How can we fix that?

Simple idea:

- Build a vocabulary (e.g. by word frequencies)
- Substitute OOV words by <UNK> (both in train and test!)
- Compute counts as usual for all tokens
- Profit!