Interpretese

vs

Translationese

Interpretese???

- what are the distinct characteristics of simultaneously interpreted language compared to batch-translated language?
- what strategies are used by human interpreters?
- can machines learn to apply tactics of human interpreters?

General strategy

- Not enough just to look at “output” of interpretation

- Corpus:
  - Japanese interpreted into English
  - Post-hoc elicited batch translations for the same J

- Differences between E^{interp} and E^{batch} will show difference between interpretese and translationese

Corpus

- Spoken monologues and interpretations (Matsubara et al., 2012)
- Lectures covering technology themes
- Two-to-four interpreters
- Post-hoc translated via Gengo
- 1684 segment pairs, 33 tokens/segment

  (about 56k tokens)

Example

つまり例えばこの表現一は認識できますが二から四は認識できない

Interp
The phrase number one only is accepted and phrases two, three, four were not accepted.

Batch
They might recognize expression one but not expression two to four.

General diffs of Interp vs Batch

- Inversion
  - Segmentation into multiple sentences
  - Passivization of single sentence
- Word generalization
  - (lower retrieval time)
- Summarization and omission
  - (to catch up)

Example (gen + segment)

(S) この日本語の待遇表現の特徴ですが英語から日本語へ直訳しただけでは表現できないといった特徴があります

(T) One of the characteristics of **honorific** Japanese is that it can not be **adequately** expressed when using a direct translation from English to Japanese.

(I) Now let me talk about the characteristic of the Japanese **polite** expressions. `<segment/>` And such expressions can not be expressed **enough** just by translating directly.

Example (gen + passivize)

(S) 以上のお話をまとめますと自然な発話というものを扱うことができる音声対話の方法ということを考案しました .

(T) In summary, we have **devised** a way for voice interaction systems to handle natural speech.

(I) And this is the summary of what I have so far stated. The spontaneous speech can be dealt with by the speech dialog method and that method was **proposed**.

Example (gen + summarize)

(S) で三番目の特徴としてはですねえ出来る限り自然な日本語の話言葉とてその出力をするといったような特徴があります .

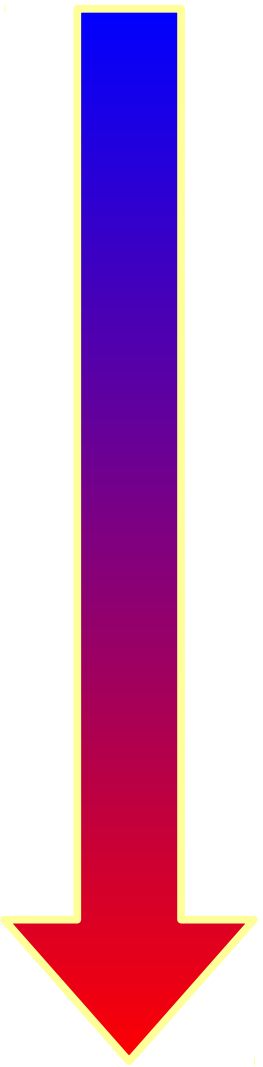
(T) Its third **characteristic** is that its output is, as much as possible, in the natural language of spoken ((Japanese)).

(I) And the third **feature** is that the translation could be produced in a very natural spoken language.

Distinguishing interp from batch

- coordinating conjunctions
- repeated content words
- demonstratives
- sentence boundaries
- proper nouns

  ...
- number of content
- high token/type ratio
- high stem/type ratio
- lots of pronouns
- passive sentences



Distinguishing interp from batch

Indicative of Interpretese

- coordinating conjunctions
- repeated content words
- demonstratives
- sentence boundaries
- proper nouns

Segmentation

...

- number of content words
- high token/type ratio
- high stem/type ratio

Generalization

- lots of pronouns
- passive sentences

Passivization

Indicative of Translationese

Stuff we're doing now

- Better understanding of interpretese (more complex features, etc.)
- Automatically predicting when to passivize or segment (use as actions in Jordan's RL framework)
- Question: generalization is useful, but hurts us at evaluation time?
- Question: should we be trying to get good batch translations faster, or good simultaneous interpretations?