Can Your Model Tell a Negation from an Implicature? Unravelling Challenges With Intent Encoders

Yuwei Zhang[1,2] **Siffi Singh**[1] **Sailik Sengupta**[1] **Igor Shalyminov**[1] **Hang Su**[1] **Hwanjun Song**[1,3] **Saab Mansour**[1]

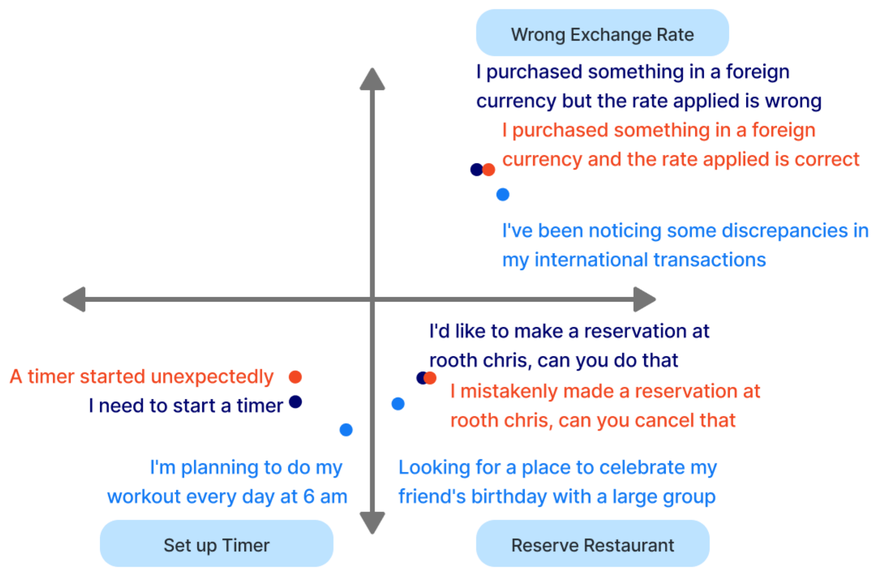
[1]aws AI Labs [2]University of California, San Diego [3]KAIST, Republic of Korea

Large embedding models don't understand semantics 😱

But, we can improve them! 📈

🧐 Flaws in Understanding

Intent classification and clustering tasks often require embedding models to understand subtle distinctions in semantics. We noticed that embedding models, such as `instructor-large`), embed negation utterances ($\mathbf{x}, \neg\mathbf{x}$) closer to, and implied utterances ($\mathbf{x} \rightarrow \mathbf{y}$) further away from the original utterance $\mathbf{x}$.



🛠️ Model Improvement

Given triplets u, u_i^p, u_i^n, we finetune the embedding models using a loss function similar to Su et. al. 2023:

$$l_i = \frac{\exp(s(f(u_i), f(u_i^p))/\gamma)}{\Sigma_{j\in\mathcal{B}} \exp(s(f(u_i), f(u_j^n))/\gamma)} + \frac{\exp(s(f(u_i^p), f(u_i))/\gamma)}{\Sigma_{j\in\mathcal{B}} \exp(s(f(u_i^p), f(u_j^n))/\gamma)}$$

To generate triplets, we first use $252,744$ unique and unlabelled utterances from various dialogue data-sets (see `Disable LLM` 👉). Then, we add more u^p and u^n utterances that are synthetically generated using less-capable LLMs (see `Our best` 👉).

Note that the data-sets and models used for synthetic data generation are different at train and test time.

🌡️ Benchmark - Intent Semantics Toolkit

Consider a tuple $\{u_i, u_i^p, u_i^n\}$ where u^p denotes implicature and u^n denotes negation.

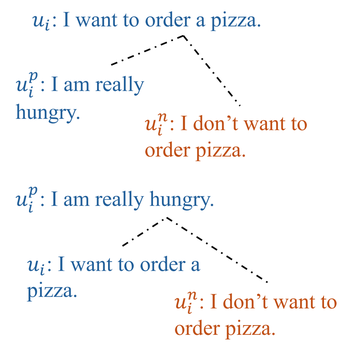
Triplet Task

$$\frac{1}{N_T} \Sigma_{i=1}^{N_T} \mathbb{I}(D(f(u_i), f(u_i^p)) < D(f(u_i), f(u_i^n)))$$

Binary Classification

$$\frac{1}{N_B} \Sigma_{i=1}^{N_B} \mathbb{I}(D(f(u_i), f(i)) < D(f(u_i), f(\neg i)))$$

where $\neg i$ denotes 'not intent i'.
Beyond these, we also consider clustering and multi-class (intent) classification tasks. The data generation workflow constitutes *existing data-sets → use of SoTA LLMs to generate phenomena-specific data-points → quality control with human evaluation.*



Model	Original				Intent Semantic Toolkit										
	Clustering		Multi-class		Triplet (Ori-Ori)		Triplet (Ori-Imp)		Binary Classification			Clustering		Multi-class	
	KM	Agg	0-shot	10-shot	T_{hard}	T_{easy}	T_{hard}	T_{easy}	Ori	Imp	Neg	KM	Agg	0-shot	10-shot
paraphrase	81.7	83.5	61.1	83.3	22.6	84.8	**3.9**	68.9	77.6	57.3	82.4	57.2	58.8	22.2	28.2
IAE	83.4	84.7	66.6	84.7	**24.0**	84.3	3.2	67.3	86.6	70.1	79.6	58.3	59.9	25.4	30.1
instructor-base	83.8	84.9	67.5	85.8	19.1	86.1	2.0	68.0	89.1	67.3	78.4	57.9	59.2	26.2	30.9
instructor-large	**84.3**	**86.0**	**67.6**	**86.2**	23.4	**87.5**	3.6	**71.0**	**89.6**	**73.5**	**87.4**	**59.1**	**61.4**	**28.8**	**34.3**

Our toolkit reveals a lack of understanding on negation and implicature utterances.

🔎 Results

Model	Original				Intent Semantic Toolkit										
	Clustering		Multi-class		Triplet (Ori-Ori)		Triplet (Ori-Imp)		Binary Classification			Clustering		Multi-class	
	KM	Agg	0-shot	10-shot	T_{hard}	T_{easy}	T_{hard}	T_{easy}	Ori	Imp	Neg	KM	Agg	0-shot	10-shot
Baseline	84.3	86.0	67.6	86.2	23.4	87.5	3.3	70.6	89.6	**73.9**	79.4	62.2	64.4	29.2	34.1
Disable LLM	84.1	85.0	71.7	85.5	39.3	86.2	8.0	50.8	89.2	53.0	**84.8**	65.5	66.9	31.7	34.3
Ours best	**84.6**	**86.8**	**73.4**	**87.2**	**51.1**	**93.7**	**20.4**	**77.6**	**94.0**	73.6	83.1	**65.9**	**68.2**	**33.9**	**37.2**

Synthetic data generation with a particular prompting strategy (labeled $P^4, N^{1,3}$ in the paper) & fine-tuning improves all the models (table shows results on `instructor-large`).

We also observe negative co-relation between the triplet task and the binary classification task that makes designing a single embedding model that improves on both tasks a challenging future work!

Scan me

amazon | science