Prosody in Cascade and Direct Speech-to-Text Translation: a case study on Korean *Wh*-Phrases

Anonymous ACL submission

Abstract

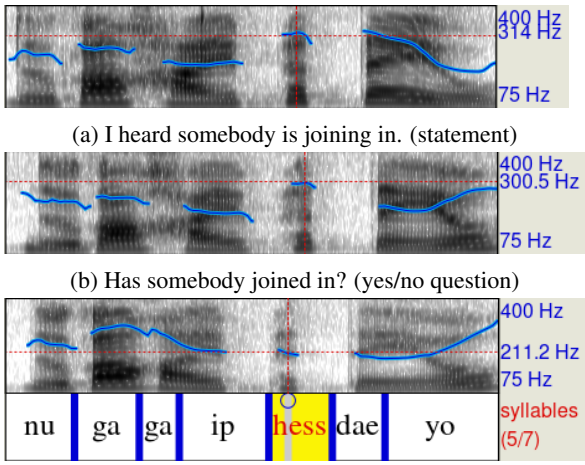
Speech-to-Text Translation (S2TT) has typically been addressed with *cascade* systems, where speech recognition systems generate a transcription that is subsequently passed to a translation model. While there has been a growing interest in developing *direct* speech translation systems to avoid propagating errors and losing non-verbal content, prior work in direct S2TT has struggled to conclusively establish the advantages of integrating the acoustic signal directly into the translation process. This work proposes using contrastive evaluation to quantitatively measure the ability of direct S2TT systems to disambiguate utterances where prosody plays a crucial role. Specifically, we evaluated Korean-English translation systems on a test set containing *wh*-phrases, for which prosodic features are necessary to produce translations with the correct intent, whether it's a statement, a yes/no question, a *wh*-question, and more. Our results clearly demonstrate the value of direct translation systems over cascade translation models, with a notable 12.9% improvement in overall accuracy in ambiguous cases, along with up to a 15.6% increase in F1 scores for one of the major intent categories. To the best of our knowledge, this work stands as the first to provide quantitative evidence that direct S2TT models can effectively leverage prosody.

1 Introduction

Speech-to-Text Translation (S2TT) is the task of automatically generating a text translation in a target language given an input speech signal. Traditionally, S2TT has been achieved by concatenating two systems: one in charge of generating an intermediate transcription of the source speech signal and one of translating the intermediate text into a target language. Although such a pipeline, known as *"cascade"* architecture, remains the dominant technology in Speech-to-Text Translation, it has some shortcomings. Firstly, it is affected by error propagation for which errors in the transcription phase are carried over and amplified in the translation phase. Secondly, some information is lost as non-verbal content (e.g. prosody) is discarded from the text. As a potential solution to these issues, "*direct*" systems that can perform translation directly from speech signals without needing intermediate transcriptions have emerged in the last few years. Bentivogli et al. (2021) claim direct systems have an advantage over the cascade architecture by modelling prosody during the translation process. However, there is no conclusive evidence to support this claim as both types of systems have similar overall performances, and current datasets do not regularly include instances where speech signals are necessary to disambiguate the meaning of an utterance, making quantitative analysis on the effect of prosody in S2TT particularly challenging (Sperber and Paulik, 2020; Bentivogli et al., 2021).

The aim of this paper is to investigate the potential of direct S2TT to effectively leverage non-lexical information, particularly prosody, and quantify their impact. Since identifying ambiguous utterances that rely on prosody for disambiguation is nontrivial, especially in English where sentence structure typically carries more weight than prosodic cues, we focus on Korean *wh*-phrases where the presence of a prosodic boundary distinguishes *wh*-interrogatives from *wh*-indefinites (e.g., 어디갔어요 (eodi gasseoyo) → where did you go?/did you go somewhere?), as well as other interpretations.

In this paper, we (i) introduce a new contrastive evaluation framework for Korean-English S2TT systems, designed for ranking translations of ambiguous utterances containing *wh*-particles; (ii) quantitatively demonstrate the capacity of direct S2TT systems to effectively model prosodic cues from the input, yielding an overall improvement over cascade models of 12.9% in accuracy for ambiguous utterances, and up to a 15.6% increase in



(a) I heard somebody is joining in. (statement)

(b) Has somebody joined in? (yes/no question)

(c) Who joined in? (*wh*-question)

Figure 1: An example in the ProSem dataset: Based on the intent, the transcription "누가가입했대요" (nuga gaiphessdaeyo) can be mapped to a different pair of recording and translation, see (a), (b), and (c). The "blue" lines on the spectrogram, i.e., the recording, are the pitch (F0) contours.

F1 scores within one of the major intent types; (iii) highlight the limitations of punctuations in disambiguating certain intent types despite being strong signals in distinguishing questions from statements.

2 Korean Prosody and Wh-Particles

Prosody refers to the acoustic features that are exhibited across multiple phonetic segments, also known as suprasegmental features (Lehiste and Lass, 1976). These suprasegmental features can take shape in a multitude of ways. For example, by stressing a single word in a phrase (phrasal stress), by adding pauses or modifying the length of syllables (boundary cues) or by varying the tonal and stress patterns in the utterance (metre) (Gerken and McGregor, 1998). In an intonational language like Korean, the intended meaning of an utterance is often conveyed via intonation and rhythm instead of lexical pitch accents or tones (Jun, 2005; Jeon, 2015). While prosodic structures in Korean utterances are still debated, there are at least two levels of prosody above the word: the Accentual Phrase (AP) and the Intonation Phrase (IP). The AP is the basic unit for prosodic analysis marked by a tonal pattern THLH which consists of variations of the pitch between low (L) and high (H), with T being either L or H depending on the phrase's initial segment, while the IP consists of one or more APs and a boundary tone on the right edge of the phrase.

Korean *wh*-particles are an example of a linguistic phenomenon where the tonal patterns and IP boundary tones are necessary to disambiguate the meaning of the utterance, as otherwise they can be interpreted as both interrogative particles or indefinite pronouns (e.g. 누구" (nugu) → "who" / "somebody"). Figure 1 shows the pitch contours for the recordings of the utterance "누가가입했대요" (nuga gaiphessdaeyo). By varying the boundary tone H+L%, H+LH%, and L+H%, the utterance can be interpreted as a statement, yes/no question or *wh*-question respectively.

3 Contrastive Evaluation

Contrastive evaluation is an automatic accuracy-based evaluation technique that measures the capability of a system to distinguish correct from incorrect outputs. This is achieved by asking a generative model θ to score and rank a set of predefined outputs containing each a correct and a contrastive utterance (e.g., *"the cat sleeps"* vs. *"the cat sleep"* (Linzen et al., 2016)). Following previous work (Sennrich, 2017; Vamvas and Sennrich, 2021), we define the score of an utterance as the sum of the target token log probabilities normalised by the length of the full target sequence Y:

$$\text{score}(Y|X,\theta) = \frac{1}{|Y|}\sum_{i=1}^{|Y|} \log p_\theta(y_i|X, y_{<i})$$

where X is the input signal, $|Y|$ the target sequence length and θ the evaluated model.

In this work, we perform contrastive evaluation of cascade and direct S2TT systems on Korean *wh*-phrases. Since multiple prosodic realisations can occur per utterance (as in Figure 1), in contrast to previous work where only one contrastive utterance per example was available, we consider a model having correctly identified the intended translation only if its score is higher compared to the score of all the possible incorrect translations. In addition to the general accuracy of the model in identifying the correct translation, we report contrastive precision, recall and F1 scores of the systems on the various *wh*-phrases' intent types.

4 Experimental Setting

In our experiment, we adopted the ProSem corpus (Cho et al., 2019) as the contrastive evaluation test set. Originally designed for Spoken Language Understanding, this corpus consists of 3552 utterances recorded by two Korean native speakers of a different gender. All the utterances make use of one of the six Korean *wh*-particles and are further classi-

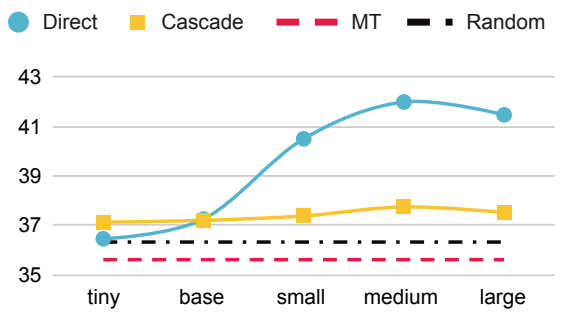
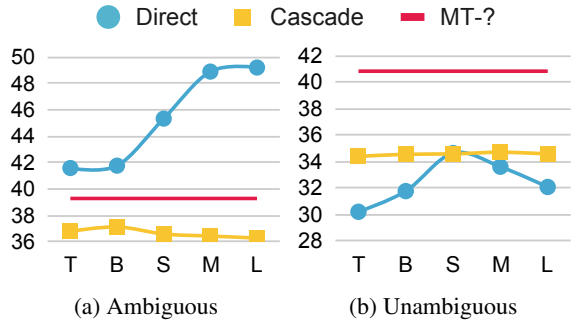


Figure 2: Contrastive evaluation accuracy ↑ scores on ProSem for direct (blue) and cascade (yellow) S2TT systems by varying the size of the Whisper model, along with Random selection (black) and an MT system that has gold transcriptions as input (red).



(a) Ambiguous (b) Unambiguous

Figure 3: Contrastive evaluation accuracy ↑ for direct (blue), cascade (yellow) S2TT systems with different Whisper model sizes, and MT with gold transcriptions augmented with question marks (MT-?, red) on ambiguous and unambiguous contrastive sets.

fied into seven intent categories: statements, yes/no questions, *wh*-questions, rhetorical questions, commands, requests, and rhetorical commands, with the first three categorised as major intent types. In the dataset, there are a total of 1292 distinct transcriptions, each associated with up to 4 utterances of a different intent. Consequently, our test set is composed of 3552 contrastive sets per speaker, each featuring a correct translation linked to a specific recording and at least one incorrect translation.

Each recorded utterance in the dataset is thus paired to a gold translation, as well as a number of incorrect ones that are associated with recordings of the same transcription (but with different prosody). For example, in the recording in Figure 1a the correct translation is *"I heard somebody is joining in."* while the incorrect/contrastive ones are *"Has somebody joined in?"* and *"Who joined in?"* .

For our experiments, we utilise state-of-the-art pretrained models. Specifically, we use Open AI's Whisper models (Radford et al., 2022) for both the S2TT direct systems and the ASR components in the cascade systems, reporting results obtained from all the provided multilingual models. As for the MT component in the cascade systems, we make use of the Korean-English baseline model provided for the Tatoeba challenge (Tiedemann, 2020), trained on approximately 34.5M Opus MT parallel data (Tiedemann and Thottingal, 2020).

5 Results

Contrastive Evaluation Accuracy Figure 2 shows the results of the contrastive evaluation, along with the average accuracy of randomly selecting one of the 2-4 potential translations. As expected, the performance of both cascade and direct systems exhibits an upward trend with increasing model size. Notably, the direct systems outperformed both the MT and cascade systems, with the "medium" direct system exhibiting an improvement of 6.4% and 4.3% in accuracy respectively.

In contrast, the MT model with gold transcription as input failed to surpass random selection in performance due to its inability to distinguish between different translations effectively when presented with the same transcription. On the other hand, despite relying on the aforementioned MT model, the cascade systems managed to achieve scores surpassing random selection, with an improvement of up to 2.1% observed in the Whisper "medium" system. This improvement can be attributed to the inclusion of punctuation marks in the transcriptions, which are absent in the gold transcriptions, that aid in disambiguating questions from statements.

Effect of Punctuation To better understand the disparity in performance between direct and cascade systems, we conducted an analysis to assess the role of punctuation within the MT inputs. To do so, we added question marks to the ProSem gold transcriptions based on the intents of the correct translations. Subsequently, we categorised the contrastive sets into two distinct groups: "Ambiguous" and "Unambiguous", where the latter are the ones where punctuation alone is sufficient to discern the correct intention among the options considered. In total, we identified 1602 unambiguous and 1950 ambiguous sets (more details in Appendix D).

Figure 3a shows that, on ambiguous contrastive sets, all direct systems consistently outperform their cascade counterparts and even surpass the MT

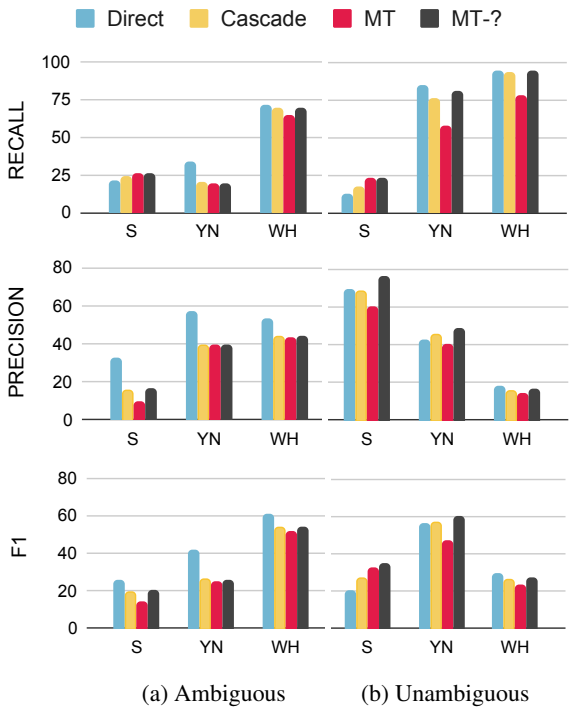


Figure 4: Contrastive evaluation recall, precision and F1 ↑ scores on ambiguous and unambiguous contrastive pairs for each intent major type: statements (S), yes/no questions (YN), *wh*-questions (WH). Direct and cascade systems based on Whisper "medium", and MT systems with and without gold question marks.

system, which has access to gold transcriptions. The gap between the direct and cascade systems is notably wider compared to the overall performance shown in Figure 2, with differences reaching up to 12.9% for the "large" model, supporting the hypothesis that direct models are capable of modelling acoustic signals to handle ambiguous utterances effectively. On the other hand, Figure 3b shows that the augmented gold MT model, which serves as an upper bound for the cascade systems, outperforms the best-performing direct model by 6.2% in accuracy, illustrating that punctuation is an effective convoy for certain prosodic information. The effectiveness of punctuation is reflected in the performance of cascade systems themselves, which, except for the "small" model, outperform the direct systems. It's worth noting that all systems, despite their strengths, did not achieve the anticipated levels of performance on the unambiguous contrastive sets. This can be attributed to the fact that while question marks are commonly used in modern Korean, they are not mandatory, which leaves ambiguity in utterances lacking them.

Intent Disambiguation While Figure 2 and 3a demonstrate the advantages of preserving acoustic signals during the translation process, it's important to note that the overall accuracy achieved by all systems remains relatively low. Figure 4 reveals a significant challenge common to all systems when it comes to disambiguating statements, as they achieve a recall score of less than 25% in this category. In contrast, the highest recall scores are consistently observed in the *wh*-questions intent category. The low recall score for yes/no questions and the subpar precision for *wh*-questions, two intent types that are indistinguishable for MT-based systems, indicate a distinct bias towards the *wh*-question type. This bias can be attributed to the primary use of *wh*-particles in the Korean language for forming *wh*-questions.

Overall, on ambiguous contrastive sets, the direct model outperforms the other two systems in terms of F1 scores across all major intent categories, achieving improvements of up to 15.5% in the case of yes/no questions. However, on unambiguous sets, the direct model's performance is comparable to cascade models in question categories but falls short on statements, where its recall is notably low. This performance gap on statements may be due to the inherent challenge of accurately capturing the nuanced prosody and context associated with statements, which direct models may struggle to discern effectively. Full results and confusion matrices are reported in Appendix C.

6 Conclusion

The objective of this paper was to test whether direct S2TT systems could take advantage of the prosodic information contained in the speech signal. To achieve this, we conducted quantitative analyses focused on Korean *wh*-particles which can represent either *wh*-interrogatives or *wh*-indefinites encompassing a range of intents in accordance with the input acoustic features. Our contrastive evaluation results provide compelling evidence that the direct S2TT systems outperform the cascade systems in overall accuracy and F1 score across all the major intent types on ambiguous utterances. Cascade systems perform better than random primarily thanks to the inclusion of punctuation in the transcriptions. However, it's essential to underscore that while punctuation marks play a valuable role in aiding disambiguation, they are not sufficient to resolve all types of intents, emphasizing the importance of considering prosody in S2TT systems.

7 Limitations

While our study has yielded positive results, it is essential to acknowledge several limitations. Firstly, the contrastive evaluation approach in this study diverges from previous work in that it was not conducted with minimally different utterances. The set of possible translations used here differs significantly in structure and, to some extent, vocabulary. This variation may potentially influence the resulting scores, despite being normalised. Secondly, the findings of this research may not be readily generalisable beyond the specific context of Korean *wh*-particles. To examine different linguistic phenomena in various language pairs, specific contrastive datasets will need to be meticulously crafted. As previously discussed, this process poses a significant challenge. Lastly, despite employing state-of-the-art models, the overall accuracy observed in the contrastive evaluation remains relatively low. This suggests that there is substantial room for improvement within speech translation systems, reflecting the ongoing development needs in this field.

References

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Won Ik Cho, Jeonghwa Cho, Jeemin Kang, and Nam Soo Kim. 2019. Prosody-semantics interface in seoul korean: Corpus for a disambiguation of wh-intervention. In *Proceedings of the 19th international congress of the phonetic sciences (icphs 2019)*, pages 3902–3906.

Won Ik Cho, Seok Min Kim, Hyunchang Cho, and Nam Soo Kim. 2021. kosp2e: Korean Speech to English Translation Corpus. In *Proc. Interspeech 2021*, pages 3705–3709.

LouAnn Gerken and Karla McGregor. 1998. An overview of prosody and its role in normal and disordered child language. *American Journal of Speech-Language Pathology*, 7(2):38–48.

Jung-Woo Ha, Kihyun Nam, Jingu Kang, Sang-Woo Lee, Sohee Yang, Hyunhoon Jung, Eunmi Kim, Hyeji Kim, Soojin Kim, Hyun Ah Kim, et al. 2020. Clovacall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers. *arXiv preprint arXiv:2004.09367*.

Hae-Sung Jeon. 2015. Prosody. *The handbook of Korean linguistics*, pages 41–58.

Sun-Ah Jun. 2005. Prosody in sentence processing: Korean vs. english. *UCLA Working Papers in Phonetics*, 104:26–45.

Ilse Lehiste and Norman J Lass. 1976. Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, 225:239.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Jungyeul Park, Jeen-Pyo Hong, and Jeong-Won Cha. 2016. Korean language resources for everyone. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 49–58, Seoul, South Korea.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Wh-Particle		Interrogative	Indefinite
뭐	mwo	what	something
누구	nugu	who	someone
언제	eonje	when	some time
어디	eodi	where	some place
어떻게	eotteohge	how	somehow
몇	myeot	how many	some

Table 1: Korean *wh*-Particles and English *wh*-interrogatives/indefinite pronouns in the ProSem dataset.

Intent	#
Statement	1085
Yes/no Q	1047
Wh-Q	849
Rhetorical Q	302
Commands	175
Requests	56
Rhetorical C	38

Wh-particle	#
Who	1,895
What	877
Where	199
When	172
How	163
How many	246

Table 2: Number of utterances in Prosem per *wh*-particle and intent type.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

A Dataset Statistics

Table 1 shows Korean *wh*-particles and their English translations. The particle 왜 (wae, why) is not present in the ProSem dataset as it is rarely used as a quantifier. On the other hand, 몇 (myeot, how many) is used instead despite not being technically a *wh*-particle. Table 2 shows the number of utterances per intent type and *wh*-particle in the Prosem dataset, with most of the utterances belonging to one of the three major intent types (statement, yes/no questions, *wh*-questions).

Model	Size	kosp2e	ProSem
Direct	T	1.0	5.3
	B	4.7	10.2
	S	13.0	17.4
	M	19.4	**21.4**
	L	**21.1**	19.6
Cascade	T	10.6 (16.2)	10.9 (27.0)
	B	12.3 (12.1)	12.2 (22.3)
	S	13.9 (9.1)	13.3 (16.3)
	M	14.9 (7.3)	14.1 (**13.9**)
	L	15.2 (**6.6**)	14.3 (**13.9**)
MT		14.2	7.2 / 15.0

Table 3: BLEU ↑ scores for Whisper-S2TT (Direct), Whisper-ASR+MT (Cascade) and MT with gold transcriptions on the kosp2e and ProSem (without and with additional punctuation) test sets. Model sizes: tiny (T), base (B), small (S), medium (M) and large (L). CER ↓ for Whisper-ASR in brackets.

B General Performance

We present the SacreBLEU[1] (Post, 2018) score and the Character Error Rate (Morris et al., 2004, CER) of the systems to assess their general performance in the translation and transcription tasks respectively. In addition to the results on the ProSem testset, we provide general performance on the kosp2e (Cho et al., 2021) test set. As shown in Table 3, the results align with expectations, demonstrating that Whisper's performance improves with model size for both translation and recognition tasks on both test sets. The direct systems perform well on both test sets with BLEU scores up to 21.1 and 21.4 on the kosp2e and ProSem test sets respectively. As for the cascade systems, it is worth noting that the MT on gold transcription serves as an upper benchmark for the performance of the cascade systems. However, we can see that all the cascade systems achieve a higher BLEU score on ProSem compared to the base MT model. As discussed in Section 5, this is mainly due to the lack of punctuation in the transcription. By augmenting the model with question marks, we can see a drastic increase in BLEU score reaching 15.0, outperforming the cascade systems. Moreover, by comparing the CER scores on the two test sets, we observe that they are generally higher on the ProSem test set. This suggests that the utterances in the ProSem test set may be considered out-of-domain compared to more general test sets, contributing to the higher CER scores.

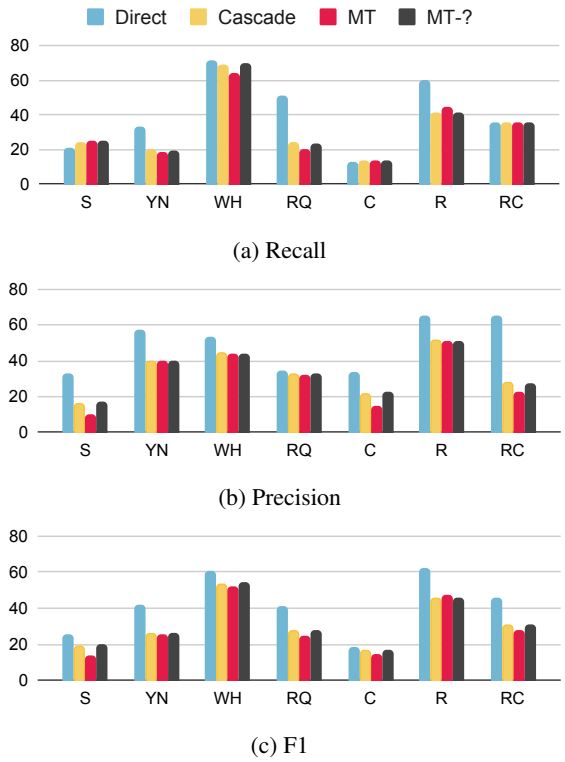


Figure 5: Contrastive evaluation recall, precision and F1 ↑ scores con *ambiguous* sets for direct and cascade Whisper "medium", and Machine Translation systems, for each intent type.

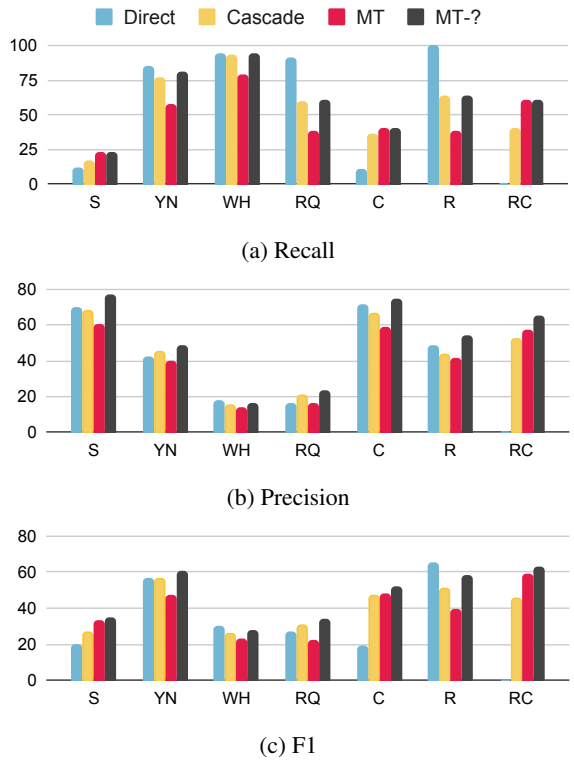


Figure 6: Contrastive evaluation recall, precision and F1 ↑ scores con *unambiguous* sets for direct and cascade Whisper "medium", and Machine Translation systems, for each intent type.

C Full Intent Disambiguation Results

Figure 5, 6 and 7 shows the recall, precision and f1 scores for the models on all the intent types (statements (S), yes/no questions (YN), wh-questions (WH), rhetorical questions (RQ), commands (C), requests (R), and rhetorical commands (RC)). In the context of ambiguous contrastive sets (Figure 5), the direct system consistently outperforms other models across all intent types, showcasing superior performance across all metrics. On unambiguous sets, the direct systems excel primarily in achieving high recall scores for questions (yes/no questions, *wh*-questions, rhetorical commands, and requests). However, for non-question intent types, the direct systems exhibit recall scores often below 12%, plummeting as low as 0% for rhetorical commands. This differentiation is reflected in the overall results (Figure 7), where the direct system surpasses text-based models in terms of F1 scores specifically for questions.

Figure 8 offers a closer look at the confusion matrices for the systems during the intent disambiguation task in contrastive evaluation. As detailed in Section 5, it's evident that all models display a notable bias toward the *wh*-question intent type, a tendency that is particularly pronounced in cascade and MT systems. Notably, the MT model, when not augmented with additional punctuation, exhibits a stronger inclination toward interpreting utterances as statements, especially evident in requests, where the incorrect selection of statements significantly decreases when punctuation is added (from 34% to 16%). Overall, the confusion matrices shed light on the challenges faced by text-based systems in effectively disambiguating intent, indicating a preference for interpreting utterances as one of the three major intent types.

D More on Punctuation

In Section 5, we categorised the ProSem test set into two distinct groups: ambiguous and unambiguous. The ambiguous category consists of contrastive sets where punctuation alone fails to definitively determine the correct intent. Figure 1 illustrates examples for both ambiguous and unambiguous contrastive sets. The contrastive set where the statement (Figure 1a) is the correct translation is

[1] nrefs:var|case:mixed|tok:13a|smooth:exp|version:1.5.1

	Direct	Cascade medium		MT		Random	Wh-Q
	medium	W/O	W	W/O	W		Random
Ambiguous	48.9	36.4	39.2	36.5	39.3	32.3	42.8
Unambiguous	33.6	34.7	36.0	34.6	40.8	41.3	28.6

Table 4: Contrastive evaluation accuracy ↑ scores on ambiguous and unambiguous contrastive sets for systems without (W/O) and with (W) question marks in the input, and pure and *wh*-question biased random selection.

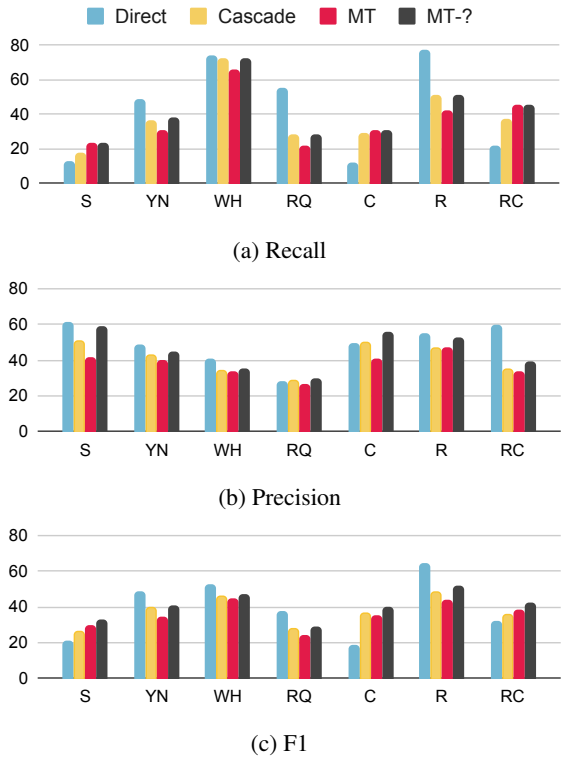


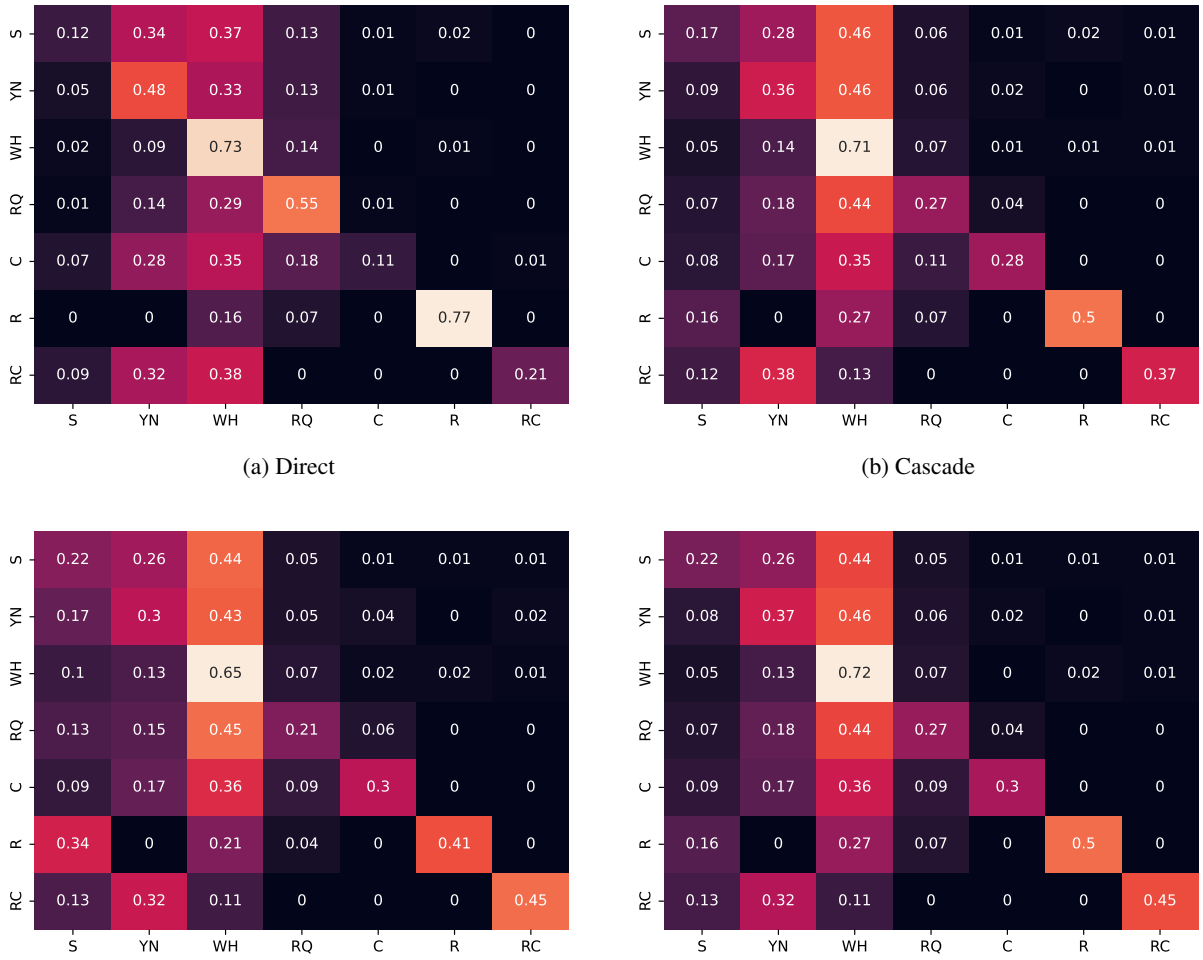
Figure 7: Overall contrastive evaluation recall, precision and F1 ↑ scores on ProSem for direct and cascade Whisper "medium", and Machine Translation systems, for each intent type.

an example of an unambiguous set because the absence of a question mark in the transcription "누가가입했대요" is sufficient to identify the correct intent as a "statement" as both yes/no and *wh*-questions contain question marks. On the other hand, in the scenario where the correct translation corresponds to the utterance with *wh*-question intent type (Figure 1c), the set becomes ambiguous, as ambiguity arises because both yes/no and *wh*-questions share the same transcription "누가가입했대요?".

Table 4 shows results for systems with and without question marks in the input. This was achieved either by removing question marks from the ASR transcription or by augmenting the gold transcription. Besides the accuracy for pure random selection ("Random"), we introduced an additional random baseline which is biased towards selecting the *wh*-question intent type (i.e., choosing a *wh*-question if it's an option, and selecting randomly otherwise) to simulate better the behaviour of the systems ("Wh-Q Bias"). We can observe that, as anticipated, none of the MT-based systems outperform Wh-Q Bias for ambiguous sets. This is expected since the input transcription does not provide sufficient information to disambiguate the correct intent. In contrast, the direct S2TT model significantly outperforms random in these cases, showcasing its effectiveness in handling ambiguity. On the other hand, for unambiguous examples, scores show a significant improvement when question marks are introduced or used in the MT input. Notably, the MT system with gold transcription outperforms the direct S2TT model in handling these examples. Overall, these results consistently align with our prior findings. Direct S2TT models hold an advantage over text-based systems because they can leverage prosodic information from the input signal to disambiguate ambiguous sentences. While punctuation aids in disambiguating questions from statements, it remains insufficient to resolve all instances of ambiguity.

E Vanilla Models

In this section, we report the results for smaller direct and cascade S2TT systems trained from scratch. To train our models, we used three distinct datasets: kosp2e (Cho et al., 2021), Korean Parallel corpora (Park et al., 2016) and ClovaCall (Ha et al., 2020). The kosp2e dataset was used to train all the systems as it contains speech signals, transcriptions and translation required to train direct S2TT, ASR and MT models. ClovaCall was used with kosp2e to train ASR systems, while the Korean Parallel corpora were used for MT systems as described in Section 4. Table 5 shows the statistics of the datasets used for training the systems. We used *fairseq S2T* (Wang et al., 2020) implementations for the S2TT and ASR models, with *"s2t transformer"* architectures and default training



(a) Direct

(b) Cascade

(c) MT

(d) MT-?

Figure 8: Normalised confusion matrices for Whisper "medium" direct and cascade, and Machine Translation (MT) systems with and without additional punctuation. Classes: statements (S), yes/no questions (YN), *wh*-questions (WH), rhetorical questions (RQ), commands (C), requests (R), and rhetorical commands (RC).

Dataset	Split	#	hs
ProSem	test	7104	7
kosp2e	train	106653	257
	dev	1266	2
	test	2320	4
ClovaCall	train	59662	50
Korean Parallel Corpora	train	125226	
	dev	1720	
S2TT	train	106652	257
	dev	1266	2
ASR	train	166315	307
	dev	1266	2
MT	train	231879	
	dev	2986	

Table 5: Datasets sizes in number of utterances/parallel sentences and recordings time in hours. Bottom half shows the data sizes used for training the direct S2TT, ASR and MT systems.

Model	Size	kosp2e	ProSem
Direct	S	2.0	0.7
	M	2.1	0.5
Direct+ASR init	S	9.1	1.6
	M	8.8	1.6
Cascade	S	0.2 (88.9)	0.1 (125.6)
	M	0.2 (88.6)	0.1 (127.4)
MT		**19.7**	9.5 / **11.4**

Table 6: BLEU ↑ scores and CER ↓ (in brackets) for direct and cascade Speech-to-Text Translation systems trained from scratch with architecture small (S) and medium (M), and MT models (without/with gold punctuation on the ProSem test set).

settings. In addition, we report results for a direct S2TT model with an ASR-initialised encoder. All results are the average of four different seeds.

E.1 Results

Results in Table 6 show the general performance of the direct and cascade systems trained from scratch. Compared to the results for whisper-based models in Section 5, the base direct and cascade systems could not provide satisfactory outputs on either test sets. However, despite the poor performance of the ASR models (CER > 88%), when used to initialise the direct S2TT models, they improved drastically the latter's performance, with an increase of 7.1 and 6.7 points in BLEU for the small and medium models respectively on the kosp2e test set. It's worth noting that the MT system, despite being trained on a notably smaller dataset compared to the OpusMT model, managed to achieve a high BLEU score on the kosp2e test set. This can be attributed to its training on in-domain data, underlining the impact of domain-specific training in enhancing performance.

Model	Accuracy
Random	36.3
MT	35.4
MT-?	39.4
Cascade	36.4
Direct	36.1
Direct+ASR init	**39.8**

Table 7: Contrastive evaluation accuracy ↑ scores on ProSem for Machine Translation (MT), cascade and direct Speech-to-Text Translation systems trained from scratch, as well random selection accuracy.

Table 7 shows the contrastive evaluation overall accuracies for non-Whisper translation systems on the ProSem test set. The cascade model was not able to perform better than random, achieving a similar score but a higher score to the base gold MT. The base direct S2TT system could not outperform the cascade model, as its performance was weak overall as previously shown. In contrast, the ASR-initialised direct S2TT system outperformed the other systems, achieving an accuracy increase of 3.4% over the cascade system. Although the overall accuracy remains modest, this observation lends credence to the hypothesis that direct S2TT systems effectively capture prosodic cues to disambiguate syntactically complex utterances.