Translating Conversational Speech to Standard Linguistic Form

Darren Scott Appling[1], Nick Campbell[2]

[1]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[2]NiCT/ATR-SLC, National Institute of Information and Communications Technology, Keihanna Science City, Kyoto 619-0288, Japan
darren.scott.appling@gatech.edu, nick@nict.go.jp

Abstract

This paper describes the so-called ill-formed nature of spontaneous conversational speech as observed from the study of a 1500-hour corpus of recorded dialogue speech. We note that the structure is quite different from that of more formal speech or writing and propose a Statistical Machine Translation approach for mapping between the spoken and written forms of the language as if they were two entirely separate languages. We further posit that the particular nature of the spoken language is especially well suited for the display of affective states, inter-speaker relationships and discourse management information. In summary, both modes of communication appear to be particularly suited to their pragmatic function, neither is ill-formed, and it appears possible to map automatically between the two. This mapping has applications in speech technology for the processing of conversational speech.

Index Terms: natural speech, transcription, ill-formed, written language, translation, statistical mapping

10.21437/Interspeech.2007-650

1. Introduction

Non-verbal behavior is often thought of as being limited to gesture and facial expression, but in this paper we show that *non-verbal speech features* are also extremely common. [1] has shown that `non-lexical' fragments are extremely common in conversational speech. From analysis of 150,000 transcribed conversational utterances, recorded from one speaker over a period of four years, it was found that 49% (almost half) were best described as `non-lexical'; i.e., that their intended meaning could not be adequately understood from an examination of a transcription of their transcribed text alone. They had to be heard for their intended effect to be understood. An example in English might be “Oh yeah”, where the words can be uttered as a challenge (“Do you really think so - I don't!”) or as confirmation (“yes”) or as indicating a sudden recollection (“yes, now I remember”), etc., and their meaning can sometimes be inferred from context (if only text is available) but are immediately obvious from the speech.

Table 1 provides detailed figures of occurrences of such utterances in a large Japanese corpus of transcribed conversational speech, and Table 2 shows some examples. Very few of these utterance types can be found as a lexical entry in a standard language dictionary, yet it has been experimentally confirmed that the intended meanings of many of these non-verbal utterances (or conversational `grunts') can be perceived consistently by listeners even when presented in isolation without any discourse context information [2]. In many cases, the intentions underlying the utterances can be appropriately and consistently paraphrased even by listeners of completely different cultural and linguistic backgrounds. This paper extends the analysis to include inarticulate fragments in longer utterances, and shows how a mapping can be performed between the spoken and written forms of the language using SMT, and offers an explanation for the so-called `ill-formed' nature of spontaneous speech.

Table 1. Occurrences of various utterances in the conversational speech corpus.

number of unique ‘lexical’ utterances	75242
number of ‘non-lexical’ utterances	73480
number of ‘non-lexical’ utterance types	4492
total number of utterances transcribed	148772
proportion of ‘non-lexical’ utterances	49.4%

2. Wrappers and Fillers

[3] has suggested that spontaneous conversational speech is better thought of as including both *A-type* and *I-type* components. The former functioning primarily for the expression of affect, speaker relationships, and discourse management, and the latter (I-type) functioning primarily to convey propositional content (or linguistic meaning).

The following explanation was proposed to account for the supposedly ill-formed nature of spontaneous colloquial speech:

“Whereas in written communication the word sequences are usually carefully deliberated and well-formed, in the case of spontaneous-speech the flow is generated in real-time and a stream of words and phrases might typically (in colloquial English) appear as follows:

“*... erm, anyway, you know what I mean, ..., it's like, er, sort of* **a stream of** *... er ...* **words, and phrases,** *all* **strung together,** *if you know what I mean, you know ...* "

where the words in bold-font form the content (or the *filling* of the utterance) and the italicized words form the *wrapping* or decoration around the content."

“Here the term `filler' is used to describe the I-type content (the text which would normally be included in a cleaned-up orthographic transcription of the utterance), and the term 'wrapper' is used to describe the A-type (affect displaying) portions of the utterance, that are often considered as ill-formed. This usage is in (deliberate) contrast to the usual interpretation of a `filler' as something which occupies a `gap' or a supposed empty space in a discourse. On the contrary, this paper suggests that by their very frequency, these non-propositional and often non-verbal speech sounds provide not just time for processing the spoken utterance but also a regular base for the comparison of fluctuations in voice-quality and speaking-style.”

“The biggest difference that is usually immediately apparent between I-type and A-type utterances is their length. Although some longer utterances such as `Good morning', `How are you today?', and `Did you see the game

last night?!' can be considered as primarily phatic, and hence A-type, the transfer of propositional information that defines I-type utterances usually requires more words to be strung together in a longer sequence. [4].”

Table 2. The most common complete utterances in the corpus, (data from one speaker, numbers show occurrence frequency). Note the highly repetitive nature of these common expressions

48038	うん	1733	で	829	ま
15555	あ	1675	ほんで	800	んんん
10961	ふん	1550	うんうん	787	まあ
8408	うーん	1535	もう	751	わかった
7769	え	1428	でも	737	や
5796	ああ	1422	ふんで	730	ありがとう
4891	ほんま	1412	はあ	713	あれ
4610	あー	1370	ええ	703	そうそうそう
3704	んん	1329	そう	692	は
3608	はい	1299	ふんん	692	そうなんや
3374	なんか	1291	ほんまあ	687	あたし
3164	ん	1246	うんうんうん	679	んんーん
3010	いや	1227	あのう	674	はいはい
2942	ふーん	1206	ううん	673	そうそうそうそう
2860	あの	1118	これ	658	フフ
2246	ふうん	1108	そうそう	645	せやな
2238	なあ	1085	おん	623	ほんなら
1871	そうなん	1079	まあな	599	うんうんうんうん
1761	な	903	あああ	588	ほん
1736	うんん	871	だから	583	よいしょ

3. Translating Conversational Speech

For the present study, we consider the corpus transcriptions of conversational speech to be a set of samples of a certain language, similar to but in specific ways different from the standard written form of the Japanese language. Parallels might be drawn (regarding the degree of difference only) between Spanish and Italian, or Middle English and current English. From this approach, we were able to use techniques developed in statistical machine translation for mapping between one `language' and the other; in this case between standard Japanese and its colloquial “street” equivalent.

We first produced a dictionary of frequent wrappers, without resource to linguistic knowledge, using a ‘longest-common-substring’ algorithm to identify the most frequent symbol sequences occurring at utterance-initial or utterance final positions in the transcribed corpus. As training data, we used the set of transcribed utterances having a length of between 20 and 40 kana characters (η=43,186). A kana symbol in the Japanese phonetic alphabet approximately corresponds to a syllable. By setting a threshold of 10 repetitions as a minimum criterion for inclusion, and then sorting the utterances and matching characters from left-to-right to obtain the longest common substring, we obtained 899 frequently occurring utterance-initial forms, and then by matching right-to-left (i.e., by sorting the reversed strings) we obtained 957 frequent utterance-final forms.

These “edge-pattern” wrapper sequences were then matched wherever they occurred utterance-internally and were used as further segmentation points to divide the longer utterances into `wrapper' and `filler' sections, with the edge patterns being defined as wrappers and the intervening sections assumed to be `fillers'. Figure 1 illustrates the result of this two-stage process. The `words' in bold font being the common (typically non-lexical) `wrappers'. Even to those who cannot read Japanese, it will be apparent from the figure that these are very frequent. Note that lines starting with a “#” are manually-produced transliterations and rarely include such terms. In a hand-checked subset of 1000 utterances we counted 2337 wrappers; an average of 2.34 per utterance. Note that single-character (single-syllable) wrappers are difficult to detect automatically without recourse to a morphological analysis of the transcription, so the actual number of occurrences may be much higher.

4. Machine Translation Approach

From the conversational speech corpus we were able to make use of about 1000 hand `translated' sentence pairs to train and test a statistical model. Statistical machine translation is an automatic method to do translation of a source language into a target language given a bilingual corpus of aligned sentences. [5] introduces work on the original IBM word-based models and later [6] proposed a method for creating models at the phrase level where a phrase is a grouping of contiguously aligned words. The model is based on the older noisy channel approach where we seek the source sentence that maximizes the probability of translation for a target sentence:

$$\phi(\bar{s}|\bar{t}) = \frac{count(\bar{s},\bar{t})}{\sum_{\bar{s}} count(\bar{s},\bar{t})} \quad (1)$$

Using Bayes' rule we can decompose the translation probability into a translation model and a language model. The model translation component $p(s|t)$ is estimated from the relative frequency counts of phrases.

$$argmax_t p(t|s) = argmax_t p(s|t)p(t) \quad (2)$$

$p(t)$ is a language model trained on the target side of the corpus. In addition the to translation probabilities these two features we use the lexical weight translation probabilities as a measure of how well each phrase's words translate to each other. For our experiments we used the freely available Pharaoh decoder which implements the phrase-based model described in [6] and also includes some additional model components: a simple distortion model based on the start position of a source and target phrase, a word penalty weight used to control the length of the target translation output. In order to train a phrase based model the IBM word alignments are needed and were trained using the GIZA++ toolkit [7].

Due to the small size of the corpus we use 10-fold cross validation as a way to generalize over the quality of the corpus and translations produced. We partitioned the corpus into sets of 105 sentences each and 113 for the last set. For each experiment a new corpus was created by removing the fold from the corpus and creating a phrase translation model. We use the default parameters of the Pharaoh decoder and as a method of evaluation the IBM BLEU metric was used to measure the quality of the translations of each fold test set. Afterwards, to increase the quality of the translations, minimum error rate training was used to optimize the model's feature weights, where each of the features described above has a weight component describing how much the feature should be relied upon when deciding the score of a translation.

あ,もしもし,あのちょっとけいやくのないようへんこうしていただきいんですけど
(もしもし。契約の内容を変更していただきたいのですが)
しらんゆう**ねんな**,ひつこい**ねんもう**ぴかぴかぴかぴかひかってるからきになってさあ
(知らないと言っているのに、しつこいぴかぴか光っているので気になって)
たべれん**ねん**で, たべれん**ねんけど**きもちわるいし,まだまだしんどいし, **みたいな**
(多分食べられます。食べられるのですが、気持ちが悪いしまだしんどい、と言った感じで)
だかほんまあんまたたかんでいいらしい**ねんけど**, **ま**,ちょっとほこりおとすていど
(だから本当に、あまり叩かなくて良いらしいです。少し埃を落とす程度で)
うんうんうん,でもさ,どうせさ,いろいろあつめんねやったら,これをしってたら
(どうせ色々集めるのなら、これを知っていれば)
うん,**そら,**こまま,こおりやまのほうちょっとまっすぐいったところ **やねん けどな**
(はい。このまま郡山の方へまっすぐ言ったところなのですが)
まあはんなどうろあるから**なあ**,やっぱりつうこうりょうすくないかもしれん**よなあ**
(まあ、阪奈道路があるからやっぱり交通量は少ないかもしれませんね)
それもかんがえようよな,**なんか**ほんまにきんてつでぜんぶすんねやったらいいけど
(それも考えようですよ。本当に近鉄で全部するのならいいけれど)
あるくのいたい,**む,なんか**どっちかはんぶんがすごいしびれてあるかれへん**ねん**て
(歩くのが痛い。どちらか半分がとても痺れて歩けないのです)
なんかめんどくさいな,おかしつねにかっとかなあかん**やんとか**おもっとってんけど
(何か面倒くさいなあ。お菓子は常に買っておかなければならないと思っていたのですが)
なんかさあ,あのかたちがちゃんとなってへんからはきにくいすりっぱってある**やん**

Figure 1. Sample utterances of Japanese conversational speech, selected at random from those having a length of between 20 and 40 mora in the corpus. Each utterance is followed by its equivalent transliteration in standard Japanese for comparison. Bold font shows the automatically-detected 'wrappers' in these utterances.

Table 3. Original BLEU scores and MERT optimized BLEU scores for the 10-fold cross validation experiment.

Fold	Original BLEU	MERT Optimized BLEU
1	34.06	64.84
2	37.96	64.91
3	37.83	68.38
4	34.78	56.65
5	32.62	43.98
6	29.23	32.24
7	68.48	100
8	31.71	60.55
9	28.97	37.14
10	30.00	31.29

Since the content words between colloquial speech and standard representation are highly similar there is a considerable and significant increase in accuracy after MERT training in most cases. Using SMT models for colloquial speech to formal speech translation is an interesting way to observe the similarity of grammar construction. It gives us the ability to answer questions about how common phrases become shortened or elongated and how content words are positioned in utterances.

5. Discussion

The special structure of spoken language has often been described as “ill-formed” but we maintain that it is ideally suited to the simultaneous expression of (a) propositional content (*i.e., linguistic information*) and (b) speaker-state, discourse management cues, and speaker-listener-relationships (*i.e., affective information*). By the frequent insertion of so-called “fillers” and other repetitive fragments such as laughs, grunts, etc., the speaker provides the listener with constant reference points for evaluating affective states, as displayed by subtle changes in voice-quality information. In this way, the supposedly ``ill-formed'' structure of spontaneous speech actually provides a mechanism whereby the speaker can express both propositional content *and* affective information simultaneously and in parallel within the same utterance.

However, for the synthesis or recognition of conversational speech, we need to be able to map between a more standard representation of an utterance and the more colloquial versions. In tightly-controlled specific task domains, people can be constrained to use the standard language when interfacing with a spoken dialogue system, but for ubiquitous computing environments where machines `co-exist' with human beings, it will be necessary to process the less well-formed utterance types typical of street-speech.

For speech synthesis in toys, or domestic appliances which interact with humans through speech, there may be no need for the extreme colloquialism as found in our corpus, but when robots, customer-care systems, and translation devices take part in a dialogue with a human, there may be a need for such non-standard speech - it is more effective in certain situations, and certainly more expressive.

This paper has presented results of an analysis of a 1,500-hour corpus of transcribed natural conversations and has shown that the forms of speech revealed by the transcriptions, while being very different from these described by a standard grammar of the language, are actually well suited to the simultaneous transmission of affective as well as linguistic information. The paper proposed a segmentation of these spoken utterances into wrappers and fillers, where the filling is the linguistic content, and the wrapping is the affective display, and proposed techniques for the automatic mapping form the standard forms of the language to the colloquial forms as observed in the corpus. This mapping uses tools developed for statistical machine translation and treats the spoken and written versions as two entirely different languages.

6. Future Work

Future work remains to refine automatic methods for classifying wrappers and fillers. Such a method could be used directly for translation purposes and for speech applications. In machine translation when dealing with speech corpora, information about wrappers and prosaic information, such as the way a wrapper is uttered can be used to produce a speech context model for creating the best translation not only off their lexical translations but also as important with speech, based on the way words and phrases are heard. From this we may be able to produce a technology that maps between the users stylized perception of their language and the real-world usage as described by the corpus. In the speech domain a future synthesizer may offer a `familiarity' slider option, then map from the same standard input (orthography) to an appropriate speaking style to express the content of the utterance in an appropriate manner.

7. Acknowledgments

This work is supported by the Japan Science & Technology Corporation (JST), the National Institute of Information and Communications Technology (NiCT), and the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan (SCOPE). The second author is especially grateful to the management of ATR for their continuing encouragement and support.

8. References

[1] Campbell, N., 2004. “Extra-Semantic Protocols; Input Requirements for the Synthesis of Dialogue Speech” in *Affective Dialogue Systems, Lecture Notes in Artificial Intelligence*, vol. 3068 Eds Andre, E.; Dybkjaer, L.; Minker, W.; Heisterkamp, P., New York, Springer. 221-228.

[2] Campbell, N., & Erickson, D., 2004. “What do people hear? A study of the perception of non-verbal affective information in conversational speech”, in *Jnl Phonetic Society of Japan.*

[3] Campbell, N., 2006. "On the structure of spoken language", in Proc Speech Prosody, Dresden, Germany, May 2006.

[4] Campbell, N., 2007, “Conversational Speech Synthesis and the need for some laughter”, IEEE Transactions on Audio, Speech, and Language Processing, Vol 14, No. 4, 1171-1178, July 2006.

[5] Brown, P.F., Della Pietra , S. A., Della Pietra , V. J., and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. Comp. Linguistics, 19(2):263–311.

[6] Koehn, Philipp, Och, Franz, and Marcu, Daniel. 2003. “Statistical phrase-based translation”. In Proceedings of HLT/NAACL, pages 127–133.

[7] Och, Franz Josef, & Ney, Hermann. 2003 "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.