Characterizing Spoken Dialog Corpora using Interaction Style Dimensions

Nigel G. Ward
Computer Science, University of Texas at El Paso
500 West University Avenue, El Paso, Texas 79968, USA
nigelward@acm.org

Abstract
The construction of dialog systems today relies heavily on appropriate corpora, but corpus selection is more an art than a science. This paper describes initial explorations with a new technique that may help: characterizing spoken dialog corpora in terms of a dimensional model of interaction styles. After discussing the need for more informative characterization of spoken dialog corpora and surveying related work, this paper presents the model, illustrates its potential utility by applying it to subsets of the Switchboard corpus, and outlines some next steps.

Keywords: corpus characterization, corpus similarity, corpus selection, vector-space model, dialog behaviors

1. Motivation
Today the process of selecting corpora in support of dialog systems training or tuning is rarely systematic. While dialog systems developers rely heavily on machine learning from corpora to acquire the various knowledge and parameters needed for effective systems, existing methods for corpus comparison rely mostly on lexical and topic overlap, e.g. (Pavlick and Nenkova, 2015), so it is difficult to predict how well other types of knowledge will transfer.

In particular, we are interested in parameters of interaction style, as these vary across domains, dialog purposes, and user demographics. Getting the style right is an essential part of providing a high quality user experience (Marge et al., in press 2021). This is no longer a distant goal, as core speech components now enable us to to implement situation-appropriate turn taking, politeness behaviors, rapport building strategies, and so on (Metcalf et al., 2019). Because our fundamental knowledge in these areas is still spotty, so developers rely on discovery or learning from corpora. Indeed, it is still common for a new development project to start with the collection of a task-specific corpus. However most would prefer to be able to better exploit existing corpora (Kashyap et al., 2021). One recent success was a socially well-behaved recommendation system for movies, created by discovering behaviors from a suitable subset of Switchboard data (Pecune et al., 2019). Selection of the subset was easy because Switchboard was designed around topics, and in particular the “movies” tag was available. However, in general, we would like to be able to rapidly delineate relevant corpus subsets, even when annotations are lacking. Another use for corpus characterization is for corpus design, selection, and mixing. For example, with the increasing interest in pretrained models, derived from large generic corpora but then tunable on smaller specific data to meet specific needs, there is an increasing need to make informed choices regarding the corpora on which to pretrain. Today this is largely based on convenience and hunches. For example, several recent large efforts have aimed to collect dialogs to exemplify a sweet-spot style that is simultaneously natural for humans and implementable with current technology (Budzianowski et al., 2018; Byrne et al., 2019). These efforts have used clever manipulations such as having dialogs be written down and then read, or interposing a text-to-speech engine between the confederate and the subjects. But we currently lack any ability to model how such manipulations affect style, let alone quantify in what respects such datasets match the needs of systems designers and developers. A good model of how corpora vary would help us better design such corpora or better supplement them to make the pretrained models more suitable for practical purposes.

Thus we need better ways to describe spoken dialog corpora, both to to help collect better corpora, and to help make better use of existing resources.

2. Precursor Work

2.1. Data-Driven Corpus Categorization
Biber, in his landmark contribution both to style description and to systematic corpus description, investigated what he termed “conversation text types” (Biber, 2004). Using transcripts as data and a text-based feature set (a few automatically-parsed grammatical structure elements and many lexical features, including a selection of discourse markers, parts of speech, and semantic classes), computed over diverse corpora, he used Principal Component Analysis to derive three dimensions of variation: information-focused vs interactive discourse, stance-focused vs context-focused discourse, and narrative-focused vs present-tense discourse. He showed how different conversations could be automatically located in this space.

This method has been very influential in the compar-
son of diverse text corpora, and also occasionally for speech corpora, albeit only using information in the transcripts (Shen and Kikuchi, 2014). However these models generally seem to have low explanatory power; for example, Biber’s three dimensions accounted for only 36% of the variance. They also generally have failed to exploit acoustic-prosodic features, although these potentially provide much more information than text. One exception is the work of Siegert et al, who demonstrated the value of acoustic-prosodic information for corpus selection, although unfortunately only for the narrow problem of training emotion recognizers (Siegert et al., 2018). Overall, work in this tradition appears not to have found practical use.

2.2. Describing Individual Style Variation

Style has long been a topic of interest in linguistics, with a very diverse set of style properties identified and studied (Troiano et al., 2021 submitted). Most work has focused on sentence- or utterance-level styles alone, and most computational work on style has focused on style and style transfer in text generation or speech synthesis alone. Styles at wider timescales and in speaker interactions have been less studied, although their importance is seen in our rich folk vocabulary for describing it, including terms like agreeable, fast-paced, playful, chatty, soft-spoken, businesslike, domineering, and many more. Much work in this area has been built on Tannen’s seminal observations on “conversational styles” (Tannen, 1989; Tannen, 1980). Despite many insightful observations and findings, work in this tradition has been mostly non-quantitative, and thus of limited practical utility.

More recently, computational models have been developed to study style in dialog (Grothendieck et al., 2011; Laskowski, 2016; Yamamoto et al., 2020). These works have used turn-taking and other features, to derive models of how individuals vary. However such approaches have not previously been applied to corpus characterization.

2.3. A Dimensional Model of Interaction Styles

The explorations reported in this paper use a recent model of interaction styles, originally developed for studying individual differences. The key ideas in this model, following previous work, are that it is dimensional, and that it was developed using speech-aware, interaction-related features.

The properties and development of this model are not the focus of the current paper, as they are presented in (Ward, 2021a) and described in full detail in the documentation at (Ward, 2021b), with the code available at (Ward, 2022). This subsection provides a brief overview, as a preliminary to considering its utility for corpus characterization in the next section.

This model takes as input any American English conversation, conversation fragment, corpus, or subcorpus, and represents its style as a vector of length 8: that is, it maps dialogs into a vector space representation of interaction styles.

The mapping process happens in eight steps.

1. Low-level (frame level) prosodic features are computed, specifically the raw pitch, intensity, and cepstral coefficients.
2. These are normalized by track (Ward, 2021c).
3. Filters and aggregation processes are applied to obtain mid-level features over various temporal spans, including estimates of intensity, speaking rate, phoneme lengthening, creakiness, enunciation or reduction, and the extent to which the pitch is high or low, or wide or narrow.
4. These mid-level features are normalized using parameters that brought each to mean 0 and standard deviation 1 on the training data.
5. The match of these normalized features to 12 meaningful temporal configurations is computed every 20 milliseconds. These meaningful temporal configurations represent specific American English prosodic constructions (Ward, 2019), which convey meanings and activities such as a turn switch, topic closing, enthusiasm, positive assessment, and contrast. These provide an inclusive set to track behaviors relating to a wide range of dialog states, activities, behaviors and interactive events, including many of those often considered most important in human interaction (Couper-Kuhlen and Selting, 2018).
6. The match values are binned and pooled across 30-second fragments. There are 7 bins per configuration, thus there are bins for when a speaker is expressing a strong, mild, or weak contrast, or managing an ambiguous, clear, or strong turn switch, and so on.
7. The resulting 84 values are rotated to a dimensional representation where the top dimensions capture most of the variance.
8. The top 8 dimensions are retained.

Further, each of the eight dimensions has an interpretation, as summarized in Table I. Although the derived from only prosodic information, the resulting dimensions turned out to have interesting lexical and topic correlates. The interpretations were accordingly based on subjective impressions and on examination of dimension loadings, lexical tendencies, LIWC category tendencies. They are revised slightly from those previously reported (Ward, 2021a), after examination of more data. While often useful, the interpretations are of course not needed for uses such as similarity estimation.

It is worth noting in passing that the model was created from data. The details are outside the scope of this
oung words like so and anyway often marks extreme topic shifts, appearing often in these fragments the associated breathy constriction. This is generally used for off-topic utterances, since interaction styles are neither instantaneous nor constant over long times. We chose to take these from Switchboard because it is large, has good metadata, and is well-studied (Godfrey et al., 1992; Godfrey et al., 1997). We developed the model using a subset of 33022 fragments, including 335 speakers. We applied steps 1–6 above to each fragment in the training data to get an initial representation, and then applied Principal Component Analysis to obtain a concise 8-dimensional representation. We chose eight dimensions were chosen because they explained over half (53%) of the variance, and because each was meaningfully interpretable.

The rest of this section briefly explains the interpretations of the dimensions. Evidence and further discussion appears at the companion website (Ward, 2021b).

Dimension 1 relates simply to the amount of shared engagement. Dimension 2 is very high or low when one speaker versus the other is taking an active speaking role and the other an active listening role. Dimension 3 involves expressing positive assessment, for example when talking about the speaker’s dog, a good fishing day, or a favorite football team, versus expressing negative feelings, for example about income tax, lawn problems, the futility of oversees aid, or time flying by. Dimension 4 is very high or low when one speaker versus the other is being confident and/or dominant as they talk about something they know well, such as their woodworking hobby or school funding, while the other is acknowledging the other as an expert on the topic.

Dimension 5 deserves a little more discussion, as the relevance to interaction style features is not obvious. The positive pole of Dimension 5 involves a thoughtful style. Among the prosodic feature loadings, the highest loading is for the Topic Closing Construction, with the associated low pitch typically used in these fragments to express a stance of calm rationality, as the speaker describes something they know well, such as how a network is set up or how security cameras work. Here there was apparently a low tolerance for silence, with the speakers often rattling on and often buying time with repetitions of words or phrases.

Dimension 6 relates to a resigned attitude, for example when taking about a corporate promotion policy or the prevalence of gun-safety carelessness versus a positive, change-oriented outlook, for example when discussing new exercise regimens, changes in women’s roles, or medical research advances. Dimension 7 relates to stating and justifying opinions, for example regarding general claims about dealing with people or situations versus finding common ground, for example when talking about similar experiences with catalog shopping, making hamburger, or drug testing. Dimension 8 involves the continuum between talk about remote or currently unimportant and half-understood or half-remembered ideas or events versus expressing strong opinions, for example regarding people or practices that are strongly disliked or strongly admired.

3. Illustrations of the Potential Value

Thus we have a model which can compute a reduced-dimensionality representation of the interaction style for any corpus, subcorpus, conversation, or conversational fragment. This section describes some initial explorations of how this can be used for corpus characterization and specific practical purposes.

3.1. Corpus Characterization

This model supports visualization of subcorpus differences. As an example, if we view Switchboard as a collection of subcorpora, one per topic, we can map out roughly where each topic lies, by plotting the average interaction style of all fragments within that topic. In the Appendix, Figures 1 and 2 show two single projections of interaction style space, and Figure 3 shows two additional dimensions, after a little preprocessing (specifically using the averages of the absolute values,
3.2. Similarity Estimation

The model supports similarity estimation \cite{Kilgariff1998}, for now by simply using the Euclidean distance in the 8-dimensional space. For example, considering Switchboard’s 20 most distinctive topics, the closest pair was: “politics,” at \([1.0, 2.6, 0.1, 1.6, 0.4, 0.1, 0.1, -0.2]\) and “capital punishment,” at \([1.1, 2.7, 0.3, 1.6, 0.5, 0.0, 0.0, -0.0]\). The other most similar pairs were “baseball” and “football,” “weather/climate” and “vacation spots,” and “movies” and “TV programs.” Such similarity estimates could be used to support targeted data augmentation. Considering again the scenario of seeking data to train a movie recommendation system, dialogs about movies have an average style of \([-1.6, -0.0, 0.5, 0.0, -0.7, -0.5, 0.3, -1]\), to which the most similar subcorpora were “TV programs” were “clothing and dress,” “football,” and “baseball.” This indicates what subcorpora would be potentially most compatible as supplementary data.

3.3. Identifying Outliers

The similarity metric can of course be used not only to find close corpora, but also distant corpora or subcorpora, as suggested by the figures. Moreover, the distance measure can support data cleaning. For example, many conversations in Switchboard have the “movies” tag, but not all fragments are equally relevant, either to the topic or to the general style. Identifying marginal fragments can be useful, and the model can identify atypical fragments as those most distant from the average interaction style. For the movies topic, examination of the five most distant fragments revealed that these were indeed mostly atypical — two involved strong moral judgments, and one was mostly about audience behavior — and would be good candidates for exclusion from the training set for a normal, upbeat movie recommending system.

If interaction-style similarity were no more than a reflection of topic similarity, its utility would be low, but these examples illustrate how it provides added value.

4. Next Steps

These initial explorations suggest that the model could have immediate practical value. Nevertheless many fundamental questions remain.

This model was built using only data from Switchboard, and thus only casual conversations. While the dimensions obtained are quite general, and it is possible that Switchboard collection is diverse enough to serve as a microcosm for all of American English, future work should explore this, and should try applying these methods to larger, multi-genre data collections.

This model was built using only automatically computed features and simple mathematical procedures. While the result seems related to human perceptions of important aspects of style, this needs to be checked. A proper examination would require, for example, a collection of human judgments of various dimensions, plus similarity judgments for calibration of distance estimates. Future work should also explore how style perceptions and preferences vary among individuals. This model at least provides a starting place for such investigations.

5. Summary

A priority for our field is enabling better use of existing language resources. Spoken data is fundamentally richer than text data, and this brief report has proposed a new way to exploit this, using a dimensional model of interaction styles in spoken dialog. There are many open questions and avenues for deeper investigation, but also reasons to expect this approach and toolkit to have immediate practical utility.

6. Acknowledgments

I thank Jonathan E. Avila for helping refine the dimension interpretations.
7. Bibliographical References


8. Language Resource References


A. Appendix
Figure 1: Average Interaction Styles of Some Topics in Switchboard, Projected to Interaction Style Dimensions 1 and 3. (0,0) is the global average style. The axis units are standard deviations computed over all conversation fragments. The topic names shown are just the mnemonics for the actual sentence-length prompts given to the participants.
Figure 2: As above, Dimensions 5 and 6.
Figure 3: As above, averages of absolute values, Dimensions 2 and 4.