Corpus Analysis of Simultaneous Interpretation Data for Improving Real Time Speech Translation

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932, USA
vkumar,jchen,adc,srini@research.att.com

Abstract

Real-time speech-to-speech (S2S) translation of lectures and speeches require simultaneous translation with low latency to continually engage the listeners. However, simultaneous speech-to-speech translation systems have been predominantly repurposing translation models that are typically trained for consecutive translation without a motivated attempt to model incrementality. Furthermore, the notion of translation is simplified to translation plus simultaneity. In contrast, human interpreters are able to perform simultaneous interpretation by generating target speech incrementally with very low ear-voice span by using a variety of strategies such as compression (paraphrasing), incremental comprehension, and anticipation through discourse inference and expectation of discourse redundancies. Exploiting and modeling such phenomena can potentially improve automatic real-time translation of speech. As a first step, in this work we identify and present a systematic analysis of phenomena used by human interpreters to perform simultaneous interpretation and elucidate how it can be exploited in a conventional simultaneous translation framework. We perform our study on a corpus of simultaneous interpretation of Parliamentary speeches in English and Spanish. Specifically, we present an empirical analysis of factors such as time constraint, redundancy and inference as evidenced in the simultaneous interpretation corpus.

Index Terms: simultaneous speech-to-speech translation, simultaneous interpretation, paraphrasing, time lag

1. Introduction

Voice translation based mobile applications are beginning to be used routinely by a large population of real-world users. Speech translation in such systems is typically either a single turn translation or in some rare cases a turn-segmented, multi-utterance dialog interaction. In contrast, real-time translation of spoken language input in monologues such as lectures, meetings and speeches has received far less attention in the translation research community. Unlike in dialog translation, human translators cannot afford to wait for the completion of an utterance to translate. Thus, in simultaneous interpretation (SI), as opposed to consecutive translation, an interpreter plays the role of receiver and sender concurrently with the main objective of producing a fluent, faithful rendition of the source speech in the target language with minimal delay. Human interpreters, in order to minimize the delay, resort to a diverse set of strategies which include: removing redundancies, anticipation through discourse inference, compression (paraphrasing) and incremental comprehension.

Computational models for simultaneous translation that were limited to speech-to-text translation have been addressed in some projects (e.g., TC-STAR) in the past [1, 2, 3]. More recently, the IWSLT evaluation campaign has included a track for the offline speech translation of TED talks [4]. However, most of the models were repurposing consecutive translation models without a motivated attempt to model incrementality inherent in simultaneous interpretation. Besides, these systems seldom addressed incremental speech production of the target language. In this work, we perform a systemic study of the processes involved in simultaneous interpretation and contrast it with a real-time speech-to-speech (S2S) translation system. Our objective is to understand the various inference, compression and anticipation techniques used by human interpreters in SI and develop models for exploiting such strategies in conventional S2S translation systems.

First, we present an overview of some key strategies used in simultaneous interpretation and motivate their potential use in S2S translation. Second, we present an empirical analysis on simultaneous interpretation data in English-Spanish and Spanish-English. We elucidate several strategies with a systematic analysis of the EPIC [5] corpus with respect to strategies such as simultaneity, compression, hesitations and inference. We perform the investigation by aligning the speech on both sides with corresponding text using speech recognition and subsequently inducing a word-level alignment across the languages. To our knowledge, this is the first large scale analysis of simultaneous speech interpretation data using automatic tools beyond the work performed in [5, 6].

2. Theory of Simultaneous Interpretation

Simultaneous interpretation is an extremely complex process where the interpreters use a wide range of cognitive, linguistic and prosodic phenomena to facilitate an inferential mode of communication. In contrast, state-of-the-art speech translation still focuses only on rendering the exact source language content in the target language through appropriate lexical choice and reordering. As a consequence, conventional translation methodologies for translating real-time conversation such as talks and lectures introduce undesirable noise and latency, thereby increasing the cognitive load on a target language participant. In this section, we describe some key strategies used by interpreters in SI that can potentially be used to improve real-time automatic speech translation. A more comprehensive account of SI strategies can be found in [7].

2.1. Time Constraint

The central tenet of SI is that the interpreter is simultaneously receiving and rendering the message from the source language to the target language. There are two factors that control the lag between the source and target speech. The simultaneous

interpreter needs to keep the delay between a source language chunk and its corresponding target language chunk minimal in order to continually engage the listeners. Studies have shown that this measure, known as *ear-voice span* [8], is about the average latency (4-5 seconds) that can be tolerated by a listener in cross-lingual communication. Another factor that determines the limit on the lag is the short-term memory capacity of the interpreter. Empirical analysis of SI speeches indicate the registered spread for this lag to range from 200ms to 10 or 15 seconds [9] with longer lags being associated with losses and errors. Regardless of the varying lag, interpreters have been observed to maintain their own rate of speaking [7] which augurs well for using text-to-speech (TTS) systems that typically adhere to a standard speech rate and have difficulty in dynamically altering the speech rate.

It is clear that an automatic speech interpretation system must be able to generate target hypotheses with low latency to reduce the ear-voice span. However, machines can potentially alleviate the short-term memory constraint in a significant manner. Furthermore, if one were to keep the rate of target speech synthesis using a TTS system more or less constant, the burden of keeping up pace with the source language speaker in a S2S system will be transferred to the source text segmentation and machine translation components. One can use a variety of word order and syntax based strategies to achieve such a constraint in machine translation. For example, during translation from English-Spanish, Spanish offers the syntactic flexibility of allowing a direct object to open a neutral utterance; the unmarked structure i.e. Subject – Verb – Direct Object – Indirect Object, may be changed by placing any element in the initial position, thus, promoting incrementality in translation.

2.2. Compression

Simultaneous interpreters use many strategies to compress the source speaker's message in order to satisfy the time constraint property discussed previously. The compression can be in terms of linguistic, syntactic or semantic content. Lexical compression refers to the use of fewer words to express the same idea and has been extensively studied as a paraphrasing problem [10, 11]. Syntactic compression results from choice of a shorter construction than the original. Examples include breaking a sentence with involved clause structure into several simpler sentences, using a prepositional phrase instead of a clause, substituting a single word for a word combination, etc. Semantic compression strategies consist of several means to mark co-reference such as use of the same word throughout the discourse, or of its synonyms and paraphrases; shifting a word component to a different part-of-speech; use of pro-forms for semantically self-contained words (nouns, adjectives, verbs and adverbs), etc.

Conventional machine translation systems typically do not address any of the above compression techniques. It is conceivable that some of these strategies are automatically learned during the phrase table construction (in phrase-based translation). However, almost all systems are completely faithful to the entire source message and potentially generate verbose translations.

2.3. Linguistic and Cognitive Inference

Inference techniques can be considered complementary to compression techniques. Simultaneous interpreters often use linguistic inferences in establishing co-reference. For example, in the extract "*My delegation...we...we...members of the council...we...we...*", the interpreter infers co-reference between the pronoun *we* and noun phrase *my delegation*, until the occurrence of *members of the council*. Interpreters also iterate semantic components to form a coherent discourse. Cognitive inference occurs when the interpreter uses his/her background knowledge on the subject to add emphasis, contrast. Expansion of pronouns to the original referents is also a common form of inference in SI. For example, the interpreter may choose to replace *we* with the original noun phrase in case it is used often.

While machines are still far from being able to perform cognitive inference soundly, it is possible to perform linguistic inference in speech translation. For instance, chunking the source language text appropriately can help encapsulate concepts within a semantically independent unit. Appropriate expansion of pronouns can also be performed in traditional translation to facilitate improved understanding.

2.4. Discourse context

Interpreters often use their memory of the conversation to gather appropriate context, disambiguate and . It has been shown that the maximum number of units that can be processed simultaneously in human memory is seven-plus-or-minus-two [12]. Within this span, the signal tends to weaken and disappears within 30 seconds. Such a constraint does not exist in written text translation.

Machines are much better at preserving this long term discourse context. However, current translation systems use context mainly in terms of a n-gram window used in the language model. To avoid sparsity, the phrases themselves are not context dependent. Thus, exploiting the translation history through appropriate models can offer significant improvement in the translation quality of long talks.

3. Simultaneous Interpretation Data

We use the European Parliamentary interpretation corpus (EPIC) for performing our analysis. The EPIC corpus [5] is a parallel corpus of European Parliamentary speeches and their corresponding simultaneous interpretations. The source speeches are either in English (81), Spanish (21) or Italian (17) and each source speech is simultaneously interpreted in two other languages. We extracted the audio from the video clips of each source language speaker while the audio for the interpreted target speeches was already provided. The corpus also contains the transcripts of all the speeches. The speeches are also classified according to the style of delivery as *read*, *impromptu* and *mixed*. The English-Spanish portion contains 43, 24 and 14 speeches in the *read*, *impromptu* and *mixed* categories while the Spanish-English portion is comprised of 9, 5 and 7 speeches in these categories.

As a first step in our analysis we forced aligned the English and Spanish speeches independently using generic acoustic models. The English acoustic model was trained on about 600 hours of TED talks while the Spanish acoustic model was trained on close to 1000 hours of speech collected through smartphones. Both the acoustic models were trained using minimum phone error (MPE) criterion using the AT&T WATSON^{SM} speech recognizer [13]. The resulting word alignments contained the start and end duration for each word as well as silences (with duration). Subsequently, we aligned the transcripts in the parallel speeches at the sentence level using dynamic programming with an English-Spanish dictionary.

3.1. Inducing word alignment

In simultaneous interpretation, interpreters use a variety of techniques to anticipate, compress and render the message in the

target language. As a result, inducing word correspondence using automatic word alignment is quite difficult. Unlike parallel text used in building word and phrase-based machine translation models, SI texts maybe non-parallel and even non-comparable. Hence, we used a custom algorithm for aligning the words across the two languages. The matching was facilitated by a dictionary obtained through automatic alignment [14] of a large English-Spanish parallel corpus comprising of about 8 million sentence pairs. The resulting dictionary was filtered such that only top 10 target translations (sorted by posterior probability) of each source word was preserved in the final dictionary. As a first step, we performed automatic sentence alignment of the source and target speeches using dynamic programming alignment scheme described in [15].

The word alignment procedure takes as input an English sentence and its corresponding aligned Spanish sentence. It links each English word with its closest matching Spanish word if there is one according to heuristics. Specifically, the input consists of a sequence of English words $\langle e_1, e_2, \ldots, e_m \rangle$ and a corresponding sequence of Spanish words $\langle f_1, f_2, \ldots, f_n \rangle$. In addition, there is a function TIME maps an English or Spanish word to its start time and another function STOP maps an English or Spanish word to *true* if it is a stopword and *false* otherwise. Lastly, it is assumed that translation probabilities $P(f|e)$ are available.

The procedure takes three parameters. δl and δr define the left and right part of the time window in which the Spanish word f corresponding to the English word e is taken to appear. t is a probability threshold that forbids a Spanish word f from linking to an English word e when $P(f|e) < t$. For these experiments, $\delta l = 1$ second, $\delta r = 6$ seconds, and $t = 0.008$. The procedure tries to link each English word $e \in \{e_1, \ldots, e_m\}$ to a Spanish word as follows. First, a candidate set F_e of Spanish words is constructed such that $f \in \{f_1, \ldots, f_n\}$ is placed in F_e if and only if the following criteria hold:

- TIME$(e) - \delta l \leq$ TIME$(f) \leq$ TIME$(e) + \delta r$
- STOP$(e) \wedge$ STOP(f) or $\neg$STOP$(e) \wedge \neg$STOP(f)
- $P(f|e) \geq t$

Second, f^* is output where $f^* = \arg\max_{f \in F_e} P(f|e)$.

3.2. Syntactic tagging

Freeling [16] was used to parse the parallel English and Spanish sides of the corpus. It assigns Penn Treebank style parts of speech (POS) to the English text, which is then chunked into one of 35 different chunk types. In contrast, it assigns EAGLES [17] style parts of speech to the Spanish text. Because it is designed to capture many morphosyntactic distinctions, these are quite fine grained; there are more than 200 types of Spanish POS tags whereas there are less than 50 types of Penn Treebank POS tags. Subsequently, Freeling chunks the Spanish text into one of 174 chunk types. For both English and Spanish, Freeling assigns a head constituent to each chunk that it creates. The constituent can be a word or a nested chunk inside the target chunk.

In order to link English chunks to Spanish chunks, a conversion table was manually created that mapped both the English chunk types and the Spanish chunk types into the same kind of "merged" chunk types. There were 19 of these merged chunk types encompassing fairly general categories such as np (noun phrase) or vg (verb group). Let MERGE be a function that maps a node representing a chunk, either English or Spanish, into its corresponding merged chunk type.

The chunk alignment procedure assumes that the word alignment procedure has already produced links from English words to Spanish words. Given that, the chunk alignment procedure proceeds as follows. For each English word e that is linked to a Spanish word f, the maximal projection e_p of e in the English chunk structure is obtained. Now, nodes along the head path of f in the Spanish chunk structure are searched for the highest node f_h such that MERGE(f_h) =MERGE(e_p). If there exists such a node f_h, then e_p is linked to f_h.

4. Empirical Analysis

In this section, we perform a variety of empirical analyses to evince the theories regarding simultaneous interpretation presented in Section 2.

4.1. Time Constraint

We measure the simultaneity in the interpretation corpus by computing overlapping intervals of SL speech - TL pause (S/P); SL speech - TL speech (S/S) and SL pause - TL speech (P/S). The case of SL pause - TL pause (P/P) is not of interest as it implies no SI is in progress. We measure the overlap for each of the three genres, read, spontaneous and mixed source speech. The overlapping intervals were computed from the word alignments generated by the speech recognizer. The results are presented in Table 1. While the percentage of time does not vary for S/P across all the genres in English-Spanish and Spanish-English, the percentage of S/P and P/S vary significantly. Specifically, during the interpretation from Spanish to English, the interpreters seem to spend more time talking when the source speaker is silent. This can imply two things: the interpreters need to catch up more with the Spanish speakers rate or the syntactical rendering of content from Spanish into English is more difficult and hence the interpreters need to be more active. The analysis clearly shows the need to look beyond consecutive translation strategies that typically have no S/S overlap.

4.2. Hesitations

We also analyzed the use of hesitations in SI. We simply counted the occurrences of *ehm, mhm, mmm* in the transcriptions. The idea behind using hesitations on the target side by the interpreter is to gain more time to see a verb or think about the appropriate translation of a content word. From Table 1 one can see that on an average the interpreters use fewer hesitations when interpreting from English-to-Spanish in contrast with Spanish-to-English. This can be attributed to the possibility of free word order in Spanish that may facilitate the interpreter to use hesitations less frequently in comparison with English that has more rigid syntax.

4.3. Compression

As mentioned in Section 2.2, compression or paraphrasing can be witnessed in SI at the sentence, phrase, word or even syllable level. We analyze the notion of compression at the sentence level by looking at the number of sentences that were aligned to NULL during the dynamic programming sentence alignment. The results are shown in Table 1. It is clear that the simultaneous interpreters almost always omit sentences during SI. While this is more pronounced for read speech as the interpreter has to process information at a rapid rate from the source side, it is less pronounced for impromptu speech. The result also indicate that interpreters sometimes, though less often, introduce new sentences perhaps as glue to aid in SI.

Phenomenon		en→es			es→en		
		read	impromptu	mixed	read	impromptu	mixed
Time constraint (%)	S/S	11.1	13.0	16.3	12.2	11.0	13.4
	S/P	42.5	39.7	42.0	37.4	25.0	33.2
	P/S	46.4	47.3	41.7	50.4	64.0	53.4
	total	100.0	100.0	100.0	100.0	100.0	100.0
Hesitations	# source hesitations/speech	13	12	13	2	9	1
	# target hesitations/speech	6	4	5	11	11	7
Redundancy	#source sent aligned to NULL/speech	58	17	28	15	5	5
	# target sent aligned to NULL/speech	9	2	4	1	2	1
	% linked across words	64.5	60.1	66.6	66.5	62.9	70.6
	% linked across chunks	70.2	67.8	72.7	89.2	90.5	92.1
Lag	Lag (msec) across words	2078.3	1627.8	2097.3	2554.0	2612.4	1504.2
	Lag (msec) across chunks	2086.8	1680.2	2114.2	2636.6	2685.3	1525.3

Table 1: Empirical analysis of time constraint, redundancy and hesitations in simultaneous interpretation as evidenced in the EPIC corpus.

4.3.1. Sub-Sentence Level Redundancy

Based on the results of word alignment and chunk alignment algorithms, we measure the percentage of cases where a source word (chunk) was successfully aligned to a target word (chunk). The results are shown in Table 1. They bolster the results of earlier sections that indicate that simultaneous interpretation from English to Spanish is "easier" than the same task in the opposite direction, as there are more linked words and chunks for the former case.

4.4. Lag

We measure the time interval between when a source word is spoken and its target word is spoken, according to the results of word alignment (Section 3.1). We do the same for aligned source and target chunks according to the results of chunk alignment (Section 3.2). The results are shown in Table 1. They show that simultaneous interpretation from English to Spanish is "easier" than the same task in the other direction, since the lag is shorter in the former case. This may be explained by the fact that English is a restricted word order language in comparison to Spanish, which has a freer word order. Consequently, a simultaneous interpretation system would need to have special strategies to deal with the case of translating to a language with a more restricted word order.

4.5. Discourse Analysis

Simultaneous interpreters often introduce deictic expressions such as pronouns even when they are not used by the source speaker [7]. This is because it takes a shorter amount of time to pronounce these expressions instead of their full form equivalents. Here, we explore several ways deictic expressions can be employed and measure their prevalence in the EPIC corpus.

One way that a simultaneous interpreter can employ deictic expressions is to use them to refer to entities. In this case, the interpreter would use use *anaphora* such as pronoun (e.g. "him") or a common noun (e.g. "the leader") to refer to an *antecedent* which is a proper noun (e.g. "President Barack Obama"). One way to measure the amount of this phenomena occurring in the EPIC corpus is to compute the ratio of pronouns and common nouns to proper nouns in either the source speech or the target speech. Subsequently, we might expect that the ratio should be higher in the target speech, if the simultaneous interpreter has a tendency to translate proper nouns as pronouns or common nouns. Of course, this is only a rough approximation, as a more exact comparison would use co-reference chains linking each anaphor to its antecedent, which we do not have for this corpus. In any case, the results, shown in the upper part of Table 2 do provide some evidence that the simultaneous interpreter is employing deictic expressions more often than the source speaker to refer to entities.

	en → es	es → en
$\frac{\text{pron,common noun}}{\text{proper noun}}$ **(en)**	2.687	4.092
$\frac{\text{pron,common noun}}{\text{proper noun}}$ **(es)**	2.957	3.776
$\frac{\text{verb}\rightarrow\text{noun,pron}}{\text{verb}\rightarrow *}$	0.1020	0.1514
$\frac{\text{noun,pron}\rightarrow\text{verb}}{\text{noun,pron}\rightarrow *}$	0.0251	0.0543

Table 2: Ratios of various link types in EPIC corpus show that simultaneous interpreters often use deictic expressions to compress their speech.

Another way that a simultaneous interpreter can employ deictic expressions is to use them to refer to events. According to [7], in this case a typical approach that an interpreter might use would be to nominalize a mention of an event. For example, if the source speaker says "Israeli troops attacked the bridge," the interpreter might say "the attack on the bridge by Israeli troops." The benefit of this reformulation is that later on in the same speech, the interpreter can refer to the same event succinctly using a pronoun. We determine the prevalence of this strategy in the EPIC corpus by computing the fraction of source verbs that are aligned to target nouns or pronouns. The results are shown in the lower part of Table 2. They show that there is a non-insignificant number of cases where this strategy is being used.

5. Conclusion

We presented an overview of strategies used in simultaneous interpretation and their applicability in simultaneous speech-to-speech translation systems. We presented an empirical analysis on the simultaneous interpretation data as part of the EPIC corpus with respect to SI strategies such as simultaneity, compression and discourse inference. Our work is the first large scale analysis of SI data using automatic tools (e.g., word alignments obtained through speech recognition). We performed the analysis on English-Spanish and Spanish-English portions of the EPIC corpus and demonstrated several key aspects that are specific to simultaneous interpretation. We are working on a similar analysis on the EPIC corpus within a simultaneous speech-to-speech translation system as part of our current work.

6. References

[1] D. Vilar, E. Matusov, S. Hasan, R. Zens, and H. Ney, "Statistical machine translation of European parliamentary speeches," in *Proceedings of MT Summit*, 2005.

[2] C. Fügen, A. Waibel, and M. Kolss, "Simultaneous translation of lectures and speeches," *Machine Translation*, vol. 21, pp. 209–252, 2007.

[3] O. Hamon, C. Fügen, D. Mostefa, V. Arranz, M. Kolss, A. Waibel, and K. Choukri, "End-to-end evaluation in simultaneous translation," in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, March 2009.

[4] M. Federico, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2011 evaluation campaign," in *Proceedings of IWSLT*, 2011.

[5] C. Bendazzoli and A. Sandrelli, "An approach to corpus-based interpreting studies," in *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation*, Saarbrücken, 2005.

[6] C. Bendazzoli, "The European Parliament as a Source of Material for Research into Simultaneous Interpreting: Advantages and Limitations," in *Translationswissenschaft - Stand und Perspektiven. Innsbrucker Ringvorlesungenzur Translationswissenschaft VI (Forum Translaitonswissenschaft, Band 12)*, N. L. Zybatow, Ed. Frankfurt: Peter Lang, 2010, pp. 51–68.

[7] G. V. Chernov, *Inference and anticipation in simultaneous interpreting*. John Benjamins, 2004.

[8] M. Lederer, "Simultaneous interpretation: units of meaning and other features," in *Language interpretation and communication*, D. Gerver and H. W. Sinaiko, Eds. Plenum Press, New York, 1978, pp. 323–332.

[9] B. Moser, "Simultaneous interpretation: A hypothetical model and its practical application," in *Language interpretation and communication*, D. Gerver and H. W. Sinaiko, Eds. Plenum Press, New York, 1978, pp. 353–368.

[10] K. R. McKeown, "Paraphrasing using given and new information in a question-answer system," in *Proceedings of ACL*, La Jolla, California, 1979.

[11] D. Ravichandran and E. Hovy, "Learning surface text patterns for a question answering system," in *Proceedings of ACL*, Philadelphia, Pennsylvania, 2002.

[12] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, 1956.

[13] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tür, A. Ljolje, and S. Parthasarathy, "The AT&T Watson Speech Recognizer," Tech. Rep., September 2004.

[14] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[15] V. K. Rangarajan Sridhar, L. Barbosa, and S. Bangalore, "A scalable approach to building a parallel corpus from the Web," in *Proceedings of Interspeech*, 2011.

[16] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA, May 2012.

[17] G. Leech, "Eagles – recommendations for the syntactic annotation of corpora," 1996.