A German Dialogue Dataset for Tech Support
COMP 551
Applied Machine Learning

Thomas Page — thomas.page@mail.mcgill.ca — 26XXXXXXX
David Tamrazov — david.tamrazov@mail.mcgill.ca — 26XXXXXXX
Alexander Wong — alexander.wong4@mail.mcgill.ca — 260602944

September 21, 2017

URL: www.oururl.com

1 Overview

Our dataset is gathered from Apple Support Community discussion boards, and would be useful for training machine learning systems specifically for the task of tech support in german. The applicability of this dataset is definitely most geared towards supporting Apple-related questions, but probably can be generalized to most technical customer service platforms.

2 Dataset Description

As the data is collected from tech support discussion boards, it contains conversations specific

3 Discussion

One of the key design decisions made was how were we going to define a conversation and therefore what discussion threads would make the cut for the dataset? Our initial intuition was to only include discussions with at least one instance of a response from the original poster (OP), in addition to their original post. We then revised this to at least one instance of non-consecutive posts by any user, for the inclusion of the OP does not preclude a thread from being a conversation among others.

However, we eventually settled on the current definition of a conversation, which simply includes any thread with at least one response. While this may seem to disregard the objective of collecting data on conversations, the primary objective for this data set is to collect examples of good technical support, and one may argue that perhaps the best tech support takes the form of a singular, concise response (complexity and obscurity of the original question being a difference maker in this regard).

4 Statement of Contributions

Thomas got the ball rolling by introducing us to the BeautifulSoup python library that played a large roll in simplifying the task from the coding side. He also wrote the initial scripts to test feasibility, and suggested sourcing our data from apple tech support pages. Upon agreeing on our objective, we then worked cooperatively to create a program that extracted the data we wanted from the pages, without double counting or skipping posts. Dave and Alex worked on optimizing Thomas's initial program, and implemented the logic for testing each discussion thread for our criteria of what defined a conversation.

Thomas ran the program from home (Xhrs), and all three of us did some data analytics to include in the report. Alex wrote the first draft of the report, and from there everyone contributed to finalizing it for submission.

We hereby state that all the work presented in this report is that of the authors.