Server-Side Proxy Data for Entity Selections
Smart Text Selection, which is the same technology behind [Smart Linkify](https://ai.googleblog.com/2018/08/the-machine-learning-behind-android.html), does not predict arbitrary selections, but focuses on well-defined entities, such as addresses or phone numbers, and tries to predict the selection bounds for those categories. In the absence of multi-word entities, the model is trained to only select a single word in order to minimize the frequency of making multi-word selections in error.

The Smart Text Selection feature was originally trained using proxy data sourced from web pages to which [schema.org](http://schema.org/) annotations had been applied. These entities were then embedded in a selection of random text, and the model was trained to select just the entity, without spilling over into the random text surrounding it.

While this approach of training on schema.org-annotations worked, it had several limitations. The data was quite different from text that we expect users see on-device. For example, websites with schema.org annotations typically have entities with more proper formatting than what users might type on their phones. In addition, the text samples in which the entities were embedded for training were random and did not reflect realistic context on-device.