

A EXAMPLES

Table 6 shows examples of predictions on the HWU corpus using both example-driven and observers. These examples show that semantically similar example utterances are identified, particularly when using observers. Furthermore, the examples in Table 6 show that explicitly reasoning over examples makes intent classification models more interpretable.

Utterance: It is too loud. Decrease the volume
Intent: audio-volume-down
Model: CONVBERT + MLM + <i>Example</i>
Predicted Intent: audio-volume-up
Nearest Examples:
Make sound louder (audio-volume-up)
Your volume is too high, please repeat that lower (audio-volume-down)
Too loud (audio-volume-down)
Can you speak a little louder (audio-volume-up)
Model: CONVBERT + MLM + <i>Example + Observers</i>
Predicted Intent: audio-volume-down
Nearest Examples:
It's really loud can you please turn the music down (audio-volume-down)
Up the volume the sound is too low (audio-volume-up)
Too loud (audio-volume-down)
Decrease the volume to ten (audio-volume-down)
Utterance: Please tell me about the historic facts about India
Intent: qa-factoid
Model: CONVBERT + MLM + <i>Example</i>
Predicted Intent: general-quirky
Nearest Examples:
How has your life been changed by me (general-quirky)
Is country better today or ten years ago? (general-quirky)
What happened to Charlie Chaplin? (general-quirky)
How does production and population affect us? (general-quirky)
Model: CONVBERT + MLM + <i>Example + Observers</i>
Predicted Intent: qa-factoid
Nearest Examples:
Tell me about Alexander the Great (qa-factoid)
Give me a geographic fact about Vilnius (qa-factoid)
Tell me about Donald Trump (qa-factoid)
I want to know more about the upcoming commonwealth games (qa-factoid)

Table 5: Examples of predictions on the HWU corpus with both example-driven and observers.

B ABLATIONS

We carry out ablations over the number of observers used to train and evaluate the models. Furthermore, we vary the number of examples seen at *inference time*, as a percentage of the set of training examples. The results shown in Table 6 demonstrate that while having more observers helps, even a single observer provides benefits. This suggests that the observed performance gain (shown in Table 1) is primarily a consequence of the disentangled attention rather than averaging over multiple observers.

The ablation over the number of examples used at inference time demonstrates that the models perform reasonably well with much fewer examples (e.g., 5% is <1000 examples or approximately

Setting	BANKING77	CLINC150	HWU64
OBSERVERS = 20; EXAMPLES = 100%	93.83	97.31	93.03
OBSERVERS = 10; EXAMPLES = 100%	93.60	97.62	92.01
OBSERVERS = 5; EXAMPLES = 100%	93.37	97.38	92.19
OBSERVERS = 1; EXAMPLES = 100%	93.83	97.33	92.57
OBSERVERS = 20; EXAMPLES = 50%	93.83	97.31	93.03
OBSERVERS = 20; EXAMPLES = 10%	92.86	97.24	92.38
OBSERVERS = 20; EXAMPLES = 5%	92.82	96.95	92.57
OBSERVERS = 20; EXAMPLES = 1%	80.40	68.37	73.79

Table 6: Ablation over the number of observers (during both training and testing) and the number of examples (only during testing) used for the CONVBERT + MLM + EXAMPLE-DRIVEN + OBSERVERS model. The percentage of examples refers to the proportion of the *training set* that is used as examples for the model at evaluation time.

5 per intent). The performance drop in the few-shot experiments suggests that it is important to train with more data, however the results in Table 6 demonstrate that it not necessarily important to have all of the examples at inference time.