Pinyin to Hanzi Translation

Lucas Freitas（傅恭澤）
Cynthia Meng (孟笙寒)

The Problem

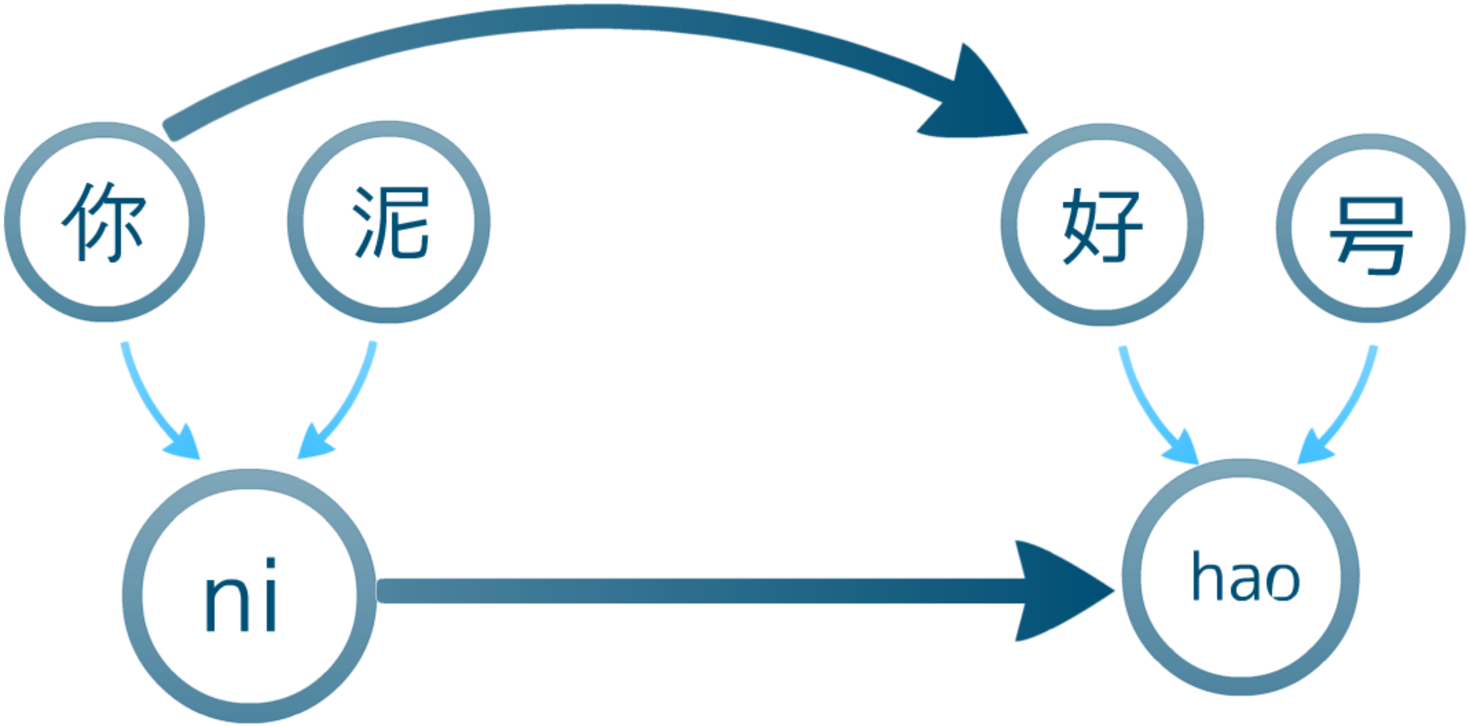
- Unsegmented Pinyin might have many different potential meanings
- Examples:
  - si for 死 (death) and 四 (four)
  - ni 你 (you) and 泥 (dirt)
- In addition to that, a given Chinese character can have more than one pronunciation (and thus, more than one Pinyin spelling)
- 了 can be pronounced "liao" or "le" depending on the context

Design Decision

- We contemplated using different methods for pinyin segmentation.
- First, we obviously considered TANGO
- Second, we debated between a Hidden Markov Model and a Segmented Hidden Markov Model
- We finally decided on the HMM because of efficiency reasons
- Test data: Alice in Wonderland (Chinese version)

What is an HMM?

The gist: we have an output, but we do not know the exact input (they are hidden)



Demo

- We implemented a UI as well, in the hopes of making our program more accessible to other users
- Basically, the input is a string of unsegmented pinyin, and the output are Hanzi characters

Demo

As of 05/06, precision was around 0.67, but we want to optimize our code to get closer to 0.78 like in the Literature. We want to change this slide before presenting.