Interpethes vs. Translationes: The Uniqueness of Human Strategies in Simultaneous Interpretation

He He
Computer Science
University of Maryland
hhe@cs.umd.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Jordan.Boyd.Graber@colorado.edu

Hal Daumé III
Computer Science and UMIACS
University of Maryland
hal@cs.umd.edu

Abstract

Computational approaches to simultaneous interpretation are stymied by how little we know about the tactics human interpreters use. We produce a parallel corpus of translated and simultaneously interpreted text and study differences between them through a computational approach. Our analysis reveals that human interpreters regularly apply several effective tactics to reduce translation latency, including sentence segmentation and passivization. In addition to these unique, clever strategies, we show that limited human memory also causes other idiosyncratic properties of human interpretation such as generalization and omission of source content.

1 Human Simultaneous Interpretation

Although simultaneous interpretation has a key role in today’s international community, it remains under-explored within machine translation (MT). One key challenge is to achieve a good quality/speed trade-off: deciding when, what, and how to translate. In this study, we take a data-driven, comparative approach and examine: (i) What distinguishes simultaneously interpreted text (Interpethes) from batch-translated text (Translationes)? (ii) What strategies do human interpreters use?

Most previous work focuses on qualitative analysis (Bendazzoli and Sandrelli, 2005; Camayd-Freixas, 2011; Shimizu et al., 2014) or pattern counting (Tohyama and Matsubara, 2006; Sridhar et al., 2013). In contrast, we use a more systematic approach based on feature selection and statistical tests. In addition, most work ignores translated text, making it hard to isolate strategies applied by interpreters as opposed to general strategies needed for any translation. Shimizu et al. (2014) are the first to take a comparative approach; however, they directly train MT systems on the interpretation corpus without explicitly examining interpretation tactics. While some techniques can be learned implicitly, the model may also learn undesirable behavior such as omission and simplification: byproducts of limited human working memory (Section 4).

Prior work studies simultaneous interpretation of Japanese↔English (Tohyama and Matsubara, 2006; Shimizu et al., 2014) and Spanish↔English (Sridhar et al., 2013). We focus on Japanese↔English interpretation. Since information required by the target English sentence often comes late in the source Japanese sentence (e.g., the verb, the noun being modified), we expect it to reveal a richer set of tactics. Our contributions are three-fold. First, we collect new human translations for an existing simultaneous interpretation corpus, which can benefit future comparative research. Second, we use classification and feature selection methods to examine linguistic characteris-

1 Unlike consecutive interpretation (speakers stop after a complete thought and wait for the interpreter), simultaneous interpretation has the interpreter to translate while listening to speakers.

2 Language produced in the process of translation is often considered a dialect of the target language: “Translationes” (Baker, 1993). Thus, “Interpethes” refers to interpreted language.

3 The tactics are consistent with those discovered on other language pairs in prior work, with additional ones specific to head-final to head-initial languages.

https://github.com/hhexiy/interpretese
tics comparatively. Third, we categorize human interpretation strategies, including word reordering tactics and summarization tactics. Our results help linguists understand simultaneous interpretation and help computer scientists build better automatic interpretation systems.

2 Distinguishing Translationese and Interpretese

In this section, we discuss strategies used in Interpretese, which we detect automatically in the next section. Our hypothesis is that tactics used by interpreters roughly fall in two non-exclusive categories: (i) delay minimization, to enable prompt translation by arranging target words in an order similar to the source; (ii) memory footprint minimization, to avoid overloading working memory by reducing communicated information.

Segmentation Interpreters often break source sentences into multiple smaller sentences (Camayd-Freixas, 2011; Shimizu et al., 2013), a process we call segmentation. This is different from what is commonly used in speech translation systems (Fujita et al., 2013; Oda et al., 2014), where translations of segments are directly concatenated. Instead, humans try to incorporate new information into the precedent partial translation, e.g., using “which is” to put it in a clause (Table 1, Example 3), or creating a new sentence joined by conjunctions (Table 1, Example 5).

Passivization Passivization is useful for interpreting from head-final languages (e.g., Japanese, German) to head-initial languages (e.g., English, French) (He et al., 2015). Because the verb is needed early in the target sentence but only appears at the end of the source sentence, an obvious strategy is to wait for the final verb. However, if the interpreter uses passive voice, they can start translating immediately and append the verb at the end (Table 1 Examples 4–5). During passivization, the subject is often omitted when obvious from context.

Generalization Camayd-Freixas (2011) and Al-Khanji et al. (2000) observe that interpreters focus on delivering the gist of a sentence rather than duplicating the nuanced meaning of each word. More frequent words are chosen as their retrieval time is faster (Dell and O’Séaghdha, 1992; Cuetos et al., 2006) (e.g., “honorific” versus “polite” in Table 1 Example 1). Although Volansky et al. (2013) show that generalization happens in translation too, it is likely more frequent in Interpretese given the severe time constraints.

Summarization Faced with overwhelming information, interpreters need efficient ways to encode meaning. Less important words, or even a whole sentence can drop, especially when the interpreter falls behind the speaker. In Table 1 Example 2, the literal translation “as much as possible” is reduced to “very”, and the adjective “Japanese” is omitted.

Before we study these characteristics quantitatively in the next section, we visualize Interpretese and Translationese by a word cloud in Figure 1. The size of each word is proportional to the difference between its frequencies in Interpretese and Translationese (Section 3). The word color indicates whether it is more frequent in Interpretese (black) or Translationese (gold). “the” is over-represented in Interpretese, a phenomenon also occurs in Translationese vs. the original text (Etemadi and Toutanova, 2014). More conjunction words (e.g., “and”, “so”, “or”, “then”) are used in Interpretese, likely for segmentation, whereas “that” is more frequent in Translationese—a sign of clauses. In addition, the pronoun “I” occurs more often in Translationese while “be” and “is” occur more often in Interpretese, which is consistent with our passivization hypothesis.
### Table 1: Examples of tactics used by interpreters to cope with divergent word orders, limited working memory, and the pressure to produce low-latency translations. We show the source input (S), translated sentences (T), and interpreted sentences (I). The tactics are listed in the rightmost column and marked in the text: more general translations are highlighted in italics; (||) marks where new clauses or sentences are created; and passivized verbs in translation are underlined. Information appearing in translation but omitted in interpretation are in (parentheses). Summarized expressions and their corresponding expression in translation are underlined by wavy lines.

<table>
<thead>
<tr>
<th>Source (S), translation (T) and interpretation (I) text</th>
<th>Tactic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S) この日本語の待特表現の特徴ですか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. (</td>
<td></td>
</tr>
<tr>
<td>(S) で三番目の特徴としてはですね出来る限り自然な日本語の話葉ととその出をと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.</td>
<td>generalize summarize (omit)</td>
</tr>
<tr>
<td>(S) つまり例えばこの表現は認識できますが四回は認識できない。 (T) In sum, we've conducted an analysis on the Japanese language, using a grammar different from school grammar, called derivational grammar. (As a result,) we were able to produce a simpler analysis (than the conventional method). (I) So, we are doing Japanese analysis based on derivational grammar, (</td>
<td></td>
</tr>
<tr>
<td>(S) つまり例えばこの表現は認識できますが四回は認識できない。 (T) They might recognize expression one but not expressions two to four. (I) The phrase number one only is accepted (</td>
<td></td>
</tr>
<tr>
<td>(S) 以上のお話ををまとめますと自然な発話というものをなうことができる音声対話の方法ということができました。 (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 (</td>
<td></td>
</tr>
</tbody>
</table>

3 Classification of Translationese and Interpeterese

We investigate the difference between Translationese and Interpretese by creating a text classifier to distinguish between them and then examining the most useful features. We train our classifier on a bilingual Japanese-English corpus of spoken monologues and their simultaneous interpretations (Matsubara et al., 2002). To obtain a three-way parallel corpus of aligned translation, interpretation, and their shared source text, we first align the interpreted sentences to source sentences by dynamic programming following Ma (2006). This step results in 1684 pairs of text chunks, with 33 tokens per chunk on average. We then collect human translations from Gengo for each source text chunk (one translator per monologue). The original corpus has four interpreters per monologue. We use all available interpretation by copying the translation of a text chunk for its additional interpretation.

3.1 Discriminative Features

We use logistic regression as our classifier. Its job is to tell, given a chunk of English text, which translation produced it. We add \( \ell_1 \) regularization to select the non-zero features that best distinguish Interpretese from Translationese. We experiment with three dif-
different sets of features: (1) **POS**: $n$-gram features of POS tags (up to trigram); (2) **LEX**: word unigrams; (3) **LING**: features reflecting linguistic hypotheses (Section 2), most of which are counts of indicator functions normalized by length of the chunk (Appendix A).

The top linguistic features listed in Table 3 are consistent with our hypotheses. The most prominent ones—which are also revealed by POS and LEX—are the segmentation features, including counts of conjunction words (CC), content words (nouns, verbs, adjectives, and adverbs) that appear more than once (repeated), demonstratives (demo) such as *this, that, these, those*, segmented sentences (sent), and proper nouns (NNP). More conjunction words and more sentences in a text chunk are signs of segmentation. Repeated words and the frequent use of demonstratives come from transforming clauses to independent sentences. Next are the passivization features, indicating more passivized verbs (passive) and fewer pronouns (pronoun) in Interpretese. The lack of pronouns may be results of either subject omission during passivization or general omission. The last group are the vocabulary features, showing fewer numbers of stem types, token types, and content words in Interpretese, evidence of word generalization. In addition, a smaller number of content words suggests that interpreters may use more function words to manipulate the sentence structure.

### 3.2 Classification Results

Recall that our goal is to understand Interpretese, not to classify Interpretese and Translationese; however, the ten-fold cross validation accuracy of LING, POS, LEX are 0.66, 0.85, and 0.94. LEX and POS yield high accuracy as some features are overfitting, e.g., in this dataset, most interpreters used “parsing” for “構文解析” while the translator used “syntactic analysis”. Therefore, they do not reveal much about the characteristics of Interpretese except for frequent use of “and” and CC, which indicates segmentation. Similarly, Volansky et al. (2013) and Etemadi and Toutanova (2014) also find lexical features very effective but not generalizable for detecting Translationese and exclude them from analysis. One reason for the relatively low accuracy of LING may be inconsistent use of strategies among humans (Section 4).

<table>
<thead>
<tr>
<th><strong>LING</strong></th>
<th><strong>POS</strong></th>
<th><strong>LEX</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>+</td>
<td>(S) CC +</td>
</tr>
<tr>
<td>repeated</td>
<td>+</td>
<td>. CC +</td>
</tr>
<tr>
<td>demo</td>
<td>+</td>
<td>(S) CC IN +</td>
</tr>
<tr>
<td>sent</td>
<td>+</td>
<td>NN CC PR +</td>
</tr>
<tr>
<td>passive</td>
<td>+</td>
<td>(S) CC DT +</td>
</tr>
<tr>
<td>pronoun</td>
<td>–</td>
<td>CC RB DT +</td>
</tr>
<tr>
<td>NNP</td>
<td>+</td>
<td>, RB DT +</td>
</tr>
<tr>
<td>stem type</td>
<td>–</td>
<td>. CC DT +</td>
</tr>
<tr>
<td>tok type</td>
<td>–</td>
<td>NN FW NN +</td>
</tr>
<tr>
<td>content</td>
<td>–</td>
<td>NN CC RB –</td>
</tr>
</tbody>
</table>

Table 3: Top 10 highest-weighted features in each model. The sign shows whether it is indicative of Interpretese (+) or Translationese (–).

### 4 Strategy Analysis

To better understand under what situations these tactics are used, we apply two-sample t-tests to compare the following quantities between Interpretese and Translationese: (1) number of inversions (non-monotonic translations) on all source tokens (inv-all), verbs (inv-verb) and nouns (inv-noun); (2) number of segmented sentences; (3) number of natural passivization (pass-st), meaning copying a passive construction in the source sentence into the target sentence, and intentional passivization (pass-t), meaning introducing passivization into the target sentence when the source sentence is in active voice; (4) number of omitted words on the source side and inserted words on the target side; (5) average word frequency given by Microsoft Web $n$-gram—higher means more common. For all pairs of samples, the null hypothesis $H_0$ is that the means on Interpretese and Translationese are equal; the alternative hypotheses and results are in Table 2.

As expected, segmentation and intentional passivization happen more often during interpretation. Interpretese has fewer inversions, especially for verbs; reducing word order difference is important for delay minimization. Since there are two to four different interpretations for each lecture, we further analyze how consistent humans are on these decisions. All interpreters agree on segmentation 73.7% of the time, while the agreement on passivization is...
Table 2: Two-sample t-tests for Interpretese and Translationese. The test statistics are bolded when we reject $H_0$ at the 0.05 significance level (two-tailed).

<table>
<thead>
<tr>
<th>Sample</th>
<th>inv-all</th>
<th>inv-verb</th>
<th>inv-noun</th>
<th>segment</th>
<th>pass-t</th>
<th>pass-st</th>
<th>omit</th>
<th>insert</th>
<th>word freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>-1.55</td>
<td>-3.81</td>
<td>-2.13</td>
<td>4.21</td>
<td>5.67</td>
<td>1.41</td>
<td>16.16</td>
<td>10.66</td>
<td>7.88</td>
</tr>
<tr>
<td>t-stat</td>
<td>.12</td>
<td>&lt;.001</td>
<td>.03</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5 Conclusion

While we describe how Translationese and Interpretese are different and characterize how they differ, the contribution of our work is not just examining an interesting, important dialect. Our work provides opportunities to improve conventional simultaneous MT systems by exploiting and modeling human tactics. He et al. (2015) use hand-crafted rules to decrease latency; our data-driven approach could yield additional strategies for improving MT systems. Another strategy—given the scarcity and artifacts of interpretation corpus—is to select references that present delay-minimizing features of Interpretese from translation corpus (Axelrod et al., 2011). Another future direction is to investigate cognitive inference (Chernov, 2004), which is useful for semantic/syntactic prediction during interpretation (Grissom II et al., 2014; Oda et al., 2015).

A Feature Extraction

We use the Berkeley aligner (Liang et al., 2006) for word alignment, the Stanford POS tagger (Toutanova et al., 2003) to tag English sentences, and Kuromorii[10] to tokenize, lemmatize and tag Japanese sentences. Below we describe the features in detail.

Inversion: Let $\{A_i\}$ be the set of indexes of target words to which each source word $w_i$ is aligned. We count $A_i$ and $A_j$ ($i < j$) as an inverted pair if $\max(A_i) > \min(A_j)$. This means that we have to wait until the $j$th word to translate the $i$th word.

Segmentation: We use the punkt sentence segmenter (Kiss and Strunk, 2006) from NLTK to detect sentences in a text chunk.

Passivization: We compute the number of passive verbs normalized by the total number of verbs. We detect passive voice in English by matching the following regular expression: a be verb (be, are, is, was, were etc.) followed by zero to four non-verb words and one verb in its past participle form. We detect passive voice in Japanese by checking that the dictionary form of a verb has the suffix “れる”.

Vocabulary To measure variety, we use $V_t/N$ and $V_s/N$, where $V_t$ and $V_s$ are counts of distinct tokens and stems, and $N$ is the total number of tokens. To measure complexity, we use word length, number of syllables per word, approximated by vowel sequences; and unigram and bigram frequency from Microsoft Web N-gram.

Summarization We use the sentence compression ratio, sentence length, number of omitted source words, approximated by counts of unaligned words, and number of content words.

Acknowledgments

We thank CIAIR (Nagoya University, Japan) for providing the interpretation data which formed the foundation of this research. We also thank Alvin Grissom II, Naho Orita and the reviewers for their insightful comments. This work was supported by NSF grant IIS-1320538. Boyd-Graber is also partially supported by NSF grants CCF-1409287 and NCSE-1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

only 57.1%—passivization is an acquired skill; not all interpreters use it when it can speed interpretation.

The tests also confirm our hypotheses on generalization and omission. However, these tactics are not inherent to the task of simultaneous interpretation. Instead, they are a byproduct of humans’ limited working memory. Computers can load much larger resources into memory and weigh quality of different translations in an instant, thus potentially rendering the speaker’s message more accurately. Therefore, directly learning from corpus of human interpretation may lead to suboptimal results (Shimizu et al., 2014).
References


