Naturalization of Text

Parth Vipul Shah
Department of Computer Science
University of Southern California
pvshah@usc.edu

Ryan Luu
Department of Computer Science
University of Southern California
rluu@usc.edu

Bowman Brown
Department of Computer Science
University of Southern California
bnbrown@usc.edu

Ira Deshmukh
Department of Computer Science
University of Southern California
ideshmuk@usc.edu

Alfianto Widodo
Department of Computer Science
University of Southern California
awidodo@usc.edu

1 Project Domain & Goals

A major factor that discriminates spontaneous speech from written text is the presence of paralinguistic features such as filled pauses (fillers), false starts, laughter, disfluencies and discourse markers that are beyond the framework of formal grammars. There are multiple ways of achieving natural speech, but the focus of our research are speech disfluencies.

A speech disfluency is any of various breaks, irregularities, or non-lexical vocables which occur within the flow of otherwise fluent speech. This can include filler grunts and noises such as ”uh” or ”um”. Speakers use ”uh” and ”um” to announce that they are initiating what they expect to be a minor or major delays in speaking. A speaker can use these announcements in turn to implicate that they are searching for a word, are deciding what to say next, want to keep the floor, or want to cede the floor. Understanding these cues and their impact on the context of a given conversation can be said to be considered a natural part of human dialogue.

Therefore we aim to transform raw text documents of spoken dialogue / speech into its most natural-sounding version by augmenting these documents with disfluencies. Using various NLP techniques such as bi-grams and context analysis, we aim to create a model that can detect the most appropriate places in the document to insert these disfluencies and which form of it to insert. A successful implementation of the ideas examined in this proposal can aid in the efforts of naturalizing speech synthesis e.g. Google Duplex, Google Home, Alexa.

2 Related Works

Recent innovations in natural language processing have significantly improved the ability for artificial intelligence applications to both process and create structurally and logically correct text. While generated text may read naturally, there is a definitive gap when it comes to natural sounding artificial speech. Approaches to naturalizing any kind of generated language are few and far between, despite users deeming less artificial sounding voices as better (Fischer et. al.) and more intelligible (Pisoni et. al.).

Contrary to this papers proposal to introduce disfluencies, current approaches to creating natural sounding speech attempts to generate speech waveforms directly instead of adding additional natural sounding elements to text that can be processed (Tan et. al. , Kong et. al and Li et. al.). Furthermore, many of the works involving disfluencies and natural human speech patterns is centered around removing disfluencies and naturalizations in order to better process input speech (Wang et. al. , etc…) as challenges may arise when attempting to process disfluencies downstream. Other attempts to naturalize synthetic voices involve rephrasing speech commands (Einolghozati et. al.) or generating the rhythms that are common in natural human speech (Kharitonov et. al. and Lee)

There is very little work that involves including disfluencies as an efficient means of naturalizing text. A recent example, LARD (Passali et. al.) allows for the addition of “repetitions, replacements and restarts” to the data for simple disfluency generation, but is intended to modify sets of training data in order to allow for better disfluency detection (for later removal). (Wester et. al.) takes a psychological approach and works to add disfluencies as a way to measure perceptions of synthetic voices. Our method is efficient and makes use of a psychological understanding of disfluency generation (Bakti) in human speaking in order to create a dialogue script that can efficiently allow lower quality synthetic voices to sound more human.

3 Datasets

As a result of seeking to introduce disfluencies into conversations, we are looking to capture the closest imitation of natural human speech in the form of a text file. Meaning we are looking for a spoken corpora that can correlate natural spoken dialogue to the level of possibly containing individual intonation units. Towards that end, we have identified datasets such as (but not limited to) the (Santa Barbara Corpus,), (SRI American Express travel agent dialogue corpus,), and (Marilyn A. Walker, 2012) corpus of film dialogue as suitable starting points of data from which to facilitate our experiment. More potential corpuses that we have identified can be found in the references segment of this proposal. Here is an example line of dialogue from the Santa Barbara Corpus:

LENORE: ... So you don't need to go ... borrow equipment from anybody, ... to – do the feet? ... [Do the hooves?]

Since our data captures intonations, disfluencies, and other various forms of nonstandard formal grammar, we are able to extract those lines in an effort to train our naturalization text model. It is worth it to admit that there will exist inconsistency across our data due to us using dialogue from different environments such movie scripts, interviews, and natural occurring conversations. Therefore, much of the preprocessing and cleaning up of our data will have to done by hand by our team.

As a basis, our data preprocessing steps will currently involve or consider:

1) Extracting the lines containing disfluencies such as pauses and ”uh”, ”um” to form an intermediate corpus. The lines should be properly tagged with their respective POS or relevant marker. This can be done using tools such as NLTK or regex.
2) Extraction of the lines where intonation units are present. While not standardized, they are typically represented by special characters such as ”#” within a given corpus.
3) Consideration of special cases where we will have to identify and extract multiple variations of the above disfluencies such as ”uhh” and ”uhm”.

4 Technical Challenges

There are multiple ways to naturalize sentences. It can be achieved by the insertion of pauses and filler words or by intonating. Furthermore, there are multiple natural sounding variations of the same input sentence. This makes evaluation particularly challenging.

Multiple studies (Dall et. al., 2014) have employed humans to evaluate machine translated text. However, humans are inconsistent when it comes to naturalizing speech. The same speaker may transform the same sentence differently on different occasions based on context. However, this can be allayed by employing the MOS (Mean Opinion Score) test or the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test while conducting the survey. Human evaluation is time consuming, expensive and inconsistent.

Other studies (Hu et. al., 2022) have proposed using automated evaluation to quantitatively measure the transformed text against high quality, annotated training data. Cosine Similarity, Word Overlap, BLEU and source-BLEU are some of the metrics that can be employed for evaluation. Automated evaluation may yield poor scores that do not accurately represent the quality of naturalization.

The final challenge is obtaining high quality datasets with pauses and filler words inserted for training. Not only do speaker turns (Heldner and Edlund, 2010) influence pauses but filler words ”uh” and ”um” conform to phonology, prosody, syntax, semantics and pragmatics just like other English words. (Clark and Fox Tree, 2002) This makes insertions context specific and more complex due to the interplay of these words and plain sentences.

5 Division of Labor

Since we have multiple different corpora that we would like to work with, each individual team member will be tasked with preprocessing their own corpora of choice using a standardized schematic that we all agree upon beforehand. After that, using some versioning tool such as github, we can all contribute to building the model. The assigning of roles to perform tasks such as splitting the data, evaluation, testing, writing up the reports, etc., can all be decided at a later date after the main task of preprocessing is completed.

References

[Wikipedia 2022] Wikipedia Speech Disfluency

https://en.wikipedia.org/wiki/Speech_disfluency

[Clark and Fox Tree 2002] Alfred V. Aho and Jeffrey D. Ullman. 2002. *Using uh and um in spontaneous dialog. Cognition.*

[Heldner and Edlund 2010] Mattias Heldner and Jens Edlund. 2010. *Pauses, gaps and overlaps in conversations*

[Hu et. al. 2022] Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal and Aston Zhang. 2022. *Text Style Transfer: A Review and Experimental Evaluation*

[Dall et. al. 2014] Rasmus Dall, Marcus Tomalin, Mirjam Wester, William Byrne and Simon King. 2014. *Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis*

[Santa Barbara Corpus] Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

www.linguistics.ucsb.edu/research/santa-barbara-corpus

[Kaggle Movie Dialouge Corpus] https://www.kaggle.com/datasets/Cornell-University/movie-dialog-corpus

[SRI American Express travel agent dialogue corpus] www.ai.sri.com/ communic/amex/amex.html

[TRAINS Dialogue Corpus] www.cs.rochester.edu/research/speech/trains.html

[Collection of Dialogue Corpora] martinweisser.org/corpora_site/dialogue_corpora.html

[Marilyn A. Walker 2012] Marilyn A. Walker, Grace I. Lin, Jennifer E. Sawyer. ”An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style.” In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012.

[Kühne K, Fischer MH, Zhou Y. 2020] The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. Front Neurorobot. 2020 Dec 16;14:593732. doi: 10.3389/fnbot.2020.593732. PMID: 33390923; PMCID: PMC7772241.]

[Pisoni DB, Manous LM, Dedina MJ 1987] Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. Comput Speech Lang. 1987 Sep;2(3-4):303-320. doi: 10.1016/0885-2308(87)90014-3. PMID: 23226919; PMCID: PMC3515065.]

[Tan et. al. 2022] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, Tie-Yan Liu 2022. *NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality*

[Kong et. al. 2020] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae 2020. *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech SynthesisQuality*

[Li et. al. 2021] Yinghao Aaron Li, Ali Zare, Nima Mesgarani 2021. *StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion*

[Wang et. al. 2022] Wang, Shaolei and Wang, Zhongyuan and Che, Wanxiang and Zhao, Sendong and Liu, Ting 2022. *Combining Self-supervised Learning and Active Learning for Disfluency Detection*

[Einolghozati et. al. 2020] Arash Einolghozati and Anchit Gupta and Keith A. Diedrick and S. Gupta 2020. *Sound Natural: Content Rephrasing in Dialog Systems*

[Kharitonov et. al. 2022] Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, Wei-Ning Hsu 2022. *Text-Free Prosody-Aware Generative Spoken Language Modeling*

[Lee et. al. 2021] Jaeryoung Lee 2021. *Generating Robotic Speech Prosody for Human Robot Interaction: A Preliminary Study*

[Passali et. al. 2022] T. Passali, T. Mavropoulos, G. Tsoumakas, G. Meditskos, S. Vrochidis 2022. *LARD: Large-scale Artificial Disfluency Generation*

[Wester et. al. 2015] Mirjam Wester, Matthew Aylett, Marcus Tomalin, Rasmus Dall 2015. *Artificial Personality and Disfluency*

[Bakti et. al. 2009] Maria Bakti 2009. *Speech Disfluencies in Simultaneous Interpretation*