Intership subject for the Research in Computer Science (SIF) master

Make text look like speech: disfluency generation using sequence-to-sequence neural networks

Supervisor

Gwénolé Lecorvé (gwenole.lecorve@irisa.fr)

Hosting structure

Laboratory : IRISA
Team : Expression
City : Lannion
Description of the team : The team Expression focuses on expressivity in human languages. It has strong skills in text-to-speech and frequently takes part in the international speech synthesis challenges.

Keywords

Natural language and speech processing ; natural language generation ; sequence-to-sequence (recurrent) neural networks ; deep learning.

Appreciated skills

- Interest in artificial intelligence, natural language, speech, machine learning and/or human-machine interaction
- Skills in Python, Linux and shell scripting.

Context of the intership

Speech disfluencies can be defined as a phenomenon which interrupts the flow of speech and does not add any propositional content [1] For instance, the utterance "*She came from Boston on uh from Denver on Monday*" is interrupted by the token "*uh*" before the speaker made a mistake, which he/she just repaired in a second part [2]. Disfluencies mainly occur when the speed of speaking becomes faster than the speed of thinking, which is particularly frequent in spontaneous (i.e. unprepared) speech.

Despite the lack of propositional or semantic content, disfluencies have several communicative values. They facilitate synchronization of speech between addressees in conversations, and improve listening comprehension by creating delays in speech and signaling the complexity of the upcoming message. In spite of this, most current speech synthesis systems only partially integrate disfluencies. As a consequence, human-machine interactions are often perceived as relatively unnatural and inexpressive. Therefore, being able to automatically generate disfluencies is crucial to propose advanced interactive applications (video games, intelligent personal assistants, etc.).

A first exploratory method for the automatic insertion of disfluencies has been recently proposed within team [3, 4]. This method relies on a formalization where disfluencies are decomposed into revisions, repetitions and pauses, and defines disfluency insertion as a theoretical process where these elementary transformations are iteratively composed. A proof of concept of the proposed method has been validated through an implementation based on conditional random fields and language models on an English corpus. This work differs from most of the previous work which either focus on the detection or cleaning of disfluencies in speech transcripts [5], or solely concentrate on pause insertion and signal processing issues [6, 7].

Description of the internship

The goal of the internship is to improve the proof-of-concept method already developped in the team by training advanced neural networks and refining the formalization of the disfluency phenomenon. More precisely, the objectives are :

- Propose neural models to turn out the problem as a sequence-to-sequence task (sequential labelling, machine translation, etc.)
- If needed, enhance the formalization of the disfluency generation process
- Train models on large corpora using a dedicated GPU server
- Test the models on English and/or French
- Propose evaluation metrics
- Evaluate and compare the proposed models.

Bibliography

[1] J. E. F. Tree, “The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech,” Journal of Memory and Language, vol. 34, 1995.

[2] E. E. Shriberg, “Preliminaries to a theory of speech disfluencies,” Ph.D. dissertation, University of California, 1994.

[3] R. Qader. "Pronunciation and disfluency modelling for spontaneous speech synthesis". Ph.D. dissertation, University of Rennes 1, 2017.

[4] R. Qader, G. Lecorvé, D. Lolive, P. Sébillot. "Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept". Proc. of TALN, 2017.

[5] H. Hassan, L. Schwartz, D. Hakkani-Tür, and G. Tür, “Segmentation and disfluency removal for conversational speech translation.” in Proc. of Interspeech, 2014.

[6] R. Dall, M. Tomalin, M. Wester, W. J. Byrne, and S. King, “Investigating automatic & human filled pause insertion for speech synthesis.” in Proc. of Interspeech, 2014.

[7] S. Betz, P. Wagner, and D. Schlangen, “Micro-structure of disfluencies: Basics for conversational speech synthesis,” Proc. of Interspeech, 2015.