（約１時間分）のコールセンター担当者部分について認識実験を行ったところ以下の結果を得た．

表 2　認識実験

カバレージ	98.9%
パープレキシティ	113
文字誤認識率（CER）	19.02%

5.2 固有表現抽出実験

表 1 に示したコーパスの約半分を用いた言語モデルを作成し，残りのデータの中から 2 万発話（約 173K 単語）を固有表現抽出実験に用いた．実験データは以下の 3 種類を作成した．

A. 人手で書き起こしたテキストデータ
B. 音声認識により認識した結果の 1 位候補からなるテキスト
C. 音声認識により認識した結果の *N*-best 候補

A の場合，コーパス作成時に付与した句読点はすべて削除する一方フィラーについてはそのまま利用した．B，C は無音を示す認識結果を削除後，文字単位で近似文字列照合を行って A との対応関係を求め，（A に付与された）固有表現ラベルを付与，実験データとした． また，CRF の学習にあたっては L2 正則化を実施している[f].
結果を表 3 に示す．実験は上記データを 10 等分し，90%で学習，残りでテストを実施する 10 fold 法によった．ただし実験 1 および 4 ではテストデータとして A 自身ではなく B の該当部分を用いている．

MI class (*/-)は相互情報量に基づくクラス特徴のある/なしを示し，Companies, Products，Money はそれぞれ法人名，金融商品名，金額表現を意味する．精度（F 値）算出にあたっては，固有表現の開始・終端および中間を区別することなく，単語ごとに当該ラベルが付与されたかどうかで判断している．これらの結果から見ると以下のことが読み取れる．

表 3　固有表現抽出結果

No	Training Data	Test Data	MI class	**Accuracy** (upper: F-measure, lower: Precision/Recall))		
				Companies	Products	Money
1	Human Transcript (A)	B	*	**0.561**	**0.740**	0.868
				0.687/0.475	0.887/0.634	0.893/0.844
2	ASR output 1-best (B)	B	*	0.508	0.736	0.880
				0.664/0.411	0.868/0.639	0.90/0.861
3	ASR output *N*-best (C)	C	*	0.509	0.730	**0.893**
				0.680/0.407	0.836/0.648	0.897/0.888
4	Human Transcript (A)	B	-	0.246	0.692	0.890
				0.729/0.148	0.923/0.553	0.903/0.878
5	ASR output 1-best (B)	B	-	0.263	0.674	0.887
				0.717/0.168	0.923/0.530	0.904/0.871
6	Human Transcript (A)	A	*	0.814	0.921	0.957
				0.814/0.814	0.939/0.903	0.958/0.956
	# of entities in test data (B,C)			346	553	3,337

e 正確には予備実験で得られた相互情報量の分布に基づき，その値を０～1 にスケーリングしている．

f 正則化の任意パラメータは最初に学習データを 2 分し，F 値のクロスバリデーションを実施することにより最適な値に決定後，再度全体から学習したモデルを作成した．

A. 相互情報量に基づくクラス（MI Class）は法人名，金融商品名の抽出精度向上において役立っている．一方，金額表現においては効果が見られない．
B. 金額表現では音声認識結果を学習データとすることでわずかながら精度向上が見られた．
C. 法人名・金融商品名においては音声認識結果を学習データとした結果，精度が低下している．

A.は MI Class が単語自身が持つ情報を平滑化し，補完するもので新たな情報を追加するものではないため，学習データに応じて効果が決まる，つまりサンプル数が多い金額表現では効果がなく，よりサンプル数が少ない法人名，商品名で効果的であったと解釈できる．一方「音声認識結果の利用」については解釈が難しいが，データがより多い金融商品や金額表現の結果から推察すると音声認識結果，さらに *N*-best 結果を用いることが再現率を向上させる一方適合率を低下させ，いずれの効果が優るかによって F 値が決まるように見受けられる．このように音声認識結果を学習データとすることが「適合率」を有意に低下させることは Zhai[15]の結果にも見られることから，より精度向上のためにはさまざまな学習データから作成したモデルで抽出を実施し，その結果から最終判定を行うといった手法が必要であろう．

5.3 関連研究

須藤[11]は日本語の音声認識データに対して SVM による固有表現抽出を提案し，F 値 0.69 を得ている（8 カテゴリの固有表現，データは主として新聞コーパスの読み上げ）．また Surdeane[17]は Switchboard（英語の電話音声コーパス）についてやはり SVM で最高 0.75 の固有表現抽出精度を報告しているが，こちらは人手による書き起こしで認識結果は用いていない．本研究の結果が絶対値として良好であるかどうであるかは評価が難しいが，これらと比較して期待できる結果であると考える．

6. おわりに

本研究ではコールセンターにおける自然な会話データについて，音声認識を実施し，その結果から固有表現抽出することを試みた．大語彙，自由発話，電話音声という厳しい条件の下でありながら，金融商品で 0.74，金額表現で 0.89 の F 値を得ることができた．使用する学習データの違いや相互情報量に基づくクラス特徴の効果については，必ずしも一貫しない結果となったが，今後これらについてさらに精査したいと考えている．

謝辞 条件付確率場についてご教示いただき，またツールを使用させていただいた東京基礎研究所ナレッジ･インフラストラクチャーグループの坪井祐太氏に深謝する．

参考文献

1) Tomoyosi Akiba, Kiyoaki Aikawa, Yoshiaki Itoh, Tatsuya Kawahara, Hiroaki Nanjo, Hiromitsu Nishizaki, Norihito Yasuda, Yoichi Yamashita, Katunobu Itou: Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data, IPSJ Journal, Vol.50, No.2, pp.501-513 (2009).
2) Schuster, M.: Japanese Voice Search, IPSJ 第 82 回音声言語処理研究会 (2010).
3) リックテレコム編: コールセンター白書 2009 (2009).
4) 竹内 広宜，那須川 哲哉，渡辺 日出雄: コールセンターにおける目的を持ったビジネス会話のモデリングと会話マイニングへの応用，人工知能学会論文誌，Vol.23, No.6, pp. 384-391 (2008).
5) David, C. C., Miller, D.: The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text Proceedings 4th International Conference on Language Resources and Evaluation , pp. 69-71 (2004).
6) Chen, S. F. et al.: Advances in Speech Transcription at IBM Under the DARPA EARS Program, IEEE trans. Audio, Speech, and Language Processing, Vol.14, No.5, pp. 1596-1608 (2006).
7) 前川喜久雄: 日本語話し言葉コーパスの概観, Ver.1.1, http://www.kokken.go.jp/katsudo/seika/corpus/releaseinfo/040/overview.pdf から入手可能．
8) 西村雅史，伊東伸泰: 講義コーパスを用いた自由発話の大語彙連続音声認識，電子情報通信学会論文誌, Vol.83-D-II, No.11, pp. 2473-2480 (2000).
9) 伊東伸泰，西村雅史，荻野紫穂，山崎一孝: 単語単位による日本語言語モデルの検討，自然言語処理, Vol.6, No.2 (1999).
10) David D., Palmer and Ostendorf, M.: Improving Information Extraction by Modeling Errors in Speech Recognizer Output, Proceeding of HLT, pp. 156–160 (2001).
11) Sudoh, K., Tsukada, H., and Isozaki, H.: Incorporating Speech Recognition Confidence into Discriminative Named Entity Recognition of Speech Data, Proceeding of 44th Annual Meeting of the ACL, pp. 617–624 (2006).
12) 内元清貴，馬青，村田真樹，小作浩美，内山将夫，伊佐原均: 最大エントロピー法と書き換え規則に基づく日本語固有表現抽出，自然言語処理, Vol.7, No.2, pp. 63-90 (2000).
13) 坪井祐太，鹿島久嗣，工藤拓: 言語処理における識別モデルの発展 – HMM から CRF まで，言語処理学会第 12 回年次大会（NLP2006）チュートリアル (2006).
14) http://www.cnts.ua.ac.be/conll2003/ner/ .
15) Zhai, L., Fung, P., Schwartz, R., Carpuat, M., and Wu, D.: Using N-best List for Named Entity Recognition from Chines Speech, Proceedings of HLT-NAACL, pp. 37-40 (2004).
16) Mangu, L., Brill, B., and Stolcke, A.: Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network, Computer Speech and Language, Vol.14, No.4, pp. 373-400 (2000).
17) Surudeanu, M., Turmo, J., and Comelles, E.: Named Entity Recognition from Spontaneous Open-Domain Speech, Proceedings of the 9th International Conference on Interspeech (2005).