データベース名称：　実証実験コーパスを用いた言語モデルおよび辞書

I データの概要

平成 21 年度に全国 5 地方で訪日観光分野を対象とした大規模音声翻訳実証実験を実施した。この実証実験は、観光施設などに端末を設置して 1〜3 ヶ月の期間にわたって実施された。このときに収集された日英中韓 4 ヶ国語の実利用音声データを対象として書き起こし作業を行った。本データは、この書き起こしテキスト約 17 万発話を形態素解析処理したものから作成した N グラム頻度（4 グラム）データである。また、音声認識に用いるための発音辞書も同時に提供する。

観光分野タスクの言語モデルを作成するときには、その内容に関連する書籍の文章や WEB 収集したコーパスを利用するのが一般的であるが、実利用時にはズレがあることが考えられる。本データを活用することが、そのズレを少しでも解消するための一助になれば幸いである。

II 関連発表

磯谷亮輔・松田繁樹・林輝昭・河合恒・中村哲, ″全国音声翻訳実証実験の実施と実利用データを用いた音声認識のモデル適応″,電子情報通信学会論文誌 D, Vol.J96-D, No.1, pp.209-220

III データ内容

(1) ファイル構成

(lang=ja,en,zh,ko)

hosei2009/$lang/

all_hosei.lm4.$lang.punc.count	4 グラム頻度ファイル（句読点有）
all_hosei.lm4.$lang.nopunc.count	4 グラム頻度ファイル（句読点無）
lex.$lang.nopunc.txt	発音辞書
propn.$lang.txt	発音辞書中の固有名詞候補単語

文字コードは、UTF8 である。

(2) 各言語の学習セット例

学習セットの書式の例を以下に示す。

日本語、韓国語は発音情報を含む形態素になっている。

[ja]

句読点有

お客|オキャク 様|サマ 、 忘れ物|ワスレモノ は|ワ 何|(ナニ$ナン) です|デス か|カ ？

みたらし団子|ミタラシダンゴ の|ノ 匂い|ニオイ が|ガ し|シ ます|マス 。 いくら|イクラ です|デス か|カ ？

句読点無

お客|オキャク 様|サマ 忘れ物|ワスレモノ は|ワ 何|(ナニ$ナン) です|デス か|カ

みたらし団子|ミタラシダンゴ の|ノ 匂い|ニオイ が|ガ し|シ ます|マス <sb> いくら|イクラ です|デス か|カ

[en]

句読点有

hello how are you today ? we're traveling on a train .

where is the hotel ?

句読点無

hello how are you today <sb> we're traveling on a train

where is the hotel

[zh]

句読点有

嗯 ， 大家 好 。 我 从 中国 上海 来 ， 非常 高兴 见到 你们 。

谢谢 您 的 光临 ， 再见 。

句読点無

嗯 大家 好 <sb> 我 从 中国 上海 来 非常 高兴 见到 你们

谢谢 您 的 光临 再见

[ko]

句読点有

열차|yeol#cha 가|ga <pb> 지연|ji#yeon 됐|dwaet 습니다|sseum#ni#da .

어떤|eo#tteon <pb> 온천|on#cheo 입니까|nim#ni#kka ?

句読点無

열차|yeol#cha 가|ga 지연|ji#yeon 됐|dwaet 습니다|sseum#ni#da

어떤|eo#tteon 온천|on#cheo 입니까|nim#ni#kka

<sb>は文境界記号、<pb>は韓国語の節境界記号である。

日本語形態素解析器は、茶筌 2.4.5、IPA 辞書 2.6.3 を使用している。

中国語の形態素解析器は NICT 内製を使用している。

韓国語の形態素解析器は POSTEC 製を使用している。

(3) コーパス諸元および言語モデル諸元

全コーパスのデータから約 2%ずつテストセット、開発セットを選び、残りを学習した言語モデルのテストセット Perplexity を求めた。スムージングは mKN でカットオフはしていない。

コーパス諸元と言語モデル諸元を示す。

言語	句読点	コーパス諸元				言語モデル諸元		
		全発話数	単語数	語彙数	テストセット発話数	2-gram ppl	3-gram ppl	oov rate
ja	有	99669	609913	13516	1967	19.7	14.7	1.55%
ja	無	99669	503621	13513	1964	31	22	1.82%
en	有	30045	152574	5067	621	27.5	23.4	2.30%
en	無	30045	119810	5065	618	48	39.6	2.78%
zh	有	26107	123481	6638	553	36.1	34.8	4.41%
zh	無	26107	94352	6636	538	72	67.9	4.56%
ko	有	14611	103631	4856	309	18.9	14.8	3.93%
ko	無	14611	66978	4854	307	44.3	41.9	6.06%

(4) 発音辞書

辞書の発音表記は各言語つぎのようになっている。

日本語:カタカナ

英語 :CMU39 音素

中国語:ピンイン

韓国語:ローマ字